446

ASTÉRISQUE

2023

SÉMINAIRE BOURBAKI
VOLUME 2022/2023
EXPOSÉS 1197–1210

SOCIÉTÉ MATHÉMATIQUE DE FRANCE
Publié avec le concours du CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Association des collaborateurs de Nicolas Bourbaki
Institut Henri Poincaré, 11 rue Pierre-et-Marie-Curie
75231 Paris Cedex 05 (France) https://www.bourbaki.fr

Mots clés et classification mathématique

Exposé nº 1197. — Modular forms, congruence subgroups, Fourier coefficients, unbounded denominators, algebraicity theorems, uniformisation of Riemann surfaces, Nevanlinna theory – 11F11, 11F30, 14G99, 20H05, 30D35, 30F35, 33C05.

Exposé nº 1198. — Théorie cinétique des gaz, équation de Boltzmann, limite de Boltzmann–Grad, cumulants, équation de Hamilton–Jacobi, grandes déviations, fluctuations – 82C22, 35Q20, 35Q70, 82B40.

Exposé nº 1199. — Pointwise ergodic theory, discrete harmonic analysis, Jean Bourgain – 37A46.

Exposé nº 1200. — Automorphisme de Frobenius asymptotique, théorie de l'intersection, automorphisme générique – 14G17, 14C17, 12L12.

Exposé nº 1201. — Conjectures homologiques, anneaux et espaces perfectoides, algèbres de Cohen–Macaulay – 13C14, 13H10, 13D22, 13D02, 14G45.

Exposé nº 1202. — Free probability, random permutations, random matrices, random graphs, expander graphs, strong asymptotic freeness, non-backtracking operators – 60B20, 15B52, 46L54, 05C48.

Exposé nº 1203. — Algèbres de von Neumann, problème de plongement de Connes, complexité algorithmique, corrélations quantiques, MIP*=RE – 46L06, 46L10, 68Q10, 81P05, 81P45.

Exposé nº 1204. — Topologie symplectique, homologie de Floer, groupes d'homéomorphismes, groupes de difféomorphismes – 37B55, 53D40, 37A05, 37C85, 57S25.

Exposé nº 1205. — Carleson problem, Schrödinger operator, multilinear harmonic analysis, Fourier restriction – 42B10, 42B37, 81U30.

Exposé nº 1206. — Hyperbolic groups, exponential growth rates, limit groups – 20F67, 20E08, 57K32, 20F69.

Exposé nº 1207. — Set theory, model companionship, continuum problem, forcing axioms, large cardinals – 03E35, 03E50, 03E57, 03C10, 00A30, 03A05.

Exposé nº 1208. — Équations de Navier–Stokes, solutions de Leray, non-unicité, instabilité spectrale – 35Q30, 76D03, 76D05, 35P15.

Exposé nº 1209. — Catégories tensorielles, caractéristique positive – 18M05, 18M25, 20C20.

Exposé nº 1210. — Percolation, universality, rotation invariance, star-triangle – 60K35.

SÉMINAIRE BOURBAKI

Volume 2022/2023
Exposés 1197–1210

doi : 10.24033/ast.1205

Résumé. — Ce 74e volume du Séminaire Bourbaki contient les textes des quatorze exposés présentés pendant l'année 2022/2023 : conjecture des dénominateurs non bornés, validité de la théorie cinétique des gaz, théorie ergodique ponctuelle, théorème de Lang–Weil tordu, conjecture du facteur direct, permutations aléatoires et graphes de Ramanujan, algèbres de von Neumann et corrélations quantiques, structure du groupe des homéomorphismes de la sphère de dimension 2, convergence ponctuelle pour l'équation de Schrödinger, croissance exponentielle dans les groupes hyperboliques, axiomes de forcing forts et hypothèse du continu, non-unicité des solutions de Leray de l'équation de Navier–Stokes, catégories tensorielles en caractéristique positive, invariance par rotation pour la percolation planaire.

Abstract. — This 74th volume of the Bourbaki Seminar gathers the texts of the fourteen lectures delivered during the year 2022/2023: unbounded denominators conjecture, validity of the kinetic theory of gases, pointwise ergodic theory, twisted Lang–Weil theorem, direct factor conjecture, random permutations and Ramanujan graphs, von Neumann algebras and quantum correlations, structure of the group of homeomorphisms of the 2-dimensional sphere, pointwise convergence for the Schrödinger equation, exponential growth in hyperbolic groups, strong forcing axioms and the continuum problem, non-uniqueness of Leray solutions to the Navier–Stokes system, tensor categories in positive characteristic, rotation invariance for planar percolation.

SOCIÉTÉ MATHÉMATIQUE DE FRANCE 2023

TABLE DES MATIÈRES

Résumé des exposés

Javier Fresán. — *The unbounded denominators conjecture* [*after F. Calegari, V. Dimitrov, and Y. Tang*]
Let f be a modular form for a finite index subgroup Γ of $\mathrm{SL}_2(\mathbf{Z})$ whose Fourier coefficients are algebraic numbers. It follows from the classical theory of modular forms that these coefficients have bounded denominators when Γ is a congruence subgroup. In the late 1960s, Atkin and Swinnerton-Dyer conjectured that, conversely, a form with bounded denominators is always modular for a congruence subgroup. I will explain a recent proof of this conjecture by Calegari, Dimitrov, and Tang. It relies on beautiful interactions between a new algebraicity theorem for power series, Nevanlinna theory for explicit uniformisations of the complex plane minus the roots of unity, and the fact that $\mathrm{SL}_2(\mathbf{Z}[1/p])$ has the congruence subgroup property.

François Golse. — *Validité de la théorie cinétique des gaz : au-delà de l'équation de Boltzmann* [*d'après T. Bodineau, I. Gallagher, L. Saint-Raymond et S. Simonella*]
L'obtention d'une justification rigoureuse de la théorie cinétique des gaz à partir du principe fondamental de la dynamique, dû à Newton, pour un grand nombre de sphères identiques interagissant par collisions binaires élastiques, est un problème formulé par Hilbert en 1900 (6^{e} problème). En 1975, Lanford a démontré la validité de l'équation de Boltzmann sur un intervalle de temps très court, de l'ordre d'une fraction du laps de temps moyen entre deux collisions successives subies par une même particule. Ce résultat de Lanford peut être interprété comme une sorte de « loi des grands nombres » lorsque le nombre de particules tend vers l'infini. Ce point de vue pose plusieurs questions.

D'abord, le cœur de l'argument utilisé par Boltzmann pour aboutir à l'équation portant son nom est l'hypothèse que deux particules sur le point d'entrer en collision sont presque indépendantes statistiquement. Ceci suggère d'examiner la validité de cette hypothèse en étudiant la dynamique des corrélations entre particules. D'autre part, l'interprétation de l'équation de Boltzmann comme loi des grands nombres conduit à étudier précisément les fluctuations de la mesure empirique dans l'espace des phases autour de sa moyenne (dont l'évolution est décrite par l'équation de Boltzmann). Une série d'articles récents de T. Bodineau, I. Gallagher, L. Saint-Raymond et S. Simonella répond à ces diverses questions et permet d'aller au-delà de l'équation de Boltzmann dans la compréhension de la théorie cinétique des gaz.

Ben Krause. — *Pointwise Ergodic Theory: Examples and Entropy* [*after Jean Bourgain*]
In this talk, we will explain how Bourgain combined elementary computations with a deep understanding of the entropic method to prove his pointwise ergodic theorem. The focus throughout will be on the intuition and heuristic which led him to his proof.

Silvain Rideau-Kikuchi. — *Sur un théorème de Lang–Weil tordu* [*d'après E. Hrushovski, K. V. Shuddhodan et Y. Varshavsky*]
Le théorème d'approximation de Lang–Weil donne une estimation du nombre de points dans une variété V (géométriquement intègre) sur un corps fini F : il y en a de l'ordre de $|F|^d$ où d est la dimension de la variété V. Puisque F est le corps fixé d'un automorphisme de Frobenius ϕ, cette question peut se reformuler comme celle d'estimer le nombre de points dans l'intersection de la diagonale de V^2 avec le graphe de ϕ. Dans cet exposé, nous considérerons une généralisation, due à Hrushovski, de cet énoncé à d'autres variétés que la diagonale et nous exposerons les ingrédients d'une preuve récente par Shuddhodan et Varshavsky.

Nous exposerons aussi certaines des nombreuses conséquences de cet énoncé en dynamique algébrique, ainsi qu'en théorie des modèles. L'une d'entre elle, particulièrement frappante, est que, de même qu'Ax avait pu, grâce aux estimations de Lang–Weil, donner une caractérisation de la « théorie des corps finis », ces estimations tordues permettent de caractériser la « théorie des automorphismes de Frobenius » et de montrer que c'est la théorie d'un automorphisme générique.

Gabriel Dospinescu. — *La conjecture du facteur direct* [*d'après Y. André et B. Bhatt*]
La conjecture du facteur direct de Hochster (énoncée dans les années 70) est un énoncé d'algèbre commutative de nature apparemment anodine : si B est une extension finie d'un anneau commutatif noethérien régulier A, alors A est un facteur direct de B en tant que A-module. Cette conjecture fait partie d'un faisceau de conjectures connues sous le nom de « conjectures homologiques », avec des implications frappantes en géométrie algébrique. Après la percée de Raymond C. Heitmann en 2002, qui a démontré la conjecture pour $\dim A \leqslant 3$, Yves André a démontré la conjecture du facteur direct en 2016. Peu de temps après Bhargav Bhatt a fourni une preuve plus simple. Les deux démonstrations utilisent de manière cruciale la théorie des espaces perfectoïdes de Peter Scholze, et le but de l'exposé est d'expliquer les principaux ingrédients de la preuve, ainsi que les raffinements obtenus ultérieurement par André et Bhatt.

Mylène Maïda. — *Strong convergence of the spectrum of random permutations and almost-Ramanujan graphs* [*after C. Bordenave and B. Collins*]
A finite graph is said to be *Ramanujan* if the spectral gap of its adjacency matrix is maximal, which makes it an excellent expander graph. From a family of random permutations, Bordenave and Collins construct a sequence of random graphs that are *almost-Ramanujan*. This property can in this case be reformulated in terms of *strong convergence* in free probability theory. The talk will be an opportunity to present known results on strong convergence and some of their applications. We will also insist on an important tool for their proof, the *non-backtracking* operator associated to the weighted adjacency operator of a graph. We will explain the link between the spectrum of the two operators and discuss the use of the *non-backtracking* operator in the study of random graphs.

Mikael de la Salle. — *Algèbres de von Neumann, produits tensoriels, corrélations quantiques et calculabilité* [*d'après Ji, Natarajan, Vidick, Wright et Yuen*]
En 1976, Connes demande si toute algèbre de von Neumann finie se plonge dans un ultraproduit d'algèbres de matrices. En 1980, Tsirelson demande si, dans la formulation mathématique de la mécanique quantique, autoriser des espaces de Hilbert de dimension infinie change fondamentalement le modèle. En 1993, Kirchberg conjecture que le produit tensoriel de deux copies de la C^*-algèbre pleine du groupe libre de rang infini dénombrable peut être muni d'une unique norme de C^*-algèbre. De manière surprenante et non triviale, ces trois problèmes sont en fait équivalents, c'est maintenant bien compris. Ces problèmes viennent d'être résolus, par la négative, avec des méthodes d'informatique : calculabilité, complexité, et informatique quantique. Je ferai de mon mieux pour raconter les grandes lignes de cette très longue preuve.

Étienne Ghys. — *Le groupe des homéomorphismes de la sphère de dimension* 2 *qui respectent l'aire et l'orientation n'est pas un groupe simple* [*d'après D. Cristofaro-Gardiner, V. Humilière et S. Seyfaddini*]
Depuis la fin des années 1970, on sait que la composante neutre du groupe des difféomorphismes à support compact d'une variété connexe est un groupe simple. Dans le cas des difféomorphismes qui préservent une forme volume ou une forme symplectique, on dispose d'un résultat analogue : il a alors un sous-groupe distingué « évident » qui est simple. Pour les homéomorphismes qui respectent le volume, la situation est comprise lorsque la dimension est supérieure ou égale à 3. Le cas des surfaces, et tout particulièrement de la sphère de dimension 2, a résisté à de nombreux efforts depuis une quarantaine d'années. Le théorème de D. Cristofaro-Gardiner, V. Humilière et S. Seyfaddini est une surprise : le groupe des homéomorphismes de la sphère de dimension 2 qui respectent l'aire et l'orientation n'est *pas* un groupe simple. La démonstration est un tour de force et fait largement usage de l'homologie de Floer périodique. J'essaierai de présenter le contexte ainsi que les grandes lignes de ce beau résultat.

Jonathan Hickman. — *Pointwise convergence for the Schrödinger equation* [*after Xiumin Du and Ruixiang Zhang*]
For an initial datum $f \in L^2(\mathbf{R}^n)$, consider the linear Schrödinger equation

$$\begin{cases} iu_t - \Delta_x u = 0, \\ u(x,0) = f(x) \end{cases} \qquad (x,t) \in \mathbf{R}^n \times \mathbf{R}.$$

In 1980, Carleson asked which additional conditions on f guarantee

$$\lim_{t\to 0} u(x,t) = f(x) \qquad \text{for almost every } x \in \mathbf{R}^n. \tag{$\star$}$$

More precisely, what is the minimal Sobolev regularity index s such that $(\star)$ holds whenever $f \in H^s(\mathbf{R}^n)$?

Whilst the $n = 1$ case was fully understood by the early 1980s, in higher dimensions the situation is much more nuanced. Nevertheless, a recent series of dramatic developments brought about an almost complete resolution of the problem. First Bourgain (2016) produced a subtle counterexample demonstrating that pointwise convergence can fail for certain $f \in H^s(\mathbf{R}^n)$ with $s < \frac{n}{2(n+1)}$. Complementing this, convergence was then shown to hold for $s > \frac{n}{2(n+1)}$ when $n = 2$ in a landmark paper of Du, Guth and Li (2017) and later in all dimensions in equally important work of Du and Zhang (2019).

This seminar will explore the positive result of Du and Zhang (2019). The argument combines sophisticated modern machinery from harmonic analysis such as the multilinear Strichartz estimates of Bennett, Carbery and Tao (2006) and the ℓ^2 decoupling theory of Bourgain and Demeter (2015). However, equally important are a variety of elementary guiding principles, rooted in Fourier analysis, which govern the behaviour of solutions to the Schrödinger equation. The talk will focus on these basic Fourier analytic principles, building intuition and presenting a powerful toolbox for tackling problems in modern PDE and harmonic analysis.

Clara Löh. — *Exponential growth rates in hyperbolic groups* [*after Koji Fujiwara and Zlil Sela*]
A classical result of Jørgensen and Thurston shows that the set of volumes of finite volume complete hyperbolic 3-manifolds is a well-ordered subset of the real numbers of order type ω^ω; moreover, they showed that each volume can only be attained by finitely many isometry types of hyperbolic 3-manifolds.

Fujiwara and Sela established a group-theoretic companion of this result: If Γ is a non-elementary hyperbolic group, then the set of exponential growth rates of Γ is well-ordered, the order type is at least ω^ω, and each growth rate can only be attained by finitely many finite generating sets (up to automorphisms).

In this talk, I will outline this work of Fujiwara and Sela and discuss related results.

Matteo Viale. — *Strong forcing axioms and the continuum problem* [*after Asperó's and Schindler's proof that* $\mathbf{MM}^{++}$ *implies Woodin's Axiom* $(*)$]
A topological approach to forcing axioms considers them as strong forms of the Baire category theorem; an algebraic approach describes certain properties of "algebraic closure" for the universe of sets that can be derived from them. Our goal is to show how the theorem of Aspéro and Schindler links the geometric and algebraic points of view. Drawing on Gödel's program, we connect these mathematical results to the philosophical debate on what could constitute a viable solution of the continuum problem.

Anne-Laure Dalibard. — *Non-unicité des solutions du système de Navier–Stokes avec terme source* [*d'après Dallas Albritton, Elia Brué et Maria Colombo*]
La dynamique des fluides visqueux incompressibles est décrite par les équations de Navier–Stokes, pour lesquelles on dispose principalement de deux façons de construire des solutions en dimension trois. La première, due à Leray et étendue par Hopf, repose sur une méthode de compacité, et conduit à l'existence de solutions dites « faibles », globales (c'est-à-dire définies pour tout temps). La seconde, due initialement à Fujita et Kato et généralisée ensuite, consiste à construire des solutions dites « fortes » par une méthode de point fixe, dans un espace fonctionnel à forte régularité. Les solutions fortes ainsi obtenues sont naturellement uniques, mais sont a priori locales. Cette dichotomie conduit naturellement à la question suivante, restée ouverte pendant presque un siècle : les solutions de Leray–Hopf sont-elles uniques ?

Récemment, Dallas Albritton, Elia Brué et Maria Colombo ont apporté une réponse négative à cette question fondamentale, en considérant le cas d'un fluide initialement au repos et soumis à une force extérieure. Leur preuve repose sur la construction d'un profil linéairement instable dans des variables auto-similaires et s'inspire d'un résultat de Vishik pour l'équation d'Euler, ainsi que des travaux de Sverak et de ses collaborateurs.

Daniel Juteau. — *Catégories tensorielles symétriques en caractéristique positive* [*d'après Kevin Coulembier, Pavel Etingof, Victor Ostrik...*]
Le formalisme tannakien a d'abord été développé par l'école de Grothendieck pour les besoins de la théorie des motifs. L'idée principale est que se donner un groupe (algébrique affine sur un corps k, disons algébriquement clos) est essentiellement équivalent à se donner sa catégorie de représentations en tant que catégorie monoïdale symétrique, munie du foncteur d'oubli (dit foncteur fibre) vers la catégorie des espaces vectoriels : une catégorie pré-tannakienne (monoïdale symétrique, et vérifiant des conditions nécessaires naturelles) admettant un « foncteur fibre » est forcément équivalente à la catégorie des représentations du groupe des automorphismes tensoriels du foncteur fibre.

Dans le cas de la caractéristique 0, Deligne a montré en 1990 qu'une catégorie pré-tannakienne $\mathscr{C}$ admet un foncteur fibre (*i.e.* est tannakienne) si et seulement si tout objet a une puissance alternée qui est nulle. En 2002, il a montré un résultat plus général : si on suppose seulement que $\mathscr{C}$ est à croissance modérée (pour tout objet V, la longueur de $V^{\otimes n}$ est sous-exponentielle), alors $\mathscr{C}$ a une sorte de foncteur fibre, non pas vers les espaces vectoriels a priori, mais vers les super espaces vectoriels (espaces vectoriels $\mathbb{Z}/2\mathbb{Z}$-gradués).

L'extension de ces résultats au cas où k est de caractéristique $p > 0$ a été un problème ouvert pendant une vingtaine d'années, mais de grands progrès ont été faits récemment. En particulier, Ostrik a identifié une catégorie de Verlinde Ver_p comme but naturel des « foncteurs fibres » en caractéristique p. Plus récemment, Coulembier,

Etingof et Ostrik ont donné une certaine réponse à notre question : ils ont caractérisé les catégories pré-tannakiennes admettant un foncteur tensoriel symétrique vers Ver_p comme celles qui sont Frobenius-exactes et de croissance modérée (cette dernière condition pouvant être remplacée par : tout objet est annulé par une puissance alternée). Un cas particulier, qui est aussi une étape importante dans la preuve, est une caractérisation des catégories tannakiennes en caractéristique p. Nous donnerons un aperçu de ces résultats, ainsi que des exemples d'applications aux représentations modulaires.

Vincent Tassion. — *Rotation invariance for planar percolation* [*after Hugo Duminil-Copin, Karol Kajetan Kozlowski, Dmitry Krachun, Ioan Manolescu, and Mendes Oulamara*]
We consider critical percolation on the square lattice $\mathbb{Z}^2$, seen as a graph: For each edge, we flip a coin, the edge is kept with probability $1/2$, otherwise it is deleted, independently of the other edges. This gives rise to a random subgraph of $\mathbb{Z}^2$. The law of this random subgraph is invariant under $\pi/2$-rotation, because it inherits the symmetries of the lattice. But if we consider the large connected components, new symmetries emerge: Hugo Duminil-Copin, Karol Kajetan Kozlowski, Dmitry Krachun, Ioan Manolescu, and Mendes Oulamara have shown that the law of these components is asymptotically invariant under all rotations. This result constitutes a major advance towards the understanding of critical phenomena in planar statistical mechanics: the main conjecture in the field is that the law of large connected components is in fact invariant under conformal transformations, and it satisfies a universality principle: this asymptotic law does not depend on the underlying lattice. In this talk we will give a rigorous meaning to these statements, and then discuss some essential aspects: what role does the parameter $1/2$ play? What heuristic reasons justify the emergence of symmetries? What are the main ideas for invariance by rotation?

Séminaire BOURBAKI Novembre 2022
74e année, 2022–2023, no 1197, p. 1 à 27
doi: 10.24033/ast.1206

THE UNBOUNDED DENOMINATORS CONJECTURE
[*after* F. Calegari, V. Dimitrov, and Y. Tang]

by **Javier Fresán**

Introduction

This written account of my talk at the Bourbaki seminar surveys some of the ideas in the beautiful proof by Calegari, Dimitrov, and Tang (2021) of the unbounded denominators conjecture, a long standing open problem in the theory of modular forms that gives a simple criterion to decide whether a modular form with algebraic Fourier coefficients at infinity is "invariant" under a congruence subgroup of $\mathrm{SL}_2(\mathbf{Z})$ or not.

Throughout, we write $\overline{\mathbf{Q}}$ for the algebraic closure of $\mathbf{Q}$ in $\mathbf{C}$ and $\overline{\mathbf{Z}} \subset \overline{\mathbf{Q}}$ for the subring of algebraic integers. We let $\mathfrak{H} = \{\tau \in \mathbf{C} \mid \mathrm{Im}(\tau) > 0\}$ denote the upper half-plane and[(1)] $q = \exp(\pi i \tau)$. Recall that $\mathrm{SL}_2(\mathbf{Z})$ acts on $\mathfrak{H}^* = \mathfrak{H} \cup \mathbb{P}^1(\mathbf{Q})$ by Möbius transformations and that *congruence subgroups* of $\mathrm{SL}_2(\mathbf{Z})$ are those containing

$$\Gamma(M) = \ker\big(\mathrm{SL}_2(\mathbf{Z}) \to \mathrm{SL}_2(\mathbf{Z}/M\mathbf{Z})\big) = \{A \in \mathrm{SL}_2(\mathbf{Z}) \mid A \equiv I \bmod M\}$$

for some integer $M \geqslant 1$. The *unbounded denominators conjecture*, originating from work of Atkin and Swinnerton-Dyer (1971), is now the following theorem:

Theorem 0.1 (Calegari–Dimitrov–Tang, 2021). *Let $f(\tau)$ be a holomorphic function on the upper-half plane $\mathfrak{H}$ such that*

(*a*) *there exists an integer k and a subgroup $\Gamma \subset \mathrm{SL}_2(\mathbf{Z})$ of finite index such that*

$$f\left(\frac{a\tau + b}{c\tau + d}\right) = (c\tau + d)^k f(\tau) \tag{1}$$

holds for all matrices $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ in Γ;

(*b*) *f locally extends to a meromorphic function around each point of $\mathbb{P}^1(\mathbf{Q})$;*

(*c*) *f admits a Fourier expansion in $\overline{\mathbf{Z}}[\![q^{1/N}]\!]$ for some integer $N \geqslant 1$.*

Then the equality (1) *holds for all matrices in a congruence subgroup of* $\mathrm{SL}_2(\mathbf{Z})$.

(1) One reason for choosing this unusual convention for q will be explained in remark 0.2 below.

In what follows, we will refer to functions f satisfying assumptions (a) and (b) simply as *modular forms* of weight k, or *modular functions* if $k = 0$, for the group Γ. Note that every subgroup of finite index of $\mathrm{SL}_2(\mathbf{Z})$ contains a matrix of the shape $\left(\begin{smallmatrix} 1 & 2N \\ 0 & 1 \end{smallmatrix}\right)$ for some integer $N \geqslant 1$. More precisely, the *width* of each cusp $\zeta \in \mathbb{P}^1(\mathbf{Q})$ is defined as the smallest integer $m_\zeta \geqslant 1$ such that the stabiliser of ζ under the action of Γ on $\mathbb{P}^1(\mathbf{Q})$ contains, up to conjugation in $\mathrm{SL}_2(\mathbf{Z})$, one of the matrices $\pm \left(\begin{smallmatrix} 1 & m_\gamma \\ 0 & 1 \end{smallmatrix}\right)$. The assumption that f has a Fourier expansion in $\overline{\mathbf{Z}}[\![q^{1/N}]\!]$ implies that the width of the cusp at infinity divides $2N$. For $k = 0$, in the conclusion of the theorem we can take a congruence subgroup containing $\Gamma(L(\Gamma))$, where $L(\Gamma)$ stands for the lowest commun multiple of the widths of all cusps, a generalisation of the notion of level for non-congruence subgroups.

Let us explain the name of the conjecture. If the coefficients of $f \in \overline{\mathbf{Q}}[\![q^{1/N}]\!]$ have *bounded denominators*, which amounts to saying that f lies in the subspace

$$\overline{\mathbf{Z}}[\![q^{1/N}]\!] \otimes_{\overline{\mathbf{Z}}} \overline{\mathbf{Q}} = \mathbf{Z}[\![q^{1/N}]\!] \otimes_{\mathbf{Z}} \overline{\mathbf{Q}} \subset \overline{\mathbf{Q}}[\![q^{1/N}]\!],$$

then we can apply theorem 0.1 to an integral multiple of f. The contrapositive statement then says the following:

> *Let $f(\tau)$ be a modular form for a subgroup of finite index of $\mathrm{SL}_2(\mathbf{Z})$ with a Fourier expansion in $\overline{\mathbf{Q}}[\![q^{1/N}]\!]$. If f is not modular for any congruence subgroup, then the Fourier coefficients of f have unbounded denominators.*

By contrast, all modular forms f for congruence subgroups have bounded denominators by the theory of Hecke operators (Shimura, 1971, Theorem 3.52). In a nutshell, after multiplying f by a large enough power of the modular discriminant to turn it into a holomorphic cusp form, we can write it as a linear combination of Hecke eigenforms, and the Fourier coefficients of those are algebraic integers since they are polynomial expressions with integer coefficients in the Hecke eigenvalues[(2)]. Thus, the condition of having bounded denominators completely distinguishes congruence and non-congruence modular forms among all modular forms with algebraic Fourier coefficients at infinity.

By a theorem of Mennicke (1965) and Bass, Lazard, and Serre (1964), the group $\mathrm{SL}_n(\mathbf{Z})$ has the *congruence subgroup property* for each $n \geqslant 3$, meaning that all its subgroups of finite index contain a congruence subgroup. The same is true for other arithmetic groups such as $\mathrm{SL}_2(\mathbf{Z}[1/p])$ for each prime number p. Most subgroups of finite index of $\mathrm{SL}_2(\mathbf{Z})$, however, are *not* congruence. For example, given an integer $g \geqslant 0$, there is only a finite number of congruence subgroups Γ such that the

[(2)] This argument fails for non-congruence modular forms. Although there is still a way to define Hecke operators, their action is trivial on those forms that do not come from the smallest congruence subgroup containing Γ by results of Serre, presented in Thompson (1989), and Berger (1994).

curve $X(\Gamma) = \mathfrak{H}^*/\Gamma$ has genus g (Dennin, 1975), whereas there is an infinite number of non-congruence subgroups with the same property (Jones, 1979). Some explicit examples of non-congruence subgroups will be given in section 0.2 below.

One reason to care about modular forms for non-congruence subgroups is *Belyi's theorem*, according to which every smooth projective curve defined over $\overline{\mathbf{Q}}$ can be realised as a cover of the projective line $\mathbb{P}^1$ that is only ramified at $0, 1, \infty$ (such coverings are often called *Belyi maps*). Taking the isomorphism $\mathfrak{H}/\Gamma(2) \simeq \mathbb{P}^1 \setminus \{0, 1, \infty\}$ given by the modular lambda function into account, any such curve is hence isomorphic to $X(\Gamma)$ for a subgroup $\Gamma \subset \Gamma(2)$ of finite index. As we will see below, theorem 0.1 provides us with a criterion to decide whether Γ is a congruence subgroup or not in terms of the integrality properties of the associated Belyi map.

0.1. First reductions

It will be enough to prove the theorem under the assumption that f is a modular *function* with *integer* coefficients. We first explain the reduction to the case $k = 0$. For this, consider the q-series expansions

$$\begin{aligned} \frac{\lambda(\tau)}{16} &= q\prod_{n=1}^{\infty}\left(\frac{1+q^{2n}}{1+q^{2n-1}}\right)^8 = q - 8q^2 + 44q^3 - \cdots, \\ \eta\left(\tfrac{\tau}{2}\right)^2 &= q^{1/12}\prod_{n=1}^{\infty}(1-q^n)^2 = q^{1/12} - 2q^{13/12} - q^{25/12} + \cdots, \end{aligned} \tag{2}$$

which define a modular function for the group $\Gamma(2)$ and a modular form of weight 1 for $\Gamma(12)$ respectively. The first one induces an isomorphism[(3)]

$$\mathfrak{H}/\Gamma(2) \xrightarrow{\sim} \mathbb{P}^1 \setminus \{0, 1/16, \infty\}$$

that extends to a map sending the cusp at infinity to 0. The second one, a 12th root of the modular discriminant $\Delta(\tau/2)$, does not vanish on the upper half-plane and has the property that its inverse has integer Fourier coefficients at infinity. Therefore,

$$F(\tau) = \left(\frac{\lambda(\tau)}{16}\right)^k \frac{f(\tau)}{\eta(\frac{\tau}{2})^{2k}} \in \overline{\mathbf{Z}}[\![q^{1/N}]\!]$$

satisfies the assumptions of theorem 0.1 with the weight $k = 0$ and the subgroup $\Gamma \cap \Gamma(12)$ of $\mathrm{SL}_2(\mathbf{Z})$. If F is a modular function for a congruence subgroup, then f is a modular form for a congruence subgroup. Note that the first factor is there to kill the pole at $q = 0$ introduced by η, thus keeping the condition that $f(\tau)$ is holomorphic at infinity. This operation could, however, create new poles at other cusps; this explains the lack of symmetry between infinity and the other cusps in the statement of the theorem.

[(3)] One says that λ is a *Hauptmodul* for $\Gamma(2)$.

Remark 0.2. One explanation for the normalisations $x = \lambda/16$ and $q = \exp(\pi i \tau)$ is that they allow for the identity $\mathbf{Z}[\![q]\!] = \mathbf{Z}[\![x]\!]$ coming from the expressions

$$x = q - 8q^2 + 44q^3 + \cdots, \quad q = x + 8x^2 + 84x^3 + \cdots$$

of x and q as power series with *integer* coefficients in q and x respectively.

Let us now explain how to reduce to the case $f \in \mathbf{Z}[\![q^{1/N}]\!]$ following a suggestion of John Voight (Calegari, Dimitrov, and Tang, 2021, Remark 6.3.2). Let Γ be a finite index subgroup of $\mathrm{SL}_2(\mathbf{Z})$. By Belyi's theorem, the curve $X(\Gamma)$, its cusp at infinity, the uniformiser $q^{1/N}$, and the covering $X(\Gamma) \to \mathbb{P}^1$ are defined over some number field K. Moreover, the algebro-geometric interpretation of modular functions as rational functions on the curve $X(\Gamma)$ shows that they carry a natural structure of a K-vector space, corresponding to those modular functions whose q-expansion at infinity has coefficients in K. After enlarging K to its Galois closure if necessary, an element σ of the Galois group $\mathrm{Gal}(K/\mathbf{Q})$ transforms the covering $X(\Gamma) \to \mathbb{P}^1$ into a covering $X(\Gamma_\sigma) \to \mathbb{P}^1$ for possibly another subgroup Γ_σ of finite index, that we may conjugate so that the cusp at infinity maps again to ∞. Since the Galois action on q-expansions is given by applying σ to the coefficients, the conjugate of a modular function will still be modular for a subgroup of finite index. Now, the modularity assumption on $f \in \overline{\mathbf{Z}}[\![q^{1/N}]\!]$ implies that there exists a number field L, with ring of integers $\mathscr{O}_L$, such that f lies in $\mathscr{O}_L[\![q^{1/N}]\!]$. If $\alpha_1, \ldots, \alpha_d$ is a $\mathbf{Z}$-basis of $\mathscr{O}_L$, then $f_i(\tau) = \mathrm{Tr}_{L/\mathbf{Q}}(\alpha_i f(\tau))$ lies in $\mathbf{Z}[\![q^{1/N}]\!]$ and is still modular for a finite index subgroup of $\mathrm{SL}_2(\mathbf{Z})$ by the above. By the special case of theorem 0.1 in which the function is assumed to have integer coefficients, each of these functions is modular for a congruence subgroup Γ_i, so f is modular for $\Gamma_1 \cap \cdots \cap \Gamma_d$.

To summarise, we are reduced to proving the following statement:

Theorem 0.3. *Let $N \geqslant 1$ be an integer and let $f(\tau) \in \mathbf{Z}[\![q^{1/N}]\!]$ be a holomorphic function on $\mathfrak{H}$ that locally extends to a meromorphic function around each point of $\mathbb{P}^1(\mathbf{Q})$ and is invariant under the action of a subgroup $\Gamma \subset \Gamma(2)$ of finite index. Then f is a modular function for a congruence subgroup.*

0.2. An interpretation in terms of Belyi maps

In the notation of theorem 0.3, let $Y(\Gamma) = \mathfrak{H}/\Gamma$ and consider the diagram

$$\begin{array}{ccc} Y(\Gamma) & \xrightarrow{\quad f \quad} & \mathbf{C} \\ \Big\downarrow{\scriptstyle \pi} & \nearrow & \\ Y(2) \simeq \mathbb{P}^1 \setminus \{0, 1/16, \infty\} & & \end{array}$$

where π is an étale cover and $Y(2)$ and $\mathbb{P}^1 \setminus \{0, 1/16, \infty\}$ are identified through the isomorphism $\lambda/16$. We can then think of f as a multivalued *algebraic* function of the variable

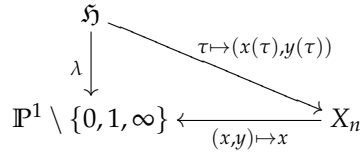
$\lambda/16$ ramified at the points $0, 1/16, \infty$, as indicated by the dotted arrow. By expanding it as a Puiseux series at a branch above 0, the theorem can be rephrased as saying that

f lies in $\mathbf{Z}[\![\frac{\lambda(\tau/m)}{16}]\!] \otimes \mathbf{C}$ for some $m \geqslant 1$ if and only if Γ is a congruence subgroup.

Example 0.4 (Fermat curves). Let $n \geqslant 1$ be an integer and consider the Fermat curve X_n with affine equation $x^n + y^n = 1$. Since the modular lambda function λ does not take the values 0 and 1, there exist holomorphic functions $x, y \colon \mathfrak{H} \to \mathbf{C}$ satisfying

$$x(\tau)^n = \lambda(\tau) \quad \text{and} \quad y(\tau)^n = 1 - \lambda(\tau).$$

The diagonal arrow in the diagram



factors through an isomorphism $\mathfrak{H}/\Phi(n) \simeq X_n$, where the *Fermat group* $\Phi(n)$ is defined as the kernel of the composition $\Gamma(2) \to \Gamma(2)^{\mathrm{ab}} \to \Gamma(2)^{\mathrm{ab}}/n$. Explicitly, $\Phi(n)$ is generated by the n-th powers of the matrices $A = \left(\begin{smallmatrix} 1 & 2 \\ 0 & 1 \end{smallmatrix}\right)$ and $B = \left(\begin{smallmatrix} 1 & 0 \\ 2 & 1 \end{smallmatrix}\right)$, and by the commutator $[\Delta, \Delta]$ of the subgroup $\Delta = \langle A, B \rangle = \Gamma(2)/\{\pm I\}$ that they generate. It is a classical result of Klein and Fricke (2017, page 534) that $\Phi(n)$ is a congruence subgroup if and only if $n \in \{1, 2, 4, 8\}$. This property is reflected by the fact that the modular functions $x(\tau)$ and $y(\tau)$ have unbounded denominators unless n takes one of those values, or yet by the fact that the coefficients of the power series[(4)]

$$\sqrt[n]{1-x} = \sum_{m=0}^{\infty} \frac{16^m \left(\frac{-1}{n}\right)_m}{m!} \left(\frac{x}{16}\right)^m \in \mathbf{Q}\left[\!\!\left[\frac{x}{16}\right]\!\!\right]$$

have bounded denominators if and only if $n \in \{1, 2, 4, 8\}$, in which case they are all integers. Indeed, writing the m-th coefficient as

$$a_m = (-16)^m \frac{(n-1)(2n-1)\cdots[(m-1)n+1]}{n^m m!},$$

we see that, for each odd prime number p dividing n, the p-adic valuation $v_p(a_m)$ is smaller than $-v_p(m!)$, which tends to $-\infty$ as $m \to +\infty$. If 2 divides n, then $v_2(a_m)$ is equal to $4m - mv_2(n) - v_2(m!)$, so that again it tends to $-\infty$ as soon as $v_2(n) \geqslant 4$ but is non-negative for $n \in \{2, 4, 8\}$ since $v_2(m!) = \sum_{k=1}^{\infty} \lfloor m/2^k \rfloor \leqslant m$. Finally, $v_p(a_m) \geqslant 0$ for all primes p not dividing n, as can be seen by choosing $r \geqslant v_p(m!)$ and replacing the 1s in the numerator of a_m with $1 = un + vp^r$ for some integers u, v.

(4) Here, $(\alpha)_m = \alpha(\alpha+1)\cdots(\alpha+m-1)$ denotes the Pochhammer symbol of a complex number α.

1. Algebraicity theorems

A key ingredient in the proof by Calegari, Dimitrov, and Tang of the unbounded denominators conjecture is a generalisation of an algebraicity theorem for power series with integer coefficients due to André (2004), which the authors call the *arithmetic holonomicity theorem* (theorem 1.6). Before stating it and giving a sketch of one of its many proofs, we briefly overview the history of this kind of results and glimpse at their applications by Harbater (1988), Ihara (1994), Bost (1999), and others to the study of fundamental groups of arithmetic surfaces.

1.1. A few rationality theorems

A toy example of the statements that will be considered is the remark that, if the radius of convergence R of a power series $f \in \mathbf{Z}[\![x]\!]$ is strictly larger than 1, then f is a polynomial. Indeed, write $f(x) = \sum_{n=0}^{\infty} a_n x^n$ and choose $1 < \eta < R$ and $C \geqslant 0$ such that $|f(x)| \leqslant C$ on the disc of radius η. Using the Cauchy residue formula

$$a_n = \frac{1}{2\pi i} \int_{|x|=\eta} \frac{f(x)}{x^{n+1}} dx,$$

we find the estimate $|a_n| \leqslant C/\eta^n$ for all $n \geqslant 0$. Since $\eta > 1$, the right-hand side has limit 0 as n tends to infinity, and this implies $a_n = 0$ for large enough n because a non-zero integer has absolute value at least 1. All subsequent proofs of rationality or algebraicity theorems will follow, in a more or less sophisticated manner, this path of creating a tension between two estimates coming from an integral representation (*Cauchy-like bound*) and from the arithmetic nature of the coefficients (*Liouville-like bound*). A first generalisation of this toy example is a celebrated theorem by Émile Borel (1894), in which f is only assumed to be meromorphic on the disc $D(0, R)$.

Theorem 1.1 (Borel, 1894). *If a power series $f \in \mathbf{Z}[\![x]\!]$ can be written as a quotient of convergent power series with complex coefficients on a disc of radius $R > 1$, then f represents a rational function.*

The proof relies on a characterisation of rational functions in terms of the vanishing of a Hankel determinant $\det(a_{n+i+j})_{0 \leqslant i,j \leqslant N}$ for all large enough n.

So far, we have only taken *archimedean* information into account. Working at all places allows one to relax the integrality assumption on the coefficients while still getting rationality. For this, we consider the p-adic absolute value $|\cdot|_p$, normalised as $|p|_p = 1/p$ so that the product formula holds.

Theorem 1.2 (Dwork, 1960). *A power series $f = \sum_{n=0}^{\infty} a_n x^n \in \mathbf{Q}[\![x]\!]$ represents a rational function if and only if*

(*a*) *there exists a finite set S of prime numbers such that a_n lies in $\mathbf{Z}[1/S]$ for all n;*

(*b*) *there exist real numbers R_∞ and $(R_p)_{p\ prime}$ satisfying $R_\infty \prod_p R_p > 1$ such that f is a quotient of convergent power series with complex coefficients on the closed disc of radius R_∞ and a quotient of convergent power series with $\mathbf{C}_p$-coefficients on the p-adic closed disc of radius R_p.*

This theorem generalises Borel's, since all p-adic radii R_p can be taken equal to 1 when the coefficients a_n are integers. Also known as the *Borel–Dwork criterion*, it was first proved in (Dwork, 1960, Theorem 3), where it was famously exploited to establish the rationality of the zeta function of an algebraic variety over a finite field. A further generalisation by Pólya and Bertrandias allows one to consider domains of meromorphy more general than the disc, with the radius replaced by the transfinite diameter.

Note that both condition (a) and the strict inequality in condition (b) are necessary, as witnessed by the following examples borrowed from Harbater (1988, page 856):

- the series $\sum_{n=0}^{\infty}(2^n/q_n)x^n$, where q_n is the smallest prime number bigger than 2^n, is not a rational function, despite the equality $R_\infty \prod_p R_p = 2$;
- the series with integer coefficients $\sum_{n=0}^{\infty} x^{n!}$ is not a rational function (here R_∞ and all R_p are equal to 1).

Harbater observed that for *algebraic* power series $f \in \mathbf{Q}[\![x]\!]$, condition (b) readily implies rationality thanks to Eisenstein's theorem on the growth of the denominators of their coefficients. Under this assumption, he could then weaken the inequality.

Theorem 1.3 (Harbater, 1988). *Let $f \in \mathbf{Q}[\![x]\!]$ be an algebraic power series. If conditions (a) and (b) with $R_\infty \prod_p R_p \geqslant 1$ in Dwork's theorem hold, then f is rational.*

This statement is proved in (Harbater, 1988, Proposition 2.1). In the next section we will present an application to the study of fundamental groups which was one of the catalysers of the proof of the unbounded denominators conjecture.

1.2. An application to fundamental groups

Let us show how Harbater's rationality theorem 1.3 can be used to prove that the arithmetic surface $\mathbb{P}^1_{\mathbf{Z}} \setminus \{0, 1, \infty\}$ is simply connected, *i.e.* that there are no non-trivial finite covers of the projective line over $\mathbf{Z}$ only ramified at $0, 1, \infty$. In other words, there are no Belyi maps with integer coefficients. This result was first obtained by T. Saito as an application of Abhyankar's lemma (Ihara, 1994, Appendix). In his original

paper, Harbater (1988, Example 3.1) only dealt with covers that are étale over 0, but then Ihara noticed that one can reduce to this case by taking a pullback by $z \mapsto z^N$. This trick will reappear in the proof of the unbounded denominators conjecture.

Theorem 1.4. *The étale fundamental group of* $\mathbb{P}^1_{\mathbf{Z}} \setminus \{0, 1, \infty\}$ *is trivial*[5].

Proof. Let $X \to \mathbb{P}^1_{\mathbf{Z}} \setminus \{0, 1, \infty\}$ be a finite étale cover and let N be the ramification index over 0. The pullback by the map $z \mapsto z^N$ from $\mathbb{P}^1_{\mathbf{Z}} \setminus \{0, \mu_N, \infty\}$ to $\mathbb{P}^1_{\mathbf{Z}} \setminus \{0, 1, \infty\}$ then extends to a finite étale cover $Y \to S = \mathbb{P}^1_{\mathbf{Z}} \setminus \{\mu_N, \infty\}$. Let f be a regular function on Y. Since $\mathrm{Spec}(\mathbf{Z})$ is simply connected, the section 0 of S lifts to a section 0 of Y around which f can be developed into a power series with integer coefficients and radius of convergence equal to 1 since there are no branch points over the open unit disc. By theorem 1.3, such a power series is rational, and hence all functions on Y come from the base S, which means precisely that this étale cover is trivial. □

This theme was further developed by Bost (1999) to obtain arithmetic analogues of the Lefschetz theorem on the fundamental group of a hyperplane section on a smooth projective variety based on the use of Arakelov's intersection theory on arithmetic surfaces. Under a positivity assumption akin to ampleness, he proves that the étale fundamental group of an arithmetic surface over the ring of integers $\mathscr{O}_K$ of a number field K is isomorphic to that of $\mathrm{Spec}(\mathscr{O}_K)$; this applies for example to an integral model of the modular curve $X(6)$. Calegari, Dimitrov, and Tang (2021, Theorem 7.4.4) extend it to all modular curves of level N over a number field containing $\mathbf{Q}(\mu_N)$ using the unbounded denominators conjecture. Note that there exist examples of number fields whose ring of integers has infinite fundamental group. By contrast, Bost and Charles (2022, Corollary 9.3.8) have recently proved that the arithmetic modular curves $\mathscr{Y}(N)$ over $\mathbf{Z}$ have *finite* étale fundamental group. We do not seem to know a single value of N for which this group is non-trivial.

1.3. The arithmetic holonomicity theorem

We now turn to algebraicity, as opposed to rationality, theorems. The following is a particular case of André (2004, Théorème 5.4.3), which more generally allows one to consider power series with rational coefficients by imposing conditions at all places (as in Dwork's theorem in comparison with Borel's).

Theorem 1.5 (André, 2004)**.** *Let* $f \in \mathbf{Z}[\![x]\!]$*. Assume that there exists a holomorphic function*[6] $\varphi\colon \overline{D(0,1)} \to \mathbf{C}$ *satisfying* $\varphi(0) = 0$ *and* $|\varphi'(0)| > 1$ *and such that* $f(\varphi(z)) \in \mathbf{C}[\![z]\!]$ *is holomorphic on* $|z| < 1$*. Then* f *is an algebraic function.*

[5] More generally, the étale fundamental group of $\mathbb{P}^1_{\mathscr{O}_K} \setminus \{0, 1, \infty\}$ is isomorphic to that of $\mathrm{Spec}(\mathscr{O}_K)$ for each number field K with ring of integers $\mathscr{O}_K$.

[6] By which we mean that φ is holomorphic on an open neighborhood of $D(0,1)$.

For the map $\varphi(z) = Rz$, the condition is that f is holomorphic on a disc of radius strictly larger than 1, and in that case f is even a polynomial by Borel's theorem. André (2004) and Bost (2001) apply this theorem to establish new cases of Grothendieck's p-curvature conjecture; see also the account by Chambert-Loir (2002) in this seminar.

It is worth noting that in the limit case $|\varphi'(0)| = 1$, the function f might be transcendental. An example is provided by Gauss's hypergeometric function

$$f(x) = {}_2F_1\left(\begin{matrix}\frac{1}{2} & \frac{1}{2} \\ 1\end{matrix} \mid 16x\right) = \sum_{n=0}^{\infty} \binom{2n}{n}^2 x^n \in \mathbf{Z}[\![x]\!],$$

which is transcendental despite the fact that the uniformisation $\varphi(z) = \lambda(z)/16$ brings it, by the classical Jacobi formula, into the holomorphic function on the unit disc $f(q) = (\sum_{n\in\mathbf{Z}} q^{n^2})^2$, which is a modular form of weight 1 for $\Gamma(2)$. Another limit case, in which algebraicity is known from the beginning[(7)] , is the map

$$\varphi(z) = \sqrt[N]{\lambda(z^N)/16}, \tag{3}$$

which turns the modular function $f(\tau) \in \mathbf{Z}[\![q^{1/N}]\!]$ into a holomorphic function on the unit disc. In that case, we are interested in bounding the algebraicity degree of f in terms of each function φ satisfying the assumptions of theorem 1.5, in order to get a bound as sharp as possible by making a better choice than (3).

Theorem 1.6 (Calegari–Dimitrov–Tang, 2021)**.** *Consider the following data:*

- ▷ *a non-constant rational function $p(x) \in \mathbf{Q}(x)$ without pole at $x = 0$;*
- ▷ *a formal power series $x(t) \in t + t^2\mathbf{Q}[\![t]\!]$ such that $p(x(t))$ has integer coefficients;*
- ▷ *a holomorphic function $\varphi\colon \overline{D(0,1)} \to \mathbf{C}$ satisfying $\varphi(0) = 0$ and $|\varphi'(0)| > 1$ and such that $p(\varphi(z))$ is holomorphic on $\overline{D(0,1)}$.*

Let $\mathscr{H}(x(t), \mathbf{Z})$ be the $\mathbf{Q}(p(x))$-vector space generated by formal power series $f \in \mathbf{Q}[\![x]\!]$ such that $f(x(t))$ has integer coefficients and $f(\varphi(z))$ is holomorphic on $\overline{D(0,1)}$. Then

$$\dim_{\mathbf{Q}(p(x))} \mathscr{H}(x(t), \mathbf{Z}) \leqslant e\, \frac{\int_{|z|=1} \log^+ |p \circ \varphi|\, \mu_{\mathrm{Haar}}}{\log|\varphi'(0)|}, \tag{4}$$

where $e = \exp(1)$ and $\log^+ = \log\max(1, -)$.

Some remarks are in order before we move into the proof.

(7) As explained in Calegari, Dimitrov, and Tang (2021, Theorem 7.2.1), the difference between these two examples is that the hypergeometric function is a solution of a differential equation with *infinite* local monodromy around $x = 0$.

- ▷ If f belongs to $\mathscr{H}(x(t), \mathbf{Z})$, then so do all its powers f^n. The finite-dimensionality of the space $\mathscr{H}(x(t), \mathbf{Z})$ then implies that f is algebraic. Taking $p(x) = x$ and $x(t) = t$, one hence recovers André's theorem 1.5, with an extra bound on the degree of f over $\mathbf{Q}(x)$.
- ▷ The reason for the name "arithmetic holonomicity theorem" is that Calegari, Dimitrov, and Tang (2021, Corollary 2.0.5) apply it to functions which are solutions of a non-zero differential operator in $\overline{\mathbf{Q}}(x)[d/dx]$ with trivial local monodromy around each point in the image of φ; by Cauchy's analyticity theorem on the solutions of ordinary differential equations with analytic coefficients, this is a way to guarantee that the function $f(\varphi(z))$ is holomorphic.
- ▷ A bound involving $\sup_{|z|=1} \log|p \circ \varphi|$ instead of the integrated term is easier to obtain, but will not be enough to make the leveraging argument in the proof of the unbounded denominators conjecture work (see section 3).

1.4. Proof of theorem 1.6

There are at least five different proofs of this theorem. Calegari, Dimitrov, and Tang present three in their paper, all of them relying on diophantine approximation techniques in a high number of variables that goes to infinity at the end of the argument. A more conceptual one-variable proof is given by Bost and Charles (2022, section 8.3.2) as an application of their theory of formal-analytic arithmetic surfaces in the framework of Arakelov geometry; it can be thought of as an arithmetic counterpart of a result by Nori bounding the degree of a dominant morphism between surfaces by a quotient of self-intersections of a divisor on the source and its direct image on the target. Inspired by this approach, Calegari, Dimitrov, and Tang later found a fifth proof using Bost's slope method. Both of these proofs actually yield the slightly better bound

$$\frac{\int_{|z|=|w|=1} \log|p(\varphi(z)) - p(\varphi(w))| \ \mu_{\mathrm{Haar}}(z)\mu_{\mathrm{Haar}}(w)}{\log|\varphi'(0)|},$$

which has the advantage of not involving the intriguing factor e anymore. The proof I sketch below relies on André's remark that considering the lowest monomial for the lexicographic order instead of the highest one simplifies the first proof given by the authors by avoiding the use of Bilu's equidistribution theorem.

Sketch of proof. Let $\mathbf{x} = (x_1, \dots, x_d)$. We use the standard multi-index notation

$$\mathbf{x}^{\mathbf{j}} = x_1^{j_1} \cdots x_d^{j_d} \quad \text{and} \quad p(\mathbf{x}) = (p(x_1), \dots, p(x_d)).$$

Let $f_1, \dots, f_m$ be $\mathbf{Q}(p(x))$-linearly independent elements of $\mathscr{H}(x(t), \mathbf{Z})$. Our goal is to show that the number m is bounded by the right-hand side of (4).

Step 1 (*Construction of an auxiliary function*). Let $d, \alpha \geqslant 1$ be integers and $\kappa \in (0,1)$. A standard application of Siegel's lemma[8] yields a *non-zero* auxiliary power series

$$F(\mathbf{x}) = \sum_{\substack{\mathbf{i}\in\{1,\dots,m\}^d \\ \mathbf{k}\in\{0,\dots,D-1\}^d}} a_{\mathbf{i},\mathbf{k}}\, p(\mathbf{x})^{\mathbf{k}} \prod_{s=1}^{d} f_{i_s}(x_s) \in \mathbf{Q}[\![\mathbf{x}]\!]$$

such that F vanishes to order $\geqslant \alpha$ at $\mathbf{x} = 0$, all $a_{\mathbf{i},\mathbf{k}}$ are integers bounded in absolute value by $\exp(\kappa C\alpha + o(\alpha))$ for some $C \in \mathbf{R}$ that only depends on $p(x)$ and φ, and

$$D \leqslant \frac{1}{(d!)^{1/d}} \frac{1}{m} \left(1 + \frac{1}{\kappa}\right)^{\frac{1}{d}} \alpha + o(\alpha). \tag{5}$$

In both estimates, the meaning of the asymptotic notation is that $o(\alpha)/\alpha$ has limit 0 as $\alpha \to \infty$ while d and κ are fixed. The idea is to express the vanishing condition as a system of $\binom{\alpha+d}{d} \sim \alpha^d/d!$ linear equations in the $(mD)^d$ variables $a_{\mathbf{i},\mathbf{k}}$. These equations have a priori rational coefficients, but the integrality conditions on $p(x(t))$ and $f_i(x(t))$ imply that there exists an integer M such that $f_i(x)$ lies in $\mathbf{Z}[\![x/M]\!]$. Moreover, there exists some radius $\rho > 0$ such that φ induces an analytic isomorphism from the connected component of $\varphi^{-1}(D(0,\rho))$ containing 0 to $D(0,\rho)$, and then all $f_i(x)$ converge on that disc. The constant C is given by $e^C = M/\rho$. The choice (5) guarantees that there are more equations that variables, so that Siegel's lemma yields a non-zero integral solution for the coefficients $a_{\mathbf{i},\mathbf{k}}$. If the function F were identically zero, then the f_i would be $\mathbf{Q}(p(x))$-linearly dependent.

Step 2 (*Cauchy-like bound*). Let $G(\mathbf{z})$ be a non-zero holomorphic function on the polydisc $|\mathbf{z}| \leqslant 1$, and let $c\mathbf{z}^{\mathbf{n}}$ be the smallest monomial for the lexicographic order in its power series representation. Then the following inequality holds:

$$\log|c| \leqslant \int_{T^d} \log|G| \mu_{\mathrm{Haar}},$$

where $T^d = \{\mathbf{z} \in \mathbf{C}^d \mid |z_i| = 1\}$. For $d = 1$, this follows from Jensen's formula

$$\log|c| = \int_T \log|G|\mu_{\mathrm{Haar}} + \sum_{\substack{w_i \in \overline{D(0,1)}\setminus\{0\} \\ G(w_i)=0}} \log|w_i|,$$

since the second term in the right-hand side is smaller than or equal to 0. One then performs an induction on d, by writing $\mathbf{z} = (z_1, \mathbf{z}')$, $\mathbf{n} = (n_1, \mathbf{n}')$, and $G(\mathbf{z}) = z_1^{n_1} H(\mathbf{z})$.

[8] Recall that *Siegel's lemma* is the following statement. Let $L > M$ and let $A = (a_{ij})$ be a non-zero $M \times L$ matrix with integer coefficients satisfying $|a_{ij}| \leqslant B$. Then the equation $A\mathbf{x} = 0$ has a non-zero integral solution with $\max|x_i| \leqslant \lfloor (LB)^{M/(L-M)} \rfloor$; see (Bombieri and Gubler, 2006, Lemma 2.9.1).

The assumption that $c\mathbf{z}^{\mathbf{n}}$ is the smallest monomial for the lexicographic order implies that $H(\mathbf{z})$ is holomorphic and the smallest monomial for the lexicographic order of $H(0,\mathbf{z}')$ is $c\mathbf{z}'^{\mathbf{n}'}$; see (Calegari, Dimitrov, and Tang, 2021, Lemma 2.4.1).

Let us apply this inequality to $G(\mathbf{z}) = F(\varphi(z_1),\dots,\varphi(z_d))$, which is a holomorphic function on $\overline{D(0,1)}$ because so are $p(\varphi(z))$ and $f_i(\varphi(z))$ by assumption. We get

$$\begin{aligned}\log|c| &\leqslant \int_{T^d} \log|F(\varphi(z_1),\dots,\varphi(z_d))|\mu_{\mathrm{Haar}}\\ &\leqslant dD\int_T \log^+|p\circ\varphi|\mu_{\mathrm{Haar}} + \kappa C\alpha + o(\alpha).\end{aligned}$$

The second inequality follows from integrating over T^d the pointwise bound

$$\log|F(\varphi(z_1),\dots,\varphi(z_n))| \leqslant D\sum_{i=1}^{d}\log^+|p(\varphi(z_i))| + \kappa C\alpha + o(\alpha),$$

which follows from the properties of the auxiliary function constructed in Step 1, on noting that the sum consists of $(mD)^d = \exp(o(\alpha))$ terms.

Step 3 (*Liouville-like bound*). From the integrality properties of $p(x(t))$ and $f_i(x(t))$, it follows that $F(x(t_1),\dots,x(t_d))$ is a non-zero power series in $\mathbf{t} = (t_1,\dots,t_d)$ with integer coefficients. Let β be the exact order of vanishing of $F(\mathbf{x})$ at $\mathbf{x} = 0$. From the assumptions $x(t) \in t + t^2\mathbf{Q}[\![t]\!]$ and $\varphi(z) = \varphi'(0)z + z^2\mathbf{C}[\![z]\!]$, we see that the coefficient c of the lowest monomial for the lexicographic order of $G(\mathbf{z})$ is the product of $\varphi'(0)^\beta$ and the corresponding coefficient of $F(x(\mathbf{t}))$. Being a non-zero element of the set $\varphi'(0)^\beta\mathbf{Z}$, it satisfies

$$\log|c| \geqslant \beta\log|\varphi'(0)| \geqslant \alpha\log|\varphi'(0)|.$$

Step 4 (*End of proof*). Putting the bounds from Step 2 and Step 3 together we get

$$\log|\varphi'(0)| \leqslant \frac{dD}{\alpha}\int_T \log^+|p\circ\varphi|\mu_{\mathrm{Haar}} + \kappa C + \frac{o(\alpha)}{\alpha}.$$

As $\alpha\to\infty$, the term $o(\alpha)/\alpha$ has limit 0 and the term dD/α is bounded above by

$$\frac{d}{(d!)^{1/d}}\frac{1}{m}\left(1+\frac{1}{\kappa}\right)^{\frac{1}{d}},$$

which has limit e/m by Stirling's formula as $d\to\infty$ first, then $\kappa\to 0$. Finally, since C is independent of d and κ, the extra term κC also disappears in the limit. □

2. Strategy of proof

In this section, we explain how the arithmetic holonomicity theorem 1.6 can be used to give a bound on the dimension of the space of modular functions with bounded denominators and cusp widths dividing $2N$ over the space of modular forms for the congruence subgroup $\Gamma(2N)$. The rough idea is to use Ihara's trick to get rid of the branch point of f at 0, and then apply the theorem to a big disc in the universal cover of $\mathbf{C} \setminus 16^{1/N}\mu_N$. After the computations of sections 4 and 5, the resulting upper bound will be sharp enough to make the leveraging argument of section 3 work.

2.1. The level of a non-congruence subgroup

Let $\Gamma \subset \mathrm{SL}_2(\mathbf{Z})$ be a subgroup of finite index. Since $\mathrm{SL}_2(\mathbf{Z})$ acts transitively on $\mathbb{P}^1(\mathbf{Q})$, the stabiliser of ∞ consists of all matrices $\pm \left(\begin{smallmatrix} 1 & m \\ 0 & 1 \end{smallmatrix}\right)$ with $m \in \mathbf{Z}$, and Γ has finite index, each point $\zeta \in \mathbb{P}^1(\mathbf{Q})$ is fixed by a non-trivial element of Γ, which is of the form $\pm M \left(\begin{smallmatrix} 1 & m \\ 0 & 1 \end{smallmatrix}\right) M^{-1}$ for some $m \in \mathbf{Z}$ and some $M \in \mathrm{SL}_2(\mathbf{Z})$ satisfying $M\infty = \zeta$. The smallest integer $m \geqslant 1$ with this property is called the *width* of the cusp[(9)].

Definition 2.1 (Wohlfahrt)**.** The level of Γ is the lowest common multiple of the widths at all cusps. We denote it by $L(\Gamma)$.

We will only consider the level of subgroups containing $E = \{\pm I\}$. According to Wohlfahrt (1964, Theorem 2), this definition generalises the usual notion of level of a congruence subgroup, in the sense that the level of a congruence subgroup Γ containing E is the *smallest* integer N for which $\Gamma \subset \langle E, \Gamma(N)\rangle$.

2.2. Modular forms with bounded denominators

For each even integer $N \geqslant 2$, we consider the $\mathbf{Q}(\lambda)$-vector spaces

$$M_2 \subset M_N \subset R_N \subset \mathbf{Q}[\![q^{1/N}]\!][1/q]$$

defined as follows:

- M_N is the field of rational functions on the modular curve $Y(N) = \mathfrak{H}/\Gamma(N)$. In particular, $M_2 = \mathbf{Q}(\lambda)$.
- R_N is generated by holomorphic modular functions for a subgroup of finite index of $\mathrm{SL}_2(\mathbf{Z})$, with rational coefficients with bounded denominators at infinity and widths dividing N at all cusps $\zeta \in \mathbb{P}^1(\mathbf{Q})$.

[(9)] The width of a cusp only depends on its Γ-orbit. Geometrically, each orbit defines a ramification point above ∞ of the quotient map $\mathfrak{H}/\Gamma \to \mathfrak{H}/\mathrm{SL}_2(\mathbf{Z})$ and its width is equal to the ramification index.

The inclusion $M_N \subset R_N$ comes from the argument using Hecke operators sketched in the introduction; with this notation, the unbounded denominators conjecture is the statement that this inclusion is an equality. For $N \geqslant 4$, we know the exact value[(10)] of the dimension of M_N over M_2, namely:

$$[M_N \colon M_2] = \frac{1}{2}[\Gamma(2) \colon \Gamma(N)] = \frac{N^3}{12} \prod_{p|N} \left(1 - \frac{1}{p^2}\right). \tag{6}$$

That R_N is finite-dimensional over M_2 follows from a crude application of theorem 1.6, as we will explain in the next section. Since the level of the intersection of two finite index subgroups containing E of level dividing N still divides N by Calegari, Dimitrov, and Tang (2021, Lemma 4.1.3), the space R_N is actually a ring, and hence a field since $\mathbf{Q}[\![q^{1/N}]\!][1/q]$ is an integral domain. Its degrees over M_N and M_2 are compared by means of (6). For example, bounding the finite product in that equality by the infinite Euler product of $1/\zeta(2)$, we get the inequality

$$[R_N \colon M_N] \leqslant \frac{12\zeta(2)}{N^3}[R_N \colon M_2]. \tag{7}$$

2.3. Bounding the degree $[R_N \colon M_2]$

Considering the coordinate $t = q^{1/N}$, we will apply the arithmetic holonomicity theoorem 1.6 to the polynomial $p(x) = x^N$ and the power series

$$x(t) = (\lambda(\tau)/16)^{1/N} = t - \tfrac{8}{N}t^{N+1} + \cdots \in t + t^2\mathbf{Q}[\![t]\!],$$

which satisfies $p(x(t)) \in \mathbf{Z}[\![t]\!]$ by the first equality in (2). For the same reason, $f(x(t))$ is a power series with integral coefficients for every $f \in \mathbf{Z}[\![q^{1/N}]\!]$.

Let $F_N \colon D(0,1) \to \mathbf{C} \setminus \mu_N$ be a universal covering map satisfying $F_N(0) = 0$, and

$$\varphi_r \colon \overline{D(0,1)} \longrightarrow \mathbf{C} \setminus 16^{-1/N}\mu_N, \quad \varphi_r(z) = 16^{-1/N}F_N(rz)$$

for some $r < 1$. We claim that the space R_N has a $\mathbf{Q}$-basis consisting of modular forms $f \in \mathbf{Z}[\![q^{1/N}]\!]$ such that $f(\varphi_r(z))$ is holomorphic on the closed unit disc $\overline{D(0,1)}$.

Indeed, if f is invariant under a subgroup $\Gamma \subset \Gamma(2)$ of finite index, then f descends to a regular function on the curve $Y = Y(\Gamma)$. Consider the diagram

$$\begin{array}{ccccc} Y & \hookrightarrow & Y' & \hookleftarrow & X \\ \downarrow & & \downarrow & & \downarrow \\ Y(2) & \hookrightarrow & \mathbf{C} \setminus \{1/16\} & \xleftarrow[z \mapsto z^N]{} & U = \mathbf{C} \setminus 16^{-1/N}\mu_N, \end{array}$$

(10) The factor $1/2$ comes from the fact that $-I$ belongs to $\Gamma(2)$ but not to $\Gamma(N)$.

where Y' denotes the curve Y with all the cusps above $0 \in Y(2)$ filled in, under the usual identification of $Y(2)$ with $\mathbf{C} \setminus \{0, 1/16\}$ via $\lambda/16$, and X is the fibre product. As in Ihara's trick, the condition that all cusps of Y have widths dividing N implies that $X \to U$ is a covering map (*i.e.* not ramified over $0 \in U$). Therefore, the universal covering map $\varphi_r \colon \overline{D(0,1)} \to U$ factors as

$$\overline{D(0,1)} \longrightarrow X \longrightarrow U,$$

hence a holomorphic map $\overline{D(0,1)} \to Y'$. Up to multiplying f by a high enough power of λ, we can assume that f is holomorphic at all cusps of Y' above 0. Then f is holomorphic on Y' and $f(\varphi_r(z))$ is nothing but the composition with $\overline{D(0,1)} \to Y'$.

The holomorphic function $\sqrt[N]{\lambda(z^N)/16} \colon D(0,1) \to U$ factors through the universal covering map $D(0,1) \to U$ and has derivative equal to 1 at $z = 0$; since the derivative at 0 of a holomorphic function $D(0,1) \to D(0,1)$ that maps 0 to 0 and is not a rotation has modulus < 1 by the Schwarz lemma, this implies $|\varphi_r'(0)| > 1$. The assumptions of theorem 1.6 are thus in force, hence the bound

$$[R_N \colon M_2] \leqslant eN \, \frac{\int_{|z|=1} \log^+ |\varphi_r| \, \mu_{\mathrm{Haar}}}{\log |\varphi_r'(0)|}, \tag{8}$$

which already shows that R_N is finite-dimensional.

Proposition 2.2. *Let $F_N \colon D(0,1) \to \mathbf{C} \setminus \mu_N$ be an analytic universal covering map with $F_N(0) = 0$. Assume that the following properties hold:*

(*a*) *there exists a real number $A > 0$ such that*

$$|F_N'(0)| \gg 16^{1/N}\big(1 + \frac{A}{N^3}\big)$$

holds as N goes to infinity ;

(*b*) *for each $B > 0$,*

$$\int_{|z|=1-\frac{B}{N^3}} \log^+ |F_N| \mu_{Haar} \ll_B \frac{\log N}{N}$$

holds as N goes to infinity.

Then there exists a real number C such that $[R_N \colon M_2] \leqslant CN^3 \log N$.

This follows from (8) by choosing $r = 1 - AN^{-3}/2$ and $B = A/2$.

3. The leveraging step

Before proving that there actually exists a universal covering map satisfying the assumptions of proposition 2.2, we explain how to derive the equality $R_N = M_N$ from the existence of a bound of the shape $CN^3 \log N$.

Theorem 3.1 (Calegari–Dimitrov–Tang). *Assume that there exists an integer $N \geqslant 1$ for which the inequality $[R_N \colon M_N] > 1$ holds. Then the inequality*

$$[R_{Np} \colon M_{Np}] \geqslant 2[R_N \colon M_N]$$

holds for all prime numbers p which do not divide N.

Corollary 3.2. *Assume that there exists a real number C such that the inequality*

$$[R_N \colon M_2] \leqslant CN^3 \log N$$

holds for all even integers N. Then $R_N = M_N$.

Proof. By contradiction, let us assume that R_N is strictly larger than M_N. By applying repeatedly theorem 3.1, we get the estimate

$$[R_{N\prod_{p\in S} p} \colon M_{N\prod_{p\in S} p}] \geqslant 2^{|S|}[R_N \colon M_N],$$

where S denotes the set of prime numbers smaller than some fixed X and not dividing N. On the other hand, the general bound (7) along with the assumption on $[R_N \colon M_N]$ give

$$[R_{N\prod_{p\in S} p} \colon M_{N\prod_{p\in S} p}] \leqslant 12C\zeta(2) \log N + 12C\zeta(2) \sum_{p\in S} \log p.$$

By the prime number theorem, for large enough X, there exists $\varepsilon > 0$ such that S has cardinal between $(1-\varepsilon)X/\log X$ and $(1+\varepsilon)X/\log X$. Since $2^{(1-\varepsilon)X/\log X}$ grows faster than $12C\zeta(2)(1+\varepsilon)X$, the lower and the upper bound contradict each other. □

3.1. Sketch of proof of theorem 3.1

Let p be a prime number not dividing N and $R_N M_{Np}$ the compositum of the fields R_N and M_{Np}. Using the multiplicativity of degrees in the tower of field extensions

$$M_N \subset M_{Np} \subset R_N M_{Np} \subset R_{Np}$$

and the fact that the intersection of R_N and M_{Np} is equal to M_N, one finds

$$[R_{Np} \colon M_{Np}] = [R_{Np} \colon R_N M_{Np}][R_N \colon M_N].$$

It is hence enough to prove that, if R_N is strictly larger than M_N, then R_{Np} is *not* generated by R_N and M_{Np}. By contradiction, assume that it is.

Choose a form $f(\tau) \in \mathbf{Z}[\![q^{1/N}]\!]$ in the complement $R_N \setminus M_N$. The finite-dimensionality of R_N over M_N and the properties of the level imply that such a form is invariant under a normal subgroup $G \subset \langle E, \Gamma(N)\rangle$ of level N. Indeed, every element in a $\mathbf{Q}(\lambda)$-basis of R_N is invariant under a finite index subgroup containing E whose level divides N, and one defines G as the largest normal subgroup of $\mathrm{SL}_2(\mathbf{Z})$ contained in the intersection of all those. We let $\Gamma_0(p)$ and $\Gamma^0(p)$ denote the subgroups of $\mathrm{SL}_2(\mathbf{Z})$ consisting of matrices $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ such that $c \equiv 0$ and $b \equiv 0 \bmod p$, respectively. The matrix $A = \left(\begin{smallmatrix} p & 0 \\ 0 & 1 \end{smallmatrix}\right)$ conjugates them, in the sense that $A\Gamma_0(p)A^{-1} = \Gamma^0(p)$ holds. Consider the form

$$f(p\tau) \in \mathbf{Z}[\![q^{1/N}]\!],$$

which is invariant under the subgroup $A^{-1}GA \cap \mathrm{SL}_2(\mathbf{Z})$. Elementary manipulations, as performed in (Calegari, Dimitrov, and Tang, 2021, Lemma 4.1.7), show that the level $L(A^{-1}GA \cap \mathrm{SL}_2(\mathbf{Z}))$ divides Np, and hence that $f(p\tau)$ belongs to R_{Np}. By our assumption that R_{Np} is generated by R_N and M_{Np}, the form $f(p\tau)$ is also invariant under $G \cap \Gamma(Np)$. In view of theorem 3.3 below, this implies that $f(p\tau)$ is invariant under $\langle E, \Gamma(N)\rangle \cap \Gamma_0(p)$. Hence, the original form $f(\tau)$ is invariant under

$$A\big(\langle E, \Gamma(N)\rangle \cap \Gamma_0(p)\big)A^{-1} = \langle E, \Gamma(N)\rangle \cap \Gamma^0(p),$$

which contains $\Gamma(Np)$. In other words, f belongs to M_{Np}. From $R_N \cap M_{Np} = M_N$, we then get the contradiction that f belongs to M_N.

Theorem 3.3. *The subgroup generated by E, $G \cap \Gamma(Np)$, and $A^{-1}GA \cap \Gamma_0(p)$ contains a congruence subgroup. More precisely,*

$$\langle E, G \cap \Gamma(Np), A^{-1}GA \cap \Gamma_0(p)\rangle = \langle E, \Gamma(N)\rangle \cap \Gamma_0(p). \tag{9}$$

The proof follows Serre's argument in the letter reproduced in (Thompson, 1989), and its adaptation from $\mathrm{SL}_2(\mathbf{Z})$ to $\Gamma(N)$ by Berger (1994), who proved that the subgroup generated by $G \cap \Gamma_0(p)$ and $A^{-1}GA \cap \Gamma_0(p)$ is equal to $\langle E, \Gamma(N)\rangle \cap \Gamma_0(p)$. The key idea is to introduce the finite group $S = \langle E, \Gamma(N)\rangle / G$ and the subgroup $B \subset \mathrm{SL}_2(\mathbf{F}_p)$ of upper triangular matrices, along with the quotient map $\pi\colon \langle E, \Gamma(N)\rangle \to S$ and the reduction map $r\colon \langle E, \Gamma(N)\rangle \cap \Gamma_0(p) \to B$, and then prove that the homomorphism

$$\begin{aligned} h = (\pi_1, \pi_2, r)\colon \langle E, \Gamma(N)\rangle \cap \Gamma_0(p) &\longrightarrow S \times S \times B \\ x &\longmapsto (\pi(x), \pi(AxA^{-1}), r(x)) \end{aligned} \tag{10}$$

is surjective. Using $\ker((\pi_1, r)) = \langle E, G \cap \Gamma(Np)\rangle$ and $\ker(\pi_2) = A^{-1}GA \cap \Gamma_0(p)$, this suffices to conclude. Indeed, thanks to the surjectivity of the map, the image of

each element $x \in \langle E, \Gamma(N)\rangle \cap \Gamma_0(p)$ can be written as

$$h(x) = (\pi_1(x), 0, r(x)) + (0, \pi_2(x), 0) = h(y) + h(z)$$

for some $y \in \ker(\pi_2)$ and $z \in \ker((\pi_1, r))$, and hence x lies in $\langle \ker((\pi_1, r)), \ker(\pi_2)\rangle$.

The proof of the surjectivity of

$$(\pi_1, \pi_2)\colon \langle E, \Gamma(N)\rangle \cap \Gamma_0(p) \longrightarrow S \times S$$

relies on two properties of the group $\mathrm{SL}_2(\mathbf{Z}[1/p])$: Ihara's theorem that realises it as an amalgam, as explained in (Serre, 1980, page 80), and the theorem by Mennicke (1967) according to which all of its subgroups of finite index are congruence. This is how we use them. By Goursat's lemma on the subgroups of a product, if (π_1, π_2) is not surjective, then there exist a non-trivial group T and surjective morphisms $h_i\colon S \to T$ satisfying $h_1 \circ \pi_1 = h_2 \circ \pi_2$. By construction, the maps

$$g_1\colon \langle E, \Gamma(N)\rangle \longrightarrow T, \quad g_2\colon A^{-1}\langle E, \Gamma(N)\rangle A \longrightarrow T$$

given by $g_1(x) = h_1(\pi(x))$ and $g_2(A^{-1}xA) = h_2(\pi(x))$ agree on the intersection

$$\langle E, \Gamma(N)\rangle \cap A^{-1}\langle E, \Gamma(N)\rangle A = \langle E, \Gamma(N)\rangle \cap \Gamma_0(p),$$

and hence induce a surjective map on the amalgam

$$\langle E, \Gamma(N)\rangle *_{\langle E, \Gamma(N)\rangle \cap \Gamma_0(p)} A^{-1}\langle E, \Gamma(N)\rangle A \longrightarrow T.$$

Inside $\mathrm{SL}_2(\mathbf{Z}[1/p]) = \mathrm{SL}_2(\mathbf{Z}) *_{\Gamma_0(p)} \mathrm{SL}_2(\mathbf{Z})$, the source of this map is isomorphic to the congruence subgroup consisting of matrices congruent to I or $-I$ modulo N. Since T is finite, the kernel of this map is a subgroup of finite index of $\mathrm{SL}_2(\mathbf{Z}[1/p])$, and hence contains a congruence subgroup. The same holds for its restriction to $\langle E, \Gamma(N)\rangle$, and this implies that the kernel of the non-zero map g_1 is a congruence subgroup of $\mathrm{SL}_2(\mathbf{Z})$ containing G but strictly smaller than $\langle E, \Gamma(N)\rangle$. This contradicts the fact that G has level N, since $\langle E, \Gamma(N)\rangle$ is the smallest congruence subgroup with this property.

Essentially these same arguments lead Ribet (1984) to prove a result on the group cohomology of $\Gamma(N)$ with coefficients in a finite field known as *Ihara's lemma*. To state it, assume $N \geqslant 3$, let ℓ be a prime number, and consider the map

$$\mathrm{H}^1(\Gamma(N), \mathbf{F}_\ell)^{\oplus 2} \longrightarrow \mathrm{H}^1(\Gamma(N) \cap \Gamma_0(p), \mathbf{F}_\ell) \tag{11}$$

that sends a pair (ψ, ϕ) of cocycles $\Gamma(N) \to \mathbf{F}_\ell$ to the difference of the restriction $\psi|_{\Gamma(N)\cap\Gamma_0(p)}$ and the cocycle $A\varphi\colon \Gamma(N) \cap \Gamma_0(p) \to \mathbf{F}_\ell$ given by $g \mapsto \varphi(AgA^{-1})$. Then Ihara's lemma is the statement that all classes in the kernel of (11) restrict to zero on some congruence subgroup of $\Gamma(N)$. The link with the above is that,

after identifying $\mathrm{H}^1(\Gamma(N), \mathbf{F}_\ell)$ with $\mathrm{Hom}(R, \mathbf{F}_\ell)$, for an elementary ℓ-abelian quotient R of $\Gamma(N)$, the kernel of (11) is in duality with the cokernel of a morphism $\Gamma(N) \cap \Gamma_0(p) \to R \times R$.

Letting $\Gamma_1(p)$ denote the subgroup of $\mathrm{SL}_2(\mathbf{Z})$ of matrices congruent to $\left(\begin{smallmatrix} 1 & * \\ 0 & 1 \end{smallmatrix}\right)$ mod p, Calegari, Dimitrov, and Tang (2021, Lemma 4.6.3) prove that the conclusion of Ihara's lemma still holds for the kernel of the composition of (11) with the restriction map to $\mathrm{H}^1(\Gamma(N) \cap \Gamma_1(p), \mathbf{F}_\ell)$, and deduce from this the surjectivity of (10). □

4. The uniformization radius of $\mathbf{C} \setminus \mu_N$

Let us assume $N \geqslant 2$. Then $\mathbf{C} \setminus \mu_N$ is a projective line punctured at least 3 times, and it is hence uniformised by the upper half-plane $\mathfrak{H}$. All analytic universal covering maps realising $\mathbf{C} \setminus \mu_N$ as the quotient of $\mathfrak{H}$ by a Fuchsian subgroup of $\mathrm{PSL}_2(\mathbf{R})$ are conjugate to each other, so that there is a unique uniformisation map

$$\widetilde{F}_N \colon \mathfrak{H} \longrightarrow \mathbf{C} \setminus \mu_N$$

once one enforces the normalisations $\widetilde{F}_N(i) = 0$ and $\lim_{r\to\infty} \widetilde{F}_N(ir) = 1$.

Definition 4.1. We let $F_N \colon D(0,1) \to \mathbf{C} \setminus \mu_N$ denote the composition of $\widetilde{F}_N$ with the standard conformal isomorphism $D(0,1) \to \mathfrak{H}$ that maps 0 to i. That is,

$$F_N(x) = \widetilde{F}_N\left(i\frac{1+x}{1-x}\right).$$

This maps hence satisfies $F_N(0) = 0$ and $\lim_{r\to 1^-} F_N(r) = 1$.

Our goal is to compute $|F_N'(0)|$, which is called the *uniformisation radius* of $\mathbf{C} \setminus \mu_N$. The result was first obtained by Kraus and Roth (2016, Remark 5.1). We will rather follow the self-contained approach by Calegari, Dimitrov, and Tang (2021).

Theorem 4.2 (Kraus–Roth, 2016)**.** *The uniformization radius of $\mathbf{C} \setminus \mu_N$ is equal to*

$$|F_N'(0)| = 16^{1/N} \frac{\Gamma\left(1+\frac{1}{2N}\right)^2 \Gamma\left(1-\frac{1}{N}\right)}{\Gamma\left(1-\frac{1}{2N}\right)^2 \Gamma\left(1+\frac{1}{N}\right)}, \tag{12}$$

and hence admits an asymptotic expansion as $N \to \infty$ of the form

$$|F_N'(0)| = 16^{1/N}\left(1 + \frac{\zeta(3)}{2N^3} + O(N^{-5})\right).$$

Therefore, the map F_N satisfies condition (a) from section 2.3. From the equality (12), the asymptotic expansion follows readily by using the classical identity of the gamma function

$$\Gamma(1+s) = \exp\left(-\gamma s + \sum_{k=2}^{\infty} (-1)^k \frac{\zeta(k)}{k} s^k \right) \quad \text{for} \quad |s| < 1,$$

where γ stands for Euler's constant.

4.1. Sketch of proof of theorem 4.2

The strategy to find the uniformisation radius follows Poincaré's original approach to the uniformisation theorem, as beautifully described in the collective book by Saint-Gervais (2010). Namely, we will express the local analytic inverse map ψ_N of F_N with $\psi_N(0) = 0$ as the quotient of two linearly independent solutions of an explicit second order linear differential equation. For a general punctured projective line, these solutions are not expected to be expressible in terms of classical special functions, but the symmetries of roots of unity will bring hypergeometric functions into the picture.

Recall that the *schwarzian derivative* of a holomorphic function f of the variable z is defined at a point $z = z_0$ with $f'(z_0) \neq 0$ by the formula

$$\{f, z_0\} = \left[\left(\frac{f''}{f'} \right)' - \frac{1}{2} \left(\frac{f''}{f'} \right)^2 \right] (z_0).$$

As a function of z, it satisfies the chain rule

$$\{g \circ f, z\} = \{g, f(z)\} f'(z)^2 + \{f, z\}.$$

From this and a straightforward computation showing that $\{T, z\}$ vanishes for all Möbius transformations T, it follows that f and $T \circ f$ have the same schwarzian derivative. Up to Möbius transformations, the second order differential equation

$$\frac{d^2y}{dz} + \frac{1}{2}\{f, z\} y = 0$$

admits $\eta_1 = f(f')^{-1/2}$ and $\eta_2 = (f')^{-1/2}$ as two linearly independent solutions, so that one recovers the original function f as their quotient.

Since F_N and $\widetilde{F}_N$ are related by a Möbius transformation, we can work with the latter. Let $z = \widetilde{F}_N(\tau)$ and $z_0 \in \mathbf{C} \setminus \mu_N$. For each preimage τ_0 of z_0 in $\mathfrak{H}$, we may view τ as an analytic function of z in a small neighbourhood of z_0 satisfying $\tau(z_0) = \tau_0$ and $\tau'(z_0) \neq 0$. Since the functions resulting from different choices of a preimage differ by a Möbius transformation, the value of the schwarzian $\{\tau, z_0\}$ is well-defined. It turns out to be easier to work with the reciprocal function

$$1/\widetilde{F}_N \colon \mathfrak{H} \to \mathbb{P}^1(\mathbf{C}) \setminus \{0, \mu_N\},$$

whose schwarzian is related to the previous one by the identity $\{\tau, z\} = z^4 \{\tau, 1/z\}$.

Set $p_0 = 0$ and $p_k = \xi_N^k$ for $k = 1, \dots, N$, where $\xi_N = \exp(2\pi i/N)$ is the standard primitive Nth root of unity. By Hempel (1988, Theorem 3.1), the schwarzian of an analytic local inverse of a universal covering of $\mathbf{C} \setminus \{p_0, \dots, p_N\}$ takes the form

$$\{\tau, 1/\widetilde{F}_N\} = \frac{1}{2}\sum_{k=0}^{N} \frac{1}{(z-p_k)^2} + \sum_{k=0}^{N} \frac{m_k}{z-p_k}$$

for certain complex numbers m_k, the so-called *accessory parameters*, satisfying

$$\sum_{k=0}^{N} m_k = 0, \quad \sum_{k=0}^{N} (2m_k p_k + 1) = 0, \quad \sum_{k=0}^{N} (m_k p_k^2 + p_k) = 0. \tag{13}$$

(These constraints express the vanishing to order 4 of the schwarzian at infinity.) Moreover, writing the non-zero accessory parameters in the form $m_k = 1/(q_k - p_k)$, a Möbius transformation T with $Tp_k = P_k$ and $Tq_k = Q_k$ turns them into $M_k = 1/(Q_k - P_k)$ in the corresponding expression for the schwarzian $\{\tau, Tz\}$ by Lemma 3.2 of *loc. cit.*

In the case at hand, the fact that $\mathbf{C} \setminus \{0, \mu_N\}$ is stable under the Möbius transformation $Tz = \xi z$ implies the equalities $m_0 = 0$ and $m_k = k\xi_N^{-k}$ for some constant c, which is then seen to be equal to $c = -1/2 - 1/N$ using the second constraint in (13). Putting everything together, one then finds the equality

$$\{\tau, \widetilde{F}_N\} = \frac{(N^2-1)z^{N-2} + z^{2N-2}}{2(z^N-1)^2},$$

and hence that the analytic local inverse map of F_N is a quotient $\psi_N = \eta_1/\eta_2$ of two linearly independent solutions η_1 and η_2 of the differential equation

$$4(z^N-1)^2 \frac{d^2y}{dz^2} + [(N^2-1)z^{N-2} + z^{2N-2}]y = 0. \tag{14}$$

In fact, the unique solutions η_1 and η_2 satisfying the initial conditions

$$\eta_1(0) = 0, \quad \eta_1'(0) = 1, \quad \eta_2(0) = 1, \quad \eta_2'(0) = 0$$

are linearly independent and, since $\eta_1(z)/\eta_2(z) = z + O(z^2)$, yield $|F_N'(0)| = 1/|\psi_N'(0)|$.

To compute this quantity, it will be more convenient to work with the closely related function $G_N \colon D(0,1) \to \mathbf{C} \setminus \{1\}$ given by

$$G_N(x) = F_N(x^{1/N})^N,$$

which is well defined since $F_N(\xi x) = \xi F_N(x)$ holds for each Nth root of unity ξ by Calegari, Dimitrov, and Tang (2021, Lemma 5.1.2), and satisfies $|G_N'(0)| = |F_N'(0)|^N$. The functions $\phi_i(z) = \eta_i(z^{1/N})$ are then solutions to the differential equation

$$z(z-1)^2 \frac{d^2y}{dz^2} + \left(1 - \frac{1}{N}\right)(z-1)^2 \frac{dy}{dz} + \left(\frac{1}{4} + \frac{z-1}{4N^2}\right)y = 0,$$

which is essentially of hypergeometric type. They are explicitly given by

$$\phi_1 = \sqrt{1-z}\cdot z^{1/N}\cdot {}_2F_1\left(\begin{matrix}\frac{N+1}{2N} & \frac{N+1}{2N}\\ 1+\frac{1}{N}\end{matrix}\mid z\right),\quad \phi_2 = \sqrt{1-z}\cdot {}_2F_1\left(\begin{matrix}\frac{N-1}{2N} & \frac{N-1}{2N}\\ 1-\frac{1}{N}\end{matrix}\mid z\right),$$

which results into the expression

$$\psi_N(z) = |F_N'(0)|^{-1} z \frac{{}_2F_1\left(\begin{matrix}\frac{N+1}{2N} & \frac{N+1}{2N}\\ 1+\frac{1}{N}\end{matrix}\mid z^N\right)}{{}_2F_1\left(\begin{matrix}\frac{N-1}{2N} & \frac{N-1}{2N}\\ 1-\frac{1}{N}\end{matrix}\mid z^N\right)}. \tag{15}$$

Besides, F_N being a covering map of $\mathbf{C}\setminus\mu_N$, its local inverse is naturally defined on the whole unit disc $D(0,1)$ and has $\lim_{z\to 1^-}\psi_N(z)=1$ as $z\in D(0,1)$ approaches 1 by our particular normalisation $\lim_{r\to 1^-}F_N(r)=1$. The final expression for the uniformisation radius of $\mathbf{C}\setminus\mu_N$ then comes from the asymptotic formula

$$\lim_{z\to 1^-}\frac{{}_2F_1\left(\begin{matrix}a & a\\ 2a\end{matrix}\mid z\right)}{\log(1-z)} = \frac{\Gamma(2a)}{\Gamma(a)^2} \tag{16}$$

and the relation $\Gamma(s+1)\Gamma(s+1/2) = 4^{-s}\sqrt{\pi}\Gamma(2s+1)$ obtained from the duplication formula and the functional equation of the gamma function. □

5. Mean growth estimate of F_N^N

Let us assume $N\geqslant 2$. Recall the universal covering map $F_N\colon D(0,1)\to\mathbf{C}\setminus\mu_N$ from definition 4.1. In order to complete the strategy to prove the unbounded denominators conjecture laid out in section 2, it remains to find a good uniform bound for the mean growth of the function F_N. This is accomplished by the following theorem:

Theorem 5.1 (Calegari–Dimitrov–Tang). *The estimate for the mean growth*

$$\int_{|z|=r}\log^+|F_N|\mu_{\mathrm{Haar}} \ll \frac{1}{N}\log\left(\frac{N}{1-r}\right)$$

holds uniformly in $N\geqslant 2$ and $r\in(0,1)$.

This is an improvement, possible thanks to the exceptional symmetry of roots of unity, of a theorem by Tsuji (1952), which for *fixed* N gives a general asymptotic

$$\int_{|z|=r}\log^+|F|\mu_{\mathrm{Haar}} = \frac{1}{N-1}\log\left(\frac{1}{1-r}\right) + O_{p_1,\dots,p_N}(1)\quad \text{as } r\to 1^-$$

for any universal covering map $F\colon D(0,1)\to\mathbf{C}\setminus\{p_1,\dots,p_N\}$ with $F(0)=0$.

5.1. Tools from Nevanlinna theory

Very roughly speaking, Nevanlinna theory aims at measuring "how many" values close to a given point $a \in \mathbb{P}^1(\mathbf{C})$ a meromorphic function $f\colon \overline{D(0,R)} \to \mathbb{P}^1(\mathbf{C})$ takes. It gives, for example, a quantitative refinement of the little Picard theorem. For $a = \infty$, this is done through the following three quantities, defined for each $0 \leqslant r \leqslant R$:

▷ the *mean proximity function*

$$m(r,f) = \int_{|z|=r} \log^+ |f| \mu_{\mathrm{Haar}} \in [0,\infty);$$

▷ the *counting function*

$$N(r,f) = \sum_{0<|\rho|<r} \mathrm{ord}_\rho^-(f) \log \frac{r}{|\rho|} + \mathrm{ord}_0^-(f) \log r,$$

where $\mathrm{ord}_z^-(f)$, a non-negative integer, stands for the order of the pole z of f;

▷ the *characteristic function*

$$T(r,f) = m(r,f) + N(r,f),$$

which is the best behaved among these three quantities.

Let $c(f,a)$ denote the first non-zero coefficient in the Laurent power series expansion of f around a. From the Poisson–Jensen formula we get

$$T(r,f) - T(r,1/f) = \log |c(f,0)| \tag{17}$$

and from the triangle inequality

$$|T(r,f) - T(r,f-a)| \leqslant \log^+ |a| + \log 2 \tag{18}$$

for all $a \in \mathbf{C}$, which can then be combined into the inequality

$$\big|T(r,f) - T(r,1/(f-a))\big| \leqslant \log^+ |a| + \log 2 + \log |c(f,a)|,$$

in which the right-hand side is independent of the radius r. Therefore, the "number of times" that f takes the value ∞ or any other value $a \in \mathbf{C}$ are equivalent for big enough r. This is usually referred to as the *first main theorem* in Nevanlinna theory.

The *second main theorem* is the statement that, for distinct points $a_1, \dots, a_n \in \mathbf{C}$, the sum of the mean proximity functions at the a_i's is bounded by

$$\sum_{i=1}^n m\left(r, \frac{1}{f-a_i}\right) \leqslant 2T(r,f) + \text{small error term}.$$

We will not need the full strength of this theorem, but only the elementary manipulations with the characteristic function that enter its proof and the following so-called *lemma on the logarithmic derivative*:

Proposition 5.2. *Let $f\colon \overline{D(0,R)} \to \mathbf{C}$ be a nowhere vanishing holomorphic function satisfying $f(0) = 1$. For all $0 < r < R$, the following holds*

$$m\left(r, \frac{f'}{f}\right) < \log^+\left(\frac{m(R,f)}{r}\frac{R}{R-r}\right) + \log 2 + \frac{1}{e}.$$

A short proof is given in (Calegari, Dimitrov, and Tang, 2021, Lemma 6.1.5).

5.2. Sketch of proof of theorem 5.1

Since the characteristic function of a holomorphic function $f\colon D(0,1) \to \mathbf{C}$ coincides with its mean proximity, we can reformulate the theorem as the estimate

$$T(r, F_N) \ll \frac{1}{N}\log\left(\frac{N}{1-r}\right).$$

On noting that the function $F_N^N - 1$ does not vanish on $D(0,1)$, one easily derives

$$NT(r,F^N) - \log 4 \leqslant T\left(r, \frac{F_N^N}{F_N^N - 1}\right) \leqslant NT(r,F^N) + \log 4 \tag{19}$$

from the relations (17) and (18). Therefore, it is equivalent to prove the estimate

$$T\left(r, \frac{F_N^N}{F_N^N - 1}\right) \ll \log\left(\frac{N}{1-r}\right).$$

One advantage of working with this function is that, up to a factor, it is the quotient of the logarithmic derivatives of F_N and $f = 1 - F_N^N$, namely

$$\frac{F_N^N}{F_N^N - 1} = \frac{1}{N}\frac{F_N}{F_N'}\frac{f'}{f}, \tag{20}$$

so that we will be able to exploit the bounds from proposition 5.2 (note that F_N' is nowhere vanishing since F_N is an étale analytic map). By elementary manipulations, performed in Corollaries 6.2.6 and 6.2.8 of *loc. cit.*, we get

$$m\left(r, \frac{f'}{f}\right) \ll \sup_{|z|=(1+r)/2} \log^+\log|F_N| + \log^+\left(\frac{N}{1-r}\right),$$
$$m\left(r, \frac{F_N}{F_N'}\right) \ll T(r,F_N) + O\left(\sup_{|z|=(1+r)/2} \log^+\log|F_N| + \log^+\left(\frac{N}{1-r}\right)\right).$$

From the identity (20), along with (19), we then get

$$T(r,F_N) \ll \frac{1}{N}\log^+\left(\frac{N}{1-r}\right) + \frac{1}{N}\sup_{|z|=(1+r)/2} \log^+\log|F_N|.$$

It remains to bound $\log|F_N|$ over circles, for which the following suffices:

Lemma 5.3. *For $r \in (0,1)$ and large enough N, the estimate*

$$\sup_{|z|=r} \log |F_N| \ll \frac{N}{1-r}$$

holds, with absolute implicit constants.

A better bound is due to Kraus and Roth (2016, Theorems 1.2 and 1.10). Calegari, Dimitrov, and Tang (2021) prove the lemma by exploiting the action of the fuchsian group of $\mathbf{C} \setminus \mu_N$ to reinterpret the statement in terms of the asymptotic behaviour of $\widetilde{F}_N$ near the cusp at infinity, which can then be studied by means of the explicit formula (15) for the local inverse and a refinement of formula (16). □

Acknowledgement. — I first learnt about the proof of the unbounded denominators conjecture during a reading seminar organised at the École Normale Supérieure by Nicolas Bergeron and François Charles in the Fall 2021. Many thanks to them, as well as to all the speakers and participants of that seminar. This was also the origin of long email exchanges with Vesselin Dimitrov, who patiently answered all my questions, provided enlightening explanations, and helped me improving a first draft of this text. I would also like to thank Yves André, Jean-Benoît Bost, José Ignacio Burgos Gil, Antoine Chambert-Loir, Christophe Margerin, Jean Lannes, and John Voight for useful discussions.

References

André, Y. (2004). "Sur la conjecture des p-courbures de Grothendieck–Katz et un problème de Dwork", in: *Geometric aspects of Dwork theory*. Ed. by Adolphson A. *et al.* Vol. I. 2. Walter de Gruyter, pp. 55–112.

Atkin, A. O. L. and Swinnerton-Dyer, H. P. F. (1971). "Modular forms on noncongruence subgroups", in: *Combinatorics*. Vol. XIX. Proc. Sympos. Pure Math. American Mathematical Society, pp. 1–26.

Bass, H., Lazard, M., and Serre, J.-P. (1964). "Sous-groupes d'indice fini dans SL(n, **Z**)", *Bull. Amer. Math. Soc.* **70**, pp. 385–392.

Berger, G. (1994). "Hecke operators on noncongruence subgroups", *C. R. Acad. Sci. Paris Sér. I Math.* **319** (9), pp. 915–919.

Bombieri, E. and Gubler, W. (2006). *Heights in Diophantine geometry*. New Mathematical Monographs 4. Cambridge University Press.

Borel, É. (1894). "Sur une application d'un théorème de M. Hadamard", *Bulletin des sciences mathématiques* **18**, pp. 22–25.

Bost, J.-B. (1999). "Potential theory and Lefschetz theorems for arithmetic surfaces", *Ann. Sci. École Norm. Sup.* **32** (2), pp. 241–312.

Bost, J.-B. (2001). "Algebraic leaves of algebraic foliations over number fields", *Publ. Math. Inst. Hautes Études Sci.* **93**, pp. 161–221.

Bost, J.-B. and Charles, F. (2022). "Quasi-projective and formal-analytic arithmetic surfaces", arXiv: `math/2206.14242v2`.

Calegari, F., Dimitrov, V., and Tang, Y. (2021). "The unbounded denominators conjecture", arXiv: `math/2109.09040v2`.

Chambert-Loir, A. (2002). "Théorème d'algébricité en géométrie diophantienne [d'aprés J.-B. Bost, Y. André, D. & G. Chudnovsky]", in: *Séminaire Bourbaki, 53e année, 2000–2001*. Vol. 282. Astérisque. Soc. Math. France, pp. 175–209.

Dennin, J. B. (1975). "The genus of subfields of $K(n)$", *Proc. Amer. Math. Soc.* **51**, pp. 282–288.

Dwork, B. (1960). "On the rationality of the zeta function of an algebraic variety", *Amer. J. Math.* **82**, pp. 631–648.

Harbater, D. (1988). "Galois covers of an arithmetic surface", *Amer. Math. J.* **110** (5), pp. 849–885.

Hempel, J. A. (1988). "On the uniformization of the n-punctured sphere", *Bull. London Math. Soc.* **20**, pp. 97–115.

Ihara, Y. (1994). "Horizontal divisors on arithmetic surfaces associated with Belyĭ uniformizations", in: *The Grothendieck theory of dessins d'enfants* (*Luminy, 1993*). Vol. 200. London Math. Soc. Lecture Note Ser. Cambridge Univ. Press, pp. 245–254.

Jones, G. A. (1979). "Triangular maps and noncongruence subgroups of the modular group", *Bull. London Math. Soc.* **11** (2), pp. 117–123.

Klein, F. and Fricke, R. (2017). *Lectures on the theory of elliptic modular functions*. Vol. 1. Classical Topics in Mathematics. Translated from the German original by Arthur M. DuPre. Beijing: Higher Education Press.

Kraus, D. and Roth, O. (2016). "Sharp lower bounds for the hyperbolic metric of the complement of a closed subset of the unit circle and theorems of Schwartz–Pick-, Schottky- and Landau-type for analytic functions", *Constr. Approx.* **43** (1), pp. 47–69.

Mennicke, J. L. (1965). "Finite factor groups of the unimodular group", *Ann. of Math.* **81**, pp. 31–37.

——— (1967). "On Ihara's modular group", *Invent math.* **4**, pp. 202–228.

Ribet, K. A. (1984). "Congruence relations between modular forms". In: *Proceedings of the International Congress of Mathematicians* (*Warsaw, 1983*). Vol. 1, pp. 503–514.

Saint-Gervais, H. P. de (2010). *Uniformisation des surfaces de Riemann. Retour sur un théorème centenaire.* ENS Éditions, Lyon.

Serre, J.-P. (1980). *Trees*. Translated from the French original by John Stillwell. Springer-Verlag.

Shimura, G. (1971). *Introduction to the arithmetic theory of automorphic functions*. Princeton University Press.

Thompson, J. G. (1989). "Hecke operators and noncongruence subgroups", in: *Group theory* (*Singapore, 1987*). Including a letter from Jean-Pierre Serre. De Gruyter, Berlin, pp. 215–224.

Tsuji, M. (1952). "Theory of Fuchsian groups", *Jpn. J. Math.* **21**, pp. 1–27.

Wohlfahrt, K. (1964). "An extension of F. Klein's level concept", *Illinois J. Math.* **8**, pp. 529–535.

Javier Fresán

Sorbonne Université and Université Paris Cité
CNRS, IMJ-PRG
F-75005 Paris, France

E-mail: javier.fresan@imj.prg.fr

Séminaire BOURBAKI Novembre 2022
74e année, 2022–2023, no 1198, p. 29 à 86
doi : 10.24033/ast.1207

VALIDITÉ DE LA THÉORIE CINÉTIQUE DES GAZ : AU-DELÀ DE L'ÉQUATION DE BOLTZMANN

[*d'après* T. Bodineau, I. Gallagher, L. Saint-Raymond, S. Simonella]

par **François Golse**

Introduction

On doit à Maxwell (1860, 1867) et Boltzmann (1872, 1964) la fondation de la théorie cinétique des gaz. Quoique l'idée d'une description particulaire de la matière remonte à l'Antiquité, Maxwell est le premier à faire appel à des notions de statistique afin de mettre cette description en équation. Le caractère novateur de cette entreprise apparaît très clairement à la lecture de l'introduction de (Maxwell, 1867) : la liste exhaustive des précurseurs cités par Maxwell est brève — Lucrèce [(1)] (Ier siècle av. J.-C.), D. Bernoulli (1738), Le Sage (1761), Herapath (1847), avant les travaux évidemment fondamentaux de Joule (1848) et Clausius (1857). Mais ce n'est qu'au XXe siècle que l'on comprend que la théorie cinétique des gaz peut être déduite rigoureusement des équations de la mécanique classique appliquées à un système de particules sphériques identiques interagissant lors de collisions élastiques dans une certaine limite asymptotique. Hilbert est le premier à avoir formulé cette question comme un problème mathématique (cité dans son sixième problème sur l'« axiomatisation de la physique » (Hilbert, 1902)). Mais le texte fondamental identifiant clairement le régime asymptotique à considérer est l'article (Grad, 1949). Après les articles précurseurs (Gallavotti, 1969) et (Cercignani, 1972), Lanford (1975) donne la première justification rigoureuse de l'équation de Boltzmann comme conséquence du principe fondamental de la dynamique (c'est-à-dire la deuxième loi de Newton) appliqué à chaque molécule de gaz.

[(1)] *Cf.* par exemple « Nam quoniam per inane uagantur, cuncta necessest | Aut gravitate sua ferri primordia rerum | Aut ictu forte alterius. Nam < cum > cita saepe | Obuia conflixere, fit ut diuersa repente | Dissiliant [...] » (Car puisqu'il errent à travers le vide, il faut que les principes des choses soient tous emportés soit par leur propre poids, soit encore par le choc d'un autre atome). De Rerum Natura II, v. 83–87, trad. A. Ernout, Les Belles Lettres, Paris, 2009. Ces vers décrivent précisément le processus physique sous-jacent à l'équation de Boltzmann (19) ci-dessous. Voici ce qu'en dit Maxwell (1867) : « [...] he describes the atoms as all moving downwards with equal velocities, which, at quite uncertain times and places, suffer an imperceptible change, just enough to allow of occasional collisions taking place between the atoms. »

Le théorème de Lanford a été étendu dans diverses directions depuis 1975 : voir par exemple (Ayi, 2017 ; Gallagher, Saint-Raymond et Texier, 2013 ; Illner et Pulvirenti, 1989 ; King, 1975). Toutefois, un certain nombre de questions essentielles demeuraient ouvertes même après les plus récents de ces travaux. Ces questions sont principalement de deux types bien distincts. D'une part, sur le plan mathématique, le théorème de Lanford exprime qu'une certaine quantité converge vers une solution de l'équation de Boltzmann. Peut-on alors estimer l'erreur entre la limite, c'est-à-dire la solution de l'équation de Boltzmann, et la quantité qui l'approche ? D'autre part, la convergence démontrée par Lanford a lieu sur un intervalle de temps assez restreint — on trouvera une discussion assez détaillée de ce point par Lanford lui-même dans (Lanford, 1976) — alors que l'équation de Boltzmann possède des solutions globales en temps pour des données initiales extrêmement générales (voir (DiPerna et Lions, 1989), ainsi que l'exposé (Gérard, 1988) dans ce même séminaire). En pratique, les spécialistes de gaz raréfiés utilisent d'ailleurs l'équation de Boltzmann ou certains de ses avatars pour des simulations numériques sur des plages de temps beaucoup plus longues que celle prédite par (Lanford, 1975). Peut-on alors démontrer la validité de l'équation de Boltzmann sur des intervalles de temps arbitrairement longs ?

Il semble probable qu'une réponse complète à cette dernière question devrait mettre en jeu de nouvelles idées par rapport à celles de (Lanford, 1975) — voir toutefois la section 6.

En revanche, quant aux questions du premier type, des progrès spectaculaires ont été accomplis récemment dans une série de travaux importants (Bodineau, Gallagher, Saint-Raymond et Simonella, 2018, 2020a, 2023). Citons également les articles (Bodineau, Gallagher, Saint-Raymond et Simonella, 2020b, 2022c), qui étendent la théorie des fluctuations étudiée dans (Bodineau, Gallagher, Saint-Raymond et Simonella, 2018, 2020a, 2023) par delà le temps de Lanford, mais dans une situation particulière, celle des fluctuations autour d'un état d'équilibre. On évoquera brièvement ces derniers travaux dans la section 6 de cet exposé.

Il est évidemment impossible de rendre compte en quelques dizaines de pages de l'ensemble des résultats obtenus et du détail des méthodes mathématiques employées dans ces articles, qui sont malheureusement d'une grande technicité compte tenu de la difficulté du problème.

Toutefois, dans les contributions de Bodineau, Gallagher, Saint-Raymond et Simonella au problème de la justification rigoureuse de la théorie cinétique des gaz, un rôle essentiel semble dévolu à une nouvelle équation de type Hamilton–Jacobi « fonctionnelle », équation satisfaite par une notion de « fonction génératrice des cumulants » dans la même limite que celle étudiée par Lanford.

C'est donc sur cette équation de Hamilton–Jacobi, et sur ses applications à la description statistique de la dynamique des gaz, que l'on a choisi de centrer cet exposé. On trouvera une présentation sensiblement différente de ces mêmes travaux dans la

conférence plénière de Laure Saint-Raymond au Congrès international des mathématiciens (2022) : voir (Bodineau, Gallagher, Saint-Raymond et Simonella, 2022b) — présentation dont on s'est toutefois inspiré ici pour évoquer certaines des méthodes de démonstration utilisées dans les travaux cités plus haut.

Les questions étudiées dans le présent exposé font évidemment appel à plusieurs notions fondamentales relatives à la justification rigoureuse par Lanford de l'équation de Boltzmann, dont un compte-rendu très précis et détaillé se trouve dans (Gallagher, Saint-Raymond et Texier, 2013) (voir également (Cercignani, Illner et Pulvirenti, 1994)), et qui a déjà fait l'objet d'une présentation à ce même séminaire (Golse, 2014). On a essayé autant que possible d'éviter les redites entre ce précédent exposé et le texte qui va suivre. Un certain nombre de questions, comme par exemple les « paradoxes » liés à l'irréversibilité, ont déjà été décrits et commentés dans (Golse, 2014) ; on a délibérément choisi de ne pas y revenir, et d'y renvoyer le lecteur chaque fois que cela était possible.

Je tiens à remercier Thierry Bodineau, Isabelle Gallagher, Laure Saint-Raymond et Sergio Simonella de m'avoir communiqué une première version de leur article (Bodineau, Gallagher, Saint-Raymond et Simonella, 2022b), ainsi que de leurs suggestions pendant la préparation de cet exposé.

1. Limite de Boltzmann–Grad et théorème de Lanford

Commençons par rappeler dans cette section quelques notions de base, déjà présentées dans (Golse, 2014), mais indispensables pour la suite.

1.1. La dynamique moléculaire

Considérons un gaz monoatomique, que l'on voit comme un système de N molécules qui sont des boules de diamètre $\varepsilon \in]0, \frac{1}{2}[$. La position et la vitesse de la k-ième boule à l'instant t sont notées respectivement $x_k(t)$ et $v_k(t) \in \mathbf{R}^3$ où $k = 1, \dots, N$. Dans toute la suite, on supposera pour simplifier que l'évolution du gaz est spatialement périodique, de sorte que $x_k(t) \in \mathbf{T}^3 = \mathbf{R}^3/\mathbf{Z}^3$. En l'absence de force extérieure agissant sur les molécules, la deuxième loi de Newton écrite pour chaque molécule de gaz est [(2)]

$$\frac{dx_k}{dt}(t) = v_k(t)\,, \quad \frac{dv_k}{dt}(t) = 0\,, \qquad \text{si } \operatorname{dist}(x_k(t), x_l(t)) > \varepsilon \text{ pour tout } l \neq k\,. \tag{1}$$

[(2)] Pour $x, y \in \mathbf{T}^3$, on note $\operatorname{dist}(x,y) = \min\{|X - Y| \text{ t.q. } X, Y \in \mathbf{R}^3\,,\ X = x \text{ et } Y = y \text{ mod. } \mathbf{Z}^3\}$.

Au cours d'une collision entre la k-ième et la l-ième molécule à un instant t^*, les positions de ces molécules varient continûment en temps, c'est-à-dire que

$$\mathrm{dist}(x_k(t^*-0), x_l(t^*-0)) = \varepsilon \implies x_k(t^*+0) = x_k(t^*-0) \quad \text{et} \quad x_l(t^*+0) = x_l(t^*-0)\,, \tag{2}$$

tandis que leurs vitesses varient de façon discontinue comme suit :

$$\begin{aligned} v_k(t^*+0) &= v_k(t^*-0) - ((v_k(t^*-0) - v_l(t^*-0)) \cdot n_{kl}(t^*))\, n_{kl}(t^*)\,, \\ v_l(t^*+0) &= v_l(t^*-0) + ((v_k(t^*-0) - v_l(t^*-0)) \cdot n_{kl}(t^*))\, n_{kl}(t^*)\,, \end{aligned} \tag{3}$$

en notant [3] $n_{kl}(t^*) := (x_l(t^* \pm 0) - x_k(t^* \pm 0))/\varepsilon$. On notera dans la suite de cet exposé

$$\Lambda_N^\varepsilon := \{(x_1, \dots, x_N) \in (\mathbf{T}^3)^N \text{ t.q. } \mathrm{dist}(x_k(t), x_l(t)) > \varepsilon \text{ pour } k, l = 1, \dots, N\,,\ k \neq l\}$$

— il s'agit de l'ensemble des positions physiquement admissibles pour les molécules, qui ne peuvent s'interpénétrer — et $\Gamma_N^\varepsilon := \Lambda_N^\varepsilon \times (\mathbf{R}^3)^N$, l'espace des phases à N particules. On suppose connues les positions et les vitesses de chaque molécule à l'instant initial $t = 0$, soit

$$x_k(0) = x_k^{in}\,, \quad v_k(0) = v_k^{in}\,, \qquad\qquad k = 1, \dots, N \tag{4}$$

avec $(x_1^{in}, v_1^{in}, \dots, x_N^{in}, v_N^{in}) \in \Gamma_N^\varepsilon$, et on s'intéresse aux solutions

$$t \mapsto (x_1(t), v_1(t), \dots, x_N(t), v_N(t)) \in \Gamma_N^\varepsilon$$

de ce problème de Cauchy (1)-(2)-(3) avec la condition initiale (4). Notons m_N la mesure de Lebesgue sur $(\mathbf{T}^3 \times \mathbf{R}^3)^N$.

Proposition 1.1. *Soient $N \geqslant 2$ et $\varepsilon \in]0, \frac{1}{2}[$ fixés. Il existe $E \subset \overline{\Gamma_N^\varepsilon}$ tel que $m_N(E) = 0$ et vérifiant la propriété suivante : pour tout $(x_1^{in}, v_1^{in}, \dots, x_N^{in}, v_N^{in}) \in \overline{\Gamma_N^\varepsilon} \setminus E$, le problème de Cauchy (1)-(2)-(3)-(4) admet une unique solution*

$$t \mapsto (x_1(t), v_1(t), \dots, x_N(t), v_N(t)) =: S_t^{N,\varepsilon}(x_1^{in}, v_1^{in}, \dots, x_N^{in}, v_N^{in})$$

définie pour tout $t \in \mathbf{R}$. Ceci définit $S_t^{N,\varepsilon}$ comme flot sur $\overline{\Gamma_N^\varepsilon} \setminus E$: pour tout $t \in \mathbf{R}$, on a $S_t^{N,\varepsilon}(\overline{\Gamma_N^\varepsilon} \setminus E) \subset \overline{\Gamma_N^\varepsilon} \setminus E$ et $S_{t+s}^{N,\varepsilon} = S_t^{N,\varepsilon} \circ S_s^{N,\varepsilon}$. D'autre part, la mesure m_N est invariante sous l'action de $S_t^{N,\varepsilon}$, c'est-à-dire que $m_N(S_t^{N,\varepsilon}(A)) = m_N(A)$ pour tout $A \subset \overline{\Gamma_N^\varepsilon}$ mesurable et tout $t \in \mathbf{R}$.

Voir (Alexander, 1976), ou encore le chapitre 4 (en particulier la Proposition 4.1.1) de (Gallagher, Saint-Raymond et Texier, 2013). Ce résultat est décrit avec un peu plus de détails dans la section 3.1 de (Golse, 2014).

Dans la suite, il sera commode de noter $z_k^{in} := (x_k^{in}, v_k^{in})$ et $z_k(t) := (x_k(t), v_k(t))$.

[3] Soient $x, y \in \mathbf{T}^3$ tels que $r = \mathrm{dist}(x, y) < \frac{1}{2}$. Il existe un unique vecteur unitaire n dans $\mathbf{R}^3$ tel que $y = x + rn$. Ce vecteur sera noté $n = (y - x)/r$ ou $(y - x)/|y - x|$ dans la suite de cet exposé.

1.2. Ensemble grand-canonique et loi d'échelle de Boltzmann–Grad

La théorie cinétique des gaz est obtenue à partir du système (1)-(2)-(3) dans un régime asymptotique très particulier, connu sous le nom de loi d'échelle de Boltzmann–Grad. On y suppose que le diamètre des molécules ε est très petit par rapport au diamètre du domaine $\mathbf{T}^3$ où elles sont confinées, tandis que leur nombre N est très grand. La quantité ε^2 est donc homogène à une surface — $\frac{1}{4}\pi\varepsilon^2$ est la surface de la section équatoriale d'une molécule — de sorte que la quantité $\ell := \mathrm{Vol}(\mathbf{T}^3)/(N\varepsilon^2)$ est homogène à une longueur, proportionnelle au libre parcours moyen — c'est-à-dire à la longueur moyenne séparant deux collisions subies par la même molécule typique dans le gaz. La loi d'échelle de Boltzmann–Grad postule que cette longueur est du même ordre de grandeur que le diamètre du domaine spatial $\mathbf{T}^3$. En particulier, le volume $(N-1)\frac{4}{3}\pi\varepsilon^3$ de l'espace dans lequel aucun des points $x_k(t)$ ne peut pénétrer (dit « volume exclu », rempli par les $N-1$ autres molécules) est $O(\varepsilon)$ et donc négligeable dans la limite de Boltzmann–Grad. C'est pourquoi la théorie cinétique des gaz de Maxwell et Boltzmann, obtenue dans cette limite, ne peut décrire que des gaz parfaits.

Contrairement au cadre considéré dans (Cercignani, Illner et Pulvirenti, 1994 ; Gallagher, Saint-Raymond et Texier, 2013 ; Lanford, 1975) ainsi que dans (Golse, 2014), où N est un paramètre entier que l'on fait tendre vers l'infini [(4)], on supposera ici que N, ainsi que les positions et les vitesses initiales des N molécules sont des variables aléatoires. Soit une suite $(\phi_n)_{n\geqslant 0}$ où $\phi_n \in C_b((\mathbf{T}^3\times\mathbf{R}^3)^n)$ pour tout $n \geqslant 1$, et où $\phi_0 \in \mathbf{R}$. De façon équivalente, on considère la fonction définie sur l'espace grand-canonique Ω par

$$\Phi : \Omega := \bigcup_{n\geqslant 0} \{n\}\times(\mathbf{T}^3\times\mathbf{R}^3)^n \ni (N, z_1, \dots, z_N) \mapsto \sum_{n\geqslant 0} \mathbf{1}_{N=n}\phi_n(z_1,\dots,z_n)\,, \quad (5)$$

et on définit sa moyenne grand-canonique par la formule

$$\mathbb{E}_\varepsilon(\Phi) := \frac{1}{\mathscr{Z}_\varepsilon}\sum_{n\geqslant 0}\frac{\mu_\varepsilon^n}{n!}\int_{(\mathbf{T}^3\times\mathbf{R}^3)^n}\phi_n(Z_n)\mathbb{F}_n^{in}(Z_n)dZ_n\,,$$

où on a noté $Z_n := (z_1,\dots,z_n)$, où $\mu_\varepsilon > 0$ est un paramètre qui sera précisé plus loin, et où

$$\mathbb{F}_n^{in}(Z_n) := \prod_{i=1}^n f^{in}(z_i)\prod_{1\leqslant j<k\leqslant n}\mathbf{1}_{\mathrm{dist}(x_j,x_k)>\varepsilon}\,. \quad (6)$$

On définit de la sorte une mesure de probabilité borélienne $\mathbb{P}_\varepsilon$ sur Ω, et le terme « ensemble grand-canonique » désigne le couple $(\Omega, \mathbb{P}_\varepsilon)$.

[(4)] Formalisme dit de l'« ensemble canonique ». Dans (Cercignani, Illner et Pulvirenti, 1994 ; Gallagher, Saint-Raymond et Texier, 2013 ; Golse, 2014 ; Lanford, 1975), seules les positions et les vitesses initiales des N molécules sont distribuées aléatoirement.

Dans cette formule, f^{in} est la fonction de distribution (à une molécule) qui joue le rôle de condition initiale dans l'équation de Boltzmann. Sans perte de généralité, on supposera que f^{in} est une densité de probabilité sur l'espace des phases $\mathbf{T}^3 \times \mathbf{R}^3$. Notons que

$$\tfrac{4}{3}\pi n\varepsilon^3 > 1 \implies \prod_{1\leqslant j<k\leqslant n} \mathbf{1}_{\mathrm{dist}(x_j,x_k)>\varepsilon} = 0\,,$$

de sorte que la somme définissant $\mathbb{E}_\varepsilon(\Phi)$ ne comporte qu'un nombre fini de termes non nuls. Le réel $\mathscr{Z}_\varepsilon$ (nommé « fonction de partition ») est défini par la condition de normalisation $\mathbb{E}_\varepsilon(1) = 1$:

$$\mathscr{Z}_\varepsilon := \sum_{n\geqslant 0} \frac{\mu_\varepsilon^n}{n!} \int_{(\mathbf{T}^3\times\mathbf{R}^3)^n} \mathbb{F}_n^{in}(Z_n)dZ_n\,.$$

Lemme 1.2. *Supposons que la densité de probabilité f^{in} appartient à l'espace $L^\infty(\mathbf{T}^3; L^1(\mathbf{R}^3))$ et que $\varepsilon^3\mu_\varepsilon \to 0$ quand $\varepsilon \to 0^+$. Alors*

$$\mathbb{E}_\varepsilon(N) \sim \mu_\varepsilon \qquad \textit{lorsque } \varepsilon \to 0^+\,,$$

où N désigne la fonction Φ sur l'espace grand-canonique Ω associée par la formule (5) *à la suite de fonctions $(\phi_n)_{n\geqslant 0}$ telle que $\phi_n(z_1,\dots,z_n) = n$ pour tout $n \geqslant 0$ et tous $z_1,\dots,z_n \in \mathbf{T}^3 \times \mathbf{R}^3$.*

Démonstration. Posons $C := \|f^{in}\|_{L^\infty(\mathbf{T}^3;L^1(\mathbf{R}^3))}$, et

$$\begin{aligned} u_n &:= \int_{(\mathbf{T}^3\times\mathbf{R}^3)^n} \mathbb{F}_n^{in}(Z_n)dZ_n \\ &\leqslant \int_{(\mathbf{T}^3\times\mathbf{R}^3)^n} \mathbb{F}_{n-1}^{in}(Z_{n-1})dZ_{n-1} \int_{\mathbf{T}^3\times\mathbf{R}^3} f^{in}(z_n)dz_n \quad \leqslant u_{n-1} \leqslant 1\,. \end{aligned}$$

D'autre part

$$\begin{aligned} u_n &\geqslant \int_{(\mathbf{T}^3\times\mathbf{R}^3)^{n-1}} \left(\int_{\mathbf{T}^3\times\mathbf{R}^3} \left(1 - \sum_{j=1}^{n-1} \mathbf{1}_{|x_n-x_j|\leqslant\varepsilon} \right) f^{in}(z_n)dz_n \right) \mathbb{F}_{n-1}^{in}(Z_{n-1})dZ_{n-1} \\ &\geqslant u_{n-1}(1 - \tfrac{4}{3}\pi C(n-1)\varepsilon^3)\,. \end{aligned}$$

Donc

$$\mathbb{E}_\varepsilon(N) \leqslant \frac{1}{\mathscr{Z}_\varepsilon} \sum_{n\geqslant 1} \frac{\mu_\varepsilon^n}{n!} n = \frac{\mu_\varepsilon}{\mathscr{Z}_\varepsilon} \sum_{n\geqslant 1} \frac{\mu_\varepsilon^{n-1}}{(n-1)!} = \mu_\varepsilon\,,$$

tandis que, comme $u_{n-1} \leqslant u_{n-2}$

$$\begin{aligned} \mathbb{E}_\varepsilon(N) &\geqslant \frac{\mu_\varepsilon}{\mathscr{Z}_\varepsilon} \sum_{n\geqslant 0} \frac{\mu_\varepsilon^{n-1}}{n!} n u_{n-1}(1 - \tfrac{4}{3}\pi C(n-1)\varepsilon^3) \\ &\geqslant \frac{\mu_\varepsilon}{\mathscr{Z}_\varepsilon} \sum_{n\geqslant 1} \frac{\mu_\varepsilon^{n-1}u_{n-1}}{(n-1)!} - \tfrac{4}{3}\pi C\varepsilon^3 \frac{\mu_\varepsilon^2}{\mathscr{Z}_\varepsilon} \sum_{n\geqslant 2} \frac{\mu_\varepsilon^{n-2}u_{n-2}}{(n-2)!} = \mu_\varepsilon(1 - \tfrac{4}{3}\pi C\varepsilon^3\mu_\varepsilon)\,. \end{aligned}$$

□

Remarque 1.3. En supposant que $\mu_\varepsilon \to +\infty$ lorsque $\varepsilon \to 0^+$ et sous les mêmes hypothèses que dans le Lemme 1.2, on a $\mathscr{Z}_\varepsilon \sim e^{\mu_\varepsilon}$. La démonstration de ce fait, laissée au lecteur, suit de près celle du lemme.

Dans le formalisme grand-canonique, on réalisera donc la loi d'échelle de Boltzmann–Grad en posant

$$\mu_\varepsilon = \varepsilon^{-2}\,, \quad \text{de sorte que } \varepsilon^2 \mathbb{E}_\varepsilon(N) \to 1 \text{ lorsque } \varepsilon \to 0^+\,.$$

Observons d'ailleurs que

$$\mathbb{P}_\varepsilon(\{N=n\}) = \mathbb{E}_\varepsilon(\mathbf{1}_{N=n}) = \mathscr{Z}_\varepsilon^{-1} \frac{\mu_\varepsilon^n}{n!} \int_{(\mathbf{T}^3\times\mathbf{R}^3)^n} \mathbb{F}_n^{in}(Z_n) dZ_n\,,$$

formule qui évoque bien sûr la loi de Poisson. (Si on fait $\varepsilon = 0$ dans cette formule, et que l'on pose $\mu_0 = \lambda$ — ce qui contredit évidemment la loi d'échelle de Boltzmann–Grad — on trouve en effet que $\mathscr{Z}_0 = e^\lambda$, puis que $\mathbb{P}_0(\{N=n\}) = e^{-\lambda}\frac{\lambda^n}{n!}$, ce qui est la définition de la loi de Poisson de paramètre λ).

Le lecteur familier des énoncés usuels du théorème de Lanford dans (Cercignani, Illner et Pulvirenti, 1994) ou (Gallagher, Saint-Raymond et Texier, 2013) pourra être surpris du choix de l'ensemble grand-canonique dans cette étude. En effet, le processus collisionnel considéré ici, à savoir des collisions binaires entre sphères dures, laisse invariant le nombre de molécules. N'est-il donc pas artificiel de considérer le nombre de molécules comme aléatoire ? En réalité, le fait de prescrire le nombre total de molécules introduit nécessairement une corrélation entre ces molécules qui ne provient ni de la dynamique, ni du choix de la condition initiale. Le formalisme grand-canonique permet précisément d'éviter cela (voir la remarque (6) dans la section 2.4.1 de (Pulvirenti et Simonella, 2017)).

1.3. Mesure(s) empirique(s) et corrélations

Pour tout $\varepsilon > 0$, posons

$$\rho^\varepsilon[N, Z_N] := \frac{1}{\mu_\varepsilon} \sum_{j=1}^N \delta_{z_j}\,, \qquad Z_N := (z_1, \dots, z_N) \in \Gamma_N^\varepsilon\,.$$

Il s'agit d'une fonction sur l'espace grand-canonique Ω à valeurs dans l'espace des mesures de Radon sur $\mathbf{T}^3 \times \mathbf{R}^3$. Elle vérifie $\rho^\varepsilon[N, Z_N] \geqslant 0$ et $\|\rho^\varepsilon[N, Z_N]\|_{VT} = N/\mu_\varepsilon$. Pour tout $k \geqslant 1$, on pose de même

$$\rho_k^\varepsilon[N, Z_N] := \frac{1}{\mu_\varepsilon^k} \sum_{\substack{j:\{1,\dots,k\}\to\{1,\dots,N\} \\ \text{injective}}} \delta_{z_{j(1)}} \otimes \dots \otimes \delta_{z_{j(k)}} \geqslant 0\,,$$

fonction définie sur Ω à valeurs dans l'espace des mesures de Radon sur $(\mathbf{T}^3 \times \mathbf{R}^3)^k$. Elle vérifie $\rho_k^\varepsilon \geqslant 0$ ainsi que $\|\rho_k^\varepsilon[N, Z_N]\|_{VT} = N(N-1)\dots(N-k+1)/\mu_\varepsilon^k$. On notera enfin

$$\rho_t^\varepsilon := \rho^\varepsilon[N, S_t^{N,\varepsilon} Z_N]\,, \qquad \rho_{k,t}^\varepsilon := \rho_k^\varepsilon[N, S_t^{N,\varepsilon} Z_N]\,. \tag{7}$$

À partir de là, on définit la suite des corrélations $F_k^\varepsilon(t,\cdot)$ entre k molécules à l'instant t pour l'ensemble grand-canonique par la formule

$$\int_{(\mathbf{T}^3\times\mathbf{R}^3)^k} F_k^\varepsilon(t,Z_k)h_k(Z_k)dZ_k := \mathbb{E}_\varepsilon(\langle \rho_{k,t}^\varepsilon, h_k\rangle)\,, \quad k\geqslant 1\,, \tag{8}$$

pour toute fonction test $h_k \in C_b((\mathbf{T}^3\times\mathbf{R}^3)^k)$ symétrique en ses k variables.

Ce formalisme diffère de la présentation de la « hiérarchie BBGKY » adoptée dans (Golse, 2014). Posons

$$\mathbb{F}_{n:k}(t,Z_k) := \int_{(\mathbf{T}^3\times\mathbf{R}^3)^{n-k}} \mathbb{F}_n^{in}(S_{-t}^{n,\varepsilon}Z_n)dz_{k+1}\dots dz_n\,.$$

Cette formule vaut pour $n>k\geqslant 0$; on convient par ailleurs que $\mathbb{F}_{n:n}=\mathbb{F}_n=\mathbb{F}_n^{in}\circ S_{-t}^{n,\varepsilon}$ et que $\mathbb{F}_{n:k}=0$ lorsque $k>n$.

Une remarque importante s'impose : la fonction de distribution à n corps initiale $\mathbb{F}_n^{in}$ est évidemment symétrique en les variables $z_1,\dots,z_n$ comme le montre la formule (6), et cette symétrie est propagée par la dynamique $S_t^{n,\varepsilon}$, de sorte que $\mathbb{F}_n(t,\cdot)$ est également symétrique en les variables $z_1,\dots,z_n$. Du point de vue de la physique, cette symétrie traduit le fait que les molécules sont indistinguables.

Le point de vue de la hiérarchie BBGKY décrit dans (Golse, 2014) consistait à étudier $\mathbb{F}_{n:1}$ dans la limite où $n\to+\infty$ avec $\varepsilon\to 0^+$ vérifiant la condition de Boltzmann–Grad $n\varepsilon^2=1$. La traduction entre ce formalisme et celui de l'ensemble grand-canonique considéré ici découle du lemme ci-dessous.

Lemme 1.4. *Pour tout $\varepsilon>0$ et tout $t\geqslant 0$,*

$$F_k^\varepsilon(t,\cdot)=\frac{1}{\mathscr{Z}_\varepsilon}\sum_{n\geqslant k}\frac{\mu_\varepsilon^{n-k}}{(n-k)!}\mathbb{F}_{n:k}(t,\cdot)\,,\qquad k\geqslant 1\,.$$

Démonstration. En effet, pour toute fonction test $h_k\in C_b((\mathbf{T}^3\times\mathbf{R}^3)^k)$ symétrique,

$$\begin{aligned}
\mathbb{E}_\varepsilon(\langle\rho_{k,t}^\varepsilon,h_k\rangle) &= \frac{1}{\mathscr{Z}_\varepsilon}\sum_{n\geqslant k}\frac{\mu_\varepsilon^n}{n!}\int_{(\mathbf{T}^3\times\mathbf{R}^3)^{n-k}}\langle\rho_k^\varepsilon[n,S_t^{n,\varepsilon}Z_n],h_k\rangle\mathbb{F}_n^{in}(Z_n)dZ_n\\
&= \frac{1}{\mathscr{Z}_\varepsilon}\sum_{n\geqslant k}\frac{\mu_\varepsilon^n}{n!}\int_{(\mathbf{T}^3\times\mathbf{R}^3)^{n-k}}\langle\rho_k^\varepsilon[n,Z_n],h_k\rangle\mathbb{F}_n^{in}(S_{-t}^{n,\varepsilon}Z_n)dZ_n\\
&= \frac{1}{\mathscr{Z}_\varepsilon}\sum_{n\geqslant k}\frac{\mu_\varepsilon^{n-k}}{n!}\int_{(\mathbf{T}^3\times\mathbf{R}^3)^{n-k}}\sum_{\substack{j:\{1,\dots,k\}\to\{1,\dots,N\}\\ \text{injective}}}h_k(z_{j(1)},\dots,z_{j(k)})\mathbb{F}_n^{in}(S_{-t}^{n,\varepsilon}Z_n)dZ_n\\
&= \frac{1}{\mathscr{Z}_\varepsilon}\sum_{n\geqslant k}\frac{\mu_\varepsilon^{n-k}}{n!}n(n-1)\dots(n-k+1)\int_{(\mathbf{T}^3\times\mathbf{R}^3)^{n-k}}h_k(z_1,\dots,z_k)\mathbb{F}_n^{in}(S_{-t}^{n,\varepsilon}Z_n)dZ_n\\
&= \frac{1}{\mathscr{Z}_\varepsilon}\sum_{n\geqslant k}\frac{\mu_\varepsilon^{n-k}}{(n-k)!}\int_{(\mathbf{T}^3\times\mathbf{R}^3)^{n-k}}h_k(z_1,\dots,z_k)\mathbb{F}_{n:k}(t,Z_n)dZ_n\,.
\end{aligned}$$

La mesure de Lebesgue sur Γ_n^ε est invariante par $S_t^{n,\varepsilon}$, d'où la seconde égalité. La quatrième égalité découle de la symétrie des fonctions h_k et $\mathbb{F}_n(t,\cdot)$. □

1.4. Équation de Liouville et corrélations

La formule $\mathbb{F}_n(t,Z_n) = \mathbb{F}_n^{in}(S_{-t}^{\varepsilon,n}Z_n)$ montre que $\mathbb{F}_n$ est constante sur les courbes intégrales de (1)-(2)-(3), d'où l'on déduit l'équation de Liouville

$$\begin{aligned} &\partial_t \mathbb{F}_n(t,Z_n) + \sum_{i=1}^n v_i \cdot \nabla_{x_i} \mathbb{F}_n(t,Z_n) = 0\,, \qquad Z_n \in \Gamma_n^\varepsilon\,, \\ &\mathbb{F}_n(t,Z_n) = \mathbb{F}_n(t,\hat{Z}_n[i,j]) \qquad \text{si } \mathrm{dist}(x_i(t),x_j(t)) = \varepsilon\,, \end{aligned} \tag{9}$$

où

$$\hat{Z}_n[i,j] := (x_1,v_1,\dots,x_i,v_i',\dots,x_j,v_j',\dots,x_n,v_n)\,, \qquad \text{lorsque } \mathrm{dist}(x_i(t),x_j(t)) = \varepsilon\,,$$
$$v_i' := v_i - ((v_i-v_j)\cdot n_{ji})n_{ji}\,, \quad v_j' := v_j + ((v_i-v_j)\cdot n_{ji})n_{ji}\,, \qquad \text{où } n_{ji} := \tfrac{x_i-x_j}{\varepsilon}\,.$$

En particulier, la condition aux limites ajoutée à l'équation de Liouville découle de (2)-(3). Il sera commode de remplacer le problème aux limites ci-dessus pour l'équation de Liouville, posé sur le domaine Γ_n^ε, par une équation au sens des distributions sur $(\mathbf{T}^3\times\mathbf{R}^3)^n$. Par définition $\mathbb{F}_n(t,Z_n) = 0$ dans $(\mathbf{T}^3\times\mathbf{R}^3)^n \setminus \Gamma_n^\varepsilon$, de sorte que

$$\begin{aligned} \partial_t \mathbb{F}_n(t,Z_n) + \sum_{i=1}^n v_i\cdot\nabla_{x_i}\mathbb{F}_n(t,Z_n) &= \sum_{1\leqslant i<j\leqslant n} \mathbb{F}_n\big|_{\partial^+\Gamma_n^\varepsilon}(v_j-v_i)\cdot n_{ij}\delta_{\mathrm{dist}(x_i,x_j)=\varepsilon} \\ &= \sum_{1\leqslant i<j\leqslant n} \mathbb{F}_n\big|_{\partial^+\Gamma_n^\varepsilon}((v_j-v_i)\cdot n_{ij})_+\delta_{\mathrm{dist}(x_i,x_j)=\varepsilon} \\ &\quad - \sum_{1\leqslant i<j\leqslant n} \mathbb{F}_n\big|_{\partial^+\Gamma_n^\varepsilon}((v_j-v_i)\cdot n_{ij})_-\delta_{\mathrm{dist}(x_i,x_j)=\varepsilon} \end{aligned} \tag{10}$$

au sens des distributions sur $(\mathbf{T}^3\times\mathbf{R}^3)^n$. Pour tout réel r, on pose $r_+ = \max(r,0)$ et $r_- = \max(-r,0)$, tandis que la trace interne de $\mathbb{F}_n$ sur $\partial\Gamma_n^\varepsilon$ est notée

$$\mathbb{F}_n\big|_{\partial^+\Gamma_n^\varepsilon} := \lim_{\eta\to 0^+} \mathbb{F}_n\big|_{\partial\Gamma_n^{\varepsilon+\eta}}\,.$$

Enfin $\delta_{\mathrm{dist}(x_i,x_j)=\varepsilon}$ désigne la distribution de simple couche de densité 1 portée par l'hypersurface de $(\mathbf{T}^3)^n$ d'équation $\mathrm{dist}(x_i,x_j)=\varepsilon$. (Pour la justification de (10), voir la formule (20) de (Golse, 2014), ainsi que la formule (II.3.1) de (Schwartz, 1966)).

On utilise alors la condition aux limites de (9) pour exprimer $\mathbb{F}_n\big|_{\partial^+\Gamma_n^\varepsilon}$ aux endroits où $(v_j - v_i)\cdot n_{ij} > 0$ (précaution absolument essentielle, comme on le verra) :

$$\begin{aligned}
&\partial_t \mathbb{F}_n(t,Z_n) + \sum_{i=1}^n v_i\cdot\nabla_{x_i}\mathbb{F}_n(t,Z_n)\\
&= \sum_{1\leqslant i<j\leqslant n} \mathbb{F}_n(t,\hat{Z}_n[i,j])\big|_{\partial^+\Gamma_n^\varepsilon}((v_j-v_i)\cdot n_{ij})_+\delta_{\mathrm{dist}(x_i,x_j)=\varepsilon}\\
&\quad - \sum_{1\leqslant i<j\leqslant n} \mathbb{F}_n(t,Z_n)\big|_{\partial^+\Gamma_n^\varepsilon}((v_j-v_i)\cdot n_{ij})_-\delta_{\mathrm{dist}(x_i,x_j)=\varepsilon}\,.
\end{aligned} \tag{11}$$

En intégrant chaque membre de cette égalité par rapport aux variables $z_2,\dots,z_n$, on aboutit à

$$\partial_t \mathbb{F}_{n:1}(t,z_1) + v_1\cdot\nabla_{x_1}\mathbb{F}_{n:1}(t,z_1) = (n-1)\varepsilon^2\mathscr{B}_\varepsilon^{12}(\mathbb{F}_{n:2})(t,z_1) \tag{12}$$

où

$$\begin{aligned}
\mathscr{B}_\varepsilon^{12}(\mathbb{F}_{n:2})(t,z_1) := &\iint_{\mathbf{R}^3\times\mathbf{S}^2}\mathbb{F}_{n:2}(t,x_1,v_1',x_1-\varepsilon\omega,v_2')((v_1-v_2)\cdot\omega)_+dv_2d\omega\\
&-\iint_{\mathbf{R}^3\times\mathbf{S}^2}\mathbb{F}_{n:2}(t,x_1,v_1,x_1+\varepsilon\omega,v_2)((v_1-v_2)\cdot\omega)_+dv_2d\omega\,,
\end{aligned}$$

où on a posé

$$v_1' := v_1 - ((v_1-v_2)\cdot\omega)\omega\,,\qquad v_2' := v_2 + ((v_1-v_2)\cdot\omega)\omega\,.$$

On renvoie le lecteur à (Golse, 2014), tout particulièrement à l'argument permettant de passer de (10), c'est-à-dire de l'égalité (20) de (Golse, 2014), à (12), autrement dit à l'égalité (22) et à la formule (23) de (Golse, 2014).

Multiplions maintenant chaque membre de (12) par $\mu_\varepsilon^{n-1}/(n-1)!$, et sommons les expressions ainsi obtenues pour $n\geqslant 1$. Comme $\varepsilon^2\mu_\varepsilon = 1$, il vient

$$\begin{aligned}
(\partial_t + v_1\cdot\nabla_{x_1})F_1^\varepsilon(t,z_1) =&(\partial_t + v_1\cdot\nabla_{x_1})\frac{1}{\mathscr{Z}_\varepsilon}\sum_{n\geqslant 1}\frac{\mu_\varepsilon^{n-1}}{(n-1)!}\mathbb{F}_{n:1}(t,z_1)\\
=&\frac{1}{\mathscr{Z}_\varepsilon}\sum_{n\geqslant 2}\frac{\mu_\varepsilon^{n-2}}{(n-2)!}\mathscr{B}_\varepsilon^{12}(\mathbb{F}_{n:2})(t,z_1) = \mathscr{B}_\varepsilon^{12}(F_2^\varepsilon)(t,z_1)\,.
\end{aligned} \tag{13}$$

1.5. L'hypothèse de chaos moléculaire et le théorème de Lanford

L'égalité (13) n'est pas vraiment une équation pour F_1^ε, puisqu'elle fait intervenir F_2^ε.

L'idée clé de Boltzmann lui permettant d'arriver à l'équation portant son nom est que deux molécules *sur le point d'entrer en collision* sont statistiquement indépendantes, hypothèse dite du « chaos moléculaire » (voir par exemple les sections 8 et

11 de (Grad, 1958)). Évidemment, deux molécules *venant juste d'entrer en collision* ne peuvent pas être statistiquement indépendantes. Considérons alors l'intégrande du terme $\mathscr{B}_\varepsilon^{12}(F_2^\varepsilon)$, à savoir

$$(F_2^\varepsilon(t,x_1,v_1',x_1-\varepsilon\omega,v_2')-F_2^\varepsilon(t,x_1,v_1,x_1+\varepsilon\omega,v_2))((v_1-v_2)\cdot\omega)_+\,.$$

Dans le second terme de cette différence, on a

$$(v_1-v_2)\cdot\omega=-\tfrac{1}{\varepsilon}(v_1-v_2)\cdot(x_1-x_2)>0\,,$$

de sorte que la molécule située en x_1 de vitesse v_1 s'approche de la molécule située en $x_2=x_1+\varepsilon\omega$ et de vitesse v_2 (voir la figure 1). Comme ces deux molécules sont sur le point d'entrer en collision, l'hypothèse du chaos moléculaire entraîne que

$$F_2^\varepsilon(t,x_1,v_1,x_1+\varepsilon\omega,v_2)\mathbf{1}_{(v_1-v_2)\cdot\omega>0}\simeq F_1(t,x_1,v_1)F_1(t,x_1,v_2)\mathbf{1}_{(v_1-v_2)\cdot\omega>0}\,,$$

où

$$F_1(t,x_1,v_1):=\lim_{\varepsilon\to0^+}F_1^\varepsilon(t,x_1,v_1)\,.$$

De même, dans le premier terme de cette différence, on a

$$(v_1-v_2)\cdot\omega=\tfrac{1}{\varepsilon}(v_1-v_2)\cdot(x_1-x_2)=-\tfrac{1}{\varepsilon}(v_1'-v_2')\cdot(x_1-x_2)>0\,,$$

de sorte que la molécule située en x_1 de vitesse v_1' s'approche de la molécule située en $x_2=x_1-\varepsilon\omega$ et de vitesse v_2' (voir la figure 1). Ces deux molécules sont donc elles aussi sur le point d'entrer en collision, et l'hypothèse du chaos moléculaire entraîne que

$$F_2^\varepsilon(t,x_1,v_1',x_1-\varepsilon\omega,v_2')\mathbf{1}_{(v_1-v_2)\cdot\omega>0}\simeq F_1(t,x_1,v_1')F_1(t,x_1,v_2')\mathbf{1}_{(v_1-v_2)\cdot\omega>0}\,.$$

En passant formellement à la limite dans (13), et en tenant compte des implications de l'hypothèse du chaos moléculaire de Boltzmann mentionnées ci-dessus, on trouve que F_1 est solution de l'équation de Boltzmann

$$(\partial_t+v\cdot\nabla_x)F_1(t,z)=\mathscr{B}(F_1)(t,z)\,,\tag{14}$$

où $\mathscr{B}(F_1)$ est l'intégrale de collision de Boltzmann, soit

$$\mathscr{B}(F_1)(t,z):=\int_{\mathbf{R}^3\times\mathbf{S}^2}(F_1(t,x,v')F_1(t,x,v_*')-F_1(t,x,v)F_1(t,x,v_*))((v-v_*)\cdot\omega)_+dv_*d\omega,\tag{15}$$

avec la notation

$$v'\equiv v'(v,v_*,\omega):=v-((v-v_*)\cdot\omega)\omega\,,\quad v_*'\equiv v_*'(v,v_*,\omega):=v_*+((v-v_*)\cdot\omega)\omega\,.\tag{16}$$

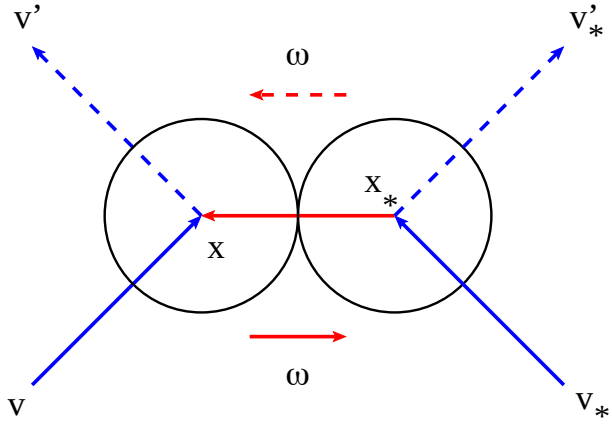


Figure 1 – Collision binaire. Avant la collision $(v - v_*) \cdot (x - x_*) < 0$, après collision $(v' - v'_*) \cdot (x - x_*) = -(v - v_*) \cdot (x - x_*) > 0$. On pose $\omega = -\frac{x - x_*}{|x - x_*|}$ avant collision, de sorte que $x_* = x + \varepsilon\omega$ et $(v - v_*) \cdot \omega > 0$, et $\omega = \frac{x - x_*}{|x - x_*|}$ après collision, de sorte que $x_* = x - \varepsilon\omega$ et $(v - v_*) \cdot \omega > 0$.

Le raisonnement ci-dessus montre tout l'intérêt d'avoir utilisé la condition aux limites de (9) dans la section précédente pour exprimer le terme $\mathbb{F}_n\big|_{\partial^+\Gamma_n^\varepsilon}$ là où $(v_j - v_i) \cdot n_{ij} > 0$. Si l'on néglige cette étape, on aboutit à l'équation

$$(\partial_t + v_1 \cdot \nabla_{x_1}) F_1^\varepsilon(t, z_1) = \int_{\mathbf{R}^3 \times \mathbf{S}^2} F_2^\varepsilon(t, x_1, v_1, x_1 - \varepsilon\omega, v_2)(v_1 - v_2) \cdot \omega dv_2 d\omega \,,$$

au lieu de (13). Si l'on suppose maintenant que

$$F_2^\varepsilon(t, x_1, v_1, x_1 - \varepsilon\omega, v_2) \simeq F_1(t, x_1, v_1) F_1(t, x_1, v_2)$$

dans l'intégrale au membre de droite *indépendamment du signe de* $(v_1 - v_2) \cdot \omega$, on trouve, en passant formellement à la limite dans l'équation ci-dessus, que

$$(\partial_t + v_1 \cdot \nabla_{x_1}) F_1(t, z_1) = \int_{\mathbf{R}^3} F_1(t, x_1, v_1) F_1(t, x_1, v_2) \left(\int_{\mathbf{S}^2} (v_1 - v_2) \cdot \omega d\omega \right) dv_2 = 0 \,.$$

Autrement dit, l'hypothèse de factorisation de $F_2^\varepsilon(t, x_1, v_1, x_1 - \varepsilon\omega, v_2)$ *indépendamment du signe de* $(v_1 - v_2) \cdot \omega$, et donc en particulier pour des molécules venant juste d'entrer en collision, par conséquent fortement corrélées, nous amènerait à la conclusion inintéressante — et expérimentalement fausse — que F_1 évolue suivant l'équation de transport libre (sans intégrale de collision).

Cette discussion montre la subtilité de l'hypothèse de Boltzmann, ainsi que la difficulté à la formaliser mathématiquement. La limite $F_1^\varepsilon(t, z_1) \to F_1(t, z_1)$ lorsque

$\varepsilon \to 0^+$ doit avoir lieu dans une topologie permettant de déduire que

$$F_2^\varepsilon(t,x_1,v_1,x_1+\varepsilon\omega,v_2)((v_1-v_2)\cdot\omega)_+ \to F_1(t,x_1,v_1)F_1(t,x_1,v_2)((v_1-v_2)\cdot\omega)_+\,,$$

mais pas que

$$F_2^\varepsilon(t,x_1,v_1,x_1-\varepsilon\omega,v_2)((v_1-v_2)\cdot\omega)_+ \to F_1(t,x_1,v_1)F_1(t,x_1,v_2)((v_1-v_2)\cdot\omega)_+$$

lorsque $\varepsilon \to 0^+$. Comme ces convergences ont lieu sur des ensembles de mesure nulle, la convergence p.p. pour la mesure de Lebesgue de $(\mathbf{T}^3\times\mathbf{R}^3)^2$ est insuffisante.

On renvoie le lecteur aux textes de Grad (1949, 1958) et à l'appendice A1 du livre (Sone, 2007) pour une analyse plus détaillée de ces questions.

Passons maintenant à l'énoncé du théorème de Lanford, qui est la justification rigoureuse de l'équation de Boltzmann (14) à partir du système (1)-(2)-(3).

Théorème 1.5 (Lanford). *Soit $f^{in}\in C^1(\mathbf{T}^3\times\mathbf{R}^3)$, densité de probabilité telle que*

$$f^{in}(x,v)+|\nabla_x f^{in}(x,v)|\leqslant C_0 e^{-\beta_0|v|^2}\,,\qquad x\in\mathbf{T}^3\,,\ v\in\mathbf{R}^3\,, \tag{17}$$

où $C_0,\beta_0>0$. Considérons la famille des fonctions de corrélations $(F_k^\varepsilon)_{k\geqslant 1,\varepsilon>0}$ définies par (8) *pour l'ensemble grand-canonique où $\mu_\varepsilon=\varepsilon^{-2}$, selon la loi d'échelle de Boltzmann–Grad. Alors, il existe $T_0=T[C_0,\beta_0]>0$ tel que, lorsque $\varepsilon\to 0^+$*
(1) $F_1^\varepsilon(t,\cdot)$ converge uniformément sur tout compact de $\mathbf{T}^3\times\mathbf{R}^3$ vers $f(t,\cdot)$ pour tout $t\in[0,T_0]$, où f est l'unique solution de l'équation de Boltzmann (14) *sur $\mathbf{T}^3\times\mathbf{R}^3$ vérifiant la condition initiale*

$$f(0,x,v)=f^{in}(x,v)\,,\qquad x\in\mathbf{T}^3\,,\ v\in\mathbf{R}^3\,; \tag{18}$$

(2) pour tous $k\geqslant 2$ et $t\in[0,T_0]$, la famille $F_k^\varepsilon(t,Z_k)$ des fonctions de corrélation à k molécules converge pour presque tout $Z_k\in(\mathbf{T}^3\times\mathbf{R}^3)^k$ vers

$$f(t,\cdot)^{\otimes k}(Z_k):=\prod_{j=1}^k f(t,z_j)\,.$$

Cet énoncé appelle une remarque importante : si l'on compare les points (1) et (2) du théorème de Lanford, on voit que la notion de convergence utilisée pour F_k^ε avec $k\geqslant 2$ est plus faible que celle utilisée pour F_1^ε. Ce point particulier avait été prévu par Grad (1958) (voir section 11, p. 223, dernier paragraphe) bien avant la démonstration de Lanford.

On trouve notamment dans (Grad, 1958) la phrase suivante « [...] it is possible to specify the exceptional set on which $F_2^\varepsilon(t,z_1,z_2)$ does not converge to $F_1(t,z_1)F_1(t,z_2)$ rather precisely [...] » . Ce point particulier est étudié en détail par Bodineau, Gallagher, Saint-Raymond et Simonella (2018), et nous y renvoyons le lecteur intéressé par cette question.

1.6. Ce qu'on ne doit pas ignorer à propos de l'équation de Boltzmann

Bien que le but de cet exposé soit d'aller « au-delà de l'équation de Boltzmann », nous utiliserons à plusieurs reprises certaines propriétés mathématiques de base de cette équation, propriétés que nous allons rappeler ici.

L'équation de Boltzmann s'écrit en toute généralité

$$(\partial_t + v\cdot\nabla_x)f(t,x,v) + \mathbf{a}(t,x)\cdot\nabla_v f(t,x,v) = \mathscr{B}(f)(t,x,v)\,, \qquad (x,v)\in \mathbf{T}^3\times\mathbf{R}^3\,, \tag{19}$$

où $\mathbf{a}$ est un champ d'accélération provenant d'une force extérieure — comme la gravité par exemple — tandis que $\mathscr{B}(f)$ désigne l'intégrale des collisions de Boltzmann et décrit la variation en temps de la population de molécules de vitesse v due à des collisions avec des molécules de vitesses différentes. L'inconnue f est la « fonction de distribution » [(5)] des molécules, à savoir la densité par rapport à la mesure de Lebesgue sur l'espace des phases $\mathbf{T}^3\times\mathbf{R}^3$ du nombre de molécules situées au point x à l'instant t et animées de la vitesse v. Dans tout cet exposé, on négligera systématiquement l'effet de la force extérieure, de sorte que $\mathbf{a}=0$, et que

$$(\partial_t + v\cdot\nabla_x)f(t,x,v) = \mathscr{B}(f)(t,x,v)\,, \qquad (x,v)\in \mathbf{T}^3\times\mathbf{R}^3\,.$$

Nous aurons besoin des propriétés essentielles suivantes de l'intégrale des collisions de Boltzmann. Comme le montre la définition (15), cette intégrale des collisions n'agit que sur la dépendance en v de la fonction de distribution. Il suffit donc de l'étudier sur des fonctions constantes en (t,x).

Formulation faible. Pour tout $\phi\in C(\mathbf{R}^3)$ à décroissance rapide, $\mathscr{B}(\phi)\in C(\mathbf{R}^3)$ est à décroissance rapide, et, pour toute fonction test $\psi\in C(\mathbf{R}^3)$ à croissance polynomiale à l'infini, on a les identités suivantes (dont la preuve sera esquissée plus loin, et où v' et v'_* sont donnés en fonction de v, v_* et ω par (16))

$$\begin{aligned}
\int_{\mathbf{R}^3}\mathscr{B}(\phi)(v)\psi(v)dv &= \int_{\mathbf{R}^3\times\mathbf{R}^3\times\mathbf{S}^2}(\psi(v')-\psi(v))\phi(v)\phi(v_*)((v-v_*)\cdot\omega)_+ dvdv_*d\omega\\
&= \int_{\mathbf{R}^3\times\mathbf{R}^3\times\mathbf{S}^2}(\psi(v'_*)-\psi(v_*))\phi(v)\phi(v_*)((v-v_*)\cdot\omega)_+ dvdv_*d\omega\\
&= \tfrac12\int_{\mathbf{R}^3\times\mathbf{R}^3\times\mathbf{S}^2}(\psi(v')+\psi(v'_*)-\psi(v)-\psi(v_*))\phi(v)\phi(v_*)((v-v_*)\cdot\omega)_+ dvdv_*d\omega\\
&= \tfrac14\int_{\mathbf{R}^3\times\mathbf{R}^3\times\mathbf{S}^2}(\psi(v)+\psi(v_*)-\psi(v')-\psi(v'_*))(\phi(v')\phi(v'_*)-\phi(v)\phi(v_*))\\
&\qquad\times((v-v_*)\cdot\omega)_+ dvdv_*d\omega\,.
\end{aligned}$$

[(5)] La littérature physique anglo-saxonne utilise indifféremment, pour désigner cette quantité, les termes de « distribution function » ou de « velocity distribution function ».

Lois de conservation locales. Pour tout $\phi \in C(\mathbf{R}^3)$ à décroissance rapide,

$$\int_{\mathbf{R}^3} \mathscr{B}(\phi)(v)dv = 0 \quad \text{(conservation locale de la masse)},$$
$$\int_{\mathbf{R}^3} \mathscr{B}(\phi)(v)\tfrac{1}{2}|v|^2 dv = 0 \quad \text{(conservation locale de l'énergie)},$$

ainsi que

$$\int_{\mathbf{R}^3} \mathscr{B}(\phi)(v)v dv = 0 \quad \text{(conservation locale de l'impulsion)}.$$

Ces lois de conservation se déduisent de la formulation faible en observant que

$$\psi(v) = \alpha_0 + \sum_{i=1}^{3} \alpha_i v_i + \alpha_4 |v|^2 \implies \psi(v') + \psi(v'_*) - \psi(v) - \psi(v_*) = 0$$
$$\text{pour tous } v, v_* \in \mathbf{R}^3 \text{ et } \omega \in \mathbf{S}^2,$$

où les vitesses v' et v'_* sont données en fonction de v, v_*, ω par les relations (16).

Évidemment, si $\phi \equiv \phi(t,x,v)$ est une solution de classe C^1 de l'équation de Boltzmann sur $[0,T[\times\mathbf{T}^3 \times \mathbf{R}^3$ à décroissance rapide en v ainsi que ses dérivées premières, on déduit des formules ci-dessus que

$$\partial_t \int_{\mathbf{R}^3} \phi(t,x,v)dv + \nabla_x \cdot \int_{\mathbf{R}^3} v\phi(t,x,v)dv = 0\,,$$
$$\partial_t \int_{\mathbf{R}^3} v_i\phi(t,x,v)dv + \nabla_x \cdot \int_{\mathbf{R}^3} vv_i\phi(t,x,v)dv = 0\,,$$
$$\partial_t \int_{\mathbf{R}^3} \tfrac{1}{2}|v|^2\phi(t,x,v)dv + \nabla_x \cdot \int_{\mathbf{R}^3} v\tfrac{1}{2}|v|^2\phi(t,x,v)dv = 0\,.$$

Il s'agit bien de lois de conservation locales, puisque chacune se met sous la forme $\mathrm{div}_{t,x} V(t,x) = 0$, où V est un champ de vecteurs sur $]0,T[\times\mathbf{T}^3$.

Théorème H de Boltzmann. Soit $\phi \in C(\mathbf{R}^3)$ à décroissance rapide telle que $\phi > 0$ et $\ln\phi$ soit à croissance polynomiale (par exemple $\phi(v) = e^{-P(|v|^2)}$ où P est une fonction polynomiale de coefficient dominant strictement positif). Alors

$$\int_{\mathbf{R}^3} \mathscr{B}(\phi)(v)\ln\phi(v)dv = -\tfrac{1}{4}\int_{\mathbf{R}^3\times\mathbf{R}^3\times\mathbf{S}^2} (\phi(v')\phi(v'_*) - \phi(v)\phi(v_*))\ln\left(\frac{\phi(v')\phi(v'_*)}{\phi(v)\phi(v_*)}\right)$$
$$\times((v-v_*)\cdot\omega)_+ dvdv_*d\omega \leqslant 0\,,$$

puisque ln est une fonction croissante. D'autre part

$$\mathscr{B}(\phi) = 0 \iff \int_{\mathbf{R}^3} \mathscr{B}(\phi)(v)\ln\phi(v)dv = 0 \iff \phi \text{ est une maxwellienne},$$

c'est-à-dire que ϕ est de la forme

$$\mathscr{M}_{\rho,u,\theta}(v) := \frac{\rho}{(2\pi\theta)^{3/2}} e^{-|v-u|^2/2\theta} . \tag{20}$$

De nouveau, si $\phi \equiv \phi(t,x,v) > 0$ est une solution de classe C^1 de l'équation de Boltzmann sur $[0,T[\times\mathbf{T}^3 \times \mathbf{R}^3$ à décroissance rapide en v ainsi que ses dérivées premières, telle que $\ln\phi$ soit à croissance polynomiale en v, on en déduit que

$$\frac{d}{dt}\int_{\mathbf{T}^3\times\mathbf{R}^3} \phi(t,x,v)\ln\phi(t,x,v)dxdv \leqslant 0 .$$

Ceci permet en particulier de majorer la quantité

$$\int_{\mathbf{T}^3\times\mathbf{R}^3} \phi(t,x,v)\ln\phi(t,x,v)dxdv$$

en fonction de sa donnée initiale. Cette propriété est évidemment l'une des clés pour obtenir des solutions globales de l'équation de Boltzmann sans restriction de taille sur les données initiales (DiPerna et Lions, 1989 ; Gérard, 1988).

Le seul point un peu difficile dans les énoncés ci-dessus est la caractérisation des maxwelliennes par l'équation fonctionnelle $\ln(\phi(v)\phi(v_*)) = \ln(\phi(v')\phi(v'_*))$: voir par exemple le Théorème 3.1.1 dans (Cercignani, Illner et Pulvirenti, 1994). Tous les autres énoncés se déduisent de la formulation faible, dont nous allons maintenant expliquer la preuve.

D'abord, ϕ étant continue à décroissance rapide et ψ étant à croissance polynomiale, la fonction

$$(v,v_*) \mapsto |\phi(v')| + |\phi(v'_*)|$$

est à décroissance rapide, tandis que la fonction

$$(v,v_*) \mapsto |\psi(v')| + |\psi(v'_*)|$$

est à croissance polynomiale pour tout $\omega \in \mathbf{S}^2$, sachant que les vitesses v', v'_* sont données en fonction de v, v_* et ω par (16).

Puis, pour tout $\omega \in \mathbf{S}^2$, l'application $(v,v_*) \mapsto (v',v'_*)$ est une isométrie linéaire involutive de $\mathbf{R}^3 \times \mathbf{R}^3$, et conserve donc la mesure de Lebesgue $dvdv_*$. Donc

$$\begin{aligned}
&\int_{\mathbf{R}^3\times\mathbf{R}^3\times\mathbf{S}^2} \psi(v)\phi(v')\phi(v'_*)((v-v_*)\cdot\omega)_+ dvdv_*d\omega \\
&= \int_{\mathbf{R}^3\times\mathbf{R}^3\times\mathbf{S}^2} \psi(v')\phi(v)\phi(v_*)(-(v-v_*)\cdot\omega)_+ dvdv_*d\omega \\
&= \int_{\mathbf{R}^3\times\mathbf{R}^3\times\mathbf{S}^2} \psi(v')\phi(v)\phi(v_*)((v-v_*)\cdot\omega)_+ dvdv_*d\omega ,
\end{aligned}$$

où la première égalité utilise que $(v' - v'_*) \cdot \omega = -(v - v_*) \cdot \omega$, tandis que la seconde est basée sur le fait que le changement de variables $\omega \mapsto -\omega$ laisse v' et v'_* invariants. Ceci démontre la première égalité de la formulation faible. La deuxième découle du fait que, pour tout $\omega \in \mathbf{S}^2$ fixé, la symétrie $v \mapsto v_*$ échange v' et v'_*. Les deux dernières égalités de la formulation faible en découlent aussitôt. On se reportera à la section 3.1 de (Cercignani, Illner et Pulvirenti, 1994) pour un traitement plus détaillé de ces propriétés.

1.7. Le théorème de Lanford comme loi des grands nombres

Après cette digression sur les propriétés mathématiques de l'équation de Boltzmann, revenons au théorème de Lanford.

Le fait que la fonction de corrélation à deux molécules F_2^ε se factorise dans la limite de Boltzmann–Grad, ce qui est le point (2) du théorème de Lanford, donne une information sur la famille ρ_t^ε des mesures empiriques définies en (7).

Corollaire 1.6. *Soit $h \in C_c(\mathbf{T}^3 \times \mathbf{R}^3)$. Sous les hypothèses du théorème de Lanford, pour tout $\eta > 0$*

$$\lim_{\varepsilon \to 0^+} \mathbb{P}_\varepsilon \left(\left\{ \left| \langle \rho_t^\varepsilon, h \rangle - \int_{\mathbf{T}^3 \times \mathbf{R}^3} h(z) f(t,z) dz \right| > \eta \right\} \right) = 0 \,.$$

Démonstration. L'inégalité de Bienaymé–Tchebychev dit que

$$\begin{aligned} &\mathbb{P}_\varepsilon \left(\left\{ \left| \langle \rho_t^\varepsilon, h \rangle - \int_{\mathbf{T}^3 \times \mathbf{R}^3} h(z) f(t,z) dz \right| > \eta \right\} \right) \\ &\leqslant \frac{1}{\eta^2} \mathbb{E}_\varepsilon \left(\left| \langle \rho_t^\varepsilon, h \rangle - \int_{\mathbf{T}^3 \times \mathbf{R}^3} h(z) f(t,z) dz \right|^2 \right) . \end{aligned}$$

Or

$$\begin{aligned} \mathbb{E}_\varepsilon \left(\left| \langle \rho_t^\varepsilon, h \rangle - \int_{\mathbf{T}^3 \times \mathbf{R}^3} h(z) f(t,z) dz \right|^2 \right) &= \mathbb{E}_\varepsilon (\langle \rho_t^\varepsilon, h \rangle^2) + \left(\int_{\mathbf{T}^3 \times \mathbf{R}^3} h(z) f(t,z) dz \right)^2 \\ &\quad - 2 \int_{\mathbf{T}^3 \times \mathbf{R}^3} h(z) F_1^\varepsilon(t,z) dz \int_{\mathbf{T}^3 \times \mathbf{R}^3} h(z) f(t,z) dz \,, \end{aligned}$$

et un calcul simple montre que

$$\langle \rho_t^\varepsilon, h \rangle^2 = \langle \rho_t^\varepsilon \otimes \rho_t^\varepsilon, h \otimes h \rangle = \frac{1}{\mu_\varepsilon} \langle \rho_t^\varepsilon, h^2 \rangle + \langle \rho_{2,t}^\varepsilon, h \otimes h \rangle \,,$$

de sorte que

$$\mathbb{E}_\varepsilon (\langle \rho_t^\varepsilon, h \rangle^2) = \frac{1}{\mu_\varepsilon} \int_{\mathbf{T}^3 \times \mathbf{R}^3} h(z)^2 F_1^\varepsilon(t,z) dz + \int_{(\mathbf{T}^3 \times \mathbf{R}^3)^2} h(z_1) h(z_2) F_2^\varepsilon(t, Z_2) dZ_2 \,.$$

Comme h est à support compact, le point (1) du théorème de Lanford montre que

$$\frac{1}{\mu_\varepsilon}\int_{\mathbf{T}^3\times\mathbf{R}^3} h(z)^2 F_1^\varepsilon(t,z)dz - 2\int_{\mathbf{T}^3\times\mathbf{R}^3} h(z)F_1^\varepsilon(t,z)dz\int_{\mathbf{T}^3\times\mathbf{R}^3} h(z)f(t,z)dz \\ \to -2\left(\int_{\mathbf{T}^3\times\mathbf{R}^3} h(z)f(t,z)dz\right)^2,$$

tandis que le point (2) du théorème de Lanford et une borne sur F_2^ε sur laquelle on reviendra plus tard (voir la Proposition 2.4 et le paragraphe qui la suit) montrent que

$$\int_{(\mathbf{T}^3\times\mathbf{R}^3)^2} h(z_1)h(z_2)F_2^\varepsilon(t,Z_2)dZ_2 \to \left(\int_{\mathbf{T}^3\times\mathbf{R}^3} h(z)f(t,z)dz\right)^2$$

lorsque $\varepsilon \to 0^+$, d'où le résultat annoncé. □

Rappelons la loi (faible et forte) des grands nombres dans l'énoncé suivant.

Théorème 1.7 (Loi des grands nombres). *Soit $(Y_n)_{n\geqslant 1}$, suite de variables aléatoires indépendantes et identiquement distribuées, telles que $\mathbb{E}(|Y_1|) < +\infty$. Alors*

$\frac{1}{n}(Y_1 + \ldots + Y_n) \to \mathbb{E}(Y_1)$ *en probabilité et presque sûrement lorsque $n \to +\infty$.*

Soit $\mathscr{Y} := (\mathbf{T}^3 \times \mathbf{R}^3)^{\mathbf{N}^*}$ muni de la mesure de probabilité borélienne définie par

$$\mathrm{Prob}(\{\mathbf{z} \in \mathscr{Y} \text{ t.q. } (z_1,\ldots,z_n) \in A_1 \times \ldots \times A_n\}) = \prod_{i=1}^n \int_{A_i} w(z)dz\,,$$

où w est une densité de probabilité borélienne sur $\mathbf{T}^3 \times \mathbf{R}^3$. Posons $Y_n(\mathbf{z}) := h(z_n)$; on vérifie sans peine que les variables aléatoires Y_n sont indépendantes et de même distribution, puisque

$$\mathrm{Prob}(\{\mathbf{z} \in \mathscr{Y} \text{ t.q. } Y_n(\mathbf{z}) > y\}) = \int_{\mathbf{T}^3\times\mathbf{R}^3} \mathbf{1}_{h(z)>y} w(z)dz\,.$$

Enfin $\mathbb{E}(|Y_1|) \leqslant \|h\|_{L^\infty(\mathbf{T}^3\times\mathbf{R}^3)} < +\infty$. D'après la loi faible des grands nombres

$$\lim_{n\to+\infty} \mathrm{Prob}\left(\left\{\mathbf{z} \in \mathscr{Y} \text{ t.q. } \left|\left\langle \frac{1}{n}\sum_{i=1}^n \delta_{z_i}, h\right\rangle - \int_{\mathbf{T}^3\times\mathbf{R}^3} h(z)w(z)dz\right| > \eta\right\}\right) = 0\,.$$

Cet exemple évoque évidemment la situation décrite dans le Corollaire 1.6. Pour autant, il s'agit une analogie plutôt que d'une véritable application de la loi faible des grands nombres telle qu'énoncée dans le Théorème 1.7. En effet

(a) les positions initiales des N particules ne peuvent en aucun cas être considérées comme indépendantes pour $\varepsilon > 0$ à cause du facteur d'exclusion $\prod_{1\leqslant i<j\leqslant N} \mathbf{1}_{\mathrm{dist}(x_i,x_j)>\varepsilon}$ dans

$$\mathbb{F}_N^{in}(Z_N) := \prod_{1\leqslant i<j\leqslant N} \mathbf{1}_{\mathrm{dist}(x_i,x_j)>\varepsilon} (f^{in})^{\otimes N}(Z_N)\,;$$

(b) au fur et à mesure de l'évolution du gaz, le nombre de collisions binaires entre molécules augmente, de sorte que les positions et les vitesses de ces N molécules deviennent « de moins en moins indépendantes ».

Néanmoins, il est utile de penser au théorème de Lanford comme à une sorte de loi des grands nombres appliquée à la suite de variables aléatoires $(h(z_n(t)))_{n\geqslant 1}$, où $z_n(t) = (x_n(t), v_n(t))$ est le couple position-vitesse de la n-ième molécule à l'instant t.

2. Cumulants et équation de Hamilton–Jacobi

Pour aller plus loin, deux voies se présentent :
(1) étendre l'intervalle de temps $[0, T[C_0, \beta_0]]$ sur lequel le théorème de Lanford permet de déduire l'équation de Boltzmann (14) de la dynamique moléculaire (1)-(2)-(3) ;
(2) puisque le théorème de Lanford peut s'interpréter comme une loi des grands nombres, transplanter dans le cadre de la limite de Boltzmann–Grad certains des théorèmes limites de la théorie des probabilités précisant la loi des grands nombres.

La démarche (1) présenterait un intérêt considérable. On sait que $T[C_0, \beta_0]$ est de l'ordre d'une fraction (1/5 pour être précis) du laps de temps moyen entre deux collisions successives subies par une même molécule prise au hasard dans le gaz : voir (Lanford, 1976) p. 117, Remarque 3 p. 132. Comme expliqué dans (Golse, 2014) à la fin de la section 4.1, le temps de validité du théorème de Lanford a été étendu à $+\infty$ par Illner et Pulvirenti (1989) dans le cas d'un domaine spatial $\mathbf{R}^3$ au lieu de $\mathbf{T}^3$, et pour des données initiales correspondant à un degré de raréfaction du gaz tel que le libre parcours moyen est très grand, et augmente même au cours de l'évolution grâce à l'effet de dispersion dû à l'opérateur de transport libre $\partial_t + v \cdot \nabla_x$. À ce jour, l'extension du temps de validité du théorème de Lanford pour une classe assez générale de données initiales, même au temps d'existence d'une solution classique maximale de l'équation de Boltzmann, demeure un problème ouvert, peut-être impossible à résoudre avec les méthodes de démonstrations employées jusqu'ici.

C'est pourquoi une partie importante des travaux de Bodineau, Gallagher, Saint-Raymond et Simonella adoptent la démarche (2). On verra toutefois que cela leur permet d'apporter quelques réponses partielles (restreintes au cas des fluctuations autour de solutions d'équilibre), mais d'un grand intérêt, au problème (1).

2.1. Grandes déviations et théorème central limite

Revenons au cadre idéal du Théorème 1.7. La loi des grands nombres nous dit que la moyenne « empirique » des variables aléatoires $Y_1, \dots, Y_n$ tend vers leur espérance mathématique commune $\mathbb{E}(Y_1)$. Une question naturelle consiste donc à étudier les fluctuations de la moyenne empirique autour de sa limite, c'est-à-dire

$$\sqrt{n}\left(\frac{Y_1+\ldots+Y_n}{n}-\mathbb{E}(Y_1)\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i-\mathbb{E}(Y_i))\,.$$

Le choix de dilater les fluctuations par le facteur $\sqrt{n}$ est justifié par le théorème central limite, rappelé ci-dessous.

Théorème 2.1 (Théorème central limite). *Soit $(Y_n)_{n\geqslant 1}$, suite de variables aléatoires réelles indépendantes et identiquement distribuées, telles que $\mathbb{E}(Y_1^2)<+\infty$. Alors*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i-\mathbb{E}(Y_i))\to\mathcal{N}(0,\sigma^2)\quad\text{ en loi lorsque } n\to+\infty\,,$$

où $\mathcal{N}(0,\sigma^2)$ est la loi normale centrée de variance $\sigma^2:=\mathbb{E}(Y_1^2)-\mathbb{E}(Y_1)^2$, de densité gaussienne $e^{-y^2/2\sigma^2}/\sigma\sqrt{2\pi}$ par rapport à la mesure de Lebesgue sur $\mathbf{R}$.

(Voir par exemple le chapitre IV, section 4.3 de (Malliavin, 1995), où ce résultat porte le nom de « Théorème de Laplace »).

Comme on l'a dit plus haut, le problème de la limite de Boltzmann–Grad ne peut pas se réduire à l'étude d'une suite de variables aléatoires indépendantes. Toutefois, le théorème central limite suggère d'étudier les fluctuations de la mesure empirique autour de la fonction F_1^ε, dont le théorème de Lanford montre qu'elle converge vers la solution de l'équation de Boltzmann. Spécifiquement, on considèrera

$$\zeta_t^\varepsilon:=\sqrt{\mu_\varepsilon}(\rho_t^\varepsilon-F_1^\varepsilon(t,\cdot))\,,\qquad t\geqslant 0\,,\quad \varepsilon>0\,,\tag{21}$$

que l'on peut réécrire sous la forme

$$\langle\zeta_t^\varepsilon,h\rangle:=\sqrt{\mu_\varepsilon}\left(\langle\rho_t^\varepsilon,h\rangle-\mathbb{E}_\varepsilon(\langle\rho_t^\varepsilon,h\rangle)\right)\,,\qquad t\geqslant 0\,,\quad \varepsilon>0\,,\tag{22}$$

pour toute fonction $h\in C_b(\mathbf{T}^3\times\mathbf{R}^3)$. La fluctuation ζ_t^ε décrit une correction d'ordre supérieur $1/\sqrt{\mu_\varepsilon}=\varepsilon$ à la quantité F_1^ε qui converge vers une solution f de l'équation de Boltzmann d'après le théorème de Lanford (Théorème 1.5 ci-dessus).

Revenons encore au cadre du Théorème 1.7. Comme le théorème central limite porte sur la convergence en loi des fluctuations de la moyenne empirique des variables aléatoires $Y_1,\dots,Y_n$ autour de $\mathbb{E}(Y_1)$, on va étudier W_n, la loi de cette mesure empirique $\frac{1}{n}(Y_1+\ldots+Y_n)$. Notons w la loi commune aux variables aléatoires Y_i pour

$i \geqslant 1$. L'indépendance des Y_i entraîne que W_n est la mesure image du produit de convolution à n termes $w \star \ldots \star w$ par l'homothétie de rapport $1/n$ (voir par exemple le Corollaire 4.2.2 de (Malliavin, 1995)).

Théorème 2.2 (Théorème de Cramér). *Soit $(Y_n)_{n\geqslant 1}$ une suite de variables aléatoires réelles indépendantes et identiquement distribuées de loi w, telles que, pour tout $\tau \in \mathbf{R}$, l'on ait $\mathbb{E}(\exp(\tau Y_1)) < +\infty$. Pour tout entier $n \geqslant 1$, notons W_n la loi de la moyenne empirique $\frac{1}{n}(Y_1 + \ldots + Y_n)$. Alors, pour tout ouvert $\mathscr{O}$ et tout fermé $\mathscr{F}$ de $\mathbf{R}$*

$$\overline{\lim_{n\to+\infty}} \tfrac{1}{n} \ln W_n(\mathscr{F}) \leqslant - \inf_{y\in\mathscr{F}} K(y) \quad \text{et} \quad \underline{\lim_{n\to+\infty}} \tfrac{1}{n} \ln W_n(\mathscr{O}) \geqslant - \inf_{y\in\mathscr{O}} K(y)\,,$$

où K est la transformée de Legendre du logarithme de la transformée de Laplace de w, à savoir

$$K(\theta) := \sup_{\tau\in\mathbf{R}} (\theta\tau - \ln \mathbb{E}(\exp(\tau Y_1))) = \sup_{\tau\in\mathbf{R}} \left(\theta\tau - \ln \int_{\mathbf{R}} e^{\tau y} w(dy) \right).$$

Le théorème de Cramér est l'un des énoncés fondamentaux de la théorie des « grandes déviations » (voir par exemple (Varadhan, 1984) pour plus de détails, ainsi que pour une preuve de ce théorème). Soit $\bar{Y} := \mathbb{E}(Y_1)$; l'inégalité de Jensen montre que

$$\ln \mathbb{E}(\exp(\tau Y_1)) \geqslant \tau \bar{Y} \quad \text{pour tout } \tau \in \mathbf{R}\,,$$

de sorte que $K(\bar{Y}) \leqslant 0$. D'autre part la fonction $\tau \mapsto \theta\tau - \ln \mathbb{E}(\exp(\tau Y_1))$ s'annule pour $\tau = 0$, de sorte que $K(\theta) \geqslant 0$ pour tout $\theta \in \mathbf{R}$ par définition, si bien que

$$K(\bar{Y}) = 0 = \min_{\theta\in\mathbf{R}} K(\theta)\,.$$

Comme la fonction K est convexe par construction, elle décroît sur $]-\infty, \bar{Y}]$ et croît sur $[\bar{Y}, +\infty[$. Ainsi

$$a > \bar{Y} \implies \lim_{n\to+\infty} \tfrac{1}{n} \ln W_n(]a, +\infty[) = \lim_{n\to+\infty} \tfrac{1}{n} \ln W_n([a, +\infty[) = -K(a)\,,$$
$$a < \bar{Y} \implies \lim_{n\to+\infty} \tfrac{1}{n} \ln W_n(]-\infty, a[) = \lim_{n\to+\infty} \tfrac{1}{n} \ln W_n(]-\infty, a]) = -K(a)\,.$$

Autrement dit, la probabilité que la moyenne empirique s'écarte de $\bar{Y}$ d'une distance $\eta > 0$ décroît exponentiellement lorsque $n \to +\infty$ dès que $K(a+\eta) + K(a-\eta) > 0$.

Le théorème de Cramér apporte donc une information quant à la précision de l'approximation de la moyenne empirique par l'espérance dans la loi des grands nombres.

Bien que le problème de la limite de Boltzmann–Grad ne puisse pas se réduire à l'étude d'une suite de variables aléatoires indépendantes comme expliqué plus haut, le théorème de Cramér suggère donc cependant de considérer la quantité

$$\mathscr{K}_\varepsilon(t,h) := \frac{1}{\mu_\varepsilon} \ln \mathbb{E}_\varepsilon(\exp(\mu_\varepsilon \langle \rho_t^\varepsilon, h\rangle))\,, \quad h \in C_b(\mathbf{T}^3 \times \mathbf{R}^3)\,,\ t \geqslant 0\,. \tag{23}$$

2.2. Cumulants

La famille des cumulants d'ordre k est définie à partir de la famille des fonctions de corrélations F_k^ε comme suit :

$$f_k^\varepsilon := \mu_\varepsilon^{k-1} \sum_{j=1}^{k} \sum_{\sigma \in \mathscr{P}_k^j} (-1)^{j-1}(j-1)! F_\sigma^\varepsilon \,, \tag{24}$$

où $\mathscr{P}_k^j$ désigne l'ensemble des partitions $\sigma := \{\sigma_1, \dots, \sigma_j\}$ de $\{1, \dots, k\}$ en j sous-ensembles, et où

$$F_\sigma^\varepsilon(t, Z_k) := \prod_{i=1}^{j} F_{\sigma_i}^\varepsilon(t, Z_{\sigma_i}) \quad \text{avec } F_{\sigma_i}^\varepsilon(t, Z_{\sigma_i}) := F_{|\sigma_i|}^\varepsilon(t, z_{l_1}, \dots, z_{l_{|\sigma_i|}}) \,,$$

les indices $l_1, \dots, l_{|\sigma_i|}$ étant définis par l'égalité $\sigma_i = \{l_1, \dots, l_{|\sigma_i|}\}$. On vérifie que

$$f_1^\varepsilon = F_1^\varepsilon \,, \quad f_2^\varepsilon(t, \cdot) = \mu_\varepsilon(F_2^\varepsilon(t, \cdot) - F_1^\varepsilon(t, \cdot) \otimes F_1^\varepsilon(t, \cdot))$$

puis que

$$\begin{aligned} f_3^\varepsilon(t, \cdot) =& \mu_\varepsilon^2(F_3^\varepsilon(t, \cdot) - F_{\{1,2\}}^\varepsilon(t, \cdot) F_{\{3\}}^\varepsilon(t, \cdot) - F_{\{2.3\}}^\varepsilon(t, \cdot) F_{\{1\}}^\varepsilon(t, \cdot) \\ & - F_{\{1,3\}}^\varepsilon(t, \cdot) F_{\{2\}}^\varepsilon(t, \cdot) + 2F_1^\varepsilon(t, \cdot) \otimes F_1^\varepsilon(t, \cdot) \otimes F_1^\varepsilon(t, \cdot)) \,. \end{aligned}$$

Le second cumulant f_2^ε est évidemment crucial pour la limite de Boltzmann–Grad, puisque l'hypothèse du chaos moléculaire de Boltzmann, qui est la clé du raisonnement permettant de déduire l'équation de Boltzmann comme approximation de l'identité (13) reliant F_1^ε à F_2^ε, s'écrit simplement

$$\lim_{\varepsilon \to 0^+} \frac{1}{\mu_\varepsilon} f_2^\varepsilon(t, x, v_1, x + \varepsilon\omega, v_2) = 0 \quad \text{lorsque } |\omega| = 1 \text{ et } (v_2 - v_1) \cdot \omega < 0 \,.$$

Un travail important de Pulvirenti et Simonella (2017) montre d'ailleurs comment des estimations sur les cumulants permettent de quantifier l'erreur correspondant à l'hypothèse de chaos moléculaire, et utilise ces estimations pour préciser la limite de Boltzmann–Grad (voir les Théorèmes 2.4 et 2.5 de (Pulvirenti et Simonella, 2017)).

Dans ce qui va suivre, la notion de cumulant va être utilisée de manière différente. Le lemme ci-dessous permet d'abord de montrer que la notion de cumulant est reliée à la fonctionnelle $\mathscr{K}^\varepsilon$ dont l'étude est suggérée par le théorème de Cramér.

Lemme 2.3. *Pour tout $\varepsilon > 0$, tout $t \geqslant 0$ et tout $h \in C_b(\mathbf{T}^3 \times \mathbf{R}^3)$, on a*

$$\mathscr{K}^\varepsilon(t, h) = \sum_{k \geqslant 1} \frac{1}{k!} \left\langle f_k^\varepsilon(t, \cdot), (e^h - 1)^{\otimes k} \right\rangle \,.$$

Démonstration. D'abord

$$\begin{aligned}\mathbb{E}_\varepsilon(\exp(\mu_\varepsilon\langle\rho_t^\varepsilon,h\rangle))-1&=\sum_{m\geqslant 1}\frac{\mu_\varepsilon^m}{m!}\mathbb{E}_\varepsilon(\langle\rho_t^\varepsilon,h\rangle^m)\\&=\sum_{m\geqslant 1}\frac{1}{m!}\mathbb{E}_\varepsilon\left(\sum_{j_1,\dots,j_m=1}^{N}h(z_{j_1}(t))\dots h(z_{j_m}(t))\right)\\&=\sum_{m\geqslant 1}\frac{1}{m!}\sum_{n=1}^{m}\mu_\varepsilon^n\sum_{\sigma\in\mathscr{P}_m^n}\int_{(\mathbf{T}^3\times\mathbf{R}^3)^n}\prod_{j=1}^{n}h(z_j)^{|\sigma_j|}F_n^\varepsilon(t,Z_n)dZ_n\,.\end{aligned}$$

Or le nombre de partitions de $\{1,\dots,m\}$ en n ensembles à $\nu_1,\dots,\nu_n$ éléments vaut

$$\frac{1}{n!}\binom{m}{\nu_1}\binom{m-\nu_1}{\nu_2}\dots\binom{m-\nu_1-\dots-\nu_{n-2}}{\nu_{n-1}}=\frac{1}{n!}\frac{m!}{\nu_1!\dots\nu_n!}\,,$$

de sorte que

$$\sum_{m\geqslant n}\frac{1}{m!}\sum_{\sigma\in\mathscr{P}_m^n}\prod_{j=1}^{n}h(z_j)^{|\sigma_j|}=\frac{1}{n!}\prod_{j=1}^{n}\sum_{\nu_j\geqslant 1}\frac{h(z_j)^{\nu_j}}{\nu_j!}=\frac{1}{n!}\prod_{j=1}^{n}(e^{h(z_j)}-1)\,.$$

En échangeant l'ordre des sommations en m et en n, on trouve donc que

$$\begin{aligned}\mathbb{E}_\varepsilon(\exp(\mu_\varepsilon\langle\rho_t^\varepsilon,h\rangle))-1&=\sum_{n\geqslant 1}\mu_\varepsilon^n\int_{(\mathbf{T}^3\times\mathbf{R}^3)^n}\sum_{m\geqslant n}\frac{1}{m!}\sum_{\sigma\in\mathscr{P}_m^n}\prod_{j=1}^{n}h(z_j)^{|\sigma_j|}F_n^\varepsilon(t,Z_n)dZ_n\\&=\sum_{n\geqslant 1}\frac{\mu_\varepsilon^n}{n!}\langle F_n^\varepsilon(t,\cdot),(e^h-1)^{\otimes n}\rangle\,.\end{aligned}$$

Puis

$$\begin{aligned}\ln\mathbb{E}_\varepsilon(\exp(\mu_\varepsilon\langle\rho_t^\varepsilon,h\rangle))&=\sum_{p\geqslant 1}\frac{(-1)^{p-1}}{p}\left(\sum_{n\geqslant 1}\frac{\mu_\varepsilon^n}{n!}\langle F_n^\varepsilon(t,\cdot),(e^h-1)^{\otimes n}\rangle\right)^p\\&=\sum_{p\geqslant 1}\frac{(-1)^{p-1}}{p}\sum_{n_1,\dots,n_p\geqslant 1}\frac{\mu_\varepsilon^{n_1+\dots+n_p}}{n_1!\dots n_p!}\prod_{k=1}^{p}\langle F_{n_k}^\varepsilon,(e^h-1)^{\otimes n_k}\rangle\,.\end{aligned}$$

Grâce au dénombrement des partitions à p éléments de $\{1,\dots,n\}$ rappelé ci-dessus

$$\begin{aligned}\ln\mathbb{E}_\varepsilon(\exp(\mu_\varepsilon\langle\rho_t^\varepsilon,h\rangle))&=\sum_{n\geqslant 1}\frac{\mu_\varepsilon^n}{n!}\sum_{p=1}^{n}\frac{(-1)^{p-1}}{p}p!\sum_{\sigma\in\mathscr{P}_n^p}\prod_{k=1}^{p}\langle F_{|\sigma_k|}^\varepsilon,(e^h-1)^{\otimes|\sigma_k|}\rangle\\&=\mu_\varepsilon\sum_{n\geqslant 1}\frac{1}{n!}\langle f_k^\varepsilon,(e^h-1)^{\otimes k}\rangle\,,\end{aligned}$$

d'où le résultat annoncé. □

Le Lemme 2.3 explique donc pourquoi la fonctionnelle $\mathscr{K}^\varepsilon$ est appelée « fonction génératrice des cumulants ».

2.3. Equation de Hamilton–Jacobi fonctionnelle : approche formelle

L'un des résultats principaux de Bodineau, Gallagher, Saint-Raymond et Simonella (2023) est que, dans la limite de Boltzmann–Grad, c'est-à-dire lorsque $\varepsilon \to 0^+$, la fonction génératrice des cumulants est solution d'une équation de Hamilton–Jacobi « fonctionnelle ». Le but de cette section est de présenter un calcul purement formel, inspiré de la section 3.3 de (Bodineau, Gallagher, Saint-Raymond et Simonella, 2022b) aboutissant à cette équation. Les énoncés rigoureux feront l'objet de la section suivante.

2.3.1. *Dérivées fonctionnelles de $\mathscr{K}^\varepsilon(t,\cdot)$ d'ordre un et deux.* — Soient trois fonctions test $h, \phi, \psi \in C_b(\mathbf{T}^3 \times \mathbf{R}^3)$. On commence par calculer

$$\left\langle \frac{\partial \mathscr{K}^\varepsilon}{\partial h}(t,h), \phi \right\rangle = \lim_{\eta\to 0} \frac{\mathscr{K}^\varepsilon(t,h+\eta\phi) - \mathscr{K}^\varepsilon(t,h)}{\eta}$$
$$= \frac{1}{\mu_\varepsilon \mathbb{E}_\varepsilon(\exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h\rangle))} \lim_{\eta\to 0} \mathbb{E}_\varepsilon\left(\frac{\exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h+\eta\phi\rangle - \exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h\rangle))}{\eta} \right)$$
$$= \frac{\mathbb{E}_\varepsilon(\langle \rho_t^\varepsilon, \phi\rangle \exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h\rangle))}{\mathbb{E}_\varepsilon(\exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h\rangle))} .$$

Passons maintenant au calcul de la dérivée seconde :

$$\frac{\partial^2 \mathscr{K}^\varepsilon}{\partial h^2}(t,h)\cdot(\phi,\psi) = \lim_{\eta\to 0} \frac{1}{\eta}\left(\left\langle \frac{\partial \mathscr{K}^\varepsilon}{\partial h}(t,h+\eta\phi), \psi \right\rangle - \left\langle \frac{\partial \mathscr{K}^\varepsilon}{\partial h}(t,h), \psi \right\rangle \right)$$
$$= \lim_{\eta\to 0} \frac{1}{\eta}\left(\frac{\mathbb{E}_\varepsilon(\langle \rho_t^\varepsilon, \psi\rangle \exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h+\eta\phi\rangle))}{\mathbb{E}_\varepsilon(\exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h+\eta\phi\rangle))} - \frac{\mathbb{E}_\varepsilon(\langle \rho_t^\varepsilon, \psi\rangle \exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h\rangle))}{\mathbb{E}_\varepsilon(\exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h\rangle))} \right)$$
$$= \lim_{\eta\to 0} \frac{1}{\eta}\left(\frac{\mathbb{E}_\varepsilon(\langle \rho_t^\varepsilon, \psi\rangle \exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h+\eta\phi\rangle)) - \mathbb{E}_\varepsilon(\langle \rho_t^\varepsilon, \psi\rangle \exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h\rangle))}{\mathbb{E}_\varepsilon(\exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h+\eta\phi\rangle))} \right)$$
$$+\mathbb{E}_\varepsilon(\langle \rho_t^\varepsilon, \psi\rangle \exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h\rangle)) \lim_{\eta\to 0} \frac{1}{\eta}\left(\frac{1}{\mathbb{E}_\varepsilon(\exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h+\eta\phi\rangle))} - \frac{1}{\mathbb{E}_\varepsilon(\exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h\rangle))} \right)$$
$$= \mu_\varepsilon \frac{\mathbb{E}_\varepsilon(\langle \rho_t^\varepsilon, \phi\rangle\langle \rho_t^\varepsilon, \psi\rangle \exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h\rangle))}{\mathbb{E}_\varepsilon(\exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h\rangle))}$$
$$-\mu_\varepsilon \frac{\mathbb{E}_\varepsilon(\langle \rho_t^\varepsilon, \psi\rangle \exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h\rangle))\mathbb{E}_\varepsilon(\langle \rho_t^\varepsilon, \phi\rangle \exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h\rangle))}{\mathbb{E}_\varepsilon(\exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h\rangle))^2} ,$$

ce que l'on peut réécrire sous la forme

$$\frac{1}{\mu_\varepsilon}\frac{\partial^2 \mathscr{K}^\varepsilon}{\partial h^2}(t,h)\cdot(\phi,\psi) = \frac{\mathbb{E}_\varepsilon(\langle \rho_t^\varepsilon, \phi\rangle\langle \rho_t^\varepsilon, \psi\rangle \exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h\rangle))}{\mathbb{E}_\varepsilon(\exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h\rangle))}$$
$$- \left\langle \frac{\partial \mathscr{K}^\varepsilon}{\partial h}(t,h), \phi \right\rangle \left\langle \frac{\partial \mathscr{K}^\varepsilon}{\partial h}(t,h), \psi \right\rangle ,$$

ou encore, de manière plus condensée,

$$\frac{1}{\mu_\varepsilon}\frac{\partial^2 \mathscr{K}^\varepsilon}{\partial h^2}(t,h) + \frac{\partial \mathscr{K}^\varepsilon}{\partial h}(t,h) \otimes \frac{\partial \mathscr{K}^\varepsilon}{\partial h}(t,h) = \frac{\mathbb{E}_\varepsilon(\exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h\rangle)\langle \rho_t^\varepsilon \otimes \rho_t^\varepsilon, \cdot\rangle)}{\mathbb{E}_\varepsilon(\exp(\mu_\varepsilon\langle \rho_t^\varepsilon, h\rangle))}\,.$$

2.3.2. *Évolution de $\mathscr{K}^\varepsilon(t,\cdot)$.* — Compte-tenu de la définition (23) de $\mathscr{K}^\varepsilon$ (qui est un logarithme), il sera commode de calculer

$$\begin{aligned}\exp(\mu_\varepsilon \mathscr{K}^\varepsilon(t,h))\mu_\varepsilon \partial_t \mathscr{K}^\varepsilon(t,h) &= \partial_t \exp(\mu_\varepsilon \mathscr{K}^\varepsilon(t,h)) = \partial_t \mathbb{E}_\varepsilon(\exp\langle \mu_\varepsilon \rho_t^\varepsilon, h\rangle)\\ &= \frac{1}{\mathscr{Z}_\varepsilon}\sum_{n\geqslant 1}\frac{\mu_\varepsilon^n}{n!}\partial_t \int_{(\mathbf{T}^3\times\mathbf{R}^3)^n} \exp\left(\sum_{i=1}^n h(z_i(t))\right)\mathbb{F}_n^{in}(Z_n)dZ_n\\ &= \frac{1}{\mathscr{Z}_\varepsilon}\sum_{n\geqslant 1}\frac{\mu_\varepsilon^n}{n!}\int_{(\mathbf{T}^3\times\mathbf{R}^3)^n} \exp\left(\sum_{i=1}^n h(z_i)\right)\partial_t \mathbb{F}_n(t,Z_n)dZ_n\,.\end{aligned}$$

On supposera d'autre part que la fonction test h est suffisamment régulière — par exemple, on pourra se restreindre au cas où $h \in C_c^\infty(\mathbf{T}^3\times\mathbf{R}^3)$, mais on verra plus loin qu'il n'est pas nécessaire que h soit indéfiniment différentiable.

On exprime alors $\partial_t \mathbb{F}_n(t,\cdot)$ au moyen de l'équation de Liouville écrite au sens des distributions sous la forme (11) :

$$\begin{aligned}\partial_t \mathbb{F}_n(t,Z_n) &= -\sum_{i=1}^n v_i\cdot\nabla_{x_i}\mathbb{F}_n(t,Z_n)\\ &+\sum_{1\leqslant i<j\leqslant n}\mathbb{F}_n(t,\hat{Z}_n[i,j])\big|_{\partial^+\Gamma_n^\varepsilon}((v_j-v_i)\cdot n_{ij})_+\delta_{\mathrm{dist}(x_i,x_j)=\varepsilon}\\ &-\sum_{1\leqslant i<j\leqslant n}\mathbb{F}_n(t,Z_n)\big|_{\partial^+\Gamma_n^\varepsilon}((v_j-v_i)\cdot n_{ij})_-\delta_{\mathrm{dist}(x_i,x_j)=\varepsilon}\,.\end{aligned}$$

Donc

$$\begin{aligned}\exp(\mu_\varepsilon \mathscr{K}^\varepsilon(t,h))\mu_\varepsilon\partial_t\mathscr{K}^\varepsilon(t,h) &= -\frac{1}{\mathscr{Z}_\varepsilon}\sum_{n\geqslant 1}\frac{\mu_\varepsilon^n}{n!}\left\langle \sum_{i=1}^n v_i\cdot\nabla_{x_i}\mathbb{F}_n(t,\cdot), \exp\left(\sum_{i=1}^n h(z_i)\right)\right\rangle\\ &+\frac{1}{\mathscr{Z}_\varepsilon}\sum_{n\geqslant 1}\frac{\mu_\varepsilon^n}{n!}\left\langle \sum_{1\leqslant j<k\leqslant n}\mathbb{F}_n(t,\hat{Z}_n[j,k])\big|_{\partial^+\Gamma_n^\varepsilon}((v_k-v_j)\cdot n_{jk})_+\delta_{\mathrm{dist}(x_j,x_k)=\varepsilon}, e^{\sum_{i=1}^n h(z_i)}\right\rangle\\ &-\frac{1}{\mathscr{Z}_\varepsilon}\sum_{n\geqslant 1}\frac{\mu_\varepsilon^n}{n!}\left\langle \sum_{1\leqslant j<k\leqslant n}\mathbb{F}_n(t,\cdot)\big|_{\partial^+\Gamma_n^\varepsilon}((v_k-v_j)\cdot n_{jk})_-\delta_{\mathrm{dist}(x_j,x_k)=\varepsilon}, e^{\sum_{i=1}^n h(z_i)}\right\rangle\\ &=: T_1+T_2-T_3.\end{aligned}$$

Le terme T_1 se réécrit comme suit :

$$\begin{aligned}
T_1 =& \frac{1}{\mathscr{Z}_\varepsilon}\sum_{n\geqslant 1}\frac{\mu_\varepsilon^n}{n!}\left\langle \mathbb{F}_n(t,\cdot), \sum_{i=1}^n v_i\cdot\nabla_{x_i}\exp\left(\sum_{i=1}^n h(z_i)\right)\right\rangle\\
=& \frac{1}{\mathscr{Z}_\varepsilon}\sum_{n\geqslant 1}\frac{\mu_\varepsilon^n}{n!}\left\langle \mathbb{F}_n(t,\cdot), \exp\left(\sum_{i=1}^n h(z_i)\right)\sum_{i=1}^n v_i\cdot\nabla_x h(z_i)\right\rangle\\
=& \mathbb{E}_\varepsilon(\exp(\langle\mu_\varepsilon\rho_t^\varepsilon,h\rangle)\langle\mu_\varepsilon\rho_t^\varepsilon, v\cdot\nabla_x h\rangle)\\
=& \mathbb{E}_\varepsilon(\exp(\mu_\varepsilon\langle\rho_t^\varepsilon,h\rangle))\mu_\varepsilon\left\langle\frac{\partial\mathscr{K}^\varepsilon}{\partial h}(t,h), v\cdot\nabla_x h\right\rangle\\
=& \exp(\mu_\varepsilon\mathscr{K}^\varepsilon(t,h))\mu_\varepsilon\left\langle\frac{\partial\mathscr{K}^\varepsilon}{\partial h}(t,h), v\cdot\nabla_x h\right\rangle.
\end{aligned}$$

Passons au terme T_3 :

$$\begin{aligned}
T_3 =& \frac{1}{\mathscr{Z}_\varepsilon}\sum_{n\geqslant 1}\frac{\mu_\varepsilon^n}{n!}\left\langle \sum_{1\leqslant j<k\leqslant n}\mathbb{F}_n(t,\cdot)\big|_{\partial^+\Gamma_n^\varepsilon}((v_k-v_j)\cdot n_{jk})_-\delta_{\mathrm{dist}(x_j,x_k)=\varepsilon+0}, e^{\sum_{i=1}^n h(z_i)}\right\rangle\\
=& \tfrac{1}{2}\mu_\varepsilon^2\mathbb{E}_\varepsilon(\exp(\langle\mu_\varepsilon\rho_t^\varepsilon,h\rangle)\langle\rho_t^\varepsilon\otimes\rho_t^\varepsilon, ((v-v_*)\cdot\tfrac{x-x_*}{\varepsilon})_+\delta_{\mathrm{dist}(x,x_*)=\varepsilon+0}\rangle).
\end{aligned}$$

Quant au terme T_2, il vient :

$$\begin{aligned}
T_2 =& \frac{1}{\mathscr{Z}_\varepsilon}\sum_{n\geqslant 1}\frac{\mu_\varepsilon^n}{n!}\left\langle \sum_{1\leqslant j<k\leqslant n}\mathbb{F}_n(t,\hat{Z}_n[j,k])\big|_{\partial^+\Gamma_n^\varepsilon}((v_k-v_j)\cdot n_{jk})_+\delta_{\mathrm{dist}(x_j,x_k)=\varepsilon}, e^{\sum_{i=1}^n h(z_i)}\right\rangle\\
=& \frac{1}{2\mathscr{Z}_\varepsilon}\sum_{n\geqslant 1}\frac{\mu_\varepsilon^n}{n!}\left\langle \sum_{j,k=1}^n\mathbb{F}_n(t,\cdot)((v_k'-v_j')\cdot n_{jk})_+\delta_{\mathrm{dist}(x_j,x_k)=\varepsilon+0}, e^{\mathbf{D}h(z_j,z_k,n_{kj})+\sum_{i=1}^n h(z_i)}\right\rangle\\
=& \tfrac{1}{2}\mu_\varepsilon^2\mathbb{E}_\varepsilon(\exp(\langle\mu_\varepsilon\rho_t^\varepsilon,h\rangle)\langle\rho_t^\varepsilon\otimes\rho_t^\varepsilon, e^{\mathbf{D}h(z,z_*,\frac{x-x_*}{\varepsilon})}((v-v_*)\cdot\tfrac{x-x_*}{\varepsilon})_-\delta_{\mathrm{dist}(x,x_*)=\varepsilon+0}\rangle),
\end{aligned}$$

où on a noté

$$\mathbf{D}h(z,z_*,\omega) := h(x,v-((v-v_*)\cdot\omega)\omega)+h(x_*,v_*+((v-v_*)\cdot\omega)\omega)-h(z)-h(z_*). \tag{25}$$

Posons, en observant que $\mathbf{D}h(z,z_*,\omega)=\mathbf{D}h(z,z_*,-\omega)$,

$$J_\varepsilon[h](z,z_*) := \int_{\mathbf{S}^2}\left(e^{\mathbf{D}h(z,z_*,\omega)}-1\right)((v-v_*)\cdot\omega)_+\delta_{\varepsilon\omega}(x-x_*)d\omega. \tag{26}$$

Observons que, pour toute fonction test ψ continue,

$$\begin{aligned}
\int_{\mathbf{T}^3\times\mathbf{T}^3}\psi(x,x_*)\delta_{\mathrm{dist}(x,x_*)=\varepsilon} &= \int_{\mathbf{T}^3\times\mathbf{S}^2}\psi(x,x+\varepsilon\omega)\varepsilon^2 dxd\omega\\
&= \int_{\mathbf{T}^3\times\mathbf{T}^3}\psi(x,x_*)\left(\varepsilon^2\int_{\mathbf{S}^2}\delta_{\varepsilon\omega}(x-x_*)d\omega\right)dxdx_*.
\end{aligned}$$

On ramène ainsi les expressions ci-dessus à des intégrales sur la sphère unité, ce qui fait émerger le facteur ε^2 que l'on retrouve dans la loi d'échelle de Boltzmann–Grad. Donc, en utilisant la condition $\mu_\varepsilon = 1/\varepsilon^2$, il vient

$$
\begin{aligned}
&T_2 - T_3 \\
&= \tfrac{1}{2}\mu_\varepsilon^2 \mathbb{E}_\varepsilon(\exp(\langle \mu_\varepsilon \rho_t^\varepsilon, h\rangle))\langle \rho_t^\varepsilon \otimes \rho_t^\varepsilon, (e^{\mathbf{D}h(z,z_*,\frac{x-x_*}{\varepsilon})}-1)((v-v_*)\cdot\tfrac{x-x_*}{\varepsilon})_- \delta_{\mathrm{dist}(x,x_*)=\varepsilon+0}\rangle) \\
&= \tfrac{1}{2}\mu_\varepsilon^2 \mathbb{E}_\varepsilon(\exp(\langle \mu_\varepsilon \rho_t^\varepsilon, h\rangle)) \left(\frac{1}{\mu_\varepsilon}\frac{\partial^2 \mathscr{K}^\varepsilon}{\partial h^2}(t,h) + \frac{\partial \mathscr{K}^\varepsilon}{\partial h}(t,h) \otimes \frac{\partial \mathscr{K}^\varepsilon}{\partial h}(t,h) \right) \cdot \varepsilon^2 J_\varepsilon[h] \\
&= \tfrac{1}{2}\exp(\mu_\varepsilon \mathscr{K}^\varepsilon(t,h))\mu_\varepsilon \left(\frac{1}{\mu_\varepsilon}\frac{\partial^2 \mathscr{K}^\varepsilon}{\partial h^2}(t,h) + \frac{\partial \mathscr{K}^\varepsilon}{\partial h}(t,h) \otimes \frac{\partial \mathscr{K}^\varepsilon}{\partial h}(t,h) \right) \cdot J_\varepsilon[h]\,.
\end{aligned}
$$

On trouve donc que

$$
\begin{aligned}
&\exp(\mu_\varepsilon \mathscr{K}^\varepsilon(t,h))\mu_\varepsilon \partial_t \mathscr{K}^\varepsilon(t,h) = \exp(\mu_\varepsilon \mathscr{K}^\varepsilon(t,h))\mu_\varepsilon \left\langle \frac{\partial \mathscr{K}^\varepsilon}{\partial h}(t,h), v\cdot\nabla_x h \right\rangle \\
&+\tfrac{1}{2}\exp(\mu_\varepsilon \mathscr{K}^\varepsilon(t,h))\mu_\varepsilon \left(\frac{1}{\mu_\varepsilon}\frac{\partial^2 \mathscr{K}^\varepsilon}{\partial h^2}(t,h) + \frac{\partial \mathscr{K}^\varepsilon}{\partial h}(t,h) \otimes \frac{\partial \mathscr{K}^\varepsilon}{\partial h}(t,h) \right) \cdot J_\varepsilon[h]\,,
\end{aligned}
$$

c'est-à-dire que

$$
\begin{aligned}
\partial_t \mathscr{K}^\varepsilon(t,h) &= \left\langle \frac{\partial \mathscr{K}^\varepsilon}{\partial h}(t,h), v\cdot\nabla_x h \right\rangle \\
&+\tfrac{1}{2}\left(\frac{1}{\mu_\varepsilon}\frac{\partial^2 \mathscr{K}^\varepsilon}{\partial h^2}(t,h) + \frac{\partial \mathscr{K}^\varepsilon}{\partial h}(t,h) \otimes \frac{\partial \mathscr{K}^\varepsilon}{\partial h}(t,h) \right) \cdot J_\varepsilon[h]\,.
\end{aligned}
\tag{27}
$$

Passons à la limite formellement dans cette équation, en supposant que $\mathscr{K}^\varepsilon(t,h)$ converge vers $\mathscr{K}(t,h)$ lorsque $\varepsilon \to 0^+$: il vient

$$
\partial_t \mathscr{K}(t,h) = \left\langle \frac{\partial \mathscr{K}}{\partial h}(t,h), v\cdot\nabla_x h \right\rangle + \tfrac{1}{2}\frac{\partial \mathscr{K}}{\partial h}(t,h) \otimes \frac{\partial \mathscr{K}}{\partial h}(t,h) \cdot J_0[h]\,, \tag{28}
$$

avec

$$
J_0[h] := \left(\int_{\mathbf{S}^2} \left(e^{\mathbf{D}h(z,z_*,\omega)} - 1 \right) ((v-v_*)\cdot\omega)_+ d\omega \right) \delta_0(x-x_*)\,. \tag{29}
$$

L'équation vérifiée par $\mathscr{K}$ est une équation aux dérivées partielles « fonctionnelle » (puisque la « variable » h appartient à un espace fonctionnel de dimension infinie), de type Hamilton–Jacobi, puisque l'équation obtenue fait intervenir une fonctionnelle non linéaire ne dépendant que de la dérivée première $\partial\mathscr{K}/\partial h$.

Terminons cette section par une mise en garde : la dérivée seconde $\partial^2 \mathscr{K}^\varepsilon(t,h)/\partial h^2$ multipliée par le petit paramètre $1/\mu_\varepsilon$ dans (27) pourrait évoquer une limite à « viscosité » petite, et faire de la limite (formelle) de (27) vers (28) un résultat analogue à la limite des solutions u_ε de l'équation de Burgers

$$
\partial_t u_\varepsilon + u_\varepsilon \partial_x u_\varepsilon = \varepsilon \partial_x^2 u_\varepsilon
$$

vers les solutions « entropiques » de l'équation de Hopf

$$\partial_t u + \partial_x \left(\frac{u^2}{2} \right) = 0 \, .$$

Voir par exemple (Hopf, 1950). La définition (23) est d'ailleurs formellement tout à fait analogue à la transformation de Cole-Hopf, formule (3) de (Hopf, 1950). Mais le « coefficient » $J_\varepsilon[h]/\mu_\varepsilon$ de la dérivée seconde $\partial^2 \mathscr{K}^\varepsilon / \partial h^2(t,h)$ n'est ni positif, ni même de signe constant, puisque $\mathbf{D}h$ prend en général des valeurs positives et négatives. En effet, on vérifiera facilement que

$$\mathbf{D}h(x,v,x,v_*,\omega) > 0 \implies \mathbf{D}h(x,v',x,v'_*,-\omega) = -\mathbf{D}h(x,v,x,v_*,\omega) < 0 \, ,$$

par le même argument que celui utilisé dans la section 1.6 pour démontrer la formulation faible de l'intégrale de collision de Boltzmann.

2.4. L'équation de Hamilton–Jacobi : principaux résultats

Soient $t, \varepsilon > 0$. Dans toute la suite, on notera $\mathbf{Z}_n([0,t])$ la restriction à l'intervalle $[0,t]$ du chemin $\mathbf{R} \ni \tau \mapsto Z_n(\tau) := S_\tau^{n,\varepsilon}(Z_n^{in}) \in (\mathbf{T}^3 \times \mathbf{R}^3)^n$. On notera $\mathbf{z}_j([0,t])$ la j-ième composante de $\mathbf{Z}_n([0,t])$. On notera également $\mathbf{D}_n([0,t])$ l'espace des chemins càdlàg [6] définis sur $[0,t]$ à valeurs dans $(\mathbf{T}^3 \times \mathbf{R}^3)^n$, muni de la topologie de Skorokhod ((Billingsley, 1999), chapitre 3, section 12).

Commençons par donner une borne sur la fonction génératrice des cumulants.

Proposition 2.4. *On suppose que la fonction de distribution initiale f^{in} vérifie la borne* (17). *Il existe $C, T_0 > 0$ tels que, pour tout $h : D_1([0,+\infty[) \to \mathbf{R}$ continue vérifiant*

$$h((x,v)([0,t])) \leqslant \alpha + \tfrac{1}{4}\beta_0 \sup_{0 \leqslant s \leqslant t} |v(s)|^2 \, ,$$

l'on a

$$\left| \frac{1}{\mu_\varepsilon} \ln \mathbb{E}_\varepsilon \left(\exp \left(\sum_{j=1}^N h(\mathbf{z}_j([0,t])) \right) \right) \right| \leqslant \frac{CC_0 e^\alpha}{\beta_0^2} \sum_{n \geqslant 1} \left(\frac{CC_0 e^\alpha}{\beta_0^2} \right)^{n-1} (t+\varepsilon)^{n-1}$$
$$= \frac{CC_0 e^\alpha}{\beta_0^2 - CC_0 e^\alpha (t+\varepsilon)}$$

pour $t + \varepsilon < \min \left(T_0, \frac{\beta_0^2 e^{-\alpha}}{CC_0} \right) := T_\alpha[C_0, \beta_0]$.

[6] Acronyme de « continu à droite, [avec] limite à gauche ».

Cette borne supérieure est démontrée dans le chapitre 8 de (Bodineau, Gallagher, Saint-Raymond et Simonella, 2023) (Théorème 10, voir aussi le Théorème 4 du chapitre 4 de cette même référence). Elle implique une borne sur F_2^ε utilisée dans la preuve du Corollaire 1.6.

À partir de là, on va modifier légèrement la définition de la fonctionnelle $\mathscr{K}^\varepsilon$ de façon à intégrer le terme de transport libre $\left\langle \frac{\partial \mathscr{K}}{\partial h}(t,h), v\cdot\nabla_x h \right\rangle$. Pour cela, on considère des fonctions test h de la forme

$$h(z([0,t])) := g(t,z(t)) - \int_0^t (\partial_t + v\cdot\nabla_x) g(s,z(s))ds\,.$$

Soit

$$\begin{aligned}\mathbf{B}_\alpha := \{ g \in C^1([0,T_\alpha]\times\mathbf{T}^3\times\mathbf{R}^3;\mathbf{C}) \text{ t.q. } |g(t,z)| \leqslant (1-\tfrac{t}{2T_\alpha})(\alpha+\tfrac{1}{8}\beta_0|v|^2) \\ \text{et } \sup_{0\leqslant t\leqslant T_\alpha} |(\partial_t + v\cdot\nabla_x)g(t,z)| \leqslant \tfrac{1}{2T_\alpha}(\alpha+\tfrac{1}{8}\beta_0|v|^2)\}\,.\end{aligned}$$

Évidemment, si $g\in\mathbf{B}_\alpha$, on a

$$|h(z([0,t]))| \leqslant |g(t,z(t))| + \int_0^t |(\partial_t+v\cdot\nabla_x)g(s,z(s))|ds \leqslant (\alpha+\tfrac{1}{8}\beta_0|v|^2)\,,$$

de sorte que la borne uniforme de la proposition ci-dessus s'applique.

Proposition 2.5. *On suppose que la fonction de distribution initiale f^{in} vérifie la borne* (17). *Soit $\alpha>0$; pour tout $t\in[0,T_\alpha[$ (où T_α est comme dans la Proposition 2.4) et tout $g\in\mathbf{B}_\alpha$, on pose*

$$\mathbf{K}(t,g) := \lim_{\varepsilon\to 0^+} \frac{1}{\mu_\varepsilon} \ln \mathbb{E}_\varepsilon \left(\exp\left(\sum_{j=1}^N h(\mathbf{z}_j([0,t])) \right) \right)\,,$$

où

$$h(z([0,t])) = g(t,z(t)) - \int_0^t (\partial_t+v\cdot\nabla_x)g(s,z(s))ds\,.$$

Pour tous $t\in[0,T_\alpha[$ et $g\in\mathbf{B}_\alpha$, la dérivée fonctionnelle $\partial\mathbf{K}(t,g)/\partial g(t)$ s'identifie à une fonction continue sur $\mathbf{T}^3$ à valeurs dans l'espace des mesures de Radon sur $\mathbf{R}^3$, et pour tout $t\in[0,T_\alpha[$, il existe $C(t)>0$ tel que

$$\sup_{0\leqslant s\leqslant t}\sup_{x\in\mathbf{T}^3} \left\| (1+|v|)\exp(\tfrac{1}{8}\beta_0|v|^2)\frac{\partial\mathbf{K}(s,g)}{\partial g(s)}(x,\cdot)\right\|_{VT} \leqslant C(t)\,.$$

Les énoncés ci-dessus nous permettent alors de préciser l'équation de Hamilton–Jacobi obtenue à la section précédente par un calcul formel.

Théorème 2.6. *On suppose que la fonction de distribution initiale f^{in} vérifie la borne* (17). *Soit $\alpha > 0$; la fonction $\mathbf{K}$ définie sur $[0, T_\alpha[\times \mathbf{B}_\alpha$ est une solution de la forme intégrée en t de l'équation de Hamilton–Jacobi*

$$\begin{cases} \partial_t \mathbf{K}(t,g) = \mathscr{H}\left(\dfrac{\partial \mathbf{K}(t,g)}{\partial g(t)}, g(t) \right), & (t,g) \in [0, T_\alpha[\times \mathbf{B}_\alpha, \\ \mathbf{K}(0,g) = \displaystyle\int_{\mathbf{T}^3 \times \mathbf{R}^3} (e^{g(0,z)} - 1) f^{in}(z) dz, \end{cases} \tag{30}$$

où le hamiltonien $\mathscr{H}$ est défini par la formule suivante :

$$\mathscr{H}(p,q) := \tfrac{1}{2} \int_{\mathbf{T}^3 \times (\mathbf{R}^3)^2 \times \mathbf{S}^2} \left(e^{\mathbf{D}q(x,v,x_*v_*,\omega)} - 1 \right) ((v - v_*) \cdot \omega)_+ d\omega \, p(x,dv) p(x,dv_*) dx \tag{31}$$

pour toute fonction $q \in C(\mathbf{T}^3 \times \mathbf{R}^3)$ vérifiant

$$\sup_{x \in \mathbf{T}^3} |q(x,v)| \leqslant c + \tfrac{1}{8} \beta_0 |v|^2, \qquad v \in \mathbf{R}^3,$$

et tout $p \in C(\mathbf{T}^3; \mathscr{M}(\mathbf{R}^3))$ vérifiant

$$\sup_{x \in \mathbf{T}^3} \left\| (1 + |v|) \exp(\tfrac{1}{8} \beta_0 |v|^2) \frac{\partial \mathbf{K}(s,g)}{\partial g(s)}(x, \cdot) \right\|_{VT} < +\infty.$$

On rappelle enfin que la notation $\mathbf{D}q$ est définie dans (25).

Ce théorème est l'un des résultats majeurs de Bodineau, Gallagher, Saint-Raymond et Simonella (2023). Nous verrons plus loin qu'il a des applications importantes, d'une part à l'étude des fluctuations autour de l'équation de Boltzmann, et d'autre part à celle des grandes déviations dans la limite de Boltzmann–Grad. On pourrait s'interroger sur l'intérêt de remplacer la dynamique à N corps (1)-(2)-(3), posée sur l'espace des phases $(\mathbf{T}^3 \times \mathbf{R}^3)^N$ de dimension $6N$, grande mais finie, par l'équation (30), posée sur un espace de dimension infinie. Comme elle est obtenue après passage à la limite $\varepsilon \to 0^+$ et que le processus limite est markovien, on peut toutefois espérer que la dynamique sous-jacente à (30) soit plus simple que la dynamique à N corps.

2.5. De l'équation de Hamilton–Jacobi à l'équation de Boltzmann

Une première observation est que l'équation de Boltzmann peut se déduire de l'équation de Hamilton–Jacobi (30).

Proposition 2.7. *On suppose que la fonction de distribution initiale f^{in} vérifie la borne* (17). *Soit*

$$f(t,\cdot) := \frac{\partial \mathbf{K}(t,g)}{\partial g(t)}\Big|_{g=0} \in C(\mathbf{T}^3; \mathscr{M}(\mathbf{R}^3)).$$

Alors f est solution faible de l'équation de Boltzmann sous forme intégrée en temps avec donnée initiale f^{in}.

Commençons par un calcul formel à partir de l'équation (28).

Argument formel. Revenons à la fonctionnelle $\mathscr{K}$. Dérivons chaque membre de (28) en h dans la direction ϕ :

$$\begin{aligned}\partial_t \left\langle \frac{\partial \mathscr{K}}{\partial h}(t,h), \phi \right\rangle =& \frac{\partial^2 \mathscr{K}}{\partial h^2}(t,h)\cdot(\phi, v\cdot\nabla_x h) + \left\langle \frac{\partial \mathscr{K}}{\partial h}(t,h), v\cdot\nabla_x\phi \right\rangle \\ &+ \tfrac{1}{2}\frac{\partial^2 \mathscr{K}}{\partial h^2}(t,h)\otimes\frac{\partial \mathscr{K}}{\partial h}(t,h)\cdot(\phi, J_0[h]) \\ &+ \tfrac{1}{2}\frac{\partial \mathscr{K}}{\partial h}(t,h)\otimes\frac{\partial^2 \mathscr{K}}{\partial h^2}(t,h)\cdot(J_0[h], \phi) \\ &+ \tfrac{1}{2}\frac{\partial \mathscr{K}}{\partial h}(t,h)\otimes\frac{\partial \mathscr{K}}{\partial h}(t,h)\cdot\left(\frac{dJ_0}{dh}[h]\cdot\phi\right),\end{aligned}$$

puis faisons $h = 0$ dans cette équation. Comme $J_0[0] = 0$, il vient

$$\begin{aligned}\partial_t \left\langle \frac{\partial \mathscr{K}}{\partial h}(t,0), \phi \right\rangle =& \left\langle \frac{\partial \mathscr{K}}{\partial h}(t,0), v\cdot\nabla_x\phi \right\rangle \\ &+ \tfrac{1}{2}\frac{\partial \mathscr{K}}{\partial h}(t,0)\otimes\frac{\partial \mathscr{K}}{\partial h}(t,0)\cdot\left(\frac{dJ_0}{dh}[0]\cdot\phi\right).\end{aligned}$$

Un calcul simple montre que

$$\frac{dJ_0}{dh}[0]\cdot\phi = \left(\int_{\mathbf{S}^2} \mathbf{D}\phi(z, z_*, \omega)((v-v_*)\cdot\omega)_+ d\omega\right)\delta_0(x-x_*).$$

Posons $m(t,\cdot) := \frac{\partial \mathscr{K}}{\partial h}(t,0)$, que nous allons manipuler comme une fonction continue sur $\mathbf{T}^3$ à valeurs dans les mesures de Radon sur $\mathbf{R}^3$. Alors

$$\begin{aligned}&\left\langle m(t)\otimes m(t), \frac{dJ_0}{dh}[0]\cdot\phi \right\rangle \\ &= \int_{\mathbf{T}^3\times(\mathbf{R}^3)^2\times\mathbf{S}^2} \mathbf{D}\phi(x,v,x,v_*,\omega)((v-v_*)\cdot\omega)_+ d\omega m(t,x,dv)m(t,x,dv_*)dx.\end{aligned}$$

On est arrivé ainsi à l'identité satisfaite par $m(t,\cdot)$ pour toute fonction test ϕ dans un espace bien choisi (par exemple $C_c^\infty(\mathbf{T}^3\times\mathbf{R}^3)$) :

$$\begin{aligned}&\langle \partial_t m(t), \phi\rangle = \langle m(t), v\cdot\nabla_x\phi\rangle \\ &+\tfrac{1}{2}\int_{\mathbf{T}^3\times(\mathbf{R}^3)^2\times\mathbf{S}^2} \mathbf{D}\phi(x,v,x,v_*,\omega)((v-v_*)\cdot\omega)_+ d\omega m(t,x,dv)m(t,x,dv_*)dx,\end{aligned}$$

qui n'est rien d'autre que la formulation faible de l'équation de Boltzmann (14). □

Passons à l'argument rigoureux basé sur le Théorème 2.6. Commençons par une observation élémentaire, basée sur l'idée suivante : dans la fonctionnelle $\mathbf{K}(t,g)$, on doit considérer $g(t,\cdot)$ et $(\partial_t + v\cdot\nabla_x)g(\tau,\cdot)\big|_{\tau\in[0,t]}$ comme deux « variables » indépendantes — on reviendra d'ailleurs plus loin sur ce point précis.

Lemme 2.8. *Soient $\alpha > 0$ et $\sigma,\tau \in [0,T_\alpha]$. L'application $\mathbf{B}_\alpha \ni g \mapsto g(s,\cdot) \in \mathbf{C}$ est de classe C^1, et sa dérivée partielle par rapport à $g(\tau,\cdot)$ « à $(\partial_t + v\cdot\nabla_x)g$ constant » est donnée par la formule* [7]

$$\left(\frac{\partial g(\sigma,\cdot)}{\partial g(\tau,\cdot)}\right)_{(\partial_t+v\cdot\nabla_x)g} = A_{\sigma-\tau}\,,$$

où $A_\theta := e^{-\theta v\cdot\nabla_x}$ est le groupe à un paramètre engendré par l'opérateur d'advection, donné par la formule $(A_\theta\psi)(x,v) := \psi(x-\theta v,v)$.

Démonstration. Exprimons $g(\sigma,\cdot)$ en fonction de $g(\tau,\cdot)$ et de $(\partial_t+v\cdot\nabla_x)g(\tau,\cdot)\big|_{t\in[0,T_\alpha]}$ en observant que

$$g(\sigma,x,v) = g(t,x+(\tau-\sigma)v,v) - \int_\sigma^\tau (\partial_t + v\cdot\nabla_x)g(\theta,x+(\theta-\sigma)v,v)d\theta\,,$$

c'est-à-dire que

$$g(\sigma,\cdot) = A_{\sigma-\tau}g(\tau,\cdot) - \int_\sigma^\tau A_{\sigma-\theta}(\partial_t + v\cdot\nabla_x)g(\theta,\cdot)d\theta\,.$$

La formule annoncée en découle aussitôt. □

Démonstration de la Proposition 2.7. On rappelle la condition initiale de (30) :

$$\mathbf{K}(0,g) = \int_{\mathbf{T}^3\times\mathbf{R}^3} (e^{g(0,z)} - 1)f^{in}(z)dz\,,$$

d'où l'on tire immédiatement que

$$\left\langle \frac{\partial\mathbf{K}(0,g)}{\partial g(0)}, \psi \right\rangle = \int_{\mathbf{T}^3\times\mathbf{R}^3} e^{g(0,z)}\psi(z)f^{in}(z)dz\,.$$

Puis, d'après le lemme ci-dessus et la règle de dérivation des fonctions composées

$$\left\langle \frac{\partial\mathbf{K}(0,g)}{\partial g(t)}, \psi \right\rangle = \left\langle \frac{\partial\mathbf{K}(0,g)}{\partial g(0)}, A_{-t}\psi \right\rangle = \int_{\mathbf{T}^3\times\mathbf{R}^3} e^{g(0,z)}A_{-t}\psi(z)f^{in}(z)dz\,. \tag{32}$$

Ensuite, on écrit la forme intégrée en t de (30) :

$$\mathbf{K}(t,g) = \mathbf{K}(0,g) + \int_0^t \mathscr{H}\left(\frac{\partial\mathbf{K}(s,g)}{\partial g(s)}, g(s)\right) ds\,,$$

[7] On s'est résolu faute de mieux à employer ici la notation regrettable en usage dans les traités classiques de thermodynamique, à savoir $(\partial f/\partial x)_y$ pour désigner la dérivée partielle de f par rapport à la variable x, la variable y restant constante dans la prise de dérivée.

et on dérive chaque membre de cette égalité par rapport à $g(t,\cdot)$ dans la direction ψ, en utilisant de nouveau la règle de dérivation des fonctions composées et le Lemme 2.8 pour trouver que

$$\begin{aligned}\left\langle \frac{\partial \mathbf{K}(t,g)}{\partial g(t)}, \psi \right\rangle &= \int_{\mathbf{T}^3\times\mathbf{R}^3} e^{g(0,z)} A_{-t}\psi(z) f^{in}(z)dz \\ &+ \int_0^t \left\langle \frac{\partial \mathscr{H}}{\partial p}\left(\frac{\partial \mathbf{K}(s,g)}{\partial g(s)}, g(s)\right), \frac{\partial^2 \mathbf{K}(s,g)}{\partial g(t)\partial g(s)}\cdot(\psi,\cdot)\right\rangle ds \\ &+ \int_0^t \left\langle \frac{\partial \mathscr{H}}{\partial q}\left(\frac{\partial \mathbf{K}(s,g)}{\partial g(s)}, g(s)\right), A_{s-t}\psi\right\rangle ds\,.\end{aligned} \tag{33}$$

La formule (31) implique que

$$\left\langle \frac{\partial \mathscr{H}}{\partial p}(p,q), \tilde{p}\right\rangle = \int_{\mathbf{T}^3\times(\mathbf{R}^3)^2\times\mathbf{S}^2} \left(e^{\mathbf{D}q(x,v,x_*v_*,\omega)}-1\right)((v-v_*)\cdot\omega)_+ d\omega\, p(x,dv)\tilde{p}(x,dv_*)dx \tag{34}$$

d'où l'on tire en particulier que

$$\left\langle \frac{\partial \mathscr{H}}{\partial p}(p,0), \tilde{p}\right\rangle = 0\,. \tag{35}$$

D'autre part

$$\begin{aligned}&2\left\langle \frac{\partial \mathscr{H}}{\partial q}(p,q), \psi\right\rangle \\ &= \int_{\mathbf{T}^3\times(\mathbf{R}^3)^2\times\mathbf{S}^2} e^{\mathbf{D}q(x,v,x_*v_*,\omega)}\mathbf{D}\psi(x,v,x_*v_*,\omega)((v-v_*)\cdot\omega)_+ d\omega\, p(x,dv)p(x,dv_*)dx,\end{aligned} \tag{36}$$

d'où

$$\left\langle \frac{\partial \mathscr{H}}{\partial q}(p,0), \psi\right\rangle = \tfrac{1}{2}\int_{\mathbf{T}^3\times(\mathbf{R}^3)^2\times\mathbf{S}^2} \mathbf{D}\psi(x,v,x_*v_*,\omega)((v-v_*)\cdot\omega)_+ d\omega\, p(x,dv)p(x,dv_*)dx\,.$$

Spécialisons (33) à $g=0$: il vient

$$\langle f(t,\cdot),\psi\rangle = \left\langle \frac{\partial \mathbf{K}(t,g)}{\partial g(t)}\Big|_{g=0}, \psi\right\rangle + \int_0^t \left\langle \frac{\partial \mathscr{H}}{\partial q}\left(f(s,\cdot),0\right), A_{s-t}\psi\right\rangle ds\,,$$

c'est-à-dire

$$\begin{aligned}\langle f(t,\cdot),\psi\rangle = &\int_{\mathbf{T}^3\times\mathbf{R}^3} \psi(z)A_t f^{in}(z)dz + \tfrac{1}{2}\int_0^t \int_{\mathbf{T}^3\times(\mathbf{R}^3)^2\times\mathbf{S}^2} \mathbf{D}\psi(x,v,x_*v_*,\omega) \\ &\times A_{t-s}f(s,x,v)A_{t-s}f(s,x,v_*)((v-v_*)\cdot\omega)_+ d\omega\, dvdv_*dx\,.\end{aligned}$$

On reconnaît dans cette identité une formulation faible intégrée en temps de l'équation de Boltzmann — cette écriture utilise la troisième formulation faible de l'intégrale des collisions de Boltzmann rappelée dans la section 1.6, ainsi que la formule de Duhamel pour traiter l'opérateur d'advection dans l'équation de Boltzmann (14). □

3. Grandes déviations

Dans cette section, on va voir comment la fonctionnelle **K** et l'équation de Hamilton–Jacobi qu'elle vérifie interviennent dans l'étude des grandes déviations pour un système de sphères dures dans la limite de Boltzmann–Grad.

On va considérer une variante à poids de l'équation de Boltzmann :

$$\begin{aligned}(\partial_t + v\cdot\nabla_x)\phi(t,x,v)\\ = \int_{\mathbf{R}^3\times\mathbf{S}^2}\Big(\phi(t,x,v'_*)\phi(t,x,v')e^{-\mathbf{D}q(t,x,v,x,v_*,\omega)} - \phi(t,x,v_*)\phi(t,x,v)e^{\mathbf{D}q(t,x,v,x,v_*,\omega)}\Big)\\ \times((v-v_*)\cdot\omega)_+ dv_* d\omega\,,\end{aligned} \tag{37}$$

avec condition initiale

$$\phi(0,x,v) = f^{in}(x,v)e^{q(0,x,v)}\,. \tag{38}$$

On supposera que la fonction $q \equiv q(t,x,v)$ est lipschitzienne sur $[0,T]\times\mathbf{T}^3\times\mathbf{R}^3$. Pour tous $r, T > 0$, on considère

$$\begin{aligned}\mathscr{R}_{r,T} := \{\phi :\ [0,T]\times\mathbf{T}^3\times\mathbf{R}^3 \to \mathbf{R}_+ \text{ solution forte de (37)-(38) pour une fonction}\\ q \text{ telle que } \|q\|_{W^{1,\infty}([0,T]\times\mathbf{T}^3\times\mathbf{R}^3)} \leqslant r\}\,.\end{aligned}$$

Maintenant, on va regarder la mesure empirique $t \mapsto \rho_t^\epsilon$ comme élément de l'espace $D([0,T],\mathscr{M}_+^1(\mathbf{T}^3\times\mathbf{R}^3))$ des chemins càdlàg à valeurs dans l'espace des mesures positives sur $\mathbf{T}^3\times\mathbf{R}^3$ de masse finie muni de la topologie faible. L'espace $D([0,T],\mathscr{M}_+^1(\mathbf{T}^3\times\mathbf{R}^3))$ est muni de la topologie de Skorokhod. Soient γ_1 et γ_2 deux éléments de $D([0,T],\mathscr{M}_+^1(\mathbf{T}^3\times\mathbf{R}^3))$; posons

$$d_{[0,T]}(\gamma_1,\gamma_2) := \inf_{\alpha\in\mathscr{A}_T}\max\left(\sup_{0\leqslant t\leqslant T}|\alpha(t)-t|,\ \sup_{0\leqslant t\leqslant T}\mathbf{d}(\gamma_1(t),\gamma_2(\alpha(t)))\right)$$

où $\mathscr{A}_T$ est l'ensemble des bijections croissantes de $[0,T]$ dans lui-même, et où

$$\mathbf{d}(m_1,m_2) = \sum_{n\geqslant 1} 2^{-n}|\langle m_1 - m_2, \chi_n\rangle|\,,$$

sachant que $(\chi_n)_{n\geqslant 1}$ est une suite dense dans l'espace $C_0(\mathbf{T}^3\times\mathbf{R}^3)$ des fonctions continues tendant vers 0 à l'infini. La distance $d_{[0,T]}$ définit la topologie de Skorokhod sur l'espace $D([0,T],\mathscr{M}_+^1(\mathbf{T}^3\times\mathbf{R}^3))$ (voir (Billingsley, 1999), chapitre 3, section 12).

Le principaux résultats concernant les grandes déviations de la mesure empirique d'un système de sphères dures dans la limite de Boltzmann–Grad sont résumés dans le théorème suivant, qui regroupe les Théorèmes 3 et 9 de Bodineau, Gallagher, Saint-Raymond et Simonella (2023).

Théorème 3.1. *Supposons que f^{in} vérifie* (17). *Pour tout $r > 0$, il existe $\alpha \equiv \alpha[r, \beta_0, C_0] > 0$ et $T_r \in]0, T_\alpha[$ (où T_α est défini dans la Proposition 2.4) tels que, dans la limite de Boltzmann–Grad*
(*a*) *pour tout fermé $\mathscr{G} \subset D([0,T_r], \mathscr{M}^1_+(\mathbf{T}^3 \times \mathbf{R}^3))$ pour la distance $d_{[0,T_r]}$*

$$\overline{\lim_{\varepsilon\to 0^+}} \frac{1}{\mu_\varepsilon} \ln \mathbb{P}_\varepsilon(\rho^\varepsilon \in \mathscr{G}) \leqslant - \inf_{\phi\in\mathscr{G}} \mathbf{K}^*(T_r, \phi)\,,$$

(*b*) *pour tout ouvert $\mathscr{O} \subset D([0,T_r], \mathscr{M}^1_+(\mathbf{T}^3 \times \mathbf{R}^3))$ pour la distance $d_{[0,T_r]}$*

$$\underline{\lim_{\varepsilon\to 0^+}} \frac{1}{\mu_\varepsilon} \ln \mathbb{P}_\varepsilon(\rho^\varepsilon \in \mathscr{O}) \geqslant - \inf_{\phi\in\mathscr{O}\cap\mathscr{R}_{r,T}} \mathbf{K}^*(T_r, \phi)\,.$$

Ici $\mathbf{K}^*$ est la transformée de Legendre de $\mathbf{K}$: pour tout $\phi \in D([0,T], \mathscr{M}^1_+(\mathbf{T}^3 \times \mathbf{R}^3))$

$$\mathbf{K}^*(T,\phi) := \sup_{g\in\mathbf{B}_\alpha} \left(\langle \phi(T,\cdot), g(T,\cdot)\rangle - \langle\langle \phi, (\partial_t + v\cdot\nabla_x) g\rangle\rangle - \mathbf{K}(T,g)\right)\,,$$

où on a noté

$$\langle\langle \phi, q\rangle\rangle := \int_0^T \langle \phi(t,\cdot), q(t,\cdot)\rangle dt\,.$$

En particulier, pour tout $\phi \in \mathscr{R}_{r,T_r}$

$$\lim_{\eta\to 0^+} \overline{\lim_{\varepsilon\to 0^+}} \frac{1}{\mu_\varepsilon} \ln \mathbb{P}_\varepsilon \left(d_{[0,T_r]}(\rho^\varepsilon, \phi) \leqslant \eta \right) = -\mathbf{K}^*(T_r,\phi)\,,$$
$$\lim_{\eta\to 0^+} \underline{\lim_{\varepsilon\to 0^+}} \frac{1}{\mu_\varepsilon} \ln \mathbb{P}_\varepsilon \left(d_{[0,T_r]}(\rho^\varepsilon, \phi) < \eta \right) = -\mathbf{K}^*(T_r,\phi)\,.$$

Ce résultat fait l'objet du chapitre 7 de (Bodineau, Gallagher, Saint-Raymond et Simonella, 2023). Sa démonstration met en jeu quelques propriétés importantes sur l'équation de Hamilton–Jacobi de la section précédente, dont nous allons donner une idée.

Précisons que ce type de résultat avait été conjecturé auparavant par Rezakhanlou dans (Rezakhanlou et Villani, 2008) et par Bouchet (2020) — avec toutefois une formule apparemment différente de celle du Théorème 3.1 pour la fonctionnelle des grandes déviations $\mathbf{K}^*(T,\phi)$: on y reviendra dans la suite (voir la Proposition 3.4).

3.1. Méthode des caractéristiques pour l'équation de Hamilton–Jacobi

Soit $t \in [0, T_\alpha]$. Considérons le système hamiltonien

$$\begin{aligned} &(\partial_s + v\cdot\nabla_x)\phi_t(s,\cdot) = \frac{\partial \mathscr{H}}{\partial q}(\phi_t(s,\cdot), q_t(s,\cdot))\,, \qquad \phi_t(0,\cdot) = f^{in} e^{q(0,\cdot)}\,, \\ &(\partial_s + v\cdot\nabla_x)(q_t - g)(s,\cdot) = -\frac{\partial \mathscr{H}}{\partial p}(\phi_t(s,\cdot), q_t(s,\cdot))\,, \quad q_t(t,\cdot) = g(t,\cdot)\,. \end{aligned} \tag{39}$$

Grâce à la formule (36), on vérifie sans peine que la première de ces deux équations n'est autre que (37)-(38) : ce point sera établi plus loin (voir section 3.2).

Proposition 3.2. *Soit $\alpha > 0$ et $g \in \mathbf{B}_\alpha$. Soit (ϕ_t, q_t) solution du système* (39). *Posons*

$$\hat{\mathbf{K}}(t,g) := \langle f^{in}, e^{q_t(0,\cdot)} - 1\rangle + \langle\langle \phi_t, (\partial_s + v\cdot\nabla_x)(q_t - g)\rangle\rangle + \int_0^t \mathscr{H}(\phi_t(s,\cdot), q_t(s,\cdot))ds\,. \tag{40}$$

Alors $\hat{\mathbf{K}}$ est solution de (30) *sur $[0,t]$, et vérifie en outre*

$$\frac{\partial \hat{\mathbf{K}}(t,g)}{\partial g(t)} = \phi_t(t)\,, \qquad \frac{\partial \hat{\mathbf{K}}(t,g)}{\partial(\partial_s g + v\cdot\nabla_x g)} = -\phi_t\,.$$

La théorie classique de l'équation de Hamilton–Jacobi dit que le graphe de la différentielle de la solution de l'équation de Hamilton–Jacobi en la variable d'espace est invariant par le flot des équations de Hamilton. C'est bien ce qu'exprime la proposition ci-dessus, à condition de considérer que les soi-disant variables $g(t,\cdot)$ et $(\partial_s + v\cdot\nabla_x)g$ sont non seulement indépendantes comme on l'a dit plus haut, mais encore conjuguées au sens de la théorie des systèmes hamiltoniens.

Cette proposition suggère d'étudier (a) l'unicité de la solution de l'équation d'Hamilton–Jacobi (30), et (b) l'existence et unicité pour le système des équations de Hamilton (39) (*cf.* section 7.2 dans le chapitre 7 de (Bodineau, Gallagher, Saint-Raymond et Simonella, 2023)). Cette étude est largement basée sur une variante du théorème de Cauchy-Kowalevski de Nirenberg (1972) et Ovsjannikov (1971) (voir Appendice de (Bodineau, Gallagher, Saint-Raymond et Simonella, 2023)), déjà utilisée dans la preuve du théorème de Lanford (voir section 5.3 dans (Golse, 2014)). Au terme de cette étude, on aboutit à l'énoncé suivant.

Proposition 3.3. *Soit $\alpha > 0$. Il existe $T_\alpha^* \in]0, T_\alpha]$ tel que la fonctionnelle $\hat{\mathbf{K}}$ soit définie sur $[0, T_\alpha^*] \times \mathbf{B}_\alpha$, et*

$$\mathbf{K}(t,g) = \hat{\mathbf{K}}(t,g)\,, \quad \text{pour tout } (t,g) \in [0, T_\alpha^*] \times \mathbf{B}_\alpha\,.$$

3.2. Transformée de Legendre

Soit $\bar\phi$, solution de (37) avec condition initiale (38) pour une fonction poids lipschitzienne $\bar q$ telle que $\|\bar q\|_{W^{1,\infty}([0,T_0]\times\mathbf{T}^3\times\mathbf{R}^3)} < r$. Rappelons que, d'après la formule (36),

$$\begin{aligned}
&\left\langle \frac{\partial \mathscr{H}}{\partial q}(p,q), \psi \right\rangle \\
&= \tfrac{1}{2}\int_{\mathbf{T}^3\times(\mathbf{R}^3)^2\times\mathbf{S}^2} e^{\mathbf{D}q(x,v,x,v_*,\omega)} \mathbf{D}\psi(x,v,x,v_*,\omega)((v-v_*)\cdot\omega)_+ d\omega p(x,v)p(x,v_*)dvdv_*dx \\
&= \int_{\mathbf{T}^3\times(\mathbf{R}^3)^2\times\mathbf{S}^2} e^{\mathbf{D}q(x,v,x,v_*,\omega)} (\psi(x,v') - \psi(x,v))((v-v_*)\cdot\omega)_+ d\omega p(x,v)p(x,v_*)dvdv_*dx \\
&= \int_{\mathbf{T}^3\times(\mathbf{R}^3)^2\times\mathbf{S}^2} \left(e^{-\mathbf{D}q(x,v,x,v_*,\omega)} p(x,v')p(x,v'_*) - e^{\mathbf{D}q(x,v,x,v_*,\omega)} p(x,v)p(x,v_*)\right) \\
&\qquad \times((v-v_*)\cdot\omega)_+ d\omega dv_* \psi(x,v) dvdx.
\end{aligned}$$

La deuxième égalité ci-dessus s'obtient par la même démonstration que celle de la troisième égalité dans la formulation faible de l'intégrale des collisions de Boltzmann de la section 1.6 (par échange de v et v_*). La troisième égalité s'obtient comme la seconde égalité dans la formulation faible de l'intégrale des collisions de Boltzmann de la section 1.6 (en utilisant la transformation $(v, v_*) \mapsto (v', v'_*)$ à ω fixé définie par (16)). Ainsi

$$\frac{\partial \mathscr{H}}{\partial q}(p,q) = \int_{\mathbf{R}^3\times\mathbf{S}^2} \left(e^{-\mathbf{D}q(x,v,x,v_*,\omega)} p(x,v') p(x,v'_*) - e^{\mathbf{D}q(x,v,x,v_*,\omega)} p(x,v) p(x,v_*) \right) \times ((v-v_*)\cdot\omega)_+ d\omega dv_* ,$$

ce qui montre que l'équation de Boltzmann à poids (37) avec condition initiale (38) coïncide avec la première équation du système hamiltonien (39), comme annoncé plus haut.

En utilisant de nouveau une variante du théorème de Cauchy-Kowalevski de Nirenberg (1972) et Ovsjannikov (1971), on montre que le problème de Cauchy (37)-(38) admet une unique solution sur l'intervalle $[0, T_0 e^{-5r}]$, vérifiant la borne

$$\sup_{0\leqslant t\leqslant T_0 e^{-5r}} \|\bar\phi(t,\cdot)\exp(\tfrac{1}{4}\beta_0|v|^2)\|_{L^\infty(\mathbf{T}^3\times\mathbf{R}^3)} \leqslant 4C_0 e^r .$$

On vérifie de même que pour l'équation de Boltzmann originale (14) que les solutions de l'équation de Boltzmann à poids (37) sont de masse, d'impulsion et d'énergie constantes :

$$\langle (\partial_s + v\cdot\nabla_x)\bar\phi(s,\cdot), 1\rangle = \langle (\partial_s + v\cdot\nabla_x)\bar\phi(s,\cdot), v_i\rangle = \langle (\partial_s + v\cdot\nabla_x)\bar\phi(s,\cdot), |v|^2\rangle = 0$$

pour $i = 1,2,3$. La démonstration suit de près celle des lois de conservation locales établies dans la section 1.6 pour l'équation de Boltzmann originale.

Proposition 3.4. *Supposons que f^{in} vérifie* (17). *Pour tout $r > 0$, il existe $\alpha \equiv \alpha[r,\beta_0,C_0] > 0$ et $T_r \in]0, T_\alpha[$ tels que*

$$\mathbf{K}^*(t,\phi) = \int_{\mathbf{T}^3\times\mathbf{R}^3} \left(\phi(0,z)\ln\left(\frac{\phi(0,z)}{f^{in}(z)}\right) - \phi(0,z) + f^{in}(z)\right) dz + \sup_{q\in L^\infty([0,t]\times\mathbf{T}^3\times\mathbf{R}^3)} \left(\langle\langle (\partial_s + v\cdot\nabla_x)\phi, q\rangle\rangle - \int_0^t \mathscr{H}(\phi(s,\cdot), q(s,\cdot))ds \right) .$$

Or la première équation du système hamiltonien (39) exprime précisément le fait que la fonction $\bar q$, c'est-à-dire la fonction poids dans (37)-(38), est un point critique pour le problème variationnel de la proposition ci-dessus. Comme d'autre part la solution $\bar\phi$ vérifie $\bar\phi \geqslant 0$, on déduit de la formule (31) que

$$q \mapsto \mathscr{H}(\bar\phi(s,\cdot), q)$$

est convexe. Par conséquent

$$\sup_{q\in L^\infty([0,t]\times\mathbf{T}^3\times\mathbf{R}^3)}\left(\langle\langle(\partial_s+v\cdot\nabla_x)\bar\phi,q\rangle\rangle-\int_0^t\mathscr{H}(\bar\phi(s,\cdot),q(s,\cdot))ds\right)$$
$$=\langle\langle(\partial_s+v\cdot\nabla_x)\bar\phi,\bar q\rangle\rangle-\int_0^t\mathscr{H}(\bar\phi(s,\cdot),\bar q(s,\cdot))ds\,.$$

Cette formule permet donc de calculer « explicitement » la fonctionnelle des grandes déviations intervenant dans le Théorème 3.1 par la méthode des caractéristiques, c'est-à-dire en résolvant le système hamiltonien (39).

La Proposition 3.4 montre que la fonctionnelle des grandes déviations obtenue dans le Théorème 3.1 coïncide bien avec celle conjecturée dans les travaux de Rezakhanlou (Conjecture 4 du texte de Rezakhanlou dans (Rezakhanlou et Villani, 2008)), et de Bouchet (formule (27) de (Bouchet, 2020)).

Terminons par une remarque intéressante — qui joue un rôle dans la démonstration des résultats de cette section (voir (Bodineau, Gallagher, Saint-Raymond et Simonella, 2023), chapitre 7, section 7.2).

Il est commode de changer les variables (p,q) du hamiltonien (31) en posant $P=pe^{-q}$ et $Q=e^q$, d'où

$$H(p,q)=\tfrac12\int_{\mathbf{T}^3\times(\mathbf{R}^3)^2\times\mathbf{S}^2}\big(Q(x,v')Q(x,v'_*)-Q(x,v)Q(x,v_*)\big)\\ \times((v-v_*)\cdot\omega)_+d\omega\,P(x,dv)P(x,dv_*)dx=H'(P,Q)\,. \tag{41}$$

Un calcul simple montre que

$$\frac{\partial H'}{\partial P}(P,Q)=\int_{\mathbf{R}^3\times\mathbf{S}^2}\big(Q(x,v')Q(x,v'_*)-Q(x,v)Q(x,v_*)\big)\,P(x,v_*)((v-v_*)\cdot\omega)_+dv_*d\omega\,,$$
$$\frac{\partial H'}{\partial Q}(P,Q)=\int_{\mathbf{R}^3\times\mathbf{S}^2}\big(P(x,v')P(x,v'_*)-P(x,v)P(x,v_*)\big)\,Q(x,v_*)((v-v_*)\cdot\omega)_+dv_*d\omega\,.$$

On peut alors réécrire le système (39) au moyen des nouvelles fonctions inconnues $\gamma_t:=\phi_te^{-q_t}$ et $\eta_t=e^{q_t}$: on trouve sans difficulté [8] que

$$(\partial_s+v\cdot\nabla_x)\gamma_t+\gamma_t(\partial_s+v\cdot\nabla_x)g=\frac{\partial H'}{\partial Q}(\gamma_t,\eta_t)$$
$$=+\int_{\mathbf{R}^3\times\mathbf{S}^2}(\gamma'_t\gamma'_{t*}-\gamma_t\gamma_{t*})\eta_{t*}((v-v_*\cdot\omega))_+dv_*d\omega\,,$$
$$(\partial_s+v\cdot\nabla_x)\eta_t-\eta_t(\partial_s+v\cdot\nabla_x)g=-\frac{\partial H'}{\partial P}(\gamma_t,\eta_t)$$
$$=-\int_{\mathbf{R}^3\times\mathbf{S}^2}(\eta'_t\eta'_{t*}-\eta_t\eta_{t*})\gamma_{t*}((v-v_*\cdot\omega))_+dv_*d\omega\,,$$

[8] En notant $f'=f(t,x,v')$ et $f'_*=f(t,x,v'_*)$, tandis que $f_*=f(t,x,v_*)$, et que les vitesses v',v'_* sont données par (16) en fonction de v,v_* et ω.

système posé sur $[0,t] \times \mathbf{T}^3 \times \mathbf{R}^3$ avec les conditions aux limites

$$\gamma_t(0,\cdot) = f^{in}\,, \qquad \eta_t(t,\cdot) = e^{g(t,\cdot)}\,.$$

La structure de ce nouveau système hamiltonien est intéressante : il s'agit de deux équations de Boltzmann à poids, mais *avec des sens opposés de propagation du temps*. Comme il s'agit de prouver l'existence et l'unicité de solutions à des équations de type Boltzmann sans utiliser la flèche du temps « naturelle » , à savoir celle pour laquelle

$$t \mapsto \iint_{\mathbf{T}^3\times\mathbf{R}^3} \phi(t,x,v)\ln\phi(t,x,v)dxdv \text{ est décroissante,}$$

il n'est pas très étonnant que la démonstration repose de nouveau sur le théorème de Cauchy-Kowalevski abstrait de Nirenberg (1972) et Ovsjannikov (1971) : voir la Proposition 7.2.3 de (Bodineau, Gallagher, Saint-Raymond et Simonella, 2023).

4. Fluctuations

Dans cette section, on va s'intéresser au champ de fluctuations (21) de la mesure empirique ρ_t^ε dans la limite de Boltzmann–Grad. Rappelons-en la définition :

$$\begin{aligned}\langle \zeta_t^\varepsilon, \phi\rangle :=& \sqrt{\mu_\varepsilon}\left(\langle \rho_t^\varepsilon,\phi\rangle - \mathbb{E}_\varepsilon(\langle \rho_t^\varepsilon,\phi\rangle)\right)\\ =& \frac{1}{\sqrt{\mu_\varepsilon}}\left(\sum_{i=1}^N \phi(\mathbf{z}_i(t)) - \mu_\varepsilon \int_{\mathbf{T}^3\times\mathbf{R}^3} F_1^\varepsilon(t,z)\phi(z)dz\right)\,.\end{aligned}$$

Puisqu'il s'agit de décrire les fluctuations de la mesure empirique autour de F_1^ε qui converge vers une solution de l'équation de Boltzmann, nous aurons évidemment besoin des opérateurs linéarisés (adjoints) associés à l'équation de Boltzmann (14) autour d'une solution f définie sur $[0,T]\times\mathbf{T}^3\times\mathbf{R}^3$:

$$\begin{aligned}\mathscr{L}_t^*\phi(z) :=& v\cdot\nabla_x\phi(z) + \mathbf{L}_t^*\phi(z)\,, \quad \text{où on a posé}\\ \mathbf{L}_t^*\phi(z) :=& \int_{\mathbf{R}^3\times\mathbf{S}^2} \mathbf{D}\phi(x,v,x,v^*,\omega) f(t,x,v_*)((v-v_*)\cdot\omega)_+ dv_*d\omega\,.\end{aligned}$$

Théorème 4.1. *Supposons que f^{in} vérifie* (17)*. Il existe $T^* \equiv T^*[C_0,\beta_0] > 0$ tel que, dans la limite de Boltzmann–Grad, ζ_t^ε converge en loi vers ζ_t pour tout $t\in[0,T^*]$ lorsque $\varepsilon\to 0^+$, où ζ_t est le processus gaussien centré solution de*

$$d\zeta_t = \mathscr{L}_t\zeta_t dt + db_t\,, \qquad 0<t<T^*\,, \tag{42}$$

et où b_t est un bruit gaussien delta-corrélé en t,x, de covariance

$$\begin{aligned}\mathbf{Cov}(t,\phi,\psi) :=& \tfrac{1}{2}\int_{\mathbf{T}^3\times(\mathbf{R}^3)^2\times\mathbf{S}^2} \mathbf{D}\phi(x,v,x,v_*,\omega)\mathbf{D}\psi(x,v,x,v_*,\omega)\\ &\times f(t,x,v)f(t,x,v_*)((v-v_*)\cdot\omega)_+ dxdvdv_*d\omega\,.\end{aligned}$$

De plus ζ_0 est le champ gaussien centré de covariance

$$\mathbf{E}(\langle\zeta_0,\phi\rangle\langle\zeta_0,\psi\rangle) = \int_{\mathbf{T}^3\times\mathbf{R}^3} \phi(z)\psi(z)f^{in}(z)dz\,.$$

Spohn avait conjecturé ce résultat (et démontré la convergence de la covariance) dans (Spohn, 1981) : voir l'équation (4.1) de cette référence, ainsi que l'article antérieur (Van Beijeren et al., 1980) dans le cas où la solution de l'équation de Boltzmann considérée est un équilibre maxwellien (20).

4.1. Solutions faibles de (42)

Précisons la notion de solution de l'équation de Boltzmann aux fluctuations (42). Soit $U(t,s)$ le flot engendré par l'opérateur de Boltzmann linéarisé $\mathscr{L}_t$: autrement dit

$$\partial_t U(t,s)\phi - \mathscr{L}_t U(t,s)\phi = 0\,, \qquad U(s,s)\phi = \phi\,, \qquad 0\leqslant s\leqslant t\leqslant T^*\,.$$

En dérivant par rapport à t_2 la relation de flot $U(t_1,t_3) = U(t_1,t_2)U(t_2,t_3)$, il vient

$$\partial_s U(t,s)\phi + U(t,s)\mathscr{L}_s\phi = 0\,, \qquad U(t,t)\phi = \phi\,, \qquad 0\leqslant s\leqslant t\leqslant T^*\,.$$

Puis, par passage à l'adjoint, on trouve que

$$\begin{aligned}
&\partial_s U(t,s)^*\phi + \mathscr{L}_s^* U(t,s)^*\phi = 0\,, \qquad U(t,t)^*\phi = \phi\,, \qquad 0\leqslant s\leqslant t\leqslant T^*\,,\\
&\partial_t U(t,s)^*\phi - U(t,s)^*\mathscr{L}_t^*\phi = 0\,, \qquad U(s,s)^*\phi = \phi\,, \qquad 0\leqslant s\leqslant t\leqslant T^*\,.
\end{aligned}$$

Formellement, la solution de (42) vérifie

$$\langle\zeta_t,\phi\rangle = \langle\zeta_0, U(t,0)^*\phi\rangle + \int_0^t \langle db_s, U(t,s)^*\phi\rangle$$

pour toute fonction test ϕ. Donc

$$\begin{aligned}
\hat{\mathscr{C}}(t,s,\psi,\phi) := \mathbb{E}(\langle\zeta_t,\psi\rangle\langle\zeta_s,\phi\rangle) =& \mathbb{E}(\langle\zeta_0, U(t,0)^*\psi\rangle\langle\zeta_0, U(s,0)^*\phi\rangle)\\
&+ \mathbb{E}\left(\int_0^t\int_0^s \langle db_\tau, U(t,\tau)^*\phi\rangle\langle db_\sigma, U(s,\sigma)^*\phi\rangle\right)\\
&+ \mathbb{E}\left(\langle\zeta_0, U(t,0)^*\psi\rangle\int_0^s \langle db_\sigma, U(s,\sigma)^*\phi\rangle\right)\\
&+ \mathbb{E}\left(\langle\zeta_0, U(s,0)^*\psi\rangle\int_0^t \langle db_\tau, U(t,\tau)^*\phi\rangle\right)\,,
\end{aligned}$$

ce qui se réécrit

$$\hat{\mathscr{C}}(t,s,\psi,\phi) = \mathbb{E}(\langle\zeta_0, U(t,0)^*\psi\rangle\langle\zeta_0, U(s,0)^*\phi\rangle) + \int_0^s \mathbf{Cov}(\sigma, U(t,\sigma)^*\psi, U(s,\sigma)^*\phi)d\sigma\,. \tag{43}$$

Définition 4.2. Une solution faible de (42) est un processus gaussien centré dont la covariance vérifie l'identité ci-dessus. (On rappelle en effet que la loi d'un processus gaussien centré est complètement déterminée par sa covariance).

Dérivons chaque membre de l'identité ci-dessus par rapport à t :

$$\begin{aligned}\partial_t \hat{\mathscr{C}}(t,s,\psi,\phi) &= \mathbb{E}(\langle \zeta_0, U(t,0)^* \mathscr{L}_t^* \psi\rangle \langle \zeta_0, U(s,0)^*\phi\rangle) \\ &\qquad + \int_0^s \mathbf{Cov}(\sigma, U(t,\sigma)^*\mathscr{L}_t^*\psi, U(s,\sigma)^*\phi) d\sigma \\ &= \hat{\mathscr{C}}(t,s,\mathscr{L}_t^*\psi,\phi)\,, \quad t>s\,.\end{aligned}$$

D'autre part, pour $t=s$

$$\begin{aligned}\frac{d}{dt}\hat{\mathscr{C}}(t,t,\psi,\phi) &= \big(\partial_t \hat{\mathscr{C}}(t,s,\psi,\phi) + \partial_s \hat{\mathscr{C}}(t,s,\psi,\phi)\big)\Big|_{t=s} \\ &= \mathbb{E}(\langle \zeta_0, U(t,0)^*\mathscr{L}_t^*\psi\rangle\langle\zeta_0, U(t,0)^*\phi\rangle) \\ &\quad + \int_0^t \mathbf{Cov}(\sigma, U(t,\sigma)^*\mathscr{L}_t^*\psi, U(t,\sigma)^*\phi)d\sigma \\ &\quad + \mathbb{E}(\langle \zeta_0, U(t,0)^*\psi\rangle\langle\zeta_0, U(t,0)^*\mathscr{L}_t^*\phi\rangle) \\ &\quad + \int_0^t \mathbf{Cov}(\sigma, U(t,\sigma)^*\psi, U(t,\sigma)^*\mathscr{L}_t^*\phi)d\sigma + \mathbf{Cov}(t,\psi,\phi)\,,\end{aligned}$$

de sorte que

$$\frac{d}{dt}\hat{\mathscr{C}}(t,t,\psi,\phi) = \hat{\mathscr{C}}(t,t,\mathscr{L}_t^*\psi,\phi) + \hat{\mathscr{C}}(t,t,\psi,\mathscr{L}_t^*\phi) + \mathbf{Cov}(t,\psi,\phi)\,.$$

Évidemment, il s'agit là d'un calcul formel (en toute rigueur, il faudrait revenir à la formulation intégrale en temps équivalente). Toutefois, nous allons, dans la section suivante, faire le lien entre ces formules et la solution de l'équation de Hamilton–Jacobi fonctionnelle (30).

Avant cela, on va modifier très légèrement le calcul ci-dessus de façon à absorber la dérivée en temps le long des caractéristiques de l'opérateur d'advection $v\cdot\nabla_x$. Rappelons la notation $A_t = e^{-tv\cdot\nabla_x}$, c'est-à-dire que $A_t\phi(x,v) = \phi(x-tv,v)$.

Comme $(\phi,\psi)\mapsto\hat{\mathscr{C}}(t,s,\psi,\phi)$ est une forme bilinéaire, on calcule facilement

$$\begin{aligned}\frac{d}{dt}\hat{\mathscr{C}}(t,t,A_t\psi,A_t\phi) &= \hat{\mathscr{C}}(t,t,\mathscr{L}_t^*A_t\psi,A_t\phi) + \hat{\mathscr{C}}(t,t,A_t\psi,\mathscr{L}_t^*A_t\phi) + \mathbf{Cov}(t,A_t\psi,A_t\phi) \\ &\quad - \hat{\mathscr{C}}(t,t,v\cdot\nabla_x A_t\psi, A_t\phi) - \hat{\mathscr{C}}(t,t,A_t\psi, v\cdot\nabla_x A_t\phi) \\ &= \hat{\mathscr{C}}(t,t,\mathbf{L}_t^*A_t\psi,A_t\phi) + \hat{\mathscr{C}}(t,t,A_t\psi,\mathbf{L}_t^*A_t\phi) + \mathbf{Cov}(t,A_t\psi,A_t\phi)\,.\end{aligned}$$

D'autre part, pour $t>s$, on a

$$\begin{aligned}\frac{d}{dt}\hat{\mathscr{C}}(t,s,A_t\psi,\phi_s) &= \hat{\mathscr{C}}(t,s,\mathscr{L}_t^*A_t\psi,\phi_s) - \hat{\mathscr{C}}(t,s,v\cdot\nabla_x A_t\psi,\phi_s) \\ &= \hat{\mathscr{C}}(t,s,\mathbf{L}_t^*A_t\psi,\phi_s)\,,\end{aligned}$$

de sorte que

$$\hat{\mathscr{C}}(t,s,A_t\psi,\phi_s) = \hat{\mathscr{C}}(s,s,A_s\psi,\phi_s\sigma) + \int_s^t \hat{\mathscr{C}}(\sigma,s,\mathbf{L}_\sigma^* A_\sigma\psi,\phi_s)\,,$$

identité dont on intègre ensuite chaque membre par rapport à s, pour aboutir à

$$\int_0^t \hat{\mathscr{C}}(t,s,A_t\psi,\phi_s)ds = \int_0^t \left(\hat{\mathscr{C}}(s,s,A_s\psi,\phi_s) + \int_s^t \hat{\mathscr{C}}(\sigma,s,\mathbf{L}_\sigma^* A_\sigma\psi,\phi_s)\right) ds\,.$$

En remplaçant ψ par $A_{-t}\psi$, on trouve finalement que

$$\int_0^t \hat{\mathscr{C}}(t,s,\psi,\phi_s)ds = \int_0^t \left(\hat{\mathscr{C}}(s,s,A_{s-t}\psi,\phi_s) + \int_s^t \hat{\mathscr{C}}(\sigma,s,\mathbf{L}_\sigma^* A_{\sigma-t}\psi,\phi_s)d\sigma\right) ds\,.$$

En résumé, la solution ζ_t de (42) est un processus gaussien centré dont la covariance $\hat{\mathscr{C}}(t,s,\psi,\phi)$ est solution du système d'équations intégrales

$$\begin{cases} \hat{\mathscr{C}}(t,t,\psi,\phi) = \hat{\mathscr{C}}(0,0,A_{-t}\psi,A_{-t}\phi) + \displaystyle\int_0^t \mathbf{Cov}(s,A_{s-t}\psi,A_{s-t}\phi) \\ \qquad\qquad + \displaystyle\int_0^t \left(\hat{\mathscr{C}}(s,s,\mathbf{L}_s^* A_{s-t}\psi,A_{s-t}\phi) + \hat{\mathscr{C}}(s,s,A_{s-t}\psi,\mathbf{L}_s^* A_{s-t}\phi)\right) ds\,, \\ \displaystyle\int_0^t \hat{\mathscr{C}}(t,s,\psi,\phi_s)ds = \int_0^t \left(\hat{\mathscr{C}}(s,s,A_{s-t}\psi,\phi_s) + \int_s^t \hat{\mathscr{C}}(\sigma,s,\mathbf{L}_\sigma^* A_{\sigma-t}\psi,\phi_s)d\sigma\right) ds\,. \end{cases} \tag{44}$$

On y ajoute l'information provenant de la condition initiale :

$$\hat{\mathscr{C}}(0,0,\psi,\phi) = \int_{\mathbf{T}^3\times\mathbf{R}^3} \phi(z)\psi(z)f^{in}(z)dz\,. \tag{45}$$

Or on va voir que l'équation d'Hamilton–Jacobi permet justement de montrer que la limite de la covariance du champ de fluctuation ζ_t^ε lorsque $\varepsilon \to 0^+$ vérifie le même système (44) d'équations intégrales que ci-dessus. Ce résultat, que nous allons présenter dans la section suivante, est évidemment une étape fondamentale dans la démonstration du Théorème 4.1.

4.2. Covariance limite

La covariance du champ de fluctuations est définie comme suit : pour toute paire de fonctions test $\phi,\psi \in C_b(\mathbf{T}^3\times\mathbf{R}^3)$, on pose

$$\mathscr{C}_\varepsilon(t,s,\psi,\phi) := \mathbb{E}_\varepsilon(\langle\zeta_t^\varepsilon,\psi\rangle\langle\zeta_s^\varepsilon,\phi\rangle)\,, \qquad 0 \leqslant s \leqslant t\,.$$

L'équation de Hamilton–Jacobi (30) va nous permettre de caractériser la covariance limite du champ de fluctuations (limite de $\mathscr{C}_\varepsilon(t,s,\psi,\phi)$ lorsque $\varepsilon \to 0^+$).

Proposition 4.3. *Supposons que f^{in} vérifie* (17)*, et soit f la solution du problème de Cauchy* (14)-(18) *sur $[0,T^*]\times\mathbf{T}^3\times\mathbf{R}^3$ où $T^*\equiv T^*[C_0,\beta_0]>0$. Pour toute paire de fonctions test $\phi,\psi\in C_b(\mathbf{T}^3\times\mathbf{R}^3)$ et tous $s\leqslant t$ dans $[0,T_0]$,*

$$\mathscr{C}_\varepsilon(t,s,\psi,\phi)\to\mathscr{C}(t,s,\psi,\phi)\quad\text{ lorsque }\varepsilon\to 0^+\,.$$

De plus, la covariance limite $\mathscr{C}$ est solution du système (44) *d'équations intégrales, et vérifie la condition initiale* (45)*.*

Démonstration. Observons pour commencer que

$$\frac{\partial^2\mathbf{K}(t,g)}{\partial g(t)^2}\Big|_{g=0}\cdot(\psi,\phi)=\mathscr{C}(t,t,\psi,\phi)\,.$$

Partons de (32) ; en dérivant une fois de plus par rapport à $g(t,\cdot)$, il vient

$$\frac{\partial^2\mathbf{K}(0,g)}{\partial g(t)^2}\Big|_{g=0}\cdot(\psi,\phi)=\int_{\mathbf{T}^3\times\mathbf{R}^3}A_{-t}\psi(z)A_{-t}\phi(z)f^{in}(z)dz=\mathscr{C}(0,0,A_{-t}\psi,A_{-t}\phi)\,.$$

Revenons ensuite à l'identité (33), dont on dérive chaque membre au point $g=0$ par rapport à $g(t,\cdot)$ dans la direction ψ, pour trouver

$$\begin{aligned}\frac{\partial^2\mathbf{K}(t,g)}{\partial g(t)^2}\cdot(\phi,\psi)&=\frac{\partial^2\mathbf{K}(0,g)}{\partial g(t)^2}\cdot(\phi,\psi)\\
&+\int_0^t\frac{\partial^2\mathscr{H}}{\partial p^2}\left(\frac{\partial\mathbf{K}(s,g)}{\partial g(s)},g(s)\right)\cdot\left(\frac{\partial^2\mathbf{K}(s,g)}{\partial g(t)\partial g(s)}\cdot(\phi,\cdot),\frac{\partial^2\mathbf{K}(s,g)}{\partial g(t)\partial g(s)}\cdot(\psi,\cdot)\right)ds\\
&+\int_0^t\frac{\partial^2\mathscr{H}}{\partial q\partial p}\left(\frac{\partial\mathbf{K}(s,g)}{\partial g(s)},g(s)\right)\cdot\left(A_{s-t}\phi,\frac{\partial^2\mathbf{K}(s,g)}{\partial g(t)\partial g(s)}\cdot(\psi,\cdot)\right)ds\\
&+\int_0^t\left\langle\frac{\partial\mathscr{H}}{\partial p}\left(\frac{\partial\mathbf{K}(s,g)}{\partial g(s)},g(s)\right),\frac{\partial^3\mathbf{K}(s,g)}{\partial g(t)^2\partial g(s)}\cdot(\phi,\psi,\cdot)\right\rangle ds\\
&+\int_0^t\frac{\partial^2\mathscr{H}}{\partial p\partial q}\left(\frac{\partial\mathbf{K}(s,g)}{\partial g(s)},g(s)\right)\cdot\left(\frac{\partial^2\mathbf{K}(s,g)}{\partial g(t)\partial g(s)}\cdot(\phi,\cdot),A_{s-t}\psi\right)ds\\
&+\int_0^t\frac{\partial^2\mathscr{H}}{\partial q^2}\left(\frac{\partial\mathbf{K}(s,g)}{\partial g(s)},g(s)\right)\cdot(A_{s-t}\phi,A_{s-t}\psi)\,ds\,.\end{aligned}$$

En utilisant la règle de dérivation des fonctions composées et le Lemme 2.8, on a

$$\frac{\partial^2\mathbf{K}(s,g)}{\partial g(t)\partial g(s)}\cdot(\psi,\cdot)=\frac{\partial^2\mathbf{K}(s,g)}{\partial g(s)^2}\cdot(A_{s-t}\psi,\cdot)\,.$$

D'autre part

$$\mathscr{H}(p,0)=0\,,\quad\text{d'où}\quad\frac{\partial\mathscr{H}}{\partial p}(p,0)=0\quad\text{et}\quad\frac{\partial^2\mathscr{H}}{\partial p^2}(p,0)=0\,,$$

de sorte qu'en faisant $g = 0$ dans l'identité ci-dessus, il vient

$$\begin{aligned}\mathscr{C}(t,t,\phi,\psi) &= \mathscr{C}(0,0,A_{-t}\phi,A_{-t}\psi)\\ &+\int_0^t \frac{\partial^2\mathscr{H}}{\partial q\partial p}\,(f(s,\cdot),0)\cdot\left(A_{s-t}\phi,\frac{\partial^2\mathbf{K}(s,0)}{\partial g(s)^2}\cdot(A_{s-t}\psi,\cdot)\right)ds\\ &+\int_0^t \frac{\partial^2\mathscr{H}}{\partial p\partial q}\,(f(s,\cdot),0)\cdot\left(\frac{\partial^2\mathbf{K}(s,0)}{\partial g(s)^2}\cdot(A_{s-t}\phi,\cdot),A_{s-t}\psi\right)ds\\ &+\int_0^t \frac{\partial^2\mathscr{H}}{\partial q^2}\,(f(s,\cdot),0)\cdot(A_{s-t}\phi,A_{s-t}\psi)\,ds\end{aligned}$$

puisqu'on sait, d'après la Proposition 2.7, que $\frac{\partial\mathbf{K}(s,0)}{\partial g(s)} = f(s,\cdot)$, où f est la solution de l'équation de Boltzmann (14).

Ensuite, la formule (31) montre que

$$\begin{aligned}\frac{\partial^2\mathscr{H}}{\partial q^2}\,(f(s,\cdot),0)\cdot(\Phi,\Psi) &= \tfrac12\int_{\mathbf{T}^3\times(\mathbf{R}^3)^2\times\mathbf{S}^2}\mathbf{D}\Phi(s,x,v,x,v_*,\omega)\mathbf{D}\Psi(s,x,v,x,v_*,\omega)\\ &\times f(s,x,v)f(s,x,v_*)((v-v_*)\cdot\omega))_+ d\omega dv dv_* dx = \mathbf{Cov}(s,\Phi,\Psi)\,,\end{aligned}$$

tandis que

$$\begin{aligned}&\frac{\partial^2\mathscr{H}}{\partial p\partial q}\,(f(s,\cdot),0)\cdot(\Phi,\Psi)\\ &= \tfrac12\int_{\mathbf{T}^3\times(\mathbf{R}^3)^2\times\mathbf{S}^2}\mathbf{D}\Psi(t,x,v,x,v_*,\omega)(\Phi(s,x,v)f(s,x,v_*)+\Phi(s,x,v_*)f(s,x,v))\\ &\qquad\times((v-v_*)\cdot\omega))_+ d\omega dv dv_* dx\\ &= \int_{\mathbf{T}^3\times(\mathbf{R}^3)^2\times\mathbf{S}^2}(\Phi(x,v')f(s,x,v'_*)+\Phi(x,v'_*)f(s,x,v')-\Phi(x,v)f(s,x,v_*)\\ &\qquad-\Phi(x,v_*)f(s,x,v))\Psi(x,v)((v-v_*)\cdot\omega))_+ d\omega dv dv_* dx\\ &= \int_{\mathbf{T}^3\times\mathbf{R}^3}\Psi(x,v)\mathbf{L}_t\Phi(x,v)dvdx = \int_{\mathbf{T}^3\times\mathbf{R}^3}\Phi(x,v)\mathbf{L}_t^*\Psi(x,v)dvdx\,.\end{aligned}$$

Par conséquent

$$\begin{aligned}\mathscr{C}(t,t,\phi,\psi) =& \mathscr{C}(0,0,A_{-t}\phi,A_{-t}\psi)+\int_0^t\mathscr{C}(s,s,A_{s-t}\psi,\mathbf{L}_t^*(A_{s-t}\phi))ds\\ &+\int_0^t\mathscr{C}(s,s,A_{s-t}\phi,\mathbf{L}_t^*(A_{s-t}\psi))ds+\int_0^t\mathbf{Cov}(s,A_{s-t}\phi,A_{s-t}\psi)ds\,.\end{aligned}$$

On vient donc de déduire de l'équation de Hamilton–Jacobi (30) le fait que la limite de $\mathscr{C}_\varepsilon(t,t,\cdot,\cdot)$ vérifie la première équation du système (44), tout comme la covariance $\hat{\mathscr{C}}(t,t,\cdot,\cdot)$ de la solution de (42). Que la limite de $\mathscr{C}_\varepsilon(t,s,\cdot,\cdot)$ vérifie également la seconde équation du système (44) s'obtient par un calcul analogue, quoiqu'un peu plus

long, que nous ne ferons donc pas ici, mais pour lequel nous renvoyons le lecteur à la fin de la preuve de la Proposition 5.5.2 dans (Bodineau, Gallagher, Saint-Raymond et Simonella, 2023). □

Comme annoncé plus haut, il suffit maintenant de démontrer l'unicité des solutions de (44) pour en déduire que la corrélation limite $\mathscr{C}$ du champ de fluctuations de la mesure empirique ζ_t^ε autour de la solution de l'équation de Boltzmann coïncide avec la corrélation $\hat{\mathscr{C}}$ de la solution de (42).

Après quoi, une fois démontré que ζ_t^ε converge en loi vers un champ gaussien centré, on en déduira que ce champ gaussien est précisément la solution ζ_t de (42), puisqu'un champ gaussien centré est caractérisé par sa covariance. Le caractère gaussien du champ limite utilise la petitesse des cumulants d'ordre 3 : voir Proposition 2.4.

Ces deux étapes bien distinctes de la démonstration du Théorème 4.1 (unicité et convergence vers un champ gaussien centré) ne reposent pas sur l'équation de Hamilton–Jacobi, et nous renvoyons le lecteur intéressé à en connaître les démonstrations aux sections 6.2 et 6.3, chapitre 6 de (Bodineau, Gallagher, Saint-Raymond et Simonella, 2023).

4.3. Formule de Spohn pour la covariance limite

Spohn (1981) propose encore une autre formule pour la covariance limite $\mathscr{C}(t,s,\psi,\phi)$.

Proposition 4.4. *Sous les mêmes hypothèses qu'à la Proposition 4.3, et pour tous* $0 \leqslant s \leqslant t \leqslant T^*$, *l'on a*

$$\mathscr{C}(t,s,\psi,\phi) = \int_{\mathbf{T}^3\times\mathbf{R}^3} U(t,s)^*\psi(z)\phi(z)f(s,z)dz$$
$$+\int_0^t \int_{(\mathbf{T}^3\times\mathbf{R}^3)^2} R^{1,2}(f(\tau,\cdot),f(\tau,\cdot))(z,z_*)(U(t,\tau)^*\psi)(z)(U(s,\tau)^*\phi)(z_*)dzdz_*\,,$$

où $R^{1,2}$ *est l'opérateur de recollision, défini comme suit :*

$$R^{1,2}(g,g)(z_1,z_2) := \delta_0(x_1-x_2)\int_{\mathbf{S}^2}(g(z_1')g(z_2') - g(z_1)g(z_2))((v_1-v_2)\cdot\omega)_+ d\omega\,.$$

Avant de donner la démonstration de cette proposition, expliquons la terminologie d'« opérateur de recollision » désignant $R^{1,2}$. Pour cela, revenons à l'équation

de Liouville (10), et écrivons l'équation vérifiée par $\mathbb{F}_{n:2}$. On intègre donc chaque membre de (10) par rapport aux variables $z_3, \dots, z_n$, pour trouver que

$$\begin{aligned}
&(\partial_t + \sum_{i=1}^{2} v_i \cdot \nabla_{x_i})\mathbb{F}_{n:2}(t,z_1,z_2) \\
&= (n-2)\varepsilon^2 \mathscr{B}_\varepsilon^{1,3}(\mathbb{F}_{n:3})(t,z_1,z_2) + (n-2)\varepsilon^2 \mathscr{B}_\varepsilon^{2,3}(\mathbb{F}_{n:3})(t,z_1,z_2) \\
&\quad + \mathbb{F}_{n:2}(t,z_1',z_2')\Big|_{\mathrm{dist}(x_1,x_2)=\varepsilon+0}((v_1-v_2)\cdot n_{12})_+ \delta_{\mathrm{dist}(x_1,x_2)=\varepsilon} \\
&\quad - \mathbb{F}_{n:2}(t,z_1,z_2)\Big|_{\mathrm{dist}(x_1,x_2)=\varepsilon+0}((v_1-v_2)\cdot n_{12})_- \delta_{\mathrm{dist}(x_1,x_2)=\varepsilon}\,.
\end{aligned} \tag{46}$$

Les intégrales de collision sur la seconde ligne de (46) sont données par

$$\begin{aligned}
&\mathscr{B}_\varepsilon^{1,3}(\mathbb{F}_{n:3})(t,z_1,z_2) \\
&= \int_{\mathbf{R}^3\times\mathbf{S}^2} (\mathbb{F}_{n:3}(t,z_1',z_2,x_1-\varepsilon\omega,v_3') - \mathbb{F}_{n:3}(t,z_1,z_2,x_1+\varepsilon\omega,v_3))((v_1-v_3)\cdot\omega)_+ d\omega dv_3\,, \\
&\mathscr{B}_\varepsilon^{2,3}(\mathbb{F}_{n:3})(t,z_1,z_2) \\
&= \int_{\mathbf{R}^3\times\mathbf{S}^2} (\mathbb{F}_{n:3}(t,z_1,z_2',x_2-\varepsilon\omega,v_3') - \mathbb{F}_{n:3}(t,z_1,z_2,x_2+\varepsilon\omega,v_3))((v_1-v_3)\cdot\omega)_+ d\omega dv_3\,.
\end{aligned}$$

Les deux derniers termes sur les troisièmes et quatrièmes lignes de (46) traduisent les collisions entre les molécules n$^{\mathbf{os}}$ 1 et 2. Comme ces molécules sont celles que l'on a choisies parmi les n en considérant la marginale à deux corps $F_{n:2}(t,z_1,z_2)$, ces collisions correspondent à des corrélations entre les molécules en question, ce qui semble contredire l'hypothèse de chaos moléculaire de Boltzmann, utilisée pour factoriser

$$\mathbb{F}_{n:2}(t,z_1',z_2')\Big|_{\mathrm{dist}(x_1,x_2)=\varepsilon+0}\mathbf{1}_{(v_1-v_2)\cdot n_{12}>0} \quad \text{et} \quad \mathbb{F}_{n:2}(t,z_1,z_2)\Big|_{\mathrm{dist}(x_1,x_2)=\varepsilon+0}\mathbf{1}_{(v_1-v_2)\cdot n_{12}<0}$$

dans le formalisme canonique utilisé dans (Cercignani, Illner et Pulvirenti, 1994; Gallagher, Saint-Raymond et Texier, 2013; Golse, 2014; Lanford, 1975), ou, de façon équivalente, la corrélation F_2^ε dans le formalisme grand-canonique utilisé ici, afin d'arriver à l'équation de Boltzmann.

Intégrons donc ces deux derniers termes contre une fonction test $\phi \equiv \phi(t,z_1,z_2)$ et faisons tendre ε vers 0 en appliquant l'hypothèse de chaos moléculaire en supposant que $\mathbb{F}_{n:2}$ se factorise en $g(t,\cdot)\otimes g(t,\cdot)$ pour les particules sur le point de collisionner : il vient

$$\begin{aligned}
&\frac{1}{\varepsilon^2}\int_{(\mathbf{T}^3\times\mathbf{R}^3)^2} \mathbb{F}_{n:2}(t,z_1',z_2')\Big|_{\mathrm{dist}(x_1,x_2)=\varepsilon+0}((v_1-v_2)\cdot n_{12})_+ \delta_{\mathrm{dist}(x_1,x_2)=\varepsilon}\,\phi(z_1,z_2)dz_1dz_2 \\
&\qquad \to \int_{\mathbf{T}^3\times(\mathbf{R}^3)^2\times\mathbf{S}^2} g(t,x,v_1')g(t,x,v_2')((v_1-v_2)\cdot\omega)_+\phi(x,v_1,x,v_2)dxdv_1dv_2d\omega\,,
\end{aligned}$$

et

$$\frac{1}{\varepsilon^2}\int_{(\mathbf{T}^3\times\mathbf{R}^3)^2}\mathbb{F}_{n:2}(t,z_1,z_2)\Big|_{\mathrm{dist}(x_1,x_2)=\varepsilon+0}((v_1-v_2)\cdot n_{12})_-\delta_{\mathrm{dist}(x_1,x_2)=\varepsilon}\phi(z_1,z_2)dz_1dz_2$$
$$\to\int_{\mathbf{T}^3\times(\mathbf{R}^3)^2\times\mathbf{S}^2}g(t,x,v_1)g(t,x,v_2)((v_1-v_2)\cdot\omega)_+\phi(x,v_1,x,v_2)dxdv_1dv_2d\omega\,.$$

Donc

$$\frac{1}{\varepsilon^2}\int_{(\mathbf{T}^3\times\mathbf{R}^3)^2}\mathbb{F}_{n:2}(t,z_1',z_2')\Big|_{\mathrm{dist}(x_1,x_2)=\varepsilon+0}((v_1-v_2)\cdot n_{12})_+\delta_{\mathrm{dist}(x_1,x_2)=\varepsilon}\phi(z_1,z_2)dz_1dz_2$$
$$-\frac{1}{\varepsilon^2}\int_{(\mathbf{T}^3\times\mathbf{R}^3)^2}\mathbb{F}_{n:2}(t,z_1,z_2)\Big|_{\mathrm{dist}(x_1,x_2)=\varepsilon+0}((v_1-v_2)\cdot n_{12})_-\delta_{\mathrm{dist}(x_1,x_2)=\varepsilon}\phi(z_1,z_2)dz_1dz_2$$
$$\to\int_{(\mathbf{T}^3\times\mathbf{R}^3)^2}R^{1,2}(g,g)(t,z_1,z_2)\phi(z_1,z_2)dz_1dz_2\,,$$

ce qui montre le lien entre les recollisions des molécules n$^{\mathbf{os}}$ 1 et 2 et l'opérateur $R^{1,2}$.

La formule de Spohn dans la proposition ci-dessus est donc particulièrement intéressante car elle met en évidence la contribution des recollisions, qui sont des événements rares (noter qu'on a dû diviser leur contribution par ε^2 pour les mettre à une échelle observable dans la limite de Boltzmann–Grad), à la covariance limite. Quoique rares, ces recollisions ont un effet sur l'instabilité de la dynamique et sur les fluctuations, puisqu'elles correspondent à des collisions entre des sphères de rayon négligeable — rappelons que dans la dynamique du billard, plus le rayon des molécules est petit, plus grande est l'instabilité de la dynamique.

Démonstration de la Proposition 4.4. On va se limiter au cas où $s = t$ — le cas $s < t$ se traitant de façon analogue. La formule est évidemment correcte pour $t = s = 0$, puisqu'on sait que

$$\mathscr{C}(0,0,\phi,\psi)=\mathbb{E}(\langle\zeta_0,\phi\rangle\langle\zeta_0,\psi\rangle)=\int_{\mathbf{T}^3\times\mathbf{R}^3}\phi(z)\psi(z)f^{in}(z)dz\,.$$

Considérons l'opérateur

$$\Sigma_t\psi(z_1):=-\delta_0(x_1-x_2)\int_{\mathbf{R}^3\times\mathbf{R}^3}(f(t,z_1)f(t,z_2)+f(t,z_1')f(t,z_2'))\mathbf{D}\psi(z_1,z_2,\omega)$$
$$\times((v_1-v_2)\cdot\omega)_+dz_1d\omega\,.$$

Lemme 4.5. *L'opérateur Σ_t représente la covariance du bruit* $\mathbf{Cov}(t,\cdot,\cdot)$*, au sens où*

$$\mathbf{Cov}(t,\phi,\psi)=\int_{\mathbf{T}^3\times\mathbf{R}^3}\phi(z)\Sigma_t\psi(z)dz\,.$$

Cet opérateur vérifie

$$\Sigma_t\phi(z)=-(f(t,\cdot)\mathscr{L}_t^*+\mathscr{L}_tf(t,\cdot))\phi(z)+\phi(z)\partial_tf(t,z)$$
$$+\int_{\mathbf{T}^3\times\mathbf{R}^3}R^{1,2}(f(t,\cdot),f(t,\cdot))(z,z_*)\phi(z_*)dz_*\,.$$

Admettons ce lemme, dont la démonstration résulte d'un calcul élémentaire laissé au lecteur. On part de la formule (43) avec $s = t$, soit

$$\mathscr{C}(t,t,\phi,\psi) = \mathbb{E}(\langle\zeta_0, U(t,0)^*\phi\rangle\langle\zeta_0, U(t,0)^*\psi\rangle) + \int_0^t \mathbf{Cov}(\tau, U(t,\tau)^*\phi, U(t,\tau)^*\psi)d\tau\,,$$

formule que l'on transforme comme suit :

$$\begin{aligned}\mathscr{C}(t,t,\phi,\psi) = &\int_{\mathbf{T}^3\times\mathbf{R}^3} U(t,0)^*\phi(z)U(t,0)^*\psi(z)f^{in}(z)dz\\ &+\int_0^t\left(\int_{\mathbf{T}^3\times\mathbf{R}^3}\phi(z)U(t,\tau)\Sigma_t U(t,\tau)^*\psi(z)dz\right)d\tau\,.\end{aligned}$$

Utilisons maintenant la deuxième formule du lemme, que l'on écrit sous la forme

$$\begin{aligned}U(t,\tau)\Sigma_t U(t,\tau)^*\psi(z) &= \partial_\tau\left(U(t,\tau)f(\tau,\cdot)U(t,\tau)^*\right)\psi(z)\\ +U(t,\tau)&\int_{\mathbf{T}^3\times\mathbf{R}^3} R^{1,2}(f(t,\cdot),f(t,\cdot))(z,z_*)U(t,\tau)^*\phi(z_*)dz_*\,,\end{aligned}$$

d'où l'on tire que

$$\begin{aligned}\mathscr{C}(t,t,\phi,\psi) = &\int_{\mathbf{T}^3\times\mathbf{R}^3} U(t,0)^*\phi(z)U(t,0)^*\psi(z)f^{in}(z)dz\\ &+\int_0^t\int_{\mathbf{T}^3\times\mathbf{R}^3}\phi(z)\partial_\tau\left(U(t,\tau)f(\tau,\cdot)U(t,\tau)^*\right)\psi(z)dzd\tau\\ +\int_0^t\left(\int_{(\mathbf{T}^3\times\mathbf{R}^3)^2}\right.&\left.(U(t,\tau)^*\phi)(z)R^{1,2}(f(t,\cdot),f(t,\cdot))(z,z_*)(U(t,\tau)^*\psi)(z_*)dz\right)d\tau\,.\end{aligned}$$

Par intégration en τ, la seconde intégrale au membre de droite se combine avec le premier terme au membre de droite pour donner ce qui est bien la formule annoncée dans le cas $s = t$. □

5. Esquisse de preuve pour le Théorème 2.6

On se limitera dans cette section à quelques trop brèves indications. On espère qu'elles inciteront les lecteurs à étudier en détail (Bodineau, Gallagher, Saint-Raymond et Simonella, 2023), et leur serviront de guide de lecture.

Une première remarque s'impose : comme pour la preuve du théorème de Lanford, la démonstration du Théorème 2.6 passe par une représentation « assez explicite » des cumulants de tous ordres, plutôt que par une analyse des solutions de l'équation (30) — sans même parler d'un passage à la limite pour $\varepsilon \to 0^+$ dans (27) pour en déduire l'équation (28).

5.1. Représentation des corrélations

Commençons par la solution de l'équation de Boltzmann (14)

$$(\partial_t + v_1 \cdot \nabla_{x_1}) F_1(t, z_1) = \mathscr{B}_+(F_1)(t, z_1) - \mathscr{B}_-(F_1)(t, z_1)\,,$$

où l'intégrale des collisions est décomposée en termes de perte $\mathscr{B}_-$ et de gain $\mathscr{B}_+$:

$$\mathscr{B}_-(F_1)(t, z_1) = F_1(t, x_1, v_1) \int_{\mathbf{R}^3 \times \mathbf{S}^2} F_1(t, x_1, v_2)((v_1 - v_2) \cdot \omega)_+ dv_2 d\omega\,,$$
$$\mathscr{B}_+(F_1)(t, z_1) = \int_{\mathbf{R}^3 \times \mathbf{S}^2} F_1(t, x_1, v_1') F_1(t, x_1, v_2')((v_1 - v_2) \cdot \omega)_+ dv_2 d\omega\,,$$

avec

$$v_1' = v_1 - ((v_1 - v_2) \cdot \omega)\omega\,, \qquad v_2' = v_2 + ((v_1 - v_2) \cdot \omega)\omega\,.$$

On exprime ensuite F_1 par la formule de Duhamel

$$F_1(t, z_1) = F_1(0, x_1 - tv_1, v_1) + \int_0^t \mathscr{B}_+(F_1)(t - s, x - sv_1, v_1) ds - \int_0^t \mathscr{B}_-(F_1)(t - s, x - sv_1, v_1) ds\,,$$

puis l'on itère cette formule en exprimant à nouveau F_1 par cette même formule dans les termes de perte $\mathscr{B}_-$ et de gain $\mathscr{B}_+$. On aboutit ainsi à représenter F_1 par une série de la forme

$$F_1(t, z_1) = \sum_{n \geqslant 0} \sum_{\mathbb{A}_{1,n}} \int_{([0,t_1] \times \mathbf{R}^3 \times \mathbf{S}^2)^{n-1}} (f^{in})^{\otimes n}(\Psi_{1,n}(0)) S(\Psi_{1,n}) dT_{2,n} dV_{2,n} d\Omega_{2,n}\,, \quad (47)$$

où $V_{2,n} := (v_2, \ldots, v_n)$, avec des définitions similaires de $\Omega_{2,n}$ et de $T_{2,n}$, à ceci près que l'on suppose en plus $0 < t_n < t_{n-1} < \ldots < t_2 < t_1 = t$. Chaque terme dans cette série est représenté par un arbre [(9)] $\mathbb{A}_{1,n}$ enraciné [(10)] en 1 avec $n - 1$ points de ramification $a_i \in \{2, \ldots, n\}$ correspondant à des instants de collision $t_2 > \ldots > t_n > 0$ pour $i = 2, \ldots, n$, comme sur la figure 2. La pseudo-trajectoire $\Psi_{1,n}$ est définie par la règle suivante :
(a) sur $]t_i, t_{i-1}[$, on fait évoluer un groupe de $i - 1$ particules par la dynamique du transport libre $Z_{i-1} \mapsto (X_{i-1} - (t_{i-1} - t)V_{i-1}, V_{i-1})$;
(b) au temps $t_i + 0$, on ajoute la particule n° i à la position $x_{a_i}(t_i)$ avec la vitesse v_i ;
(c) si $v_{a_i}(t_i + 0) - v_i) \cdot \omega_i > 0$, on effectue la collision en remplaçant $(v_{a_i}(t_i + 0), v_i)$ par $(v_{a_i}'(t_i + 0), v_i')$ au moyen des relations de collision paramétrées par ω_i, *i.e.*

$$(v_{a_i}'(t_i + 0), v_i') = (v_{a_i}(t_i + 0) - ((v_{a_i}(t_i + 0) - v_i) \cdot \omega_i)\omega_i, v_i + ((v_{a_i}(t_i + 0) - v_i) \cdot \omega_i)\omega_i)$$

[(9)] Pour ce qui est des arbres et des graphes, on utilisera la terminologie de l'annexe du chapitre IV de (Bourbaki, 1981).
[(10)] C'est-à-dire avec un sommet distingué.

et on itère le processus ci-dessus en posant

$$(v_{a_i}(t_i-0), v_i(t_i-0)) = (v'_{a_i}(t_i+0), v'_i)\,;$$

si $(v_{a_i}(t_i+0)-v_i)\cdot\omega_i > 0$, on poursuit de même sans changer les vitesses $v_{a_i}(t_i+0)$ et v_i (ce dernier cas correspondant à la contribution des termes de perte $\mathscr{B}_-$).

Ces règles définissent ainsi une pseudo-trajectoire $\Psi_n(t)$ issue de z_1 à l'instant t, et dont la valeur à $t=0$ est $\Psi_n(0) \in (\mathbf{T}^3\times\mathbf{R}^3)^n$.

Dans la série ci-dessus, on pose enfin

$$S(\Psi_{1,n}) := \prod_{i=1}^{n} \pm((v-v_{a_i}(t_i+0))\cdot\omega_i)\,,$$

où le préfacteur $\pm$ vaut $+$ en cas de collision avec changement de vitesse, et $-$ dans le cas contraire.

Une représentation analogue existe pour F_1^ε avant le passage à la limite $\varepsilon\to 0^+$:

$$F_1^\varepsilon(t,z_1) = \sum_{n\geqslant 0}\sum_{\mathbb{G}_{1,n}} \int_{\mathscr{E}^\varepsilon} F_n^\varepsilon(\Psi_{1,n}^\varepsilon(0)) S^\varepsilon(\Psi_{1,n}) dT_{2,n} dV_{2,n} d\Omega_{2,n}\,. \tag{48}$$

Les différences avec le calcul de F_1 sont les suivantes :
(a') sur $]t_i, t_{i-1}[$, on fait évoluer un groupe de $i-1$ particules par la dynamique du billard à $i-1$ corps (1)-(2)-(3), au lieu du transport libre ; il y a donc en général des recollisions, comme expliqué dans la section précédente, de sorte que cette représentation est paramétrée par un graphe $\mathbb{G}_{1,n}$ pouvant présenter des circuits (chaque recollision créant un circuit : voir figure 2), et non plus par un arbre ;
(b') au temps t_i, on ajoute la particule n° i à la position $x_{a_i}(t_i)+\varepsilon\omega_i$ au lieu de $x_{a_i}(t_i)$ avec la vitesse v_i ;
(c') on intègre sur les paramètres admissibles $\mathscr{E}^\varepsilon$ au lieu de $([0,t_1]\times\mathbf{R}^3\times\mathbf{S}^2)^{n-1}$ avec la seule contrainte $0<t_n<\ldots<t_2<t$, car les particules ne doivent se recouvrir à aucun instant. En particulier $\Psi_n^\varepsilon(0)\in\Gamma_n^\varepsilon$.

Évidemment, il existe une représentation analogue à (48) pour les corrélations d'ordre supérieur F_k^ε. Au lieu d'un seul arbre, elle met en jeu k graphes de sommets marqués $1,2,\ldots,k$, avec $n_1,\ldots,n_k$ arêtes, et ces graphes peuvent bien sûr interagir.

5.2. Représentation des cumulants

Examinons le cas du second cumulant $f_2^\varepsilon := \mu_\varepsilon(F_2^\varepsilon - F_1^\varepsilon\otimes F_1^\varepsilon)$. Comme expliqué plus haut, la représentation de F_2^ε met en jeu deux graphes avec chacun un sommet marqué, notés 1 et 2. Deux cas se présentent alors : (i) ces deux graphes sont disconnectés, c'est-à-dire restent à une distance supérieure à ε sur l'intervalle de temps $[0,t]$, (ii) à un (ou plusieurs) instants dans l'intervalle de temps $[0,t]$, ces deux graphes subissent une recollision *externe*, mettant en jeu une arête de chacun des deux graphes.

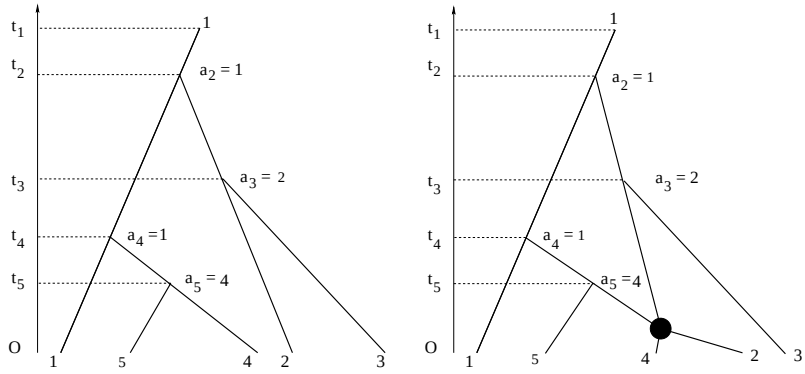


Figure 2 – À gauche un arbre de type $\mathbb{A}_{1,5}$ à 5 molécules correspondant à la dynamique de l'équation de Boltzmann. À droite, un graphe de type $\mathbb{G}_{1,5}$ à 5 molécules correspondant à la dynamique (1)-(2)-(3). L'intersection entourée en rouge désigne une « recollision » entre les molécules n^{os}2 et 4 avant que celles-ci n'entrent en collision avec la molécule n^{o}1. Lorsqu'à l'instant $t_4 - 0$ la particule n^{o}4 est ajoutée aux particules n^{os}1, 2, 3, elle n'en est pas indépendante car elle a déjà subi une collision avec la molécule n^{o}2 auparavant (dans le cas présent, avant l'instant t_5). Le triangle de sommets les points de ramification a_2, a_4 et le disque noir correspondant à la recollision entre molécules n^{os} 2 et 4 est un circuit dans le graphe $\mathbb{G}_{1,5}$.

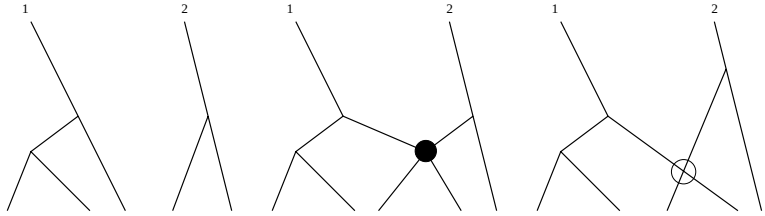


Figure 3 – À gauche les arbres (1) et (2) sont disconnectés, ce qu'on note (1) $\not\sim$ (2) ; au milieu, les arbres (1) et (2) subissent une recollision externe (disque noir), ce qu'on note (1) $\sim_r$ (2) ; enfin, à droite, les arbres (1) et (2) sont en situation de recouvrement (cercle), ce qu'on note (1) $\sim_o$ (2).

Les graphes disconnectés ne sont pas indépendants, car ils sont corrélés par la relation d'exclusion (le fait d'être à distance supérieure à ε sur $[0,t]$). De même que, d'après (6)

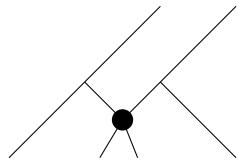
$$\mathbb{F}_2^{in}(z_1,z_2) = f^{in}\otimes f^{in}(z_1,z_2) - f^{in}(z_1)f^{in}(z_2)\mathbf{1}_{|x_1-x_2|\leqslant\varepsilon}\,,$$

(où on peut penser aux points z_1 et z_2 comme à deux arbres triviaux à un seul sommet), on peut représenter la contribution des graphes disconnectés comme différence entre le carré tensoriel de F_1^ε et celle où les deux graphes de sommets marqués 1 et 2 se trouvent à une distance inférieure à ε à un (ou plusieurs) instants dans l'intervalle de temps $[0,t]$ sans transformation de vitesse, situation qu'on désigne du nom de « recouvrement ». Autrement dit, la contribution des graphes disconnectés est égale à la différence de la contribution des graphes vus comme indépendants et de la contribution des graphes en situation de recouvrement.

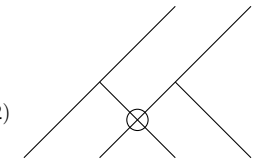
À partir de là, on arrive à une représentation du second cumulant de la forme

$$f_2^\varepsilon = \mu_\varepsilon(F_2^\varepsilon - F_1^\varepsilon\otimes F_1^\varepsilon)$$

$$= \mu_\varepsilon \sum_{(1)\sim_r(2)}$$



$$- \mu_\varepsilon \sum_{(1)\sim_o(2)}$$



+terme obtenu par propagation du second cumulant à $t=0$

(voir dans la section 4.4.2 de (Bodineau, Gallagher, Saint-Raymond et Simonella, 2023) la formule suivant (4.4.1), ou encore la figure 10 de (Bodineau, Gallagher, Saint-Raymond et Simonella, 2022b), où le terme provenant de la condition initiale n'est pas mentionné, comme expliqué dans la Remarque 3.1 de cette dernière référence).

Partons de z_1, z_2 à l'instant t tels que $|(x_1-x_2)-(t-s)(v_1-v_2)| \leqslant \varepsilon$ pour au moins une valeur de $s\in[0,t]$, et considérons le cas très particulier de deux arbres sans point de ramification (autrement dit sans collisions entre $[0,t]$). La contribution d'un tel terme à $F_2^\varepsilon - F_1^\varepsilon\otimes F_1^\varepsilon$ sera génériquement d'ordre 1 en norme L^∞. Ce qu'on peut donc espérer, dans le meilleur des cas, est que $F_2^\varepsilon - F_1^\varepsilon\otimes F_1^\varepsilon$ soit petit en variation totale, comme terme d'ordre un intégré sur un ensemble de mesure petite — dans ce cas précis très particulier, sur un cylindre de volume $\pi\varepsilon^2\times t|v_1-v_2|$. Ceci justifie heuristiquement la mise à l'échelle de $F_2^\varepsilon - F_1^\varepsilon\otimes F_1^\varepsilon$, que l'on doit multiplier par μ_ε pour compenser l'évanescence de ce terme dans la limite de Boltzmann–Grad.

Ces remarques très fragmentaires donnent un aperçu de comment calculer et estimer le cumulant d'ordre 2. On comprend qu'il existe une représentation analogue pour les cumulants d'ordre quelconque. En gros, les contributions au cumulant d'ordre k sont d'ordre μ_ε^{k-1} (provenant de la mise à l'échelle dans la définition (24) de f_k^ε), multiplié par le nombre de graphes sur lesquels on doit sommer, terme que l'on multiplie à son tour par $O((\varepsilon^2 \times t(|v_1| + \ldots + |v_k|))^{k-1})$, qui est la mesure du domaine d'intégration, multipliée par la taille de la fonction à intégrer, pour laquelle on utilise la décroissance gaussienne de la condition initiale, afin de compenser la contribution de la norme des vitesses relatives à la taille du domaine d'intégration. Quant au nombre de graphes sur lesquels sommer, on se ramène à k^{k-2} (nombre d'arbres non orientés à k sommets étiquetés, d'après la formule de Cayley — *cf.* (Aigner et Ziegler, 2010), chapitre 30), multiplié par 2^{k-1} (correspondant au choix d'un signe $\pm$ à chaque sommet excepté la racine). Ce raisonnement — joint à un calcul combinatoire un peu subtil — conduit à la formule (3.4) de (Bodineau, Gallagher, Saint-Raymond et Simonella, 2022b).

Le lecteur intéressé à connaître le détail de ces calculs et des estimations qui en découlent est invité à lire les chapitres 3 et 4 de (Bodineau, Gallagher, Saint-Raymond et Simonella, 2023), qui précisent les formules évoquées ci-dessus de représentation des cumulants de tous ordres en termes de graphes, ainsi que le chapitre 8, pour la démonstration des bornes sur les cumulants de tous ordres (contenant le résultat énoncé plus haut comme Proposition 2.4). La limite en $\varepsilon \to 0^+$ conduisant à la fonction $\mathbf{K}(t,h)$ et au Théorème 2.6 est traitée dans le chapitre 9 de (Bodineau, Gallagher, Saint-Raymond et Simonella, 2023).

6. Épilogue : par-delà le temps de Lanford

Tous les résultats dont il a été question jusqu'ici dans cet exposé portaient sur des intervalles de temps courts, de l'ordre du « temps de Lanford », c'est-à-dire de $T_0 = T_0[C_0, \beta_0]$ dans l'énoncé du Théorème 1.5. Rappelons que cette restriction avait été annoncée dans le troisième paragraphe de la section 2. Mais la dernière phrase de ce paragraphe faisait également espérer que l'on puisse également obtenir des informations sur la limite de Boltzmann–Grad après le temps de Lanford.

Rappelons que Bodineau, Gallagher et Saint-Raymond (2016) ont réussi à pousser la limite de Boltzmann–Grad par-delà le temps de Lanford dans des régimes particuliers, conduisant à l'équation de Boltzmann linéaire puis à une asymptotique de diffusion, ou au voisinage d'un équilibre maxwellien, puis dans une limite hydrodynamique conduisant aux équations de Stokes bidimensionnelles dépendant du temps (Bodineau, Gallagher et Saint-Raymond, 2017). Sur le premier de ces deux résultats, on pourra consulter également (Golse, 2014).

L'idée de pousser l'analyse des fluctuations de la mesure empirique autour de la solution de l'équation de Boltzmann s'impose alors comme particulièrement naturelle. On peut en effet espérer que ce cadre particulier offre la possibilité de justifier une équation aux fluctuations sur des plages de temps tendant vers l'infini avec $1/\varepsilon$. Voici le théorème obtenu par Bodineau, Gallagher, Saint-Raymond et Simonella (2020b, 2022c) dans cette direction (Théorème 4.1 de (Bodineau, Gallagher, Saint-Raymond et Simonella, 2022b)).

Notons $M(v) := \frac{1}{(2\pi)^{3/2}} e^{-|v|^2/2}$ la maxwellienne de densité 1, de vitesse moyenne nulle et de température 1, autrement dit la gaussienne centrée réduite. Notons également

$$\begin{aligned}
\mathbf{L}_M\phi(t,x,v) := \int_{\mathbf{R}^3\times\mathbf{S}^2} &(M(v')\phi(t,x,v'_*) + M(v'_*)\phi(t,x,v') \\
&- M(v)\phi(t,x,v_*) - M(v_*)\phi(t,x,v))((v-v_*)\cdot\omega)_+ dv_* d\omega\,, \\
\mathbf{L}^*_M\phi(t,x,v) := \int_{\mathbf{R}^3\times\mathbf{S}^2} &(\phi(t,x,v'_*) + \phi(t,x,v') \\
&- \phi(t,x,v_*) - \phi(t,x,v))((v-v_*)\cdot\omega)_+ M(v_*) dv_* d\omega\,,
\end{aligned}$$

les intégrales de collision directe et adjointe linéarisées autour de M.

Théorème 6.1. *Supposons que $f^{in} = M$. Dans la limite de Boltzmann–Grad $\mu_\varepsilon = \varepsilon^{-2}$ pour le formalisme grand-canonique, le champ de fluctuations ζ^ε_t de (21) converge en loi sur tout intervalle de temps de la forme $[0, T_\varepsilon]$ avec $T_\varepsilon = O(\ln\ln\ln\mu_\varepsilon)$ vers le processus gaussien centré solution de l'équation de Boltzmann aux fluctuations*

$$d\zeta_t = (-v\cdot\nabla_x + \mathbf{L}_M)\zeta^\varepsilon_t + db_t\,,$$

où b_t est un bruit gaussien delta-corrélé en t, x, de covariance

$$\begin{aligned}
\mathbf{Cov}(t,\phi,\psi) := \tfrac{1}{2}\int_{\mathbf{T}^3\times(\mathbf{R}^3)^2\times\mathbf{S}^2} &\mathbf{D}\phi(x,v,x,v_*,\omega)\mathbf{D}\psi(x,v,x,v_*,\omega) \\
&\times M(v)M(v_*)((v-v_*)\cdot\omega)_+ dx dv dv_* d\omega\,,
\end{aligned}$$

et où ζ_0 est le champ gaussien centré de covariance

$$\mathbf{E}(\langle\zeta_0,\phi\rangle\langle\zeta_0,\psi\rangle) = \int_{\mathbf{T}^3\times\mathbf{R}^3}\phi(z)\psi(z)M(v)dz\,.$$

On ne laissera de côté la démonstration de ce dernier résultat. Ce qui en est dit dans la section 4.3 de (Bodineau, Gallagher, Saint-Raymond et Simonella, 2022b) fait comprendre qu'elle est particulièrement complexe. Signalons la formule suivante pour la covariance dans ce cas :

$$\mathscr{C}(t,0,h,g^{in}) = \int_{\mathbf{T}^3\times\mathbf{R}^3} g(t,x,v)h(x,v)M(v)dxdv\,,$$

où $g(t,\cdot) = e^{t(-v\cdot\nabla_x + \mathbf{L}_M^*)} g^{in}$. Cette formule se déduit facilement de celle de Spohn (Proposition 4.4), au moins sur le laps de temps bref (de l'ordre du temps de Lanford) sur lequel elle avait été établie jusqu'aux travaux de Bodineau, Gallagher, Saint-Raymond et Simonella (2020b, 2022c). Il suffit en effet d'utiliser l'identité immédiate $M\mathbf{L}_M^*(g) = L_M(Mg)$, et d'observer que $R^{1,2}(M,M) = 0$.

Enfin, comme la longueur de l'intervalle de temps sur lequel l'équation de Boltzmann aux fluctuations autour de l'équilibre maxwellien est établie tend vers l'infini avec $1/\varepsilon$, on peut également étudier ces fluctuations dans un régime hydrodynamique convenablement linéarisé, pour obtenir un système de Stokes-Fourier aux fluctuations — sur ce dernier point, voir l'article très récent (Bodineau, Gallagher, Saint-Raymond et Simonella, 2022a).

Références

Aigner, M. et Ziegler, G. M. (2010). *Proofs from The Book*. Fourth edition. Springer-Verlag, Berlin, p. viii+274.

Alexander, R. (1976). « Time evolution for infinitely many hard spheres », *Comm. Math. Phys.* **49** (3), p. 217-232.

Ayi, N. (2017). « From Newton's law to the linear Boltzmann equation without cut-off », *Comm. Math. Phys.* **350** (3), p. 1219-1274.

Billingsley, P. (1999). *Convergence of probability measures*. Second edition. Wiley Series in Probability and Statistics : Probability and Statistics. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, p. x+277.

Bodineau, T., Gallagher, I. et Saint-Raymond, L. (2016). « The Brownian motion as the limit of a deterministic system of hard-spheres », *Invent. Math.* **203** (2), p. 493-553.

——— (2017). « From hard sphere dynamics to the Stokes-Fourier equations : an L^2 analysis of the Boltzmann-Grad limit », *Ann. PDE* **3** (1), Paper No. 2, 118.

Bodineau, T., Gallagher, I., Saint-Raymond, L. et Simonella, S. (2018). « One-sided convergence in the Boltzmann-Grad limit », *Ann. Fac. Sci. Toulouse Math.* (*6*) **27** (5), p. 985-1022.

——— (2020a). « Fluctuation theory in the Boltzmann-Grad limit », *J. Stat. Phys.* **180** (1-6), p. 873-895.

——— (2020b). « Long-time correlations for a hard sphere gas at equilibrium », *arXiv :2012.03813* [*math.AP*], *to appear in Comm. on Pure and Appl. Math.*

——— (2022a). « Dynamics of Dilute Gases at Equilibrium : From the Atomistic Description to Fluctuating Hydrodynamics », *arXiv 2210 :11812* [*math.AP*].

——— (2022b). « Dynamics of Dilute Gases : A Statistical Approach », *prépublication arXiv :2201.10149v2* [*math.AP*].

Bodineau, T., Gallagher, I., Saint-Raymond, L. et Simonella, S. (2022c). « Long-time derivation at equilibrium of the fluctuating Boltzmann equation », *arXiv :2201.04514* [*math.AP*].

——— (2023). « Statistical Dynamics of a Hard Sphere Gas : Fluctuating Boltzmann Equation and Large Deviations », *Ann. of Math.* **198** (3), p. 1047-1201.

Boltzmann, L. (1872). « Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen », *Sitzungberichte der Kaiserliche Akad. Wissenschafte* **66**, p. 275-370.

——— (1964). *Lectures on gas theory*. Translated by Stephen G. Brush. University of California Press, Berkeley-Los Angeles, Calif., p. ix+490.

Bouchet, F. (2020). « Is the Boltzmann equation reversible ? A large deviation perspective on the irreversibility paradox », *J. Stat. Phys.* **181** (2), p. 515-550.

Bourbaki, N. (1981). *Éléments de mathématique*. Groupes et algèbres de Lie. Chapitres 4, 5 et 6. [Lie groups and Lie algebras. Chapters 4, 5 and 6]. Masson, Paris, p. 290.

Cercignani, C. (1972). « On the Boltzmann equation for rigid spheres », *Transport Theory Statist. Phys.* **2** (3), p. 211-225.

Cercignani, C., Illner, R. et Pulvirenti, M. (1994). *The mathematical theory of dilute gases*. T. 106. Applied Mathematical Sciences. Springer-Verlag, New York, p. viii+347.

DiPerna, R. J. et Lions, P.-L. (1989). « On the Cauchy problem for Boltzmann equations : global existence and weak stability », *Ann. of Math.* (2) **130** (2), p. 321-366.

Gallagher, I., Saint-Raymond, L. et Texier, B. (2013). *From Newton to Boltzmann : hard spheres and short-range potentials*. Zurich Lectures in Advanced Mathematics. European Mathematical Society (EMS), Zürich, p. xii+137.

Gallavotti, G. (1969). « Divergences and the Approach to Equilibrium in the Lorentz and the Wind-Tree Models », *Phys. Rev.* **185**, p. 308-322.

Gérard, P. (1988). « Solutions globales du problème de Cauchy pour l'équation de Boltzmann (d'après R. J. DiPerna et P.-L. Lions) », *Astérisque* (161-162). Séminaire Bourbaki, Vol. 1987/88, Exp. No. 699, 5, 257-281 (1989).

Golse, F. (2014). « De Newton à Boltzmann et Einstein : validation des modèles cinétiques et de diffusion (d'après T. Bodineau, I. Gallagher, L. Saint-Raymond, B. Texier) », *Astérisque* (367-368). Séminaire Bourbaki, Vol. 2013/14, Exp. No. 1083, ix, 285-326 (2015).

Grad, H. (1949). « On the kinetic theory of rarefied gases », *Comm. Pure Appl. Math.* **2**, p. 331-407.

——— (1958). *Principles of the kinetic theory of gases*. Handbuch der Physik, Bd. 12, Thermodynamik der Gase. Herausgegeben von S. Flügge. Springer-Verlag, Berlin-Göttingen-Heidelberg, p. 205-294.

Hilbert, D. (1902). « Mathematical problems », *Bull. Amer. Math. Soc.* (*N.S.*) **8**. (Reprinted in Bull. Amer. Math. Soc. **37** (2000), 407-436), p. 437-479.

Hopf, E. (1950). « The partial differential equation $u_t + uu_x = \mu u_{xx}$ », *Comm. Pure Appl. Math.* **3**, p. 201-230.

Illner, R. et Pulvirenti, M. (1989). « Global validity of the Boltzmann equation for two- and three-dimensional rare gas in vacuum. Erratum and improved result : "Global validity of the Boltzmann equation for a two-dimensional rare gas in vacuum" [Comm. Math. Phys. **105** (1986), no. 2, 189–203] and "Global validity of the Boltzmann equation for a three-dimensional rare gas in vacuum" [ibid. **113** (1987), no. 1, 79–85] by Pulvirenti », *Comm. Math. Phys.* **121** (1), p. 143-146.

King, F. (1975). « BBGKY hierarchy for positive potentials ». Thèse de doct. Dept of Math., Univ. of California, Berkeley.

Lanford III, O. E. (1975). « Time evolution of large classical systems », in : *Dynamical systems, theory and applications* (*Rencontres, Battelle Res. Inst., Seattle, Wash., 1974*). Lecture Notes in Phys., Vol. 38. Springer, Berlin, p. 1-111.

——— (1976). « On a derivation of the Boltzmann equation », in : *International Conference on Dynamical Systems in Mathematical Physics* (*Rennes, 1975*). Astérisque, No. 40. Soc. Math. France, Paris, p. 117-137.

Malliavin, P. (1995). *Integration and probability*. T. 157. Graduate Texts in Mathematics. With the collaboration of Hélène Airault, Leslie Kay and Gérard Letac, Edited and translated from the French by Kay, With a foreword by Mark Pinsky. Springer-Verlag, New York, p. xxii+322.

Maxwell, J. C. (1860). « Illustrations of the Dynamical Theory of Gases », *Philosophical Magazine* (*4*) **19**, p. 19-32.

——— (1867). « On the Dynamical Theory of Gases », *Philosophical Trans. Roy. Soc. London* **147**, p. 49-88.

Nirenberg, L. (1972). « An abstract form of the nonlinear Cauchy-Kowalewski theorem », *J. Differential Geometry* **6**, p. 561-576.

Ovsjannikov, L. V. (1971). « A nonlinear Cauchy problem in a scale of Banach spaces », *Dokl. Akad. Nauk SSSR* **200**, p. 789-792.

Pulvirenti, M. et Simonella, S. (2017). « The Boltzmann-Grad limit of a hard sphere system : analysis of the correlation error », *Invent. Math.* **207** (3), p. 1135-1237.

Rezakhanlou, F. et Villani, C. (2008). *Entropy methods for the Boltzmann equation*. T. 1916. Lecture Notes in Mathematics. Lectures from a Special Semester on Hydrodynamic Limits held at the Université de Paris VI, Paris, 2001, Edited by François Golse and Stefano Olla. Springer, Berlin, p. xii+107.

Schwartz, L. (1966). *Théorie des distributions*. Publications de l'Institut de Mathématique de l'Université de Strasbourg, IX-X. Nouvelle édition, entièrement corrigée, refondue et augmentée. Hermann, Paris, p. xiii+420.

Sone, Y. (2007). *Molecular gas dynamics*. Modeling and Simulation in Science, Engineering and Technology. Theory, techniques, and applications. Birkhäuser Boston Inc., Boston, MA, p. xiv+658.

Spohn, H. (1981). « Fluctuations around the Boltzmann equation », *J. Statist. Phys.* **26** (2), p. 285-305.

Van Beijeren, H. et al. (1980). « Equilibrium time correlation functions in the low-density limit », *J. Statist. Phys.* **22** (2), p. 237-257.

Varadhan, S. R. S. (1984). *Large deviations and applications*. T. 46. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial et Applied Mathematics (SIAM), Philadelphia, PA, p. v+75.

François Golse

École polytechnique, CMLS
Route de Saclay, 91128 Palaiseau Cedex

E-mail : francois.golse@polytechnique.edu

Séminaire BOURBAKI Novembre 2022
74e année, 2022–2023, no 1199, p. 87 à 120
doi: 10.24033/ast.1208

POINTWISE ERGODIC THEORY: EXAMPLES AND ENTROPY
[*after* **Jean Bourgain**]

by **Ben Krause**

Overview

Pointwise ergodic theory, the motivation for discrete harmonic analysis, has at its roots the classical theorem of Birkhoff (1931), which can be described as follows:

> For every ergodic —that is, "sufficiently randomizing"— measure-preserving transformation, τ, of a probability space, (X, μ), and any integrable function $f \in L^1(X, \mu)$, μ-almost surely, one can recover the mean of f by considering the Cesáro sums
>
> $$\frac{1}{N} \sum_{n \leqslant N} f(\tau^n x) \to \int_X f \, d\mu \ \mu - \text{a.e.}$$

Informally, this theorem says that one can recover the *spatial mean* of f,

$$\int_X f \, d\mu,$$

by considering the *temporal means*

$$\{\frac{1}{N} \sum_{n \leqslant N} f(\tau^n x)\},$$

formed by "sampling" the function f at the "times" $\{\tau^n x\}$ and taking the appropriate average.[(1)]

A classical question in pointwise ergodic theory concerned the almost-everywhere existence of limiting behavior of averages

$$\frac{1}{N} \sum_{n=1}^{N} \tau^{a_n} f \tag{1}$$

[(1)] Even in the case when τ is not ergodic, the temporal means $\{\frac{1}{N} \sum_{n \leqslant N} \tau^n f(x)\}$ still converge μ-almost everywhere.

where $\{a_n\}$ is "sparse"; as is custom, here and throughout we use $\tau^k f$ to denote the function

$$x \mapsto f(\tau^k x).$$

When the lower density of the sequence $\{a_n\}$ is bounded away from zero

$$\liminf \frac{|\{n : a_n \leqslant N\}|}{N} > 0,$$

detecting convergence becomes more straightforward, and the classical question concerned the existence of sequences $\{a_n\}$ with zero density,

$$\lim \frac{|\{n : a_n \leqslant N\}|}{N} = 0,$$

for which the averages (1) converged almost everywhere. In Bellow and Losert (1984), such a sequence was constructed; it consisted of taking long blocks of natural numbers, followed by much longer gaps, followed by slightly longer blocks, followed by even longer gaps, etc. In particular, this sequence had an upper Banach density of 1

$$d^*(\{a_n\}) := \limsup_{|I|\to\infty \text{ an interval}} \frac{|\{a_n\} \cap I|}{|I|} = 1.$$

The question remained, however, whether or not there existed upper Banach density-zero sequences, $\{a_n\}$ with $d^*(\{a_n\}) = 0$, for which the almost-everywhere convergence of the averages (1) could be proved. In particular, the classical question, explicitly posed first by Furstenberg, see also Bellow (1982), was whether or not the averages along the squares

$$\frac{1}{N} \sum_{n=1}^{N} \tau^{n^2} f$$

converged pointwise almost everywhere, initially for $f \in L^2(X)$. In breakthrough work, Bourgain (1988b,c, 1989b) answered this question affirmatively, and proved the almost everywhere convergence of (1) for any polynomial sequence,

$$\{a_n = P(n)\}, \quad P \in \mathbb{Z}[\cdot],$$

and any $f \in L^p(X)$, $p > 1$, for any σ-finite measure space X; this result was later proven to be sharp (Buczolich and Mauldin, 2007; Lavictoire, 2011).

Theorem 0.1. *Suppose that (X, μ) is a σ-finite measure space, $\tau : X \to X$ is a measure-preserving transformation, and $P \in \mathbb{Z}[\cdot]$ is a polynomial with integer coefficients. Then for each $1 < p < \infty$*

$$\frac{1}{N} \sum_{n=1}^{N} \tau^{P(n)} f$$

converges μ-a.e.

Although the issue of pointwise convergence is qualitative, Bourgain's insight was to quantify the rate at which convergence occurred – and then to use an abstract transference argument first due to Calderón (1968) to deduce these quantitative estimates from a single "universal" measure preserving system. By considering sequences of the form

$$\mathbb{Z} \ni n \mapsto \tau^n f(x), \quad x \in X \text{ fixed}$$

and using the measure-preserving nature of τ, Bourgain was able to reduce matters to proving estimates in the case of the integers with counting measure and the shift $(\mathbb{Z}, |\cdot|, \tau : x \mapsto x - 1)$.

In particular, Bourgain was after quantitative estimates on the oscillation of the averaging operators

$$\frac{1}{N} \sum_{n=1}^{N} f(x - P(n)), \tag{2}$$

applied first to $\ell^2(\mathbb{Z})$-functions. A natural perspective on (2) is as a convolution of f and

$$K_N(x) := \frac{1}{N} \sum_{n=1}^{N} \delta_{P(n)}(x)$$

where δ_m denotes the point-mass at $m \in \mathbb{Z}$; as this problem is $\ell^2(\mathbb{Z})$-based, the Fourier transform method is naturally employed, and the key to the analysis is an understanding of the exponential sums

$$\frac{1}{N} \sum_{n \leqslant N} e^{-2\pi i \beta \cdot P(n)},$$

which is accomplished via the *circle method* from analytic number theory; the interplay between the "soft" analytic issue of pointwise convergence and "hard" analytic estimates on the integers/Euclidean space via analytic-number-theoretic means is characteristic of the fields of pointwise ergodic theory and discrete harmonic analysis.

I first came to understand Bourgain's work by reading Thouvenot (1990), which I think explains Theorem 0.1 beautifully; the goal of these notes is to complement Thouvenot (1990) by trying to explain the motivation behind Bourgain's argument.

Accordingly, for the sake of clarity, we will shift our focus slightly from proving Theorem 0.1, and will instead focus on the related maximal estimate, in the representative case of $L^2(X)$.

Theorem 0.2. *Suppose that (X, μ) is a σ-finite measure space, $\tau : X \to X$ is a measure-preserving transformation, and $P \in \mathbb{Z}[\cdot]$ is a polynomial with integer coefficients. Then there exists an absolute constant $\mathbf{C}$, independent of (X, μ, τ), so that*

$$\| \sup_N \Big| \frac{1}{N} \sum_{n=1}^{N} \tau^{P(n)} f \Big| \|_{L^2(X)} \leqslant \mathbf{C} \cdot \|f\|_{L^2(X)}.$$

By Calderón's transference principle, Theorem 0.2 follows from the analoguous estimate of the integers: if we define

$$\mathscr{M} f(x) := \sup_N \Big| \frac{1}{N} \sum_{n=1}^{N} f(x - P(n)) \Big|, \tag{3}$$

then our focus turns to establishing the following estimate

Theorem 0.3. *For any $P \in \mathbb{Z}[\cdot]$, the following norm inequality holds: there exists an absolute constant* **C** *so that*

$$\|\mathscr{M} f\|_{\ell^2(\mathbb{Z})} \leqslant \mathbf{C} \cdot \|f\|_{\ell^2(\mathbb{Z})}.$$

Below, following the lead of Thouvenot (1990), we will restrict to the case where

$$P(n) = n^d,$$

as this eliminates some number-theoretic technicality while still capturing the essence of the problem.

Notation. — Here and throughout we abbreviate the complex exponential $e(t) := e^{2\pi i t}$, so that we may express the Fourier transform in Euclidean space, and on the integers, respectively as

$$\hat{f}(\xi) = \int_{\mathbb{R}} f(x) \cdot e(-\xi x)\, dx, \quad g^{\vee}(x) = \int_{\mathbb{R}} g(\xi) \cdot e(\xi x)\, d\xi$$
$$\hat{f}(\beta) = \sum_n f(n) \cdot e(-\beta n), \quad g^{\vee}(n) = \int_{\mathbb{T}} g(\beta) \cdot e(\beta n)\, d\beta.$$

We will let

$$\phi_k(t) := 2^{-k} \cdot \phi(2^{-k} \cdot t)$$

denote the usual L^1-normalized dyadic dilations, and for frequencies θ, we let

$$\mathrm{Mod}_\theta g(x) := e(\theta x) \cdot g(x) \tag{4}$$

so that

$$\widehat{\mathrm{Mod}_\theta g}(\beta) = \hat{g}(\beta - \theta),$$

and recall the Hardy–Littlewood Maximal operator

$$M_{\mathrm{HL}} f(x) := \sup_{r>0} \frac{1}{2r} \int_{-r}^{r} |f(x-t)|\, dt \quad \text{or} \quad := \sup_{N \geqslant 0} \frac{1}{2N+1} \sum_{n=-N}^{N} |f(x-n)|;$$

although we use the same notation to refer to both continuous and discrete maximal operator, it will be clear from context which formulation we use.

We will let $[N] := \{1, \dots, N\}$, and abbreviate $\sum_{n \leqslant N} := \sum_{n=1}^{N}$. We will use the symbol $\mathbf{c}$ to denote suitably small constants, which remain bounded away from zero, and $\mathbf{C}$ to denote suitably large constants, which remain bounded above. If we need these constants to depend on parameters, we use subscripts, thus $\mathbf{c}_d$ is a constant that is small depending on d. We use $X = O(Y)$ to denote the statement that $|X| \leqslant \mathbf{C} \cdot Y$, and analogously define $X = O_d(Y)$.

Finally, we will use the heuristic notation

$$f \quad \text{“}=\text{”} \quad g$$

to denote moral equivalence: up to tolerable errors, f and g exhibit the same type of behavior.

1. Discrete Complications

Before beginning our discussion of Theorem 0.3, let us explain why we might expect this to be a challenging problem.

For problems with a “linear” flavor, the discrete theory essentially mirrors the continuous theory

$$\sup_r \frac{1}{r} \int_0^r |f(x-t)|\, dt \quad \text{“}=\text{”} \quad \sup_N \frac{1}{N} \sum_{n=1}^{N} |f(x-n)|$$

as can be seen by experimenting with functions of the form $F(\lfloor x \rfloor)$ and using dilation invariance of the real-variable maximal function to reduce attention to real variable functions that are constant on unit scales.

The problems become dramatically more complicated once linearity is destroyed. In this case, we consider the simple example of the Hardy–Littlewood maximal function along the curve $t \mapsto t^d$. The continuous maximal function

$$Mf := M_d f := \sup_r \big| \frac{1}{r} \int_0^r f(x-t^d)\, dt \big| = \sup_r \big| \frac{1}{r} \int_0^{r^d} f(x-t)\, \frac{1}{dt^{1-1/d}}\, dt \big|, \tag{5}$$

is just a weighted version of M_{HL} via the pointwise majorization

$$\begin{aligned} \frac{1}{r} \int_0^{r^d} |f(x-t)| \frac{1}{dt^{1-1/d}}\, dt &\leqslant \sum_{j=1}^{\infty} 2^{-j/d} \cdot \big(\frac{2^{j/d}}{r} \int_{2^{-j} \cdot r^d}^{2^{1-j} \cdot r^d} |f(x-t)| \frac{1}{dt^{1-1/d}}\, dt \big) \\ &\leqslant \mathbf{C}/d \cdot \sum_{j=1}^{\infty} 2^{-j/d} \cdot \big(\frac{2^j}{r^d} \int_{2^{-j} \cdot r^d}^{2^{1-j} \cdot r^d} |f(x-t)|\, dt \big) \leqslant \mathbf{C}/d \cdot \sum_{j=1}^{\infty} 2^{-j/d} \cdot M_{\mathrm{HL}} f(x) \\ &\leqslant \mathbf{C} \cdot M_{\mathrm{HL}} f(x). \end{aligned} \tag{6}$$

On the other hand, no such trick is available in the study of

$$\mathscr{M}f(x) := \mathscr{M}_d f(x) := \sup_N \Big| \frac{1}{N} \sum_{n \leqslant N} f(x - n^d) \Big|,$$

due to the presence of a smallest scale – there is no real analogue for an infinitesimal change of variables in the discrete setting.

Passing to the Fourier side actually highlights this difference. We can express both M and $\mathscr{M}$ as a maximal operator taken over a lacunary sequence of Fourier multipliers, after exploiting non-negativity. Let us begin with M:

$$Mf(x) := \sup_k \Big| \big(V_k(\xi)\hat{f}(\xi)\big)^\vee(x) \Big|,$$

where

$$V_k(\xi) := \int_0^1 e(-\xi 2^{dk} t^d)\, dt, \tag{7}$$

so that

$$V_k(\xi) = \int_0^1 e(-\xi 2^{dk} t) \frac{1}{dt^{1-1/d}}\, dt =: \widehat{\mu}(2^{dk}\xi) = \begin{cases} 1 + O\big(2^{dk}|\xi|\big) \\ O\big((2^{dk}|\xi|)^{-1/d}\big), \end{cases} \tag{8}$$

as can be seen by Taylor expanding the exponential around the origin and using the principle of stationary phase (cleverly integrating by parts) for the second estimate. Above, we set

$$\mu(t) := \mu_d(t) := \frac{1}{dt^{1-1/d}} \cdot \mathbf{1}_{(0,1]}. \tag{9}$$

What this analysis says is that the multipliers V_k try very hard to look like $\widehat{\varphi_{dk}}$ for, say, a Schwartz function $\varphi \geqslant 0$ with $\widehat{\varphi}(0) = 1$, as in this case, one has similar estimates:

$$\widehat{\varphi_{dk}}(\xi) = \begin{cases} 1 + O\big(2^{dk}|\xi|\big) \\ O\big((2^{dk}|\xi|)^{-100}\big) \end{cases} \tag{10}$$

(say); compare to (7). Now, by replacing the weaker ℓ^∞_k-norm of $\{(\mu_{dk} - \varphi_{dk}) * f)\}$ with the stronger ℓ^2_k-norm, we arrive at

$$\begin{aligned} Mf &\leqslant \sup_k |\varphi_{dk} * f| + \sup_k |(\mu_{dk} - \varphi_{dk}) * f| \\ &\leqslant \mathbf{C} \cdot M_{\mathrm{HL}} f + \big(\sum_k |(\mu_{dk} - \varphi_{dk}) * f|^2\big)^{1/2} \\ &=: \mathbf{C} \cdot M_{\mathrm{HL}} f + Sf, \end{aligned} \tag{11}$$

where Sf is a so-called *square function*, which is highly-tailored to study L^2-based problems. Indeed, we use Plancherel to bound

$$\begin{aligned}\|Sf\|_{L^2(\mathbb{R})}^2 &= \|\big(\sum_k |(\mu_{dk}-\varphi_{dk})*f|^2\big)^{1/2}\|_{L^2(\mathbb{R})}^2 = \sum_k \|(\mu_{dk}-\varphi_{dk})*f\|_{L^2(\mathbb{R})}^2 \\ &= \sum_k \|(V_k - \widehat{\varphi_{dk}})\cdot \hat{f}\|_{L^2(\mathbb{R})}^2 = \int \sum_k |V_k(\xi)-\widehat{\varphi_{dk}}(\xi)|^2\cdot|\hat{f}(\xi)|^2\,d\xi \\ &\leqslant \sup_\xi \sum_k |V_k(\xi)-\widehat{\varphi_{dk}}(\xi)|^2\cdot\|\hat{f}\|_{L^2(\mathbb{R})}^2 \qquad (12)\\ &\leqslant \mathbf{C}\cdot\sup_\xi \sum_k \min\{2^{kd}|\xi|,(2^{kd}|\xi|)^{-1/d}\}^2\cdot\|f\|_{L^2(\mathbb{R})}^2 \\ &\leqslant \mathbf{C}_d\cdot\|f\|_{L^2(\mathbb{R})}^2,\end{aligned}$$

using the fact that $\widehat{\varphi_{dk}}(\xi)$ satisfies the same estimates as V_k, namely (7), so that for $|\xi|\leqslant \mathbf{C}\cdot 2^{-dk}$

$$V_k(\xi)-\widehat{\varphi_{dk}}(\xi) = \big(1+O(2^{dk}|\xi|)\big)-\big(1+O(2^{dk}|\xi|)\big) = O(2^{dk}|\xi|)$$

and when $|\xi|>\mathbf{C}\cdot 2^{-dk}$

$$V_k(\xi),\ \widehat{\varphi_{dk}}(\xi) = O((2^{dk}|\xi|)^{-1/d}).$$

If we try the same trick with the discrete operator $\mathscr{M}$,

$$\mathscr{M}f(x) = \sup_k |K_k * f(x)|$$

where

$$K_k(x) := \frac{1}{2^k}\sum_{n\leqslant 2^k}\delta_{n^d}(x), \qquad (13)$$

we can similarly express $\mathscr{M}$ as a maximal multiplier operator

$$\mathscr{M}f(x) = \sup_{k\geqslant 0}|\big(\widehat{K_k}(\beta)\hat{f}(\beta)\big)^\vee(x)|,$$

where the multipliers $\widehat{K_k}$ are of a different form than the $\{V_k\}$:

$$\big\{\widehat{K_k}(\beta) := \frac{1}{2^k}\sum_{m\leqslant 2^k} e(-\beta m^d)\big\}_{k\geqslant 0}.$$

Each multiplier is a *Weyl sum*, and requires the so-called circle method of Hardy and Littlewood to analyze. As we will see below, each multiplier

$$\widehat{K_k}(\beta)$$

is large and interesting whenever β is "k-close" to a rational number with a "k-small" denominator, *i.e.* β lives in a so-called "k-major arc", and is "k-negligible" otherwise, when β lives in the complementary "k-minor arc." In particular, we see subtle *arithmetic* issues that arise as we seek to analyze the relevant multipliers; contrast this to the Euclidean situation, where we were able to understand the multipliers purely according to the *magnitude* of the frequency variable. In other words, whereas the analysis in the Euclidean setting is entirely dictated by the distance from the frequency variable to the distinguished zero-frequency – multi-frequency issues arise as we seek to understand the multipliers $\widehat{K_k}(\beta)$. Essentially, the main work in bounding

$$\|\mathscr{M} f\|_{\ell^2(\mathbb{Z})} \leqslant \mathbf{C} \cdot \|f\|_{\ell^2(\mathbb{Z})},$$

boils down to overcoming these multi-frequency complications.

2. Examples

In what follows, we can and will assume that k is sufficiently large depending on d.

To come to grips with

$$\mathscr{M} f := \sup_k |K_k * f|,$$

we first build some intuition by studying some examples:

Whereas the dilation invariance of the real line allows one to study (5) or M_{HL} using examples that live at unit scales, there is no such dilation invariance on $\mathbb{Z}$. Rather, a rough analogue of "zooming in" is passing to an arithmetic progression. Of course, this analogy is not precise, as arithmetic progressions are characterized by both gap size and diameter. Accordingly, we begin by analyzing the behavior of (13) when applied to functions

$$\varphi_{Q,N} := \mathbf{1}_{Q\mathbb{Z}} \cdot \varphi(\cdot/N) \tag{14}$$

where φ is a smooth bump function, and we think of $Q \leqslant N^{1/2}$; note the approximation

$$\|\varphi_{Q,N}\|_{\ell^2(\mathbb{Z})} \approx (N/Q)^{1/2}. \tag{15}$$

A common simplifying assumption when passing to arithmetic progressions is that the gap size be prime, as this eliminates various arithmetic technicalities, so we will do so below.

With these reductions in mind, we begin to compute.

2.1. Example

For technical reasons, we will replace the full convolution operator K_k, with its smooth "top half," in that for a smooth $\mathbf{1}_{[1,2]} \leqslant \phi \leqslant \mathbf{1}_{[1/2,4]}$, we consider

$$K_k' := \sum_n \phi_k(n) \cdot \delta_{n^d}. \tag{16}$$

Using convexity, arguing as in (6), we can bound

$$\sup_k |K_k * f| \leqslant \mathbf{C} \cdot \sup_k |K_k' * f|,$$

so there is no harm in this replacement.

So, we will be interested in understanding

$$K_k' * \varphi_{Q,N}. \tag{17}$$

There are some scaling considerations that we quickly note: Since

$$|n^d - (n-1)^d| \geqslant 2^{k(d-1)}$$

for $2^{k-1} < n \leqslant 2^{k+2}$, (17) becomes trivial if $N \leqslant 2^{k(d-1)}$, as in this case each element of the sum set

$$\{n^d : 2^{k-1} < n \leqslant 2^{k+2}\} + \{Qj : j \leqslant N/Q\}$$

has $O(1)$ representations of the form $n^d + Qj$. On the other hand since K_k' is supported on $[2^{dk+2}]$, we can assume that $N \leqslant 2^{dk+2}$, as convolution with K_k' acts independently on intervals separated by $> 2^{dk+2}$. In particular, by translation invariance we can and will restrict to $|x| \leqslant \mathbf{C} \cdot 2^{dk}$, and assume that

$$N \approx 2^{k(d-1+\delta)} \tag{18}$$

for some $0 < \delta \leqslant 1$.

If we use Fourier inversion, we may express

$$(17) = \int \widehat{K_k'}(\beta) \cdot \widehat{\varphi_{Q,N}}(\beta) \cdot e(\beta x)\, d\beta. \tag{19}$$

To determine the Fourier transform of $\varphi_{Q,N}$, we express the indicator function of $Q\mathbb{Z}$ as an exponential sum,

$$\mathbf{1}_{Q\mathbb{Z}}(n) = \frac{1}{Q} \sum_{A=1}^{Q} e(A/Q \cdot n),$$

and compute

$$\sum_n \frac{1}{Q} \sum_{A=1}^{Q} e(A/Q \cdot n) \cdot \varphi(n/N) \cdot e(-n\beta) = \frac{1}{Q} \sum_{A=1}^{Q} N\widehat{\varphi}(N(\beta - A/Q)) \tag{20}$$

by applying Poisson summation to the Schwartz function

$$t \mapsto \frac{1}{Q} \sum_{A=1}^{Q} e(A/Q \cdot t) \cdot \varphi(t/N) \cdot e(-t\beta)$$

In particular, up to Schwartz-tail considerations, we are only interested in

$$\beta \in \mathbb{Z}/Q\mathbb{Z} + O(N^{\mathbf{c}_{d,\delta}-1}),$$

as in the opposite case

$$\big|\sum_n \frac{1}{Q} \sum_{A=1}^{Q} e(A/Q \cdot n) \cdot \varphi(n/N) \cdot e(-n\beta)\big| \leqslant \mathbf{C}_{d,\delta} \cdot N^{-100}$$

using the Schwartz decay of $\hat{\varphi}$, see (20). So, for such β, decomposing

$$\beta = A/Q + \eta, \quad |\eta| \leqslant \mathbf{C} \cdot N^{\mathbf{c}_{d,\delta}-1},$$

and $n = pQ + r$, we find that

$$\begin{aligned} \beta n^d &= (A/Q + \eta) \cdot (pQ + r)^d \\ &\equiv A/Q \cdot r^d + \eta \cdot (pQ)^d + O(|\eta| \cdot 2^{(d-1)k} \cdot Q) \mod 1, \end{aligned}$$

so that for such β

$$\begin{aligned} \widehat{K_k'}(\beta) &= \sum_n \phi_k(n) \cdot e(-\beta n^d) \\ &= \sum_{pQ+r} \phi_k(pQ+r) \cdot e(-A/Q \cdot r^d) \cdot e(-\eta \cdot (pQ)^d) + O\big(\frac{2^{k(d-1)} \cdot Q}{N^{1-\mathbf{c}_{d,\delta}}}\big) \\ &= \frac{1}{Q} \sum_{r=1}^{Q} e(-A/Q \cdot r^d) \cdot \sum_{pQ} Q \cdot \phi_k(pQ) \cdot e(-\eta \cdot (pQ)^d) + O\big(\frac{2^{k(d-1)} \cdot Q}{N^{1-\mathbf{c}_{d,\delta}}}\big), \quad (21) \end{aligned}$$

using the smoothness of ϕ. To drop this error terms, we stipulate that $Q \leqslant 2^{k\delta/2}$, see (18), so that for $|\beta - A/Q| \leqslant \mathbf{C} \cdot N^{\mathbf{c}_{d,\delta}-1}$ we may express

$$K_k'(\beta) = S(A/Q) \cdot \sum_{pQ} Q \cdot \phi_k(pQ) \cdot e(-(\beta - A/Q) \cdot (pQ)^d) + \widehat{\mathscr{E}_k}(\beta),$$

where $S(A/Q)$ are *complete Weyl sums*, and $\mathscr{E}_k$ is an error term with small Fourier coefficients. Explicitly:

$$S(A/Q) := \frac{1}{Q} \sum_{n \leqslant Q} e(-A/Q \cdot n^d) = \frac{1}{Q} \sum_{m \leqslant Q} e(-A/Q \cdot m) \cdot |\{n \leqslant Q : n^d \equiv m \mod Q\}|$$

precisely captures the equidistribution properties of $n^d \mod Q$, quantified via the upper bound,

$$|S(A/Q)| \leqslant \mathbf{C}_\varepsilon \cdot Q^{\varepsilon - \frac{1}{d}}, \quad (A,Q) = 1, \quad \varepsilon > 0; \tag{22}$$

see Hua (1982). And, $\mathscr{E}_k$ is a negligible error term, in that

$$\|\widehat{\mathscr{E}_k}\|_{L^\infty(\mathbb{T})} \leqslant \mathbf{C} \cdot 2^{-k\delta/4}$$

(provided $\mathbf{c}_{d,\delta}$ has been chosen appropriately), so that

$$\|\mathscr{E}_k * \varphi_{Q,N}\|_{\ell^2} = \|\widehat{\mathscr{E}_k} \cdot \widehat{\varphi_{Q,N}}\|_{L^2(\mathbb{T})} \leqslant \mathbf{C} \cdot 2^{-k\delta/4} \cdot \|\widehat{\varphi_{Q,N}}\|_{L^2(\mathbb{T})} = \mathbf{C} \cdot 2^{-k\delta/4} \cdot \|\varphi_{Q,N}\|_{\ell^2(\mathbb{Z})};$$

in what follows, we will discard $\mathscr{E}_k$ from consideration.

By a Riemann summation argument, comparing

$$\begin{aligned} Q \cdot \phi_k(Qp) \cdot e(-(pQ)^d(\beta - A/Q)) = \int_p^{p+1} Q \cdot \phi(Qt) \cdot e(-(\beta - A/Q) \cdot (tQ)^d)\, dt \\ + O\big(2^{-\mathbf{c}_{d,\delta}k} \cdot 2^{-k} \cdot Q \cdot (1 + 2^{-k} \cdot |Qp|)^{-100}\big) \end{aligned}$$

we approximate, up to pointwise errors of the order $2^{-\mathbf{c}_{d,\delta}k}$

$$\begin{aligned} \widehat{K_k'}(\beta) &= S(A/Q) \cdot \int \phi(t) \cdot e(-2^{dk}(\beta - A/Q) \cdot t^d)\, dt + O(2^{-\mathbf{c}_{d,\delta}k}) \\ &= S(A/Q) \cdot \int \phi'(s) \cdot e(-2^{dk}(\beta - A/Q) \cdot s)\, ds + O(2^{-\mathbf{c}_{d,\delta}k}), \qquad \phi'(s) := \frac{\phi(s^{1/d})}{ds^{1-1/d}} \\ &= S(A/Q) \cdot \widehat{\phi'}(2^{dk}(\beta - A/Q)) + O(2^{-\mathbf{c}_{d,\delta}k}) \end{aligned}$$

where ϕ' is Schwartz as well, see (16). Consequently

$$\begin{aligned} \text{(19)} \quad \text{“}=\text{”} \quad & \frac{1}{Q} \sum_{A \leqslant Q} S(A/Q) \int N\widehat{\varphi}(N(\beta - A/Q)) \cdot \widehat{\phi'}(2^{dk}(\beta - A/Q)) \cdot e(\beta x)\, d\beta, \\ &= \frac{1}{Q} \sum_{A \leqslant Q} e(A/Qx) \cdot S(A/Q) \cdot \Phi(x), \end{aligned}$$

where we consolidate

$$\Phi(x) := \int \varphi((x - 2^{dk}s)/N) \cdot \phi'(s)\, ds$$

so that

$$\hat{\Phi}(\beta) = N\hat{\varphi}(N\beta) \cdot \widehat{\phi'}(2^{dk}\beta),$$

and thus $\|\Phi\|_{\ell^2(\mathbb{Z})} \approx \frac{N}{2^{dk/2}}$. Summing, we find that

$$\begin{aligned}\|K_k' * \varphi_{Q,N}\|_{\ell^2(\mathbb{Z})}^2 " &= " \sum_x \big|\frac{1}{Q}\sum_{A\leqslant Q} e(A/Qx)\cdot S(A/Q)\big|^2 \cdot |\Phi(x)|^2 \\ &= \frac{1}{Q^2}\sum_{A,B\leqslant Q} S(A/Q)\cdot\overline{S(B/Q)}\cdot\sum_x e((A/Q-B/Q)x)\cdot|\Phi(x)|^2 \\ &= \frac{1}{Q^2}\sum_{A,B\leqslant Q} S(A/Q)\cdot\overline{S(B/Q)}\cdot\widehat{|\Phi|^2}(A/Q-B/Q). \end{aligned} \tag{23}$$

Since

$$\widehat{|\Phi|^2} = \hat{\Phi} * \hat{\Phi}^*, \quad \text{where} \quad g^*(x) := \overline{g(-x)}$$

is essentially supported inside $\{|\xi| \leqslant \mathbf{C}\cdot N^{-1}\}$, we have

$$\widehat{|\Phi|^2}(A/Q - B/Q) = \delta_{A=B}\cdot\|\Phi\|_{\ell^2(\mathbb{Z})}^2 + O((N/Q)^{-100}) \tag{24}$$

as whenever $A \neq B$, $|A/Q - B/Q| \geqslant Q^{-1} \gg N^{-1}$. Substituting (24) into (23), we find that

$$\begin{aligned}\|K_k' * \varphi_{Q,N}\|_{\ell^2(\mathbb{Z})}^2 " &= " \frac{1}{Q^2}\sum_{A,B\leqslant Q} S(A/Q)\cdot\overline{S(B/Q)}\cdot\delta_{A=B}\cdot\|\Phi\|_{\ell^2(\mathbb{Z})}^2 \\ &= \frac{1}{Q^2}\sum_{A\leqslant Q}|S(A/Q)|^2\cdot\|\Phi\|_{\ell^2(\mathbb{Z})}^2.\end{aligned}$$

By Hua's estimate (22), using the fact that Q is prime, we bound

$$\frac{1}{Q}\sum_{A\leqslant Q}|S(A/Q)|^2 = \frac{1}{Q} + \frac{1}{Q}\sum_{A\leqslant Q-1}|S(A/Q)|^2 \leqslant \mathbf{C}_\varepsilon\cdot(1/Q + Q^{\varepsilon-2/d})$$

so that we find

$$\begin{aligned}\|K_k' * \varphi_{Q,N}\|_{\ell^2(\mathbb{Z})} &\leqslant \mathbf{C}_\varepsilon\cdot Q^{\varepsilon-1/d}\cdot Q^{-1/2}\cdot\frac{N}{2^{dk/2}} \\ &= \mathbf{C}_\varepsilon\cdot Q^{\varepsilon-1/d}\cdot(N/2^{dk})^{1/2}\cdot(N/Q)^{1/2} \\ &\leqslant \mathbf{C}_\varepsilon\cdot Q^{\varepsilon-1/d}\cdot(N/2^{dk})^{1/2}\cdot\|\varphi_{Q,N}\|_{\ell^2(\mathbb{Z})}\end{aligned}$$

The prefactor $(N/2^{dk})^{1/2}$ comes from scaling considerations; if we are interested in an estimate that is independent of scale, we arrive at the bound

$$\|K_k' * \varphi_{Q,N}\|_{\ell^2(\mathbb{Z})} \leqslant \mathbf{C}_\varepsilon\cdot Q^{\varepsilon-1/d}\cdot\|\varphi_{Q,N}\|_{\ell^2(\mathbb{Z})}.$$

In particular, quantitatively, the lower bound

$$\|K_k' * \varphi_{Q,N}\|_{\ell^2(\mathbb{Z})} \geqslant \delta\cdot\|\varphi_{Q,N}\|_{\ell^2(\mathbb{Z})}$$

automatically forces a bound on the "arithmetic complexity" of $\varphi_{Q,N}$ via the estimate

$$Q \leqslant \mathbf{C}_\varepsilon\cdot\delta^{-d-\varepsilon}.$$

In particular, we arrive at the following heuristic:

Heuristic 2.1. *The only obstruction to*

$$\|K_k' * f\|_{\ell^2(\mathbb{Z})} \ll \|f\|_{\ell^2(\mathbb{Z})}$$

are "low arithmetic complexity" considerations.

2.2. The Take-Away

By an application of Weyl's Lemma, a special case of which is stated below, Bourgain was able to make the previous Heuristic 2.1 rigorous, concluding that the above range of examples were typical: if we set

$$\Pi_k(\beta) := \sum_{(A,Q)=1, Q\leqslant 2^{\mathbf{c}k}} \widehat{\chi}(2^{(d-\mathbf{c})k}(\beta - A/Q))$$

for a Schwartz function χ with

$$\mathbf{1}_{[-1/4,1/4]} \leqslant \widehat{\chi} \leqslant \mathbf{1}_{[-1/2,1/2]}$$

then

$$\widehat{K_k}(\beta) = \widehat{K_k}(\beta) \cdot \Pi_k(\beta) + O(2^{-\mathbf{c}'k}), \tag{25}$$

and similarly for K_k'. In particular, whenever $\Pi_k(\beta) \neq 1$, then necessarily the conclusion of Weyl's Lemma holds.

Lemma 2.2 (Weyl's Lemma, Special Case). *Suppose* $|\beta - a/q| \leqslant \frac{1}{q\cdot N^{d-\mathbf{c}}}$ *with*

$$N^{\mathbf{c}} \leqslant q \leqslant N^{d-\mathbf{c}}.$$

Then there exists some $\mathbf{c}_d > 0$ *so that*

$$|\frac{1}{N}\sum_{n\leqslant N} e(-\beta n^d)| \leqslant \mathbf{C}_d \cdot N^{-\mathbf{c}_d}.$$

At this point, by recycling the reasoning from the previous example, one arrives at the physcial-space approximation

$$K_k' \quad \text{“}=\text{”} \quad L_k' := \sum_{(A,Q)=1,\ Q\leqslant 2^{\mathbf{c}k}} S(A/Q)\cdot \mathrm{Mod}_{A/Q}(\chi_{(d-\mathbf{c})k} * \phi_{dk}')$$

in that

$$\|\widehat{K_k' - L_k'}\|_{L^\infty(\mathbb{T})} \leqslant \mathbf{C}_d \cdot 2^{-\mathbf{c}_d k} \tag{26}$$

and so the maximal function is bounded on $\ell^2(\mathbb{Z})$

$$\| \sup_k |(K_k' - L_k') * f| \|_{\ell^2(\mathbb{Z})}^2 \leqslant \sup_\beta \sum_k |\widehat{K_k'}(\beta) - \widehat{L_k'}(\beta)|^2 \cdot \|f\|_{\ell^2(\mathbb{Z})}^2 \leqslant \mathbf{C}_d \cdot \|f\|_{\ell^2(\mathbb{Z})}^2, \quad (27)$$

by arguing as in (12), inserting the quantitative bound (26) for the final inequality.

Following Heuristic 2.1, it makes sense to decompose L_k' according to the approximate level-sets of the Gauss sums, and seek sufficient decay in s on the $\ell^2(\mathbb{Z})$-norms of maximal functions

$$\sup_{k \geqslant \mathbf{C}s} |L_{k,s}' * f|,$$

where

$$L_{k,s}' := \sum_{A/Q \in \mathscr{R}_s} S(A/Q) \cdot \mathrm{Mod}_{A/Q}(\chi_{(d-\mathbf{c})k} * \phi_{dk}')$$

for

$$\mathscr{R}_s := \{(A,Q) = 1,\ 2^{s-1} \leqslant Q < 2^s\} : \quad (28)$$

one bounds

$$\sup_k |L_k' * f| = \sup_k |\sum_{s \leqslant \mathbf{c}k} L_{k,s}' * f| \leqslant \sum_{s=1}^\infty \sup_{k \geqslant \mathbf{C}s} |L_{k,s}' * f|.$$

After a little slight of hand, using Plancherel's theorem to morally extract a geometrically decacying prefactor,

$$L_{k,s}'(x) \quad \text{“}=\text{”} \quad 2^{-\mathbf{c}_d s} \cdot \sum_{A/Q \in \mathscr{R}_s} \mathrm{Mod}_{A/Q}(\chi_{(d-\mathbf{c})k} * \phi_{dk}')(x)$$

it suffices to prove the following maximal inequality (possibly for a slightly different choice of χ):

$$\| \sup_{k \geqslant \mathbf{C}s} |\big(\sum_{A/Q \in \mathscr{R}_s} \mathrm{Mod}_{A/Q} \chi_k \big) * f| \|_{\ell^2(\mathbb{Z})} \leqslant \mathbf{C}_\varepsilon \cdot 2^{\varepsilon s} \cdot \|f\|_{\ell^2(\mathbb{Z})}, \quad \varepsilon > 0;$$

by averaging over translations, exploiting the smoothness of $\{\chi_k : k \geqslant \mathbf{C}s\}$ at physical scales $2^{\mathbf{C}s}$, it suffices to prove the analogous real-variable inequality:

$$\| \sup_{k \geqslant \mathbf{C}s} |\big(\sum_{A/Q \in \mathscr{R}_s} \mathrm{Mod}_{A/Q} \chi_k \big) * f| \|_{L^2(\mathbb{R})} \leqslant \mathbf{C}_\varepsilon \cdot 2^{\varepsilon s} \cdot \|f\|_{L^2(\mathbb{R})}, \quad \varepsilon > 0;$$

finally, by exploiting the dilation invariance of $\mathbb{R}$, matters at last reduce to establishing the following multi-frequency maximal estimate, see Bourgain (1989b):

Proposition 2.3. *Suppose that* $\Theta := \{\theta_1, \dots, \theta_N\}$ *are* 1*-separated,*

$$\textit{i.e.} \quad |\theta_i - \theta_j| > 1, \quad i \neq j.$$

Then

$$\|\mathscr{M}_\Theta f\|_{L^2(\mathbb{R})} := \| \sup_{k \geqslant \mathbf{C}} |\sum_{n \leqslant N} (\mathrm{Mod}_{\theta_n} \chi_k) * f| \|_{L^2(\mathbb{R})} \leqslant \mathbf{C}_\varepsilon \cdot N^\varepsilon \cdot \|f\|_{L^2(\mathbb{R})}, \quad \varepsilon > 0. \quad (29)$$

The proof of Proposition 2.3, which we will presently establish with a bound on the right side (29) of the form $\log^2 N$, combines ideas from harmonic analysis, probability theory, and Banach space geometry, and was a creative novelty, having further applications to problems in pointwise ergodic theory (Bourgain, 1990; Demeter, 2007; Demeter, Lacey, et al., 2008) and to problems in *time frequency analysis*, for instance Demeter, Tao, and Thiele (2008) and Lacey (2000). On the other hand, in some ways, the proof technique was highly constrained: there are only so many ways to control a maximal function on L^2, as we will explore below.

3. The Multi-Frequency Problem

3.1. Preliminary Observations

For what is to follow, we introduce that notation

$$\Xi_k f := \sum_{n \leqslant N} (\mathrm{Mod}_{\theta_n} \chi_k) * f, \tag{30}$$

so that we can express

$$\mathscr{M}_\Theta f = \sup_k |\Xi_k f|;$$

here, as above, χ is a Schwartz function with

$$\mathbf{1}_{[-1/4,1/4]} \leqslant \hat{\chi} \leqslant \mathbf{1}_{[-1/2,1/2]}$$

While the Ξ_k have oscillatory kernels, they admit a natural projection structure, in that

$$\Xi_k \Xi_l = \Xi_l, \quad k \geqslant l + 2,$$

as can be seen by passing to Fourier space, see (31) below; to avoid needless technicality, we will henceforth sparsify our set of scales into parity classes, and restrict our attention to a single class, so that whenever $k > k'$, we necessarily have $k \geqslant k' + 2$.

As establishing Proposition 2.3 is an L^2-based problem, to better understand these convolution operators, we pass to Fourier space, and compute

$$\widehat{\Xi_k f}(\xi) = \sum_{n \leqslant N} \widehat{\chi_k}(\xi - \theta_n) \cdot \hat{f}(\xi) \tag{31}$$

so that

$$\widehat{\Xi_k}(\xi) = \sum_{n \leqslant N} \widehat{\chi_k}(\xi - \theta_n),$$

after conflating the operator with its kernel, so that we can alternatively represent

$$\begin{aligned}\Xi_k f(x) &= \sum_{n\leqslant N} e(\theta_n x)\int \widehat{\chi_k}(\xi)\hat{f}(\xi+\theta_n)e(\xi x)\,d\xi\\ &= \sum_{n\leqslant N} e(\theta_n x)\cdot\big(\chi_k * (\mathrm{Mod}_{-\theta_n} f)\big)(x)\\ &= \sum_{n\leqslant N} e(\theta_n x)\cdot\big(\chi_k * (\chi * \mathrm{Mod}_{-\theta_n} f)\big)(x)\\ &=: \sum_{n\leqslant N} e(\theta_n x)\cdot(\chi_k * f_{\theta_n})(x), \end{aligned}\tag{32}$$

using the fact that $k \geqslant \mathbf{C}$ and a brief argument with the Fourier transform to arrive at the reproducing identity

$$\chi_k * \chi = \chi.$$

The advantage to passing to the formulation involving $\{f_{\theta_n}\}$ is that the smoothing effect of convolution with χ_k has been "factored" out from the oscillatory exponentials $\{e(\theta_n x) : n\}$. In particular, heuristically, on intervals of bounded size $\mathbf{C}$, as k gets large only the exponentials should vary: if $|I| = \mathbf{C}$ is an interval of appropriate length, then whenever $x \in I$ and $2^k \gg \mathbf{C}$

$$I \ni x \mapsto \sum_{n\leqslant N} e(\theta_n x)\cdot \phi_k * f_{\theta_n}(x) \quad \text{“}=\text{”} \quad \sum_{n\leqslant N} e(\theta_n x)\cdot \phi_k * f_{\theta_n}(x_I)\tag{33}$$

for any x_I. In particular, if we subdivide $\mathbb{R}$ into (dyadic) intervals $\{I : |I| = \mathbf{C}\}$ then on each interval we can estimate

$$\begin{aligned}\int_I |\Xi_k f(x)|^2\,dx \quad \text{“}=\text{”} \quad \min_{x_I\in I}\int_I |\sum_n e(\theta_n x)\cdot\chi_k * f_{\theta_n}(x_I)|^2\,dx\\ \leqslant \mathbf{C}\cdot\min_{x_I\in I}\sum_{n\leqslant N}|\chi_k * f_{\theta_n}(x_I)|^2\cdot|I|\end{aligned}\tag{34}$$

as can be seen by bounding

$$\begin{aligned}\|\sum_{n\leqslant N} e(\theta_n x)\cdot a_n\|^2_{L^2(I)} \leqslant \|\sum_{n\leqslant N} e(\theta_n x)\cdot a_n\cdot\psi_I(x)\|^2_{L^2(\mathbb{R})} &= \|\sum_{n\leqslant N} a_n\cdot\widehat{\psi_I}(\xi-\theta_n)\|^2_{L^2(\mathbb{R})}\\ = \sum_{n\leqslant N} a_n\overline{a_m}\cdot\langle\widehat{\psi_I}(\cdot-\theta_n),\widehat{\psi_I}(\cdot-\theta_m)\rangle &= \sum_{n\leqslant N}|a_n|^2\cdot\|\widehat{\psi_I}\|^2_{L^2(\mathbb{R})}\\ \leqslant \mathbf{C}\cdot\sum_{n\leqslant N}|a_n|^2\cdot|I|,\end{aligned}$$

for $\mathbf{1}_I \leqslant |\psi_I| \leqslant \mathbf{C}\cdot(1+\mathrm{dist}(\cdot,I)/|I|)^{-100}$ with a Fourier transform compactly supported inside $[-1/2,1/2]$; this support constrain ensures that

$$\psi_I(\xi-\theta_n)\cdot\psi_I(\xi-\theta_m)\equiv 0, \quad n\neq m,$$

and since $|I| \geqslant \mathbf{C}$, the uncertainty principle is satisfied and such a ψ_I can be chosen.

Seeking uniformity, if we set

$$\mathscr{F}_\Theta(x)^2 := \mathscr{F}_{\Theta,\mathbf{C}}(x)^2 := \sum_{n\leqslant N} \sup_{k\geqslant \mathbf{C}} |\chi_k * f_{\theta_n}(x)|^2, \tag{35}$$

then we have a uniform *norm* bound

$$\sup_{k\geqslant \mathbf{C}} \|\Xi_k f\|_{L^2(I)} \leqslant \mathbf{C}\cdot \min_{x_I\in I} \mathscr{F}_\Theta(x_I)\cdot |I|^{1/2} \leqslant \|\mathscr{F}_\Theta\|_{L^2(I)}, \tag{36}$$

and our task is to control

$$\|\mathscr{M}_\Theta f\|_{L^2(\mathbb{R})} = \Big(\sum_{|I|=\mathbf{C}} \|\sup_{k\geqslant \mathbf{C}} |\Xi_k f|\|_{L^2(I)}^2\Big)^{1/2}, \tag{37}$$

where the previous calculation motivates us to split up the real line into intervals of "small" length and treat the contribution of $\mathscr{M}_\Theta f$ on each interval individually. The problem, therefore, boils down to controlling a supremum on L^2: we can localize and handle the contribution of each individual Ξ_k via the bound

$$\begin{aligned}
\|\mathscr{F}_\Theta\|_{L^2(\mathbb{R})}^2 &= \sum_{n\leqslant N} \int \sup_{k\geqslant \mathbf{C}} |\chi_k * f_{\theta_n}(x)|^2 \leqslant \mathbf{C}\cdot \sum_{n\leqslant N} \int |f_{\theta_n}(x)|^2\,dx = \mathbf{C}\cdot \sum_{n\leqslant N} \int |\widehat{f_{\theta_n}}(\xi)|^2\,d\xi \\
&= \mathbf{C}\cdot \sum_{n\leqslant N} \int |\hat\chi(\xi)|^2\cdot |\widehat{f}(\xi+\theta_n)|^2\,d\xi = \mathbf{C}\cdot \int \sum_{n\leqslant N} |\hat\chi(\xi-\theta_n)|^2\cdot |\hat f(\xi)|^2\,d\xi \\
&\leqslant \mathbf{C}\cdot \sup_\xi \sum_{n\leqslant N} |\hat\chi(\xi-\theta_n)|^2\cdot \|\hat f\|_{L^2(\mathbb{R})}^2 \leqslant \mathbf{C}\cdot \|f\|_{L^2(\mathbb{R})}^2,
\end{aligned} \tag{38}$$

using the separation of the frequencies $|\theta_n - \theta_m| > 1$, $n \neq m$, and the Hardy–Littlewood Maximal function in the first inequality. Our task is to pass from *uniform* control of the $\{\Xi_k : k\}$ to *simultaneous* control, via $\mathscr{M}_\Theta$. This is a task that arises frequently – but is often constrained, as we pause to explore.

3.2. Bounding a Supremum on L^2

Suppose that $\{F_k\} \in L^2(X)$ is a collection of functions on a measure space, and we are interested in controling

$$\|F_*\|_{L^2(X)} := \|\sup_k |F_k|\|_{L^2(X)}. \tag{39}$$

To the best of my knowledge, there are essentially four ways to control F_* on L^2:

- ▷ Martingale/stopping time methods, like those used to prove Doob's Maximal Inequality from martingale theory, or the closely linked Hardy–Littlewood Maximal Inequality;

- ▷ Semigroup methods, like those used in the Hopf–Dunford–Schwartz Maximal Theorem, a special case of which implies dimension independent bounds on the maximal function $\sup_t |e^{t\triangle} f|$;
- ▷ TT^* orthogonality methods, in which the supremum F_* is realized as a particular linear operator,

$$T\{F_k\} := \sum_t \mathbf{1}_{E_k} F_k$$

for $\{E_k\}$ a disjoint partition of X, and then T is composed with its adjoint, to efficiently compute

$$\|T\|_{L^2(X)\to L^2(X)} = \|TT^*\|^{1/2}_{L^2(X)\to L^2(X)};$$

this technique is common in oscillatory integral situations; and

- ▷ Entropy arguments, which leverage vestigial smoothness in the map $k \mapsto F_k(x)$ to control F_*.

Of the four methods, the oscillatory nature of the averages $\{\Xi_k : k\}$ precludes a direct argument involving the first method, which gives a privileged role to the zero frequency (expectation); the serious failure of the identity

$$\Xi_k \Xi_l \neq \Xi_{k+l}$$

precludes the second method.

As for the TT^* approach, if we linearize our supremum and consider the operator

$$Tf(x) = \sum_k \mathbf{1}_{E_k}(x) \cdot \int \sum_{n\leqslant N} e(\theta_n x) \int \chi_k(x-y) e(-\theta_n y) f(y)\, dy,$$

then the dual operator, T^*, is given by

$$T^* g(x) = \sum_r \sum_{n\leqslant N} e(\theta_n y) \cdot \int e(-\theta_n x) \cdot (g \cdot \mathbf{1}_{E_r})(x) \overline{\chi_k}(x-y)\, dx,$$

and nothing is really gained by composition.

Accordingly, we turn our attention to the entropic approach to bounding a supremum on L^2.

4. From Bourgain's Toolkit: The Entropic Method

This section reviews material over which Bourgain had total command at the time of Bourgain (1989b); see Bourgain (1986, 1988a,b, 1989a) or even Bourgain (1989b, §3)

for representative examples, and Tao (2021, §6) for an excellent summary. In particular, I imagine that the information Bourgain gleaned from the above Subsection §3.1 was enough to guide him directly to the below Section §5. While the implementation of this approach in studying $\mathscr{M}_\Theta$ seems magical upon first reading Bourgain (1989b), or in my case the exposition of Thouvenot (1990), my hope is that after fully digesting the following material, the reader is able to understand the intution behind the way Bourgain came to his argument.

The basic mechanism behind the entropic approach is to leverage "size" and "smoothness," or rather "stickiness," in the parameter space to control a supremum. In terms of our problem at hand, we have uniform control over each average Ξ_k via (36), and we search for some notion of smoothness/stickiness to complement this uniformity.

To show off this interplay, we review the following example.

Lemma 4.1 (Sobolev Embedding Lemma). *Suppose that I is an interval, and that $F(x,\cdot)$ is absolutely continuous for almost every x with an L^2 density. Then the following pointwise estimate holds:*

$$F_I(x) := \sup_{t\in I} |F(x,t)| \leqslant \mathbf{C}\cdot|F(x,t_I)| + \mathbf{C}\cdot\left(\int_I |F(x,t)|^2\,dt\right)^{1/4}\cdot\left(\int_I |\partial_t F(x,t)|^2\,dt\right)^{1/4}$$

for any $t_I \in I$. In particular, if

$$\sup_{t\in I} \|F(x,t)\|_{L^2_x} \leqslant A \quad \textit{and} \quad \sup_{t\in I} \|\partial_t F(x,t)\|_{L^2_x} \leqslant a \tag{40}$$

then

$$\big\| \sup_{t\in I} |F(x,t)| \big\|_{L^2_x} \leqslant \mathbf{C}\cdot\big(A + (Aa|I|)^{1/2}\big)$$

Proof. For any $t \in I$, we may bound

$$F(x,t)^2 = F(x,t_I)^2 + \int_{[t_I,t]} \partial_s\big(F(x,s)^2\big)\,ds,$$

so

$$\begin{aligned} |F(x,t)|^2 &\leqslant |F(x,t_I)|^2 + 2\int_I |F(x,t)|\cdot|\partial_t F(x,t)|\,dt \\ &\leqslant |F(x,t_I)|^2 + 2\left(\int_I |F(x,t)|^2\,dt\right)^{1/2}\cdot\left(\int_I |\partial_t F(x,t)|^2\,dt\right)^{1/2} \end{aligned} \tag{41}$$

The right-hand side of (41) is independent of t, so we can take the supremum in t over the left-hand side of (41) and then integrate in x, applying Cauchy–Schwarz to handle the L^2-based t-averages. □

While Lemma 4.1 is very cheap, it is surprisingly robust, and is very useful in studying maximal multiplier operators of the form

$$\sup_t |(\hat{f} \cdot \mathbf{m}(t\cdot))^\vee|$$

for bounded $m \in \mathscr{C}^1(\mathbb{R} \setminus \{0\})$, see Bourgain (1986, Lemma 3).

It is helpful to discretize this argument: for each $v \geqslant 1$, define

$$\Lambda_v := \big(2^{-v} \cdot \mathbb{Z}\big) \cap I$$

and define the parent of $t \in \Lambda_v$, $\rho(t) \in \Lambda_{v-1}$ to be the minimal element so that

$$B(t, 2^{-v}) \cap B(\rho(t), 2^{1-v}) \neq \emptyset, \qquad B(x,s) := \{y : |x-y| < s\}.$$

Given x-a.e. continuity in $t \mapsto F(x,t)$, to study F_I, it suffices to bound

$$\sup_{t \,\in\, \bigcup_{v \geqslant 1} \Lambda_v} |F(x,t)|;$$

by monotone convergence, it suffices to estimate, uniformly in finite subsets $T \subset \bigcup_{v \geqslant 1} \Lambda_v$,

$$F_T(x) := \sup_{t \in T} |F(x,t)|.$$

To do so, for each $t \in T$, we may telescope

$$t = (t - \rho(t)) + (\rho(t) - \rho^2(t)) + \cdots + \rho_0(t)$$

where ρ^j is the jth composition of ρ, and $\rho_0(t)$ is the appropriate composition so that $\rho_0(t) \in \Lambda_{v_0}$ for some v_0 to be determined below.

Note that the number of increments required to arrive at a representative $\rho_0 \in \Lambda_{v_0}$ is uniformly bounded, since T is finite. We bound

$$\begin{aligned} F_T(x) &\leqslant \sup_{t \in \Lambda_{v_0}} |F(x,t)| + \sum_{v > v_0} \sup_{t \in \Lambda_v} |F(x,t) - F(x, \rho(t))| \\ &\leqslant \Big(\sum_{t \in \Lambda_{v_0}} |F(x,t)|^2 \Big)^{1/2} + \sum_{v > v_0} \Big(\sum_{t \in \Lambda_v} |F(x,t) - F(x, \varrho(t))|^2 \Big)^{1/2}, \end{aligned}$$

noting that all sums are in fact finite, and take L^2_x-norms, before optimizing over $v_0 \geqslant 0$ to derive the desired upper bound:

$$A \cdot |\Lambda_{v_0}|^{1/2} + a \cdot \sum_{v > v_0} |\Lambda_v|^{1/2} \cdot 2^{-v} \leqslant \mathbf{C} \cdot \big(A + (Aa|I|)^{1/2} \big);$$

see (40).

In both of these arguments, we relied upon smoothness in the map $t \mapsto F(x,t)$. Really, though, we were relying on decaying contributions from

$$\Lambda_v \ni t \mapsto |F(x,t) - F(x,\rho(t))| \tag{42}$$

as v grows, and the controlled *entropy estimate*

$$|\Lambda_v| \leqslant \mathbf{C} \cdot 2^v \cdot |I|;$$

from the metric perspective, this estimate is measuring the extent to which elements in I adhere to each other —"stick together"— at scales 2^{-v}. Estimates like

$$\sup_{t\in I} \|\partial_t F(x,t)\|_{L^2_x} \leqslant a$$

allow us to capture the smallness in (42) in an L^2-average sense. But, we may also pointwise approximate $\{F(x,t) : t \in I\}$ more directly using a similar telescoping mechanism.

For T as above, consider the set

$$X(x) := X_T(x) := \big\{F(x,t) : t \in T\big\}, \tag{43}$$

and for each v so that $2^{-v} \leqslant 2 \cdot \operatorname{diam}(X(x))$, define $\Lambda_v(x) \subset T$ to be a collection of times t so that

$$X(x) \subset \bigcup_{t\in\Lambda_v(x)} B\big(F(x,t), 2^{-v}\big) \tag{44}$$

subject to the constraint that $|\Lambda_v(x)|$ is minimal; the cardinality is essentially the 2^{-v}-*entropy* of the set.

Now, let V be so large that each element of T is separated by $> 2^{1-V}$, so that $T = \Lambda_V(x)$. And define the parent of $t \in \Lambda_v(x)$, $\varrho(t) \in \Lambda_{v-1}(x)$ to be the minimal element so that

$$B\big(F(x,t), 2^{-v}\big) \cap B\big(F(x,\varrho(t)), 2^{1-v}\big) \neq \emptyset. \tag{45}$$

For any $s \in T$, we may similarly bound

$$\begin{aligned} F_T(x) &\leqslant |F(x,s)| + \sum_v \sup_{t\in\Lambda_v(x)} |F(x,t) - F(x,\varrho(t))| \\ &\leqslant |F(x,s)| + \sum_v \Big(\sum_{t\in\Lambda_v(x)} |F(x,t) - F(x,\varrho(t))|^2 \Big)^{1/2} \\ &\leqslant |F(x,s)| + \mathbf{C} \cdot \sum_v 2^{-v} \cdot |\Lambda_v(x)|^{1/2}, \end{aligned} \tag{46}$$

as we may bound

$$|F(x,t) - F(x,\rho'(t))| < 2^{-v} + 2^{1-v} < 2^{2-v}$$

for each $t \in \Lambda_v(x)$ by (45). It is convenient to change perspectives and bound

$$|\Lambda_v(x)| \leqslant N_{2^{-v}}(x) \tag{47}$$

where

$$N_\lambda(x) := \sup\Big\{K : \text{ there exists a sequence of times } t_0 < t_1 < \cdots < t_K : |F(x,t_i) - F(x,t_{i-1})| > \lambda\Big\}$$

is a so-called (greedy) *jump-counting function* at altitude $\lambda > 0$, which measures the extent to which $\{F(x,t) : t\}$ "stick together" at the scale λ:

$$N_\lambda(x) < \infty \text{ for all } \lambda > 0 \iff \{F(x,t) : t\} \text{ converges} \iff \{F(x,t) : t\} \text{ "stick together" at all scales.}$$

To establish (47), one majorizes the left hand side and minorizes the right hand by the 2^{1-v}-*entropy* of the set: the size of the largest set of 2^{1-v}-separated points inside of $\{F(x,t) : t \in I\}$.

The reverse bound

$$N_{2^{-v}}(x) \leqslant |\Lambda_{v+1}(x)|$$

is simpler, so there is nothing lost quantitatively from this change, as indeed

$$\sum_v 2^{-v} \cdot |\Lambda_v(x)|^{1/2} \leqslant \sum_v 2^{-v} \cdot N_{2^{-v}}(x)^{1/2} \leqslant \mathbf{C} \cdot \sum_v 2^{-v} \cdot |\Lambda_v(x)|^{1/2}.$$

In many special examples, one is able to prove a uniform bound

$$\sup_v \|2^{-v} \cdot N_{2^{-v}}^{1/2}\|_{L^2} \leqslant \mathbf{C} \cdot A, \tag{48}$$

which says that in an L^2-averaged sense

$$N_{2^{-v}} \quad \text{" } \leqslant \text{ "} \quad \mathbf{C} \cdot A^2 \cdot 2^{2v}$$

i.e. that it costs a quadratically growing price to cover the collection of data $\{F(x,t) : t\}$ by balls of a given radius. The following examples are representative.

Entropic Example One. — Consider the (discrete-time) averaging operators,

$$F(x,t) = \mathbb{E}_k f(x) \cdot \mathbf{1}_{[2^k, 2^{k+1})}(t) \tag{49}$$

where

$$\mathbb{E}_k f(x) := \sum_{|I|=2^k \text{ dyadic}} \big(\frac{1}{|I|}\int_I f(t)\,dt\big) \cdot \mathbf{1}_I(x) \tag{50}$$

is the conditional expectation operator, projecting onto the σ-algebra generated by the dyadic intervals $\{2^k \cdot [n, n+1) : n \in \mathbb{Z}\}$. The "stopping-time" structure embedded in the definition of $N_{2^{-v}}$ allows one to neatly employ methods from dyadic harmonic analysis – secretly, martingale techniques – to establish (48).

Entropic Example Two. — To the extent that

$$\mathbb{E}_k f \quad \text{“}=\text{”} \quad \chi_k * f,$$

in that both operators "blur" at spatial scales 2^k, discarding "fine scale" information below this threshold, and preserving "coarse scale" properties that can be detected above this spatial threshold, one can combine a square function argument with further orthogonality arguments, in particular the quantitative bound

$$\|\mathbb{E}_k \psi_l - \chi_k * \psi_l\|_{L^2(\mathbb{R})} \leqslant \mathbf{C} \cdot 2^{-|k-l|/2} \cdot \|\psi_l\|_{L^2(\mathbb{R})}, \quad \psi_l := \chi_l - \chi_{l-1}, \tag{51}$$

to extend (48) to the case where

$$F(x,t) = f * \chi_k(x) \cdot \mathbf{1}_{[2^k, 2^{k+1})}(t), \tag{52}$$

and similarly with χ replaced with any other Schwartz function with $\hat{\chi}(0) = 1$. These ideas first appeared in Jones, Kaufman, et al. (1998).

4.1. The Jump-Counting Approach to Entropy

While the uniform estimate (48) is a priori insufficient to control the full supremum over $t \in T$, this entropic argument yields a remarkable strengthening over the trivial estimate

$$\|F_T\|_{L^2_x} \leqslant \|S_T\|_{L^2_x} \leqslant |T|^{1/2} \cdot A,$$

where we set

$$S_T(x)^2 := \sum_{t \in T} |F(x,t)|^2.$$

In particular, for any $t \in T$,

$$F_T(x) \leqslant |F(x,t)| + \mathbf{C} \cdot \sum_v 2^{-v} \cdot N_{2^{-v}}(x)^{1/2}$$
$$\leqslant |F(x,t)| + \frac{S_T(x)}{|T|^{1/2}} + \sum_{v:\, \frac{S(x)}{|T|^{1/2}} \leqslant 2^{-v} \leqslant 2 \cdot S(x)} 2^{-v} \cdot N_{2^{-v}}(x)^{1/2} \qquad (53)$$

so that, essentialy, the uniform bound (48) implies[(2)]

$$\|F_T\|_{L^2_x} \quad \text{“} \leqslant \text{”} \quad \mathbf{C} \cdot \log|T| \cdot A. \qquad (55)$$

In point of fact, as we will see below, (55) often holds with a $\log^2|T|$ prefactor.

4.2. Introduction to Variation

As the difficulty with the heuristic justification for (54) shows, see (53), a major problem is that, in general, we cannot expect a uniform bound on

$$x \mapsto \sup_v 2^{-v} \cdot N_{2^{-v}}(x)^{1/2},$$

see Jones and Wang (2004) or Qian (1998).

To get around this issue, one instead sacrifices the power $1/2 \to 1/r$, $r > 2$ and introduces the so-called *r-variation* of $\{F(x,t) : t \in I\}$

$$\mathscr{V}^r(x) := \mathscr{V}^r_F(x) := \sup \big(\sum_i |F(x,t_i) - F(x,t_{i-1})|^r \big)^{1/r}, \qquad (56)$$

where the supremum runs over all finite increasing subsequences inside of I. Unlike the jump counting function, the r-variation operators crucially satisfies a triangle inequality,

$$\mathscr{V}^r_{F+G} \leqslant \mathscr{V}^r_F + \mathscr{V}^r_G,$$

and one may bound

$$\sup_v 2^{-v} \cdot N_{2^{-v}}(x)^{1/r} \leqslant \mathscr{V}^r(x),$$

(2) There is a natural comparison between this estimate and the abstract Hilbert space *Rademacher–Menshov* inequality, which also states that

$$\| \sup_{n \leqslant N} |F(x,n)| \|_{L^2} \leqslant \mathbf{C} \cdot \log N \cdot A \qquad (54)$$

under orthogonality constrains on the functions $\{F(\cdot,n) : n\}$. The analogy is at the level of proof and is that of Lebesgue integration to Riemann integration: the entropy bound organizes the data $\{F(x,n) : n\}$ according to its image, while the Rademacher–Menshov inequality is proven by analogously organizing the data according to the domain of the time parameter $n \in [N]$.

which is important, as the $\mathscr{V}^r$ operators often admit a strong L^2-theory. In particular, if $|T| = N$, so that $N_\lambda \leqslant N$ for all λ, we may bound

$$2^{-v} \cdot N_{2^{-v}}^{1/2} \leqslant 2^{-v} \cdot N_{2^{-v}}^{1/r} \cdot N_{2^{-v}}^{1/2-1/r} \leqslant N^{1/2-1/r} \cdot \mathscr{V}^r \tag{57}$$

and if we set $r = 2 + \frac{\mathbf{c}}{\log N}$, then we eliminate the pre-factor of $N^{1/2-1/r}$ and end up with the bound

$$2^{-v} \cdot N_{2^{-v}}^{1/2} \leqslant \mathbf{C} \cdot \mathscr{V}^r, \qquad r = 2 + \frac{\mathbf{c}}{\log N}.$$

Substituting into (53), we bound, for any $t \in T$

$$F_T(x) \leqslant |F(x,t)| + \frac{S_T(x)}{N^{1/2}} + \sum_{v:\, \frac{S_T(x)}{N^{1/2}} \leqslant 2^{-v} \leqslant S_T(x)} \mathscr{V}^r(x) \leqslant |F(x,t)| + \frac{S_T(x)}{N^{1/2}} + \log N \cdot \mathscr{V}^r(x),$$

which says that

$$\|F_T\|_{L^2} \leqslant \mathbf{C} \cdot \big(A + \log N \cdot \|\mathscr{V}^r\|_{L^2}\big), \qquad r = 2 + \frac{\mathbf{c}}{\log N},$$

so control over the r-variation operators leads, essentially, to (55).

The relevant estimates for $\mathscr{V}^r$ derive, in many cases, from the following inequality, classically used as a convergence result in martingale theory Lépingle (1976); see Jones, Seeger, and Wright (2008) for a discussion, and Guo, Roos, and Yung (2020) or Oberlin et al. (2012) for more exotic examples.

Proposition 4.2 (Lépingle's Inequality, Special Case). *The following estimate holds in the conditional expectation case* (49)*:*

$$\|\mathscr{V}^r\|_{L^2(\mathbb{R})} \leqslant \mathbf{C} \cdot \frac{r}{r-2} \cdot A.$$

Proposition 4.2 extends similarly to the case of convolution operators (52): by combining a square function argument

$$\begin{aligned} \mathscr{V}^r(\chi_k * f : k) &\leqslant \mathscr{V}^r(\mathbb{E}_k f : k) + \mathscr{V}^r(\chi_k * f - \mathbb{E}_k f : k) \\ &\leqslant \mathscr{V}^r(\mathbb{E}_k f : k) + 2 \cdot \big(\sum_k |\chi_k * f - \mathbb{E}_k f|^2\big)^{1/2} \end{aligned} \tag{58}$$

with the estimates (51) introduced above, one can use orthogonality techniques and Proposition 4.2 to bound both terms in (58). Above, we define the discrete-time variation

$$\mathscr{V}^r(f_k : k)(x) := \sup \big(\sum_i |f_{k_i}(x) - f_{k_{i-1}}(x)|^r\big)^{1/r}$$

where the supremum runs over all finite subsequences $\{k_i\}$.

While the $\mathscr{V}^r$ operators are more delicate than the pertaining maximal functions,

$$F_T(x) \leqslant |F(x,t)| + \mathscr{V}^r(x)$$

for any $t \in T$, they are essentially of even strength, in that we have the following heuristic:

Heuristic 4.3. *In either case* (49) *or* (52)*, it is very hard for $\mathscr{V}^r$ to be large when both F_I and the square function*

$$S_I(x) := \big(\sum_{k:2^k \in I} |F(x,2^k) - F(x,2^{k+1})|^2 \big)^{1/2}$$

are small: $\mathscr{V}^r$ " $\approx$ " $\frac{r}{r-2} \cdot (F_I + S_I)$ " $\approx$ " $\frac{r}{r-2} \cdot F_I$, $\frac{r}{r-2} \cdot S_I$.[3]

Finally, and significantly, given our vector-valued perspective on studying

$$\{f_{\theta_1}, \dots, f_{\theta_N}\},$$

see (34), we observe that just as do the maximal function and square function, the $\mathscr{V}^r$ operators interact well in the vector-valued setting: for sequence-space valued functions $\vec{F} = (F_1, F_2, \dots)$

$$\mathscr{V}^r_{\vec{F}}(x) := \sup \big(\sum_i \|F_n(x,t_i) - F_n(x,t_{i-1})\|^r_{\ell^2_n} \big)^{1/r} \leqslant \|\mathscr{V}^r_{F_n}(x)\|_{\ell^2_n}, \tag{59}$$

by Minkowski's inequality for sequence spaces (as $r > 2$), where the supremum runs over finite increasing subsequences of $\{t_i\}$.

With this section in mind, we begin to see how Bourgain developed his argument.

5. The Argument Takes Shape

Bourgain's task was to establish (37), where we are only thinking about the case where 2^k is very large relative to $|I| = \mathbf{C}$. By monotone convergence, we can restrict to finitely many scales $k \in T \subset \mathbb{N} \cap [\mathbf{C}, \infty)$. We focus on the case of a single interval.

Guided by our heuristic analysis, we let $x_I \in I$ be a point to be determined later, and seek to bound

$$\| \sup_k | \sum_{n \leqslant N} e(\theta_n x) \cdot \chi_k * f_{\theta_n}(x_I) | \|_{L^2_x(I)}.$$

[3] The close link between maximal function and square function in either context (49) or (52) is classical; see e.g. Stein (1993, §1). The relationship between $\mathscr{V}^r, F_I, S_I$ in the case (49) is via the following good-λ inequality:

$$|\{\mathscr{V}^r > \mathbf{C}\lambda, F_I, S_I \leqslant \gamma\lambda\}| \leqslant \mathbf{C} \cdot \big(\frac{r}{r-2}\big)^2 \cdot \gamma^2 \cdot |\{\mathscr{V}^r > \lambda\}|, \quad r > 2;$$

see Krause (2023, §3).

As discussed above – we are essentially forced to use the entropic approach. Specifically, we set

$$X(x_I) := \{\chi_k * \vec{f_\Theta}(x_I) := \big(\chi_k * f_{\theta_1}(x_I), \dots, \chi_k * f_{\theta_N}(x_I)\big) : k\} \tag{60}$$

and let $\vec{N}_\lambda$ denote the appropriate jump-counting function at altitude λ with respect to the sequence space norm, $\ell^2([N])$,

$$\vec{N}_\lambda(x) := \sup\Big\{K : \text{ there exists a sequence of times } \mathbf{C} \leqslant k_0 < k_1 < \cdots < k_K : \\ \|\chi_{k_i} * f_{\theta_n}(x) - \chi_{k_{i-1}} * f_{\theta_n}(x)\|_{\ell^2_{n\in[N]}} > \lambda\Big\}.$$

By arguing as above we can bound

$$\begin{aligned} &\|\sup_k |\sum_{n\leqslant N} e(\theta_n x)\cdot \chi_k * f_{\theta_n}(x_I)|\|_{L^2_x(I)} \\ &\quad\leqslant |\Xi_{k_0} * f(x_I)|\cdot |I|^{1/2} + \sum_v \|\max_{k\in\Lambda_v(x_I)} |\sum_{n\leqslant N} e(\theta_n x)\cdot \big(\chi_k - \chi_{\varrho(k)}\big) * f_{\theta_n}(x_I)|\|_{L^2_x(I)} \end{aligned} \tag{61}$$

for any $k_0 \geqslant \mathbf{C}$, see (30). The first term is of a simpler nature, so we will temporarily suppress it; and for each individual v we may bound

$$\begin{aligned} &\|\max_{k\in\Lambda_v(x_I)} |\sum_{n\leqslant N} e(\theta_n x)\cdot \big(\chi_k - \chi_{\varrho(k)}\big) * f_{\theta_n}(x_I)|\|_{L^2_x(I)} \\ &\quad\leqslant \|\big(\sum_{k\in\Lambda_v(x_I)} |\sum_{n\leqslant N} e(\theta_n x)\cdot \big(\chi_k - \chi_{\varrho(k)}\big) * f_{\theta_n}(x_I)|^2\big)^{1/2}\|_{L^2_x(I)} \\ &\quad\leqslant \big(\sum_{k\in\Lambda_v(x_I)} \|\sum_{n\leqslant N} e(\theta_n x)\cdot \big(\chi_k - \chi_{\varrho(k)}\big) * f_{\theta_n}(x_I)\|^2_{L^2_x(I)}\big)^{1/2} \\ &\quad\leqslant \mathbf{C}\cdot 2^{-v}\cdot \vec{N}_{2^{-v}}(x)^{1/2}\cdot |I|^{1/2} \end{aligned} \tag{62}$$

by arguing as in (46), applying (34) to bound

$$\|\sum_{n\leqslant N} e(\theta_n x)\cdot \big(\chi_k - \chi_{\varrho(k)}\big) * f_{\theta_n}(x_I)\|_{L^2_x(I)} \leqslant \mathbf{C}\cdot 2^{-v}\cdot |I|^{1/2}$$

uniformly for $k \in \Lambda_v(x_I)$. Above, $\{\Lambda_v(x_I) : v\}$ are sets of times that are minimal with respect to the property that

$$X(x_I) \subset \bigcup_{k\in\Lambda_v(x_I)} B_{\ell^2([N])}\big(\chi_k * \vec{f_\Theta}(x_I), 2^{-v}\big),$$

where $B_{\ell^2([N])}(\vec{v}, r)$ is the ball of radius r centered at $\vec{v} \in \ell^2([N])$ with respect to the sequence-space norm $\ell^2([N])$, and the parent function, ϱ, is as above.

The issue is the potential explosion

$$\vec{N}_{2^{-v}}(x_I) \to \infty \text{ as } v \to \infty,$$

and there is no a priori way to rule out this enemy; if there were, there would be no logarithmic loss in (55). The clever insight that Bourgain had that allowed him to push past this abstract issue was just Cauchy–Schwarz: we bound

$$|\sum_{n\leqslant N} e(\theta_n x)\cdot(\chi_k - \chi_{\varrho(k)}) * f_{\theta_n}(x_I)| \leqslant N^{1/2}\cdot\big(\sum_{n\leqslant N} |(\chi_k-\chi_{\varrho(k)})*f_{\theta_n}(x_I)|^2\big)^{1/2} \leqslant \mathbf{C}\cdot N^{1/2}\cdot 2^{-v}$$

uniformly for $k \in \Lambda_v(x_I)$, which yields the cheap bound

$$\| \max_{k\in\Lambda_v(x_I)} |\sum_{n\leqslant N} e(\theta_n x)\cdot\big(\chi_k - \chi_{\varrho(k)}\big) * f_{\theta_n}(x_I)|\|_{L^2_{\vec{x}}(I)} \leqslant \mathbf{C}\cdot 2^{-v}\cdot N^{1/2}\cdot |I|^{1/2}.$$

Altogether, Bourgain had obtained the bounds

$$\| \max_{k\in\Lambda_v(x_I)} |\sum_{n\leqslant N} e(\theta_n x)\cdot\big(\chi_k - \chi_{\varrho(k)}\big) * f_{\theta_n}(x_I)|\|_{L^2_{\vec{x}}(I)} \leqslant \mathbf{C}\cdot 2^{-v}\cdot \min\{\vec{N}_{2^{-v}}(x_I)^{1/2}, N^{1/2}\}\cdot |I|^{1/2},$$

see (62), which he cleverly interpolated, as per (57),

$$\begin{aligned} 2^{-v}\cdot\min\{\vec{N}_{2^{-v}}(x_I)^{1/2}, N^{1/2}\} &\leqslant 2^{-v}\cdot\vec{N}_{2^{-v}}(x_I)^{1/r}\cdot N^{1/2-1/r} \\ &\leqslant N^{1/2-1/r}\cdot \mathscr{V}^r_{\vec{f}_\Theta}(x_I) \leqslant \mathbf{C}\cdot\mathscr{V}^r_{\vec{f}_\Theta}(x_I) \qquad r = 2+\frac{\mathbf{c}}{\log N} \end{aligned}$$

see (59) and (60), obtaining a v-independent term on the right. Inserting this bound and arguing as in the heuristic analysis (53),

$$\begin{aligned} (61) &\leqslant \mathbf{C}\cdot \sum_{v:2^{-v}\leqslant\mathscr{F}_\Theta(x_I)} 2^{-v}\cdot\min\{\vec{N}_{2^{-v}}(x_I)^{1/2}, N^{1/2}\}\cdot|I|^{1/2} \\ &\leqslant \mathbf{C}\sum_{v:2^{-v}\leqslant\mathscr{F}_\Theta(x_I)/N^{1/2}} 2^{-v}\cdot N^{1/2}\cdot|I|^{1/2} + \mathbf{C}\sum_{v:\mathscr{F}_\Theta(x_I)/N^{1/2}\leqslant 2^{-v}\leqslant 2\cdot\mathscr{F}_\Theta(x_I)} \mathscr{V}^r_{\vec{f}_\Theta}(x_I)\cdot|I|^{1/2} \\ &\leqslant \mathbf{C}\cdot\mathscr{F}_\Theta(x_I)\cdot|I|^{1/2} + \mathbf{C}\cdot\log N\cdot\mathscr{V}^r_{\vec{f}_\Theta}(x_I)\cdot|I|^{1/2}, \qquad r = 2+\frac{\mathbf{c}}{\log N}. \quad (63) \end{aligned}$$

And, at last, after choosing x_I carefully, we bound

$$\|\mathscr{M}_\Theta f\|_{L^2(I)} \leqslant \mathbf{C}\cdot\|\mathscr{F}_\Theta\|_{L^2(I)} + \mathbf{C}\cdot\log N\cdot\|\mathscr{V}^r_{\vec{f}_\Theta}\|_{L^2(I)},$$

which says that in a scale-$\mathbf{C}$, L^2-averaged sense, at all locations one has the following inequality

$$\mathscr{M}_\Theta f \quad \text{“}\leqslant\text{”} \quad \mathscr{F}_\Theta + \log N\cdot\mathscr{V}^r_{\vec{f}_\Theta}, \qquad r = 2+\frac{\mathbf{c}}{\log N}.$$

In other words, up to logarithmic error, the vector valued maximal function and the vector-valued variation control $\mathscr{M}_\Theta$. And, square-summing over $\{|I| = \mathbf{C}\}$, taking into account the related, convolution-based version of Lépingle's Inequality, Proposition 4.2, leads to the bound

$$\|\mathscr{M}_\Theta f\|_{L^2(\mathbb{R})} \leqslant \mathbf{C} \cdot (1 + \log N \cdot \frac{r}{r-2}) \cdot \|f\|_{L^2(\mathbb{R})} \leqslant \mathbf{C} \cdot (1 + \log^2 N) \cdot \|f\|_{L^2(\mathbb{R})},$$

which satisfies (32).

Guided by this intuition, we turn to the rigorous proof.

6. The Proof of Proposition 2.3, The Multi-Frequeny Maximal Inequality

Motivated by our previous outline, we will seek to prove the following estimate:

$$\|\mathscr{M}_\Theta f\|_{L^2(\mathbb{R})} \leqslant \mathbf{C} \cdot \log^2 N \cdot \|f\|_{L^2(\mathbb{R})}.$$

Accordingly we will restrict our attention only to scales $k \geqslant \mathbf{C} \log^2 N$, and just handle the complementary cases using a square function argument

$$\begin{aligned}
\| \sup_{\mathbf{C} \leqslant k \leqslant \mathbf{C} \cdot \log^2 N} |\Xi_k f| \|_{L^2(\mathbb{R})} &\leqslant \| \Big(\sum_{\mathbf{C} \leqslant k \leqslant \mathbf{C} \cdot \log^2 N} |\Xi_k f|^2 \Big)^{1/2} \|_{L^2(\mathbb{R})} \\
&\leqslant \mathbf{C} \cdot \log N \cdot \sup_k \|\Xi_k f\|_{L^2(\mathbb{R})} = \mathbf{C} \cdot \log N \cdot \sup_k \|\widehat{\Xi_k f}\|_{L^2(\mathbb{R})} \\
&\leqslant \mathbf{C} \cdot \log N \cdot \sup_k \|\widehat{\Xi_k}\|_{L^\infty(\mathbb{R})} \cdot \|\hat{f}\|_{L^2(\mathbb{R})} \leqslant \mathbf{C} \cdot \log N \cdot \|f\|_{L^2(\mathbb{R})},
\end{aligned}$$

as

$$\sup_k \sup_\xi |\widehat{\Xi_k}(\xi)| \leqslant 1,$$

see (31). We accordingly re-define $\mathscr{F}_\Theta = \mathscr{F}_{\Theta, \log^2 N}$, see (35), and observe the inherited smoothness

$$\begin{aligned}
|\mathscr{F}_\Theta(x) - \mathscr{F}_\Theta(y)| &\leqslant \mathbf{C} \cdot \sum_{|I| = 2^k \geqslant \log^2 N} \sum_{n \leqslant N} |\chi_k * f_{\theta_n}(x) - \chi_k * f_{\theta_n}(y)| \\
&\leqslant \mathbf{C} \cdot N \cdot \sum_{k \geqslant \log^2 N} (|x-y| \cdot 2^{-k}) \cdot M_{\mathrm{HL}} f(x) \leqslant \mathbf{C} \cdot |x-y| \cdot N^{-100} \cdot M_{\mathrm{HL}} f(x), \quad (64)
\end{aligned}$$

(say), which says that $\mathscr{F}_\Theta$ is very smooth at scales $|I| = \mathbf{C}$. Above, we used the bound

$$|\partial \chi_k(x)| \leqslant \mathbf{C} \cdot 2^{-k} \cdot 2^{-k} \cdot (1 + 2^{-k} \cdot |x|)^{-100}.$$

This excision of scales allows us to be a little less delicate than Bourgain in making rigorous the heuristic (33): whereas Bourgain used a so-called *best constant argument*, we will just use the following estimate, which is effective for small intervals relative to the scales $k \geqslant \mathbf{C} \cdot \log^2 N$:

$$\sum_{n \leqslant N} e(\theta_n x) \cdot \chi_k * f_{\theta_n}(x) = \sum_{n \leqslant N} e(\theta_n x) \cdot \chi_k * f_{\theta_n}(x_I) + O\big(\frac{N \cdot |I|}{2^k} \cdot M_{\mathrm{HL}} f(x)\big)$$
$$= \sum_{n \leqslant N} e(\theta_n x) \cdot \chi_k * f_{\theta_n}(x_I) + O\big(2^{-k/2} \cdot M_{\mathrm{HL}} f(x)\big)$$

for any $x_I \in I$, certainly provided that $|I| \leqslant N^{\mathbf{C}}$.

In particular, for any $x \in I$, with $|I| = \mathbf{C}$, we may bound

$$\sup_{k \geqslant \mathbf{C} \cdot \log^2 N} |\Xi_k f(x)| \leqslant \sup_{k \geqslant \mathbf{C} \cdot \log^2 N} \big| \sum_{n \leqslant N} e(\theta_n x) \cdot \chi_k * f_n(x_I)\big| + O\big(\sum_{k \geqslant \mathbf{C} \cdot \log^2 N} 2^{-k/2} \cdot M_{\mathrm{HL}} f(x)\big),$$

so that for each I we may bound

$$\| \sup_{k \geqslant \mathbf{C} \cdot \log^2 N} |\Xi_k f(x)| \|_{L^2(I)} \leqslant \mathbf{C} \cdot \min_{x_I \in I} \| \sup_{k \geqslant \mathbf{C} \cdot \log^2 N} \big| \sum_{n \leqslant N} e(\theta_n x) \cdot \chi_k * f_n(x_I)\big| \|_{L^2(I)}$$
$$+ \mathbf{C} \cdot N^{-100} \cdot \|M_{\mathrm{HL}} f\|_{L^2(I)}$$

(say). Temporarily dropping the term involving M_{HL} as inessential, we consider the first term

$$\| \sup_{k \geqslant \mathbf{C} \cdot \log^2 N} \big| \sum_{n \leqslant N} e(\theta_n x) \cdot \chi_k * f_n(x_I)\big| \|_{L^2(I)},$$

which we bound using the entropic approach, see (63),

$$\| \sup_{k \geqslant \mathbf{C} \cdot \log^2 N} \big| \sum_{n \leqslant N} e(\theta_n x) \cdot \chi_k * f_n(x_I)\big| \|_{L^2(I)}$$
$$\leqslant \mathbf{C} \cdot \big(\mathscr{F}_\Theta(x_I) \cdot |I|^{1/2} + \log N \cdot \mathscr{V}^r_{\vec{f}_\Theta}(x_I) \cdot |I|^{1/2}\big),$$
$$\leqslant \mathbf{C} \cdot \big(\|\mathscr{F}_\Theta\|_{L^2(I)} + \log N \cdot \|\mathscr{V}^r_{\vec{f}_\Theta}\|_{L^2(I)} + N^{-100} \cdot \|M_{\mathrm{HL}} f\|_{L^2(I)}\big),$$

after choosing x_I to minimize $\mathscr{V}^r_{\vec{f}_\Theta}$ on I, and using the smoothness

$$\mathscr{F}_\Theta(x_I) = \mathscr{F}_\Theta(x) + O(N^{-100} \cdot M_{\mathrm{HL}} f(x))$$

to bound

$$\mathscr{F}_\Theta(x_I) \cdot |I|^{1/2} = \|\mathscr{F}_\Theta\|_{L^2(I)} + O\big(N^{-100} \cdot \|M_{\mathrm{HL}} f\|_{L^2(I)}\big).$$

In particular, we have bounded

$$\| \sup_{k \geqslant \mathbf{C} \cdot \log^2 N} |\Xi_k f| \|_{L^2(I)} \leqslant \mathbf{C} \cdot \big(\|\mathscr{F}_\Theta\|_{L^2(I)} + \log N \cdot \|\mathscr{V}^r_{\vec{f}_\Theta}\|_{L^2(I)} + N^{-100} \cdot \|M_{\mathrm{HL}} f\|_{L^2(I)}\big)$$

where $r = 2 + \frac{\mathbf{c}}{\log N}$, so square-summing over $|I| = \mathbf{C}$ yields, at last, the bound

$$\begin{aligned}
\|\mathscr{M}_\Theta f\|_{L^2(\mathbb{R})} &\leqslant \big(\sum_{|I|=\mathbf{C}} \| \sup_{k \geqslant \mathbf{C}\cdot\log^2 N} |\Xi_k f| \|_{L^2(I)}^2 \big)^{1/2} + \mathbf{C}\cdot\log N \cdot \|f\|_{L^2(\mathbb{R})} \\
&\leqslant \mathbf{C}\cdot \big(\sum_{|I|=\mathbf{C}} \|\mathscr{F}_\Theta\|_{L^2(I)}^2 \big)^{1/2} + \mathbf{C}\cdot\log N \cdot \big(\sum_{|I|=\mathbf{C}} \|\mathscr{V}^r_{\vec{f}_\Theta}\|_{L^2(I)}^2 \big)^{1/2} \\
&\qquad + \mathbf{C}\cdot N^{-100} \cdot \big(\sum_{|I|=\mathbf{C}} \|M_{\mathrm{HL}} f\|_{L^2(I)}^2 \big)^{1/2} + \mathbf{C}\cdot\log N \cdot \|f\|_{L^2(\mathbb{R})} \\
&\leqslant \mathbf{C}\cdot \big(\|\mathscr{F}_\Theta\|_{L^2(\mathbb{R})} + \log N \cdot \|\mathscr{V}^r_{\vec{f}_\Theta}\|_{L^2(\mathbb{R})} + N^{-100} \cdot \|M_{\mathrm{HL}} f\|_{L^2(\mathbb{R})} + \log N \cdot \|f\|_{L^2(\mathbb{R})} \big) \\
&\leqslant \mathbf{C}\cdot\log^2 N \cdot \|f\|_{L^2(\mathbb{R})},
\end{aligned}$$

completing the proof.

7. Contemporary Work

Since Bourgain's work, the topic of pointwise convergence of ergodic averages along polynomial orbits was taken up and greatly advanced by Mariusz Mirek, Eli Stein, and their collaborators (Mirek, 2018; Mirek, Stein, and Trojan, 2017, 2019; Mirek, Stein, and Zorin-Kranich, 2020; Mirek and Trojan, 2016), building on breakthrough work of Ionescu and Wainger (2006). The current state of affairs was established in Mirek, Stein, and Zorin-Kranich (2020):

Theorem 7.1. *Suppose that (X, μ) is a σ-finite measure space, $\tau : X \to X$ is a measure-preserving transformation, and $P \in \mathbb{Z}[\cdot]$ is a polynomial with integer coefficients. Then for each $1 < p < \infty$, $r > 2$*

$$\begin{aligned}
&\|\mathscr{V}^r\big(\frac{1}{N}\sum_{n=1}^N \tau^{P(n)} f : N\big)\|_{L^p(X)} + \sup_{\lambda>0} \|\lambda \cdot N_\lambda\big(\frac{1}{N}\sum_{n=1}^N \tau^{P(n)} f : N\big)^{1/2}\|_{L^p(X)} \\
&\qquad \leqslant \mathbf{C}_p \cdot (1 + \frac{r}{r-2}) \cdot \|f\|_{L^p(X)}.
\end{aligned}$$

In other words, from a quantitative perspective, the rate of convergence of the abstract averages

$$\frac{1}{N}\sum_{n=1}^N \tau^{P(n)} f$$

is precisely that of our entropic examples!

The key to this argument was a combinatorial partitioning of $\mathbb{Q} \cap [0,1]$ into the so-called Ionescu–Wainger exhaustion of the rationals: one replaces

$$\mathscr{R}_s \longrightarrow \mathscr{U}_s,$$

see (28), where $\{\mathscr{U}_s : s\}$ form a disjoint partition of $\mathbb{Q} \cap [0,1]$ which captures many of the analytical properties of $\mathscr{R}_s$, namely

$$\sup_{A/Q \in \mathscr{U}_s} |S(A/Q)| \leqslant \mathbf{C}_\varepsilon \cdot 2^{(\varepsilon - 1/d)s}, \tag{65}$$

but admit much more favorable arithmetic statistics, which allows for the approximation

$$K'_k \quad \text{“}=\text{”} \quad \sum_{s \leqslant \mathbf{c}\cdot k} \sum_{A/Q \in \mathscr{U}_s} S(A/Q) \cdot \mathrm{Mod}_{A/Q}(\chi_{(d-\mathbf{c})k} * \phi'_{dk}) =: \sum_{s \leqslant \mathbf{c}\cdot k} L_{k,s}$$

to hold in $\ell^p(\mathbb{Z})$ as well.

Although Bourgain's entropic argument is less effective in general on ℓ^p, $p \neq 2$, by applying the Rademacher–Menshov inequality and arguing as in Bourgain (1988b), one is able to establish e.g. the estimate

$$\| \sup_k |L_{k,s} * f| \|_{\ell^p(\mathbb{Z})} \leqslant \mathbf{C}_\varepsilon \cdot 2^{\varepsilon s} \cdot 2^{-\mathbf{c}_{p,d} s} \cdot \|f\|_{\ell^p(\mathbb{Z})}, \qquad 1 < p < \infty,\ \mathbf{c}_{p,d} < 1/d,$$

and similarly for the jump-counting formulation. The loss in the number of frequencies is sub-exponential in s, as in the case of Bourgain's maximal function on ℓ^2; the gain of

$$2^{-\mathbf{c}_{p,d} s}$$

follows from appropriately interpolating (65).

This quantitative improvement over the sharpest estimates for $\sup_k |L'_{k,s} * f|$,

$$\| \sup_k |L'_{k,s} * f| \|_{\ell^p(\mathbb{Z})} \leqslant \mathbf{C}_{\varepsilon,p} \cdot 2^{(\varepsilon+1)s} \cdot 2^{-\mathbf{c}_{p,d} s} \cdot \|f\|_{\ell^p(\mathbb{Z})}, \quad 1 < p < \infty,\ \mathbf{c}_{p,d} < 1/d,$$

speaks to the flexibility of these arguments, which indeed extend to handle the case of the r-variation and jump-counting operators.

References

Bellow, A. (1982). *Two problems*. Measure theory, Proc. Conf., Oberwolfach 1981, Lect. Notes Math. 945, 15-23 (1982).

Bellow, A. and Losert, V. (1984). *On sequences of density zero in ergodic theory*. Contemp. Math. 26, 49-60 (1984).

Birkhoff, G. D. (1931). “Proof of the ergodic theorem.” *Proc. Natl. Acad. Sci. USA* **17**, pp. 656–660.

Bourgain, J. (1986). “On high dimensional maximal functions associated to convex bodies”, *Am. J. Math.* **108**, pp. 1467–1476.

——— (1988a). "Almost sure convergence and bounded entropy", *Isr. J. Math.* **63** (1), pp. 79–97.

——— (1988b). "On the maximal ergodic theorem for certain subsets of the integers", *Isr. J. Math.* **61** (1), pp. 39–72.

——— (1988c). "On the pointwise ergodic theorem on L^p for arithmetic sets", *Isr. J. Math.* **61** (1), pp. 73–84.

——— (1989a). "Bounded orthogonal systems and the Λ (p)-set problem", *Acta Math.* **162** (3-4), pp. 227–245.

——— (1989b). "Pointwise ergodic theorems for arithmetic sets. With an appendix on return-time sequences, jointly with Harry Furstenberg, Yitzhak Katznelson and Donald S. Ornstein", *Publ. Math., Inst. Hautes Étud. Sci.* **69**, pp. 5–45.

——— (1990). "Double recurrence and almost sure convergence", *J. Reine Angew. Math.* **404**, pp. 140–161.

Buczolich, Z. and Mauldin, R. D. (2007). "Concepts behind divergent ergodic averages along the squares", in: *Ergodic theory and related fields. Papers of the 2004–2006 Chapel Hill workshops on probability and ergodic theory, Chapel Hill, NC, USA, 2004–2006*. Providence, RI: American Mathematical Society (AMS), pp. 41–56.

Calderón, A. P. (1968). "Ergodic theory and translation-invariant operators", *Proc. Natl. Acad. Sci. USA* **59**, pp. 349–353.

Demeter, C. (2007). "Pointwise convergence of the ergodic bilinear Hilbert transform", *Ill. J. Math.* **51** (4), pp. 1123–1158.

Demeter, C., Lacey, M. T., et al. (2008). "Breaking the duality in the return times theorem", *Duke Math. J.* **143** (2), pp. 281–355.

Demeter, C., Tao, T., and Thiele, C. (2008). "Maximal multilinear operators", *Trans. Am. Math. Soc.* **360** (9), pp. 4989–5042.

Guo, S., Roos, J., and Yung, P.-L. (2020). "Sharp variation-norm estimates for oscillatory integrals related to Carleson's theorem", *Anal. PDE* **13** (5), pp. 1457–1500.

Hua, L. K. (1982). *Introduction to number theory. Transl. from the Chinese by Peter Shiu*. Berlin-Heidelberg-New York: Springer-Verlag. xviii, 572 p., 14 figs. DM 96.00; $ 42.70 (1982).

Ionescu, A. D. and Wainger, S. (2006). "L^p boundedness of discrete singular Radon transforms", *J. Am. Math. Soc.* **19** (2), pp. 357–383.

Jones, R. L., Kaufman, R., et al. (1998). "Oscillation in ergodic theory", *Ergodic Theory Dyn. Syst.* **18** (4), pp. 889–935.

Jones, R. L., Seeger, A., and Wright, J. (2008). "Strong variational and jump inequalities in harmonic analysis", *Trans. Am. Math. Soc.* **360** (12), pp. 6711–6742.

Jones, R. L. and Wang, G. (2004). "Variation inequalities for the Fejér and Poisson kernels", *Trans. Am. Math. Soc.* **356** (11), pp. 4493–4518.

Krause, B. (2023). *Discrete Analogues in Harmonic Analysis: Bourgain, Stein, and Beyond*. Vol. 224. Grad. Stud. Math. Providence, RI: American Mathematical Society (AMS).

Lacey, M. T. (2000). "The bilinear maximal functions map into L^p for $2/3 < p \leqslant 1$", *Ann. Math.* (2) **151** (1), pp. 35–57.

Lavictoire, P. (2011). "Universally L^1-bad arithmetic sequences", *J. Anal. Math.* **113**, pp. 241–263.

Lépingle, D. (1976). "La variation d'ordre p des semi-martingales", *Z. Wahrscheinlichkeitstheor. Verw. Geb.* **36**, pp. 295–316.

Mirek, M. (2018). "Square function estimates for discrete Radon transforms", *Anal. PDE* **11** (3), pp. 583–608.

Mirek, M., Stein, E. M., and Trojan, B. (2017). "$\ell^p\left(\mathbb{Z}^d\right)$-estimates for discrete operators of Radon type: variational estimates", *Invent. Math.* **209** (3), pp. 665–748.

——— (2019). "$\ell^p(\mathbb{Z}^d)$-estimates for discrete operators of Radon type: maximal functions and vector-valued estimates", *J. Funct. Anal.* **277** (8), pp. 2471–2521.

Mirek, M., Stein, E. M., and Zorin-Kranich, P. (2020). "Jump inequalities for translation-invariant operators of Radon type on $\mathbb{Z}^d$", *Adv. Math.* **365**. Id/No 107065, p. 57.

Mirek, M. and Trojan, B. (2016). "Discrete maximal functions in higher dimensions and applications to ergodic theory", *Am. J. Math.* **138** (6), pp. 1495–1532.

Oberlin, R. et al. (2012). "A variation norm Carleson theorem", *J. Eur. Math. Soc.* (*JEMS*) **14** (2), pp. 421–464.

Qian, J. (1998). "The p-variation of partial sum processes and the empirical process", *Ann. Probab.* **26** (3), pp. 1370–1383.

Stein, E. M. (1993). *Harmonic analysis: Real-variable methods, orthogonality, and oscillatory integrals. With the assistance of Timothy S. Murphy*. Vol. 43. Princeton Math. Ser. Princeton, NJ: Princeton University Press.

Tao, T. (2021). "Exploring the toolkit of Jean Bourgain", *Bull. Am. Math. Soc., New Ser.* **58** (2), pp. 155–171.

Thouvenot, J.-P. (1990). *On the almost-sure convergence of ergodic means following certain subsequences of the integers* [*after Jean Bourgain*]. Sémin. Bourbaki, Vol. 1989/90, 42ème année, Astérisque 189-190, Exp. No. 719, 133-153 (1990).

Ben Krause

King's College London
Mathematics Department
Strand, London WC2R 2LS

E-mail: ben.krause@kcl.ac.uk

Séminaire BOURBAKI Novembre 2022
74e année, 2022–2023, no 1200, p. 121 à 140
doi : 10.24033/ast.1209

SUR UN THÉORÈME DE LANG–WEIL TORDU
[*d'après* E. Hrushovski, K. V. Shuddhodan et Y. Varshavsky]

par **Silvain Rideau-Kikuchi**

Le fil directeur de cet exposé est la grande uniformité du comportement asymptotique, pour les grandes valeurs de q, du morphisme de Frobenius

$$\begin{array}{rccc} \phi_q : & K & \to & K \\ & x & \mapsto & x^q, \end{array}$$

où K est un corps (algébriquement clos) de caractéristique positive p et q est une puissance de p.

Les estimées de Lang et Weil (1954) en sont un exemple classique : étant donné une variété algébrique [(1)] X de dimension d sur $\overline{\mathbf{F}}_p$, pour toute puissance q de p suffisamment grande, X est définie par des équations à coefficients dans $\mathbf{F}_q$ et le morphisme de Frobenius induit un morphisme $\phi_{q,X} : X \to X$. Le nombre de points fixes de $\phi_{q,X}$ est alors de l'ordre de

$$q^d + O(q^{d-1/2}).$$

De plus, la constante du O ne dépend que de la complexité des équations qui définissent X — mais pas de K ou p.

Dans la mesure où les points fixes de ϕ_q^n ne sont rien d'autre que les éléments du corps fini $\mathbf{F}_{q^n}$, l'énoncé ci-dessus est, en fait, une estimation du nombre de points de X dans le corps $\mathbf{F}_{q^n}$; ce qui est la manière classique de présenter le problème.

Un théorème remarquable de Hrushovski (2004) donne une explication générale à ce comportement asymptotique uniforme, du moins pour ce qui est des propriétés que l'on peut exprimer par une formule du premier ordre. Ici, on considère des formules qui s'écrivent avec des symboles pour l'addition, la soustraction, la multiplication, les éléments 0 et 1 et un endomorphisme σ fixé et où on autorise la quantification uniquement sur les éléments des corps que l'on considère. Par exemple, la formule

$$\forall x \forall y \, \sigma(x+y) = \sigma(x) + \sigma(y) \wedge \sigma(x \cdot y) = \sigma(x) \cdot \sigma(y) \wedge \sigma(1) = 1$$

(1) Dans cet exposé, on ne considèrera que des variétés *irréductibles* sur un corps K algébriquement clos.

qui exprime que σ est un morphisme d'anneaux, ou encore les formules

$$\forall x_0 \dots \forall x_n \, (\bigwedge_{i \leqslant n} \sigma(x_i) = x_i) \to \exists y \, \sigma(y) = y \wedge y^{n+1} + \sum_{i \leqslant n} x_i y^i = 0$$

qui expriment que le corps fixé par σ est algébriquement clos.

La théorie des modèles a une riche histoire de tels résultats asymptotiques. L'un des plus anciens est un principe de transfert entre la grande caractéristique positive et la caractéristique zéro, qui découle immédiatement de la compacité de la logique du premier ordre. Pour toute formule ψ *sans le symbole d'automorphisme*,

$$\text{pour tout } p \text{ grand, la formule } \psi \text{ est vérifiée dans } \overline{\mathbf{F}}_p,$$

si et seulement si

$$\text{la formule } \psi \text{ est vérifiée dans } \mathbf{C}.$$

Le théorème de Hrushovski généralise ce principe en prenant en compte le morphisme de Frobenius ϕ_q. D'après ce théorème, celui-ci se comporte pour les grandes valeurs de q comme un automorphisme de corps générique. Plus précisément, il existe une classe de « corps avec automorphisme générique » dont on peut donner une axiomatisation explicite, habituellement notée ACFA — *cf.* proposition 1.2 pour une liste de ces axiomes. Cette classe est l'équivalent, pour les corps avec automorphisme, des corps algébriquement clos pour les corps sans automorphisme et la section 1 de cet exposé a pour but d'en donner une présentation plus exhaustive

Le résultat précis, que nous démontrerons dans la section 3, énonce alors que pour toute formule ψ,

$$\text{pour tout } q \text{ grand } {}^{(2)}, \text{ la formule } \psi \text{ est vraie de } \overline{\mathbf{F}}_q \text{ et } \phi_q,$$

si et seulement si

$$\text{la formule } \psi \text{ est une conséquence des axiomes ACFA.}$$

Ce résultat repose en grande partie sur deux résultats qui décrivent le comportement du morphisme ϕ_q quand q varie. Le premier est un théorème de théorie algébrique des nombres, le théorème de densité de Chebotarev (voir, par exemple, Fried et Jarden, 2008, théorème 6.3.1), qui décrit les liens entre le morphisme ϕ_q et les extensions cycliques des corps de nombres. Le second est un résultat de géométrie algébrique qui généralise les estimées de Lang–Weil. C'est d'ailleurs un phénomène remarquable en soi que les propriétés asymptotiques au premier ordre du morphisme de Frobenius ne dépendent que de ces deux résultats, profonds, mais qui ne semblent *a priori* couvrir qu'une petite partie des propriétés exprimables par des formules.

(2) On considère bien ici toutes les puissances de tous les nombres premiers.

Comme on l'a vu ci-dessus, on peut voir les estimées de Lang–Weil comme un énoncé de comptage du nombre de points fixes du morphisme de Frobenius. On peut naturellement se demander si ce phénomène d'uniformité reste vrai pour des questions de comptage plus générales impliquant l'automorphisme de Frobenius. Le théorème 2.1 donne une réponse positive à cette question — ainsi que son titre à l'exposé.

Pour toute variété X sur un corps K algébriquement clos, on note $X^{(q)} = X^{\phi_q}$ le changement de base de X le long de l'automorphisme de Frobenius et $\phi_{q,X} : X \to X^{(q)}$ le morphisme induit par ϕ_q. Si $X \subseteq K^n$ est le lieu des zéros des polynômes $P_1, \dots, P_m \in K[x_1, \dots, x_n]$, alors $X^{(q)}$ est le lieu des zéros des polynômes $\phi_q(P_i)$ et $\phi_{q,X}$ est l'action de ϕ_q coordonnée par coordonnée.

Les estimées de Lang–Weil tordues affirment alors que, pour toute variété X de dimension d sur K, tout q suffisamment grand et toute sous-variété $C \subseteq X \times X^{(q)}$ vérifiant certaines hypothèses techniques, l'intersection de C avec le graphe du morphisme de Frobenius $\phi_{q,X}$ contient un nombre de points de l'ordre de

$$cq^d + O(q^{d-1/2}),$$

où c est une constante qui dépend de la géométrie de C, et les différentes bornes ne dépendent que de la complexité des équations qui définissent X et C.

Ce résultat est aussi originellement dû à Hrushovski (2004) et sa preuve repose sur le développement de nouveaux outils, intéressants en eux-mêmes, introduits à cette occasion — dont le développement d'une théorie des schémas aux différences ainsi que la théorie des modèles de certains corps valués avec automorphisme. Nous exposerons, dans la section 2, une preuve récente de Shuddhodan et Varshavsky (2022) qui, pour les citer, est « purement géométrique ».

Enfin, dans la section 4, nous discuterons de certaines applications de ces résultats en dynamique algébrique et en géométrie algébrique aux différences.

Pour conclure cette introduction, je voudrais remercier Élisabeth Bouscaren, Antoine Chambert-Loir et Martin Hils pour nos discussions ainsi que leurs nombreux commentaires dont ce texte a grandement bénéficié.

1. Automorphisme générique

Commençons par introduire la notion d'automorphisme « générique » ainsi que divers outils de théorie des modèles nécessaires à sa définition.

Un objet central de ce texte est l'étude des équations dites « aux différences », c'est-à-dire les équations de la forme (pour le cas en une variable)

$$\sum_{i_j \leqslant d} a_i x^{i_0} \sigma(x)^{i_1} \cdots \sigma^n(x)^{i_n} = 0,$$

où les coefficients a_i sont dans un corps K sur lequel on a choisi un endomorphisme σ — on parle alors de corps aux différences (K, σ). Historiquement, le nom d'équations aux différences est hérité du cas où $K = k(t)$ est un corps de fonctions rationnelles sur un corps k et $\sigma(f) = f(t+1)$. Un autre exemple classique est celui des équations aux q-différences où l'on considère le morphisme $\sigma(f) = f(qt)$. Comme on peut le voir dans ces exemples, leur étude est étroitement liée à celle de la dynamique algébrique.

Pour étudier ces équations, il est utile de disposer de « domaines universels » dans lesquels toutes les équations aux différences ont des solutions — voire suffisamment de solutions pour pouvoir en détecter la structure. La théorie des modèles fournit une notion abstraite d'un tel domaine :

Définition 1.1. Un corps aux différences (K, σ) est dit *existentiellement clos* si tout système d'équations aux différences sur K (en plusieurs variables) qui a une solution dans un corps aux différences (L, τ) qui contient K — et dont l'endomorphisme τ étend σ — a une solution dans K.

En d'autres termes, pour toute formule $\psi(x_1, \ldots, x_n)$ sans quantificateurs à paramètres dans K dans laquelle les variables non quantifiées sont parmi $x_1, \ldots, x_n$, si ψ est vérifiée par des éléments $a_1, \ldots, a_n$ d'une extension de (K, σ), elle est déjà vérifiée par des élément $c_1, \ldots, c_n \in K$.

On dit que l'automorphisme σ est *générique* puisque tout comportement possible d'un automorphisme de corps se retrouve dans le corps aux différences (K, σ).

Si l'on travaille seulement dans le langage des anneaux (c'est-à-dire sans le symbole pour l'endomorphisme σ), un corps est existentiellement clos si et seulement s'il est algébriquement clos — c'est exactement ce qu'énonce le *Nullstellensatz* de Hilbert. Il suffit donc dans ce cas de considérer des équations en une seule variable.

Les corps aux différences existentiellement clos sont, par définition, ceux qui vérifient une forme du *Nullstellensatz* pour les équations aux différences. Mais il est aussi possible, dans ce cas, de caractériser quelles équations aux différences doivent avoir une solution dans un corps aux différences pour qu'il soit existentiellement clos.

Soit X une variété sur un corps aux différences (K, σ) (algébriquement clos). On note X^σ le changement de base de X le long de σ et $\sigma_X : X \to X^\sigma$ le morphisme induit par σ. Comme précédemment, si $X \subseteq K^n$ est le lieu des zéros des polynômes $P_1, \ldots, P_m \in K[x_1, \ldots, x_n]$, alors X^σ est le lieu des zéros des polynômes $\sigma(P_i)$ obtenus en faisant agir σ sur les coefficients et pour tout $(a_1, \ldots, a_n) \in X$, $\sigma_X(a) = (\sigma(a_1), \ldots, \sigma(a_n))$.

Un morphisme $f : X \to Y$ entre variétés est dit *dominant* s'il est d'image dense pour la topologie de Zariski. On peut maintenant énoncer la caractérisation suivante, isolée par Hrushovski (*cf.* Macintyre (1997, p. 172-173)), des corps aux différences existentiellement clos :

Proposition 1.2. *Un corps aux différences (K, σ) est existentiellement clos si et seulement si :*

1. *Le corps K est algébriquement clos ;*
2. *Le morphisme σ est surjectif ;*
3. *Pour toute variété affine X sur K et toute sous-variété $C \subseteq X \times X^\sigma$ telle que les projections vers X et X^σ sont dominantes, l'ensemble des couples $(x, \sigma(x))$, avec $x \in X(K)$, est Zariski dense dans C.*

Ces conditions peuvent s'exprimer par un ensemble (infini) de formules. La principale subtilité est de pouvoir quantifier sur les sous-variétés $C \subseteq X \times X^\sigma$. Pour cela, il faut vérifier que, pour toute sous-variété $X \subseteq \mathbf{A}_K^{n+m}$, l'ensemble des $y \in \mathbf{A}_K^n$ tels que la fibre $X_y \subseteq \mathbf{A}_K^m$ est (géométriquement) irréductible est donné par une formule — en d'autres termes, que le lieu d'irréductibilité (géométrique) de la famille X_y est constructible. Mais cela est bien connu, voir par exemple Grothendieck (1966, théorème 9.7.7) ou van den Dries et Schmidt (1984) pour une approche de cette question par le biais d'algèbres de polynômes non standards. La seconde difficulté est de détecter, par une formule, quand un morphisme est dominant. Mais il suffit, pour cela, de savoir que la dimension d'une fibre est continue pour la topologie constructible — voir, par exemple, Grothendieck (1966, théorème 9.5.5).

La classe des corps aux différences existentiellement clos admet donc une axiomatisation (infinie), qui est habituellement notée ACFA. Cette axiomatisation n'est pas complète, c'est-à-dire qu'il existe des corps aux différences (K, σ) et (L, τ) existentiellement clos qui ne vérifient pas les mêmes formules sans variable non quantifiée — on parle habituellement d'énoncés. Pour qu'ils vérifient les mêmes énoncés, il faudrait évidemment que K et L aient la même caractéristique. Mais il faut également que les restrictions des automorphismes à la clôture algébrique de leur corps premier soient conjuguées. Macintyre (1997, p. 173-174) montre que c'est, en fait, suffisant :

Proposition 1.3. *Soient (K, σ) et (L, τ) des corps aux différences existentiellement clos, $F \leqslant K$ un sous-corps aux différences algébriquement clos et $f : F \to L$ un morphisme de corps aux différences — c'est-à-dire un morphisme de corps tel que $\sigma \circ f = f \circ \tau$. Alors pour toute formule $\psi(x_1, \ldots, x_n)$ et tout $a \in F^n$,*

$$a \text{ réalise } \psi \text{ dans } (K, \sigma) \text{ si et seulement si } f(a) \text{ réalise } \psi \text{ dans } (L, \tau).$$

En particulier, deux corps aux différences existentiellement clos (K, σ) et (L, τ) vérifient les mêmes énoncés si et seulement si :

$$\text{les corps aux différences } (K_0, \sigma|_{K_0}) \text{ et } (L_0, \tau|_{L_0}) \text{ sont isomorphes,}$$

où K_0 (respectivement L_0) est la clôture algébrique du corps premier de K (respectivement L).

Le résultat précédent n'est pas exactement un résultat d'élimination des quantificateurs, mais il implique tout de même que les formules ont toutes une forme très simple, à équivalence près :

Corollaire 1.4. *Toute formule $\psi(x_1, \ldots, x_n)$ est équivalente, modulo les axiomes de* ACFA, *à une disjonction de formules de la forme :*

$$\exists y\, \theta(x_1, \ldots, x_n, y),$$

où θ est sans quantificateurs et, pour tous $a_1, \ldots a_n$, il n'y a qu'un nombre fini uniformément borné de y qui vérifient $\theta(a_1, \ldots, a_n, y)$.

La proposition 1.2 a aussi pour conséquence que le corps fixé par σ dans un corps aux différences existentiellement clos est un corps dit « pseudo-fini » :

Corollaire 1.5. *Soient (K, σ) un corps aux différences existentiellement clos et $K^\sigma = \{x \in K \mid \sigma(x) = x\}$ le corps fixé par σ. On a alors que :*

- ▷ *Le corps K^σ est parfait ;*
- ▷ *Le groupe de Galois absolu de K^σ est engendré par σ et est donc isomorphe à $\hat{\mathbf{Z}}$;*
- ▷ *Toute variété (géométriquement intègre) sur K^σ a un point dans K^σ.*

Cette classe doit son nom à Ax (1968) qui a démontré qu'un énoncé dans le language des corps est vrai dans les corps finis de grand cardinal si et seulement s'il est vrai dans les corps pseudo-finis. Comme les corps finis sont exactement les corps fixés par les morphismes de Frobenius ϕ_q, le résultat d'Ax est un cas particulier, ainsi qu'une des inspirations majeures, du théorème de Hrushovski (2004) dont nous discutons dans cet exposé.

La théorie des modèles des corps aux différences existentiellement clos est un sujet très riche qui va bien au-delà des quelques résultats fondamentaux, mais relativement élémentaires, que nous utilisons ici. L'un des résultats centraux de cette théorie est le théorème de trichotomie de Chatzidakis et Hrushovski (1999) et Chatzidakis, Hrushovski et Peterzil (2002) qui donne une description fine de la géométrie des ensembles de dimension 1 définissables par une formule. Ces résultats ont aussi des conséquences remarquables en dynamique algébrique, comme, par exemple, les travaux de Medvedev et Scanlon (2014). Mais leur exposition nous emporterait bien loin de notre sujet principal.

2. Estimées de Lang-Weil tordues

Revenons maintenant à la question du comportement asymptotique du morphisme de Frobenius. Soient K un corps algébriquement clos de caractéristique strictement positive p et q une puissance de p. D'après la proposition 1.2, pour montrer

que ϕ_q se comporte comme un automorphisme générique, il faut, *a minima*, montrer que la condition 3 est vraie, pour q grand. C'est le but de cette section, dans laquelle, nous montrons une estimation précise (voir théorème 2.1) du nombre de points dans une telle intersection en suivant une preuve récente « purement géométrique » due à Shuddhodan et Varshavsky (2022).

Si $f\colon X \to Y$ est un morphisme dominant entre variétés sur K de même dimension, on note $\deg(f) = [K(X) : K(Y)]$; c'est le nombre de points (comptés avec multiplicités) dans une fibre générique de f. On note aussi $\deg_{\mathrm{ins}}(f)$ son degré inséparable ; c'est le degré $[K(X) : L]$ où L est la sous-extension séparable maximale de $K(Y)$ dans $K(X)$. On note aussi $\Gamma_{q,X} \subseteq X \times X^{(q)}$ le graphe de $\phi_{q,X}$.

Théorème 2.1. *Soient X une variété de dimension d sur un corps K algébriquement clos de caractéristique p, q une puissance de p et $C \subseteq X \times X^{(q)}$ une sous-variété telle que les projections $p_1\colon C \to X$ et $p_2\colon C \to X^{(q)}$ sont dominantes et p_2 est quasi-finie. Alors, si q est suffisamment grand, $\#C \cap \Gamma_{q,X}(K)$ est fini et*

$$\#C \cap \Gamma_{q,X}(K) = \frac{\deg(p_1)}{\deg_{\mathrm{ins}}(p_2)} q^d + O(q^{d-1/2}).$$

De plus, les diverses bornes ne dépendent que de la complexité des équations qui définissent X et C. Pour être précis, si $X \subseteq \mathbf{P}^n_K$ est localement fermée, les bornes ne dépendent que de n et des degrés de la clôture $\overline{X}$ de X, de $\overline{X} \setminus X$, de la clôture $\overline{C}$ de C et de $\overline{C} \setminus C$, en particulier, elles ne dépendent pas du corps K ou de sa caractéristique.

Exemple 2.2. Dans le cas où X est définie sur $\mathbf{F}_q$, on identifie X à $X^{(q)}$. Soit alors $C = \Delta \subseteq X \times X$ la diagonale, on a

$$\Delta \cap \Gamma_{q,X}(K) = \{(x,x) : x \in X(K) \text{ et } x = \phi_q(x)\}.$$

Comme dans ce cas-là $\deg(p_1) = \deg_{\mathrm{ins}}(p_2) = 1$, on retrouve les estimées de Lang–Weil :

$$\#X(\mathbf{F}_q) = q^d + O(q^{d-1/2}).$$

Exemple 2.3. Considérons l'exemple, très simple, où X est la droite affine $\mathbf{A}^1_K$ et $C \subseteq \mathbf{A}^2_K$ est l'ensemble $\{(x,y) : x^m = y^n\}$ pour un choix d'entiers m et n. Dans ce cas-là, on a $\deg(p_1) = n$ et $\deg_{\mathrm{ins}}(p_2) = p^r$ où r est l'exposant de p dans la décomposition de m en facteurs premiers — et donc $m = p^r s$ où s est un entier premier à p. Alors, si $q \geqslant p^r$,

$$\begin{aligned}\#C \cap \Gamma_{q,X}(K) &= \#\{x \in K : x^{p^r s} = x^{qn}\} \\ &= \#\{x \in K : x^{p^r(p^{-r}qn - s)} = 1 \text{ ou } x = 0\} \\ &= \frac{n}{p^r} q - s + 1.\end{aligned}$$

Des estimations quantitatives du théorème 2.1, on déduit aisément l'énoncé qualitatif suivant, qui n'est pas sans rappeler la condition 3 de la proposition 1.2 :

Corollaire 2.4. *Soient X une variété sur K et $C \subseteq X \times X^{(q)}$ une sous-variété telle que les projections $p_1 \colon C \to X$ et $p_2 \colon C \to X^{(q)}$ sont dominantes. Alors pour tout ouvert de Zariski $U \subseteq C$ non-vide, si q est suffisamment grand,*

$$U \cap \Gamma_{q,X}(K) \neq \emptyset.$$

La borne sur q ne dépend, de nouveau, que de la complexité des équations qui définissent X, C et U — en particulier, elle ne dépend ni de K, ni de sa caractéristique.

Dans le cas où X est définie sur $\mathbf{F}_q$ et $K = \overline{\mathbf{F}}_q$ (et donc $X^{(q)}$ est naturellement identifié à X), ces résultats sont mieux compris. Par exemple, Shuddhodan (2022) démontre des résultats plus fins sur le comportement de la suite des $\#C \cap \Gamma_{q^n,X}(K)$ quand n croît :

Théorème 2.5. *Soient X une variété sur $\overline{\mathbf{F}}_q$ définie sur $\mathbf{F}_q$ et $C \subseteq X \times X$ une sous-variété telle que p_2 est quasi-finie. Pour tout n suffisamment grand, $a_n = \#C \cap \Gamma_{q^n,X}(\overline{\mathbf{F}}_q)$ est fini et la série entière*

$$\sum_n a_n t^n \in \mathbf{Z}[[t]]$$

est rationnelle — c'est-à-dire un élément de $\mathbf{Q}(t)$.

Mais nous n'aborderons pas ici ces raffinements.

2.1. Le cas projectif lisse

Pour démontrer le théorème 2.1, nous allons tout d'abord nous concentrer sur le cas où X est lisse projective et p_2 est étale. Cela nous permettra d'introduire certains des principaux outils de la preuve.

À toute variété projective lisse X de dimension d et tout $n \leqslant d$, on associe le groupe $Z^n(X)$ des cycles de X de codimension n. C'est le groupe abélien libre engendré par les sous-variétés de X de codimension n.

Étant donné deux sous-variétés Y_1 et Y_2 de X de codimensions respectives n et $d - n$, on souhaite leur associer un entier qui représente le nombre de points dans leur intersection. Quand cette intersection est finie, il est naturel de considérer le nombre de ses points (comptés avec multiplicité).

En général, on s'autorise à d'abord « déplacer » Y_1 et Y_2 pour que leur intersection soit de dimension zéro. On obtient alors un produit d'intersection

$$\cdot \colon A^n(X) \times A^{d-n}(X) \to \mathbf{Z},$$

où $A^n(X)$ est le quotient de $Z^n(X)$ par équivalence rationnelle. Pour toute sous-variété $Y \subseteq X$ de codimension n, on note $[X]$ sa classe dans $A^n(X)$.

Dans le cas qui nous intéresse, comme p_2 est étale, les sous-variétés C et $\Gamma_{q,X}$ de $X \times X^{(q)}$ s'intersectent transversalement — c'est-à-dire que l'intersection est finie et que chaque point est de multiplicité égale à un — et donc

$$\#C \cap \Gamma_{q,X}(K) = [C] \cdot [\Gamma_{q,X}]. \tag{1}$$

L'intérêt de se ramener au calcul d'un nombre d'intersection est que l'on dispose alors d'outils cohomologiques.

Soit ℓ un nombre premier différent de p. À toute variété (propre et lisse) X, on associe ses groupes de cohomologie ℓ-adique $H^i(X, \mathbf{Q}_\ell)$ — ce sont des $\mathbf{Q}_\ell$-espaces vectoriels de dimension finie $\beta_i(X)$. De plus, pour tout morphisme de variétés (propres et lisses) $f\colon X \to Y$, on a un morphisme de tiré en arrière $f^\star\colon H^i(Y, \mathbf{Q}_\ell) \to H^i(X, \mathbf{Q}_\ell)$. Si $\dim(Y) = \dim(X)$, on a aussi un morphisme de poussé en avant $f_\star\colon H^i(X, \mathbf{Q}_\ell) \to H^i(Y, \mathbf{Q}_\ell)$. On peut étendre cette fonctorialité aux sous-variétés irréductibles $C \subseteq X \times Y$ de même dimension que Y — et donc de codimension $\dim(X)$ — en posant

$$H^i([C]) = (p_2)_\star \circ p_1^\star\colon H^i(X, \mathbf{Q}_\ell) \to H^i(C, \mathbf{Q}_\ell) \to H^i(Y, \mathbf{Q}_\ell),$$

où $p_1\colon C \to X$ et $p_2\colon C \to Y$ sont les projections. On étend H^i à tous les cycles de codimension $\dim(X)$ par linéarité.

On peut alors écrire la formule de Grothendieck-Lefschetz, voir par exemple Grothendieck (1977, proposition 3.3), qui relie le nombre d'intersection à la trace de divers morphismes de cohomologie. Pour tout $\alpha \in A^d(X \times X^{(q)})$, on a

$$\alpha \cdot [\Gamma_{q,X}] = \sum_{i=0}^{2d} (-1)^i \mathrm{Tr}(\phi_{q,X}^\star \circ H^i(\alpha)). \tag{2}$$

Il nous faut donc estimer chacun des termes de la somme dans le cas $\alpha = [C]$. Pour ce qui est du degré $i = 2d$, $H^{2d}(X)$ est de dimension 1 et on peut calculer explicitement, par exemple en considérant l'image d'un cycle de dimension zéro, que

$$\phi_{q,X}^\star \circ H^{2d}([C]) = \phi_{q,X}^\star \circ (p_2)_\star \circ p_1^\star = q^d \deg(p_1)\mathrm{id}$$

et donc

$$\mathrm{Tr}(\phi_{q,X}^\star \circ H^{2d}([C])) = \deg(p_1) q^d. \tag{3}$$

Il reste à montrer que les contributions des termes de degré $i < 2d$ sont négligeables. On fixe dorénavant un plongement $\iota\colon \overline{\mathbf{Q}}_\ell \to \mathbf{C}$ et on identifie $\overline{\mathbf{Q}}_\ell$ à un sous-corps de $\mathbf{C}$. Pour tout $f\colon H^i(X, \mathbf{Q}_\ell) \to H^i(X, \mathbf{Q}_\ell)$, on note $\rho(f)$ — ou $\rho(f, H^i(X, \mathbf{Q}_\ell))$ s'il peut y avoir une ambiguïté — son rayon spectral (archimédien). C'est le maximum des normes des valeurs propres de f. On a alors

$$|\mathrm{Tr}(\phi_{q,X}^\star \circ H^i([C]))| \leqslant \beta_i(X) \rho(\phi_{q,X}^\star \circ H^i([C])) = O(\rho(\phi_{q,X}^\star \circ H^i([C]))),$$

où $\beta_i(X)$ est la dimension de $H^i(X, \mathbf{Q}_\ell)$.

Si X et C sont définis sur $\mathbf{F}_q$, on peut naturellement identifier $X^{(q)}$ à X et $\phi^\star_{q,X}$ commute alors avec $H^i([C])$. On a donc $\rho(\phi^\star_{q,X} \circ H^i([C])) \leqslant \rho(\phi^\star_{q,X}, H^i(X, \mathbf{Q}_\ell))\rho(H^i[C])$ et le théorème 2.1 découle alors du théorème de pureté de Deligne (1974) :

Théorème 2.6. *Pour toute variété projective lisse X de dimension d définie sur $\mathbf{F}_q$ et tout $i \leqslant 2d$, les valeurs propres de $\phi^\star_{q,X}$ sur $H^i(X, \mathbf{Q}_\ell)$ sont des nombres algébriques de norme complexe $q^{i/2}$. En particulier,*

$$\rho(\phi^\star_{q,X}, H^i(X, \mathbf{Q}_\ell)) = q^{i/2}.$$

Dans le cas où X n'est pas définie sur $\mathbf{F}_q$, Hrushovski (2004) introduit un dernier ingrédient : une norme sur les cycles.

Définition 2.7. Soit α un élément de $A^\star(X \times X^{(q)})$. On définit

$$|\alpha| = \min_{\sum_i a_i[Y_i]=\alpha} \sum_i |a_i| \deg(Y_i)$$

et

$$|H^i(\alpha)| = \min_{H^i(\gamma)=H^i(\alpha)} |\gamma|.$$

L'intérêt de cette norme est qu'elle est sous-multiplicative pour la composition (à renormalisation près) et qu'elle permet de borner le rayon spectral.

Proposition 2.8 (Hrushovski (2004, lemme 10.17.3))**.** *Pour toutes variétés projectives lisses X_1, X_2 et X_3 de dimension d et tout $\alpha_1 \in A^d(X_1 \times X_2)$ et $\alpha_2 \in A^d(X_2 \times X_2)$,*

$$|H^i(\alpha_2) \circ H^i(\alpha_1)| = O(|H^i(\alpha_1)| \cdot |H^i(\alpha_2)|),$$

où les bornes ne dépendent que du degré des variétés X_1, X_2 et X_3.

Quitte à ne considérer dorénavant que des variétés de degré borné, on définit

$$||H^i(\alpha)|| = N|H^i(\alpha)|,$$

avec N suffisamment grand pour que cette norme soit sous-multiplicative. Hrushovski (2004, proposition 11.11) démontre alors que, pour tout $\alpha \in A^d(X \times X)$,

$$\rho(H^i(\alpha)) \leqslant ||H^i(\alpha)||.$$

On trouve aussi une présentation alternative de ces résultats dans Shuddhodan et Varshavsky (2022, appendice B).

Nous pouvons maintenant conclure la preuve du théorème 2.1 dans le cas où $K = \overline{\mathbf{F}}_q$. Dans ce cas, il existe un entier r tel que X et C soient définis sur $\mathbf{F}_{q^r}$. On peut alors identifier $X^{(q^r)}$ à X et $C^{(q^r)}$ à C et on vérifie que

$$\left(\phi^\star_{q,X} \circ H^i([C])\right)^r = \phi^\star_{q^r,X} \circ h = h \circ \phi^\star_{q^r,X},$$

où $h = H^i([C^{(q^{r-1})}]) \circ \cdots \circ H^i([C]) = H^i([C^{(q^{r-1})}] \circ \cdots \circ [C])$, et donc

$$\rho(\phi_{q,X}^\star \circ H^i([C]))^r = \rho(\phi_{q^r,X}^\star \circ h) = \rho(h) q^{ir/2} \leqslant ||H^i([C])||^r q^{ir/2}.$$

ce qui conclut la preuve quand $K = \overline{\mathbf{F}}_q$. Le cas général s'en déduit par un argument de spécialisation.

Par linéarité, nous avons, en fait, prouvé une estimation de la trace pour tous les cycles de dimension d, qui sera utile par la suite. Nous rappelons que nous avons identifié $\overline{\mathbf{Q}}_\ell$ à un sous-corps de $\mathbf{C}$.

Proposition 2.9. *Pour toute variété projective lisse X sur K, tout $\alpha \in A^d(X \times X^{(q)})$ et tout $i \leqslant 2d$,*

$$\mathrm{Tr}(\phi_{q,X}^\star \circ H^i(\alpha)) = O(q^{i/2}) \,. \tag{4}$$

2.2. Un cas intermédiaire

Pour introduire un dernier ingrédient important de la preuve, nous allons considérer le cas où $X \subseteq \mathbf{P}_K^n$ est localement fermée et lisse. On suppose de plus que :

- ▷ la clôture $\overline{X}$ de X dans $\mathbf{P}_K^n$ est lisse ;
- ▷ le bord $\partial X = \overline{X} \setminus X$ est une union finie de diviseurs lisses $(X_i)_{i \in I}$ à croisements normaux.

Pour tout $J \subseteq I$, on note $X_J = \bigcap_{i \in J} X_i$. En suivant une construction de Pink (1992), on considère $\pi \colon \tilde{Y} \to \overline{X} \times \overline{X}^{(q)}$ l'éclatement de $\overline{X} \times \overline{X}^{(q)}$ le long de $\bigcup_i X_i \times X_i^{(q)}$. Soit $\tilde{\Gamma} \subseteq \tilde{Y}$ la clôture de $\pi^{-1}(\Gamma_{q,X})$. Pour tout $\tilde{\alpha} \in A^d(Y)$, Shuddhodan et Varshavsky (2022, lemme 2.3.1) démontrent l'égalité suivante inspirée par des travaux de Lafforgue (2002, proposition IV.6) :

$$\tilde{\alpha} \cdot [\tilde{\Gamma}] = \sum_{J \subseteq I} (-1)^{|J|} \alpha_J \cdot [\Gamma_{q,X_J}], \tag{5}$$

où les $\alpha_J \in A^{d-|J|}(X_J \times X_J^{(q)})$ sont construits explicitement, même si cela ne sera pas utile dans la suite de cet exposé (sauf dans le cas de $\alpha_\emptyset$). Pour tous $J \subseteq I$, soient $E_J = \pi^{-1}(X_J \times X_J^{(q)})$, $i_J \colon E_J \to Y$ l'inclusion et $\pi_J = \pi|_{E_J}$ la restriction. On a alors $\alpha_J = (\pi_J)_\star i_J^\star(\tilde{\alpha})$. En particulier, $\alpha_\emptyset = \pi_\star(\tilde{\alpha})$.

Soit $\tilde{C} \subseteq \tilde{Y}$ la clôture de $\pi^{-1}(C)$ dans $\tilde{Y}$. Pour conclure la preuve, il nous reste à relier $\#C \cap \Gamma_{q,X}(K)$ à $[\tilde{C}] \cdot [\tilde{\Gamma}]$. Pour cela, il nous faut faire une dernière hypothèse sur C.

Définition 2.10. Soient X une variété sur K et $C \subseteq X \times X^{(q)}$ et $Y \subseteq X$ des sous-variétés. On note $p_1 \colon C \to X$ et $p_2 \colon C \to X^{(q)}$ les projections.

- ▷ On dit que Y est C-invariante si $p_1(p_2^{-1}(Y(K))) \subseteq Y(K)$.
- ▷ On dit que Y est localement C-invariante si, pour tout $x \in Y(K)$, il existe un voisinage ouvert (de Zariski) $U \subseteq X$ de x tel que $Y \cap U$ soit $C|_U$-invariant, où $C|_U = C \cap (U \times U^{(q)})$.

Exemple 2.11. Si C est le graphe d'un morphisme $f\colon X \to X^{(q)}$, une sous-variété $Y \subseteq X$ est C-invariante si et seulement si $f(Y(K)) \subseteq Y(K)$.

Soit $\overline{C}$ la clôture de C dans $\overline{X} \times \overline{X}^{(q)}$. En supposant $\partial X = \overline{X} \setminus X$ localement $\overline{C}$-invariant et p_2 lisse, Shuddhodan et Varshavsky (2022, lemme 2.3.2) montrent que

$$\begin{aligned}
\#C \cap \Gamma_{q,X}(K) &= [\tilde{C}] \cdot [\Gamma] \\
&= \sum_{J \subseteq I} (-1)^{|J|} [\tilde{C}]_J \cdot [\Gamma_{q,X_J}] && \text{par (5)} \\
&= \sum_{J \subseteq I} (-1)^{|J|} \sum_{i=0}^{2(d-|J|)} (-1)^i \mathrm{Tr}(\phi^\star_{q,X_J} \circ H^i([\tilde{C}]_J)) && \text{par (2)} \\
&= \mathrm{Tr}(\phi^\star_{q,\overline{X}} \circ H^{2d}([\tilde{C}]_\emptyset)) + \sum_{i<2d} O(q^{i/2}) && \text{par (4)} \\
&= \deg(p_1) q^d + O(q^{d-1/2}),
\end{aligned}$$

où la dernière égalité suit du fait que $[\tilde{C}]_\emptyset = [\overline{C}]$ et de (3).

2.3. Le cas général

Le cas général s'obtient (presque) par réduction au cas précédent. Comme précédemment, on note $\overline{X}$ la clôture de Zariski de $X \subseteq \mathbf{P}^n_K$ et on note $\overline{C}$ la clôture de C dans $\overline{X} \times \overline{X}^{(q)}$. Quitte à se retreindre à des ouverts denses de X, on se ramène au cas où, *cf.* Shuddhodan et Varshavsky (2022, proposition 3.2.1) :

- ▷ la variété X est lisse ;
- ▷ le bord $\partial X = \overline{X} \setminus X$ est localement $\overline{C}$-invariant ;
- ▷ le morphisme $p_2\colon C \to X^{(q)}$ s'écrit comme la composée d'un homéomorphisme universel plat et d'un morphisme étale.

La principale difficulté est d'assurer la seconde condition, ce que le lemme suivant permet :

Lemme 2.12 (Shuddhodan et Varshavsky, 2022, proposition 1.1.7)**.** *Il existe un ouvert dense $U \subseteq X$ et un éclatement $\pi : \tilde{X} \to \overline{X}$ qui est un isomorphisme sur U tel que $\tilde{X} \setminus \pi^{-1}(U)$ est localement $\tilde{C}$-invariant, où $\tilde{C} \subseteq \tilde{X} \times \tilde{X}^{(q)}$ est la clôture de $\pi^{-1}(\overline{C}|_U)$.*

D'après un théorème de DE JONG (1996, théorème 4.1), soit alors Z une variété projective lisse de même dimension que X et $f\colon Z \to \overline{X}$ une altération — c'est-à-dire un morphisme dominant propre génériquement fini — telle que $f^{-1}(\partial X) \subseteq Z$ soit un diviseur à croisements normaux strict. Soient $\pi\colon \tilde{Y} \to Z \times Z^{(q)}$ l'éclatement de $Z \times Z^{(q)}$ le long de $f^{-1}(\partial X)$ et $\tilde{\Gamma} \subseteq \tilde{Y}$ la clôture de $\pi^{-1}(\Gamma_{q,Z})$. SHUDDHODAN et VARSHAVSKY (2022, §2.2.5) construisent alors $\tilde{\alpha} \in A^d(\tilde{Y})$ tel que :

$$\begin{aligned}\deg(f)\deg_{\mathrm{ins}}(p_2)\cdot \#C \cap \Gamma_{q,X}(K) &= \tilde{\alpha}\cdot[\tilde{\Gamma}]\\ &= \deg(f)\deg(p_1)q^d + O(q^{d-1/2}).\end{aligned}$$

3. Théorie asymptotique de l'automorphisme de Frobenius

Nous pouvons maintenant prouver le théorème de HRUSHOVSKI (2004) selon lequel la théorie asymptotique du Frobenius est celle d'un automorphisme générique. Commençons par en rappeler l'énoncé :

Théorème 3.1. *Pour tout énoncé ψ,*

$$\textit{pour tout } q \textit{ grand, } \psi \textit{ est vraie de } \overline{\mathbf{F}}_q \textit{ et } \phi_q \textit{ — ce qu'on note } (\overline{\mathbf{F}}_q, \phi_q) \models \psi,$$

si et seulement si

$$\psi \textit{ est conséquence des axiomes } \mathrm{ACFA} \textit{ — ce qu'on note } \mathrm{ACFA} \models \psi.$$

Commençons par reformuler ce résultat en terme d'ultraproduits, qui sont des objets naturels pour traiter ces questions asymptotiques.

Définition 3.2. Soit I un ensemble non vide. Un ultrafiltre non-principal sur I est un ensemble non vide $\mathfrak{U}$ de parties de I, tel que :

- ▷ L'ensemble $\mathfrak{U}$ ne contient aucun ensemble fini ;
- ▷ L'ensemble $\mathfrak{U}$ est clos par intersection finie ;
- ▷ Pour tout $X \subseteq I$, si $X \notin \mathfrak{U}$ alors $I \setminus X \in \mathfrak{U}$.

Par le lemme de Zorn, toute collection de parties de I dont les intersections finies sont infinies est incluse dans un ultrafiltre non-principal.

Définition 3.3. Soient $\mathfrak{U}$ un ultrafiltre non-principal sur I et (K_i, σ_i) des corps aux différences indexés par I. L'ultraproduit $\prod_{i\to\mathfrak{U}}(K_i, \sigma_i)$ est le corps aux différences

$$(\prod_i K_i/\mathfrak{m}, \sigma),$$

où $\mathfrak{m}$ est l'idéal (maximal) de $\prod_i K_i$ tel que, pour tout $x_i \in K_i$,

$$(x_i)_i \in \mathfrak{m} \text{ si et seulement si } \{i \in I : x_i = 0\} \in \mathfrak{U}$$

et σ est l'automorphisme induit par les σ_i :

$$\sigma((x_i)_i + \mathfrak{m}) = (\sigma_i(x_i))_i + \mathfrak{m}.$$

D'après un théorème de Łoš, pour tout énoncé ψ, on a

$$\prod_{i\to\mathfrak{U}}(K_i,\sigma_i)\models\psi \text{ si et seulement si } \{i\in I : (K_i,\sigma_i)\models\psi\}\in\mathfrak{U}. \tag{6}$$

On voit alors que si $\{q : (\overline{\mathbf{F}}_q,\phi_q)\models\psi\}$ est cofini, il est contenu dans tout ultrafiltre non-principal $\mathfrak{U}$ sur l'ensemble des puissances de nombres premiers et on a donc $\prod_{q\to\mathfrak{U}}(\overline{\mathbf{F}}_q,\phi_q)\models\psi$. Réciproquement, si $X=\{q : (\overline{\mathbf{F}}_q,\phi_q)\not\models\psi\}$ est infini, il existe un ultrafiltre non-principal $\mathfrak{U}$ tel que $X\in\mathfrak{U}$. On a alors $\prod_{q\to\mathfrak{U}}(\overline{\mathbf{F}}_q,\phi_q)\not\models\psi$. Le théorème 3.1 est donc équivalent à l'énoncé suivant :

Proposition 3.4. *Pour tout corps aux différences (K,σ), sont équivalents :*

1. *Le corps aux différences (K,σ) vérifie les axiomes de* ACFA*;*
2. *Il existe un ultrafiltre non-principal $\mathfrak{U}$ sur l'ensemble des puissances de nombres premiers tel que $\prod_{q\to\mathfrak{U}}(\overline{\mathbf{F}}_q,\phi_q)$ et (K,σ) vérifient les mêmes énoncés.*

Pour prouver que l'assertion 2 implique l'assertion 1, il faut vérifier que pour tout ultrafiltre non-principal $\mathfrak{U}$ sur les puissances de nombres premiers, l'ultraproduit $(K,\sigma)=\prod_{q\to\mathfrak{U}}(\overline{\mathbf{F}}_q,\phi_q)$ vérifie les axiomes de ACFA (voir la proposition 1.2). Tout d'abord, le corps K est algébriquement clos, et σ est surjectif, puisque cela s'exprime par des énoncés et que c'est le cas pour tous les corps aux différences $(\overline{\mathbf{F}}_q,\phi_q)$.

Soit enfin ψ_n l'énoncé qui exprime que, pour toutes variétés $X\subseteq\mathbf{A}^n$ et $C\subseteq X\times X^\sigma$ définies par au plus n équations de degré au plus n et telles que les projections de C vers X et X^σ sont dominantes, et tout ouvert de Zariski $U\subseteq C$ non vide défini par au plus n équations de degré au plus n, il existe un $x\in X$ tel que $(x,\sigma(x))\in U$. Par le corollaire 2.4, l'ensemble des q tel que $\{(\overline{\mathbf{F}}_q,\phi_q)\models\psi_n\}$ est cofini. Il est donc dans tous les ultrafiltres non-principaux $\mathfrak{U}$ et donc $\prod_{q\to\mathfrak{U}}(\overline{\mathbf{F}}_q,\phi_q)\models\psi_n$.

La réciproque est plus classique et déjà essentiellement présente dans les travaux de Ax (1968). Étant donné un corps aux différences (K,σ) existentiellement clos, d'après la proposition 1.3, il nous faut trouver un ultrafiltre non-principal $\mathfrak{U}$ sur les puissances de nombres premiers tel que les clôtures algébriques des corps premiers de K et $\prod_{q\to\mathfrak{U}}(\overline{\mathbf{F}}_q,\phi_q)$ soient isomorphes comme corps aux différences.

Si K est de caractéristique strictement positive p, c'est relativement immédiat. Une fois choisie une identification de la clôture algébrique du corps premier de K à $\overline{\mathbf{F}}_p$, pour tout entier $n\geqslant 1$, l'ensemble

$$\Sigma_n=\{m\geqslant 1 : \sigma|_{\mathbf{F}_{p^n}}=\phi_{p^m}\}$$

est infini. En effet, il existe un entier r tel que $\sigma|_{\mathbf{F}_{p^n}}=\phi_{p^r}=\phi_{p^{r+ms}}$, pour tout entier $s\geqslant 1$. De plus, si n divise m, on a $\Sigma_m\subseteq\Sigma_n$. Il existe donc un ultrafiltre non-principal $\mathfrak{U}$ sur l'ensemble des entiers strictement positifs qui contient tous les Σ_n. Par le théorème de Łoš (*cf.* (6)), l'application diagonale $\theta\colon(\overline{\mathbf{F}}_p,\sigma|_{\overline{\mathbf{F}}_p})\to\prod_{q\to\mathfrak{U}}(\overline{\mathbf{F}}_q,\phi_q)$ qui à tout

$x \in \overline{\mathbf{F}}_p$ associe la classe de $(x)_{n\geqslant 1}$ est un morphisme de corps aux différences, ce qui conclut la preuve de la proposition 3.4 dans le cas de caractéristique positive.

Si K est de caractéristique nulle, la preuve est plus complexe et repose sur un résultat classique de théorie algébrique des nombres : le théorème de densité de Chebotarev. Soit $F \leqslant K$ une extension galoisienne finie de $\mathbf{Q}$ contenue dans K. On note $\mathscr{O} \subseteq F$ la clôture intégrale de $\mathbf{Z}$ dans F. Pour tout nombre premier p et tout idéal premier $\mathfrak{p}$ de $\mathscr{O}$ au dessus de p — c'est-à-dire tel que $\mathfrak{p} \cap \mathbf{Z} = (p)$ — on définit $D_\mathfrak{p} = \{\sigma \in \mathrm{Gal}(F : \mathbf{Q}) \mid \sigma(\mathfrak{p}) = \mathfrak{p}\}$, son groupe de décomposition. Le corps $k_\mathfrak{p} = \mathscr{O}/\mathfrak{p}$ est alors une extension (galoisienne) finie de $\mathbf{F}_p$ et on a un morphisme naturel $D_\mathfrak{p} \to \mathrm{Gal}(k_\mathfrak{p}/\mathbf{F}_p)$. On dit que p est non ramifié (dans F), si c'est un isomorphisme. On note alors $\phi_\mathfrak{p} \in D_\mathfrak{p}$ la préimage de ϕ_p. Une conséquence du théorème de Chebotarev est que l'ensemble :

$$\Sigma_F = \{p \text{ non ramifié} : \sigma|_F = \phi_\mathfrak{p} \text{ pour un } \mathfrak{p} \text{ au dessus de } p\}$$

est infini — pour être précis, le théorème de densité de Chebotarev énonce que cet ensemble est de densité (naturelle ou analytique) égale à la taille de la classe de conjugaison de $\sigma|_F$ dans $\mathrm{Gal}(F/\mathbf{Q})$ divisée par le degré de l'extension.

Soit $\mathfrak{U}$ un ultrafiltre non-principal sur l'ensemble des nombres premiers qui contient Σ_F. Pour tout $p \in \Sigma_F$, on fixe un $\mathfrak{p}$ au dessus de p tel que $\sigma|_F = \phi_\mathfrak{p}$ — et si p n'est pas dans Σ_F, on fixe $\mathfrak{p}$ quelconque au dessus de p. L'application $(\mathscr{O}, \sigma|_\mathscr{O}) \to \prod_{p\to\mathfrak{U}}(\overline{\mathbf{F}}_p, \phi_p)$ qui à tout $x \in \mathscr{O}$ associe la classe de $(x + \mathfrak{p})_p$, est un morphisme injectif d'anneaux aux différences. En effet, $\Sigma_F \in \mathfrak{U}$ et pour tout $p \in \Sigma_F$,

$$\sigma(x) + \mathfrak{p} = \phi_\mathfrak{p}(x) + \mathfrak{p} = \phi_p(x + \mathfrak{p}).$$

Ce morphisme induit un morphisme de corps aux différences $(F, \sigma|_F) \to \prod_{p\to\mathfrak{U}}(\overline{\mathbf{F}}_p, \phi_p)$.

Si E est une extension galoisienne de F contenue dans K, on a $\Sigma_E \subseteq \Sigma_F$. Il existe donc un ultrafiltre non-principal $\mathfrak{U}$ qui contient tous les Σ_F. On a donc montré ci-dessus que pour ce choix de $\mathfrak{U}$, pour toute extension galoisienne finie F de $\mathbf{Q}$ contenue dans K, il existe un morphisme de corps aux différences $\theta_F : (F, \sigma|_F) \to \prod_{p\to\mathfrak{U}}(\overline{\mathbf{F}}_p, \phi_p)$.

Soit K_0 la clôture algébrique de $\mathbf{Q}$ dans K. Par compacité du groupe de Galois absolu de $\mathbf{Q}$ pour la topologie profinie (ou par la compacité de la logique du premier ordre), on peut recoller ces morphismes pour produire un morphisme de corps aux différences $\theta : (K_0, \sigma|_{K_0}) \to \prod_{p\to\mathfrak{U}}(\overline{\mathbf{F}}_p, \phi_p)$. Par la proposition 1.3, les corps aux différences (K, σ) et $\prod_{p\to\mathfrak{U}}(\overline{\mathbf{F}}_p, \phi_p)$ vérifient les mêmes énoncés, ce qui conclut la preuve de la proposition 3.4 dans le cas de caractéristique zéro.

On remarque que dans la preuve ci-dessus, si K est de caractéristique strictement positive p, on n'a besoin que d'un ultraproduit des $(\overline{\mathbf{F}}_p, \phi_{p^n})$; ce qu'on pourrait de toute manière déduire du théorème (6) de Łoš. Si K est de caractéristique

zéro, un ultraproduit des $(\overline{\mathbf{F}}_p, \phi_p)$ suffit, ce qui est plus inattendu. Si, pour tout p premier, ACFA_p est l'ensemble d'énoncés $\mathrm{ACFA} \cup \{p = 0\}$ et ACFA_0 l'ensemble $\mathrm{ACFA} \cup \{p \neq 0 : p \text{ premier}\}$, nous avons donc prouvé le résultat suivant qui précise le théorème 3.1 :

Proposition 3.5. *Pour tout énoncé ψ,*

$$(\overline{\mathbf{F}}_p, \phi_p) \models \psi \text{ pour tout } p \text{ premier grand si et seulement si } \mathrm{ACFA}_0 \models \psi$$

et

$$(\overline{\mathbf{F}}_p, \phi_{p^n}) \models \psi \text{ pour tout } n \text{ grand si et seulement si } \mathrm{ACFA}_p \models \psi.$$

4. Quelques applications

Pour finir, exposons d'autres conséquences du théorème 2.1 en dynamique algébrique et en géométrie algébrique aux différences. Sans prétendre être exhaustif, cet énoncé a aussi des conséquences en géométrie algébrique (Esnault et Mehta, 2010 ; Esnault, Srinivas et Bost, 2016) et en théorie des groupes (Borisov et Sapir, 2005), mais elles s'éloignent plus du sujet principal de cet exposé et nous ne les aborderons pas ici.

4.1. Dynamique algébrique

Les premiers résultats dont nous discuterons concernent les points périodiques des endomorphismes de variétés. Soient X une variété sur un corps K algébriquement clos et $f\colon X \dashrightarrow X$ une application rationnelle. La f-orbite d'un point x est l'ensemble des $f^i(x)$, où $i \geqslant 0$ est un entier, quand ils sont définis. Un point x de $X(K)$ est dit périodique s'il existe un entier $i \geqslant 0$ tel que $f^i(x)$ est défini et égal à x.

Il découle du théorème 2.1 que, sur $\overline{\mathbf{F}}_p$, l'ensemble des points périodiques d'une application rationnelle dominante est dense :

Proposition 4.1. *Soient X une variété sur $\overline{\mathbf{F}}_p$ et $f\colon X \dashrightarrow X$ une application rationnelle dominante. L'ensemble des points périodiques de f dans $X(\overline{\mathbf{F}}_p)$ est alors Zariski dense dans X.*

Démonstration. Soit q une puissance de p telle que X et f soient définies sur $\mathbf{F}_q$. Par le corollaire 2.4, l'ensemble $\bigcup_{n\geqslant 1} \Gamma_f \cap \Gamma_{q^n,X}(\overline{\mathbf{F}}_q)$ est Zariski dense dans Γ_f. En particulier, sa (première) projection sur X est Zariski dense dans X. Pour tout élément x dans cette projection, $f(x)$ est défini et il existe un entier $n \geqslant 1$ tel que $f(x) = \phi^n_{q,X}(x)$. Si m est tel que $x \in \mathbf{F}_{q^m}$, on a

$$f^m(x) = \phi_{q^{nm},X}(x) = x$$

et donc x est un point périodique de f. □

Par un argument de spécialisation, Fakhruddin (2003, théorème 5.1) en déduit un résultat similaire sur un corps quelconque :

Théorème 4.2. *Soient X une variété projective sur un corps algébriquement clos K, $f\colon X \to X$ un morphisme dominant. On suppose qu'il existe un fibré en droites L sur X tel que $f^\star L \otimes L^{-1}$ soit ample. Alors, l'ensemble des points périodiques de f dans $X(K)$ est Zariski dense dans X.*

Bien que cela puisse paraître au premier abord un peu contre-intuitif, Amerik (2011) déduit aussi de la proposition 4.1 un résultat de densité des points de f-orbite infinie :

Théorème 4.3. *Soient X une variété sur $\overline{\mathbf{Q}}$ et $f\colon X \dashrightarrow X$ une application rationnelle dominante qui n'est pas d'ordre fini. L'ensemble des points de f-orbite infinie dans $X(\overline{\mathbf{Q}})$ est alors Zariski dense dans X.*

L'idée de la preuve est la suivante. Soit K un corps de nombres sur lequel X et f sont définies. On applique alors la proposition 4.1 à la réduction $\overline{f}$ de f modulo un idéal premier bien choisi $\mathfrak{p}$ de l'anneau des entiers $\mathscr{O}$ de K. Quitte à remplacer K par une extension finie, on trouve donc un point périodique $x \in \mathscr{O}/\mathfrak{p}$ de $\overline{f}$ — et on peut même supposer que $\overline{f}$ est étale en x.

On considère alors l'ensemble U des éléments de $X(K_\mathfrak{p})$ qui se réduisent à x module $\mathfrak{p}$, où $K_\mathfrak{p}$ est le complété de K pour la topologie $\mathfrak{p}$-adique. Par construction, il est invariant par une puissance de f. Par une combinaison de résultats de Amerik, Bogomolov et Rovinsky (2011) et Bell, Ghioca et Tucker (2010), on démontre alors qu'il existe une borne uniforme sur la taille des f-orbites finies de points de U — et donc qu'elles sont contenues dans une sous-variété analytique propre. Le théorème découle alors du fait que les points de $X(\overline{\mathbf{Q}})$ sont denses dans U.

4.2. Géométrie algébrique aux différences

Soit (K, σ) un corps aux différences. Un polynôme aux différences $P(x_1, \dots x_n)$ à coefficients dans K est une fonction de la forme $Q((\sigma^j(x_i))_{i\leqslant n, j\geqslant 0})$ où $Q \in K[X_{i,j} : i \leqslant n, j \geqslant 0]$. L'ordre de x_i dans P est le m maximal tel que $X_{i,m}$ apparaît dans un monôme de Q avec un coefficient non nul — de manière équivalente, le m maximal tel que $\sigma^m(x_i)$ apparaît dans P. On choisit la convention que, si aucun des $\sigma^j(x_i)$ n'apparaît dans P, l'ordre de x_i est $-\infty$.

Étant donné n polynômes aux différences en n variables $P_j(x_1, \dots, x_n)$, où $1 \leqslant j \leqslant n$, à coefficients dans K, on souhaite étudier la géométrie de la « variété aux différences affine » X définie comme le lieu d'annulation des P_j. On définit la dimension totale $\mathrm{tdim}(X)$ de X comme étant le supremum du degré de transcendance de $L(\sigma^k(a_i) : i \leqslant n, k \geqslant 0)$ sur L, pour toute extension $(K, \sigma) \leqslant (L, \sigma)$ et $a \in L^n$ tel

que $P_j(a) = 0$, pour tout $j \leqslant n$. C'est un équivalent naturel de la dimension de Krull dans le cadre des variétés aux différences.

Hrushovski (2004, théorème 14.2) déduit du théorème 2.1 la borne suivante sur la dimension totale de X.

Théorème 4.4. *Soit $h_{i,j}$ l'ordre de x_i dans P_j. On a*

$$\operatorname{tdim}(X) \leqslant \max_{\theta \in \mathfrak{S}_n} \sum_{i \leqslant n} h_{i,\theta(i)}.$$

C'est l'équivalent pour l'algèbre aux différences d'une conjecture de Jacobi (non résolue) en algèbre différentielle.

La preuve consiste à prouver, d'abord, un résultat similaire pour les équations polynomiales :

Proposition 4.5. *Soient $P_j(x_1, \dots, x_n)$, où $1 \leqslant j \leqslant n$, des polynômes sur un corps algébriquement clos K. Soit X le lieu des zéros des P_j, et X_0 l'union de ses composantes irréductibles de dimension 0. Alors*

$$|X_0(K)| \leqslant \sum_{\theta \in \mathfrak{S}_n} \prod_{i \leqslant n} d_{i,\theta(i)}, \tag{7}$$

où $d_{i,j}$ est le degré de P_j en x_i.

Cette version du théorème de Bézout se démontre aisément en calculant un produit d'intersection.

On relie ensuite cette borne à la borne recherchée par un argument de spécialisation — cette fois-ci en algèbre aux différences. Soit $(D, \sigma) \leqslant (K, \sigma)$ un anneau aux différences finiment engendré, bien choisi, qui contient les coefficients des P_j. Soient $f\colon (D, \sigma) \to (\overline{\mathbf{F}}_q, \phi_q)$ un morphisme d'anneaux aux différences et $Q_j = f(P_j)$. Ce polynôme aux différences s'identifie, puisque $\phi_q(x) = x^q$, à un polynôme de degré $d_{i,j}(q)$ en x_i, avec $d_{i,j}(q)$ de l'ordre de $mq^{h_{i,j}} + O(q^{h_{i,j}-1})$, où $m \in \mathbf{Z}_{>0}$, quand q est grand. On a donc

$$\lim_{q \to \infty} \log_q d_{i,j}(q) = h_{i,j}.$$

Par ailleurs, si X_q dénote le lieu des zéros des Q_j, on peut montrer que X_q est de dimension zéro et on peut déduire du théorème 2.1, que pour q suffisamment grand,

$$|X_q(\overline{\mathbf{F}}_q)| = cq^{\operatorname{tdim}(X)} + O(q^{\operatorname{tdim}(X)-1/2}),$$

où $c \in \mathbf{Q}_{>0}$. Il s'ensuit que

$$\begin{aligned} \operatorname{tdim}(X) &= \lim_{q \to \infty} \log_q |X_q(\overline{\mathbf{F}}_q)| \\ &\leqslant \lim_{q \to \infty} \log_q \sum_{\theta \in \mathfrak{S}_n} \prod_{i \leqslant n} d_{i,\theta(i)}(q) \qquad \text{par (7)} \\ &= \max_{\theta \in \mathfrak{S}_n} \sum_{i \leqslant n} h_{i,\theta(i)}. \end{aligned}$$

Références

Amerik, E. (2011). « Existence of non-preperiodic algebraic points for a rational self-map of infinite order », *Math. Res. Lett.* **18** (2), p. 251-256.

Amerik, E., Bogomolov, F. A. et Rovinsky, M. (2011). « Remarks on endomorphisms and rational points », *Compos. Math.* **147** (6), p. 1819-1842.

Ax, J. (1968). « The elementary theory of finite fields », *Ann. of Math.* (2) **88**, p. 239-271.

Bell, J. P., Ghioca, D. et Tucker, T. J. (2010). « The dynamical Mordell-Lang problem for étale maps », *Amer. J. Math.* **132** (6), p. 1655-1675.

Borisov, A. et Sapir, M. (2005). « Polynomial maps over finite fields and residual finiteness of mapping tori of group endomorphisms », *Invent. Math.* **160** (2), p. 341-356.

Chatzidakis, Z. et Hrushovski, E. (1999). « Model theory of difference fields », *Trans. Amer. Math. Soc.* **351** (8), p. 2997-3071.

Chatzidakis, Z., Hrushovski, E. et Peterzil, Y. (2002). « Model theory of difference fields. II. Periodic ideals and the trichotomy in all characteristics », *Proc. London Math. Soc.* (3) **85** (2), p. 257-311.

de Jong, A. J. (1996). « Smoothness, semi-stability and alterations », *Inst. Hautes Études Sci. Publ. Math.* (83), p. 51-93.

Deligne, P. (1974). « La conjecture de Weil. I », *Inst. Hautes Études Sci. Publ. Math.* (43), p. 273-307.

Esnault, H. et Mehta, V. (2010). « Simply connected projective manifolds in characteristic $p > 0$ have no nontrivial stratified bundles », *Invent. Math.* **181** (3), p. 449-465.

Esnault, H., Srinivas, V. et Bost, J.-B. (2016). « Simply connected varieties in characteristic $p > 0$ », *Compos. Math.* **152** (2), p. 255-287.

Fakhruddin, N. (2003). « Questions on self maps of algebraic varieties », *J. Ramanujan Math. Soc.* **18** (2), p. 109-122.

Fried, M. D. et Jarden, M. (2008). *Field arithmetic*. Third edition. T. 11. Ergeb. Math. Grenzgeb. (3). Springer-Verlag, Berlin, p. xxiv+792.

Grothendieck, A. (1966). « Éléments de géométrie algébrique. IV. Étude locale des schémas et des morphismes de schémas. Troisième partie », *Inst. Hautes Études Sci. Publ. Math.* (28), p. 255.

——— (1977). « La classe de cohomologie associée à un cycle », in : *SGA4*$\frac{1}{2}$. Lecture Notes in Math. 569. Rédigé par Pierre Deligne. Springer, Berlin, p. 129-153.

Hrushovski, E. (2004). « The elementary theory of the Frobenius automorphism ». arXiv :math/0406514.

Lafforgue, L. (2002). « Chtoucas de Drinfeld et correspondance de Langlands », *Invent. Math.* **147** (1), p. 1-241.

Lang, S. et Weil, A. (1954). « Number of points of varieties in finite fields », *Amer. J. Math.* **76**, p. 819-827.

Macintyre, A. (1997). « Generic automorphisms of fields », *Ann. Pure Appl. Logic* **88** (2-3). Joint AILA-KGS Model Theory Meeting (Florence, 1995), p. 165-180.

Medvedev, A. et Scanlon, T. (2014). « Invariant varieties for polynomial dynamical systems », *Ann. of Math.* (2) **179** (1), p. 81-177.

Pink, R. (1992). « On the calculation of local terms in the Lefschetz-Verdier trace formula and its application to a conjecture of Deligne », *Ann. of Math.* (2) **135** (3), p. 483-525.

Shuddhodan, K. V. (2022). « The (non-uniform) Hrushovski-Lang-Weil estimates », *Adv. Math.* **410**, part B, Paper No. 108753, 59 pp.

Shuddhodan, K. V. et Varshavsky, Y. (2022). « The Hrushovski–Lang–Weil estimates », *Algebr. Geom.* **9** (6), p. 651-687.

van den Dries, L. et Schmidt, K. (1984). « Bounds in the theory of polynomial rings over fields. A nonstandard approach », *Invent. Math.* **76** (1), p. 77-91.

Silvain Rideau-Kikuchi

CNRS, École normale supérieure
45 rue d'Ulm, 75005 Paris

E-mail : silvain.rideau@ens.fr

Séminaire BOURBAKI Janvier 2023
74e année, 2022–2023, no 1201, p. 141 à 197
doi : 10.24033/ast.1210

LA CONJECTURE DU FACTEUR DIRECT
[*d'après* André et Bhatt]

par **Gabriel Dospinescu**

La conjecture du facteur direct (théorème 1.1 ci-dessous) est un énoncé d'algèbre commutative presque aussi élémentaire que le Théorème de Fermat et qui est resté ouvert pendant près de 50 ans : énoncée en 1969 (*cf.* Hochster (1973) pour la version publiée), elle a été démontrée par André en 2016 (*cf.* André (2018b) pour la version publiée). Cet énoncé fait partie d'un faisceau de conjectures, les « conjectures homologiques » dont la liste et les relations donnent un peu le tournis [(1)]. En particulier, Hochster (1975, 1983) avait prouvé que l'existence de A-algèbres (ou même seulement de A-modules) de Cohen–Macaulay (voir § 1.5), pour tout anneau local noethérien A, impliquait la plupart de ces conjectures, par exemple celle du facteur direct. Hochster et Huneke (1992) avaient aussi montré [(2)] l'existence de telles A-algèbres dans le cas d'égale caractéristique, *i.e.* quand A contient un corps.

Le but de cet exposé est d'expliquer les techniques introduites par André dans ses trois articles monumentaux (André, 2018a,b, 2020), en particulier comment les espaces perfectoïdes introduits par Scholze (2012) permettent de construire des algèbres de Cohen–Macaulay pour les anneaux locaux noethériens d'inégale caractéristique et donc de prouver la conjecture du facteur direct. [(3)] Les travaux récents et spectaculaires de Bhatt (2020) poussent encore plus loin les techniques perfectoïdes (via la théorie prismatique de Bhatt et Scholze (2022) et la correspondance de Riemann–Hilbert p-adique de Bhatt et Lurie (2023)) et établissent (théorème 1.16 ci-dessous) un analogue en inégale caractéristique d'un célèbre théorème de Hochster et Huneke (1992) (en caractéristique positive), qui implique tous les résultats exposés ici, et bien plus.

[(1)] Voir le théorème 1.12 pour un condensé loin d'être exhaustif, ainsi que Hochster (1983, 2007) et Roberts (1992).

[(2)] L'existence de A-modules de Cohen–Macaulay pour A d'égale caractéristique avait été établie bien avant, *cf.* Hochster, 1975.

[(3)] Que les espaces perfectoïdes soient un outil indispensable en théorie de Hodge p-adique ne faisait guère de doute après leurs premières applications spectaculaires (Scholze, 2012, 2013, 2015). Qu'ils puissent aussi résoudre les conjectures homologiques était moins clair : la plupart de ces problèmes concernent des anneaux noethériens, propriété quasiment jamais satisfaite par les anneaux perfectoïdes.

La preuve du résultat de Bhatt est un véritable tour de force, et tous les détails ne sont pas encore (à notre connaissance) disponibles, nous avons donc décidé de nous concentrer sur les articles d'André dans cet exposé, en fournissant des preuves complètes (autant que faire se peut) des résultats principaux des trois articles mentionnés ci-dessus, tout en utilisant des idées de Bhatt (2018) pour simplifier certains arguments. [4] Le chemin que nous avons choisi pour arriver à la preuve de la conjecture du facteur direct n'est pas le plus court (la géodésique se trouve dans l'article Bhatt, 2018) ; il nous fera visiter des résultats plus puissants que la conjecture elle-même. Pour des applications de ces idées et techniques à la théorie (naissante) des singularités en caractéristique mixte nous renvoyons aux travaux de Ma et Schwede (2018, 2021) et Ma, Schwede et al. (2022) (entre autres) et pour de vastes généralisations des résultats présentés ici le lecteur pourra consulter le livre (de longueur presque infinie...) de Gabber et Ramero (2018).

Convention : Tous les anneaux sont supposés commutatifs et unitaires, et les morphismes d'anneaux sont unitaires. Si I est un idéal d'un anneau A, on dit que A est *I-complet* si A est séparé complet pour la topologie I-adique. Pour $a \in A$ on dit que A est *a-complet* si A est aA-complet, et on note $A/a := A/aA$. On note $A[I]$ l'idéal de A des éléments annulés par tous les éléments de I. Si $a_1, \dots, a_d$ sont des éléments de A, on note $(a_1, \dots, a_d) = \sum_{i=1}^{d} a_i A$ l'idéal de A qu'ils engendrent.

Remerciements. — Mes plus vifs remerciements vont à Yves André, Bharghav Bhatt, Nicolas Bourbaki, Kęstutis Česnavičius, Pierre Colmez, Luc Illusie, Wiesława Nizioł et Olivier Taïbi. Leurs commentaires et leurs suggestions ont permis au béotien du sujet d'éviter bon nombre de pièges et ont grandement amélioré le contenu et la lisibilité de ce rapport. Je remercie tout particulièrement Yves André pour sa disponibilité, son enthousiasme et ses multiples remarques.

1. Les multiples visages de la conjecture

Le but de cette section est d'énoncer les principaux résultats d'algèbre commutative « classique » [5] démontrés par André et Bhatt, et d'expliquer les liens qu'ils entretiennent.

[4] On trouvera deux preuves de l'existence de A-algèbres de Cohen–Macaulay dans ce rapport. Elles partagent un ingrédient fondamental, le lemme de platitude d'André ; l'une utilise le théorème de presque pureté de Faltings (1988, 2002), raffiné et étendu par Scholze (2012) et par Kedlaya et Liu (2015), l'autre n'en fait pas usage.

[5] Les perfectoïdes n'apparaissent donc pas dans cette section. Le lecteur trouvera dans André (2018c) un survol des preuves fait par le maître, et qui semble impossible à dépasser en terme de présentation.

1.1. Les anneaux réguliers, sources de problèmes

Un anneau local noethérien A de dimension d est dit *régulier* si son unique idéal maximal $\mathfrak{m}$ est engendré par d éléments (c'est le nombre minimal possible de générateurs). Des exemples typiques de tels anneaux sont les anneaux locaux des variétés algébriques lisses sur un corps (ou sur un anneau de valuation discrète), ainsi que leurs complétés, par exemple $K[[X_1, \dots, X_n]]$ (K étant un corps ou un anneau de valuation discrète), mais aussi $\mathbf{Z}_p[[X, Y, Z]]/(p - X^5 - Y^7 - Z^9)$, etc.

En dépit de leur définition très simple, les anneaux locaux réguliers sont une source inépuisable de problèmes délicats, et il n'est pas facile d'établir même des propriétés très basiques comme la stabilité de la régularité par localisation en un idéal premier (cela se déduit de l'interprétation homologique de la régularité fournie par le théorème de Serre), ce qui permet de globaliser [(6)] la notion de régularité. Il n'est pas difficile de montrer que tout anneau local régulier est intègre, mais il faut se fatiguer un peu pour montrer qu'il est normal [(7)], et bien plus pour montrer qu'il est même factoriel (théorème d'Auslander–Buchsbaum).

Si $(A, \mathfrak{m})$ est un anneau local régulier, son complété $\hat{A} = \varprojlim_n A/\mathfrak{m}^n$ est un anneau local régulier complet (pour la topologie $\mathfrak{m}$-adique), et le théorème de structure de Cohen montre que $\hat{A}$ a l'une des formes suivantes, à isomorphisme près :

• ou bien $V[[X_1, \dots, X_n]]$ avec V un corps ou un anneau de valuation discrète complet et non ramifié (*i.e.* l'idéal maximal de V est engendré par un nombre premier p). On dira alors que $\hat{A}$ est *non ramifié*;

• ou bien $V[[X_1, \dots, X_n]]/(p - f)$ pour un anneau de valuation discrète complet et non ramifié V, de caractéristique résiduelle p, et un élément f dans $(p, X_1, \dots, X_n)^2$ mais pas dans $pV[[X_1, \dots, X_n]]$ (on dira que $\hat{A}$ est *ramifié* dans ce cas).

1.2. Énoncé de la conjecture du facteur direct

La conjecture du facteur direct de Hochster (1973), à laquelle cet exposé est consacré, est l'énoncé suivant, à l'air parfaitement innocent :

Théorème 1.1. *Toute extension finie d'un anneau régulier est scindée.*

Précisons l'énoncé : une *extension d'anneaux* est un morphisme injectif d'anneaux $f\colon A \to B$, elle est dite *finie* si f fait de B un A-module de type fini, et *scindée* si A est un facteur direct du A-module B, autrement dit s'il existe une application A-linéaire [(8)] $r\colon B \to A$ telle que $r(f(a)) = a$ pour tout $a \in A$.

[(6)] Un anneau noethérien est régulier si ses localisés en des idéaux premiers quelconques sont des anneaux locaux réguliers.

[(7)] Autrement dit intégralement clos dans son corps des fractions.

[(8)] On ne demande pas à r d'être un morphisme d'anneaux.

Remarque 1.2. Bhatt (2018) revisite et simplifie la preuve d'André (2018b), ce qui lui permet d'établir la version dérivée suivante du théorème 1.1, conjecturée par de Jong : si A est un anneau régulier et si $f\colon X \to \mathrm{Spec}(A)$ est un morphisme propre et surjectif, alors le morphisme $A \to \mathrm{R}\Gamma(X, \mathscr{O}_X)$ est scindé dans la catégorie dérivée $\mathrm{D}(A)$. Si A est une $\mathbf{Q}$-algèbre cela se déduit des travaux de Kovács (2000), le cas $\mathrm{car}(A) = p$ avait été traité par Bhatt (2012).

Hochster (1973) a démontré le théorème 1.1 pour les anneaux réguliers contenant un corps, et a réduit, par un argument très indirect (*cf.* théorème 6.1 de Hochster, 1983) le cas général à celui d'un anneau local régulier complet, non ramifié, de corps résiduel algébriquement clos. Une avancée spectaculaire est due à Heitmann (2002) : il a démontré la conjecture quand $\dim A = 3$ (le cas $\dim A \leqslant 2$ est une conséquence de la formule d'Auslander–Buchsbaum).

Le lien entre les techniques perfectoïdes (plus précisément les presque mathématiques et le théorème de presque pureté de Faltings (1988, 2002)) et la conjecture du facteur direct semble avoir été remarqué depuis un certain temps [(9)], mais ce n'est qu'en 2014 que Bhatt (2014a) a obtenu le premier résultat un peu général via ces techniques, en traitant le cas où $B[\frac{1}{p}]$ est étale sur $A[\frac{1}{p}]$ (et même sous des hypothèses plus faibles). C'est ce cercle d'idées qui mènera à la preuve de la conjecture, mais il a fallu attendre les travaux d'André (2018a,b) pour traiter le cas général.

Remarque 1.3. a) Une extension finie $f\colon A \to B$ d'anneaux noethériens est scindée si et seulement si l'extension induite $A_{\mathfrak{m}} \to B_{\mathfrak{m}}$ l'est pour tout idéal maximal $\mathfrak{m}$ de A : l'existence d'un scindage équivaut à la surjectivité de l'application [(10)]

$$\mathrm{ev}_1\colon \mathrm{Hom}_{\mathrm{Mod}_A}(B, A) \to A,\ r \mapsto r(1),$$

qui peut se tester en localisant en tout idéal maximal $\mathfrak{m}$ de A, or

$$\mathrm{Hom}_{\mathrm{Mod}_A}(B, A)_{\mathfrak{m}} \simeq \mathrm{Hom}_{\mathrm{Mod}_{A_{\mathfrak{m}}}}(B_{\mathfrak{m}}, A_{\mathfrak{m}})$$

puisque B est un A-module de présentation finie. De même, f est scindée si et seulement si l'extension $C \to C \otimes_A B$ l'est pour une extension fidèlement plate C de A, car on peut tester la surjectivité de ev_1 après changement de base à C.

[(9)] Par exemple, voici ce que m'écrit Wiesława Nizioł : « in 2001 Lorenzo Ramero visited me in Utah and gave a talk at the Number Theory seminar on his work with Gabber and their attempt to prove the almost purity conjecture. Paul Roberts was in the audience and was really surprised by the similarity of almost math techniques with the recent proof by Heitmann of the direct summand conjecture in dim 3. Heitmann worked in the almost setting and then at some point was able to descend to the usual setting (via some finiteness properties?). Roberts got all excited about this and we had a seminar running for a semester on almost math and commutative algebra. It did not get anywhere because, of course, we did not have the almost purity in general at that time. »

[(10)] On note Mod_A la catégorie des A-modules.

b) La preuve du théorème 1.1 se ramène au cas d'une extension finie $f\colon A \to B$ avec A local régulier complet et B intègre (donc local et complet). En effet, par a) on peut supposer que A est local, puis complet, en utilisant l'extension fidèlement plate $A \to \widehat{A}$. Si $\wp$ est un idéal premier de B tel que $\dim(B/\wp) = \dim B$, alors $\dim A/(\wp \cap A) = \dim A$, puis $\wp \cap A = \{0\}$ (car A est local et intègre), et tout scindage de l'extension finie $A \to B/\wp$ en fournit un pour $A \to B$.

c) Si A est une $\mathbf{Q}$-algèbre intègre et normale, alors toute extension finie $f\colon A \to B$ est scindée. En effet, comme dans b) on peut supposer que B est intègre. Si K et L sont les corps des fractions de A et de B, par normalité de A la trace $\mathrm{Tr}_{L/K}\colon L \to K$ envoie B dans A, et $\frac{1}{[L:K]}\mathrm{Tr}_{L/K}\colon B \to A$ fournit un scindage. Donc pour les $\mathbf{Q}$-algèbres le théorème 1.1 est trivial, et pas optimal.

d) Si $\dim A \leqslant 2$ le théorème 1.1 se déduit de la formule d'Auslander–Buchsbaum. Si A est de caractéristique positive on dispose de toute une variété de preuves pas (trop) difficiles du théorème 1.1, voir l'exemple 1.3 de Bhatt (2012) pour une preuve cohomologique, et le paragraphe 6.2 de Hochster (1983) pour une preuve courte.

1.3. Fragmenteurs

Appelons *fragmenteur* (*splinter* en anglais) un anneau intègre A tel que toute extension finie de A soit scindée. Le théorème 1.1 affirme que les anneaux réguliers sont fragmenteurs. Tout fragmenteur est normal [(11)], et la réciproque est vraie pour les $\mathbf{Q}$-algèbres intègres (remarque 1.3). La situation est nettement plus compliquée en caractéristique positive ou mixte. Hochster et Huneke (1992, 1995) ont montré que les fragmenteurs noethériens de caractéristique positive et localement excellents [(12)] sont des anneaux de Cohen–Macaulay, et Bhatt (2020) vient de montrer, dans son travail spectaculaire, que cela reste vrai en caractéristique mixte (toujours sous des hypothèses d'excellence).

Une source importante de fragmenteurs est la théorie des représentations des groupes (linéairement) réductifs : si un tel groupe G agit sur une k-algèbre R qui est un anneau régulier (k étant un corps), alors l'anneau des invariants R^G est un fragmenteur (car l'inclusion $R^G \to R$ est scindée, via l'opérateur de Reynolds, R est un fragmenteur, et un facteur direct d'un fragmenteur en est encore un).

Voici un exemple (dû à Hochster, 1973) d'anneau normal, de Cohen–Macaulay (même intersection complète), et non fragmenteur. Soient k un corps de caractéristique 2 et $R = k[X,Y,Z]/(X^3+Y^3+Z^2) = k[x,y,z]$. Le morphisme $R \to k[U,V]$ envoyant x,y,z sur U^2, V^2, U^3+V^3 est fini, injectif et non scindé : s'il était scindé on aurait $R \cap Ik[U,V] = I$ pour tout idéal I de R, or $z \notin (x,y)$ et $U^3+V^3 \in (U^2,V^2)$.

[(11)] Si $x \in \mathrm{Frac}(A)$ est entier sur A, alors $A \to A[x]$ est une extension finie et elle n'est pas scindée si $x \notin A$, puisque toute rétraction A-linéaire $r\colon A[x] \to A$ doit envoyer x sur lui-même (si $x = \frac{a}{b}$ avec $a, b \in A$ et $b \neq 0$, alors $a = r(a) = r(bx) = br(x)$).

[(12)] *i.e.* dont les localisés en tout idéal maximal sont excellents.

Un schéma S est dit *fragmenteur* si pour tout morphisme fini surjectif $f\colon X \to S$ le morphisme $\mathscr{O}_S \to f_*\mathscr{O}_X$ est scindé dans la catégorie $\mathrm{Coh}(S)$ des faisceaux cohérents sur S. On dit que S est un *D-fragmenteur* si pour tout morphisme propre surjectif $f\colon X \to S$ le morphisme $\mathscr{O}_S \to \mathrm{R}f_*\mathscr{O}_X$ est scindé dans $D(\mathrm{Coh}(S))$. Bhatt (2012) a montré qu'un $\mathbf{F}_p$-schéma noethérien est fragmenteur si et seulement s'il est D-fragmenteur. Pour comparer, pour un $\mathbf{Q}$-schéma le caractère fragmenteur est (plus ou moins, *i.e.* sous des hypothèses faibles) équivalent à la normalité, alors que le caractère D-fragmenteur est (plus ou moins) équivalent, par un théorème de Kovács (2000), au fait que les singularités de S sont au pire rationnelles, *cf.* exemples 1.1 et 1.2 de Bhatt (2012).

Voir André et Fiorot (2022) et Bhatt (2012) pour plus de détails et d'exemples concernant les fragmenteurs.

1.4. Scindage et pureté

Un morphisme d'anneaux $f\colon A \to B$ est dit *pur* s'il est *universellement injectif*, *i.e.* si le morphisme induit $C \to B \otimes_A C$ reste injectif pour toute A-algèbre C, auquel cas il reste injectif pour tout A-module C. Toute extension scindée est clairement pure. Voici deux incarnations importantes de la notion de pureté :

• d'un point de vue catégorique, un morphisme $f\colon A \to B$ est pur si et seulement si le foncteur $(-) \otimes_A B\colon \mathrm{Mod}_A \to \mathrm{Mod}_B$ est fidèle. Rappelons que f est dit *plat* (resp. *fidèlement plat*) si le foncteur $(-) \otimes_A B$ est exact (resp. exact et fidèle). Ainsi tout morphisme fidèlement plat est pur ;

• Olivier (1973) a montré qu'un morphisme $f\colon A \to B$ est pur si et seulement si la théorie de la descente fonctionne bien [13] pour les B-modules. Comme il a été remarqué dans André et Fiorot (2022), cela permet de voir les morphismes purs comme les recouvrements pour la topologie canonique [14] sur la catégorie des schémas affines, ce qui en fournit une interprétation géométrique.

La conjecture du facteur direct se réincarne (*cf.* théorème 1.7) en un énoncé de pureté grâce au résultat suivant :

Proposition 1.4. *Toute extension finie et pure d'anneaux noethériens est scindée.*

Démonstration. On peut supposer que A est local et complet (remarque 1.3). Si E est une enveloppe injective du corps résiduel de A, le morphisme $E \to E \otimes_A B$ est injectif par pureté. Puisque E est un A-module injectif l'identité de E se prolonge en un morphisme de A-modules $u\colon E \otimes_A B \to E$, d'où un morphisme A-linéaire $B \to \mathrm{End}_A(E)$. Par dualité de Matlis on a $\mathrm{End}_A(E) \simeq A$, et la composée $B \to \mathrm{End}_A(E) \simeq A$ est un scindage de f. □

[13] Autrement dit le foncteur envoyant $M \in \mathrm{Mod}_A$ sur $B \otimes_A M$ muni de sa donnée de descente canonique induit une équivalence entre Mod_A et la catégorie des B-modules N munis d'un isomorphisme $N \otimes_A B \simeq B \otimes_A N$ de $B \otimes_A B$-modules vérifiant la condition usuelle de cocycle.

[14] Il s'agit de la topologie de Grothendieck la plus fine pour laquelle tous les préfaisceaux représentables sont des faisceaux.

On peut utiliser les liens entre scindage et pureté pour obtenir des conséquences importantes du théorème 1.1. Nous allons mentionner deux telles applications. Si $f\colon A \to B$ est une extension pure, alors $A \cap IB = I$ pour tout idéal I de A (cela ne fait que traduire l'injectivité du morphisme $A/I \to B/IB = B \otimes_A A/I$). Sous des hypothèses faibles cette propriété de contraction d'idéaux caractérise la pureté : Hochster (1977) a montré qu'une extension finie $A \to B$ d'un anneau noethérien intègre et normal est scindée si $A \cap IB = I$ pour tout idéal I de A. La conjecture du facteur direct et le théorème suivant sont donc équivalents (une implication étant triviale, comme remarqué ci-dessus).

Théorème 1.5. *Si $f\colon A \to B$ est une extension finie d'un anneau régulier A, alors $A \cap IB = I$ pour tout idéal I de A.*

Pour la deuxième application, rappelons qu'un morphisme d'anneaux $f\colon A \to B$ *descend la platitude* si pour tout $M \in \mathrm{Mod}_A$ la platitude sur B de $B \otimes_A M$ force celle de M. Par exemple, toute extension pure descend la platitude (Olivier, 1973). Raynaud et Gruson (1971) ont montré, généralisant un théorème de Ferrand, que toute extension finie descend la platitude. Ils ont demandé (question 1.4.3, Seconde Partie de loc.cit.) si toute extension entière d'un anneau noethérien [(15)] A descend la platitude, et ont aussi expliqué qu'il suffit de résoudre ce problème quand A est régulier ; or dans ce cas le théorème 1.1 permet de conclure puisque toute extension entière $A \to B$ est pure, en tant que colimite filtrante d'extensions finies, donc pures. Donc la conjecture du facteur direct fournit une réponse positive à la question de Raynaud et Gruson. Ohi (1996) a montré que les théorèmes 1.6 et 1.1 sont en fait équivalents.

Théorème 1.6. *Toute extension entière d'un anneau noethérien descend la platitude.*

Si A est un anneau et si $f\colon M \to N$ et $g\colon N \to P$ sont des morphismes de A-modules tels que $g \circ f$ soit pur, alors f est clairement pur. Le théorème 1.1 devient ainsi une conséquence directe du théorème fondamental suivant (qui sera discuté dans le § 1.5) et de la proposition 1.4.

Théorème 1.7. *Pour toute extension finie $A \to B$ d'un anneau régulier A il existe un morphisme d'anneaux $B \to C$ tel que $A \to C$ soit fidèlement plat.*

Remarque 1.8. 1. Géométriquement, on peut (suivant André et Fiorot, 2022) reformuler le théorème 1.1 (resp. 1.7) comme suit : si Y est un schéma noethérien régulier, alors tout morphisme fini surjectif $f\colon X \to Y$ est un recouvrement pour la topologie canonique (respectivement pour la topologie [(16)] fpqc) sur la catégorie des schémas.

[(15)] L'hypothèse que A soit noethérien n'est pas superflue !

[(16)] Attention, on dit bien la topologie et non pas la pré-topologie (l'énoncé en question serait faux pour la pré-topologie fpqc). Un morphisme $f\colon X \to Y$ de schémas affines est un recouvrement pour la topologie fpqc s'il existe un morphisme $X' \to X$ de schémas affines tel que $X' \to Y$ soit fidèlement plat.

2. Si l'on part d'une extension finie et pure $A \to B$, il est possible qu'un anneau C comme dans le théorème 1.7 n'existe pas (même avec A normal), *i.e.* la topologie fpqc est strictement plus faible que la topologie canonique. L'exemple 5.5 d'André et Fiorot (2022) est particulièrement simple (à énoncer, pas à prouver...) : l'inclusion $A := \mathbf{C}[X,Y]^{\mathbf{Z}/2\mathbf{Z}} \to B := \mathbf{C}[X,Y]$ (l'élément non trivial de $\mathbf{Z}/2\mathbf{Z}$ agissant par $(X,Y) \mapsto (-X,-Y)$) est pure (immédiat) mais pas un recouvrement fpqc (cela utilise les constructions de Raynaud et Gruson (1971) (1.4.1.1), les premiers à avoir fourni des contre-exemples).

1.5. Algèbres de Cohen–Macaulay pour les anneaux locaux noethériens

Contrairement à la conjecture du facteur direct, le théorème 1.7 ci-dessus est très difficile même quand A est une $\mathbf{Q}$-algèbre (toutes les preuves existantes passent par une réduction, via la technique des ultrafiltres, au cas des anneaux de caractéristique positive, où le Frobenius fait des merveilles). Démontré par Hochster et Huneke (1992) quand A contient un corps, et par André (2018b) (et récemment par Bhatt (2020)) pour A d'inégale caractéristique, ce théorème est une reformulation d'une conjecture de Hochster (1979) concernant l'existence d'une A-algèbre de Cohen–Macaulay pour tout anneau local noethérien A. Pour expliquer le lien, nous devons faire quelques rappels.

Soit $(A,\mathfrak{m})$ un anneau local noethérien. Une suite $x = (x_1,\dots,x_d)$ dans $\mathfrak{m}$ est un *système de paramètres* (ou *suite sécante maximale*) si $d = \dim A$ et si $A/(x_1,\dots,x_d)$ est de dimension 0, autrement dit s'il existe $t \geqslant 1$ tel que $\mathfrak{m}^t \subset (x_1,\dots,x_d)$. D'autre part, une suite $z = (z_1,\dots,z_k)$ d'éléments de A est une *suite régulière* dans un A-module M si $M/(z_1,\dots,z_k)M \neq 0$ et si la multiplication par z_{i+1} est injective dans $M/(z_1,\dots,z_i)M$ pour tout $0 \leqslant i < k$. On dit que M est un *A-module de Cohen–Macaulay* si tout système de paramètres dans A est une suite régulière dans M. On dit qu'une A-algèbre B est une *A-algèbre de Cohen–Macaulay* si B est un A-module de Cohen–Macaulay. Aucune hypothèse de finitude n'étant imposée à B, il ne faut donc pas confondre [17] cette définition avec celle d'un anneau de Cohen–Macaulay qui se trouve être une A-algèbre [18].

Le résultat suivant (Bartijn et Strooker, 1983, théorème 1.7 et Hochster et Huneke, 1995, 2.1.d) fait le lien avec le théorème 1.7.

Proposition 1.9. *Soit $(A,\mathfrak{m})$ un anneau local noethérien et soit M un A-module.*

a) Si A possède un système de paramètres qui est une suite régulière dans M, alors le complété $\mathfrak{m}$-adique de M est un A-module de Cohen–Macaulay.

b) Si A est régulier, un A-module est de Cohen–Macaulay si et seulement s'il est fidèlement plat sur A.

[17] Les diverses appellations dans la littérature anglophone donnent un peu le tournis : ce que nous définissons ici correspond à une *balanced (big) CM A-algebra*.

[18] Le langage de l'algèbre commutative n'est pas vraiment commutatif...

Compte tenu de cette proposition, le théorème 1.7 devient une conséquence [(19)] du résultat suivant, qui répond à la conjecture [(20)] de Hochster mentionnée ci-dessus.

Théorème 1.10. *Pour tout anneau local noethérien A il existe une A-algèbre de Cohen–Macaulay.*

Les constructions d'André que l'on verra dans la suite de cet exposé font intervenir des anneaux perfectoïdes, et les A-algèbres de Cohen–Macaulay qui en sortent ne sont presque jamais noethériennes. Ce n'est pas une grande surprise, puisque Hochster a déjà montré (théorème 6.1 de Hochster, 1979) que si A est un anneau local noethérien complet, normal, non Cohen–Macaulay, et contenant $\mathbf{Q}$, alors A n'admet pas de A-algèbre de Cohen–Macaulay noethérienne. Voir aussi Bhatt (2014b) pour des obstructions cohomologiques à l'existence de « petites algèbres de Cohen–Macaulay » [(21)] en caractéristique positive.

Le théorème 1.10 était connu pour des anneaux A d'égale caractéristique ou quand $\dim A \leqslant 3$, grâce aux travaux de Hochster et Huneke (1992) et Hochster (2002) (le cas d'inégale caractéristique en dimension 3 utilise de manière cruciale les travaux de Heitmann, 2002). Le cas restant est dû à André (2018b) (voir le § 1.7 pour l'approche de Bhatt). Cependant, pour l'instant ces techniques ne semblent pas permettre des avancées vers une autre conjecture célèbre de Hochster [(22)] : *tout anneau local noethérien complet A possède un A-module de Cohen–Macaulay de type fini*. Cette conjecture a un intérêt particulier puisqu'elle entraîne la conjecture de Serre (ouverte aussi depuis environ 50 ans) sur la (stricte) positivité des multiplicités d'intersection, mais en dehors du cas de la dimension $\leqslant 2$ très peu de choses sont connues.

Remarque 1.11. Dans une direction un peu différente mentionnons la conjecture de macaulayfication de Faltings : pour tout schéma noethérien quasi-excellent X il existe un schéma de Cohen–Macaulay $\tilde{X}$ et un morphisme projectif birationnel $\pi\colon \tilde{X} \to X$ qui est un isomorphisme au-dessus du lieu (ouvert) de Cohen–Macaulay de X. Voir Česnavičius (2021) pour une preuve, même sous des hypothèses plus faibles.

[(19)] Voici l'argument, suivant André, 2018b : soit $\mathfrak{m}$ un idéal maximal de A, soit $\widehat{A}_{\mathfrak{m}} := \varprojlim_n A_{\mathfrak{m}}/\mathfrak{m}^n A_{\mathfrak{m}}$ et soit $\wp$ un idéal premier minimal de $B \otimes_A \widehat{A}_{\mathfrak{m}}$ tel que $\widehat{A}_{\mathfrak{m}} \to B' := (B \otimes_A \widehat{A}_{\mathfrak{m}})/\wp$ reste injectif et fini (*cf.* remarque 1.3, point b)). Alors B' est local, et on choisit une B'-algèbre de Cohen–Macaulay $C(\mathfrak{m})$. Alors $C(\mathfrak{m})$ est fidèlement plate sur $\widehat{A}_{\mathfrak{m}}$ (proposition 1.9). Comme A est noethérien, $C := \prod_{\mathfrak{m}} C(\mathfrak{m})$ est plate sur A, et même fidèlement plate puisque l'image de $\operatorname{Spec}(C) \to \operatorname{Spec}(A)$ est stable par générisation (par platitude) et ne contient pas de point fermé par construction.

[(20)] À l'origine elle concernait l'existence de A-modules de Cohen–Macaulay : Hochster (1979) professe un certain pessimisme concernant l'existence de A-algèbres de Cohen–Macaulay...

[(21)] *i.e.* des algèbres de Cohen–Macaulay qui sont des extensions finies de l'anneau de base.

[(22)] On m'informe que l'on conjecture désormais le contraire...

1.6. Les conjectures homologiques

On trouvera d'excellentes présentations, par exemple dans Hochster (2007) et Roberts (1992), de l'écheveau des « conjectures homologiques » et de leurs diverses imbrications, nous nous contentons dans ce paragraphe de quelques extraits, qui découlent (par les travaux de Hochster (1983), Peskine et Szpiro (1972) et Evans et Griffith (1981)) de l'existence d'algèbres (ou même seulement de modules) de Cohen–Macaulay pour les anneaux locaux noethériens.

Théorème 1.12. *Soit $(R, \mathfrak{m})$ un anneau local noethérien, soit $x \in R$ et soient M, N des modules non nuls de type fini sur R.*

a) (conjecture de M. Auslander) *Si M est de dimension projective finie et sans x-torsion, alors R est sans x-torsion.*

b) (question de Bass) *Si M est de dimension injective finie alors R est Cohen–Macaulay.*

c) (conjecture d'intersection de Peskine–Szpiro) *Si $M \otimes_R N$ est de longueur finie, alors* $\dim N \leqslant \mathrm{pd}(M)$.

d) (« improved new intersection conjecture ») *Soit C un complexe de R-modules libres de type fini, concentré en degrés $[0, d]$, et tel que $H_{>0}(C)$ soit de longueur finie. Si $H_0(C)$ possède un générateur minimal non nul c tué par une puissance de $\mathfrak{m}$, alors* $\dim R \leqslant d$.

e) (« conjecture des syzygies d'Evans–Griffith ») *Supposons que R est intègre de Cohen–Macaulay. Si M est un kième module de syzygies, de dimension projective finie et non libre, alors M est de rang $\geqslant k$.*

Tous les résultats ci-dessus étaient connus à l'époque de l'article de Hochster (1979) quand l'anneau local contient un corps ou est de dimension $\leqslant 2$, et ouverts dans les cas restants. Evans et Griffith avaient montré [(23)] que l'existence de modules de Cohen–Macaulay implique le point d), qui implique e). En combinant les travaux de Hochster (1983) et Dutta (1987), on montre que d) est équivalent à la conjecture du facteur direct, et il implique c). Le point c) avait été démontré par Peskine et Szpiro (1972) en présence d'un corps, et par Roberts (1987) en général. Peskine et Szpiro ont montré que c) implique b) et a).

Hochster (1983, théorème 6.1) a montré que la conjecture du facteur direct est équivalente à cette autre conjecture, la *conjecture monomiale*, tout aussi charmante :

Théorème 1.13. *Si $x_1, \dots, x_n$ est un système de paramètres d'un anneau local noethérien $(A, \mathfrak{m})$, alors l'équation*

$$(x_1 \cdots x_n)^k = y_1 x_1^{k+1} + \cdots + y_n x_n^{k+1}$$

n'a pas de solutions $(y_1, \dots, y_n) \in A^n$ pour $k \geqslant 1$.

[(23)] Sans le dire... voir la section 2 de Hochster (1983).

Ce théorème découle facilement du théorème 1.10, voici les grandes lignes de l'argument. Si $x_1, \dots, x_n$ est une suite régulière dans un A-module M (pour un anneau A) et si $m_1, \dots, m_n \in M$ vérifient $x_1 m_1 + \cdots + x_n m_n = 0$ alors $m_i \in (x_1, \dots, x_n)M$ pour tout i [(24)]. On en déduit [(25)] que si a_i sont des entiers positifs et si $m, m_i \in M$ vérifient

$$x_1^{a_1} \cdots x_n^{a_n} m = x_1^{a_1+1} m_1 + \cdots + x_n^{a_n+1} m_n,$$

alors $m \in (x_1, \dots, x_n)M$. En particulier, comme $M/(x_1, \dots, x_n)M \neq 0$, on ne peut pas avoir $x_1^{a_1} \cdots x_n^{a_n} \in (x_1^{a_1+1}, \dots, x_n^{a_n+1})$.

Remarque 1.14. Le théorème ci-dessus est à comparer avec l'énoncé suivant, qui découle du théorème de Briançon-Skoda : si A est un anneau régulier de dimension $\leqslant n$, alors

$$(x_1 \cdots x_{n+1})^n \in (x_1^{n+1}, \dots, x_{n+1}^{n+1})$$

pour tous $x_1, \dots, x_{n+1} \in A$ (ce n'est déjà pas facile à démontrer pour $n = 2$!).

1.7. Algèbre dans la clôture intégrale absolue

Soit A un anneau intègre, et fixons une clôture algébrique K^+ du corps des fractions K de A. La *clôture intégrale absolue* A^+ de A est la clôture intégrale de A dans K^+. Elle est bien définie à isomorphisme près, et faiblement fonctorielle en A : tout morphisme $f\colon A \to B$ d'anneaux intègres induit [(26)] un morphisme $f^+\colon A^+ \to B^+$.

Pour voir le lien avec les conjectures discutées ci-dessus, mentionnons deux énoncés sympathiques. Soit $(A, \mathfrak{m})$ un anneau local noethérien complet et intègre. Si $\dim A \leqslant 2$ alors A^+ est une A-algèbre de Cohen–Macaulay (tout anneau noethérien intègre et normal de dimension 2 est de Cohen–Macaulay). À partir de la dimension 3 (resp. 4) et en égale caractéristique nulle (resp. en inégale caractéristique) A^+ n'est plus une A-algèbre de Cohen–Macaulay.

Dans un véritable tour de force d'algèbre commutative [(27)], Hochster et Huneke (1992) ont démontré le résultat suivant, qui implique immédiatement le théorème 1.10 en caractéristique positive, et même une forme plus forte, car la construction devient faiblement fonctorielle :

[(24)] Comme x_n est non diviseur de zéro dans $M/(x_1, \dots, x_{n-1})M$ on a $m_n = \sum_{i=1}^{n-1} x_i m_i'$ pour certains $m_i' \in M$. On a donc $\sum_{i=1}^{n-1} x_i(m_i + x_n m_i') = 0$, ce qui permet de conclure par récurrence.

[(25)] Si tous les a_i sont nuls, cela est évident, soit donc i tel que $a_i \neq 0$ et posons $z = \prod_{j \neq i} x_j^{a_j}$ et $x_j' = x_j^{a_j+1}$ pour $j \neq i$. Comme $x_i^{a_i}(x_i m_i - z m) + \sum_{j \neq i} x_j' m_j = 0$ et $(x_1', \dots, x_{i-1}', x_i^{a_i}, x_{i+1}', \dots, x_n')$ est une suite régulière, on obtient $x_i m_i - z m \in (x_1', \dots, x_{i-1}', x_i^{a_i}, x_{i+1}', \dots, x_n')M$, donc $z m \in (x_1', \dots, x_{i-1}', x_i, x_{i+1}', \dots, x_n')M$. On conclut par récurrence sur le nombre de a_j non nuls.

[(26)] Le cas d'une injection étant clair, supposons que f est surjectif ; on peut trouver un idéal premier $\mathfrak{q}$ de A^+ au-dessus de $\wp := \ker f$, et alors $A/\wp$ s'injecte dans $A^+/\mathfrak{q}$, ce dernier étant isomorphe à B^+.

[(27)] Voir l'article de Huneke et Lyubeznik (2007) pour une preuve plus simple.

Théorème 1.15. *Si A est un anneau local noethérien intègre et excellent* [28] *de caractéristique $p > 0$, alors A^+ est une A-algèbre de Cohen–Macaulay.*

Cette recette ne fonctionne plus si A est une $\mathbf{Q}$-algèbre, et la preuve du théorème 1.10 (et la fonctorialité faible) passe par une réduction délicate à la caractéristique positive.

Nous finissons cette longue introduction avec le résultat spectaculaire suivant, dû à Bhatt (2020), et qui fournit un analogue du théorème de Hochster–Huneke en inégale caractéristique. Il entraîne immédiatement le théorème 1.10 en inégale caractéristique, ainsi que la fonctorialité faible des algèbres de Cohen–Macaulay. La preuve est très difficile et utilise toute la palette des développements en théorie de Hodge p-adique de ces dix dernières années, ainsi que la théorie prismatique de Bhatt et Scholze (2022) . Elle mériterait sans doute un exposé à part entière...

Théorème 1.16. *Soit $(A, \mathfrak{m})$ un anneau local noethérien intègre et excellent de caractéristique résiduelle $p > 0$. Le complété p-adique de A^+ est une A-algèbre de Cohen–Macaulay.*

Remarque 1.17.

1. Hochster (1983) a démontré que la conjecture du facteur direct est équivalente au sympathique énoncé suivant : pour tout anneau local noethérien complet et intègre $(A, \mathfrak{m})$ le A-dual de A^+ est non nul. Si A est de dimension n, par dualité locale cela équivaut à $\mathrm{H}^n_{\mathfrak{m}}(A^+) \neq \{0\}$. Cette non annulation se voit facilement à partir du théorème 1.1 (le point délicat est que l'on peut aller dans l'autre sens) : par le théorème de Cohen on peut supposer [29] que A est aussi régulier, mais alors toute extension finie $A \to B$ dans A^+ est scindée, donc le morphisme $\mathrm{H}^n_{\mathfrak{m}}(A) \to \mathrm{H}^n_{\mathfrak{m}}(B)$ est injectif, et il en est de même de $\mathrm{H}^n_{\mathfrak{m}}(A) \to \mathrm{H}^n_{\mathfrak{m}}(A^+)$, ce qui permet de conclure puisque $\mathrm{H}^n_{\mathfrak{m}}(A) \neq \{0\}$.
2. Dans la situation du théorème 1.16, on montre (c'est le coeur de l'article de Bhatt (2020)) que $\mathrm{H}^i_{\mathfrak{m}}(A^+/p)$ est nul pour $i < \dim(A/p)$ et $\mathrm{H}^i_{\mathfrak{m}}(A^+)$ est nul pour $i < \dim A$. Même en dimension 3 ce genre d'énoncé va beaucoup plus loin que ceux de Heitmann (2002) (on passe d'une presque nullité à une vraie nullité!). Cela lui permet de montrer que si A est un fragmenteur, alors A est de Cohen–Macaulay : par le même argument que ci-dessus la flèche $\mathrm{H}^i_{\mathfrak{m}}(A/p) \to \mathrm{H}^i_{\mathfrak{m}}(A^+/p)$ est injective (car A est un fragmenteur), donc $\mathrm{H}^i_{\mathfrak{m}}(A/p)$ est nul pour $i < \dim(A/p)$, et A est de Cohen–Macaulay.

2. Anneaux perfectoïdes : aspects algébriques

Le but de cette section est de mettre ensemble un certain nombre de résultats sur les anneaux perfectoïdes « entiers », qui sont éparpillés façon puzzle dans la littérature. Nous renvoyons le lecteur aux articles fondamentaux de Bhatt, Morrow

[28] Par exemple complet.
[29] Cela demande un petit peu de travail...

et Scholze (2018, section 3), Bhatt et Scholze (2022, sections 2 et 3), Česnavičius et Scholze (2019, section 2), au livre de Kedlaya et Liu (2015) et aux exposés de Fontaine (2013) et Morrow (2019) dans ce Séminaire pour plus de détails.

On fixe pour toute la suite un nombre premier p et on note $\mathrm{Perf}_{\mathbf{F}_p}$ la catégorie des $\mathbf{F}_p$-algèbres parfaites, *i.e.* celles dont le morphisme de Frobenius $\varphi\colon x \mapsto x^p$ est un automorphisme (une telle algèbre est donc réduite). La réduction modulo p et le foncteur $W(-)$ des vecteurs de Witt p-typiques induisent des équivalences quasi-inverses entre $\mathrm{Perf}_{\mathbf{F}_p}$ et la catégorie des anneaux p-complets, sans p-torsion, dont la réduction modulo p est parfaite.

2.1. Vecteurs de Witt

Pour tout $R \in \mathrm{Perf}_{\mathbf{F}_p}$ il existe une unique application multiplicative, *mais pas forcément additive* $[\ ]\colon R \to W(R)$ telle que $[a] \equiv a \pmod p$ pour tout $a \in R$. Tout $x \in W(R)$ s'écrit $x = \sum_{n\geqslant 0}[x_n]p^n$ pour une unique suite $(x_n)_{n\geqslant 0}$ dans R. Il sera très utile de considérer x comme une « fonction holomorphe de la variable p ». Il est donc tentant d'introduire la notation

$$x(0) := x_0,\ x'(0) := x_1.$$

L'application $x \mapsto x(0)$ induit un isomorphisme d'anneaux $W(R)/p \simeq R$, et un calcul immédiat montre que pour tous $x, y \in W(R)$ on a

$$(xy)'(0) = x(0)y'(0) + y(0)x'(0), \tag{1}$$

ce qui explique la notation. L'optimisme dégagé par cette observation est tempéré par le fait que $(x+y)'(0) \neq x'(0) + y'(0)$ en général. Cependant la relation ci-dessus jouera un rôle important à plusieurs reprises.

Remarque 2.1. L'anneau $W(R)$ est muni d'un relèvement du Frobenius $\varphi\colon W(R) \to W(R)$, défini par $\varphi(\sum_{n\geqslant 0}[x_n]p^n) = \sum_{n\geqslant 0}[x_n^p]p^n$. L'application $\delta\colon W(R) \to W(R)$ définie par $\delta(x) := \frac{\varphi(x)-x^p}{p}$ est une p-dérivation (au sens de Buium) sur $W(R)$, *i.e.* elle vérifie

$$\delta(xy) = x^p\delta(y) + y^p\delta(x) + p\delta(x)\delta(y),\ \delta(x+y) = \delta(x) + \delta(y) + \frac{x^p + y^p - (x+y)^p}{p}$$

pour $x, y \in W(R)$. Cette application joue un rôle crucial dans Bhatt et Scholze (2022). Notons que $x'(0)^p$ est simplement la réduction modulo p de $\delta(x)$.

2.2. Constructions de Fontaine

Rappelons rapidement quelques constructions fondamentales dues à Fontaine.

Définition 2.2. Un *élément p-puissant* d'un anneau A est une suite $(a_n)_{n\geqslant 0}$ d'éléments de A telle que $a_{n+1}^p = a_n$ pour tout n. On note $\varprojlim_{x\mapsto x^p} A$ l'ensemble des éléments p-puissants de A. On dira aussi, abusivement, qu'un élément $a \in A$ est p-puissant s'il est *muni* d'une suite $(a_n)_{n\geqslant 0} \in \varprojlim_{x\mapsto x^p} A$ telle que $a_0 = a$. On écrira $a^{p^{-n}}$ ou a^{1/p^n} au lieu de a_n et on notera $(a^{p^{-\infty}})$ l'idéal de A engendré par les a_n. Noter que l'expression « un élément p-puissant a de A » contient implicitement la donnée d'une suite de racines compatibles $(a^{p^{-n}})_{n\geqslant 0}$ de a.

Par exemple, si $R \in \mathrm{Perf}_{\mathbf{F}_p}$ alors l'application $a \mapsto ([a^{1/p^n}])_{n\geqslant 0}$ induit une bijection entre R et $\varprojlim_{x\mapsto x^p} W(R)$. Plus généralement, soit B un anneau p-complet. La réduction modulo p induit une bijection multiplicative de $\varprojlim_{x\mapsto x^p} B$ sur le *basculé* $B^\flat := \varprojlim_{x\mapsto x^p} B/p$ de B, la $\mathbf{F}_p$-algèbre parfaite des suites $(x_n)_{n\geqslant 0}$ dans B/p telles que $x_{n+1}^p = x_n$ pour tout n. L'inverse $\iota\colon B^\flat \to \varprojlim_{x\mapsto x^p} B$ est construit comme suit : si $(x_n)_{n\geqslant 0} \in B^\flat$ et si $b_n \in B$ est un relèvement quelconque de x_n, alors la suite $(b_{n+k}^{p^k})_{k\geqslant 0}$ converge p-adiquement dans B vers un élément $\tilde{b}_n$ qui ne dépend pas du choix des b_n, et $\iota((x_n)_{n\geqslant 0}) = (\tilde{b}_n)_{n\geqslant 0}$. Pour tout $x \in B^\flat$ on a

$$\iota(x) = (\sharp(x), \sharp(x^{1/p}), \ldots),$$

où $\sharp\colon B^\flat \to B$ est l'application multiplicative (mais pas additive en général) obtenue en composant ι avec la projection sur la première composante $\varprojlim_{x\mapsto x^p} B \to B$, *i.e.*

$$\sharp(x_0, x_1, \ldots) = \tilde{b}_0 = \lim_{k\to\infty} b_k^{p^k}.$$

Un des grands classiques de la théorie de Fontaine est l'application

$$\theta_B\colon W(B^\flat) \to B,\ \theta_B(\sum_{n\geqslant 0} [a_n]p^n) = \sum_{n\geqslant 0} a_n^\sharp p^n.$$

Il s'agit de l'unique morphisme d'anneaux tel que $\theta_B([b]) = b^\sharp$ pour tout $b \in B^\flat$.

Toutes ces constructions sont fonctorielles en l'anneau p-complet B et permettent de caractériser $W(R)$ (pour $R \in \mathrm{Perf}_{\mathbf{F}_p}$) dans la catégorie plus grande des anneaux p-complets, comme suit. Soit B un anneau p-complet et soit $f\colon W(R) \to B$ un morphisme d'anneaux. On obtient un morphisme d'anneaux $f^\flat\colon R \to B^\flat$ en passant aux basculés, plus concrètement en envoyant a sur la suite des réductions modulo p des $f([a^{1/p^n}])$, $n \geqslant 0$. Réciproquement, un morphisme d'anneaux $g\colon R \to B^\flat$ en induit un autre $f := \theta_B \circ W(g)\colon W(R) \to B$, tel que $f^\flat = g$. On a donc

$$f(\sum_{n\geqslant 0} [a_n]p^n) = \sum_{n\geqslant 0} g(a_n)^\sharp p^n.$$

On peut résumer cette discussion comme suit :

Proposition 2.3. *Soit $R \in \mathrm{Perf}_{\mathbf{F}_p}$. Les applications $f \mapsto f^\flat$ et $g \mapsto \theta_B \circ W(g)$ induisent une bijection fonctorielle en l'anneau p-complet B*

$$\mathrm{Hom}(W(R), B) \simeq \mathrm{Hom}(R, B^\flat).$$

2.3. Anneaux perfectoïdes

Soit $R \in \mathrm{Perf}_{\mathbf{F}_p}$. On dit que $\xi \in W(R)$ est *distingué* si $\xi'(0)$ est inversible dans R, autrement dit si $\xi = [\xi(0)] + pu$ avec u inversible dans $W(R)$. Ainsi ξ est un avatar d'une « fonction holomorphe de la variable p, biholomorphe au voisinage de 0 ».

Remarque 2.4. Soit $R \in \mathrm{Perf}_{\mathbf{F}_p}$ et soit $\xi \in W(R)$ un élément distingué.

a) Pour tout morphisme $R \to S$ dans $\mathrm{Perf}_{\mathbf{F}_p}$ l'image de ξ reste distinguée dans $W(S)$. On la notera encore (abusivement) ξ. Si R est $\xi(0)$-complet, la relation (1) montre que $u\xi$ est distingué dans $W(R)$ pour tout élément inversible u de $W(R)$.

b) L'élément ξ n'est pas un diviseur de zéro dans $W(R)$: si $\xi x = 0$ alors $\xi(0)x(0) = 0$ et « en prenant la dérivée en 0 », *i.e.* en utilisant la relation (1) on obtient $\xi'(0)x(0) + \xi(0)x'(0) = 0$. Ainsi $\xi'(0)x(0)^2 = 0$, et comme $\xi'(0)$ est inversible et R est parfait, on a $x(0) = 0$. Donc $x = px_1$ pour un $x_1 \in W(R)$, et $\xi x_1 = 0$, puis par itération $x \in \cap_{n \geqslant 1} p^n W(R) = \{0\}$. Le même argument montre que si $a, x \in W(R)$ vérifient $a \mid \xi x$, alors $x(0)^2 \in (a(0)x'(0), a(0)x(0), a'(0)x(0))$. En particulier, si $p^2 \mid \xi x$ alors $p \mid x$.

Introduisons maintenant, suivant le point de vue prismatique de Bhatt et Scholze (2022), la classe la plus générale (à ce jour...) d'*anneaux perfectoïdes* :

Définition 2.5. Un anneau A est dit *perfectoïde* s'il est isomorphe à un anneau $W(R)/(\xi)$ avec $R \in \mathrm{Perf}_{\mathbf{F}_p}$ et $\xi \in W(R)$ distingué. On note Perf la catégorie des anneaux perfectoïdes, les morphismes étant ceux d'anneaux.

Remarque 2.6. Les premiers êtres du monde perfectoïde sont les corps perfectoïdes [(30)], par exemple le complété $\mathbf{C}_p$ d'une clôture algébrique de $\mathbf{Q}_p$. Puis ont vu le jour (Scholze, 2012) les algèbres de Banach perfectoïdes sur un tel corps (certaines étaient déjà bien présentes dans les articles de Colmez et Fontaine (2000) et Colmez (2002)...). Fontaine (2013) a introduit dans son exposé une classe plus large d'anneaux perfectoïdes. Enfin, Bhatt, Morrow et Scholze (2018) introduisent la classe la plus générale d'anneaux perfectoïdes, équivalente à celle introduite ci-dessus. Voir aussi la proposition 2.12 pour le lien avec les définitions plus anciennes (et plus restreintes).

[(30)] Attention : un corps perfectoïde n'est pas la même chose qu'un anneau perfectoïde (au sens de la définition 2.5) qui est aussi un corps. Tout n'est pas parfait dans le monde perfectoïde...

Exemple 2.7. 1. Il est évident que $\mathrm{Perf}_{\mathbf{F}_p}$ s'identifie à la sous-catégorie de Perf des objets tués par p (noter que $p \in W(R)$ est distingué). En particulier l'anneau nul, $\mathbf{F}_p$, $\mathbf{F}_p[T^{1/p^\infty}] := \varinjlim_n \mathbf{F}_p[T^{1/p^n}]$ sont perfectoïdes.

2. Soit $A = W(R)/(\xi)$, avec $R \in \mathrm{Perf}_{\mathbf{F}_p}$ et $\xi \in W(R)$ distingué. L'image π de $[\xi(0)^{1/p}]$ dans A, munie de la suite de racines donnée par les images des $[\xi(0)^{1/p^{n+1}}]$ $(n \geqslant 0)$, est un élément p-puissant (définition 2.2) de A. On a une égalité $(\pi^p) = (p)$ d'idéaux de A, puisque $\xi = \xi(0) + pu$ avec u inversible. Ainsi non seulement $\mathbf{Z}_p$ n'est pas perfectoïde, il n'y a même pas de morphisme d'un anneau perfectoïde vers $\mathbf{Z}_p$. Noter que si $\xi(0) = 0$ alors $\pi = 0$ et $A \in \mathrm{Perf}_{\mathbf{F}_p}$.

3. Gardons le contexte ci-dessus. Le Frobenius $A/\pi \to A/\pi^p$ s'identifie au Frobenius $R/\xi(0)^{1/p} \to R/\xi(0)$, et c'est un isomorphisme. Ainsi pour tout anneau perfectoïde A il existe un élément p-puissant $\pi \in A$ tel que $(\pi^p) = (p)$ (la proposition 2.8 ci-dessous montre que A est π-complet) et tel que le Frobenius $A/\pi \to A/\pi^p$ soit un isomorphisme. Voir la proposition 2.12 pour une réciproque partielle.

4. Pour $R = \mathbf{F}_p[T^{1/p^\infty}]$ l'anneau $W(R)$ est isomorphe au complété p-adique de $\mathbf{Z}_p[T^{1/p^\infty}]$ (réduire modulo p pour s'en convaincre). L'élément $\xi = [T] - p$ est distingué et $W(R)/(\xi)$ est isomorphe au complété p-adique de $\mathbf{Z}_p[p^{1/p^\infty}]$. Le même genre d'argument montre que si A est un anneau perfectoïde, alors le complété p-adique $A\langle T^{1/p^\infty}\rangle$ de $A[T^{1/p^\infty}] = \varinjlim_n A[T^{1/p^n}]$ l'est aussi.

5. Comme W commute aux produits, on voit facilement qu'un produit quelconque d'anneaux perfectoïdes est perfectoïde. Il n'est pas vrai (*cf.* remarque 2.13), et ceci posera de sérieux soucis plus tard, qu'une limite projective (dans la catégorie des anneaux) d'anneaux perfectoïdes est perfectoïde.

2.4. Basculement

Le résultat suivant sera constamment utilisé par la suite.

Proposition 2.8. *Tout anneau perfectoïde est p-complet et sa torsion p-primaire est tuée par p.*

Démonstration. On peut supposer que l'anneau est de la forme $A = W(R)/(\xi)$ avec $R \in \mathrm{Perf}_{\mathbf{F}_p}$ et $\xi \in W(R)$ distingué. Pour montrer que p annule $A[p^\infty]$ il suffit de montrer que $A[p^2] = A[p]$, autrement dit que si $x \in W(R)$ vérifie $\xi \mid p^2 x$ alors $\xi \mid px$, ce qui se déduit de la dernière phrase de la remarque 2.4. Montrons ensuite que A est p-complet [(31)]. Il suffit de voir que $(\xi) \subset W(R)$ est fermé pour la topologie p-adique. Mais on vient d'établir l'égalité $(\xi) \cap p^2 W(R) = p((\xi) \cap pW(R))$, et une récurrence immédiate donne $(\xi) \cap p^{n+1}W(R) = p^n((\xi) \cap pW(R))$. □

[(31)] Une manière savante est de remarquer que A est p-complet au sens dérivé et sa torsion p-primaire étant bornée le mot « dérivé » est superflu...

Soit $R \in \mathrm{Perf}_{\mathbf{F}_p}$ et soit $\xi \in W(R)$ un élément distingué. En posant $A = W(R)/(\xi)$, l'anneau A/p s'identifie à $R/\xi(0)$, en particulier *le Frobenius est surjectif sur* A/p. Cela implique, par réduction modulo p et le caractère p-complet des anneaux en présence, la surjectivité de l'application de Fontaine $\theta_A \colon W(A^\flat) \to A$. Pour étudier son noyau on commence par décrire $A^\flat$. Puisque R est parfait et $A/p \simeq R/\xi(0)$, les morphismes $x \mapsto x^{p^n}$ induisent un isomorphisme entre les systèmes projectifs $\{A/p\}$ et $\{R/\xi(0)^{p^n}\}$, les transitions étant induites par le Frobenius pour le premier et les projections canoniques pour le second. Donc $A^\flat = \varprojlim_{x \mapsto x^p} A/p$ est isomorphe à la $\xi(0)$-complétion $\hat{R}$ de R, en particulier $A^\flat$ *est* $\xi(0)$*-complet*. Pour aller plus loin nous avons besoin du :

Lemme 2.9. *Le morphisme $R \to \hat{R}$ induit un isomorphisme $A \simeq W(\hat{R})/(\xi)$.*

Démonstration. Puisque A et $W(\hat{R})/(\xi)$ sont p-complets (proposition 2.8), il suffit de voir que le morphisme induit $A/p^n \to W(\hat{R})/(\xi, p^n)$ est un isomorphisme pour tout $n \geqslant 1$. Mais $\xi - [\xi(0)]$ engendre le même idéal que p, donc

$$A/p^n \simeq W(R)/(p^n, \xi) = W(R)/([\xi(0)]^n, \xi) \simeq (W(R)/[\xi(0)]^n)/(\xi).$$

Comme $R/\xi(0)^n \simeq \hat{R}/\xi(0)^n$, le morphisme $W(R)/[\xi(0)]^n \to W(\hat{R})/[\xi(0)]^n$ est un isomorphisme, ce qui finit la preuve (en reprenant la chaîne d'isomorphismes ci-dessus avec R remplacé par $\hat{R}$). □

L'adjonction entre basculement et vecteurs de Witt (proposition 2.3) montre que le morphisme $\theta_A \colon W(A^\flat) \to A$ s'identifie, via les isomorphismes $A^\flat \simeq \hat{R}$ et $A \simeq W(\hat{R})/(\xi)$, à la projection canonique $W(\hat{R}) \to W(\hat{R})/(\xi)$, en particulier son noyau est engendré par ξ. On obtient ainsi l'isomorphisme

$$\theta_A \colon W(A^\flat)/(\xi) \simeq A, \tag{2}$$

d'où le premier résultat fondamental de la théorie :

Théorème 2.10. *Si A est un anneau perfectoïde et si $\xi \in W(A^\flat)$ est distingué et engendre* $\ker(\theta_A)$*, alors la catégorie des A-algèbres perfectoïdes est équivalente, via $B \mapsto B^\flat$ et $R \mapsto W(R)/(\xi)$, à celle des $A^\flat$-algèbres parfaites $\xi(0)$-complètes.*

Démonstration. Il suffit de montrer que $\theta_B \colon W(B^\flat)/(\xi) \simeq B$ et que $B^\flat$ est $\xi(0)$-complète pour toute A-algèbre perfectoïde B. On vient de voir que θ_B est surjective et qu'il existe un élément distingué $\xi_1 \in W(B^\flat)$ tel que $\ker(\theta_B) = (\xi_1)$, $B^\flat$ étant $\xi_1(0)$-complet. Comme B est une A-algèbre, la fonctorialité de la construction de l'application θ montre que $\xi \in \ker(\theta_B)$, donc $\xi = \xi_1 v$ pour un $v \in W(B^\flat)$. En « dérivant » on obtient (*cf.* relation (1)) $\xi'(0) = \xi_1'(0)v(0) + \xi_1(0)v'(0)$, donc $\xi_1'(0)v(0) = \xi'(0) - \xi_1(0)v'(0)$ est inversible (puisque $\xi'(0)$ l'est et que $B^\flat$ est $\xi_1(0)$-complet). Comme $\xi_1'(0)$ est inversible, $v(0)$ l'est tout autant et donc il en est de même de v, et $(\xi) = (\xi_1)$. En particulier $B^\flat$ est aussi $\xi(0)$-complète, ce qui finit la preuve. □

La propriété de stabilité suivante jouera un rôle crucial par la suite.

Corollaire 2.11. *Soit A un anneau perfectoïde. Si B et C sont des A-algèbres perfectoïdes, alors le complété p-adique $B\widehat{\otimes}_A C$ de $B \otimes_A C$ est un anneau perfectoïde. Plus précisément, si ξ engendre le noyau de $\theta_A \colon W(A^\flat) \to A$, alors*

$$B\widehat{\otimes}_A C \simeq W(B^\flat \otimes_{A^\flat} C^\flat)/(\xi).$$

Démonstration. La $A^\flat$-algèbre $R := B^\flat \otimes_{A^\flat} C^\flat$ est parfaite puisque $A^\flat, B^\flat, C^\flat$ le sont. Comme ξ reste distingué dans $W(R)$, l'anneau $T = W(R)/(\xi)$ est perfectoïde, donc p-complet. Les morphismes $B^\flat \to R$ et $C^\flat \to R$ combinés avec les isomorphismes canoniques (théorème 2.10) $W(A^\flat)/(\xi) \simeq A$, $W(B^\flat)/(\xi) \simeq B$, $W(C^\flat)/(\xi) \simeq C$ induisent un morphisme $B \otimes_A C \to T$, qui se prolonge en un morphisme $B\widehat{\otimes}_A C \to T$ puisque T est p-complet. Pour montrer que c'est un isomorphisme il suffit de voir qu'il induit une bijection $\mathrm{Hom}(T, S) \to \mathrm{Hom}(B\widehat{\otimes}_A C, S)$ pour tout anneau p-complet S. Comme

$$\mathrm{Hom}(B\widehat{\otimes}_A C, S) = \mathrm{Hom}(B \otimes_A C, S) = \\ \mathrm{Hom}\Big(\frac{W(B^\flat)}{(\xi)}, S\Big) \times_{\mathrm{Hom}\left(\frac{W(A^\flat)}{(\xi)}, S\right)} \mathrm{Hom}\Big(\frac{W(C^\flat)}{(\xi)}, S\Big),$$

le résultat suit formellement de l'adjonction fournie par la proposition 2.3. □

2.5. Anneaux perfectoïdes et Frobenius

Le résultat fondamental suivant (lemmes 3.9 et 3.10 de Bhatt, Morrow et Scholze (2018)) fournit le critère le plus simple pour tester le caractère perfectoïde d'un anneau, sous des hypothèses faibles. Cela permet de fabriquer des tas d'anneaux perfectoïdes sans être obligé de fournir une présentation $W(R)/(\xi)$, et fait aussi le lien avec les définitions plus anciennes (*cf.* remarque 2.6).

Proposition 2.12. *Soit A un anneau contenant un non diviseur de zéro $\pi \in A$ tel que $\pi^p \mid p$, A est π-complet et le Frobenius induit un isomorphisme $A/\pi \simeq A/\pi^p$. Alors A est perfectoïde.*

Démonstration. Notons que A est p-complet car $\pi^p \mid p$ et A est π-complet. Pour montrer la surjectivité de $\theta_A \colon W(A^\flat) \to A$ il suffit donc de montrer celle du Frobenius sur A/p. Pour tout $x \in A$ la surjectivité du Frobenius modulo π^p et la π-complétude de A permettent d'écrire $x = x_0^p + \pi^p x_1^p + \pi^{2p} x_2^p + \cdots$ pour certains $x_n \in A$, et alors $x \equiv (x_0 + \pi x_1 + \pi^2 x_2 + \cdots)^p \pmod{p}$.

Ensuite, on construit un élément distingué dans le noyau de $\theta := \theta_A$. La surjectivité du Frobenius $A/\pi \to A/\pi^p$ et la π-complétude de A fournissent [(32)] un élément p-puissant (définition 2.2) $u \in A$ tel que $\pi \equiv u \pmod{\pi^p}$. Comme A est π-complet, π et u engendrent le même idéal de A, donc on peut remplacer π par u et supposer que $\pi = f^\sharp$ pour un $f \in A^\flat$. Comme θ est surjective et $\pi^p \mid p$, il existe $x \in W(A^\flat)$ tel que $p = \pi^p\theta(-x) = \theta(-x[f]^p)$. En posant $\xi = p + x[f]^p$, on a $\xi'(0) = 1 + x'(0)f^p$. On vérifie ensuite que $A^\flat$ est f-complet, donc ξ est distingué, et que $\ker(\theta) = (\xi)$. □

Remarque 2.13. On déduit facilement de la proposition 2.12 les résultats suivants :

1. Soit A un anneau intègre, sans p-torsion, p-adiquement séparé et tel que $p \in \mathrm{Rad}(A)$. Alors le complété p-adique $\widehat{A^+}$ de la clôture intégrale absolue A^+ de A est perfectoïde.
2. L'anneau $\mathbf{Z}_p$ n'est pas perfectoïde (exemple 2.7), mais le théorème d'Ax–Sen–Tate l'exhibe comme l'anneau des invariants de $\widehat{\mathbf{Z}_p^+}$ sous l'action du groupe de Galois absolu de $\mathbf{Q}_p$. On voit donc qu'une limite projective d'anneaux perfectoïdes ne l'est plus forcément (un autre exemple est fourni par les sous-anneaux $\widehat{\mathbf{Z}_p[\mu_{p^\infty}]}$ et $\widehat{\mathbf{Z}_p[p^{1/p^\infty}]}$ de $\widehat{\mathbf{Z}_p^+}$, qui sont perfectoïdes, mais dont l'intersection, égale à $\mathbf{Z}_p$, ne l'est pas).

Remarque 2.14. Le lien entre anneaux perfectoïdes et Frobenius est rendu plus clair par la théorie prismatique de Bhatt et Scholze (2022). Soit A un anneau muni d'une p-dérivation δ (*cf.* remarque 2.1), et soit $\varphi(x) = x^p + p\delta(x)$ le relèvement du Frobenius induit par δ. Si I est un idéal de A, on dit que la paire (A, I) est un *prisme* si I définit un diviseur de Cartier dans $\mathrm{Spec}(A)$, A est (p, I)-complet (au sens dérivé) et $p \in I + \varphi(I)A$. Bhatt et Scholze montrent (théorème 3.10 de loc.cit.) que la catégorie des prismes parfaits (*i.e.* pour lesquels $\varphi\colon A \to A$ est un isomorphisme) est équivalente à celle des anneaux perfectoïdes, via les foncteurs $(A, I) \mapsto A/I$ et $R \mapsto (W(R^\flat), \ker(\theta_R))$.

2.6. Anneaux perfectoïdes et p-clôtures intégrales

Les articles de Roberts (2008) et d'André (2018a) mettent en avant une relation fondamentale entre la notion de p-clôture intégrale (ou simplement p-clôture, pour raccourcir) et celle d'anneau perfectoïde.

Définition 2.15. Soit A un sous-anneau d'un anneau B. On dit que A est *p-clos dans B* si pour tout $x \in B$ vérifiant $x^p \in A$ on a $x \in A$. Il existe un plus petit sous-anneau p-clos de B contenant A, que l'on appelle la *p-clôture de A dans B*.

[(32)] Prendre une suite $(a_n)_{n\geqslant 0}$ dans A telle que $a_0 = \pi$ et $a_{n+1}^p \equiv a_n \pmod{\pi^p}$ et poser $u = \lim_{n\to\infty} a_n^{p^n}$, la suite $(a_n^{p^n})_{n\geqslant 0}$ étant de Cauchy pour la topologie π-adique.

Les observations suivantes sont dues à Roberts (2008). On fixe dans la suite de ce paragraphe un anneau A muni d'un non diviseur de zéro π tel que $\pi^p \mid p$. On note $\varphi\colon A/\pi \to A/\pi^p$ le Frobenius.

Proposition 2.16. *a) La p-clôture de A dans $A[\frac{1}{\pi}]$ est*

$$R = \Big\{ x \in A\Big[\frac{1}{\pi}\Big] \mid \exists n \geqslant 0, x^{p^n} \in A \Big\}.$$

b) L'anneau A est p-clos dans $A[\frac{1}{\pi}]$ si et seulement si $\varphi\colon A/\pi \to A/\pi^p$ est injectif.

Démonstration. a) La seule difficulté est de montrer que R est bien un sous-anneau de $A[\frac{1}{\pi}]$, et plus précisément qu'il est stable par addition. Prenons $x, y \in R$ et $n, k \geqslant 1$ tels que $x^{p^n}, y^{p^n}, \pi^k x, \pi^k y$ soient tous dans A, et montrons qu'en posant $N = 2p^n k + n$ on a $(x+y)^{p^N} \in A$, ce qui permettra de conclure que $x + y \in R$. Il suffit de voir que $\binom{p^N}{i} x^{p^N - i} y^i \in A$ pour tout $0 \leqslant i \leqslant p^N$. Cela est évident si $p^n \mid i$, supposons donc que ce n'est pas le cas. Puisque $v_p(i) < n$ on a $p^{N-n} \mid \binom{p^N}{i}$, donc $\pi^{2p^n k} \mid \binom{p^N}{i}$. Comme $x^{p^n}, y^{p^n} \in A$, il suffit de voir que $\pi^{2p^n k} x^u y^v \in A$ pour $0 \leqslant u, v < p^n$, ce qui est clair puisque $\pi^k x, \pi^k y \in A$.

b) Il est clair que φ est injectif si A est p-clos dans $A[\frac{1}{\pi}]$. Dans l'autre sens, soit $x \in A[\frac{1}{\pi}]$ tel que $x^p \in A$ et soit $n \geqslant 1$ tel que $\pi^n x \in A$. Alors $(\pi^n x)^p \in \pi^p A$, donc $\pi^n x \in \pi A$ et $\pi^{n-1} x \in A$. En itérant on obtient $x \in A$. □

Nous aurons besoin du résultat suivant dans la preuve du théorème 4.3.

Proposition 2.17. *Supposons que le Frobenius $A/\pi \to A/\pi^p$ est un isomorphisme, que $g \in A$ et π sont p-puissants (déf. 2.2) et que g est non diviseur de zéro modulo π. Pour tout $n \geqslant 1$ la p-clôture de $A[\frac{g}{\pi^n}]$ dans $A[\frac{1}{\pi}]$ est $A[\frac{g^{1/p^j}}{\pi^{n/p^j}}, j \geqslant 0]$, et son complété π-adique est perfectoïde.*

Démonstration. L'algèbre $C := A[\frac{g^{1/p^j}}{\pi^{n/p^j}}, j \geqslant 0]$ est la réunion croissante de ses sous-algèbres $A[\frac{g^{1/p^j}}{\pi^{n/p^j}}]$, et clairement contenue dans la p-clôture de $A[\frac{g}{\pi^n}]$ dans $A[\frac{1}{\pi}]$.

Notons $\varphi_R\colon R/\pi \to R/\pi^p$ le Frobenius d'une A-algèbre R. Il suffit de montrer que φ_C est bijectif : le complété π-adique de C sera perfectoïde par la proposition 2.12, et C sera p-close dans $C[\frac{1}{\pi}] = A[\frac{1}{\pi}]$ (proposition 2.16), donc égale à la p-clôture de $A[\frac{g}{\pi^n}]$ dans $A[\frac{1}{\pi}]$.

Posons $u_j = \pi^{n/p^j} X^{1/p^j} - g^{1/p^j} \in A[X^{1/p^j}]$. Puisque g n'est pas un diviseur de zéro modulo π, la remarque 2.18 ci-dessous fournit des isomorphismes $A[\frac{g^{1/p^j}}{\pi^{n/p^j}}] \simeq A[X^{1/p^j}]/(u_j)$ compatibles avec la variation de j, d'où un isomorphisme $C \simeq B/I$ avec $B = A[X^{1/p^\infty}]$ et $I = (u_0, u_1, \dots)$. Puisque φ_A est bijectif, il en est de même de φ_B. En utilisant les congruences $u_{j+1}^p \equiv u_j \pmod{\pi^p B}$, on en déduit facilement que $\varphi_{B/I}$ est un isomorphisme. □

Remarque 2.18. Soient R un anneau et $r \in R$ non diviseur de zéro. Si $s \in R$ est non diviseur de zéro modulo r, alors le morphisme naturel $R[X]/(rX-s) \to R[\frac{s}{r}] \subset R[\frac{1}{r}]$ est un isomorphisme (exercice !).

2.7. Deux résultats techniques importants

Nous travaillerons souvent avec des anneaux sans p-torsion, et le caractère perfectoïde d'un anneau a le bon goût de se propager à son quotient maximal sans p-torsion :

Proposition 2.19. *Si $A \simeq W(R)/(\xi)$ (avec $R \in \mathrm{Perf}_{\mathbf{F}_p}$ et ξ distingué) est un anneau perfectoïde alors $A/A[p^\infty] \simeq W(R/\mathrm{Ann}(\xi(0)))/\xi$ l'est aussi.*

Démonstration. Rappelons que $A[p^\infty] = A[p]$ (proposition 2.8). Écrivons $\xi = [\xi(0)] + pu$, avec u inversible dans $W(R)$ et notons que $S := R/\mathrm{Ann}(\xi(0))$ est parfait (en effet, si $x^p\xi(0) = 0$ alors $(x\xi(0))^p = 0$, donc $x\xi(0) = 0$). La projection canonique $f\colon R \to S$ induit un morphisme $f\colon W(R) \to W(S)$. Si $a \in A[p]$ se relève en $x \in W(R)$, il existe $y \in W(R)$ tel que $px = \xi y$. On a $f(y) \in pW(S)$ puisque $\xi(0)y(0) = 0$ (donc $f(y(0)) = 0$). On en déduit que $f(x) \in (\xi)W(S)$ et que le morphisme $A \to W(S)/(\xi)$ se factorise en un morphisme $A/A[p] \to W(S)/(\xi)$, clairement surjectif. Il est aussi injectif : si $a \in A$ se relève en $x \in W(R)$ et a une image nulle dans $W(S)/(\xi)$, alors $f(x) = \xi y$ pour un $y \in W(S)$. Soit $z \in W(R)$ un relèvement de y, alors $[\xi(0)](x - \xi z) = 0$, puis $(\xi - pu)(x - \xi z) = 0$ et $\xi \mid px$, donc $a \in A[p]$. □

Le résultat suivant sera systématiquement utilisé par la suite. Sa preuve est assez astucieuse.

Proposition 2.20. *Tout anneau perfectoïde A est réduit et, pour tout élément p-puissant (déf. 2.2) $a \in A$, la torsion a-primaire $A[a^\infty]$ de A est tuée par $(a^{p^{-\infty}})$.*

Démonstration. Le second point est une conséquence formelle du premier : si $a^n x = 0$ pour un $x \in A$, alors $(a^{n/p}x)^p = 0$, donc $a^{n/p}x = 0$ puisque A est réduit. Soit $\pi \in A$ un élément p-puissant tel que $(\pi^p) = (p)$ (exemple 2.7). Supposons d'abord que A est sans p-torsion. Si $a \in A$ vérifie $a^p = 0$, l'isomorphisme $A/\pi \simeq A/\pi^p$ montre que $a = \pi b$ pour un $b \in A$. Comme $(\pi^p) = (p)$ et A est sans p-torsion, on a $b^p = 0$. Par itération cela force $a \in \cap_n \pi^n A = \cap_n p^n A = \{0\}$, ce qui permet de conclure.

Dans le cas général, par la proposition 2.19 et le caractère réduit (même parfait !) de $A/(\pi^{p^{-\infty}})$, il suffit de prouver l'injectivité du morphisme naturel $A \to A/A[p] \times A/(\pi^{p^{-\infty}})$. Soit $a \in A[p] \cap (\pi^{p^{-\infty}})$ et soit $x \in W(R)$ un représentant de a. Comme $a \in (\pi^{p^{-\infty}})$, on a $x(0) \in (\xi(0)^{p^{-\infty}})$. Soit $R = A^\flat$, donc $A \simeq W(R)/(\xi)$. Puisque $pa = 0$, il existe $y \in W(R)$ tel que $px = \xi y$. On a donc $\xi(0)y(0) = 0$ et $\xi'(0)y(0) + \xi(0)y'(0) = x(0)$, donc $y(0) \in (\xi(0)^{p^{-\infty}})$. Mais $\xi(0) \in \mathrm{Ann}(y(0))$ et R est parfait, donc $(\xi(0)^{p^{-\infty}}) \subset \mathrm{Ann}(y(0))$, ce qui force $y(0)^2 = 0$, puis $y(0) = 0$. On a donc $\xi \mid x$ et $a = 0$. □

3. Algèbres perfectoïdes : aspects analytiques

On fixe dans cette section (qui emprunte les notations de la précédente) un nombre premier p (d'où une notion d'anneau perfectoïde), ainsi qu'un corps K muni d'une valeur absolue non archimédienne $|\cdot|\colon K \to \mathbf{R}_{\geqslant 0}$, dont la boule unité K^0 est un anneau perfectoïde. Comme K^0 est p-complet (proposition 2.8), on a $p \in \mathrm{Rad}(K^0)$ et donc $|p| < 1$. On fixe un élément $\pi \in K^0$ comme suit :

- si $\mathrm{car}(K) = p$ on prend n'importe quel élément non nul $\pi \in K^0$ tel que $|\pi| < 1$.
- sinon, on choisit un générateur distingué $\xi \in W((K^0)^\flat)$ de $\ker(\theta_{K^0})$ et on note

$$\pi = \theta_{K^0}([\xi(0)^{1/p}]) = (\xi(0)^{1/p})^\sharp \in K^0.$$

Ainsi $\pi^p K^0 = pK^0$ (exemple 2.7), donc $|\pi| < 1$. L'élément π muni du système de racines $(\pi^{1/p^n} := (\xi(0)^{1/p^{n+1}})^\sharp)_{n\geqslant 0}$ est p-puissant (définition 2.2).

Dans les deux cas, $K^0/\pi \simeq K^0/\pi^p$ via le Frobenius, π est p-puissant et $|\pi| < 1$.

Si A est une K^0-algèbre plate (*i.e.* sans π-torsion) on note

$$A_* = \pi^{-1/p^\infty} A := \{f \in A[\frac{1}{\pi}] \mid \pi^{1/p^n} f \in A \text{ pour tout } n \geqslant 0\},$$

une sous K^0-algèbre de $A[\frac{1}{\pi}]$ contenant A et contenue dans $\pi^{-1}A$.

Remarque 3.1. On voit facilement que $(A_*)_* = A_*$ et que si A est π-complète (respectivement p-close (définition 2.15) dans $A[\frac{1}{\pi}]$), alors A_* l'est aussi. Si $(A_i)_{i\in I}$ est un système projectif de K^0-algèbres plates, alors $(\varprojlim_{i\in I} A_i)_* \simeq \varprojlim_{i\in I}(A_i)_*$.

3.1. Algèbres de Banach uniformes

On note Ban_K la catégorie des *K-algèbres de Banach*, *i.e.* celle des K-algèbres A munies d'une norme ultramétrique $|\cdot|\colon A \to \mathbf{R}_{\geqslant 0}$ sur le K-espace vectoriel A telle que $|ab| \leqslant |a||b|$ pour tous $a, b \in A$ (et $|1| = 1$ si $A \neq 0$) et qui fait de A un espace métrique complet pour la distance induite. Les morphismes dans Ban_K sont les applications lipschitziennes qui sont aussi des morphismes de K-algèbres.

Berkovich [(33)] a construit un foncteur de Ban_K vers les espaces topologiques compacts, en associant à $(A, |\cdot|) \in \mathrm{Ban}_K$ son *spectre de Berkovich* $\mathscr{M}(A)$, *i.e.* l'ensemble des semi-normes multiplicatives [(34)] $x\colon A \to \mathbf{R}_{\geqslant 0}$ telles que $x(f) \leqslant |f|$ pour tout $f \in A$, muni de la topologie la plus faible rendant continues les évaluations $x \mapsto x(f)$ pour $f \in A$. On écrira $|f(x)|$ au lieu de $x(f)$.

[(33)] On pourrait (ou devrait...) remplacer les espaces de Berkovich par ceux de Huber dans ce qui suit. Nous avons fait ce choix pour épargner au lecteur les multiples définitions intervenant dans la théorie de Huber, qui ne joueront pas de rôle sérieux dans cet exposé.

[(34)] On demande donc que $x(f+g) \leqslant \max(x(f), x(g)), x(fg) = x(f)x(g)$ et $x(1) = 1$ si $A \neq 0$.

Soit $(A, |\cdot|) \in \mathrm{Ban}_K$. Le sous-ensemble

$$A^0 = \{f \in A \mid \sup_{n \geqslant 1} |f^n| < \infty\}$$

est une sous K^0-algèbre de A, contenant la boule unité de A, en particulier $A^0[\frac{1}{\pi}] = A$. On écrira A^0_* au lieu de $(A^0)_*$. Si l'on note

$$|f|_{\mathrm{sp}} = \lim_{n \to \infty} |f^n|^{1/n} = \inf_{n \geqslant 1} |f^n|^{1/n}$$

la *semi-norme spectrale* d'un élément $f \in A$, on dispose des formules fondamentales [(35)]

$$|f|_{\mathrm{sp}} = \max_{x \in \mathscr{M}(A)} |f(x)| \text{ et } A^0_* = \{f \in A \mid |f|_{\mathrm{sp}} \leqslant 1\}.$$

On dit que $(A, |\cdot|) \in \mathrm{Ban}_K$ est *uniforme* si A^0 est une partie bornée de A. Cela arrive si et seulement si $|\cdot|$ est équivalente à $|\cdot|_{\mathrm{sp}}$ (utiliser les inclusions $\{f \in A \mid |f|_{\mathrm{sp}} < 1\} \subset A^0 \subset \{f \in A \mid |f|_{\mathrm{sp}} \leqslant 1\}$), auquel cas [(36)]

$$A^0 = A^0_* = \{f \in A \mid |f|_{\mathrm{sp}} \leqslant 1\} = \{f \in A \mid |f(x)| \leqslant 1, \text{ pour tout } x \in \mathscr{M}(A)\} \quad (3).$$

On note Ban^u_K la sous-catégorie pleine de Ban_K des K-algèbres de Banach uniformes.

On dit que A est *spectrale* si $|f| = |f|_{\mathrm{sp}}$ pour tout $f \in A$, ce qui arrive si et seulement si A^0 est la boule unité de A, auquel cas A est uniforme.

3.2. Dictionnaire analyse-algèbre, interprétation catégorique

Le lecteur trouvera un dictionnaire exhaustif analyse-algèbre dans le paragraphe 2.3 de l'article d'André (2018a).

Proposition 3.2. *Soit $(A, |\cdot|) \in \mathrm{Ban}^u_K$. La K^0-algèbre A^0 est plate, π-complète et p-close (déf. 2.15) dans $A = A^0[\frac{1}{\pi}]$. Ainsi le Frobenius $A^0/\pi \to A^0/\pi^p$ est injectif.*

Démonstration. Le seul point non évident est le fait que A^0 est p-close, mais il suffit de contempler l'égalité (3) pour s'en convaincre. □

Soit $\mathscr{C}$ la catégorie des K^0-algèbres plates et π-complètes S, telles que S soit p-close dans $S[\frac{1}{\pi}]$ (définition 2.15) et $S_* = S$. Pour $S \in \mathscr{C}$ et $f \in A := S[\frac{1}{\pi}]$ on pose

$$|f| := \inf\{|k| \mid k \in K \setminus \{0\}, \frac{f}{k} \in S\}.$$

[(35)] La première est la *formule du rayon spectral* de Berkovich ; la seconde découle facilement des inclusions évidentes $\{f \in A \mid |f|_{\mathrm{sp}} < 1\} \subset A^0 \subset \{f \in A \mid |f|_{\mathrm{sp}} \leqslant 1\}$.

[(36)] Pour la première égalité noter que si $f \in A^0_*$ alors f^{p^n} reste dans le borné $p^{-1}A^0$ pour tout n.

Alors [37] $(A, |\cdot|)$ devient une K-algèbre de Banach spectrale (donc uniforme) telle que $A^0 = S$. On obtient ainsi :

Proposition 3.3. *Le foncteur $(A, |\cdot|) \mapsto A^0$ induit une équivalence de catégories*

$$\mathrm{Ban}_K^u \simeq \mathscr{C}.$$

La remarque 3.1 montre que $\mathscr{C}$ possède toutes les petites limites (qui se calculent comme dans la catégorie des K^0-algèbres) et toutes les colimites filtrantes : si $(S_i)_{i\in I}$ est un système inductif filtrant dans $\mathscr{C}$, sa colimite dans $\mathscr{C}$ est S_*, où S est le complété π-adique de la colimite dans la catégorie des anneaux $\varinjlim_{i\in I} S_i$.

Ainsi la catégorie Ban_K^u possède toutes les petites limites, et aussi toutes les colimites filtrantes. L'inclusion de Ban_K^u dans Ban_K admet un adjoint à gauche $A \mapsto A^u$, l'uniformisé A^u de A étant le séparé complété de A pour la semi-norme spectrale. Comme Ban_K possède des pushouts (donnés par le produit tensoriel complété), il en est de même de Ban_K^u : si $A \to B$ et $A \to C$ sont des morphismes dans Ban_K^u, le pushout correspondant est $B\widehat{\otimes}_A^u C := (B\widehat{\otimes}_A C)^u$. Le calcul de $(B\widehat{\otimes}_A^u C)^0$ n'est pas aisé, mais on verra qu'il est (presque) faisable dans le monde parfait des algèbres de Banach perfectoïdes.

3.3. Algèbres de Banach perfectoïdes

Il convient de garder en tête qu'une *K-algèbre de Banach perfectoïde* n'est *pas* la même chose qu'une K-algèbre qui est un anneau perfectoïde :

Définition 3.4. Une *K-algèbre de Banach perfectoïde* est un objet $(A, |\cdot|)$ de Ban_K^u tel que A^0 soit un anneau perfectoïde. On note Perf_K la sous-catégorie pleine de Ban_K^u dont les objets sont les K-algèbres de Banach perfectoïdes.

Vérifions la compatibilité avec la définition originelle dans Scholze (2012) :

Proposition 3.5. *Une K^0-algèbre plate et π-complète B est un anneau perfectoïde si et seulement si le Frobenius $B/\pi \to B/\pi^p$ est un isomorphisme, auquel cas B_* est un anneau perfectoïde et B est p-close dans $B[\frac{1}{\pi}]$ (déf. 2.15). En particulier $A \in \mathrm{Ban}_K^u$ est dans Perf_K si et seulement si le Frobenius $A^0/\pi \to A^0/\pi^p$ est un isomorphisme ou, de manière équivalente [38], surjectif.*

Démonstration. On peut supposer que $\mathrm{car}(K) = 0$. Si B est un anneau perfectoïde, le théorème 2.10 montre que $B \simeq W(B^\flat)/(\xi)$. Puisque $\pi = \theta_{K^0}([\xi(0)^{1/p}])$, le Frobenius $B/\pi \to B/\pi^p$ est un isomorphisme et donc (proposition 2.16) B est p-close dans

[37] Le fait que S est plate, π-complète et vérifie $S = S_*$ est suffisant pour montrer que $|\cdot|$ est une norme de K-algèbre de Banach sur A, de boule unité S. On déduit de la relation $S = \{f \in A | f^p \in S\}$ que $|f^p| = |f|^p$, puis que $|f| = \lim_{n\to\infty} |f^{p^n}|^{1/p^n} = |f|_{\mathrm{sp}}$ pour tout $f \in A$, donc A est bien spectrale.

[38] Puisque A est uniforme, A^0 est p-close dans A, donc le Frobenius $A^0/\pi \to A^0/\pi^p$ est injectif.

$B[\frac{1}{\pi}]$. La réciproque découle de la proposition 2.12. Supposons que B est un anneau perfectoïde. Comme B_* est plate et π-complète, il suffit de voir (d'après ce que l'on vient de faire) que le Frobenius $B_*/\pi \to B_*/\pi^p$ est un isomorphisme. Mais B étant p-close dans $B[\frac{1}{\pi}]$, B_* l'est dans $B_*[\frac{1}{\pi}]$, donc $B_*/\pi \to B_*/\pi^p$ est injectif. Soit $x \in B_*$, alors $\pi x \in B$ donc $\pi x = a^p + \pi^p b$ avec $a, b \in A$. Puisque $x = (\pi^{-1/p}a)^p + \pi^{p-1}b$ et B_* est p-close dans $B[\frac{1}{\pi}]$, on a $\pi^{-1/p}a \in B_*$, donc $B_* = B_*^p + \pi B_*$. Par itération et en utilisant le fait que $\pi^p \mid p$, on obtient $B_* = B_*^p + \pi^p B_*$, ce qui permet de conclure. □

Remarque 3.6. 1. Le caractère perfectoïde de B_* n'est pas totalement gratuit : si $A \in \mathrm{Perf}_K$ et si $g \in A^0$ n'est pas un diviseur de zéro, il n'est pas connu (*cf.* la question 3.5.1 d'André (2018a)) si l'algèbre $g^{-1/p^\infty}A = \cap_{n\geqslant 1} g^{-1/p^n}A \subset A[\frac{1}{g}]$ est encore dans Perf_K. Cette algèbre interviendra naturellement par la suite (*cf.* théorème 5.1 par exemple).

2. Supposons que $\mathrm{car}(K) = 0$ et posons $K^\flat := (K^0)^\flat[\frac{1}{\xi(0)}]$, un corps perfectoïde de caractéristique p. En combinant la proposition ci-dessus et le théorème 2.10, on obtient facilement l'équivalence de catégories $\mathrm{Perf}_K \simeq \mathrm{Perf}_{K^\flat}$ de Scholze (2012), induite par les foncteurs $A \mapsto A^\flat := (A^0)^\flat[\frac{1}{\xi(0)}]$ et $R \mapsto W(R)[\frac{1}{\pi}]/(\xi)$ (cette approche est celle de Fontaine (2013) et du livre de Kedlaya et Liu (2015)).

3. La catégorie Perf_K possède toutes les colimites filtrantes, qui sont les mêmes que celles dans Ban_K^u : si $(B_i)_{i\in I}$ est un système inductif filtrant dans Perf_K alors $A := (\widehat{\varinjlim_{i\in I} B_i^0})_*$ est perfectoïde (proposition 3.5). L'équivalence $\mathrm{Perf}_K \simeq \mathrm{Perf}_{K^\flat}$ et la description explicite [(39)] de $\mathrm{Perf}_{K^\flat}$ montrent que Perf_K possède aussi toutes les petites limites, *mais si* $\mathrm{car}(K) = 0$ *elles ne sont pas les mêmes que celles dans* Ban_K^u, *et cela même pour les limites finies.* Voir les exemples prophylactiques 3.8.2 et 3.8.5 d'André (2018a). Le point 1) est un autre exemple de difficulté posée par la différence entre les limites dans Ban_K^u et celles dans Perf_K quand $\mathrm{car}(K) = 0$.

3.4. Le lemme d'approximation de Scholze

Nous aurons besoin du résultat technique mais fondamental ci-dessous. La preuve originelle (Scholze, 2012 dans un contexte un peu différent) est assez technique, voir aussi la preuve du lemme 2.3.1 dans Česnavičius et Scholze (2019), qui reprend un argument de Kedlaya.

Si A est un anneau et $\pi \in A$, on note $\mathrm{Spa}_\pi(A)$ l'ensemble des valeurs absolues [(40)] $|\cdot| : A \to \Gamma \cup \{0\}$, où Γ est un groupe abélien totalement ordonné (noté multiplicativement), telles que $|f| \leqslant 1$ pour tout $f \in A$ et $\lim_{n\to\infty} |\pi^n| = 0$. Si π est non diviseur

[(39)] Il s'agit de la sous-catégorie pleine de $\mathrm{Ban}_{K^\flat}^u$ des $K^\flat$-algèbres de Banach parfaites.

[(40)] On demande que $|ab| = |a||b|$, $|a+b| \leqslant \max(|a|, |b|)$, pour $a, b \in A$, $|0| = 0$ et $|1| = 1$ si $A \neq \{0\}$.

de zéro on note $\mathrm{Spa}(A[\frac{1}{\pi}], A)$ le sous-ensemble des $|\cdot| \in \mathrm{Spa}_\pi(A)$ qui se prolongent en une valeur absolue sur $A[\frac{1}{\pi}]$.

Théorème 3.7 (lemme d'approximation). *Soit A un anneau perfectoïde et soit $\pi \in A$ tel que A soit π-complet et $\pi^p \mid p$. Pour tous $f \in A$ et $n \in \mathbf{N}$, il existe $g \in A$ p-puissant (déf. 2.2) tel que*

$$|f - g| \leqslant |p| \cdot \max(|g|, |\pi|^n), \text{ pour tout } |\cdot| \in \mathrm{Spa}_\pi(A);$$

en particulier $|f| \leqslant |\pi|^n$ si et seulement si $|g| \leqslant |\pi|^n$.

Proposition 3.8. *Soient A un anneau, $\pi \in A$, non diviseur de zéro et $x \in A[\frac{1}{\pi}]$.*

a) Si $|x| < 1$ pour tout $|\cdot| \in \mathrm{Spa}(A[\frac{1}{\pi}], A)$, alors il existe $n \geqslant 1$ tel que $x^n \in \pi A$.

b) Si $y \in A[\frac{1}{\pi}]$ vérifie $|x - y| < 1$ pour tout $|\cdot| \in \mathrm{Spa}(A[\frac{1}{\pi}] = A[x][\frac{1}{\pi}], A[x])$, alors y est dans la p-clôture (déf. 2.15) de $A[x]$ dans $A[\frac{1}{\pi}]$.

Démonstration. Le point a) est standard, voir le lemme 2.3.2 de Česnavičius et Scholze (2019) pour la preuve. Le b) s'en déduit [41]. □

3.5. Presque rappels de presque algèbre

On trouve dans le livre de Gabber et Ramero (2003) un exposé exhaustif de la théorie des presque mathématiques [42] de Faltings, et un excellent résumé des points essentiels dans la section 1 de l'article d'André (2018a).

On fixe un anneau V muni d'un élément p-puissant (définition 2.2) g, d'où un idéal idempotent $\mathfrak{m} = (g^{p^{-\infty}})$ (on dira aussi, simplement et abusivement, que l'on se place *dans le cadre* g^{1/p^∞}). La proposition 2.1.7 de Gabber et Ramero (2003) montre que $\mathfrak{m} \otimes_V \mathfrak{m}$ est un V-module plat (cela est évident si V est sans g-torsion), donc les résultats de Gabber et Ramero (2003) s'appliquent.

Définition 3.9. Soit $f\colon M \to N$ un morphisme de V-modules. On dit que

- M est *presque nul* si $\mathfrak{m}M = 0$;
- f est *presque injectif* (*resp. presque surjectif, resp. un presque isomorphisme*) si $\ker f$ (resp. $\mathrm{coker} f$, resp. les deux) est presque nul;
- M est *presque plat* si $M \otimes_V N \to M \otimes_V P$ est presque injectif pour tout morphisme injectif $N \to P$ de V-modules, ce qui équivaut à la presque nullité de $\mathrm{Tor}_i^A(M, N)$ pour tous $i > 0$ et $N \in \mathrm{Mod}_V$. On définit de manière analogue la notion de module *presque projectif*;
- M est *presque fidèlement plat* si M est presque plat et si la presque nullité de $M \otimes_V N$ entraîne celle de N pour tout $N \in \mathrm{Mod}_V$;
- M est *presque de type fini* si pour tout n il existe un sous V-module de type fini $N \subset M$ tel que $g^{1/p^n} M \subset N$.

[41] Par a) il existe $d \geqslant 0$ tel que $(x - y)^{p^d} \in \pi A[x]$.

[42] Qui semble être une manière d'écrire des $o(1)$ sans l'avouer...

Si W est une V-algèbre, g reste p-puissant dans W, donc on peut appliquer les définitions ci-dessus aux W-modules, d'où une notion de W-module presque plat, etc. Cette remarque est utilisée dans la définition suivante :

Définition 3.10. Soit $g\colon A \to B$ un morphisme de V-algèbres. On dit que

• g est *presque étale* si B est un A-module presque plat et un $B \otimes_A B$-module (via la multiplication $B \otimes_A B \to B$) presque projectif ;

• g est *presque fini étale* si g est presque étale et si B est un A-module presque de type fini et presque projectif.

3.6. Miracles perfectoïdes

Les résultats suivants de Scholze (2012) sont pour le moins surprenants : leurs analogues dans le monde des algèbres affinoïdes usuelles sont totalement faux. Dans le monde perfectoïde c'est (presque) tous les jours le printemps ...

Théorème 3.11. *Soit $A \in \mathrm{Perf}_K$ et soient $A \to B$ et $A \to C$ deux morphismes dans Perf_K. Alors $B\widehat{\otimes}_A C \in \mathrm{Perf}_K$ et le morphisme naturel* [43]

$$B^0\widehat{\otimes}_{A^0}C^0 \to (B\widehat{\otimes}_A C)^0$$

est un presque isomorphisme (dans le cadre π^{1/p^∞}).

Démonstration. On peut supposer que $\mathrm{car}(K) = 0$. Les anneaux $R = B^0\widehat{\otimes}_{A^0}C^0$ et $S := R/R[p^\infty]$ sont perfectoïdes (corollaire 2.11 et proposition 2.19). La proposition 3.5 montre que $S_* \in \mathscr{C}$ (*cf.* § 3.2), donc $T := S[1/\pi] = S_*[1/\pi]$ a une structure canonique de K-algèbre de Banach spectrale telle que $T^0 = S_*$, et on a $T \in \mathrm{Perf}_K$. Comme B et C sont uniformes, les morphismes évidents $B^0 \to R \to S$ et $C^0 \to R \to S$ s'étendent en des morphismes (dans Ban_K) $B \to T$ et $C \to T$, d'où un morphisme $B\widehat{\otimes}_A C \to T$. On construit un inverse comme suit. Soit X la boule unité de $B\widehat{\otimes}_A C$. Comme B et C sont uniformes, il existe $c \geqslant 1$ tel que les images de B^0 et C^0 dans $B\widehat{\otimes}_A C$ soient contenues dans $p^{-c}X$. Le morphisme induit $S \to B\widehat{\otimes}_A C$ a une image contenue dans $p^{-2c}X$ et s'étend ainsi en un morphisme $T \to B\widehat{\otimes}_A C$, inverse du morphisme $B\widehat{\otimes}_A C \to T$.

On en déduit que $B\widehat{\otimes}_A C \in \mathrm{Perf}_K$ et que $(B\widehat{\otimes}_A C)^0$ est isomorphe à $T^0 = S_*$. Le morphisme $R \to S \to S_*$ est un presque isomorphisme, puisque l'idéal $(\pi^{p^{-\infty}})$ annule le noyau et le conoyau de $R \to S$ (proposition 2.20) et aussi de $S \to S_*$. □

Si $B \in \mathrm{Ban}_K$ et si $f_1, \ldots, f_n, g \in B$ engendrent l'idéal unité on dispose d'une B-algèbre de Banach universelle $B\langle \frac{f_1,\ldots,f_n}{g}\rangle$ dans laquelle g est inversible et les $\frac{f_i}{g}$ sont à puissances bornées. Explicitement, $B\langle \frac{f_1,\ldots,f_n}{g}\rangle$ est le quotient de l'algèbre de Banach $B\langle T_1, \ldots, T_n\rangle = B\widehat{\otimes}_K K\langle T_1, \ldots, T_n\rangle$ par *l'adhérence* de l'idéal engendré par les $gT_i - f_i$.

[43] La complétion à gauche est π-adique.

Théorème 3.12. *Soit $A \in \mathrm{Perf}_K$ et soient $f_1, \dots, f_n, g \in A$ qui engendrent l'idéal unité de A.*

a) On a $A\langle \frac{f_1,\dots,f_n}{g} \rangle \in \mathrm{Perf}_K$.

b) Si $f_1, \dots, f_n, g$ sont p-puissants (déf. 2.2) dans A^0 et si $A^0\langle (\frac{f_1}{g})^{1/p^\infty}, \dots, (\frac{f_n}{g})^{1/p^\infty} \rangle$ est le complété π-adique de $A^0[\frac{f_i^{1/p^n}}{g^{1/p^n}},\ n \geqslant 0]$, alors le morphisme naturel

$$A^0\Big\langle (\frac{f_1}{g})^{1/p^\infty}, \dots, (\frac{f_n}{g})^{1/p^\infty} \Big\rangle \to A\langle \frac{f_1, \dots, f_n}{g} \rangle^0$$

est un presque isomorphisme (dans le cadre π^{1/p^∞}).

Démonstration. Voir la section 6 de Scholze (2012) pour les détails. Le très joli argument de « réduction au cas universel » ci-dessous est dû à André (2018a). Le point a) se déduit du point b) et du théorème 3.7. Puisque $f_i, g \in A^0$ sont p-puissants, on dispose d'un morphisme $S := K^0\langle T_1^{1/p^\infty}, \dots, T_n^{1/p^\infty}, U^{1/p^\infty} \rangle[\frac{1}{\pi}] \to A$ dans Ban_K, envoyant T^{1/p^j} sur f_i^{1/p^j} et U^{1/p^j} sur g^{1/p^j}. Soit $N \geqslant 1$ tel que $\pi^N \in (f_1, \dots, f_n, g) \subset A^0$. Alors $A \widehat{\otimes}_S S\langle \frac{\pi^N, T_1, \dots, T_n}{U} \rangle$ et $A\langle \frac{f_1,\dots,f_n}{g} \rangle$ partagent la même propriété universelle, donc ces algèbres de Banach sont isomorphes. Le théorème 3.11 ramène donc la preuve au « cas universel », *i.e.* à vérifier que $S, S\langle \frac{\pi^N, T_1, \dots, T_n}{U} \rangle \in \mathrm{Perf}_K$ et que les morphismes $K^0\langle T_1^{1/p^\infty}, \dots, T_n^{1/p^\infty}, U^{1/p^\infty} \rangle \to S^0$ et $S^0\langle (\frac{T_1}{U})^{1/p^\infty}, \dots, (\frac{T_n}{U})^{1/p^\infty}, (\frac{\pi^N}{U})^{1/p^\infty} \rangle \to S\langle \frac{\pi^N, T_1, \dots, T_n}{U} \rangle^0$ sont des presque isomorphismes, ce qui se fait par un calcul direct (assez pénible...). □

Le résultat fondamental suivant, sur lequel tout repose dans la section 5, est le *théorème de presque pureté* de Faltings, étendu par Scholze (2012) et Kedlaya et Liu (2015). Voir la section 7 de Scholze (2012) pour la preuve (fort délicate) et le paragraphe 3.4 d'André (2018a) pour des compléments.

Théorème 3.13. *Soit $(A, |\cdot|) \in \mathrm{Perf}_K$. Pour toute A-algèbre finie étale B il existe une norme de K-algèbre de Banach $||\cdot||$ sur B telle que le morphisme $(A, |\cdot|) \to (B, ||\cdot||)$ soit continu, $B \in \mathrm{Perf}_K$ et B^0 soit presque fini étale de A^0 (dans le cadre π^{1/p^∞}). Si $A \to B$ est injectif, B^0 est presque fidèlement plate sur A^0.*

4. Le lemme de platitude d'André

Cette section est consacrée au premier pilier des travaux d'André, son lemme de platitude, un énoncé fondamental qui, loin d'être une platitude, sera systématiquement utilisé dans la preuve des résultats principaux de ce rapport. On trouvera des raffinements et des généralisations, ainsi que des applications spectaculaires dans les travaux de Bhatt et Scholze (2022), Česnavičius et Scholze (2019), Dine (2022) et

dans le livre de Gabber et Ramero (2018). En particulier, on trouvera dans le paragraphe 7.3 de l'article de Bhatt et Scholze (2022) une preuve purement algébrique (mais pas facile), qui n'utilise pas la théorie des espaces perfectoïdes.

On garde les notations et les conventions introduites au début de la section 3.

4.1. Le lemme de presque platitude

Le problème qui nous occupera dans ce paragraphe est le suivant : on se donne $A \in \mathrm{Perf}_K$ et $g \in A^0$, et on cherche à construire une A-algèbre $B \in \mathrm{Perf}_K$ dans laquelle g est p-puissant (définition 2.2) et telle que B^0/π soit presque fidèlement plate sur A^0/π (dans le cadre π^{1/p^∞} (§ 3.5)).

L'idée naïve est de regarder la A-algèbre de Banach universelle munie d'un système compatible de p^∞-racines de g, *i.e.* $A\langle T^{1/p^\infty}\rangle/\overline{(T-g)}$. L'algèbre $A\langle T^{1/p^\infty}\rangle$ est perfectoïde (noter que $A\langle T^{1/p^\infty}\rangle^0 = A^0\langle T^{1/p^\infty}\rangle$ et utiliser la remarque 2.7), donc uniforme, mais en général l'uniformité ne se propage pas aux quotients. On insiste et on la remplace par son uniformisé (§ 3.2)

$$A\langle g^{1/p^\infty}\rangle := \left(A\langle T^{1/p^\infty}\rangle/\overline{(T-g)}\right)^u.$$

Le résultat un peu miraculeux d'André (2018b) [(44)] est que cela résout notre problème :

Théorème 4.1. *On a $A\langle g^{1/p^\infty}\rangle \in \mathrm{Perf}_K$ et $A\langle g^{1/p^\infty}\rangle^0/\pi$ est presque fidèlement plate sur A^0/π (dans le cadre π^{1/p^∞}).*

On a un énoncé analogue modulo π^d pour tout $d \geqslant 1$, se déduisant formellement du théorème. La preuve se fait en deux étapes, détaillées ci-dessous : on relie $A\langle g^{1/p^\infty}\rangle$ à des localisés de l'algèbre perfectoïde $A\langle T^{1/p^\infty}\rangle$, puis on analyse ces localisés en utilisant de manière intensive les résultats des § 3.4 et 3.6.

4.1.1. *Une autre description de $A\langle g^{1/p^\infty}\rangle$.* — Soient $B \in \mathrm{Ban}_K^u$, $f \in B$ et $I = \overline{fB} \subset B$. Il convient de voir l'algèbre $(B/I)^u \in \mathrm{Ban}_K^u$ comme celle des fonctions analytiques sur le fermé Zariski $V(I)$ de $\mathscr{M}(B)$. Comme $x \in \mathscr{M}(B)$ est annulé par f si et seulement si $|f(x)| \leqslant |\pi^n(x)|$ pour tout $n \geqslant 1$, $(B/I)^u$ n'est rien d'autre que la limite inductive $\varinjlim_n B_n$ (dans la catégorie Ban_K^u, *cf.* § 3.2) des algèbres

$$B_n := (B\langle f/\pi^n\rangle)^u.$$

En effet, pour $S \in \mathrm{Ban}_K^u$ la donnée d'un morphisme (dans Ban_K) de $\varinjlim_n B_n$ dans S équivaut à celle d'un morphisme $\varphi\colon B \to S$ tel que $\varphi(f/\pi^n) \in S^0$ pour tout $n \geqslant 1$, or cela signifie précisément que φ se factorise par $(B/I)^u$ (puisque S est uniforme).

[(44)] L'énoncé suivant est tiré de l'article de Bhatt (2018).

Prenons maintenant $B = A\langle T^{1/p^\infty}\rangle$ et $f = T - g$ dans la discussion ci-dessus. Puisque B est perfectoïde, chacune des algèbres $B\langle f/\pi^n\rangle$ l'est (théorème 3.12), donc $B_n = B\langle f/\pi^n\rangle$ et $C := A\langle g^{1/p^\infty}\rangle \in \mathrm{Perf}_K$, en tant que colimite filtrante (dans Ban_K^u) d'algèbres perfectoïdes (remarque 3.6). Comme $C^0 = (\widehat{\varinjlim_n B_n^0})_*$ est presque isomorphe à $\widehat{\varinjlim_n B_n^0}$, l'algèbre C^0/π est presque isomorphe à $\varinjlim_n B_n^0/\pi$. Il suffit donc de voir que chaque B_n^0/π est presque fidèlement plate sur A^0/π. On fixe par la suite $n \geqslant 1$.

4.1.2. *Analyse des localisés de $A\langle T^{1/p^\infty}\rangle$ et fin de la preuve.* — Par le lemme d'approximation (théorème 3.7) il existe un élément p-puissant $h \in B^0$ tel que $|h - f| \leqslant |p| \cdot \max(|h|, |\pi^n|)$ pour tout $|\cdot| \in \mathrm{Spa}_\pi(B^0)$. Comme $|\frac{h-f}{\pi}| < 1$ pour tout $|\cdot| \in \mathrm{Spa}(B^0[\frac{1}{\pi}], B^0)$ et B^0 est p-close (définition 2.15) dans $B^0[\frac{1}{\pi}]$ (proposition 3.2), la proposition 3.8 montre que

$$h \equiv f \pmod{\pi B^0}.$$

L'inégalité $|h - f| \leqslant |p| \cdot \max(|h|, |\pi^n|)$ montre aussi que $B_n \simeq B\langle \frac{h}{\pi^n}\rangle$, et on déduit du théorème 3.12 que B_n^0/π est presque isomorphe à $\varinjlim_j C_j/\pi$, avec $C_j = B^0[\frac{h^{1/p^j}}{\pi^{n/p^j}}] \subset B$. Nous allons montrer que chaque C_j/π est fidèlement plate sur A^0/π, ce qui permettra de conclure. Par dévissage il suffit de montrer que $C_j/\pi^{1/p^j}$ l'est sur $A^0/\pi^{1/p^j}$.

On a $B^0 = A^0\langle T^{1/p^\infty}\rangle$, donc $B^0/\pi \simeq (A^0/\pi)[T^{1/p^\infty}]$, en particulier $B^0/\pi^{1/p^j} \simeq B^0/\pi$ via $x \mapsto x^{p^j}$ (puisqu'il en est de même du morphisme $A^0/\pi^{1/p^j} \to A^0/\pi$, *cf.* proposition 3.5) et $f = T - g$ n'est pas un diviseur de zéro modulo π. Puisque $h \equiv f \pmod{\pi B^0}$, il en est de même de h et la remarque 2.18 fournit un isomorphisme $C_j \simeq B^0[X^{1/p^j}]/(u_j)$, avec $u_j = \pi^{n/p^j} X^{1/p^j} - h^{1/p^j}$. Ainsi le morphisme $x \mapsto x^{p^j}$ et la congruence $h \equiv f \pmod{\pi B^0}$ induisent des isomorphismes

$$C_j/\pi^{1/p^j} \simeq B^0[X^{1/p^j}]/(\pi^{1/p^j}, h^{1/p^j}) \simeq$$

$$B^0[X]/(\pi, h) \simeq B^0[X]/(\pi, f) \simeq ((A^0/\pi)[T^{1/p^\infty}]/(T - g))[X],$$

exhibant $C_j/\pi^{1/p^j}$ comme une colimite filtrante de $A^0/\pi^{1/p^j} \simeq A^0/\pi$-modules libres de type fini, ce qui permet de conclure.

4.2. Le lemme de platitude

Le résultat suivant de Gabber et Ramero (2018) et Bhatt et Scholze (2022) est une vaste généralisation et un raffinement du théorème 4.1. Il a des nombreuses applications, par exemple à l'étude des sous-espaces Zariski fermés d'un espace perfectoïde, à la cohomologie prismatique (Bhatt et Scholze, 2022), etc.

Théorème 4.2 (lemme de platitude d'André)**.** *Soit A un anneau perfectoïde. Il existe une A-algèbre perfectoïde B telle que B/p soit fidèlement plate sur A/p et telle que tout $x \in B$ soit p-puissant (déf. 2.2). On peut même choisir B telle que tout polynôme unitaire $P \in B[X]$ possède une racine dans B.*

Nous ferons usage seulement [(45)] du théorème 4.3 ci-dessous, qui entraîne celui ci-dessus par une itération un peu pénible (*cf.* la preuve du théorème 2.3.4 de Česnavičius et Scholze (2019), pour les détails). Pour la preuve, on suit la présentation par Česnavičius et Scholze (2019) des arguments de Gabber et Ramero. Essentiellement, le rôle des localisés dans la preuve du théorème 4.1 est pris par les p-clôtures intégrales (définition 2.15).

Théorème 4.3. *Soit R un anneau muni d'un élément p-puissant (déf. 2.2) π, non diviseur de zéro, tel que $\pi^p \mid p$ et $R/\pi \simeq R/\pi^p$ via le Frobenius. Soit $P \in R[X] \setminus R$ unitaire et soit B la p-clôture (déf. 2.15) de $R[X^{1/p^\infty}]/P$ dans $(R[X^{1/p^\infty}]/P)[1/\pi]$. La π-complétion de B est perfectoïde, sans π-torsion, et B/π est fidèlement plate sur R/π.*

Notons que, par construction, P possède une racine p-puissante dans B, et que $R[X^{1/p^\infty}]/P$ est sans π-torsion car P est unitaire, donc non diviseur de zéro modulo π.

Démonstration. Soit $A = R[X^{1/p^\infty}]$. Comme le Frobenius $A/\pi \to A/\pi^p$ est un isomorphisme, A est p-close dans $A[\frac{1}{\pi}]$ et le complété π-adique $\hat{A}$ est perfectoïde (proposition 2.12). Soit $I = PA\left[\frac{1}{\pi}\right]$ et soit $u\colon A[\frac{1}{\pi}] \to (A/P)[\frac{1}{\pi}] \simeq A[\frac{1}{\pi}]/I$ la projection canonique.

Pour $f \in A$ et $n \geqslant 1$, notons $C_{n,f}$ la p-clôture de $A[\frac{f}{\pi^n}]$ dans $A[\frac{1}{\pi}]$, et $C_{\infty,f} = \cup_{n\geqslant 1} C_{n,f}$. Alors $I = \bigcup_{n\geqslant 1} \frac{P}{\pi^n} A = \bigcup_{n\geqslant 1} \frac{P}{\pi^n} C_{\infty,P}$ est un idéal π-divisible de $C_{\infty,P}$, et la projection u induit un isomorphisme

$$C_{\infty,P}/I \simeq B,$$

et donc des isomorphismes $B/\pi^n \simeq C_{\infty,P}/\pi^n$ pour tout $n \geqslant 1$. En effet, un élément $x \in A[\frac{1}{\pi}]$ vérifie $u(x) \in B$ si et seulement s'il existe $d \geqslant 0$ tel que $x^{p^d} \in A + I$ (proposition 2.16), ce qui revient à dire que $x \in C_{\infty,P}$. Il suffit donc de montrer que le complété π-adique de $C_{n,P}$ est perfectoïde et que $C_{n,P}/\pi$ est fidèlement plat sur R/π pour tout $n \geqslant 1$.

Puisque $\hat{A}$ est perfectoïde (proposition 2.12), le théorème 3.7 fournit un élément p-puissant $g \in \hat{A}$ tel que $|P - g| \leqslant |p| \max(|g|, |\pi^n|)$ pour tout $|\cdot| \in \mathrm{Spa}_\pi(\hat{A}) \simeq \mathrm{Spa}_\pi(A)$. Soit $(q_j)_{j\geqslant 0}$ une suite dans A telle que $q_j \equiv g^{1/p^j} \pmod{\pi^{n+p}\hat{A}}$ pour tout $j \geqslant 0$. On a

$$|P - q_0| \leqslant |\pi^p| \max(|q_0|, |\pi^n|), \text{ pour tout } |\cdot| \in \mathrm{Spa}_\pi(A). \tag{4}$$

[(45)] On pourrait même se dispenser de ce paragraphe et utiliser seulement le théorème 4.1.

Lemme 4.4. *Il existe $d \geqslant 1$ tel que $q_j^{p^j} \equiv P \pmod{\pi^{1/p^d}A}$ pour tout j, et q_j n'est pas un diviseur de zéro modulo π.*

Démonstration. La relation (4) et la proposition 3.8 montrent qu'il existe $d \geqslant 1$ tel que $(P - q_0)^{p^d} \in \pi A$, donc $P - q_0 \in \pi^{1/p^d}A$ (car A est p-clos dans $A[\frac{1}{\pi}]$). On conclut en remarquant que $q_j^{p^j} \equiv g \equiv q_0 \pmod{\pi\hat{A}}$ et que P n'est pas un diviseur de zéro modulo π. □

Lemme 4.5. *On a $C_{n,P} = \cup_{j\geqslant 0}A[\frac{q_j}{\pi^{n/p^j}}]$ et sa π-complétion est perfectoïde.*

Démonstration. Notons $x_j = \frac{q_j}{\pi^{n/p^j}} \in A[\frac{1}{\pi}]$ et $y_j = \frac{g^{1/p^j}}{\pi^{n/p^j}} \in \widehat{A}[\frac{1}{\pi}]$. Alors $x_{j+1}^p - x_j \in \pi A$ (puisque $q_{j+1}^p - q_j \in \pi^{n+p}A$) et $y_{j+1}^p = y_j$, donc $T := A[x_0, x_1, \ldots] \subset A[\frac{1}{\pi}]$ (resp. $S := \widehat{A}[y_0, y_1, \ldots] \subset \widehat{A}[\frac{1}{\pi}]$) est la réunion croissante des sous-algèbres $A[x_j]$ (resp. $\widehat{A}[y_j]$), et $T \subset C_{n,q_0}$. Comme $x_j - y_j \in \pi\hat{A}$ et q_j n'est pas un diviseur de zéro modulo π (lemme 4.4), on a des isomorphismes naturels

$$A[x_j]/\pi \simeq A[X]/(\pi, \pi^{n/p^j}X - q_j) \simeq \widehat{A}[X]/(\pi, \pi^{n/p^j}X - g^{1/p^j}) \simeq \widehat{A}[y_j]/\pi,$$

qui se propagent en un isomorphisme $T/\pi \simeq S/\pi$. Mais S est perfectoïde (proposition 2.17), donc le Frobenius $T/\pi^{1/p} \to T/\pi$ est un isomorphisme. Il s'ensuit que T est p-close dans $T[\frac{1}{\pi}] = A[\frac{1}{\pi}]$, donc $T = C_{n,q_0}$, et que le complété π-adique de T est perfectoïde. On conclut en remarquant que $C_{n,P} = C_{n,q_0}$, grâce à la relation (4) et au point b) de la proposition 3.8. □

Pour finir la preuve du théorème 4.3 il suffit de voir que chaque

$$A[q_j/\pi^{n/p^j}] \simeq R[T, X^{1/p^\infty}]/(\pi^{n/p^j}T - q_j)$$

est fidèlement plate sur R modulo π. Compte tenu du lemme 4.4, il suffit de recopier la fin de la preuve du théorème 4.1. □

5. Le lemme d'Abhyankar perfectoïde d'André

Cette section est consacrée à la preuve de l'autre pilier de la stratégie d'André : son difficile lemme d'Abhyankar perfectoïde (André, 2018a). Rappelons le contexte classique : on se donne une extension finie et plate B d'un anneau local régulier A d'inégale caractéristique $(0, p)$. On suppose que l'extension est ramifiée le long d'un diviseur à croisements normaux défini par une équation $f = 0$, avec f multiple de p, et que les indices de ramification sont premiers à p. On peut alors rendre l'extension étale en adjoignant des racines de f d'ordre divisible par tous les indices de ramification, puis en passant à la clôture intégrale (mais *sans inverser* f). Dans le cadre

perfectoïde les hypothèses concernant la ramification sont superflues, mais la conclusion reste valable seulement dans le cadre f^{1/p^∞} des presque mathématiques (§ 3.5). Le lecteur trouvera des généralisations et raffinements importants dans le livre de Gabber et Ramero (2018) (§ 16.9) et dans l'article de Bhatt et Scholze (2022) (§ 10.2).

On peut voir le lemme d'Abhyankar perfectoïde comme une généralisation au cas ramifié du théorème de presque pureté de Faltings, Scholze et Kedlaya–Liu (théorème 3.13). Ce théorème jouera d'ailleurs un rôle capital dans la preuve.

On garde les notations du début de la section 3.

5.1. Le théorème d'extension de Riemann

Le résultat suivant d'André [(46)] (inspiré par ceux de la section II de Scholze (2015)) est une version perfectoïde d'un théorème de Bartenwerfer (1976) (lui-même une version rigide analytique du théorème classique d'extension de Riemann pour les variétés complexes) : si X est un affinoïde rigide analytique *normal*, toute fonction analytique bornée sur le complémentaire d'un fermé analytique nulle part dense de X s'étend de manière unique en une fonction analytique bornée sur X. On verra que dans le monde perfectoïde l'hypothèse de normalité est superflue.

Soit $B \in \mathrm{Perf}_K$ et soit $g \in B^0$ un élément p-puissant (définition 2.2), non diviseur de zéro. Soit $B_n = B\langle \frac{\pi^n}{g} \rangle \in \mathrm{Perf}_K$ l'algèbre perfectoïde (théorème 3.12) des fonctions analytiques sur le lieu $|\pi^n| \leqslant |g|$ de $\mathscr{M}(B)$, et soit $B_\infty = \varprojlim_n B_n$ la K-algèbre de Fréchet des fonctions analytiques sur le lieu de non-annulation de g. On note

$$g^{-1/p^\infty} B^0 := \bigcap_{n \geqslant 1} g^{-1/p^n} B^0$$

le sous-anneau des $f \in B[\frac{1}{g}]$ tels que $g^{1/p^n} f \in B^0$ pour tout $n \geqslant 1$.

Théorème 5.1. *a) Le sous-anneau $g^{-1/p^\infty} B^0$ de $B[\frac{1}{g}]$ est la clôture intégrale complète [(47)] de B^0 dans $B[\frac{1}{g}]$, et $B[\frac{1}{g}]$ est intégralement clos dans B_∞.*

b) Le morphisme $B[\frac{1}{g}] \to B_\infty$ est injectif et l'image de $g^{-1/p^\infty} B^0$ est $B^0_\infty := \varprojlim_n B^0_n$.

L'injectivité du morphisme $B[\frac{1}{g}] \to B_\infty$ se ramène à celle de $B \to B_\infty$. Si $f \in B$ a une image nulle dans B_∞, alors fg est nulle sur $\mathscr{M}(B)$, donc de norme spectrale nulle ; comme B est uniforme, on a $fg = 0$ puis $f = 0$. Soit C la clôture intégrale complète de B^0 dans $B[\frac{1}{g}]$. Si $f \in g^{-1/p^\infty} B^0$ alors $f^n \in g^{-1} B^0$ pour tout n, donc $f \in C$. Dans l'autre sens, soit $f \in C$. Il existe N tel que $f^n \in \frac{1}{(\pi g)^N} B^0$ pour tout $n \geqslant 1$. Comme B^0

[(46)] La démonstration, ainsi qu'une version plus forte du théorème principal, qui sera cruciale par la suite, sont empruntées de Bhatt (2018).

[(47)] C'est-à-dire l'ensemble des $f \in B[\frac{1}{g}]$ dont les puissances sont contenues dans un sous B^0-module de type fini de $B[\frac{1}{g}]$.

est p-close (définition 2.15) dans B (proposition 3.2), on a $(\pi g)^{N/p^k} f \in B^0$ pour tout $k \geqslant 1$, donc $\pi^{1/p^\ell} g^{N/p^k} f \in B^0$ pour tous $k, \ell \geqslant 1$. Ainsi $g^{N/p^k} f \in B^0_* = B^0$ pour tout k et $f \in g^{-1/p^\infty} B^0$, montrant bien que $C = g^{-1/p^\infty} B^0$.

Si $f \in B_\infty$ est entier sur $B[\frac{1}{g}]$ alors il existe N tel que $(\pi g)^N f$ soit entier sur B^0. Mais un élément x de B_n entier sur B^0 est dans B^0_n car B_n est uniforme (en effet, les puissances de x restent dans un sous B^0_n-module de type fini de B_n, qui est borné dans B_n par uniformité). Donc $(\pi g)^N f \in B^0_\infty$. Pour finir la preuve du théorème 5.1 il suffit donc de montrer (et c'est le coeur de l'affaire) que $B[\frac{1}{g}] \to B_\infty$ identifie $g^{-1/p^\infty} B^0$ et B^0_∞. Comme $B^0 = B^0_*$, il suffit de vérifier que $(\pi g)^{1/p^k} f \in B^0$ pour tous $f \in B^0_\infty$ et $k \geqslant 1$. Cela découle par passage à la limite du résultat plus fin suivant, dû à Bhatt (2018), et qui demande quelques préliminaires.

On se place dans le cadre $(\pi g)^{1/p^\infty}$ des presque mathématiques (§ 3.5). On dit qu'un système projectif $\{M_n\}_{n\geqslant 1}$ de B^0-modules est *presque nul* si pour tous $k \geqslant 0$ et $n \geqslant 1$ il existe $m \geqslant n$ tel que $(\pi g)^{1/p^k}$ annule l'image de M_m dans M_n. Il est équivalent de dire que pour tout $k \geqslant 1$ le morphisme naturel $\{M_n[(\pi g)^{1/p^k}]\}_{n\geqslant 1} \to \{M_n\}_{n\geqslant 1}$ est un isomorphisme de pro-B^0-modules. On dira qu'un morphisme $f\colon \{M_n\}_{n\geqslant 1} \to \{N_n\}_{n\geqslant 1}$ de pro-B^0-modules est un *presque isomorphisme* si le noyau et le conoyau de f sont des systèmes projectifs presque nuls.

Théorème 5.2. *Pour tout $d \geqslant 1$ le morphisme de systèmes projectifs*

$$\{B^0/\pi^d B^0\}_{n\geqslant 1} \to \{B^0_n/\pi^d B^0_n\}_{n\geqslant 1}$$

est un presque isomorphisme dans le cadre $(\pi g)^{1/p^\infty}$.

Démonstration. Par dévissage on peut supposer que $d = 1$. Supposons d'abord que g n'est pas un diviseur de zéro modulo π. Alors chaque $f_n\colon B^0/\pi B^0 \to B^0_n/\pi B^0_n$ est injectif : si $F \in B^0$ a une image πG dans B^0_n, avec $G \in B^0_n$, alors $u := \frac{Fg}{\pi} \in B$ satisfait à [(48)] $|u|_{\mathrm{sp}} \leqslant 1$, donc $Fg \in \pi B^0$, puis $F \in \pi B^0$.

Ensuite, fixons $k, n \geqslant 1$ et montrons que $(\pi g)^{1/p^k}$ annule l'image de $\mathrm{coker}(f_{n+p^k})$ dans $\mathrm{coker}(f_n)$, ce qui permettra de conclure dans le cas où g n'est pas un diviseur de zéro modulo π. Soit $i_n\colon B^0 \to B^0_n$ le morphisme canonique. On veut montrer que si $f \in B^0_{n+p^k}$ a une image F dans B^0_n, alors $(\pi g)^{1/p^k} F \in \pi B^0_n + i_n(B^0)$. En posant $S = B^0[(\frac{\pi^{n+p^k}}{g})^{1/p^j}, j \geqslant 0] \subset B^0_{n+p^k}$, le morphisme naturel $\widehat{S} := \varprojlim_d S/\pi^d \to B^0_{n+p^k}$ est un presque isomorphisme dans le cadre π^{1/p^∞} (théorème 3.12), donc $\pi^{1/p^k} f \in S + \pi B^0_{n+p^k}$. Il suffit donc de montrer que $g^{1/p^k} H \in \pi B^0_n + i_n(B^0)$ quand

[(48)] En effet, soit $x \in \mathscr{M}(B)$, alors ou bien $|\pi(x)|^n \leqslant |g(x)|$, auquel cas $|u(x)| = |G(x)| \cdot |g(x)| \leqslant 1 \cdot 1 = 1$, ou bien $|\pi(x)|^n > |g(x)|$, auquel cas $|u(x)| \leqslant |\pi(x)|^{n-1} |F(x)| \leqslant 1 \cdot 1 = 1$.

H est l'image de $h = (\frac{\pi^{n+p^k}}{g})^e$ pour un $e \in \mathbf{Z}_{\geqslant 0}[1/p]$ arbitraire, mais cela est clair puisque $\pi^{p^k e}$ est un multiple de π pour $e \geqslant 1/p^k$ et

$$g^{1/p^k} H = \begin{cases} i_n(g^{\frac{1}{p^k}-e}\pi^{(n+p^k)e}) \in i_n(B^0), \text{ si } e < 1/p^k \\ i_n(g^{1/p^k})\pi^{p^k e}(\frac{\pi^n}{g})^e \in \pi B_n^0, \text{ sinon.} \end{cases}$$

Il reste à s'affranchir de l'hypothèse sur g. Comme $g \in B^0$ est p-puissant, on dispose d'un morphisme $R := K\langle T^{1/p^\infty}\rangle \to B$, envoyant T^{1/p^j} sur g^{1/p^j}, et $B_n \simeq B\widehat{\otimes}_R R_n$, avec $R_n = R\langle \frac{\pi^n}{T}\rangle$. Le théorème 3.11 fournit un presque isomorphisme

$$B^0/\pi \otimes_{R^0/\pi} R_n^0/\pi \simeq B_n^0/\pi,$$

et comme T n'est pas diviseur de zéro modulo π dans R^0, ce qui précède montre que le morphisme $\{R^0/\pi\} \to \{R_n^0/\pi\}_{n\geqslant 1}$ est un presque isomorphisme de systèmes projectifs, ce qui permet de conclure facilement que $\{B^0/\pi\} \to \{B_n^0/\pi\}_{n\geqslant 1}$ est un presque isomorphisme. □

5.2. Le lemme d'Abhyankar perfectoïde

On garde les notations introduites juste avant le théorème 5.1. Soit C une algèbre finie étale sur $B[1/g]$. *Si $A \to B$ est un morphisme d'anneaux, on note* $\mathrm{fi}(A,B) \subset B$ *la fermeture intégrale de A dans B.* Posons

$$\tilde{C}^0 = \mathrm{fi}(g^{-1/p^\infty}B^0, C).$$

On verra (proposition 5.6) que l'algèbre $\tilde{C} := \tilde{C}^0[\frac{1}{p}] = \mathrm{fi}(g^{-1/p^\infty}B, C)$ possède une structure naturelle de K-algèbre de Banach uniforme pour laquelle $\tilde{C}^0 = (\tilde{C})^0$ (en particulier $\tilde{C}^0$ est p-complète).

On suppose que K est de caractéristique nulle et on se place dans le cadre $(\pi g)^{1/p^\infty}$ des presque mathématiques (§ 3.5). L'un des résultats principaux de l'article d'André (2018a) s'énonce (rappelons que ξ est un générateur distingué de $\ker(\theta_{K^0})$) :

Théorème 5.3 (lemme d'Abhyankar perfectoïde).

a) Le morphisme $\theta\colon W((\tilde{C}^0)^\flat)/\xi \to \tilde{C}^0$ est injectif et un presque isomorphisme. En particulier, le Frobenius est presque surjectif sur $\tilde{C}^0/p$.

b) Pour tout $n \geqslant 1$ la B^0/p^n-algèbre $\tilde{C}^0/p^n$ est presque finie étale, et elle est presque fidèlement plate si C l'est sur $B[\frac{1}{g}]$.

Remarque 5.4. Contrairement au théorème de presque pureté, il n'est pas connu si $\tilde{C}^0$ est presque de type fini sur B^0. Si c'était le cas, on pourrait en déduire que $\tilde{C}^0$ est presque finie étale sur B^0. On peut montrer (proposition 5.2.3 d'André (2018a)) que ce dernier énoncé est tout de même vrai après inversion de g.

La preuve occupe les prochains paragraphes et fait jouer un rôle important aux algèbres

$$C_n := C \otimes_{B[\frac{1}{g}]} B_n,\ C_\infty := \varprojlim_n C_n,\ C_\infty^0 := \varprojlim_n C_n^0.$$

On dispose d'isomorphismes canoniques [49]

$$C_{n+1}\widehat{\otimes}_{B_{n+1}} B_n \simeq C_{n+1} \otimes_{B_{n+1}} B_n = (C \otimes_{B[\frac{1}{g}]} B_{n+1}) \otimes_{B_{n+1}} B_n \simeq C \otimes_{B[\frac{1}{g}]} B_n = C_n.$$

Remarque 5.5. a) Comme C est localement libre de rang fini sur $B[\frac{1}{g}]$, le morphisme injectif (théorème 5.1) $B[\frac{1}{g}] \to B_\infty$ induit une injection $C \to C \otimes_{B[\frac{1}{g}]} B_\infty$, et le morphisme naturel $C \otimes_{B[\frac{1}{g}]} B_\infty \to C_\infty$ est un isomorphisme. On peut donc identifier C à une sous-algèbre de C_∞, via $c \mapsto (c \otimes 1)_{n\geqslant 1}$.

b) Puisque C_n est finie étale sur $B_n \in \mathrm{Perf}_K$, on a $C_n \in \mathrm{Perf}_K$ et C_n^0 est presque finie étale sur B_n^0 (théorème de presque pureté 3.13). Les isomorphismes $C_{n+1}\widehat{\otimes}_{B_{n+1}} B_n \simeq C_n$ combinés au théorème 3.11 montrent que le morphisme naturel $C_{n+1}^0\widehat{\otimes}_{B_{n+1}^0} B_n^0 \to C_n^0$ est un presque isomorphisme (à priori dans le cadre π^{1/p^∞}, mais donc aussi dans le cadre $(\pi g)^{1/p^\infty}$).

5.2.1. *Identification de la clôture intégrale.* — Dans ce paragraphe on montre le résultat suivant.

Proposition 5.6. *L'inclusion de C dans C_∞ identifie $\tilde{C}^0$ à C_∞^0.*

Ainsi $\tilde{C} := \tilde{C}^0[\frac{1}{p}]$ s'identifie à la limite uniforme des algèbres de Banach uniformes C_n et $(\tilde{C})^0 = \tilde{C}^0$. La preuve de la proposition demande quelques préliminaires.

Lemme 5.7. *On a $C_\infty^0 = \mathrm{fi}(B_\infty^0, C_\infty)$.*

Démonstration. Si $x = (x_n)_{n\geqslant 0} \in \mathrm{fi}(B_\infty^0, C_\infty)$, alors x_n est entier sur B_n^0, donc ses puissances restent dans un sous C_n^0-module de type fini de C_n, et $x_n \in C_n^0$ car C_n est uniforme. Ainsi $x \in C_\infty^0$. Pour l'autre inclusion, soit $x = (x_n)_{n\geqslant 0} \in C_\infty^0$ et soit $\chi_n \in B_n[X]$ le polynôme caractéristique de la multiplication par x_n dans C_n. Puisque $C_{n+1} \otimes_{B_{n+1}} B_n \simeq C_n$, il existe $\chi \in B_\infty[X]$ induisant tous les χ_n. Par Cayley–Hamilton, on a $\chi_n(x_n) = 0$ pour tout n, donc $\chi(x) = 0$. Le lemme 5.8 ci-dessous montre que $\chi_n \in B_n^0[X]$ pour tout n, donc $\chi \in B_\infty^0[X]$ et $x \in \mathrm{fi}(B_\infty^0, C_\infty)$, ce qui finit la preuve. □

Lemme 5.8. *Soit $R \to S$ un morphisme fini étale de K-algèbres de Banach uniformes. Pour tout $s \in S^0$ le polynôme caractéristique de s est à coefficients dans R^0.*

[49] Le premier est dû au fait que C_{n+1} est un B_{n+1}-module projectif de type fini.

Démonstration. Soient $a_0, \dots, a_n$ les coefficients de ce polynôme. On veut montrer que $\max_{0\leqslant i\leqslant n} |a_i|_{\mathrm{sp}} \leqslant 1$, ou encore (*cf.* § 3.1) que $\max_{0\leqslant i\leqslant n} |a_i(x)| \leqslant 1$ pour tout $x \in \mathscr{M}(S)$. Si L est le corps résiduel complété de x, on veut montrer que les $a_i(x)$ appartiennent à L^0. On peut remplacer $R \to S$ par $L \to S \otimes_R L$ et donc supposer que $R = L$, auquel cas S est un produit d'extensions finies de R ; le résultat s'en déduit. □

Lemme 5.9. *On a $C = C^0_\infty[\frac{1}{pg}]$ à l'intérieur de C_∞.*

Démonstration. Notons $T = C^0_\infty[\frac{1}{pg}]$. Comme C est entier sur $B[\frac{1}{g}] = B^0[\frac{1}{pg}]$, pour tout $x \in C$ il existe $n \geqslant 0$ tel que $(pg)^n x$ soit entier sur $B^0 \subset B^0_\infty$, et donc dans $\mathrm{fi}(B^0_\infty, C_\infty) = C^0_\infty$ (lemme 5.7). Ainsi $C \subset T$. Montrons que $T \subset C$. Par le lemme 5.7 et l'isomorphisme (théorème 5.1) $B[\frac{1}{g}] \simeq B^0_\infty[\frac{1}{pg}]$ on a $T \subset \mathrm{fi}(B[\frac{1}{g}], C \otimes_{B[\frac{1}{g}]} B_\infty)$. Puisque le morphisme $B[\frac{1}{g}] \to C$ est entier on a $\mathrm{fi}(B[\frac{1}{g}], C \otimes_{B[\frac{1}{g}]} B_\infty) = \mathrm{fi}(C, C \otimes_{B[\frac{1}{g}]} B_\infty)$. Comme $B[\frac{1}{g}] \to C$ est étale, le morphisme naturel $C \otimes_{B[\frac{1}{g}]} \mathrm{fi}(B[\frac{1}{g}], B_\infty) \to \mathrm{fi}(C, C \otimes_{B[\frac{1}{g}]} B_\infty)$ est un isomorphisme. Enfin, $\mathrm{fi}(B[\frac{1}{g}], B_\infty) = B[\frac{1}{g}]$ par le théorème 5.1, donc $T \subset C$. □

On finit la preuve de la proposition 5.6 en utilisant les lemmes 5.7 et 5.9 et l'égalité (dans B_∞, *cf.* théorème 5.1) $g^{-1/p^\infty} B^0 = B^0_\infty$:

$$\tilde{C}^0 = \mathrm{fi}(g^{-1/p^\infty} B^0, C) = \mathrm{fi}(B^0_\infty, C^0_\infty[\frac{1}{pg}]) = C^0_\infty.$$

Le lemme 5.9 a une autre conséquence importante :

Corollaire 5.10. *Le morphisme $f\colon C^0_\infty[\frac{1}{p}]\widehat{\otimes}_B B_n \to C_n$ induit par la projection canonique $C^0_\infty[\frac{1}{p}] \to C_n$ est un isomorphisme.*

Démonstration. On a (lemme 5.9)

$$C_n = C \otimes_{B[\frac{1}{g}]} B_n = C^0_\infty[\frac{1}{pg}] \otimes_{B[\frac{1}{g}]} B_n \simeq C^0_\infty[\frac{1}{p}] \otimes_B B_n,$$

d'où un morphisme $h\colon C_n \to C^0_\infty[\frac{1}{p}]\widehat{\otimes}_B B_n$ de B_n-modules (automatiquement continu puisque C_n est projectif de type fini sur B_n), tel que $f \circ h$ soit l'identité. On vérifie que $h \circ f$ est l'identité sur l'image de $C^0_\infty[\frac{1}{p}] \otimes_B B_n$ dans $C^0_\infty[\frac{1}{p}]\widehat{\otimes}_B B_n$, donc aussi sur $C^0_\infty[\frac{1}{p}]\widehat{\otimes}_B B_n$ par continuité et densité. Donc f est un isomorphisme. □

5.2.2. *Etude du système projectif $\{C^0_n/\pi^d\}$ et fin de la preuve.* — La différence entre les limites dans Ban^u_K et celles dans Perf_K (remarque 3.6) se fera pleinement sentir dans ce paragraphe.

Rappelons que ξ est un générateur distingué de $\ker(\theta_{K^0})$ et que π est l'image de $[\xi(0)^{1/p}]$. En posant $R_n = (C_n^0)^\flat$, on obtient (théorème 2.10) des isomorphismes $\theta_{C_n^0}\colon C_n^0 \simeq W(R_n)/\xi$ compatibles avec la variation de n. Comme C_n^0 est sans p-torsion, les R_n n'ont pas de $\xi(0)$-torsion, donc $R = \varprojlim_n R_n \simeq (C_\infty^0)^\flat \in \mathrm{Perf}_{\mathbf{F}_p}$ n'en a pas non plus et $S := W(R)/\xi$ est une K^0-algèbre perfectoïde plate.

Proposition 5.11. *Les morphismes naturels $C_\infty^0/\pi \to \varprojlim_n (C_n^0/\pi)$ et $S \to C_\infty^0$ sont des presque isomorphismes dans le cadre $(\pi g)^{1/p^\infty}$, et $S \to C_\infty^0$ est injectif.*

Démonstration. Le morphisme naturel $\varprojlim_n (C_n^0/\pi^d) \to \varprojlim_n (C_n^0/\pi^e)$ est presque surjectif pour tous $d > e$: par K^0-platitude des C_n il suffit de voir que $R^1 \varprojlim_n (C_n^0/\pi)$ est presque nul ; or le système projectif $\{B_n^0/\pi\}_{n\geqslant 1}$ est presque Mittag-Leffler [(50)] (conséquence directe du théorème 5.2) et le presque isomorphisme $C_{n+1}^0/\pi \otimes_{B_{n+1}^0/\pi} B_n^0/\pi \to C_n^0/\pi$ (remarque 5.5) montre qu'il en est de même du système projectif $\{C_n^0/\pi\}_{n\geqslant 1}$.

Ainsi les transitions du système projectif $\{\varprojlim_n (C_n^0/\pi^d)\}_{d\geqslant 1}$ sont presque surjectives, ayant pour conséquence la presque surjectivité du morphisme

$$C_\infty^0 \simeq \varprojlim_d (\varprojlim_n (C_n^0/\pi^d)) \to \varprojlim_n (C_n^0/\pi).$$

Puisque $C_\infty^0/\pi \to \varprojlim_n (C_n^0/\pi)$ est clairement injectif, c'est un presque isomorphisme.

Pour montrer que $S \to C_\infty^0$ est injectif et un presque isomorphisme il suffit de le vérifier modulo π (les deux algèbres étant plates sur K^0 et π-complètes). Mais $S/\pi \simeq R/\xi(0)^{1/p}$, $C_n^0/\pi \simeq R_n/\xi(0)^{1/p}$ et nous avons vu que $C_\infty^0/\pi \to \varprojlim_n (C_n^0/\pi)$ est injectif et un presque isomorphisme. Il suffit donc de vérifier que $R/\xi(0)^{1/p} \to \varprojlim_n (R_n/\xi(0)^{1/p})$ est injectif et un presque isomorphisme. L'injectivité est claire puisque les R_n n'ont pas de $\xi(0)^{1/p}$-torsion. Le fait que c'est un presque isomorphisme découle (comme ci-dessus, par $\xi(0)$-complétude des R_n) de la presque surjectivité des applications canoniques $\varprojlim_n (R_n/\xi(0)^{p^k}) \to \varprojlim_n (R_n/\xi(0)^{p^{k-1}})$. Celle-ci s'obtient en utilisant la perfection des R_n et en appliquant successivement Frobenius au morphisme presque surjectif (*cf.* premier paragraphe) $\varprojlim_n (C_n^0/\pi^p) \to \varprojlim_n (C_n^0/\pi)$, qui s'identifie à $\varprojlim_n (R_n/\xi(0)) \to \varprojlim_n (R_n/\xi(0)^{1/p})$. □

Posons $T = S[\frac{1}{p}] = S_*[\frac{1}{p}] \in \mathrm{Perf}_K$ (proposition 3.5). La proposition 5.11 fournit un presque isomorphisme $T \simeq C_\infty^0[\frac{1}{p}]$. Comme πg est inversible dans B_n,

[(50)] Un système projectif $\{M_n\}_{n\geqslant 1}$ de B^0-modules est dit presque Mittag-Leffler si pour tous k, n il existe $m \geqslant n$ tel que $(\pi g)^{1/p^k}\mathrm{Im}(M_m \to M_n) \subset \mathrm{Im}(M_t \to M_n)$ pour tout $t \geqslant m$.

cela induit un isomorphisme $T\widehat{\otimes}_B B_n \simeq C^0_\infty[\frac{1}{p}]\widehat{\otimes}_B B_n$, d'où (corollaire 5.10) un isomorphisme $T\widehat{\otimes}_B B_n \simeq C_n$. Puisque T, B, B_n sont dans Perf_K, le morphisme naturel $T^0\widehat{\otimes}_{B^0} B^0_n \to (T\widehat{\otimes}_B B_n)^0$ est un presque isomorphisme (théorème 3.11). On obtient donc des presque isomorphismes, compatibles avec la variation de n et d

$$T^0/\pi^d \otimes_{B^0/\pi^d} B^0_n/\pi^d \simeq C^0_n/\pi^d.$$

Par le théorème de presque pureté 3.13, la B^0_n/π^d-algèbre C^0_n/π^d est presque finie étale (et fidèlement plate, si C l'est sur $B[\frac{1}{g}]$). Les isomorphismes ci-dessus combinés avec le théorème 5.2 et le lemme 5.12 ci-dessous montrent que T^0/π^d est presque finie étale (et fidèlement plate, si C l'est sur $B[\frac{1}{g}]$) sur B^0/π^d. Mais $T^0 = S_*$ est presque isomorphe à S, qui est presque isomorphe à C^0_∞ (proposition 5.11), qui s'identifie enfin à $\tilde{C}^0$ par la proposition 5.6. Cela finit la preuve du théorème 5.3.

Le lemme suivant a joué un rôle important ci-dessus, permettant de transférer les propriétés des B^0_n/π^d-algèbres C^0_n/π^d (fournies par le théorème de presque pureté) à la B^0/π^d-algèbre T^0/π^d. On trouve dans la section 14.2 de Gabber et Ramero (2018) des résultats beaucoup plus précis et généraux que l'énoncé ci-dessous.

Lemme 5.12. *Soit $\mathscr{P} \in \{$plat, fidèlement plat, projectif, de type fini$\}$. Soit M un B^0/π^d-module et $M_n := M \otimes_{B^0/\pi^d} B^0_n/\pi^d$.*

a) Si le B^0_n/π^d-module M_n est presque $\mathscr{P}$ pour tout n, alors M est presque $\mathscr{P}$.

b) Si M est une B^0/π^d-algèbre et si M_n est une B^0_n/π^d-algèbre presque finie étale pour tout n, alors M est presque finie étale sur B^0/π^d (idem avec presque finie étale et fidèlement plate).

Démonstration. Il s'agit d'une conséquence formelle du fait que le système projectif $\{B^0_n/\pi^d\}$ est « presque constant de valeur B^0/π^d » (théorème 5.2) et de l'interprétation catégorique (à la Yoneda) de $\mathscr{P}$. Voir le théorème 14.2.39 de Gabber et Ramero (2018) pour les détails. □

6. Existence d'algèbres de Cohen–Macaulay

On fixe par la suite de cette section, qui est le coeur de l'exposé, un nombre premier p (d'où une notion d'anneau perfectoïde). On suppose que la caractéristique résiduelle de tous les anneaux locaux noethériens rencontrés ci-dessous est égale à p. Pour éviter des longues listes d'épithètes, introduisons :

Définition 6.1. La catégorie [(51)] CLI_p a pour objets les anneaux locaux noethériens complets, intègres, d'inégale caractéristique et de corps résiduel parfait (de caractéristique p), et pour morphismes les morphismes locaux d'anneaux locaux.

[(51)] CLI est l'abréviation de « complet local intègre », l'indice p indique la caractéristique résiduelle, ainsi que le caractère parfait du corps résiduel.

Le but de cette section est d'expliquer la preuve du théorème suivant. Combiné avec les travaux de Hochster et Huneke (1992), il implique le théorème 1.10 de l'introduction.

Théorème 6.2. *Pour tout $A \in \mathrm{CLI}_p$ il existe une A-algèbre de Cohen–Macaulay.*

Voir les théorèmes 6.18 et 6.20 pour des résultats bien plus forts et précis. Ma (2021) a remarqué qu'une construction d'André (2020) fournit une preuve relativement directe [(52)] du théorème ci-dessus, n'utilisant pas le difficile lemme d'Abhyankar perfectoïde (comme dans André, 2018b) : on traite d'abord le cas d'un anneau régulier, en construisant dans ce cas une A-algèbre de Cohen–Macaulay *perfectoïde* C (ceci est standard), puis on écrit l'anneau $A \in \mathrm{CLI}_p$ comme un quotient d'un anneau régulier A_0 et on fabrique à partir de C et du lemme de platitude d'André une A-algèbre *presque perfectoïde et presque de Cohen–Macaulay* C' (*cf.* proposition 6.14 pour l'énoncé précis, mais peu ragoûtant). Nous adaptons la méthode de Ma pour obtenir aussi une forme faible de fonctorialité de la construction, retrouvant ainsi l'un des résultats principaux d'André (2020) (on trouvera cependant dans loc.cit. des résultats bien plus forts que ceux exposés ici). Une autre construction de C' (celle d'André, 2018b) utilise plutôt le fait que A est une *extension* finie d'un anneau local régulier complet A_0, et le lemme d'Abhyankar perfectoïde combiné avec le lemme de platitude d'André pour construire C'. Les deux preuves produisent donc seulement une « presque algèbre de Cohen–Macaulay (presque perfectoïde) », mais il était connu, grâce aux travaux de Hochster (2002), que cela suffit pour conclure la preuve du théorème 6.2. Gabber a découvert un autre moyen, plus canonique et direct, de passer d'une presque algèbre de Cohen–Macaulay à une vraie telle algèbre, il est exposé dans le paragraphe 17.5 du livre de Gabber et Ramero (2018). On trouvera dans cette section une version encore plus simple et directe.

6.1. Algèbres perfectoïdes CM

On fixe une fois pour toutes une suite compatible $(p^{1/p^n})_{n\geqslant 0}$ de racines de p dans une clôture algébrique de $\mathbf{Q}_p$ et on définit le corps perfectoïde

$$K := \mathbf{Z}_p\widehat{[p^{1/p^\infty}]}[\frac{1}{p}],$$

pour lequel $K^0 = \mathbf{Z}_p\widehat{[p^{1/p^\infty}]}$. Soit $p^\flat := (p, p^{1/p}, p^{1/p^2}, \ldots) \in K^{0,\flat}$ et $\xi := p - [p^\flat]$, un élément distingué de $W(K^{0,\flat})$. On note $\mathrm{Perf}^{\mathrm{tf}}_{K^0}$ la catégorie des K^0-algèbres perfectoïdes *sans p-torsion*.

[(52)] Modulo le lemme de platitude d'André, qui est tout sauf une platitude !

Définition 6.3. Soit A un anneau local noethérien. Une *A-algèbre CM* (resp. *A-algèbre perfectoïde CM*) est une A-algèbre de Cohen–Macaulay (resp. une A-algèbre de Cohen–Macaulay qui est aussi dans $\mathrm{Perf}^{\mathrm{tf}}_{K^0}$).

Nous aurons besoin des résultats suivants par la suite :

Lemme 6.4. *Soit $(x_1, \dots, x_d)$ une suite régulière dans une algèbre $C \in \mathrm{Perf}^{\mathrm{tf}}_{K^0}$, avec $x_1 = p$. Le complété $(x_1, \dots, x_d)$-adique $\widehat{C}$ de C est dans* $\mathrm{Perf}^{\mathrm{tf}}_{K^0}$.

Démonstration. Voir la proposition 2.2.1 d'André (2020) pour les détails. Puisque $I := (x_1, \dots, x_d)$ est de type fini, $\widehat{C}$ est I-complet (donc aussi p-complet). Si $z \in C$ vérifie $pz \in I^{nd}C$, alors $z \in I^{n-1}C$. En effet, on a $I^{nd} \subset (p^n, x_2^n, \dots, x_d^n)$, donc il existe $y \in C$ tel que $p(z - p^{n-1}y) \in (x_2^n, \dots, x_d^n)C$, puis par régularité de $(p, x_2^n, \dots, x_d^n)$ on obtient $z - p^{n-1}y \in (x_2^n, \dots, x_d^n)C$ et $z \in I^{n-1}C$. On en déduit immédiatement que $\widehat{C}$ est sans p-torsion et que les idéaux $p^{1/p}\widehat{C}$ et $p\widehat{C}$ sont fermés dans $\widehat{C}$ pour la topologie I-adique, donc $\widehat{C}/p \simeq \widehat{C/p}$, $\hat{C}/p^{1/p} \simeq \widehat{C/p^{1/p}}$. On conclut en utilisant la proposition 2.12. □

Lemme 6.5. *Soit $(A, \mathfrak{m}) \in \mathrm{CLI}_p$ (déf. 6.1) et soit C une A-algèbre perfectoïde* CM, *$\mathfrak{m}$-complète. Pour tous $z_1, \dots, z_n \in A$ il existe une C-algèbre perfectoïde* CM, *$\mathfrak{m}$-complète C' dans laquelle $z_1, \dots, z_n$ sont p-puissants* (*déf. 2.2*).

Démonstration. Par le théorème 4.3 il existe une C-algèbre $P \in \mathrm{Perf}^{\mathrm{tf}}_{K^0}$ fidèlement plate sur C modulo p et dans laquelle $z_1, \dots, z_n$ sont p-puissants. Si $(x_1, \dots, x_d)$ est un système de paramètres de A avec $x_1 = p$, alors $(x_1, \dots, x_d)$ est une suite régulière dans C, et elle le reste dans P par fidèle platitude modulo p de celui-ci sur C. Par la proposition 1.9 le complété $(x_1, \dots, x_d)$-adique (qui est aussi le complété $\mathfrak{m}$-adique) C' de P est une A-algèbre CM $\mathfrak{m}$-complète, qui est dans $\mathrm{Perf}^{\mathrm{tf}}_{K^0}$ par le lemme 6.4. □

6.2. Algèbres perfectoïdes CM sur un anneau local régulier

Les théorèmes 6.6 et 6.8 ci-dessous sont des analogues en inégale caractéristique du célèbre résultat de Kunz (1969) : pour un anneau noethérien A de caractéristique p chacune des assertions suivantes est équivalente à la régularité de A :

- le Frobenius $\varphi\colon A \to A$ de A est un morphisme plat.
- il existe une A-algèbre parfaite et fidèlement plate.

Rappelons (proposition 1.9) que pour un anneau local régulier A une A-algèbre est CM si et seulement si elle est fidèlement plate sur A. En utilisant la partie facile du théorème de Kunz et le résultat suivant, on en déduit facilement que pour tout anneau régulier A avec $p \in \mathrm{Rad}(A)$ il existe une A-algèbre perfectoïde fidèlement plate.

Théorème 6.6. *Pour tout anneau local régulier $(A, \mathfrak{m})$ d'inégale caractéristique il existe une A-algèbre perfectoïde* CM, $\mathfrak{m}$*-complète.*

Dans les applications nous aurons uniquement besoin du théorème pour un anneau de la forme $W(k)[[T_1, \dots, T_d]]$, k étant un corps parfait de caractéristique p, auquel cas la preuve est parfaitement élémentaire.

Démonstration. On peut supposer que A est complet, car le complété $\hat{A}$ de A est fidèlement plat sur A. Par le théorème de structure de Cohen et la régularité de A il existe un anneau de valuation V, complet et absolument non ramifié [(53)], ainsi qu'un élément $f \in V[[X_1, \dots, X_n]]$, que l'on peut choisir de la forme $f = X_1$ si $p \in \mathfrak{m} \setminus \mathfrak{m}^2$ et dans l'idéal $(p, X_1, \dots, X_n)^2$ sinon, tels que

$$A \simeq V[[X_1, \dots, X_n]]/(p - f).$$

Supposons pour commencer que k est parfait. Posons

$$A_j = V[[X_1^{1/p^j}, \dots, X_n^{1/p^j}]]/(p - f), \ A_\infty = \varinjlim_j A_j$$

et montrons que le complété p-adique $\widehat{A}_\infty$ de A_∞ est une A-algèbre perfectoïde fidèlement plate. Les morphismes $A \to A_j$ sont locaux, injectifs, finis et plats, donc fidèlement plats. Ainsi A_∞ est une A-algèbre fidèlement plate, et le lemme ci-dessous montre que $\widehat{A}_\infty$ reste fidèlement plate sur A. Il est évident que $\widehat{A}_\infty$ est sans p-torsion et que le Frobenius sur A_∞/p est surjectif. Montrons qu'il existe $\pi \in \widehat{A}_\infty$ tel que $(\pi^p) = (p)$. Cela est évident si $f = X_1$, supposons donc que $f \in (p, X_1, \dots, X_n)^2$. Puisque $A_\infty = A_\infty^p + pA_\infty$ (donc $\mathfrak{m}_{A_\infty} = \mathfrak{m}_{A_\infty}^p + pA_\infty$) et $p \in \mathfrak{m}_{A_\infty}^2$, on obtient $p = \pi^p + py$ pour certains $\pi, y \in \mathfrak{m}_{A_\infty}$, donc $(p) = (\pi^p)$. On conclut alors en utilisant la proposition 2.12 et en remarquant que le Frobenius $A_\infty/\pi \to A_\infty/\pi^p$ est injectif, puisque A_∞ est normal (tous les A_j sont réguliers, donc normaux), donc p-clos dans $A_\infty[\frac{1}{\pi}]$.

Dans le cas général soit $\bar{k}$ une clôture algébrique de k et soit $V \to W$ une extension fidèlement plate d'anneaux de valuation absolument non ramifiés, telle que le corps résiduel de W soit $\bar{k}$. Puisque $W[[X_1, \dots, X_n]]$ est une $V[[X_1, \dots, X_n]]$-algèbre fidèlement plate (par exemple par le lemme ci-dessous), $W[[X_1, \dots, X_n]]/(p - f)$ est une extension fidèlement plate de A, avec un corps résiduel algébriquement clos. Mais $W[[X_1, \dots, X_n]]/(p - f)$ possède une algèbre perfectoïde fidèlement plate S d'après ce qui précède, et elle répond à l'appel.

Ce qui précède montre qu'il existe une A-algèbre perfectoïde CM. Son complété $\mathfrak{m}$-adique répond à l'appel par les lemmes 6.4 et 6.7. □

[(53)] Cela veut dire que l'idéal maximal de V est pV.

Nous avons utilisé le résultat suivant (*cf.* lemme 1.1.1 d'André (2018b)) :

Lemme 6.7. *Soit I un idéal d'un anneau noethérien A et soit B une A-algèbre plate. On note $\hat{X} = \varprojlim_n X/I^nX$ pour tout A-module X.*

a) Pour tout A-module de type fini M le morphisme naturel $M \otimes_A \hat{B} \to \widehat{M \otimes_A B}$ est un isomorphisme.

b) La A-algèbre $\hat{B}$ est plate, et même fidèlement plate si $I \subset \mathrm{Rad}(A)$.

Démonstration. a) Une application directe du lemme d'Artin–Rees montre que le foncteur $\mathscr{F}\colon M \mapsto \widehat{M \otimes_A B}$ est exact sur les A-modules de type fini. Puisque le morphisme $f_M\colon M \otimes_A \hat{B} \to \widehat{M \otimes_A B}$ est un isomorphisme si M est libre de type fini sur A, il l'est pour tout M de type fini sur A par un argument standard.

b) La platitude découle directement du point a) et de sa preuve. Supposons que $I \subset \mathrm{Rad}(A)$ et que B est fidèlement plat sur A. Soit M un A-module de type fini tel que $M \otimes_A \hat{B} = 0$. Alors $M \otimes_A B/IB = 0$ et donc $M/IM \otimes_A B = 0$, puis $M/IM = 0$ et (lemme de Nakayama) $M = 0$. □

La preuve du résultat suivant est nettement plus délicate et nous renvoyons le lecteur à Bhatt, Iyengar et Ma (2019) car nous n'en ferons pas usage.

Théorème 6.8. *Soit A un anneau noethérien tel que $p \in \mathrm{Rad}(A)$. S'il existe une A-algèbre perfectoïde et fidèlement plate, alors A est régulier.*

6.3. Un résultat technique crucial

On fixe dans ce paragraphe un anneau $(A, \mathfrak{m}) \in \mathrm{CLI}_p$ (déf. 6.1) et $\wp \in \mathrm{Spec}A[\frac{1}{p}]$, de hauteur $c \geqslant 1$. On a donc $A/\wp \in \mathrm{CLI}_p$.

Proposition 6.9. *Il existe $g \in A \setminus \wp$ multiple de p, ainsi qu'un système de paramètres $(f_1, \dots, f_c, x_1, \dots, x_d)$ de A avec $x_1 = p$ et*

$$g\wp \subset \sqrt{(f_1, \dots, f_c)} \subset \wp.$$

La suite $(x_1, \dots, x_d)$ est un système de paramètres de $A/\wp$.

Démonstration. Par le théorème de l'idéal principal de Krull (et un argument standard d'évitement des idéaux premiers) il existe $f_1, \dots, f_c \in \wp$ tels que $\mathrm{ht}(p, f_1, \dots, f_c) = c + 1$. Ainsi $(p, f_1, \dots, f_c)$ s'étend en un système de paramètres $(f_1, \dots, f_c, x_1, \dots, x_d)$ de A avec $x_1 = p$. Puisque $\wp$ est un idéal premier minimal de l'anneau réduit $A' := A/\sqrt{(f_1, \dots, f_c)}$, tout élément de $\wp$ est un diviseur de zéro dans A', d'où l'existence de g (que l'on peut choisir multiple de p car $p \notin \wp$). □

Définition 6.10. Soit R un anneau muni d'un élément p-puissant (déf. 2.2) g, et soit S une R-algèbre. Une suite $x = (x_1, \dots, x_d)$ d'éléments de S est dite *presque régulière* (*dans le cadre* g^{1/p^∞}) si l'idéal $(g^{p^{-\infty}})$ annule [54] $\frac{(x_1,\dots,x_i):(x_{i+1})}{(x_1,\dots,x_i)}$ pour tout i, mais n'annule pas $S/(x_1, \dots, x_d)$ (en particulier $g \neq 0$).

On fera attention au fait qu'une suite régulière dans S n'a aucune raison d'être presque régulière (!) : même si $S/(x_1, \dots, x_d)$ est non nul, il pourrait très bien être annulé par $(g^{p^{-\infty}})$. Le résultat délicat suivant (tiré et adapté d'André (2020)) montre que ce genre de canular ne se produit pas dans notre situation.

Proposition 6.11. *Soit $(f_1, \dots, f_c, x_1, \dots, x_d)$ un système de paramètres de A, avec $x_1 = p$. Soit C une A-algèbre* CM, $\mathfrak{m}$*-complète, dans laquelle $p, f_1, \dots, f_c$ sont p-puissants (déf. 2.2). Posons*

$$\bar{C} := C/(f_1^{p^{-\infty}}, \dots, f_c^{p^{-\infty}})C, \ \widehat{\bar{C}} = \varprojlim_n \overline{C}/p^n.$$

a) La suite $(x_1, \dots, x_d)$ est régulière dans $\bar{C}$ et dans $\widehat{\bar{C}}$.

b) Si $C \in \mathrm{Perf}^{\mathrm{tf}}_{K^0}$, il en est de même de $\widehat{\bar{C}}$.

c) Soit $\wp \in V(f_1, \dots, f_c)$ et soit $g \in A \setminus \wp$ un élément qui devient p-puissant dans C. La suite $(x_1, \dots, x_d)$ est presque régulière (dans le cadre g^{1/p^∞}, déf. 6.10) dans $\overline{C}$ et $\widehat{\bar{C}}$.

Démonstration. a) Comme $x_1 = p$, il suffit de voir que $(x_1, \dots, x_d)$ est régulière dans $\bar{C} = \varinjlim_n C/(f_1^{1/p^n}, \dots, f_c^{1/p^n})$, ou encore que $(x_1, \dots, x_d)$ est régulière dans $C/(f_1^{1/p^n}, \dots, f_c^{1/p^n})$ pour tout n, or cela est clair puisque la suite $(f_1^{1/p^n}, \dots, f_c^{1/p^n}, x_1, \dots, x_d)$ est régulière [55] dans C.

b) Par a) $\widehat{\bar{C}}$ n'a pas de p-torsion. En utilisant la proposition 2.12, il suffit de vérifier que le Frobenius $\overline{C}/p^{1/p} \to \overline{C}/p$ est un isomorphisme, or cela découle directement des définitions et de l'isomorphisme $C/p^{1/p} \simeq C/p$ induit par le Frobenius.

c) Ceci coûte nettement plus cher. Par a) il suffit de vérifier que $\bar{C}/(x_1, \dots, x_d)$ n'est pas annulé par $(g^{p^{-\infty}})$. Supposons que ce n'est pas le cas, donc

$$(g^{p^{-\infty}})C \subset (f_1^{p^{-\infty}}, \dots, f_c^{p^{-\infty}})C + (x_1, \dots, x_d)C,$$

en particulier $g \in \mathfrak{m}^n C + \sqrt{\wp C}$ pour tout n.

Soit $\bar{C} = C/\wp C$, $\bar{A} = A/\wp$ et notons $\bar{x}$ l'image de x dans $\bar{C}$ (resp. $\bar{A}$) si $x \in C$ (resp. $x \in A$). Soit $f\colon \bar{C} \to \bar{A}$ un morphisme de $\bar{A}$-modules. Pour tout $n \geqslant 1$ il existe $\alpha \in \mathfrak{m}^n C$ et $N \geqslant 1$ tel que $(g - \alpha)^N \in \wp C$, donc $(\bar{g} - \overline{\alpha})^N = 0$ dans $\bar{C}$. En appliquant f on obtient $\bar{g}^N f(1) - \binom{N}{1}\bar{g}^{N-1} f(\overline{\alpha}) + \cdots + (-1)^N f(\overline{\alpha}^N) = 0$, avec $f(\overline{\alpha}^k) \in \mathfrak{m}_{\bar{C}}^{nk}$. Ainsi

[54] Si I, J sont deux idéaux d'un anneau S, on note $I : J = \{s \in S \mid sJ \subset I\}$.

[55] Rappelons que si M est un module sur un anneau R et $x_1, \dots, x_n \in R$, $e_1, \dots, e_n \in \mathbf{Z}_{>0}$, alors la suite $(x_1, \dots, x_n)$ est régulière dans M si et seulement si la suite $(x_1^{e_1}, \dots, x_n^{e_n})$ l'est.

$x := f(1)\bar{g} \in \bar{A}$ vérifie une équation de la forme $x^N + a_1 x^{N-1} + \cdots + a_N = 0$ avec $a_i \in \mathfrak{m}_{\bar{A}}^{ni}$ (dépendant de n, bien entendu).

Montrons que cela force $x = 0$. La clôture intégrale de $\bar{A}$ dans son corps des fractions est un anneau noethérien normal et intègre (théorème de Nagata), donc une intersection d'anneaux de valuation discrète $(V_i)_{i\in I}$ (les localisés de cette clôture intégrale en ses idéaux premiers de hauteur 1) tels que le morphisme $\bar{A} \to V_i$ soit local. L'image y de x dans un tel V_i vérifie des équations du type $y^N + b_1 y^{N-1} + \cdots + b_N = 0$ avec $b_j \in \mathfrak{m}_{V_i}^{jn}$ (dépendant de n). Cela force $v(y^N) \geqslant \min_{1\leqslant j\leqslant N}(v(b_j) + (N-j)v(y))$, puis $v(y) \geqslant n$ pour tout n, et enfin $y = 0$ et $x = 0$.

On a donc $f(1)\bar{g} = 0$ dans $\bar{A} = A/\wp$, puis $f(1) = 0$ car $g \notin \wp$. En appliquant ceci au morphisme $c \mapsto f(tc)$ pour $t \in \bar{C}$ quelconque, on voit que $f = 0$, autrement dit $\mathrm{Hom}_{\mathrm{Mod}_{A/\wp}}(C/\wp C, A/\wp) = \{0\}$, contredisant le lemme ci-dessous. □

Lemme 6.12. *Soit A un anneau local noethérien complet et intègre et soit $\wp$ un idéal premier de A. Si C est une A-algèbre* CM*, alors* $\mathrm{Hom}_{\mathrm{Mod}_A}(C, A)$ *et* $\mathrm{Hom}_{\mathrm{Mod}_{A/\wp}}(C/\wp C, A/\wp)$ *sont non nuls.*

Démonstration. Le très joli argument suivant est dû à Hochster (1994). Soit $A_0 \to A$ une extension finie, avec A_0 local régulier et complet, de corps résiduel k, et soit $q = A_0 \cap \wp$. Si E (resp. E') est une enveloppe injective du A_0-module (resp. A_0/q-module) k, par dualité de Matlis on a $\mathrm{Hom}_{\mathrm{Mod}_{A_0}}(C, A_0) \simeq \mathrm{Hom}_{\mathrm{Mod}_{A_0}}(C \otimes_{A_0} E, E)$ et $\mathrm{Hom}_{\mathrm{Mod}_{A_0/q}}(C/qC, A_0/q) \simeq \mathrm{Hom}_{\mathrm{Mod}_{A_0/q}}(C/qC \otimes_{A_0/q} E', E')$. Ces modules sont non nuls car $C \otimes_{A_0} E$ et $C/qC \otimes_{A_0/q} E$ le sont, puisque C (resp. C/qC) est fidèlement plat sur A_0 (resp. A_0/q), en tant que A-algèbre CM (proposition 1.9).

Pour conclure, il faut passer d'un A_0-dual à un A-dual (l'argument donné ci-dessous pour $\mathrm{Hom}_{\mathrm{Mod}_A}(C, A)$ est identique pour $\mathrm{Hom}_{\mathrm{Mod}_{A/\wp}}(C/\wp C, A/\wp)$). Soit $f\colon C \to A_0$ un morphisme A_0-linéaire non nul. Pour tout $c \in C$ on dispose d'un morphisme A_0-linéaire $\varphi_c\colon A \to A_0, a \mapsto f(ac)$. L'application $F\colon C \to X := \mathrm{Hom}_{\mathrm{Mod}_{A_0}}(A, A_0), c \mapsto \varphi_c$ est non nulle et A-linéaire. Puisque le morphisme $A_0 \to A$ est fini, le A-module X est de type fini sur A et sans torsion [(56)], donc se plonge dans un A-module libre de type fini A^j. En projetant sur l'un des facteurs de A^j et en composant avec F on obtient un élément non nul de $\mathrm{Hom}_{\mathrm{Mod}_A}(C, A)$. □

6.4. Construction d'une algèbre presque perfectoïde et presque CM

Soit $P \in \mathrm{Perf}_{K^0}^{\mathrm{tf}}$ et soit $g \in P \setminus \{0\}$ un élément p-puissant (définition 2.2) et multiple de p. On ne suppose *pas* que P est sans g-torsion. Soit $\iota\colon P \to P[\frac{1}{g}]$ le morphisme

[(56)] Si $a \in A \setminus \{0\}$ et $f \in X$ vérifient $a.f = 0$ alors $f(P(a)x) = 0$ pour tout $P \in XA_0[X]$ et tout $x \in A$; il suffit de prendre P tel que $P(a) \in A_0 \setminus \{0\}$ pour conclure que $f(x) = 0$.

canonique et considérons la P-algèbre

$$Q := g^{-1/p^\infty}P = \bigcap_{n\geqslant 0}\Big\{x \in P[\frac{1}{g}] |\ g^{1/p^n}x \in \iota(P)\Big\} \subset P[\frac{1}{g}].$$

L'image de g dans Q est p-puissante, multiple de p et *non diviseur de zéro*. Le morphisme $\iota\colon P \to P[\frac{1}{g}]$ induit un presque isomorphisme $\iota\colon P \to Q$ (dans le cadre g^{1/p^∞}), car $(g^{p^{-\infty}})$ annule $P[g^\infty]$ (lemme 2.20). Si P, g sont comme ci-dessus, on dira que Q est une *algèbre presque perfectoïde*.

Le résultat peu ragoûtant suivant s'obtient en combinant ceux des paragraphes ci-dessus.

Proposition 6.13. *Soit $(A, \mathfrak{m}) \in \mathrm{CLI}_p$ (déf. 6.1), $\wp \in \mathrm{Spec}A[\frac{1}{p}]$ et soient $g \in pA \setminus \wp$ et $(f_1, \dots, f_c, x_1, \dots, x_d)$ comme dans la proposition 6.9. Pour toute A-algèbre perfectoïde* CM $\mathfrak{m}$*-complète C il existe une C-algèbre $P \in \mathrm{Perf}^{\mathrm{tf}}_{K^0}$ dans laquelle g devient p-puissant (déf. 2.2) et non nul, s'insérant dans un diagramme commutatif*

$$\begin{array}{ccc} A & \longrightarrow & A/\wp \\ \downarrow & & \downarrow \\ P & \longrightarrow & Q := g^{-1/p^\infty}P \end{array}$$

et telle que l'image de $(x_1, \dots, x_d)$ dans P soit une suite régulière et presque régulière (dans le cadre g^{1/p^∞}, déf. 6.10) d'éléments p-puissants.

Démonstration. Par le lemme 6.5 on peut trouver une A-algèbre perfectoïde CM, $\mathfrak{m}$-complète C_1 qui est une C-algèbre dans laquelle $g, f_1, \dots, f_c, x_1, \dots, x_d$ deviennent p-puissants. Soit $\tilde{C}_1 = C_1/(f_1^{p^{-\infty}}, \dots, f_c^{p^{-\infty}})C_1$. La proposition 6.11 montre que $P := \varprojlim_n \tilde{C}_1/p^n \in \mathrm{Perf}^{\mathrm{tf}}_{K^0}$ et que $(x_1, \dots, x_d)$ est une suite régulière et presque régulière (dans le cadre g^{1/p^∞}) dans P. En particulier $g \neq 0$ dans P, sinon $g^{1/p^n} = 0$ pour tout n (car P est réduit, *cf.* proposition 2.20) et donc $(g^{p^{-\infty}})$ annule $P/(x_1, \dots, x_d)$, une contradiction.

Il reste à vérifier que le morphisme $A \to P$ se factorise $A/\wp \to g^{-1/p^\infty}P$. Si $f \in \wp$ alors $gf \in \sqrt{(f_1, \dots, f_c)}$, donc gf devient nilpotent dans $\tilde{C}_1$ et donc aussi dans P. Comme P est réduit (proposition 2.20), l'image de f dans P arrive dans $P[g]$ et $A \to P$ se factorise $A/\wp \to P/P[g^\infty] \to g^{-1/p^\infty}P$. □

Proposition 6.14. *Pour tout $A \in \mathrm{CLI}_p$ (déf. 6.1) il existe un système de paramètres $(x_1, \dots, x_d)$ de A avec $x_1 = p$, un élément $g \in A$ et un objet $P \in \mathrm{Perf}^{\mathrm{tf}}_{K^0}$ tels que g soit p-puissant (déf. 2.2) et non nul dans P et que $Q := g^{-1/p^\infty}P$ soit une A-algèbre dans laquelle $(x_1, \dots, x_d)$ est une suite presque régulière (dans le cadre g^{1/p^∞}, déf. 6.10).*

Démonstration. Soit k le corps résiduel de A. Il existe n et une surjection $A_0 := W(k)[[T_1, \dots, T_n]] \to A$. Soit $\wp = \ker(A_0 \to A)$, donc $A \simeq A_0/\wp$. Le théorème 6.6 fournit une A_0-algèbre CM perfectoïde et $\mathfrak{m}_{A_0}$-complète C. On conclut en appliquant la proposition ci-dessus à ces données (avec A_0 à la place de A), et en remarquant que P et Q sont presque isomorphes dans le cadre g^{1/p^∞}. □

6.5. La construction de Gabber

Pour finir la preuve du théorème 6.2 il faut passer d'une presque algèbre de Cohen–Macaulay à une vraie. Voir la proposition 4.1.2 d'André (2018b) pour la méthode des *modifications partielles* de Hochster, nous allons présenter la construction de Gabber (un peu modifiée), qui est miraculeusement élémentaire et directe. Voir aussi la section 17.5 du livre de Gabber et Ramero (2018) pour des compléments.

Soit A un anneau et soit $g \in A$ un élément p-puissant (déf. 2.2) non nul. Soit

$$S = \{(g^{a_n})_{n \geqslant 0} \in A^{\mathbf{N}} |\ a_n \in \mathbf{N}[\frac{1}{p}],\ \lim_{n\to\infty} a_n = 0\}.$$

Il est évident que S est une partie multiplicative de $A^{\mathbf{N}}$, qui ne contient pas $0 := (0, 0, \dots)$. Si M est un A-module, alors $M^{\mathbf{N}}$ est un $A^{\mathbf{N}}$-module, et on définit

$$\mathscr{G}(M) := S^{-1}M^{\mathbf{N}}.$$

Lemme 6.15. *Si $f\colon M \to N$ est un presque isomorphisme de A-modules (dans le cadre g^{1/p^∞}), alors le morphisme induit $\mathscr{G}(f)\colon \mathscr{G}(M) \to \mathscr{G}(N)$ est un isomorphisme.*

Démonstration. Écrivons simplement (x_n) au lieu de $(x_n)_{n\geqslant 0}$. Si $\mathscr{G}(f)(\frac{(m_n)}{s}) = 0$ alors $\frac{(f(m_n))}{s} = 0$ dans $S^{-1}N^{\mathbf{N}}$, donc il existe $s' = (g^{a'_n}) \in S$ tel que $s'(f(m_n)) = 0$, autrement dit $g^{a'_n} f(m_n) = 0$ pour tout n. Puisque $(g^{p^{-\infty}})$ annule $\ker f$, on obtient $g^{a'_n + 1/p^n} m_n = 0$. Ainsi $s'' := (g^{a'_n + 1/p^n}) \in S$ et $s''(m_n) = 0$ dans $M^{\mathbf{N}}$, donc $\frac{(m_n)}{s} = 0$.

Ensuite, soit $\frac{(k_n)}{s} \in \mathscr{G}(N)$. Puisque $(g^{p^{-\infty}})$ annule $\operatorname{coker}(f)$, pour tout n il existe $m_n \in M$ tel que $g^{1/p^n} k_n = f(m_n)$. Posons $s' = s \cdot (g^{1/p^n}) \in S$, alors $\frac{(k_n)}{s} = \mathscr{G}(f)(\frac{(m_n)}{s'})$, ce qui permet de conclure. □

Si C est une A-algèbre, alors $\mathscr{G}(C)$ est aussi une A-algèbre, via le morphisme diagonal $A \to A^{\mathbf{N}}$.

Lemme 6.16. *Soit C une A-algèbre et soit $x = (x_1, \dots, x_d)$ une suite dans A. Si x est presque régulière (dans le cadre g^{1/p^∞}) dans C, alors x devient une suite régulière dans $\mathscr{G}(C)$.*

Démonstration. Montrons d'abord que $\mathscr{G}(C)/(x_1,\dots,x_d)\mathscr{G}(C) \neq \{0\}$. Sinon, il existe $\alpha_i \in S^{-1}C^{\mathbf{N}}$ tels que $1 = \sum_{i=1}^d x_i\alpha_i$, donc $1 = \sum_{i=1}^d x_i \frac{y_i}{s}$ pour certains $s \in S$ et $y_i \in C^{\mathbf{N}}$. Il existe $s' \in S$ tel que $s'(s - \sum_{i=1}^d x_i y_i) = 0$ dans $C^{\mathbf{N}}$. Si $s = (g^{a_n})_{n\geqslant 0}$ et $s' = (g^{a'_n})_{n\geqslant 0}$, on en déduit (projeter sur la n-ième composante) que $g^{a_n+a'_n} \in (x_1,\dots,x_d)C$ pour tout n. Puisque $a_n + a'_n \to 0$, il s'ensuit que $(g^{p^{-\infty}})$ annule $C/(x_1,\dots,x_d)$, contredisant le fait que x est presque régulière dans C.

Il nous reste à vérifier que si $a \in \mathscr{G}(C)$ vérifie $ax_{i+1} \in (x_1,\dots,x_i)\mathscr{G}(C)$, alors $a \in (x_1,\dots,x_i)\mathscr{G}(C)$. Il existe $s \in S$ et $b, z_k \in C^{\mathbf{N}}$ tels que $a = \frac{b}{s}$ et $ax_{i+1} = \sum_{k=1}^i x_k \frac{z_k}{s}$, et il existe $s' \in S$ tel que $s'(bx_{i+1} - \sum_{k=1}^i x_k z_k) = 0$. Si $s' = (g^{a'_n})$, alors (par projection sur la n-ième composante) $g^{a'_n} b_n x_{i+1} \in (x_1,\dots,x_i)C$. Comme $(g^{p^{-\infty}})$ annule $\frac{(x_1,\dots,x_i)C:(x_{i+1})C}{(x_1,\dots,x_i)C}$, il s'ensuit que $g^{a'_n+1/p^n} b_n \in (x_1,\dots,x_i)C$ pour tout n. On peut donc écrire

$$g^{a'_n+1/p^n} b_n = \sum_{k=1}^i x_k u_{n,k}$$

pour certains $u_{n,k} \in C$. Si l'on pose $u_k = (u_{n,k})_{n\geqslant 0} \in C^{\mathbf{N}}$ et $s'' = (g^{1/p^n})_{n\geqslant 0}$, cela s'écrit $s's''b = \sum_{k=1}^i x_k u_k$ dans $C^{\mathbf{N}}$, donc $a = \frac{b}{s} = \sum_{k=1}^i x_k \frac{u_k}{ss's''} \in (x_1,\dots,x_i)\mathscr{G}(C)$. □

Lemme 6.17. *Si C est un anneau perfectoïde, alors le complété p-adique de $\mathscr{G}(C)$ est un anneau perfectoïde.*

Démonstration. $C^{\mathbf{N}}$ est un produit d'anneaux perfectoïdes, donc un anneau perfectoïde. Il suffit donc de montrer que le complété p-adique d'un localisé d'un anneau perfectoïde est encore perfectoïde, *cf.* cor. 2.1.6 de Česnavičius et Scholze, 2019. □

6.6. Construction d'algèbres perfectoïdes de Cohen–Macaulay, fonctorialité faible

Nous pouvons maintenant mettre ensemble les résultats ci-dessus et obtenir une preuve de l'existence de A-algèbres de Cohen–Macaulay pour tout $A \in \mathrm{CLI}_p$. Ceci a été démontré pour la première fois dans l'article d'André (2018b), et raffiné ensuite dans celui de Shimomoto (2018) sous la forme suivante :

Théorème 6.18. *Pour tout $(A,\mathfrak{m}) \in \mathrm{CLI}_p$ (def. 6.1) il existe une A-algèbre perfectoïde* CM *(déf. 6.3) et $\mathfrak{m}$-complète.*

Démonstration. Soient $g, x_1,\dots,x_d, P$ et Q comme dans la proposition 6.14. Le presque isomorphisme $P \to Q$ induit un isomorphisme $\mathscr{G}(P) \simeq \mathscr{G}(Q)$ (lemme 6.15), donc $\mathscr{G}(Q) \simeq \mathscr{G}(P)$ devient un anneau perfectoïde après complétion p-adique (lemme 6.17). La suite $(x_1,\dots,x_d)$ devient régulière dans $\mathscr{G}(Q)$ (lemme 6.16), en particulier $\mathscr{G}(Q) \in \mathrm{Perf}_{K^0}^{\mathrm{tf}}$. Le lemme 6.4 permet de conclure que le complété $\mathfrak{m}$-adique de $\mathscr{G}(Q)$ est une A-algèbre perfectoïde CM et $\mathfrak{m}$-complète. □

Remarque 6.19. Il n'est pas difficile d'en déduire que tout anneau local noethérien complet $(A, \mathfrak{m})$ d'inégale caractéristique $(0, p)$ admet une A-algèbre perfectoïde CM (déf. 6.3) $\mathfrak{m}$-complète. En effet, on voit facilement (utiliser le théorème de Cohen) qu'il existe une A-algèbre locale noethérienne complète $\tilde{A}$, dont le corps résiduel est algébriquement clos, et fidèlement plate sur A. Soit $\wp$ un idéal premier minimal de $\tilde{A}$ tel que $\dim \tilde{A}/\wp = \dim \tilde{A}$. Alors $\tilde{A}/\wp \in \mathrm{CLI}_p$, donc il existe une $\tilde{A}/\wp$-algèbre C perfectoïde CM, qui est automatiquement une $\tilde{A}$-algèbre perfectoïde CM. Puisque $A \to \tilde{A}$ est fidèlement plat, tout système de paramètres de A s'étend en un système de paramètres de $\tilde{A}$, et donc devient une suite régulière dans C. Ainsi C est une A-algèbre perfectoïde CM. Il suffit de compléter C pour la topologie $\mathfrak{m}$-adique et d'appliquer le lemme 6.4 pour conclure.

La méthode employée ci-dessus est assez souple pour retrouver l'un des résultats fondamentaux de l'article d'André (2020) :

Théorème 6.20. *Soit $f : A \to A'$ un morphisme surjectif dans* CLI_p *(déf. 6.1). Pour toute A-algèbre perfectoïde* CM *et $\mathfrak{m}_A$-complète C il existe une A'-algèbre perfectoïde* CM *et $\mathfrak{m}_{A'}$-complète C' s'insérant dans un diagramme commutatif*

$$\begin{array}{ccc} A & \longrightarrow & A' \\ \downarrow & & \downarrow \\ C & \longrightarrow & C' \end{array}$$

Démonstration. On peut supposer que f est la projection canonique $A \to A/\wp$ pour un $\wp \in \mathrm{Spec} A[\frac{1}{p}]$. Il suffit de remplacer l'usage de la proposition 6.14 dans la preuve ci-dessus par celui de la proposition 6.13. En effet, les arguments utilisés dans la preuve du théorème 6.18 montrent que le complété $\mathfrak{m}$-adique C' de $\mathscr{G}(Q)$ (avec Q comme dans la proposition 6.13) répond à l'appel : c'est une A'-algèbre perfectoïde CM et $\mathfrak{m}_{A'}$-complète, et on dispose d'un morphisme $P \to C'$, donc aussi d'un morphisme $C \to C'$, qui s'insère dans un diagramme comme dans le théorème par construction. □

Remarque 6.21.

1. On trouve dans le théorème 4.1.1 d'André (2020) et dans le livre de Gabber et Ramero (2018) des formes plus raffinées et générales, mais le théorème ci-dessus contient déjà bon nombre de difficultés essentielles.
2. En utilisant des factorisations de Cohen, on peut en déduire que tout morphisme $f : A \to A'$ dans CLI_p s'insère dans un diagramme comme ci-dessus, pour une A-algèbre CM C et une A'-algèbre CM C'. Ceci a des multiples applications, voir par exemple l'article de Hochster et Huneke (1995). Mentionnons simplement une conséquence frappante : si A est un anneau local régulier, extension scindée d'un sous-anneau A', alors A' est un anneau de Cohen–Macaulay [(57)].

[(57)] Ceci avait été obtenu en inégale caractéristique par Heitmann et Ma (2018), avant l'article d'André (2020).

7. Algèbres de Cohen–Macaulay via le lemme d'Abhyankar perfectoïde

Dans cette dernière section nous présentons l'approche initiale d'André (2018b) pour construire des algèbres de Cohen–Macaulay, en utilisant le lemme d'Abhyankar perfectoïde. Il est possible de pousser cette méthode pour obtenir une preuve du théorème 6.20 (c'est ce qui est fait dans l'article d'André (2020)), mais on ne le fera pas ici.

Pour toute K^0-algèbre R notons $R^\natural = W(R^\flat)/(p-[p^\flat])$. Si $P \in \mathrm{Perf}^{\mathrm{tf}}_{K^0}$ alors $P^\natural \simeq P$ via θ_P, donc si R est une P-algèbre, alors $R^\natural$ est une $P \simeq P^\natural$-algèbre. Si R est p-complète, tout morphisme $P \to R$ se factorise canoniquement $P \to R^\natural \to R$.

7.1. La construction fondamentale

Motivés par la proposition 6.13, considérons le contexte suivant. On se donne

• un morphisme $A' \to A$ dans CLI_p (déf. 6.1) et un élément $g \in pA'$.

• une algèbre $P \in \mathrm{Perf}^{\mathrm{tf}}_{K^0}$ dans laquelle g devient p-puissant et non nul et s'insérant dans un diagramme commutatif

$$\begin{array}{ccc} A' & \longrightarrow & A \\ \downarrow & & \downarrow \\ P & \longrightarrow & Q := g^{-1/p^\infty}P \end{array}$$

Notons que l'image de g dans Q est non nulle, car P est réduit.

Soit $A \to A_1$ une extension finie, étale après inversion de g, et posons [58]

$$\mathscr{F}(A_1) = \mathrm{fi}(Q, A_1 \otimes_A Q[\frac{1}{g}]).$$

Puisque $A \to A_1$ est entier, le morphisme naturel $A_1 \to A_1 \otimes_A Q[\frac{1}{g}]$ se factorise à travers $\mathscr{F}(A_1)$, induisant ainsi un diagramme commutatif

$$\begin{array}{ccc} A & \longrightarrow & A_1 \\ \downarrow & & \downarrow \\ Q & \longrightarrow & \mathscr{F}(A_1) \end{array}$$

Considérons la P-algèbre

$$\mathscr{P}(A_1) := \mathscr{F}(A_1)^\natural.$$

[58] Rappelons que $\mathrm{fi}(A,B)$ désigne la clôture intégrale de A and B.

Il n'est pas clair que $\mathscr{F}(A_1)$ soit p-complète, mais la preuve du résultat ci-dessus montrera que c'est bien le cas, donc le morphisme $P \to Q \to \mathscr{F}(A_1)$ se factorise $P \to \mathscr{P}(A_1) \to \mathscr{F}(A_1)$.

Proposition 7.1. *Soit $A \to A_1$ une extension finie, étale après inversion de g. La P-algèbre $\mathscr{P}(A_1)$ est presque fidèlement plate (dans le cadre g^{1/p^∞}) modulo p sur P, sans g-torsion et $g^{-1/p^\infty}\mathscr{P}(A_1) = \mathscr{F}(A_1)$.*

Ce résultat est une application du lemme d'Abhyankar perfectoïde, mais la vérification des hypothèses demande quelques préliminaires.

Lemme 7.2. *Si $S \in \{Q, Q^\natural\}$ alors l'algèbre S est p-complète, sans g-torsion, et p-close dans $S[\frac{1}{p}]$.*

Démonstration. Posons $R = Q^\natural$. Comme Q est sans g-torsion et $p \mid g$, l'algèbre Q est sans p-torsion. Il en est de même de R puisque $p^\flat$ n'est pas un diviseur de zéro dans $Q^\flat$: si $x = (x_n)_{n\geqslant 0} \in Q^\flat$ vérifie $p^\flat x = 0$, et si $a_n \in Q$ est un relèvement de $x_n \in Q/p$, alors $a_n \in p^{1-1/p^n}Q$ et $a_n \equiv a_{n+1}^p \equiv 0 \pmod{pQ}$ pour tout n, donc $x = 0$.

L'anneau R est p-complet car perfectoïde (proposition 2.8). Montrons que Q est p-complet. Soit $g_n = g^{\frac{1}{p^n} - \frac{1}{p^{n+1}}}$. Le morphisme $\iota\colon P \to P[\frac{1}{g}]$ induit un isomorphisme [(59)]

$$\alpha\colon Q \to \varprojlim_{\cdot g_n} \iota(P),\ x \mapsto (g^{1/p^n}x)_{n\geqslant 0}.$$

Puisque $I := (g^{p^{-\infty}})$ annule le noyau de $\iota\colon P \to \iota(P)$ et contient les g_n, le morphisme ι induit un isomorphisme de pro-systèmes $\{P, (\cdot g_n)\} \to \{\iota(P), (\cdot g_n)\}$, d'où

$$Q \simeq \varprojlim_{\cdot g_n} \iota(P) \simeq \varprojlim_{\cdot g_n} P.$$

Comme P est p-complet et sans p-torsion, on en déduit facilement que Q est p-complet.

Montrons que Q (resp. R) est p-clos dans $Q[\frac{1}{p}]$ (resp. $R[\frac{1}{p}]$). La proposition 3.5 montre que P (resp. R) est p-clos dans $P[\frac{1}{p}]$ (resp. $R[\frac{1}{p}]$). Pour conclure il suffit de montrer que si $x \in Q$ vérifie $x^p \in pQ$, alors $\frac{x}{p^{1/p}} \in P[\frac{1}{g}]$ est dans Q, ou encore que $\alpha\beta\frac{x}{p^{1/p}} \in \iota(P)$ pour tous $\alpha, \beta \in I$. Comme $x \in Q$ et $x^p \in pQ$, il existe $a, b \in P$ tels que $\alpha x = \iota(a)$ et $(\alpha x)^p = p\iota(b)$. Alors $\beta^p(a^p - pb) = 0$ (car $a^p - pb \in \ker(\iota)$ et I annule $\ker(\iota)$), donc $(\beta a)^p \in pP$. Comme P est p-clos dans $P[\frac{1}{p}]$, cela force $\beta a \in p^{1/p}P$ et $\alpha\beta\frac{x}{p^{1/p}} = \iota(\frac{\beta a}{p^{1/p}}) \in \iota(P)$, ce qui permet de conclure. □

[(59)] L'injectivité est claire, et si $(x_n)_{n\geqslant 0} \in \varprojlim_{\cdot g_n} \iota(P)$ alors la suite $(g^{-1/p^n}x_n)_{n\geqslant 0}$ est constante dans $P[\frac{1}{g}]$, sa valeur x est dans Q par définition et $\alpha(x) = (x_n)_{n\geqslant 0}$.

Lemme 7.3. *Le morphisme $\theta_Q\colon Q^\natural \to Q$ est injectif et $(g^{p^{-\infty}})$ annule son conoyau. On a donc $Q = g^{-1/p^\infty}Q^\natural \subset Q^\natural[\frac{1}{g}]$ et $Q^\natural$ est sans g-torsion.*

Démonstration. Notons encore $R = Q^\natural$. Montrons que $\theta_Q\colon R \to Q$ est injective, en particulier g n'est pas un diviseur de zéro dans R. Comme R et Q sont sans p-torsion et p-complets par la proposition ci-dessus, il suffit de montrer l'injectivité de $R/p \simeq Q^\flat/p^\flat \to Q/p$, qui se déduit de celle du Frobenius $Q/p^{1/p} \to Q/p$, *cf.* lemme. 7.2. Puisque $\iota\colon P \to Q$ est un presque isomorphisme dans le cadre g^{1/p^∞} et se factorise $P \to R \to Q$, pour conclure il suffit de montrer que $P \to R$ est un presque isomorphisme, ou encore qu'il en est de même de $\iota^\flat\colon P^\flat \to Q^\flat$, ce qui est clair. □

Lemme 7.4. *La K-algèbre $B := Q^\natural[\frac{1}{p}]$ possède une structure de K-algèbre de Banach perfectoïde sans g-torsion, telle que $g^{-1/p^\infty}B^0 = Q$, en particulier $B[\frac{1}{g}] = Q[\frac{1}{g}] \simeq P[\frac{1}{g}]$.*

Démonstration. Par la proposition 3.5 l'algèbre $Q^\natural_* := p^{-1/p^\infty}Q^\natural$ est perfectoïde, donc $B := Q^\natural[\frac{1}{p}]$ possède une structure de K-algèbre de Banach perfectoïde sans g-torsion, telle que $B^0 = Q^\natural_*$. Comme $p \mid g$ et $Q = g^{-1/p^\infty}Q^\natural$ (lemme 7.3), on a

$$g^{-1/p^\infty}B^0 = g^{-1/p^\infty}Q^\natural_* = g^{-1/p^\infty}Q^\natural = Q,$$

ce qui permet de conclure. □

Revenons maintenant à la preuve de la proposition 7.1. Par le lemme 7.4 on peut munir $B := Q^\natural[\frac{1}{p}]$ d'une structure de K-algèbre de Banach perfectoïde sans g-torsion, telle que $g^{-1/p^\infty}B^0 = Q$, en particulier $B[\frac{1}{g}] = Q[\frac{1}{g}] \simeq P[\frac{1}{g}]$. Puisque $A \to A_1$ est étale après inversion de g, l'algèbre $C = A_1 \otimes_A B[\frac{1}{g}] = A_1 \otimes_A Q[\frac{1}{g}]$ est finie, étale et fidèlement plate sur $B[\frac{1}{g}]$. Le lemme d'Abhyankar perfectoïde [(60)] montre que $\mathscr{F}(A')$ est presque fidèlement plat sur B^0 modulo p, donc aussi sur P modulo p (car B^0 et P sont presque isomorphes dans le cadre g^{1/p^∞}), et le morphisme $\mathscr{P}(A_1) = \mathscr{F}(A_1)^\natural \to \mathscr{F}(A_1)$ est injectif, de conoyau tué par $(g^{p^{-\infty}})$. Cela permet de conclure.

7.2. Nouvelle preuve de la proposition 6.14

La proposition 7.1 et le lemme de platitude d'André fournissent une nouvelle preuve [(61)] de la proposition 6.14. Comme expliqué dans la section précédente, ceci implique le théorème 6.18 via la construction de Gabber.

Soit $A \in \mathrm{CLI}_p$ et soit $(x_1, \dots, x_d)$ un système de paramètres, avec $x_1 = p$. Le morphisme de $W(k)$-algèbres $A_0 := W(k)[[T_2, \dots, T_d]] \to A, T_i \mapsto x_i$ est fini et injectif. Soit $g \in pA_0 \setminus \{0\}$ tel que $A_0[\frac{1}{g}] \to A[\frac{1}{g}]$ soit étale, et soit $P \in \mathrm{Perf}^{\mathrm{tf}}_{K^0}$ une

[(60)] Noter que, compte tenu de la discussion ci-dessus on a $\mathscr{F}(A') = \tilde{C}^0$ dans les notations du théorème 5.3.
[(61)] Il s'agit en fait de la première preuve de ce résultat, due à André (2018b).

A_0-algèbre fidèlement plate modulo p sur A_0, dans laquelle g devient p-puissant (*cf.* théorème 6.6 et 4.3 ; le théorème 4.1 ferait aussi l'affaire).

Lemme 7.5. *La suite $(x_1, \dots, x_d)$ est régulière et presque régulière dans P (dans le cadre g^{1/p^∞}) et l'image de g dans P n'est pas nulle.*

Démonstration. Si l'image de g dans P était nulle, elle le serait modulo p^n pour tout n, or $A_0/p^n \to P/p^n$ est fidèlement plat, donc $g \in \cap_{n\geqslant 1} p^n A_0 = \{0\}$, une contradiction avec $g \in A_0 \setminus \{0\}$. Ensuite, comme P est fidèlement plat modulo p sur A_0 et $x_1 = p$, la suite est régulière dans P. Il reste à expliquer pourquoi l'inclusion $(g^{p^{-\infty}})P \subset (x_1, \dots, x_d)P$ est impossible. Si elle avait lieu, on aurait $g \in (x_1, \dots, x_d)^k P$ pour tout k, et comme P/p^n est fidèlement plat sur A_0/p^n pour tout n, cela forcerait $g \in \bigcap_k (x_1, \dots, x_d)^k$, puis $g = 0$ par le théorème d'intersection de Krull. □

En appliquant la proposition 7.1 au morphisme identité $A_0 \to A_0$, avec $A_1 = A$ on obtient une P-algèbre perfectoïde $\mathscr{P}(A)$ presque fidèlement plate modulo p sur P, dans laquelle g est non diviseur de zéro et qui s'insère dans un diagramme commutatif

$$\begin{array}{ccc} A_0 & \longrightarrow & A \\ \downarrow & & \downarrow \\ P & \longrightarrow & g^{-1/p^\infty}\mathscr{P}(A) \end{array}$$

L'algèbre $\mathscr{P}(A)$ répond à l'appel lancé par la proposition 6.14 grâce au lemme ci-dessous :

Lemme 7.6. *Soit R un anneau sans p-torsion, $g \in R$ un élément p-puissant et $x = (x_1, \dots, x_d)$ une suite dans R, avec $x_1 = p$. Si x est presque régulière dans R (dans le cadre g^{1/p^∞}), elle le reste dans toute R-algèbre S qui est presque fidèlement plate sur R modulo p.*

Démonstration. Soit $I = (g^{p^{-\infty}}) \subset R$. Puisque $S/(x_1, \dots, x_d)$ est presque fidèlement plat sur $R/(x_1, \dots, x_d)$ (car $x_1 = p$) le morphisme $R/(x_1, \dots, x_d) \to S/(x_1, \dots, x_d)$ est presque injectif, donc $S/(x_1, \dots, x_d)$ n'est pas presque nul. Pour montrer que I annule $\frac{(x_1,\dots,x_i)S : x_{i+1}S}{(x_1,\dots,x_i)S}$ on peut supposer que $i > 0$ car S est sans p-torsion. Notons $\bar{R} := R/p$, $\bar{S} := S/p$. Il suffit de montrer que I annule $\frac{(x_2,\dots,x_i)\bar{S} : x_{i+1}\bar{S}}{(x_2,\dots,x_i)\bar{S}}$. Par presque platitude de $\bar{S}$ sur $\bar{R}$ on a un presque isomorphisme [(62)]

$$(x_2, \dots, x_i)\bar{S} : x_{i+1}\bar{S} \simeq ((x_2, \dots, x_i)\bar{R} : x_{i+1}\bar{R})\bar{S},$$

ce qui permet de conclure. □

[(62)] Il suffit de tensoriser avec $\bar{S}$ la suite exacte

$$0 \to (x_2, \dots, x_i) : x_{i+1}\bar{R} \to \bar{R} \to \bar{R}/(x_2, \dots, x_i).$$

Références

André, Y. (2018a). « Le lemme d'Abhyankar perfectoide », *Publ. Math., Inst. Hautes Étud. Sci.* **127**, p. 1-70.

——— (2018b). « La conjecture du facteur direct », *Publ. Math., Inst. Hautes Étud. Sci.* **127**, p. 71-93.

——— (2018c). « Perfectoid spaces and the homological conjectures », in : *Proceedings of the international congress of mathematicians, ICM 2018, Rio de Janeiro, Brazil, August 1–9, 2018. Volume II. Invited lectures*. Hackensack, NJ : World Scientific ; Rio de Janeiro : Sociedade Brasileira de Matemática (SBM), p. 277-289.

——— (2020). « Weak functoriality of Cohen–Macaulay algebras », *J. Am. Math. Soc.* **33** (2), p. 363-380.

André, Y. et Fiorot, L. (2022). « On the canonical, fpqc, and finite topologies on affine schemes. The state of the art », *Ann. Sc. Norm. Super. Pisa, Cl. Sci.* (5) **23** (1), p. 81-114.

Bartenwerfer, W. (1976). « Der erste Riemannsche Hebbarkeitssatz im nichtarchimedischen Fall », *J. Reine Angew. Math.* **286/287**, p. 144-163.

Bartijn, J. et Strooker, J. R. (1983). *Modifications monomiales*. Semin. d'algèbre Paul Dubreil et Marie-Paule Malliavin, 35ème Année, Proc., Paris 1982, Lect. Notes Math. 1029, 192-217 (1983).

Bhatt, B. (2012). « Derived splinters in positive characteristic », *Compos. Math.* **148** (6), p. 1757-1786.

——— (2014a). « Almost direct summands », *Nagoya Math. J.* **214**, p. 195-204.

——— (2014b). « On the non-existence of small Cohen–Macaulay algebras », *J. Algebra* **411**, p. 1-11.

——— (2018). « On the direct summand conjecture and its derived variant », *Invent. Math.* **212** (2), p. 297-317.

——— (2020). « Cohen–Macaulayness of absolute integral closures ». arXiv.

Bhatt, B., Iyengar, S. B. et Ma, L. (2019). « Regular rings and perfect(oid) algebras », *Commun. Algebra* **47** (6), p. 2367-2383.

Bhatt, B. et Lurie, J. (2023). « A p-adic Riemann–Hilbert functor : $\mathbf{Z}/p^n$-coefficients ». en préparation.

Bhatt, B., Morrow, M. et Scholze, P. (2018). « Integral p-adic Hodge theory », *Publ. Math., Inst. Hautes Étud. Sci.* **128**, p. 219-397.

Bhatt, B. et Scholze, P. (2022). « Prisms and prismatic cohomology », *Ann. Math.* (2) **196** (3), p. 1135-1275.

Česnavičius, K. (2021). « Macaulayfication of Noetherian schemes », *Duke Math. J.* **170** (7), p. 1419-1455.

Česnavičius, K. et Scholze, P. (2019). « Purity for flat cohomology ». À paraître à Ann. Math.

Colmez, P. (2002). « Espaces de Banach de dimension finie », *J. Inst. Math. Jussieu* **1** (3), p. 331-439.

Colmez, P. et Fontaine, J.-M. (2000). « Construction des représentations *p*-adiques semi-stables », *Invent. Math.* **140** (1), p. 1-43.

Dine, D. (2022). « Topological spectrum and perfectoid Tate rings », *Algebra and Number Theory* **16** (6), p. 1463-1500.

Dutta, S. P. (1987). « On the canonical element conjecture », *Trans. Am. Math. Soc.* **299**, p. 803-811.

Evans, E. G. et Griffith, P. (1981). « The syzygy problem », *Ann. Math.* (2) **114**, p. 323-333.

Faltings, G. (1988). « *p*-adic Hodge theory », *J. Am. Math. Soc.* **1** (1), p. 255-299.

——— (2002). « Almost étale extensions », in : *Cohomologies p-adiques et applications arithmétiques* (*II*). T. 279. Astérisque. Paris : Société Mathématique de France, p. 185-270.

Fontaine, J.-M. (2013). « Perfectoïdes, presque pureté et monodromie-poids [d'après Peter Scholze] », in : *Séminaire Bourbaki. Volume 2011/2012. Exposés 1043–1058*. Astérisque 352. Paris : Société Mathématique de France (SMF), 509-534, ex.

Gabber, O. et Ramero, L. (2003). *Almost ring theory*. T. 1800. Lect. Notes Math. Berlin : Springer.

——— (2018). « Almost rings and perfectoid spaces ». preprint.

Heitmann, R. (2002). « The direct summand conjecture in dimension three. » *Ann. Math.* (2) **156** (2), p. 695-712.

Heitmann, R. et Ma, L. (2018). « Big Cohen–Macaulay algebras and the vanishing conjecture for maps of Tor in mixed characteristic », *Algebra Number Theory* **12** (7), p. 1659-1674.

Hochster, M. (1973). « Contracted ideals from integral extensions of regular rings », *Nagoya Math. J.* **51**, p. 25-43.

——— (1975). *Topics in the homological theory of modules over commutative rings*. T. 24. Reg. Conf. Ser. Math. American Mathematical Society (AMS), Providence, RI.

——— (1977). « Cyclic purity versus purity in excellent Noetherian rings », *Trans. Am. Math. Soc.* **231**, p. 463-488.

——— (1979). « Big and small Cohen-Macaulay modules ». In : *Module theory* (*Proc. Special Session, Amer. Math. Soc., Univ. Washington, Seattle, Wash., 1977*). T. 700. Lecture Notes in Math. Springer, Berlin, p. 119-142.

——— (1983). « Canonical elements in local cohomology modules and the direct summand conjecture », *J. Algebra* **84**, p. 503-553.

——— (1994). « Solid closure », in : *Commutative algebra : syzygies, multiplicities, and birational algebra* (*South Hadley, MA, 1992*). T. 159. Contemp. Math. Amer. Math. Soc., Providence, RI, p. 103-172.

Hochster, M. (2002). « Big Cohen–Macaulay algebras in dimension three via Heitmann's theorem. » *J. Algebra* **254** (2), p. 395-408.

——— (2007). « Homological conjectures, old and new », *Ill. J. Math.* **51** (1), p. 151-169.

Hochster, M. et Huneke, C. (1992). « Infinite integral extensions and big Cohen–Macaulay algebras », *Ann. Math.* (2) **135** (1), p. 53-89.

——— (1995). « Applications of the existence of big Cohen–Macaulay algebras », *Adv. Math.* **113** (1), p. 45-117.

Huneke, C. et Lyubeznik, G. (2007). « Absolute integral closure in positive characteristic », *Adv. Math.* **210** (2), p. 498-504.

Kedlaya, K. S. et Liu, R. (2015). *Relative p-adic Hodge theory : foundations*. T. 371. Astérisque. Paris : Société Mathématique de France (SMF).

Kovács, S. J. (2000). « A characterization of rational singularities », *Duke Math. J.* **102** (2), p. 187-191.

Kunz, E. (1969). « Characterizations of regular local rings of characteristic *p* », *Am. J. Math.* **91**, p. 772-784.

Ma, L. (2021). « A short proof of the direct summand theorem via the flatness lemma ».

Ma, L. et Schwede, K. (2018). « Perfectoid multiplier/test ideals in regular rings and bounds on symbolic powers », *Invent. Math.* **214** (2), p. 913-955.

——— (2021). « Singularities in mixed characteristic via perfectoid big Cohen–Macaulay algebras », *Duke Math. J.* **170** (13), p. 2815-2890.

Ma, L., Schwede, K. et al. (2022). « An analog of adjoint ideals and PLT singularities in mixed characteristic », *J. Algebr. Geom.* **31**, p. 497-559.

Morrow, M. (2019). « The Fargues-Fontaine curve and diamonds [d'après Fargues, Fontaine, and Scholze] », in : *Séminaire Bourbaki. Volume 2017/2018. Exposés 1136–1150*. Astérisque 414. Paris : Société Mathématique de France (SMF), 533-572, ex.

Ohi, T. (1996). « Direct summand conjecture and descent for flatness », *Proc. Am. Math. Soc.* **124** (7), p. 1967-1968.

Olivier, J.-P. (1973). « Descente de quelques propriétés élémentaires par morphismes purs », *Anais Acad. Brasil. Ci.* **45**, p. 17-33.

Peskine, C. et Szpiro, L. (1972). « Dimension projective finie et cohomologie locale. Applications à la demonstration de conjectures de M. Auslander, H. Bass et A. Grothendieck », *Publ. Math., Inst. Hautes Étud. Sci.* **42**, p. 47-119.

Raynaud, M. et Gruson, L. (1971). « Critères de platitude et de projectivité. Techniques de "platification" d'un module », *Invent. Math.* **13**, p. 1-89.

Roberts, P. (1987). « Le théorème d'intersection », *C. R. Acad. Sci., Paris, Sér. I* **304**, p. 177-180.

——— (1992). « The homological conjectures », in : *Free resolutions in commutative algebra and algebraic geometry* (*Sundance, UT, 1990*). T. 2. Res. Notes Math. Jones et Bartlett, Boston, MA, p. 121-132.

——— (2008). « The root closure of a ring of mixed characteristic ». preprint.

Scholze, P. (2012). « Perfectoid spaces », *Publ. Math., Inst. Hautes Étud. Sci.* **116**, p. 245-313.

——— (2013). « p-adic Hodge theory for rigid-analytic varieties », *Forum Math. Pi* **1**, e1, 77.

——— (2015). « On torsion in the cohomology of locally symmetric varieties », *Ann. Math.* (2) **182** (3), p. 945-1066.

Shimomoto, K. (2018). « Integral perfectoid big Cohen–Macaulay algebras via André's theorem », *Math. Ann.* **372** (3-4), p. 1167-1188.

Gabriel Dospinescu

UMPA, ENS Lyon, CNRS

E-mail : gabriel.dospinescu@ens-lyon.fr

Séminaire BOURBAKI Janvier 2023
74e année, 2022–2023, no 1202, p. 199 à 223
doi: 10.24033/ast.1211

STRONG CONVERGENCE OF THE SPECTRUM OF RANDOM PERMUTATIONS AND ALMOST-RAMANUJAN GRAPHS
[*after* **Charles Bordenave and Benoît Collins**]

by **Mylène Maïda**

1. Introduction

Consider the two following statements:

> *Independent random permutations, chosen uniformly among all permutations or all matchings of n points, are strongly asymptotically free* (*viewed as operators on the orthogonal of the constant vector* **1**).

versus

> *Random n-lifts of a fixed weighted base graph are close to being Ramanujan graphs.*

They seem to belong to different mathematical landscapes, random matrix theory and free probability for the first one, theory of expander graphs for the second one. They are nevertheless two instances of the same result, due to Bordenave and Collins (2019) and that we will present hereafter. In particular, we will try to explain the meaning of the statement in each context and why it represents an important improvement with respect to the previous results, starting with the motivation from graph theory and then moving to free probability. This is not the only example of a result dealing with strong asymptotic freeness that can be applied to a completely different context and we will describe in detail, in the last part of these notes, some other applications of this notion.

Je remercie Charles Bordenave pour son aide lors de la préparation de ces notes. Elles ont été rédigées pendant mon séjour au Centre de recherches mathématiques de Montréal (IRL CNRS 3457) que je remercie pour son hospitalité. Je remercie Thierry Lévy et Guillaume Dubach pour leur relecture attentive du manuscrit, Gilles Pisier et Mikael de la Salle pour une intéressante discussion autour du lemme de linéarisation ainsi qu'un relecteur anonyme pour son acuité orthotypographique.

2. From Ramanujan graphs to the symmetric random permutation model

The notion of *Ramanujan graph* was initially introduced by Lubotzky, Phillips, and Sarnak (1988) for d-regular graphs. The terminology *Ramanujan* comes from the fact that the construction of the regular graphs considered by Lubotzky, Phillips, and Sarnak (1988) was based on arithmetic properties of pairs of well-chosen prime numbers.

Let $G = (V, E)$ be an undirected graph, with countable vertex set V and edge set E. An edge is a subset of V with two elements (we do not allow loops nor multiple edges). The degree of a vertex $v \in V$ is defined as

$$\deg(v) := \sum_{u \in V} \mathbf{1}_{\{u,v\} \in E}.$$

If, for every $v \in V$, $\deg(v) < \infty$, the graph is said to be *locally finite* and its *adjacency operator* A is defined as follows: for any $\psi \in \ell_c(V)$, which is the subspace of $\ell^2(V)$ of vectors with finite support,

$$A\psi(v) := \sum_{u \in V; \{u,v\} \in E} \psi(u).$$

In the case when V is a finite set, A can be seen as the usual *adjacency matrix* of G. As we are dealing with undirected graphs, the adjacency operator and matrix are self-adjoint. For any integer $d \geqslant 2$, we say that G is *d-regular* if all the vertices of G have degree d. For finite d-regular graphs with n vertices, if we denote by $\lambda_0 \geqslant \lambda_1 \geqslant \cdots \geqslant \lambda_{n-1}$ the eigenvalues of A in non-increasing order, one can check that $\lambda_0 = d$ and for all $j \leqslant n-1$, $|\lambda_j| \leqslant d$. If G is connected, the eigenvalue λ_0 is always simple and if G is moreover bipartite, then λ_{n-1} is equal to $-d$ and is simple. They are often called the Perron–Frobenius eigenvalues and λ_0 is associated to the constant eigenvector $\mathbf{1}$. On the other hand, if we denote by

$$\lambda(G) := \max\{|\lambda_j| \text{ such that } |\lambda_j| < d\}$$

the largest eigenvalue in absolute value which is not equal to $\pm d$, we have the following result, known as the *Alon–Boppana bound*:

Theorem 2.1 (Alon, 1986). *Let $(G_{n,d})_{n \geqslant 1}$ be any sequence of connected d-regular graphs such that, for any $n \in \mathbb{N}^*$, $G_{n,d}$ has n vertices*[(1)]*. Then*

$$\liminf_{n \to \infty} \lambda(G_{n,d}) \geqslant 2\sqrt{d-1}.$$

[(1)] It requires that $n \geqslant d+1$ and nd is even.

This leads to the following definition:

Definition 2.2 (Ramanujan graph, d-regular case)**.** A d-regular, connected, finite graph G is called *Ramanujan* if and only if every eigenvalue λ of its adjacency operator is such that $\lambda \in \{-d, d\}$ or $|\lambda| \leqslant 2\sqrt{d-1}$.

Among connected d-regular graphs, Ramanujan graphs are the graphs with maximal spectral gap. For this reason, sequences of such graphs have very good properties as *expander graphs* (we refer to Kowalski (2019) for details on the link between spectral gap and expander properties of graphs). The question is then how to construct such sequences of graphs. In this direction, one has to mention the remarkable result of Friedman (2008):

Theorem 2.3. *Let $d \geqslant 3$ be an integer. For each $n \geqslant d+1$ such that nd is even, let G_n be a random graph chosen uniformly among d-regular graphs with n vertices. Then the sequence $(G_n)_{n\geqslant 1}$ is almost-Ramanujan*[(2)] *in the sense that, for any $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(\lambda(G_n) \geqslant 2\sqrt{d-1} + \varepsilon) = 0.$$

The notion of *random lift* is also very useful to construct almost-Ramanujan graphs. On the way of defining them, we also give a more general definition of Ramanujan graphs. Let G and H be undirected connected[(3)] graphs, with no loops and no multiple edges. A *covering map* π from H to G is a map from the set of vertices of H to the set of vertices of G such that, for every vertex h of H, π gives a bijection between the edges incident to h in H and those incident to $\pi(h)$ in G. When π is a covering map from H to G, the graph G is called the *base graph* and H is called a *covering graph* of G. Since the base graph G is connected, the cardinal of $\pi^{-1}(g)$ is the same for any vertex g of G. If this cardinal is equal to n, then H is called an *n-lift* of G. If H is a tree, it is the *universal cover* of G. In particular, the universal cover of any non-empty, connected, d-regular graph is the infinite d-regular tree $\mathbb{T}_d$. It is known that the spectrum of $\mathbb{T}_d$ is the interval $[-2\sqrt{d-1}, 2\sqrt{d-1}]$[(4)]. Therefore, for a d-regular graph, being a Ramanujan graph means that all eigenvalues, except $\pm d$, are contained inside the spectrum of its universal cover.

This led Greenberg (1995), in his thesis, to give a more general definition of a Ramanujan graph, not necessarily restricted to d-regular graphs.

Definition 2.4 (Ramanujan graph, general case)**.** A finite connected graph X is Ramanujan if the spectrum of its adjacency operator is contained in $[-\rho, \rho] \cup \{-\lambda_0, \lambda_0\}$, where λ_0 is the largest eigenvalue of the graph and ρ is the spectral radius of its universal cover.

(2) or weakly Ramanujan

(3) Covering maps can be defined in a more general framework but we only need the case of connected graphs.

(4) Its spectral measure is known as the Kesten–McKay distribution.

Let us describe a standard model, due to Amit and Linial (2002) for constructing *random lifts*. Given a base graph $G = (V_G, E_G)$ and an integer $n \geqslant 2$, to each vertex v of G we associate a set of n vertices $(v,1),\dots,(v,n)$. For each edge $\{u,v\}$ of G, we choose an orientation, say $e := (u,v)$, and a uniform permutation σ_e of $[n] := \{1,\dots,n\}$, independent of all other edges. Then, if the vertex set of H is $\{(u,i); u \in V_G, i \in [n]\}$ and the edge set is $\{\{(u,i),(v,\sigma_{(u,v)}(i))\}; \{u,v\} \in E_G, i \in [n]\}$, H is a *random n-lift* of G and its law is uniform over all possible n-lifts of G. Note that the choice of the orientations of the edges made for the construction does not change the distribution of the random lift H. Improving on Theorem 2.3, Bordenave (2020) showed that, under very general conditions on the base graph, the sequence of random n-lifts form a sequence of almost-Ramanujan graphs.

The model that we describe now is the main object studied by Bordenave and Collins (2019) and can be seen as a generalization of the notion of random lift; we will call it the *symmetric random permutation model*. Let X be a countable set. Let $\sigma_1,\dots,\sigma_d$ be d permutations of the set X. We consider $\ell^2(X)$ the Hilbert space spanned by the orthonormal basis $(\delta_x)_{x\in X}$. The identity operator on $\ell^2(X)$ is denoted by $\mathbf{1}$. A permutation σ_i acts naturally as a unitary operator S_i on $\ell^2(X)$ by $S_i(g)(x) = g(\sigma_i(x))$, for any $g \in \ell^2(X)$. Let $a_0, a_1, \dots, a_d$ be matrices of size $r \times r$. The main object of interest in Bordenave and Collins (2019) is the operator $A := a_0 \otimes \mathbf{1} + \sum_{i=1}^{d} a_i \otimes S_i$ acting on $\mathbb{C}^r \otimes \ell^2(X)$. When $X = [n]$, we denote by $\sigma_{1,n},\dots,\sigma_{d,n}$ the permutations of X, by $S_{1,n},\dots,S_{d,n}$ the corresponding operators, $\mathbf{1}^{(n)}$ the identity operator and

$$A_n := a_0 \otimes \mathbf{1}^{(n)} + \sum_{i=1}^{d} a_i \otimes S_{i,n}. \tag{1}$$

Two symmetry conditions are added, one on the matrices $a_1,\dots,a_d$ and one on the permutations $\sigma_{1,n},\dots,\sigma_{d,n}$.

Assumption 2.5 (Symmetric random permutation model)**.** We equip $[d]$ with the following involution: let $q \leqslant \frac{d}{2}$ be an integer; for any $i \in [q]$, set $i^* = i+q$, for $q+1 \leqslant i \leqslant 2q$, set $i^* = i-q$, and for $2q+1 \leqslant i \leqslant d$, set $i^* = i$. We assume that:

(H**a**) $a_0 = a_0^*$ and $\forall i \in \{1,\dots,d\}, a_{i^*} = (a_i)^*$.

(Hσ) The permutations $\{\sigma_{1,n},\dots,\sigma_{q,n}\} \cup \{\sigma_{2q+1,n},\dots,\sigma_{d,n}\}$ on $[n]$ are independent, $\{\sigma_{1,n},\dots,\sigma_{q,n}\}$ are uniformly distributed among the permutations of $[n]$ and $\{\sigma_{2q+1,n},\dots,\sigma_{d,n}\}$ are uniformly distributed among the *matchings*[(5)] of $[n]$ and for any $i \in [d]$, $\sigma_{i^*,n} = (\sigma_{i,n})^{-1}$.

[(5)] A matching (or pair matching) is a permutation for which all the cycles are of length 2, that is an involution without fixed point.

Let us explain how it can be seen as a generalization of the model of random lift. Assume that d is even, $q = d/2$, $a_0 = 0$ and the matrices $a_1, \dots, a_d$ are of the form $a_i = E_{u_i v_i}$, with $u_i, v_i \in [r]$. The base graph will be the graph G with vertex set $[r]$ and adjacency matrix $A_1 = \sum_{i=1}^{d} a_i$. Under Assumption (H**a**), it will be undirected, with $d/2$ edges. The graph H with vertex set $[n] \times [r]$ and edges of the form $\{(x, u_i), (\sigma_i(x), v_i)\}$ is a n-lift of G. If the permutations $\sigma_{1,n}, \dots, \sigma_{d,n}$ fulfill Assumption (Hσ), then the random lift we obtain has the same distribution as in the construction of Amit and Linial (2002).

Bordenave and Collins (2019) show that the $A_n, n \geqslant 1$, are the adjacency operators of an almost-Ramanujan sequence of weighted graphs, in a sense related to Definition 2.4 (we refer to Theorem 3.13 for a precise statement). But in parallel to this graph-theoretical motivation, the symmetric random permutation model is linked with asymptotic freeness properties of random permutations and we develop this point of view in the next section.

3. Freeness, asymptotic freeness and strong asymptotic freeness

In the eighties, Dan Voiculescu introduced the concept of *freeness* (or *free independence*) in the context of operator algebras and created the field of *free probability theory*. In the early nineties, he discovered that many models of random matrices were *asymptotically free*, leading to model elements in operator algebras through random matrices. Since then, there has been a constant interplay between free probability theory and *random matrix theory* (RMT). We will try to give the main lines of these fruitful interactions. Among many nice references on free probability theory, we have chosen to follow the recent book of Mingo and Speicher (2017) and the lecture notes of Speicher (2019).

3.1. The notion of freeness

Let us start with the definition of freeness.

Definition 3.1 (Freeness). Consider a unital algebra $\mathscr{A}$ over $\mathbb{C}$, equipped with a linear functional $\tau \colon \mathscr{A} \to \mathbb{C}$ such that $\tau(1) = 1$. The pair $(\mathscr{A}, \tau)$ is called a *noncommutative probability space*. Unital subalgebras $(\mathscr{A}_i)_{i \in I}$ are called *free* (or *freely independent*) in $(\mathscr{A}, \tau)$ if, for any $a_1, \dots, a_k$ such that $\forall j \in [k]$, $\tau(a_j) = 0, a_j \in \mathscr{A}_{i(j)}$ and $i(1) \neq i(2) \neq \cdots \neq i(k)$,

$$\tau(a_1 \cdots a_k) = 0.$$

An important example, which is particularly relevant in our context, is the following:

Example 3.2. Let G be a group and $\mathbb{C}G$ its group algebra, that is

$$\mathbb{C}G = \Big\{ \sum_{g\in G} \alpha_g g, \alpha_g \in \mathbb{C}, \forall g \in G, \alpha_g \neq 0 \text{ for finitely many } g \Big\}.$$

Then $\mathbb{C}G$ is a unital algebra and one can define $\tau_G : \mathbb{C}G \to \mathbb{C}$,

$$\tau_G\Big(\sum_{g\in G} \alpha_g g \Big) = \alpha_e,$$

where e is the neutral element in G, so that $(\mathbb{C}G, \tau_G)$ is a non-commutative probability space. It is related with the notion of freeness for subgroups in the algebraic sense : a family $(G_i)_{i\in I}$ of subgroups of G is free if, for any $g_1, \dots, g_k$ such that $\forall j \in [k], g_j \in G_{i(j)}$ and $i(1) \neq i(2) \neq \cdots \neq i(k)$, $g_1 \dots g_k \neq e$ whenever $g_1 \neq e, \dots, g_k \neq e$. The link between free independence of subalgebras in the sense of Definition 3.1 and freeness for subgroups in the algebraic sense is made clear by the following proposition:

Proposition 3.3. *Let $(G_i)_{i\in I}$ be subgroups of a group G. Then the following statements are equivalent:*

- ▷ *The subgroups $(G_i)_{i\in I}$ are free in G.*
- ▷ *The subalgebras $(\mathbb{C}G_i)_{i\in I}$ are freely independent in the non-commutative probability space $(\mathbb{C}G, \tau_G)$.*

It is possible to enrich the structure of a non-commutative probability space as follows:

Definition 3.4. Let $(\mathscr{A}, \tau)$ be a non-commutative probability space.

- ▷ If τ is a trace, *i.e.* if $\tau(ab) = \tau(ba)$, for all $a, b \in \mathscr{A}$, then we call $(\mathscr{A}, \tau)$ a *tracial* non-commutative probability space.
- ▷ If $\mathscr{A}$ is a $*$-algebra (resp. a C^*-algebra) and τ is positive, *i.e.* if $\tau(a^*a) \geqslant 0$ for all $a \in \mathscr{A}$, then we call τ a *state* and $(\mathscr{A}, \tau)$ a $*$-probability space (resp. a C^*-probability space).
- ▷ A state τ is faithful if for all $a \in \mathscr{A}$, $\tau(a^*a) = 0$ implies $a = 0$.

If $(\mathscr{A}, \tau)$ a $*$-probability space and $(a_1, \dots, a_m)$ is a family of m elements in $\mathscr{A}$, then the *$*$-distribution* (or simply the distribution) of $(a_1, \dots, a_m)$ is given by the collection $\{\tau(P(a_1, a_1^*, \dots, a_m, a_m^*)), P \in \mathbb{C}\langle X_1, \dots, X_{2m}\rangle\}$, where $\mathbb{C}\langle X_1, \dots, X_{2m}\rangle$ is the set of non-commutative polynomials[(6)] in $2m$ variables with complex coefficients. By analogy with classical probability theory, elements of a $*$-probability space are called *random variables* and the random variables $(a_i)_{i\in I}$ are freely independent if the subalgebras generated respectively by $\{1, a_i, a_i^*\}$ are freely independent.

[(6)] in the sense that e.g. $X_1X_2 \neq X_2X_1$

3.2. The notion of asymptotic freeness

Roughly speaking, we say that a sequence of non-commutative random variables is asymptotically free if it converges in distribution to a family of freely independent random variables. Let us give a precise meaning of this sentence.

Definition 3.5. Let $((\mathscr{A}_n, \tau_n))_{n\geqslant 1}$ be a sequence of $*$-probability spaces and $(\mathscr{A}, \tau)$ a $*$-probability space. If, for any $n \geqslant 1$, $(a_{1,n}, \dots, a_{k,n})$ is a k-tuple of random variables in $(\mathscr{A}_n, \tau_n)$ and if there exist $a_1, \dots, a_k \in \mathscr{A}$ such that, for any non-commutative polynomial P in $2k$ variables, we have

$$\tau_n(P(a_{1,n}, a_{1,n}^*, \dots, a_{k,n}, a_{k,n}^*)) \xrightarrow[n\to\infty]{} \tau(P(a_1, a_1^*, \dots, a_k, a_k^*)),$$

we say that $(a_{1,n}, \dots, a_{k,n})_{n\geqslant 1}$ *converges in $*$-distribution* to $(a_1, \dots, a_k)$. If, in addition, $(a_1, \dots, a_k)$ is a family of random variables that are freely independent (with respect to τ), we say that $(a_{1,n}, \dots, a_{k,n})_{n\geqslant 1}$ are *asymptotically free* (for $n \to \infty$).

This notion appeared in Voiculescu's work (see e.g. Voiculescu, 1991), a few years after he introduced free independence. Since random matrices are typical examples of asymptotically free random variables, it built an important bridge between operator space theory and random matrix theory. This was first identified on the most emblematic ensemble of random matrices: the Gaussian Unitary Ensemble (Gue). Before stating a precise result, let us explain how one can define an ensemble of random matrices in the framework of $*$-probability spaces.

Lemma 3.6. *For every $n \geqslant 1$, if we set $(\mathscr{A}_n, \tau_n) := (M_n(L^{\infty-}(\Omega, \mathbb{P})), \mathrm{tr}_n \otimes \mathbb{E})$, where $(\Omega, \mathbb{P})$ is a classical probability space, $L^{\infty-}(\Omega, \mathbb{P}) := \bigcap_{1\leqslant p<\infty} L^p(\Omega, \mathbb{P})$ and for any complex algebra $\mathscr{A}$, $M_n(\mathscr{A}) \simeq M_n(\mathbb{C}) \otimes \mathscr{A}$ denotes the $n \times n$ matrices with entries from $\mathscr{A}$, where $\mathbb{E}$ denotes the expectation with respect to $\mathbb{P}$ and $\mathrm{tr}_n := \frac{1}{n}\mathrm{tr}$ the normalized trace on $M_n(\mathbb{C})$, then $(\mathscr{A}_n, \tau_n)$ is a $*$-probability space.*

It means that we will consider in the sequel $n \times n$ random matrices of the form $A = (a_{ij})_{i,j\in[n]}$, with, for all $i, j \in [n]$, $a_{ij} \in L^{\infty-}(\Omega, \mathbb{P})$, and a state given by

$$\tau_n(A) = \mathrm{tr}_n \otimes \mathbb{E}(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(a_{ii}).$$

An important example is provided by Gaussian random matrices:

Example 3.7. For $n \geqslant 1$, we say that A belongs to the Gaussian unitary ensemble of dimension n, and write $A \in \mathrm{Gue}(n)$, if $A = (a_{ij})_{i,j\in[n]}$ is a random $n \times n$ matrix such that $A = A^*$ and such that the entries form a centred complex Gaussian family with covariance

$$\mathrm{Cov}(a_{ij}, a_{k\ell}) = \frac{1}{n}\delta_{i,\ell}\delta_{j,k}.$$

If we consider a sequence of matrices $(G_n)_{n\geqslant 1}$ such that, for any $n \in \mathbb{N}^*$, $G_n \in \mathrm{GUE}(n)$, then one can show that, for any $k \in \mathbb{N}$,

$$\mathbb{E}\left(\mathrm{tr}_n(G_n^k)\right) \xrightarrow[n\to\infty]{} C_k := \begin{cases} 0 & \text{if } k \text{ is odd,} \\ \frac{1}{m+1}\binom{2m}{m} & \text{if } k = 2m, m \in \mathbb{N}. \end{cases}$$

There exists a $*$-probability space $(\mathscr{A}, \tau)$ on which one can define a self-adjoint element s such that, for any k in $\mathbb{N}$, $\tau(s^k) = C_k$. Then s is called a *semi-circular element* and $(G_n)_{n\geqslant 1}$ converges in $*$-distribution to s. Now, asymptotic freeness for independent matrices from the $\mathrm{GUE}(n)$ can be stated as follows :

Theorem 3.8 (Voiculescu, 1991). *Let $\ell \in \mathbb{N}^*$ be fixed and, for every $n \in \mathbb{N}^*$, let $(A_{1,n}, \dots, A_{\ell,n})$ be ℓ independent matrices from the $\mathrm{GUE}(n)$. Then the sequence $(A_{1,n}, \dots, A_{\ell,n})_{n\geqslant 1}$ converges in $*$-distribution to a family $(s_1, \dots, s_\ell)$ of freely independent semi-circular elements. The random matrices $(A_{1,n}, \dots, A_{\ell,n})_{n\geqslant 1}$ are therefore asymptotically free.*

An important property of the $\mathrm{GUE}(n)$ is its invariance by conjugation by a unitary matrix. This remark opened the way to asymptotic freeness for other ensembles of matrices. We denote by $\mathscr{U}(n) := \{U \in M_n(\mathbb{C}), UU^* = I_n\}$ the group of unitary matrices of size n. It is a compact group and it therefore carries a unique Borel probability measure which is invariant by translations, called the Haar measure on $\mathscr{U}(n)$. If U_n is distributed according to the Haar measure on $\mathscr{U}(n)$, one can check that, for any $k \in \mathbb{Z}^*$,

$$\mathbb{E}\left(\mathrm{tr}_n(U_n^k)\right) \xrightarrow[n\to\infty]{} \delta_{k,0}.$$

By analogy, if u is a unitary element in a $*$-probability space $(\mathscr{A}, \tau)$ which satisfies $\tau(u^k) = \delta_{k,0}$, for any k in $\mathbb{Z}$, then u is called *Haar unitary.* A first asymptotic freeness property in this context is the following:

Proposition 3.9. *Let $\ell \in \mathbb{N}^*$ be fixed and, for every $n \in \mathbb{N}^*$, let $(U_{1,n}, \dots, U_{\ell,n})$ be ℓ independent random matrices distributed according to the Haar measure on $\mathscr{U}(n)$. Then the sequence $(U_{1,n}, \dots, U_{\ell,n})_{n\geqslant 1}$ converges in $*$-distribution to a family $(u_1, \dots, u_\ell)$ of freely independent Haar unitaries. The random matrices $(U_{1,n}, \dots, U_{\ell,n})_{n\geqslant 1}$ are therefore asymptotically free.*

But Haar distributed random matrices can also convey asymptotic freeness to deterministic matrices.

Theorem 3.10. *For every $n \in \mathbb{N}$, let U_n be a Haar unitary random $n \times n$-matrix, let $A_n, B_n \in M_n(\mathbb{C})$, and suppose that $(A_n)_{n\geqslant 1}$ converges in $*$-distribution to a and $(B_n)_{n\geqslant 1}$ converges in $*$-distribution to b for random variables a and b in some $*$-probability*

space, and that the sequences $(A_n)_{n\geqslant 1}$ and $(B_n)_{n\geqslant 1}$ are independent[7] *of $(U_n)_{n\geqslant 1}$. Then, $(U_nA_nU_n^*, B_n)_{n\geqslant 1}$ converges in $*$-distribution to (a, b), where a and b are freely independent. In particular, the random matrices $(U_nA_nU_n^*)_{n\geqslant 1}$ and $(B_n)_{n\geqslant 1}$ are asymptotically free from each other.*

If we now go to the group $\mathfrak{S}_n$ of permutations of $[n]$, then the uniform law on $\mathfrak{S}_n$ is the Haar measure and it is natural to ask if a result similar to Theorem 3.10 holds true in this context. A positive answer has been brought by Nica (1993):

Theorem 3.11. *Let $\ell \in \mathbb{N}^*$ be fixed and, for every $n \in \mathbb{N}^*$, let $(S_{1,n}, \dots, S_{\ell,n})$ be ℓ independent random matrices distributed uniformly over the set of permutation matrices of size n. Then the sequence $(S_{1,n}, \dots, S_{\ell,n})_{n\geqslant 1}$ converges in $*$-distribution to a family $(u_1, \dots, u_\ell)$ of freely independent Haar unitaries. The random matrices $(S_{1,n}, \dots, S_{\ell,n})_{n\geqslant 1}$ are therefore asymptotically free.*

We will in fact use an extension of this result. Beyond the uniform distribution of the set of permutations of $[n]$, we will also be interested in the uniform distribution of the set of matchings of n points, that is the subset of involutions without fixed point within the permutations of $[n]$. Obviously, if $(R_n)_{n\geqslant 1}$ is a sequence of random variables with this distribution, it converges in $*$-distribution to a free random variable r such that $\tau(r^k) = 1$ if k is even and $\tau(r^k) = 0$ if k is odd. We say that r is a *free Rademacher* variable. We have the following asymptotic freeness result:

Proposition 3.12. *With the notations of Theorem 3.11 and if $(R_{1,n}, \dots, R_{m,n})$ are m independent random matrices distributed uniformly over the set of pair-matchings of n points, independent of $(S_{1,n}, \dots, S_{\ell,n})$, the sequence $(S_{1,n}, \dots, S_{\ell,n}, R_{1,n}, \dots, R_{m,n})_{n\geqslant 1}$ converges in $*$-distribution to $(u_1, \dots, u_\ell, r_1, \dots, r_m)$, where $(r_1, \dots, r_m)$ are freely independent free Rademacher variables, freely independent from $(u_1, \dots, u_\ell)$.*

3.3. Limiting operator for the symmetric random permutation model

As explained above, we want to study the convergence of the sequence of operators $(A_n)_{n\geqslant 1}$ defined in (1), under Assumption 2.5. The goal of this section is to construct the limiting object it converges to. From Proposition 3.12, we know it will involve a family of freely independent Haar unitaries and free Rademacher variables. To describe them more concretely, we go back to our very first example 3.2 of noncommutative probability space but we will enrich the structure of the group algebra.

Let G be a discrete group. We endow the group algebra $\mathbb{C}G$ with the inner product $\langle \cdot, \cdot \rangle$ such that $\langle g, h \rangle = \delta_{g,h}$ for all $g, h \in G$. Then

$$\ell^2(G) := \Big\{ \sum_{g\in G} \alpha_g g, \sum_{g\in G} |\alpha_g|^2 < \infty \Big\}$$

[7] They can be deterministic.

is a Hilbert space. If $\mathscr{B}(\ell^2(G))$ is the space of bounded operators on $\ell^2(G)$, then we can define $\lambda\colon G \to \mathscr{B}(\ell^2(G))$ as follows: if $\sum_{g\in G} |\alpha_g|^2 < \infty$, then $\forall g \in G$, we define

$$\lambda(g)\cdot\alpha = \lambda(g)\cdot\sum_{h\in G}\alpha_h h := \sum_{h\in G}\alpha_h\, g^{-1}h.$$

The map λ is called the *left regular representation* and one can check that it is unitary. We then extend λ to $\mathbb{C}G$ by linearity. As λ is injective on G, one can extend the linear form τ_G to $\lambda(\mathbb{C}G)$ by

$$\tau_G(\lambda(g)) = \tau_G(g) = \delta_{g,e}.$$

We have $\lambda(\mathbb{C}G) \subset \mathscr{B}(\ell^2(G))$ and the closure of $\lambda(\mathbb{C}G)$ with respect to the operator norm topology is a C^*-algebra called the *reduced group C^*-algebra* of G, denoted by $C_{\mathrm{red}}(G)$.

In the context of the work of Bordenave and Collins (2019), the discrete group G (hereafter denoted by X_*) we start with takes the form of a free product:

$$X_* := \mathbb{Z}^{*q} * (\mathbb{Z}/2\mathbb{Z})^{*(d-2q)},$$

where $\mathbb{Z}^{*q}$ denotes the free product of q copies of $\mathbb{Z}$. We will simply denote τ_{X_*} by τ when there is no ambiguity.

More concretely, in our case, if $g_1,\dots,g_d$ are generators of X_* such that, for $i\in[q]$, (g_i, g_{i+q}) generates the ith copy of $\mathbb{Z}$, then $((\lambda(g_i),\lambda(g_{i+q}))_{i\in[q]}, (\lambda(g_i))_{2q+1\leqslant i\leqslant d})$ is a family of freely independent random variables. Moreover, for $i \in [2q]$, for all $k\in\mathbb{Z}$, $\tau(\lambda(g_i)^k) = \tau(g_i^k) = \delta_{k,0}$, so that these random variables are Haar unitaries, whereas when $2q+1 \leqslant i \leqslant d$, $\tau(g_i^k) = 1$ if k is even and 0 if k is odd so that these random variables are free Rademacher variables. Otherwise stated, $(\lambda(g_i))_{i\in[d]}$ is a concrete realization of the limit in $*$-distribution of the operators $(S_{1,n},\dots,S_{d,n})_{n\geqslant 1}$ satisfying Assumption (Hσ) and it is a straightforward consequence of Proposition 3.12 that the sequence of operators $(A_n)_{n\geqslant 1}$ we are interested in should converge to

$$A_* := a_0\otimes\mathbf{1} + \sum_{i=1}^{d} a_i\otimes\lambda(g_i), \tag{2}$$

acting on $\mathbb{C}^r\otimes\ell^2(X_*)$. This is an element of the unital $*$-algebra $\mathscr{A} := M_r(C_{\mathrm{red}}(X_*))$, equipped with the trace $\mathrm{tr}_r\otimes\tau$. This implies also the convergence of the sequence of operators $(A_n)_{n\geqslant 1}$ to A_* in the topology of local convergence of Benjamini and Schramm but we don't want to insist here on this aspect. Bordenave and Collins (2019) rather focus on the convergence of the spectrum.

We denote by $\sigma(T)$ the spectrum of an operator T, that is

$$\sigma(T) := \{\lambda\in\mathbb{C}, T-\lambda I \text{ is not invertible}\}.$$

The convergence in $*$-distribution of $(A_n)_{n\geqslant 1}$ to A_* implies that, if $[a,b]$ is any given interval, the expected proportion of eigenvalues of A_n in $[a,b]$ converges to the spectral measure of A_* on this interval. Therefore, $\forall \varepsilon > 0$, for n large enough, with high probability,

$$\sigma(A_*) \subset \sigma(A_n) + [-\varepsilon, \varepsilon]. \tag{3}$$

The main object of the paper of Bordenave and Collins (2019) is to establish the reverse inclusion. This is related to the notion of strong convergence, that we introduce in the next section.

3.4. Strong convergence, strong asymptotic freeness, linearization trick

The fact that $(A_n)_{n\geqslant 1}$ converges in $*$-distribution to A_* does not rule out the possibility of $o(n)$ eigenvalues staying away from the limiting spectrum. These eigenvalues are usually called *outliers*. In our context, there are obvious outliers, coming from the Perron–Frobenius eigenvalues of the operators $S_{i,n}$. For example, one can check that $S_{1,n} + S_{1,n}^* + \cdots + S_{k,n} + S_{k,n}^*$ will always have an eigenvalue equal to $2k$ associated to the constant vector $\mathbf{1}$. This will produce an outlier as long as $k \geqslant 2$. To get rid of these trivial eigenvalues, we will only consider the operators $S_{i,n}$ restricted to the orthogonal $\mathbf{1}^\perp$ of the constant vector $\mathbf{1}$, or equivalently the operator A_n restricted to $H_0 := (\mathbb{C}^r \otimes \mathbf{1})^\perp$. The main theorem of Bordenave and Collins (2019) is the following:

Theorem 3.13. *Consider a sequence of random operators $(A_n)_{n\geqslant 1}$ distributed according to the symmetric random permutation model* (1)*. Then the Hausdorff distance between the spectrum $\sigma(A_*)$ of the operator A_* defined in* (2) *and the spectrum $\sigma((A_n)_{|H_0})$ of the operator A_n restricted to H_0 converges to zero in probability as n goes to infinity. Otherwise stated, $\forall \varepsilon > 0$,*

$$\mathbb{P}\left(\sigma((A_n)_{|H_0}) \subset \sigma(A_*) + [-\varepsilon, \varepsilon]\right) \xrightarrow[n\to\infty]{} 1, \tag{4}$$

$$\mathbb{P}\left(\sigma(A_*) \subset \sigma((A_n)_{|H_0}) + [-\varepsilon, \varepsilon]\right) \xrightarrow[n\to\infty]{} 1. \tag{5}$$

The second point comes directly from (3) and the fact that the spectrum of A_* does not contain the Perron–Frobenius eigenvalue (as the constant vector $\mathbf{1}$ is not in $\ell^2(X_*)$). We will develop in detail the proof of (4). Before that, we want to relate this inclusion with the notion of *strong asymptotic freeness*. With Definition 3.5 in mind, we define *strong convergence* as follows:

Definition 3.14. Let $((\mathscr{A}_n, \tau_n))_{n\geqslant 1}$ be a sequence of C*-probability spaces and $(\mathscr{A}, \tau)$ a C*-probability space. If, for any $n \geqslant 1$, $(a_{1,n}, \ldots, a_{k,n})$ is a k-tuple of random variables in $(\mathscr{A}_n, \tau_n)$ and if there exist $a_1, \ldots, a_k \in \mathscr{A}$ such that, for any non-commutative

polynomial P in $2k$ variables, we have

$$\tau_n(P(a_{1,n}, a_{1,n}^*, \dots, a_{k,n}, a_{k,n}^*)) \xrightarrow[n\to\infty]{} \tau(P(a_1, a_1^*, \dots, a_k, a_k^*)),$$

and in operator norm we have

$$\|P(a_{1,n}, a_{1,n}^*, \dots, a_{k,n}, a_{k,n}^*)\| \xrightarrow[n\to\infty]{} \|P(a_1, a_1^*, \dots, a_k, a_k^*)\|,$$

we say that $(a_{1,n}, \dots, a_{k,n})_{n\geqslant 1}$ *converges strongly* to $(a_1, \dots, a_k)$. If $(a_1, \dots, a_k)$ is a family of random variables that are freely independent (with respect to τ), we say that $(a_{1,n}, \dots, a_{k,n})_{n\geqslant 1}$ are *strongly asymptotically free*.

The following proposition, that can be found e.g. in (Collins and Male, 2014), clarifies the link between strong convergence and control of outliers. If h is a self-adjoint element in a C^*-probability space $(\mathscr{A}, \tau)$, its spectral measure μ_h is the unique probability measure on $\mathbb{R}$ such that, for any $k \geqslant 1$, $\tau(h^k) = \int t^k \mathrm{d}\mu_h(t)$; we denote its support by $\mathrm{supp}(\mu_h)$.

Proposition 3.15. *Let $((\mathscr{A}_n, \tau_n))_{n\geqslant 1}$ be a sequence of $*$-probability spaces and $(\mathscr{A}, \tau)$ a $*$-probability space. For any $n \geqslant 1$, let $(a_{1,n}, \dots, a_{k,n})$ be a k-tuple of random variables in $(\mathscr{A}_n, \tau_n)$ and $a_1, \dots, a_k \in (\mathscr{A}, \tau)$. Then the two following statements are equivalent :*

- ▷ *$(a_{1,n}, \dots, a_{k,n})_{n\geqslant 1}$ converges strongly to $(a_1, \dots, a_k)$,*
- ▷ *for any polynomial P such that $h_n := P(a_{1,n}, \dots, a_{k,n})$ is self-adjoint, $(P(a_{1,n}, \dots, a_{k,n}))_{n\geqslant 1}$ converges in $*$-distribution to $h := P(a_1, \dots, a_k)$ and $\forall \varepsilon > 0$, for n large enough,*

$$supp(\mu_{h_n}) \subset supp(\mu_h) + [-\varepsilon, \varepsilon],$$

where μ_{h_n} and μ_h are respectively the spectral measures of h_n and h.

The strong convergence result obtained by Bordenave and Collins (2019) can be stated as follows:

Theorem 3.16. *Let $d \geqslant 1$ be fixed. For every $n \geqslant 1$, consider a d-tuple $(S_{1,n}, \dots, S_{d,n})$ of random permutations of $[n]$ satisfying $(\mathrm{H}\sigma)$. Then, for any non-commutative polynomial P, for any $\varepsilon > 0$,*

$$\mathbb{P}\left(\left|\|P((S_{1,n})_{|H_0}, \dots, (S_{d,n})_{|H_0})\| - \|P(\lambda(g_1), \dots, \lambda(g_d))\|\right| > \varepsilon\right) \xrightarrow[n\to\infty]{} 0,$$

where $\|\cdot\|$ stands for the operator norm (on the space of matrices for the first term and in the C^*-algebra $C_{\mathrm{red}}(X_*)$ for the second term).

Since $(\lambda(g_1), \dots, \lambda(g_d))$ are freely independent[8]*, we say that $((S_{1,n})_{|H_0}, \dots, (S_{d,n})_{|H_0})$ are strongly asymptotically free.*

[8] This is a slight abuse of notation: it would be more correct to say that $((\lambda(g_i), \lambda(g_{i+q}))_{i\in[q]}, (\lambda(g_i))_{2q+1\leqslant i\leqslant d})$ is a family of freely independent random variables and that the corresponding family for the $S_{i,n}$'s are strongly asymptotically free.

This result may look stronger than Theorem 3.13 as it implies convergence of the norm of any polynomial P in the operators $S_{i,n}$ whereas Theorem 3.13 only deals with polynomials of degree one (with matrix coefficients). In fact, thanks to the *linearization trick*, Theorem 3.16 can be seen as a corollary of Theorem 3.13. In the context of operator algebra, the argument is due to Pisier (1996)[(9)], although the idea of linearizing polynomial problems by going to matrices is much older and known under different names in different mathematical communities. More precisely, for P a non-commutative polynomial in d variables and $r \in \mathbb{N}^*$, a matrix $\hat{P} := \begin{pmatrix} 0 & U \\ V & Q \end{pmatrix} \in M_r(\mathbb{C}) \otimes \mathbb{C}\langle X_1, \dots, X_d \rangle$, where $Q \in M_{r-1}(\mathbb{C}) \otimes \mathbb{C}\langle X_1, \dots, X_d \rangle$ and U and V are respectively row and column vectors of size $r-1$ with entries in $\mathbb{C}\langle X_1, \dots, X_d \rangle$, is a *linearization* of P if

$$\hat{P} = b_0 \otimes \mathbf{1} + b_1 \otimes X_1 + \cdots + b_d \otimes X_d,$$

and $P = -UQ^{-1}V$.

For example, a possible linearization of a monomial $P = X_{i_1} \dots X_{i_r}$ is

$$\hat{P} = \begin{pmatrix} & & & X_{i_1} \\ & & X_{i_2} & -1 \\ & \iddots & \iddots & \\ X_{i_r} & -1 & & \end{pmatrix}.$$

One can show from there that any polynomial P admits a linearization and if P is self-adjoint, the linearization can be chosen self-adjoint.

From there, one can get a criterion which is crucial for proving strong convergence:

Proposition 3.17 (Linearization trick). *If $\boldsymbol{u} := (u_1, \dots, u_d)$ is a d-tuple of unitary elements generating a unital C^*-algebra $\mathscr{A}$ and $\boldsymbol{v} := (v_1, \dots, v_d)$ is a d-tuple of unitary elements generating a unital C^*-algebra $\mathscr{B}$, then the following are equivalent:*

- ▷ *for any non-commutative polynomial P in d variables and their adjoints, $\|P(\boldsymbol{u})\| = \|P(\boldsymbol{v})\|$,*
- ▷ *for any integer r, and $r \times r$ matrices $a_0, \dots, a_d$ such that $a_0 \otimes 1 + a_1 \otimes u_1 + \cdots + a_d \otimes u_d$ and $a_0 \otimes 1 + a_1 \otimes v_1 + \cdots + a_d \otimes v_d$ are self-adjoint, we have*

$$\|a_0 \otimes 1 + a_1 \otimes u_1 + \cdots + a_d \otimes u_d\| = \|a_0 \otimes 1 + a_1 \otimes v_1 + \cdots + a_d \otimes v_d\|.$$

This explains how Theorem 3.16 is a corollary of Theorem 3.13.

In the next section, we will develop the main lines of the proof of Theorem 3.13 and introduce in particular the notion of non-backtracking operators. We will go back to free probability, in particular strong asymptotic freeness, in the last section of these notes, where we will describe in detail some applications of this property.

[(9)] We also refer the reader to the appendix in (Bordenave and Collins, 2020).

4. The use of non-backtracking operators

4.1. Ihara–Bass formula and applications

For a matrix $H := (H_{ij})_{i,j\in[n]} \in M_n(\mathbb{C})$ with complex entries, the *non-backtracking matrix* associated to H is $B := (B_{ej})_{e,f\in[n]^2}$ defined, for $e = (i,j)$ and $f = (k,\ell)$, by

$$B_{ef} = H_{k\ell}\delta_{j,k}(1-\delta_{i,\ell}).$$

The name *non-backtracking* comes from the following interpretation: if H is the adjacency matrix of a graph G, then, for any $m \in \mathbb{N}^*$, $(H^m)_{ij}$ is the number of paths from i to j of length m in G, whereas, if $e = (i,j)$ and $f = (k,\ell)$, $(B^m)_{ef}$ is the number of non-backtracking paths of length $m+1$ starting with the edge e and ending with the edge f. A non-backtracking path may have cycles but cannot immediately go back to the vertex it comes from. Although the non-backtracking operator is non-normal even when H is self-adjoint, it is a powerful tool for the study of spectral properties of random graphs. The Ihara–Bass formula allows to link the spectrum of H with the spectrum of its non-backtracking counterpart. This statement and its proof can be found e.g. in (Benaych-Georges, Bordenave, and Knowles, 2020).

Proposition 4.1 (Ihara–Bass formula). *Let $H \in M_n(\mathbb{C})$ and B the associated non-backtracking matrix. Let $\lambda \in \mathbb{C}$ be such that $\lambda^2 \notin \{H_{ij}H_{ji}; i,j \in [n]\}$, and define, for $i,j \in [n]$,*

$$H_{ij}(\lambda) := \frac{\lambda H_{ij}}{\lambda^2 - H_{ij}H_{ji}} \text{ and } m_i(\lambda) := 1 + \sum_{k=1}^{N} \frac{H_{ik}H_{ki}}{\lambda^2 - H_{ik}H_{ki}}.$$

Then $\lambda \in \sigma(B)$ if and only if $\det(M(\lambda) - H(\lambda)) = 0$, where $M(\lambda)$ is the diagonal matrix with non-zero entries equal to $m_1(\lambda), \dots, m_n(\lambda)$.

In the model of Bordenave and Collins (2019), in the general case, it is not obvious to interpret A_n as an adjacency matrix. This can be done by extending the notion of graph and considering a weighted graph on the vertex set $[n]$, where each vertex carries a loop edge with weight a_0 and $(x,i) \in [n] \times [d]$ can be thought as an directed edge between x and $\sigma_i(x)$ with weight a_i. The associated non-backtracking operator can be written

$$B_n = \sum_{j\neq i^*} a_j \otimes S_{i,n} \otimes E_{ij},$$

with $E_{ij} \in M_d(\mathbb{C})$. This means that, if $e = (x,i)$ and $f = (y,j)$,

$$(B_n)_{ef} = a_j\delta_{y,\sigma_i(x)}(1-\delta_{j,i^*}).$$

Therefore, if $\gamma = (\gamma_1, \dots, \gamma_k)$ is a path with $\gamma_t = (x_t, i_t)$, we set $a(\gamma) := \prod_{t=1}^k a_{i_t}$. The path γ is non-backtracking if $\forall t \in [k-1], i_{t+1} \neq i_t^*$. For $e,f \in [n] \times [d]$, let Γ_{ef}^k be

the set of non-backtracking paths $(\gamma_1, \dots, \gamma_k)$ such that $\gamma_1 = e$ and $\gamma_k = f$, so that

$$(B^k)_{ef} = \sum_{\gamma \in \Gamma_{ef}^{k+1}} a(\gamma) \prod_{t=1}^{k} (S_{i,n})_{x_t x_{t+1}}. \tag{6}$$

The Ihara–Bass formula reads:

Proposition 4.2 (Ihara–Bass formula - symmetric random permutation model)**.** *Let $\lambda \notin \bigcup_{i\in[d]} \sigma(a_i a_{i^*})$, and define $A_{n,\lambda} := a_0(\lambda) \otimes \mathbf{1}^{(n)} + \sum_{i=1}^{d} a_i(\lambda) \otimes S_{i,n}$, with*

$$a_i(\lambda) := \lambda a_i (\lambda^2 - a_{i^*} a_i)^{-1},$$

and

$$a_0(\lambda) := -1 - \sum_{i=1}^{d} a_i (\lambda^2 - a_{i^*} a_i)^{-1} a_{i^*}.$$

Then $\lambda \in \sigma(B_n)$ if and only if $0 \in \sigma(A_{n,\lambda})$.

To exploit this relation for the study of the spectrum of A_n, we perform some reverse engineering: for a given μ, we want to construct an operator $A_{n,\mu}$ of the same form as A_n and its non-backtracking operator $B_{n,\mu}$ such that

$$\mu \in \sigma(A_n) \text{ if and only if } 1 \in \sigma(B_{n,\mu}). \tag{7}$$

This is in fact possible if $\mu \notin \mathrm{full}(\sigma(A_*))$. For a bounded subset D of $\mathbb{C}$, we define $\mathrm{full}(D) := \mathbb{C} \setminus U$, where U is the unique unbounded connected component of $\mathbb{C} \setminus D$[(10)]. For such a μ, we now explain the recipe to construct $A_{n,\mu}$ and therefore $B_{n,\mu}$ satisfying (7). Let $G(\mu) := (\mu - A_*)^{-1}$ be the resolvant operator of A_*. This is also an operator on $\mathbb{C}^r \otimes \ell^2(X_*)$. It can be seen as an infinite block matrix where each block is an $r \times r$ matrix indexed by a pair of indices $x, y \in X_* \times X_*$. We let

$$\hat{a}_i(\mu) := G_{ee}(\mu)^{-1} G_{eg_i}(\mu),$$

with e the neutral element in X_* and $g_1, \dots, g_d$ the generators, and define

$$A_{n,\mu} := \sum_{i=1}^{d} \hat{a}_i(\mu) \otimes S_{i,n} \text{ and } B_{n,\mu} := \sum_{j \neq i^*} \hat{a}_j(\mu) \otimes S_{i,n} \otimes E_{ij}.$$

Then (7) holds.

We also denote by

$$A_{*,\mu} := \sum_{i=1}^{d} \hat{a}_i(\mu) \otimes \lambda(g_i) \text{ and } B_{*,\mu} := \sum_{j \neq i^*} \hat{a}_j(\mu) \otimes \lambda(g_i) \otimes E_{ij}.$$

(10) $\mathrm{full}(D)$ "fills the holes" of D.

For any operator T, we denote by $\rho(T) := \sup\{|\lambda|, \lambda \in \sigma(T)\}$ its spectral radius. One can show that, for $\mu \notin \text{full}(\sigma(A_*))$, one has $\rho(B_{*,\mu}) < 1$. This leads to the following criterion for comparing the spectrum of A_n and A_*:

Theorem 4.3. *For any $\varepsilon > 0$, there exists $\delta > 0$ such that*

$$\text{if for all } \mu \in \mathbb{C}, \rho(B_{n,\mu}) < \rho(B_{*,\mu}) + \delta,$$

then $\text{full}(\sigma(A_n))$ *is in an ε-neighborhood of a slight modification*[11] *of $\sigma(A_*)$. Moreover, if we define $K_0 := (\mathbb{C}^r \otimes \mathbf{1} \otimes \mathbb{C}^d)^\perp = \{f : [n] \times [d] \to \mathbb{C} : \forall i \in [d], \sum_{x\in[n]} f(x,i) = 0,\}$, then the same holds true if we replace $B_{n,\mu}$ by $(B_{n,\mu})_{|K_0}$ and A_n by $(A_n)_{|H_0}$.*

Therefore, to get (4), it is enough to show the following:

Theorem 4.4. *For any $\varepsilon > 0$, under Assumption 2.5,*

$$\mathbb{P}\left(\forall a_i \text{ such that } \max(\|a_i\| \vee \|a_i^{-1}\|^{-1}) \leqslant 1/\varepsilon, \rho((B_n)_{|K_0}) \leqslant \rho(B_*) + \varepsilon\right) \xrightarrow[n\to\infty]{} 1.$$

4.2. A glimpse of the Füredi–Komlós moment method

The core of the proof of Theorem 4.4 consists in using a moment method. In RMT, moment methods were first used to study empirical spectral measures. This goes back at least to the pioneering work of Wigner. For a random matrix M_n, with eigenvalues $\lambda_{1,n}, \dots, \lambda_{n,n}$, we denote by $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^n \delta_{\lambda_{i,n}}$ the empirical measure of its eigenvalues. The main idea is to rewrite the moments of this spectral empirical measure $m_n(k) := \int x^k \mathrm{d}\hat{\mu}_n$ in terms of traces of powers of the matrix M_n, namely $m_n(k) = \text{tr}_n(M_n^k)$, where tr_n is the normalized trace on $M_n(\mathbb{C})$. We then expand

$$\mathbb{E}\left[\text{tr}_n(M_n^k)\right] = \frac{1}{n}\sum_{i_1,\dots,i_k\in[n]} \mathbb{E}\left(M_{i_1 i_2}\cdots M_{i_k i_1}\right),$$

look at the sequence $(i_1, \dots, i_k)$ as a path $\gamma_1 = (i_1, i_2), \dots, \gamma_k = (i_k, i_1)$ and identify the contributions to the sum according to the geometric or combinatorial properties of the paths. For example, in the case when $M_n \in \text{GUE}(n)$, the paths contributing to the sum are the contours of rooted trees, leading to limiting moments given by Catalan numbers.

Later, the method has been adapted by Füredi and Komlós (1981) to study the spectral radius or the spectral norm of the matrix M_n. The idea is the following: to capture the behavior of the largest eigenvalue, one has to compute moments of order k_n growing with the size n of the matrix. Indeed, for $\rho > 0$, if we want to show that, for any $\varepsilon > 0$,

$$\mathbb{P}(\|M_n\| \leqslant \rho(1+\varepsilon)) \xrightarrow[n\to\infty]{} 1,$$

[11] We do not detail it here. For more details, see the definition of $\text{full}(\hat{\sigma}(A_*))$ in Bordenave and Collins (2019).

we write

$$\|M_n\|^{2k_n} = \|M_n M_n^*\|^{k_n} \leqslant n\mathrm{tr}_n((M_n M_n^*)^{k_n}),$$

because the largest of n nonnegative real numbers is not larger than their sum, so that we might lose a factor n in the last bound. Now, if we choose $k_n \gg \log n$ and show that

$$\mathbb{E}\left[\mathrm{tr}_n((M_n M_n^*)^{k_n})\right] \leqslant \rho^{2k_n}(1+\varepsilon)^{2k_n},$$

then, for any $\delta > 0$,

$$\begin{aligned}\mathbb{P}(\|M_n\| \leqslant \rho(1+\varepsilon)(1+\delta)) &\geqslant 1 - \frac{n\mathbb{E}\left[\mathrm{tr}_n((M_n M_n^*)^{k_n})\right]}{(1+\delta)^{2k_n}} \\ &\geqslant 1 - \exp\left(-2k_n\log(1+\delta) + \log n\right) \xrightarrow[n\to\infty]{} 1.\end{aligned}$$

This method has been very successful and has led in particular to the first universality results for the Tracy–Widom distribution by Soshnikov (1999). However, in the symmetric permutation model, it is hopeless to apply directly this method to the operators $(A_n)_{|H_0}$ but one can circumvent this obstacle by applying it to the non-backtracking operators.

4.3. High order moments of the non-backtracking operators

The moment method for the non-backtracking operator B_n is very involved, it is the main technical part and we do not intend here to give a precise account of the proof. We rather want to explain what are the main steps and the main difficulties. In this paragraph, we will omit the n subscript to lighten the notations a bit.

Following the method we have just explained above, to prove Theorem 4.4, we need to bound the spectral radius by the spectral norm of powers of the operator. More precisely, recalling that $K_0 = (\mathbb{C}^r \otimes \mathbf{1} \otimes \mathbb{C}^d)^\perp$, we will use that

$$\rho(B_{|K_0}) = \rho(B_{|K_0}^\ell)^{1/\ell} \leqslant \sup_{g\in K_0, \|g\|_2=1} \|B^\ell g\|_2^{1/\ell}.$$

Restricting B to the subspace K_0 boils down to performing an orthogonal projection of the operators S_i onto $\mathbf{1}^\perp$. More explicitely, if for every $i \in [d]$, $\underline{S}_i := S_i - \frac{1}{n}\mathbf{1}\otimes\mathbf{1}$ is such a projection, we define

$$\underline{B} := \sum_{j\neq i^*} a_j \otimes \underline{S}_i \otimes E_{ij}$$

and one can check that, for any $g \in K_0$, $B^\ell g = \underline{B}^\ell g$.

But the matrix $\underline{B}$ won't be used directly. Before projecting on K_0, we will replace B^ℓ by a matrix $B^{(\ell)}$ that coincides with B^ℓ with high probability but has better properties.

This step is known as *removing the tangles*, and will be described right after. We will then project on K_0, that is consider $\underline{B^{(\ell)}}$. However, as K_0 is not necessarily invariant under $B^{(\ell)}$, there will be some remainder term. Let us describe these steps more precisely. We first recall from (6) that, for any $\ell \in \mathbb{N}^*$,

$$(B^\ell)_{ef} = \sum_{\gamma \in \Gamma_{ef}^{\ell+1}} a(\gamma) \prod_{t=1}^{\ell} (S_i)_{x_t x_{t+1}},$$

with $\Gamma_{ef}^{\ell+1}$ the set of non-backtracking paths such that $\gamma_1 = e$ and $\gamma_{\ell+1} = f$. For a path $\gamma \in \Gamma_{ef}^{\ell+1}$, let G_γ be the graph with vertices $V_\gamma := \{x_t, t \in [\ell+1]\}$ and edges $E_\gamma := \{[x_t, i_t, x_{t+1}], t \in [\ell]\}$, where i_t can be seen as the *color* of the edge $[x_t, i_t, x_{t+1}]$. We can now define the notion of *tangle*.

Definition 4.5 (Tangles). A graph H is *tangle-free* if it contains at most one cycle. For any $\ell \in \mathbb{N}^*$, a graph H is *ℓ-tangle-free* if, for every vertex x, $(H, x)_\ell$ contains at most one cycle, where $(H, x)_\ell$ is the subgraph of H restricted to the vertices at distance at most ℓ from x for the graph distance. We say that a path γ is tangle-free if G_γ is.

We denote by F^k (respectively F_{ef}^k) the subset of tangle-free paths in Γ^k (resp. Γ_{ef}^k). For ℓ fixed, we denote by

$$(B^{(\ell)})_{ef} := \sum_{\gamma \in F_{ef}^{\ell+1}} a(\gamma) \prod_{t=1}^{\ell} (S_i)_{x_t x_{t+1}}.$$

If the permutations $\sigma_1, \dots, \sigma_d$ satisfy Assumption (Hσ), we denote by G^σ the graph whose vertex set is $[n]$ and whose edges are $[x, i, y]$ such that $\sigma_i(x) = y$ and $\sigma_{i^*}(y) = x$. For any $\ell \in \mathbb{N}^*$, if G^σ is ℓ-tangle-free, then for all $0 \leqslant k \leqslant 2\ell$, $B^k = B^{(k)}$. As above, if we denote by

$$(\underline{B^{(\ell)}})_{ef} := \sum_{\gamma \in F_{ef}^{\ell+1}} a(\gamma) \prod_{t=1}^{\ell} (\underline{S_i})_{x_t x_{t+1}},$$

then even if G^σ is ℓ-tangle-free, we have a priori that $\underline{B}^k \neq \underline{B^{(k)}}$. Let us write down more explicitly the difference between the two quantities. If we set

$$\overline{B} := \sum_{j \neq i^*} a_j \otimes (\mathbf{1} \otimes \mathbf{1}) \otimes E_{ij},$$

one can check that

$$B^{(\ell)} = \underline{B^{(\ell)}} + \frac{1}{n} \sum_{k=1}^{\ell} \underline{B^{(k-1)}} \overline{B} B^{(\ell-k)} - \frac{1}{n} \sum_{k=1}^{\ell} R_k^{(\ell)}, \tag{8}$$

where $R_k^{(\ell)}$ can be thought of as an error term equal to

$$(R_k^{(\ell)})_{ef} = \sum_{\gamma \in F_{k,ef}^{\ell+1} \setminus F_{ef}^{\ell+1}} a(\gamma) \Big(\prod_{t=1}^{k-1} (\underline{S}_i)_{x_t x_{t+1}} \Big) \Big(\prod_{t=k+1}^{\ell} (S_i)_{x_t x_{t+1}} \Big),$$

where $F_k^{\ell+1}$ is the set of paths that can be decomposed in $\gamma' \in F^k, \gamma'' \in F^2$ and $\gamma''' \in F^{\ell-k+1}$ and $F_{k,ef}^{\ell+1} = F_k^{\ell+1} \cap \Gamma_{ef}^{\ell+1}$. Note that the concatenation of three tangle-free paths is not necessarily tangle-free.

Now, if G^σ is ℓ-tangle-free, then one can check that the second term in (8) cancels on K_0, thus, for any $g \in K_0$,

$$B^\ell g = B^{(\ell)} g = \underline{B^{(\ell)}} g - \frac{1}{n} \sum_{k=1}^{\ell} R_k^{(\ell)} g,$$

so that if G^σ is ℓ-tangle-free,

$$\rho(B_{|K_0}) \leqslant \Big(\|\underline{B^{(\ell)}}\| + \frac{1}{n} \sum_{k=1}^{\ell} \|R_k^{(\ell)}\| \Big)^{1/\ell}.$$

Remember that, in the Füredi–Komlós method, we need to choose $\ell = \ell_n$ growing with n. The next natural question to ask is: for which values of ℓ_n is the probability that G^σ is ℓ_n-tangle-free large enough ? This is a nice question on random permutations that boils down to estimating the expected number of cycles of length ℓ in G^σ. Through such an estimate, Bordenave and Collins (2019) got the following:

Lemma 4.6. *For random permutations satisfying* (Hσ), *there exists* $c > 0$ *such that for all* $1 \leqslant \ell \leqslant \sqrt{n}$,

$$\mathbb{P}\left(G^\sigma \text{ is } \ell\text{-tangled }\right) \leqslant c\ell^3 \frac{(d-1)^{4\ell}}{n}.$$

Then, for $\ell = \ell_n$ large enough, we need to show that $\|\underline{B^{(\ell)}}\|$ is close to $\rho(B_*)^\ell$ and that $\|R_k^{(\ell)}\|$ is negligible. The adequate controls can be stated as follows:

Proposition 4.7. *Let* $(a_i)_{i\in[d]}$ *be fixed, satisfying the symmetry condition* (H$\boldsymbol{a}$) *and assume that* $\max(\|a_i\| \vee \|a_i^{-1}\|^{-1}) \leqslant 1/\varepsilon$. *Then,* $\exists c, \rho_1 > 0$, *for all* $1 \leqslant \ell \leqslant \log n$,

$$\mathbb{P}\left(\|\underline{B^{(\ell)}}\| \leqslant (\log n)^{20} (\rho(B_*) + \varepsilon)^\ell \right) \geqslant 1 - c \exp\Big(-\frac{\ell \log n}{c \log\log n} \Big),$$

$$\mathbb{P}\left(\|R_k^{(\ell)}\| \leqslant (\log n)^{40} \rho_1^\ell \right) \geqslant 1 - c \exp\Big(-\frac{\ell \log n}{c \log\log n} \Big).$$

Now, by choosing $\ell_n \sim \frac{\log n}{\kappa}$, with $\kappa > 1$, satisfying $\kappa > \log\left((d-1)^4 \vee \left(\frac{4\rho_1}{\varepsilon}\right)\right)$ and using a net argument on the a_i's that we do not detail here, one can get the required bound (4). This concludes the proof of Theorem 3.13. As explained above, through the linearization trick, we get Theorem 3.16 as a corollary.

5. Applications of strong asymptotic freeness

To summarize, the construction of a sequence of almost-Ramanujan (colored, weighted) graphs for which $(A_n)_{n\geqslant 1}$ plays the role of (generalized) adjacency operators is equivalent to strong asymptotic freeness for a family of permutation matrices. We conclude this presentation by listing a few other results involving strong convergence or strong asymptotic freeness and their consequences in other domains.

5.1. Reminder on the link between strong asymptotic freeness and outliers of random matrices

In Section 3.4, we have defined strong asymptotic freeness and clarified in Proposition 3.15 its relation to outliers of random matrices. To illustrate this link, we will cite a result of Collins and Male (2014) in which they use strong asymptotic freeness to show the absence of outliers for some ensembles of random matrices.

Proposition 5.1. *Let $\mathbf{U}_n$ be a p-tuple of $n \times n$ independent Haar unitary matrices and $\mathbf{Y}_n$ a q-tuple of $n \times n$ matrices that are independent[12] of $\mathbf{U}_n$. Let $\mathbf{u}$ be a p-tuple of Haar unitaries and $\mathbf{y}$ a q-tuple of random variables, freely independent from $\mathbf{u}$ in a C^*-algebra $(\mathscr{A}, \tau)$. If $(\mathbf{Y}_n)_{n\geqslant 1}$ converges strongly to $\mathbf{y}$, then $(\mathbf{U}_n, \mathbf{Y}_n)_{n\geqslant 1}$ converges strongly to $(\mathbf{u}, \mathbf{y})$.*

As an immediate corollary, they got the following result:

Corollary 5.2. *Let A_n, B_n be two $n\times n$ independent Hermitian random matrices. Assume that:*

- *the law of one of the matrices is invariant under unitary conjugation,*
- *almost surely, the empirical spectral measure of A_n (respectively B_n) converges to a compactly supported probability measure μ (respectively ν),*
- *almost surely, for any neighborhood of the support of μ (respectively ν), for n large enough, the eigenvalues of A_n (respectively B_n) belong to the respective neighborhood.*

Then almost surely, for n large enough, the eigenvalues of $A_n + B_n$ belong to a small neighborhood of the support of $\mu \boxplus \nu$, where $\boxplus$ denotes the free additive convolution[13].

[12] They can be deterministic.

[13] We don't want to get into the detail of the definition of this operation on measures. Here, we just need to know that $\mu \boxplus \nu$ is the distribution of $a + b$, where a has distribution μ, b has distribution ν and a and b are freely independent.

If the third condition in the corollary is not fulfilled, then outliers may appear. This phenomenon has been extensively studied in the RMT framework and most results can be understood through free probability theory, but with different tools. In this direction, we strongly recommend the review paper on deformed models by Capitaine and Donati-Martin (2017).

5.2. $\mathrm{Ext}(C^*_{\mathrm{red}}(F_r))$ is not a group

As mentioned at the beginning of the paper, there has been a constant interplay between RMT and operator algebra. It is in particular striking to see that the first result explicitely involving strong convergence and strong asymptotic freeness was in a paper by Haagerup and Thorbjørnsen (2005) entitled *A new application of random matrices:* $\mathrm{Ext}(C^*_{\mathrm{red}}(F_2))$ *is not a group*. This paper is a very important step in the theory for several reasons. One of them is that it introduced in the RMT framework the linearization trick, presented above in the operator algebra framework. This strategy has been applied since then in all the strong convergence results we are aware of. Although it is a bit far from the motivation of Bordenave and Collins (2019), we will describe in detail this result, mainly for historical reasons. We first give a definition of $\mathrm{Ext}(C^*_{\mathrm{red}}(F_r))$. We start from the Hilbert space $\mathscr{H} = \ell^2(\mathbb{N})$. Its Calkin algebra is $\mathscr{C}(\mathscr{H}) := \mathscr{B}(\mathscr{H})/\mathscr{K}(\mathscr{H})$, which is the quotient of the algebra $\mathscr{B}(\mathscr{H})$ of bounded operators on $\mathscr{H}$ by the ideal $\mathscr{K}(\mathscr{H})$ of compact operators, the quotient map being denoted by $q\colon \mathscr{B}(\mathscr{H}) \to \mathscr{C}(\mathscr{H})$. We also denote by $\mathscr{U}(\mathscr{H})$ the unitary group of $\mathscr{B}(\mathscr{H})$. For $\mathscr{A}$ a C^*-algebra, if π_1 and π_2 are two one-to-one $*$-homomorphisms from $\mathscr{A}$ to $\mathscr{C}(\mathscr{H})$, we define the equivalence relation

$$\pi_1 \sim \pi_2 \iff \exists u \in \mathscr{U}(\mathscr{H}), \forall a \in \mathscr{A}, \pi_2(a) = q(u)\pi_1(a)q(u)^*.$$

Then $\mathrm{Ext}(\mathscr{A})$ is the set of equivalence classes for this equivalence relation. As $\mathscr{H} \oplus \mathscr{H} \simeq \mathscr{H}$, $(\pi_1, \pi_2) \mapsto \pi_1 \oplus \pi_2$ defines a natural semigroup structure on $\mathrm{Ext}(\mathscr{A})$. The question to know under which condition on $\mathscr{A}$ it would be a group was controversial in operator algebra in the nineties. Voiculescu obtained that $\mathrm{Ext}(\mathscr{A})$ was a unital semigroup for all separable unital C^*-algebras $\mathscr{A}$. Anderson (1978) provided the first example for which $\mathrm{Ext}(\mathscr{A})$ is not a group and, if F_r denotes the free group with r generators, the question for $\mathscr{A} = C^*_{\mathrm{red}}(F_r)$ remained open for a long time. Voiculescu (1993) gave the following very useful criterion:

Theorem 5.3. *If there exists a sequence of unitary representations* $\pi_n\colon F_r \to M_n(\mathbb{C})$ *such that* $\forall h_1, \dots, h_m \in F_r$ *and* $c_1, \dots, c_m \in \mathbb{C}$,

$$\lim_{n\to\infty} \left\| \sum_{j=1}^m c_j \pi_n(h_j) \right\| = \left\| \sum_{j=1}^m c_j \lambda(h_j) \right\|, \tag{9}$$

then $\mathrm{Ext}(C^*_{\mathrm{red}}(F_r))$ *is not a group.*

Now, let us explain how strong asymptotic freeness for independent GUE matrices, shown in (HAAGERUP and THORBJØRNSEN, 2005), implies that (9) holds true. The idea is to construct explicitly the sequence $(\pi_n)_{n\geqslant 1}$ as follows: we let

$$\varphi(t) = \begin{cases} -\pi, & \text{if } t \leqslant -2 \\ \pi, & \text{if } t \geqslant 2 \\ \int_0^t \sqrt{4-s^2}\,\mathrm{d}s & \text{if } -2 \leqslant t \leqslant 2 \end{cases}$$

and $\psi(t) := \exp(i\varphi(t))$. If $(s_i)_{i\in[r]}$ is a family of semicircular elements that are freely independent and $u_i = \psi(s_i)$, then there is an isomorphism $\Phi\colon C^*_{\mathrm{red}}(F_r) \to C^*((u_i)_{i\in[r]})$ such that $\Phi(\lambda(g_i)) = u_i$. If $(X_{1,n},\dots,X_{r,n})$ are independent $\mathrm{GUE}(n)$ matrices, and $\forall i \in [r]$, $U_{i,n}(\omega) = \psi(X_{i,n}(\omega))$, we obtain a sequence of unitary matrices and, for any $\omega \in \Omega$, there exists $\pi_{n,\omega}\colon F_r \to \mathscr{U}(M_n(\mathbb{C}))$ such that $\pi_{n,\omega}(g_i) = U_{i,n}(\omega)$. Then using strong asymptotic freeness, one can check that $\forall \omega \in \Omega$, $\forall h_1,\dots,h_m \in F_r$ and $c_1,\dots,c_m \in \mathbb{C}$,

$$\lim_{n\to\infty} \Big\| \sum_{j=1}^m c_j \pi_{n,\omega}(h_j) \Big\| = \Big\| \sum_{j=1}^m c_j \lambda(h_j) \Big\|,$$

and it is enough to choose $\pi_n = \pi_{n,\omega}$, with ω in the set of probability 1 on which this last equality holds. Some generalizations of the results of Haagerup and Thorbjørnssen have been conjectured, in relation to the Peterson–Thom conjecture (see e.g. the recent work of HAYES (2020)).

5.3. Estimation of the norm of random matrices

In this last part, we explain how strong convergence results can be used to give interesting bounds on the norm of random matrices. We will present asymptotic bounds that are in majority consequences of Proposition 5.1, but also remarkable non-asymptotic bounds that have been recently obtained by BANDEIRA, BOEDIHARDJO, and VAN HANDEL (2021). Indeed, strong convergence implies convergence of the norm of polynomials in random matrices to their free counterpart and there are several examples for which the norm of the free counterpart has been computed. In particular, AKEMANN and OSTRAND (1976) showed that, if $u_1,\dots,u_p$ are Haar unitaries that are freely independent, then, for any $a_1,\dots,a_p \in \mathbb{R}$, we have

$$\Big\| \sum_{i=1}^p a_i u_i \Big\| = \min_{t\geqslant 0} \Big\{ 2t + \sum_{i=1}^p \big(\sqrt{t^2 + |a_i|^2} - t \big) \Big\}. \tag{10}$$

In particular,

$$\Big\| \sum_{i=1}^p u_i \Big\| = 2\sqrt{p-1}.$$

We can therefore deduce that, if $U_{1,n}, \dots, U_{p,n}$ are independent Haar unitary random matrices, then almost surely,

$$\left\| \sum_{i=1}^{p} U_{i,n} \right\| \xrightarrow[n\to\infty]{} 2\sqrt{p-1},$$

which is like a unitary analogue of the Alon–Boppana bound given in Theorem 2.1. In the same vein, motivated by questions for random walks on the free group, Kesten (1959) showed that

$$\left\| \sum_{i=1}^{p} (u_i + u_i^*) \right\| = 2\sqrt{2p-1},$$

so that

$$\left\| \sum_{i=1}^{p} (U_{i,n} + U_{i,n}^*) \right\| \xrightarrow[n\to\infty]{} 2\sqrt{p-1}.$$

Note that Lehner (1999) also gave a formula for the norm of the operator A_*, which is a generalisation of (10) to the case when $a_0, \dots, a_d$ are matrices.

As a conclusion, we now present remarkable non-asymptotic bounds, that is concentration inequalities for random matrices, recently obtained Bandeira, Boedihardjo, and van Handel (2021). Their initial motivation is to understand the spectral norm of an arbitrary $d \times d$ self-adjoint random matrix with centered, jointly Gaussian entries. Such a matrix X can be written

$$X := \sum_{i=1}^{p} g_i A_i, \tag{11}$$

where A_i are deterministic self-adjoint $d \times d$ matrices and $(g_i)_{i\in[p]}$ are independent real standard Gaussian variables. It is known that, if we define $\Sigma(X) := \left\| \sum_{i=1}^{p} A_i^2 \right\|$, then we have the bound

$$c\Sigma(X) \leqslant \mathbb{E}\|X\| \leqslant C\Sigma(X)\sqrt{\log d}.$$

Their goal is to improve the upper bound. To get a free analogue of the matrix (11), a natural idea is to replace the Gaussian variables by semi-circular elements to define

$$X_{\text{free}} := \sum_{i=1}^{p} A_i \otimes s_i,$$

where $(s_i)_{i\in[p]}$ are freely independent semicircular elements. The authors establish a general bound of the form:

$$\mathbb{E}\|X\| \leqslant \|X_{\text{free}}\| + Cv(X)^{1/2}\Sigma(X)^{1/2}(\log d)^{3/4}, \tag{12}$$

with $v(X) = \|\mathrm{Cov}(X)\|$ being the spectral norm of the covariance matrix of X. As a consequence, they got an inclusion of the spectrum which is reminiscent of (4): with high probability,

$$\sigma(X) \subset \sigma(X_{\mathrm{free}}) + [-\varepsilon, \varepsilon],$$

where ε is of order $v(X)^{1/2}\Sigma(X)^{1/2}(\log d)^{3/4}$. The bound is particularly relevant when ε is small in comparison to $\|X_{\mathrm{free}}\|$. They treat a large variety of examples for which the bound (12) improves on known results (random matrices with independent entries, sparse Wigner matrices etc.) or gives new concentration inequalities (patterned random matrices, independent block matrices etc.) From these non-asymptotic bounds, they can also deduce strong asymptotic freeness for a lot of models, showing that this property is much more ubiquitous than expected. The interplay between strong asymptotic freeness and random matrix theory is certainly to be continued.

References

Akemann, C. A. and Ostrand, P. A. (1976). "Computing norms in group C^*-algebras", *Amer. J. Math.* **98** (4), pp. 1015–1047.

Alon, N. (1986). "Eigenvalues and expanders", in: vol. 6. 2. Theory of computing (Singer Island, Fla., 1984), pp. 83–96.

Amit, A. and Linial, N. (2002). "Random graph coverings. I. General theory and graph connectivity", *Combinatorica* **22** (1), pp. 1–18.

Anderson, J. (1978). "A C^*-algebra $\mathscr{A}$ for which $\mathrm{Ext}(\mathscr{A})$ is not a group", *Ann. of Math.* (2) **107** (3), pp. 455–458.

Bandeira, A. S., Boedihardjo, M. T., and van Handel, R. (2021). *Matrix Concentration Inequalities and Free Probability*. Version 1. arXiv: 2108.06312.

Benaych-Georges, F., Bordenave, C., and Knowles, A. (2020). "Spectral radii of sparse random matrices", *Ann. Inst. Henri Poincaré Probab. Stat.* **56** (3), pp. 2141–2161.

Bordenave, C. (2020). "A new proof of Friedman's second eigenvalue theorem and its extension to random lifts", *Ann. Sci. Éc. Norm. Supér.* (4) **53** (6), pp. 1393–1439.

Bordenave, C. and Collins, B. (2019). "Eigenvalues of random lifts and polynomials of random permutation matrices", *Ann. of Math.* (2) **190** (3), pp. 811–875.

——— (2020). *Strong asymptotic freeness for independent uniform variables on compact groups associated to non-trivial representations*. Version 2. arXiv: 2012.08759.

Capitaine, M. and Donati-Martin, C. (2017). "Spectrum of deformed random matrices and free probability", in: *Advanced topics in random matrices*. Vol. 53. Panor. Synthèses. Soc. Math. France, Paris, pp. 151–190.

Collins, B. and Male, C. (2014). "The strong asymptotic freeness of Haar and deterministic matrices", *Ann. Sci. Éc. Norm. Supér.* (4) **47** (1), pp. 147–163.

Friedman, J. (2008). "A proof of Alon's second eigenvalue conjecture and related problems", *Mem. Amer. Math. Soc.* **195** (910), pp. viii+100.

Füredi, Z. and Komlós, J. (1981). "The eigenvalues of random symmetric matrices", *Combinatorica* **1** (3), pp. 233–241.

Greenberg, Y. (1995). "Spectra of graphs and their covering trees". PhD thesis. Hebrew University of Jerusalem.

Haagerup, U. and Thorbjørnsen, S. (2005). "A new application of random matrices: $\mathrm{Ext}(C^*_{\mathrm{red}}(F_2))$ is not a group", *Ann. of Math.* (2) **162** (2), pp. 711–775.

Hayes, B. (2020). *A random matrix approach to the Peterson-Thom conjecture*. Version 2. arXiv: 2008.12287.

Kesten, H. (1959). "Symmetric random walks on groups", *Trans. Amer. Math. Soc.* **92**, pp. 336–354.

Kowalski, E. (2019). *An introduction to expander graphs*. Vol. 26. Cours Spécialisés [Specialized Courses]. Société Mathématique de France, Paris, pp. x+276.

Lehner, F. (1999). "Computing norms of free operators with matrix coefficients", *Amer. J. Math.* **121** (3), pp. 453–486.

Lubotzky, A., Phillips, R., and Sarnak, P. (1988). "Ramanujan graphs", *Combinatorica* **8** (3), pp. 261–277.

Mingo, J. A. and Speicher, R. (2017). *Free probability and random matrices*. Vol. 35. Fields Institute Monographs. Springer, New York; Fields Institute for Research in Mathematical Sciences, Toronto, pp. xiv+336.

Nica, A. (1993). "Asymptotically free families of random unitaries in symmetric groups", *Pacific J. Math.* **157** (2), pp. 295–310.

Pisier, G. (1996). "A simple proof of a theorem of Kirchberg and related results on C^*-norms", *J. Operator Theory* **35** (2), pp. 317–335.

Soshnikov, A. (1999). "Universality at the edge of the spectrum in Wigner random matrices", *Comm. Math. Phys.* **207** (3), pp. 697–733.

Speicher, R. (2019). *Lecture Notes on "Free Probability Theory"*. Version 1. arXiv: 1908.08125.

Voiculescu, D. (1991). "Limit laws for random matrices and free products", *Invent. Math.* **104** (1), pp. 201–220.

——— (1993). "Around quasidiagonal operators", *Integral Equations Operator Theory* **17** (1), pp. 137–149.

Mylène Maïda

Univ. Lille, CNRS, UMR 8524
Laboratoire Paul Painlevé,
F-59000 Lille, France.

E-mail: mylene.maida@univ-lille.fr

Séminaire BOURBAKI Janvier 2023
74e année, 2022–2023, no 1203, p. 225 à 249
doi : 10.24033/ast.1212

ALGÈBRES DE VON NEUMANN, PRODUITS TENSORIELS, CORRÉLATIONS QUANTIQUES ET CALCULABILITÉ
[*d'après* Ji, Natarajan, Vidick, Wright et Yuen]

par **Mikael de la Salle**

Introduction

Le but de ce texte est de présenter la résolution récente de problèmes qui sont longtemps restés ouverts en algèbres d'opérateurs. Commençons par les énoncer.

Le problème de plongement de Connes (1976) demande si tout facteur II_1 admet un plongement approximatif dans le facteur II_1 hyperfini $\mathscr{R}$. Autrement dit, s'il peut être réalisé comme une sous-algèbre de von Neumann d'un ultraproduit d'algèbres de matrices. C'est une façon précise de demander si les algèbres de matrices forment un modèle suffisant pour comprendre tous les phénomènes locaux dans les algèbres de von Neumann finies.

Le problème peut être énoncé de manière équivalente sans avoir à introduire de vocabulaire d'algèbres de von Neumann. Étant donné un groupe Γ, un *caractère* est une fonction $\varphi\colon \Gamma \to \mathbf{C}$ définie positive [(1)], invariante par conjugaison et normalisée par $\varphi(1) = 1$. Les caractères qui interviennent dans la théorie des représentations des groupes finis (de la forme $\gamma \mapsto \frac{1}{d}\mathrm{Tr}(\pi(\gamma))$ pour une représentation unitaire de dimension finie d) sont des exemples de caractères, que nous appellerons caractères de dimension finie. Mais il y en a bien d'autres pour les groupes infinis, par exemple la fonction indicatrice de $\{1\}$, ou plus généralement de tout sous-groupe distingué d'indice infini. Le problème de plongement de Connes est équivalent à la question « Tout caractère du groupe libre à deux générateurs est-il limite simple d'une suite de caractères de dimension finie ? » Si on pose la question seulement pour les caractères de la forme $\varphi = \chi_N$ pour un sous-groupe distingué N, on obtient une autre question importante (elle toujours ouverte) « Tout groupe est-il hyperlinéaire ? », un affaiblissement de la question de Gromov « Tout groupe est-il sofique ? » (qui correspond à demander si χ_N est limite simple de caractères de dimension finie provenant de représentations à valeurs dans les groupes de matrices de permutations).

(1) C'est-à-dire, pour toute famille finie $\gamma_1, \dots, \gamma_n \in \Gamma$, la matrice $(\varphi(\gamma_i^{-1}\gamma_j))_{i,j\leqslant n}$ est positive

La conjecture de Kirchberg (1993) porte sur les produits tensoriels de C^*-algèbres. Une façon courte de l'énoncer est : y a-t-il une unique norme de C^*-algèbre sur $\mathscr{C} \otimes \mathscr{C}$, où $\mathscr{C}$ est la C^*-algèbre universelle engendrée par une suite dénombrable d'unitaires. Par la propriété universelle de $\mathscr{C}$, cette conjecture a de nombreuses reformulations très différentes ; le lecteur intéressé pourra en savoir beaucoup plus sur ces formes (et beaucoup d'autres choses) dans le livre de Pisier (2020), qui est entièrement dédié à ce sujet.

Le problème de Tsirelson est lié aux fondements de la mécanique quantique et, via les inégalités de Bell, à la fameuse expérience d'Alain Aspect démontrant le phénomène d'intrication quantique et qui lui a valu le prix Nobel. Dans cette expérience, Aspect mesure des corrélations entre certaines observables entre systèmes physiques. Il observe que celles-ci sont incompatibles avec la théorie des variables cachées proposée par Einstein, Podolski et Rosen, mais qu'elles sont bien compatibles avec le formalisme mathématique de la mécanique quantique reposant sur les espaces de Hilbert. Jusqu'à preuve du contraire, le monde physique est donc bien quantique et les espaces de Hilbert sont nécessaires à sa description. Tsirelson (1980, 1993) étudie différentes variantes de ce formalisme mathématique, et notamment (il y a là un petit raccourci) une distinction entre espaces de dimension finie et espaces de dimension infinie. Tsirelson (1993) affirme, sans preuve, que, pour ces deux modèles, les corrélations possibles sont essentiellement les mêmes, dans le sens où toute corrélation dans le modèle avec des espaces de Hilbert de dimension infinie est une limite de corrélations dans le modèle avec des espaces de dimension finie. C'est en effet naturel, puisqu'un espace de Hilbert est une union filtrante de ses sous-espaces de dimension finie. Ce n'est que dans les années 2000, avec l'explosion de la théorie quantique de l'information, que ces travaux de Tsirelson ont été étudiés attentivement (Navascués, Pironio et Acín, 2008) ; Tsirelson reconnaît alors qu'il a été un peu rapide et son affirmation devient donc le problème de Tsirelson.

De manière remarquable et pas du tout évidente, ces trois problèmes sont équivalents (et sont parfois appelés *conjectures*, même si Kirchberg est le seul à avoir énoncé sa question comme une conjecture). Le chemin le plus délicat, l'équivalence entre le problème de Connes et la conjecture de Kirchberg a été démontrée par Kirchberg, 1993. L'équivalence entre le problème de Tsirelson et la conjecture de Kirchberg a été démontrée plus récemment par Fritz (2012) et Junge et al. (2011) pour une direction, et Ozawa (2013) pour l'autre.

En janvier 2020, la solution de toutes ces questions a été déposée sur arXiv par une équipe de cinq informaticiens.

Théorème 0.1 (Ji et al., 2020a). *Le problème de Connes–Kirchberg–Tsirelson a une réponse négative.*

Autrement dit, il existe un facteur II_1 qui n'est pas plongeable dans un ultraproduit d'algèbres de matrices ; il y a au moins deux normes de C^*-algèbres sur $\mathscr{C} \otimes \mathscr{C}$;

il existe des inégalités de Bell qui distinguent strictement les corrélations quantiques dans le modèle de la mécanique quantique où l'on autorise des espaces de Hilbert de dimension infinie de celles dans le modèle où seuls des espaces de Hilbert de dimension finie (mais arbitraire) sont autorisés. On peut donc imaginer qu'il existe une expérience physique qui pourrait démontrer expérimentalement que les espaces de Hilbert de dimension infinie sont indispensables pour décrire le monde physique…

Le théorème 0.1 est énoncé dans Ji et al. (2020a), mais sa longue preuve est répartie aussi dans Bavarian, Vidick et Yuen (2022) et Ji et al. (2020b, 2022). Elle repose également sur plusieurs autres travaux antérieurs, notamment Natarajan et Wright (2019). Ce travail est très long et difficile. Il repose sur des idées nouvelles, mais aussi sur des décennies d'avancées en informatique théorique, informatique quantique et théorie quantique de l'information. La première version proposée reposait d'ailleurs sur des résultats dont la preuve s'est avérée fausse, et qui ont demandé aux auteurs un travail considérable de correction (Ji et al., 2020b, 2022). Il est illusoire d'en présenter une preuve complète en quelques pages. Le but de ce texte est de présenter de manière très superficielle la stratégie générale de la preuve telle que je la comprends. On s'éloignera par moments de l'approche initiale, en tentant d'être le plus compréhensible possible pour un public de mathématiciens. Par exemple, on essaiera de ne parler de classes de complexité que quand c'est vraiment nécessaire, là où les auteurs de l'article initial concentrent leurs efforts à démontrer un énoncé de complexité (MIP*=RE) dont ils déduisent assez directement le résultat mathématique. L'article original et l'excellent survol de Vidick (2022b) sont des très bons endroits pour comprendre les aspects informatiques. Une autre différence notable est qu'on réfutera directement le problème de Connes, là où les auteurs, sans doute motivés par un sens physique dont je manque cruellement, réfutaient d'abord le problème de Tsirelson. Du point de vue mathématique, cela revient à étudier uniquement des états traciaux sur des algèbres de von Neumann là où les auteurs étudient des états arbitraires, et je pense que cela ajoute beaucoup de difficultés inutiles.

Ce texte est dédié à Eberhard Kirchberg (1946–2022) et Boris Tsirelson (1950–2020), deux grands noms de l'analyse fonctionnelle et protagonistes centraux de cette histoire, qui sont décédés peu après l'annonce de sa résolution.

Je voudrais remercier les très nombreux collègues qui ont accepté de répondre à toutes mes questions, parfois naïves. Les lister tous serait trop long, mais mentionnons au moins les auteurs Thomas Vidick et Henry Yuen, mais aussi Guillaume Aubrun, Laurent Bartoldi, Michael Chapman, Emilie Elkiaer, Omar Fawzi, Cécilia Lancien, Amine Marrakchi, Paul Meunier, Sophie Morel, Étienne Moutot, Alexander Müller-Hermes, Magdalena Musat, Pascal Koiran, Mikael Rørdam, Bruno Sévennec, Todor Tsankov… Je remercie plus particulièrement Guillaume Aubrun et Thomas Vidick pour leur relecture attentive et constructive de ce texte.

1. Algèbres de von Neumann et calculabilité

1.1. Algèbres de von Neumann traciales et approximation par des algèbres de matrices

Une algèbre de von Neumann $\mathscr{M}$ est une algèbre d'opérateurs sur un espace de Hilbert $\mathscr{H}$, stable par l'adjoint, contenant l'identité de $\mathscr{H}$ et fermée pour la topologie préfaible [(2)]. Une algèbre de von Neumann est dite finie si elle admet un *état tracial* (ou simplement une trace) $\tau\colon \mathscr{M} \to \mathbf{C}$: une forme linéaire préfaiblement continue, normalisée par $\tau(1) = 1$ et vérifiant $\tau(x^*x) = \tau(xx^*) > 0$ pour tout $x \in \mathscr{M}$ non nul. La paire $(\mathscr{M}, \tau)$ est appelée une algèbre de von Neumann traciale.

Les exemples évidents d'algèbres de von Neumann traciales sont les algèbres de matrices $(M_d(\mathbf{C}), \mathrm{tr} := \frac{1}{d}\mathrm{Tr})$. Et ces exemples permettent d'en construire beaucoup d'autres avec la technique d'ultraproduit : si $\mathscr{U}$ est un ultrafiltre sur un ensemble I, et si $d_i \in \mathbf{N}$ pour tout I, on peut définir l'ultraproduit $\prod_{\mathscr{U}}(M_{d_i}(\mathbf{C}), \mathrm{tr})$ comme le quotient de $\prod_i M_{d_i}(\mathbf{C})$, l'espace des suites bornées en norme d'opérateur, par $\{(x_i) \in \prod_i M_{d_i}(\mathbf{C}) \mid \lim_{\mathscr{U}} \mathrm{tr}(x_i^* x_i) = 0\}$. Muni de la trace $\tau((x_i)) = \lim_{\mathscr{U}} \mathrm{tr}(x_i)$, c'est une algèbre de von Neumann traciale. Le problème de plongement de Connes demande s'il y a d'autres algèbres de von Neumann traciales que les sous-algèbres de von Neumann de tels ultraproduits. La construction GNS (Bekka, Harpe et Valette, 2008, Theorem C.4.10) et le fait que toute algèbre de von Neumann traciale finiment engendrée peut être réalisée dans une algèbre de von Neumann traciable engendrée par deux unitaires justifient l'équivalence entre cette forme du problème de Connes et celle donnée dans l'introduction.

1.2. Approximation de caractères et calculabilité

Plutôt que de travailler avec des caractères sur un groupe donné (le groupe libre à deux générateurs) comme dans l'introduction, on travaillera avec une famille de groupes, indexée par les paires (n, m) d'entiers positifs, et on n'étudiera les caractères qu'en restriction à des parties finies de plus en plus grandes. Cela permettra de faire entrer en jeu des notions de calculabilité. Concrètement, notons

- $\Gamma_{n,m} = (\mathbf{Z}/n\mathbf{Z})^{*m}$ le produit libre de m copies du groupe fini cyclique d'ordre n,
- $S_{n,m} \subset \Gamma_{n,m}$ sa partie génératrice finie donnée par l'union des m copies de $\mathbf{Z}/n\mathbf{Z}$,
- pour tout entier $d \geqslant 1$, $C_d(n, m) \subset \mathbf{C}^{S_{n,m}^2}$ l'enveloppe convexe des restrictions à $S_{n,m}^2$ (la boule de rayon 2, c'est-à-dire l'ensemble des produits de deux éléments de $S_{n,m}$) de caractères de dimension $\leqslant d$ de $\Gamma_{n,m}$,

[(2)] $B(\mathscr{H})$ est le dual des opérateurs à trace sur $\mathscr{H}$

- $C_{<\infty}(n,m) = \cup_d C_d(n,m)$.
- pour tout $\varepsilon > 0$, $f_\varepsilon(n,m)$ le plus petit entier d tel que $C_\infty(n,m)$ est contenu dans le ε-voisinage [(3)] de $C_d(n,m)$.

L'énoncé suivant est une variante d'arguments de Navascués, Pironio et Acín (2008).

Proposition 1.1. *Si le problème de plongement de Connes avait une réponse positive, alors la fonction $(k,n,m) \mapsto f_{\frac{1}{k}}(n,m)$ serait majorée par une fonction calculable $\mathbf{N}^3 \to \mathbf{N}$.*

Démonstration. Supposons que le problème de Connes ait une réponse positive. En particulier, pour tout n,m, $C_{<\infty}(n,m)$ est dense dans $C(n,m)$, l'ensemble des restrictions à $S^2_{n,m}$ de caractères de $\Gamma_{n,m}$. Être un caractère, c'est donné par une famille dénombrable d'inégalités, que l'on peut énumérer explicitement. Notons $C^d(n,m)$ le polytope calculable donné comme les restrictions à $S^2_{n,m}$ des fonctions ne vérifiant que les d premières inégalités. Clairement, $C^d(n,m)$ décroît vers $C(n,m)$. On a donc une approximation *par dessus* de $C(n,m)$. En parallèle, en décrivant une partie $\frac{1}{d}$-dense des matrices unitaires de taille d et d'ordre n, on peut obtenir une suite calculable contenue dans $C_d(n,m)$ et approchant $C_{<\infty}(n,m)$ *par dessous*. Le plus petit d tel que l'approximation par dessous est $\frac{1}{k}$-dense dans l'approximation par dessus est donc fini (car on a supposé que $C_{<\infty}(n,m)$ est dense dans $C(n,m)$) et calculable. C'est clairement un majorant de $f_{\frac{1}{k}}(n,m)$. □

Le théorème principal est

Théorème 1.2 (Ji et al., 2020a)**.** *La fonction $(k,n,m) \mapsto f_{\frac{1}{k}}(n,m)$ n'est pas majorée par une fonction calculable.*

Il a toujours été clair pour moi que la raison pour laquelle le problème de plongement de Connes ou la conjecture de Kirchberg étaient difficiles est qu'il y a beaucoup de choses qu'on ne comprend pas dans les algèbres d'opérateurs de dimension infinie (et en particulier dans la notion de commutation en dimension infinie), contrairement à la dimension finie où tout est limpide. Il est frappant que le théorème 1.2, duquel la réfutation du problème de Connes découle directement, ne porte pas du tout sur les algèbres d'opérateurs de dimension infinie. Il affirme, contrairement à l'intuition, que les algèbres d'opérateurs de dimension finie sont *extrêmement compliquées*, non pas du point de vue leur description mathématique, mais du point de vue de la calculabilité.

(3) Pour fixer les idées, disons qu'on a muni $\mathbf{C}^{S^2_{n,m}}$ de la norme ℓ_∞, mais n'importe quelle autre norme raisonnable (dans le sens comparable à la norme ℓ_∞ à des constantes calculables près) conviendrait. L'existence de $f_\varepsilon(n,m)$ est immédiate par compacité.

Une des fonctions non calculables les plus élémentaires est la fonction d'arrêt des machines de Turing. La preuve du théorème 1.2 passe par une réduction à cette fonction d'arrêt. Un énoncé plus précis duquel le théorème 1.2 découle facilement (exercice) est le suivant.

Théorème 1.3 (Ji et al., 2020a). *Il existe un fonction calculable qui, évaluée en une machine de Turing M, renvoie toujours un triplet (n, m, φ) avec $\varphi \in (\mathbf{C}^{S^2_{n,m}})^*$ et vérifiant*

$$\sup_{C_{<\infty}(n,m)} \varphi \begin{cases} = 1 & \text{si } M \text{ s'arrête,} \\ \leqslant \frac{1}{2} & \text{sinon.} \end{cases}$$

La forme linéaire φ dans ce théorème n'est pas arbitraire, elle est de la forme « valeur d'un jeu ».

2. Jeux

La notion de jeu que nous considérons ici est celle de jeu à deux joueurs et une manche, introduite dans le contexte de preuve interactive à deux joueurs par Ben-Or et al. (1988). Nous ne considérerons d'ailleurs que des versions symétriques. Comme c'est la seule notion de jeu que nous considérerons, nous utiliserons simplement le mot jeu.

Dans tout ce texte, un jeu est donc la donnée de $\mathscr{G} = (\mathscr{X}, \mu, \mathscr{A}, D)$ où $\mathscr{X}$ et $\mathscr{A}$ sont des ensembles finis, μ est une mesure de probabilité sur $\mathscr{X} \times \mathscr{X}$ et $D\colon \mathscr{X} \times \mathscr{X} \times \mathscr{A} \times \mathscr{A} \to \{0, 1\}$ est une fonction symétrique.

2.1. Aparté : interprétation des jeux

La raison pour laquelle on parle de jeu est la suivante. Il faut imaginer qu'il y a trois protagonistes : un arbitre et deux joueurs. L'arbitre peut communiquer avec les joueurs, mais les joueurs ne peuvent communiquer entre eux. $\mathscr{X}$ est *l'ensemble des questions*, $\mathscr{A}$ est *l'ensemble des réponses possibles*. L'arbitre tire au hasard une paire $(x, y) \in \mathscr{X} \times \mathscr{X}$ selon la loi μ. Il communique la question x au premier joueur, et la question y au second joueur ; la question posée à chaque joueur est donc inconnue de l'autre. En retour, le premier joueur donne une réponse a parmi les réponses possibles, et le second joueur répond b. Les joueurs gagnent tous les deux si $D(x, y, a, b) = 1$, et perdent tous les deux sinon. Il s'agit donc d'un jeu coopératif ; l'intérêt des joueurs est de trouver la stratégie qui optimise la probabilité de gagner.

Là où les choses se compliquent, c'est quand on se demande quelles sont les stratégies possibles. Une *stratégie classique*, c'est une stratégie où la réponse de chaque joueur est une fonction déterministe de sa question : $a = f(x)$ et $b = g(y)$. On peut aussi imaginer des stratégies où les joueurs ont accès à une source d'aléa indépendante des questions. Mais une telle stratégie peut être vue comme une combinaison convexe de stratégies déterministes, et donc ne peut améliorer la probabilité de gagner.

Une *stratégie quantique*, c'est une stratégie où les joueurs sont autorisés à partager de l'intrication quantique. Mathématiquement, cela veut dire qu'il existe un espace de Hilbert $\mathscr{H}$, un vecteur unité $\psi \in \mathscr{H}$, et, pour tout x, deux partitions de l'unité $(p_a^x)_{a\in A}$ et $(q_b^x)_{b\in B}$ dans $B(\mathscr{H})$ tels que $[p_a^x, q_b^y] = 0$ pour tout $x, y \in \mathscr{X}$ et $a, b \in \mathscr{A}$. Alors la probabilité que les joueurs répondent a, b à x, y est $\langle p_a^x q_b^y \psi, \psi\rangle$. Le formalisme de la mécanique quantique dit précisément que c'est le genre de choses que les joueurs peuvent faire sans communiquer. La probabilité de gain de cette stratégie est donc

$$\sum_{x,y\in\mathscr{X},a,b\in\mathscr{A}} \mu(x,y)D(x,y,a,b)\langle p_a^x q_b^y \psi, \psi\rangle. \tag{1}$$

Dans la suite, on ne considèrera que des formes très particulières de stratégies, qui correspondent au cas où $\mathscr{H} = L^2(\mathscr{M}, \tau)$, ψ est l'image dans $L^2(\mathscr{M}, \tau)$ de l'identité de $\mathscr{M}$, et où il y a des partitions de l'unité $(e_a^x)_{a\in\mathscr{A}}$ dans $\mathscr{M}$ pour tout $x \in \mathscr{X}$, telles que p_a^x est la multiplication à gauche par e_a^x et q_b^y est la multiplication à droite par e_b^y. Ces stratégies sont habituellement appelées stratégies synchrones (Paulsen et al., 2016), mais comme c'est la seule forme de stratégie que l'on considèrera, on les appellera simplement stratégies.

2.2. Stratégies

Une stratégie pour le jeu $\mathscr{G}$ est la donnée de $\mathscr{S} = (\mathscr{M}, \tau, p)$ où $(\mathscr{M}, \tau)$ est une algèbre de von Neumann traciale et $p = (p_a^x)_{a\in\mathscr{A}}^{x\in\mathscr{X}}$ est une famille de projections orthogonales dans $\mathscr{M}$ telle que pour tout x, $(p_a^x)_{a\in\mathscr{A}}$ est une partition de l'unité :

$$\forall x \in \mathscr{X}, \sum_{a\in\mathscr{A}} p_a^x = 1.$$

La valeur d'une stratégie $\mathscr{S}$ pour le jeu $\mathscr{G}$ est

$$\mathrm{val}(\mathscr{G}, \mathscr{S}) = \sum_{x,y\in\mathscr{X},a,b\in\mathscr{A}} \mu(x,y)D(x,y,a,b)\tau(p_a^x p_b^y).$$

De manière évidente, $\mathrm{val}(\mathscr{G}, \mathscr{S}) \in [0,1]$, et $\mathrm{val}(\mathscr{G}, \mathscr{S}) = 1$ si et seulement si $p_a^x p_b^y = 0$ pour tout (x, y) dans le support de μ et tout a, b tel que $D(x, y, a, b) = 0$. On dira qu'une stratégie est parfaite si sa valeur est égale à 1 ; on dira qu'elle est bonne si sa valeur est proche de 1. Une partie importante de l'analyse mathématique des jeux sera de comprendre ce qui se passe pour une stratégie qui est bonne.

On note aussi

$$\mathrm{val}(\mathscr{G}, \mathscr{M}) = \sup\{\mathrm{val}(\mathscr{G}, \mathscr{S})\}$$

où le supremum est pris sur toutes les stratégies où l'algèbre de von Neumann est un coin de $\mathscr{M}$, c'est-à-dire de la forme $(q\mathscr{M}q, \frac{1}{\tau(q)}\tau)$ pour q une projection non nulle dans $\mathscr{M}$.

Par exemple, si $\mathscr{M}$ est une algèbre de von Neumann commutative, on retrouve presque [4] la valeur classique d'un jeu, à savoir

$$\mathrm{val}(\mathscr{G},\mathscr{M}) = \sup\Big\{ \sum_{x,y\in\mathscr{X}} \mu(x,y) D(x,y,f(x),f(y)) \mid f\colon \mathscr{X}\to\mathscr{A}\Big\}.$$

Enfin on notera

$$\mathrm{val}(\mathscr{G},d) = \mathrm{val}\big(\mathscr{G}, M_d(\mathbf{C})\big)$$

la valeur optimale d'une stratégie sur des algèbres de matrices de taille $\leqslant d$, et

$$\mathrm{val}(\mathscr{G},<\infty) = \sup_d \mathrm{val}(\mathscr{G},d).$$

Le lecteur savant aura remarqué que $\mathrm{val}(\mathscr{G},<\infty) = \mathrm{val}(\mathscr{G},\mathscr{R})$, si $\mathscr{R}$ est le facteur II_1 hyperfini.

Par la transformée de Fourier, se donner une partition de l'unité $(p_1,\dots,p_n)$ dans $\mathscr{M}$, c'est la même chose que ce donner un unitaire $u = \sum_k e^{2i\pi\frac{k}{n}} p_k \in \mathscr{M}$ vérifiant $u^n = 1$. Se donner une stratégie pour $\mathscr{G}$, c'est donc se donner un unitaire $u(x)$ dans $\mathscr{M}$ d'ordre $|\mathscr{A}|$ pour tout $x\in\mathscr{X}$. Il découle de la définition que $\mathrm{val}(\mathscr{G},\mathscr{S})$ est alors une fonction linéaire des corrélations $\tau(u(x)^k u(y)^\ell)$. Le résultat suivant est donc une forme plus fine du théorème 1.3.

Théorème 2.1. *Il existe une fonction calculable qui à une machine de Turing associe un jeu $\mathscr{G}(M)$ tel que*

$$\mathrm{val}(\mathscr{G}(M),<\infty)\begin{cases} = 1 & \text{si } M \text{ s'arrête,}\\ \leqslant \frac{1}{2} & \text{sinon.}\end{cases}$$

On verra que l'interprétation en termes de jeux à deux joueurs n'est peut-être pas la seule valable. Il est parfois pertinent de voir un jeu comme une manière de définir des algèbres de von Neumann traciales par générateurs et relations. Les jeux fournissent alors un contexte pertinent pour parler de manière quantitative d'approximations de telles algèbres de von Neumann.

Commençons par donner deux exemples très simples mais importants de ce phénomène. Pour ces exemples, comme souvent plus tard, il est plus agréable d'autoriser que l'ensemble des réponses possibles puisse dépendre de la question posée ; autrement dit, $\mathscr{A}$ n'est plus simplement un ensemble fini mais une collection $(\mathscr{A}(x))_{x\in\mathscr{X}}$ d'ensembles finis, et $D(x,y,a,b) \in \{0,1\}$ pour tous $x,y\in\mathscr{X}$ et $a\in\mathscr{A}(x), b\in\mathscr{A}(y)$ [5]. Pour définir un jeu, les valeurs de $D(x,y,\cdot,\cdot)$ pour (x,y) en dehors du support de μ ne jouent aucun rôle ; on se permettra donc de ne pas les définir.

[4] car ici, la stratégie des deux joueurs doivent coïncider : $f = g$.

[5] On retombe sur la notion plus restrictive en choisissant un ensemble fini avec une injection de $\mathscr{A}(x)$ pour tout x, et en déclarant $D = 0$ là où D n'était initialement pas défini.

2.3. Jeu de commutation

Le *jeu de commutation* est le jeu où

- $\mathscr{X} = \{x_1, x_2, y\}$, $\mathscr{A}(x_i) = \{-1, 1\}$ et $\mathscr{A}(y) = \{-1, 1\} \times \{-1, 1\}$,
- $\mu = \frac{1}{2}(\delta_{(x_1,y)} + \delta_{(x_2,y)})$,
- $D(x_1, y, a, (a', b')) = 1_{a=a'}$ et $D(x_2, y, b, (a', b')) = 1_{b=b'}$.

La propriété importante de ce jeu est qu'une stratégie de valeur proche de 1 doit être composée de projections qui commutent presque.

Lemme 2.2. *Le jeu de commutation admet des stratégies parfaites. Si une stratégie a valeur* $1 - \varepsilon$ *sur le jeu de commutation, alors sa restriction* (p_{-1}, p_1) *et* (q_{-1}, q_1) *à* x_1 *et* x_2 *respectivement vérifie*

$$\left\| [p_1 - p_{-1}, q_1 - q_{-1}] \right\|_2^2 \leqslant 64\varepsilon. \tag{2}$$

Démonstration. Le jeu a des stratégies parfaites en dimension 1, par exemple $p_1^{x_1} = p_1^{x_2} = p_{1,1}^{y} = 1$ (et tous les autres p_a^x nuls).

Si (p_1, p_{-1}) et (q_1, q_{-1}) est la restriction d'une stratégie de valeur $\geqslant 1 - \varepsilon$, il existe une partition de l'unité $(r_{a,b})_{(a,b)\in\{-1,1\}\times\{-1,1\}}$ telle que

$$\frac{1}{2}\sum_{a,b} \tau(p_a r_{a,b}) + \tau(q_b r_{a,b}) \geqslant 1 - \varepsilon.$$

Définissons $p'_a = r_{a,1} + r_{a,-1}$ et $q'_b = r_{1,b} + r_{-1,b}$. L'inégalité précédente devient $\eta_1 + \eta_2 \leqslant 4\varepsilon$, où $\eta_1 = \sum_a \|p_a - p'_a\|_2^2$ et $\eta_2 = \sum_b \|q_b - q'_b\|_2^2$. Autrement dit, la famille $\{p_a, q_b \mid a, b \in \{-1, 1\}\}$ est proche de la famille $\{p'_a, q'_b \mid a, b \in \{-1, 1\}\}$, qui est constituée de projections qui commutent. Elles commutent donc presque, ce qu'il fallait démontrer. □

2.4. Le jeu d'anticommutation

Le jeu d'anticommutation (ou jeu du carré magique) est un jeu à 15 questions, dont deux spécifiques x_1, x_2 dont les réponses attendues sont $\mathscr{A}(x_1) = \mathscr{A}(x_2) = \{-1, 1\}$. Ce qui est important n'est pas la définition précise du jeu, mais le fait qu'une stratégie à valeur proche de 1 force une forme d'anti-commutation :

Lemme 2.3. *Le jeu d'anticommutation admet des stratégies parfaites en dimension* 4.

Si (p_{-1}, p_1) *et* (q_{-1}, q_1) *sont les restrictions à* x_1 *et* x_2 *d'une stratégie pour le jeu d'anticommutation de valeur* $\geqslant 1 - \varepsilon$, *alors*

$$\left\| (p_1 - p_{-1})(q_1 - q_{-1}) + (q_1 - q_{-1})(p_1 - p_{-1}) \right\|_2^2 \leqslant 432\varepsilon.$$

Informellement, dans le jeu d'anticommutation, le but des joueurs est de convaincre l'arbitre qu'ils savent remplir un carré 3×3 avec des ± 1, de sorte que le produit des valeurs sur chaque ligne est 1 et sur chaque colonne est -1. Par un argument de parité, il s'agit bien sûr d'une tâche impossible, mais la force de l'intrication quantique (ou de la non-commutativité) est que les joueurs qui partagent suffisamment d'intrication peuvent convaincre l'arbitre qu'ils ont mené à bien cette tâche.

De manière plus précise, le jeu est défini de la façon suivante. L'ensemble des questions est $\mathscr{X} = C \cup L$, l'union disjointe d'un carré 3×3 noté $C = \{1,2,3\}^2$ et de l'ensemble L des lignes et colonnes qui le constituent. Les questions spécifiques sont $x_1 = (1,1) \in C$ et $x_2 = (2,2) \in C$. Définissons $\alpha(\ell) = 1$ pour chaque ligne ℓ et $\alpha(\ell) = -1$ pour chaque colonne ℓ. Pour $c \in C$, on définit $\mathscr{A}(c) = \{-1,1\}$, et pour une ligne ou colonne ℓ, on définit $\mathscr{A}(\ell) \subset \prod_{c \in \ell} \{-1,1\}$ par

$$\mathscr{A}(\ell) = \big\{ (b_c)_c \in \{-1,1\}^\ell \big| \prod_{c \in \ell} b_c = \alpha(\ell) \big\}.$$

La distribution μ est la distribution uniforme sur $\{(c,\ell) \mid c \in \ell\}$. Et $D(c,\ell,a,b) = 1_{a=b_c}$.

Preuve du lemme 2.3. Pour définir une stratégie parfaite en dimension 4, il suffit de construire, pour tout $c \in C$, un unitaire auto-adjoint $U_c \in \mathscr{U}(4)$ tel que, pour tout $\ell \in L$, les $(U_c)_{c \in \ell}$ commutent et $\prod_{c \in C} U_c = \alpha(\ell)$. En effet, les projections spectrales $p^c_{\pm 1}$ de U_c ($U_c = p^c_1 - p^c_{-1}$), et $p^\ell_b = \prod_{c \in \ell} p^c_{b_c}$ formeront alors une stratégie parfaite. Pour construire de tels U_c, on considère deux unitaires auto-adjoints de taille 2 σ^X, σ^Z tels que $\sigma^X \sigma^Z = -\sigma^Z \sigma^X$ [6] et on définit les $(U_c)_c$ comme dans la figure 1.

$\sigma^X \otimes 1$	$1 \otimes \sigma^X$	$\sigma^X \otimes \sigma^X$
$1 \otimes \sigma^Z$	$\sigma^Z \otimes 1$	$\sigma^Z \otimes \sigma^Z$
$-\sigma^X \otimes \sigma^Z$	$-\sigma^Z \otimes \sigma^X$	$\sigma^X \sigma^Z \otimes \sigma^Z \sigma^X$

Figure 1 – La stratégie parfaite du jeu d'anticommutation

Une rapide inspection de chaque ligne et chaque colonne permet de se convaincre des propriétés requises.

Soit (p^c_{-1}, p^c_1) (pour $c \in C$) et $(p^\ell_b)_{b \in \mathscr{A}(\ell)}$ (pour $\ell \in L$) une stratégie de valeur $\geqslant 1 - \varepsilon$, de sorte que $(p_{-1}, p_1) = (p^{1,1}_{-1}, p^{1,1}_1)$ et $(q_{-1}, q_1) = (p^{2,2}_{-1}, p^{2,2}_1)$. Pour tout $c \in C$, définissons $U^c = p^c_{-1} - p^c_1$. Pour tout $\ell \in L$ et tout $c \in \ell$, on pose

[6] Par exemple les matrices de Pauli $\sigma^X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ et $\sigma^Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$.

$U^\ell(c) = \sum_{b\in\mathscr{A}(\ell)} b_c p_b^\ell$. Soit

$$\eta_{c,\ell} = \|U^c - U^\ell(c)\|_2 = 2\sqrt{1 - \sum_{b\in\mathscr{A}(\ell)} \tau(p_{b_c}^c p_b^\ell)}.$$

L'hypothèse que la stratégie a valeur $\geqslant 1-\varepsilon$, implique

$$\frac{1}{18}\sum_\ell \sum_{c\in\ell} \eta_{c,\ell}^2 = \int \eta_{c,\ell}^2 d\mu(c,\ell) \leqslant 4\varepsilon.$$

Et donc, si pour $\ell \in L$ on pose $\eta_\ell = (\sum_{c\in\ell} \eta_{c,\ell}^2)^{\frac{1}{2}}$, on obtient

$$\sum_\ell \eta_\ell^2 \leqslant 24\varepsilon.$$

Les matrices U^c et U_c^ℓ sont toutes des unitaires auto-adjointes. Par définition de $\mathscr{A}(\ell)$, si c, c', c'' sont les trois points de ℓ, alors $U^\ell(c) = \alpha(\ell)U^\ell(c')U^\ell(c'')$. On en déduit que

$$\|U^c - \alpha(\ell)U^{c'}U^{c''}\|_2 \leqslant \eta_{c,\ell} + \eta_{c',\ell} + \eta_{c'',\ell} \leqslant \sqrt{3}\eta_\ell.$$

Dans ce qui suit, on dénote *hi* la *i*-ème ligne, et *vj* la *j*-ième colonne. On écrira aussi $M \simeq_\delta N$ si $\|M-N\|_2 \leqslant \sqrt{3}\delta$. On a donc

$$\begin{aligned} U^{11}U^{22} &\simeq_{\eta_{h1}+\eta_{v2}} -U^{13}U^{12}U^{12}U^{32} = -U^{13}U^{32} \\ &\simeq_{\eta_{v3}+\eta_{h3}} U^{23}U^{33}U^{33}U^{31} = U^{23}U^{31} \\ &\simeq_{\eta_{h2}+\eta_{v1}} -U^{22}U^{21}U^{21}U^{11} = -U^{22}U^{11}. \end{aligned}$$

On en déduit

$$\|U^{11}U^{22} + U^{22}U^{11}\|_2 \leqslant \sum_\ell \sqrt{3}\eta_\ell \leqslant \sqrt{18\sum_\ell \eta_\ell^2}.$$

Le lemme en découle, puisqu'on a déjà justifié que $\sum_\ell \eta_\ell^2 \leqslant 24\varepsilon$, et $18\cdot 24 = 432$. □

2.5. Stabilité

Les jeux de commutation et d'anticommutation ont des points communs, qu'on retrouvera souvent par la suite :

- ils ont un petit nombre de questions dont on se préoccupe vraiment ; les autres sont seulement là pour garantir de fortes conclusions sur les restrictions à ces questions particulières de bonnes stratégies.
- ils sont stables, dans le sens où une stratégie de valeur proche de 1 est nécessairement proche d'une stratégie parfaite.

Pour rendre la notion de stabilité précise, il faut savoir comparer des stratégies. Comme dans le cas des résultats de stabilité pour les représentations de groupes (De Chiffre, Ozawa et Thom, 2019 ; Gowers et Hatami, 2017), il est pertinent de s'autoriser à comparer des stratégies de dimensions différentes mais proches. [7]

Définition 2.4. On dit qu'une stratégie $\mathscr{S}' = (q_a^x)_{a\in\mathscr{A}}^{x\in\mathscr{X}} \subset (\mathscr{N}, \tau_{\mathscr{N}})$ pour $\mathscr{G}$ est ε-proche d'une autre stratégie $\mathscr{S} = (p_a^x)_{a\in\mathscr{A}}^{x\in\mathscr{X}} \subset (\mathscr{M}, \tau)$ si :

- ▷ Il existe une projection $e \in \mathscr{M}_\infty = \mathscr{M} \otimes B(\ell_2)$ de trace finie telle que $\mathscr{N} = e\mathscr{M}_\infty e$ avec trace $\tau_{\mathscr{N}} = \frac{1}{\tau\otimes\mathrm{Tr}(e)}\tau \otimes \mathrm{Tr}\,\big|_{e\mathscr{M}_\infty e}$.
- ▷ Il existe une isométrie partielle $w \in P\mathscr{M}_\infty(1_{\mathscr{M}} \otimes e_{1,1})$ telle que

$$\tau(1 - w^*w) \leqslant \varepsilon, \tau_{\mathscr{N}}(P - ww^*) \leqslant \varepsilon,$$

- ▷ $\mathbf{E}_x \sum_{a\in\mathscr{A}} \|P_a^x - w^*Q_a^x w\|_2^2 \leqslant \varepsilon$, où l'espérance est par rapport à $\frac{1}{2}(\mu_1 + \mu_2)$, la moyenne des deux marginales de μ.

On peut vérifier que si $\mathscr{S}$ et $\mathscr{S}'$ sont ε-proches, alors $\mathrm{val}(\mathscr{G}, \mathscr{S}) - \mathrm{val}(\mathscr{G}, \mathscr{S}') = O(\sqrt{\varepsilon})$.

Définition 2.5. Un jeu $\mathscr{G}$ est stable avec module $\delta\colon [0,1] \to [0,1]$ si toute stratégie pour $\mathscr{G}$ de valeur $\geqslant 1 - \varepsilon$ est $\delta(\varepsilon)$-proche d'une stratégie parfaite.

Une inspection de la preuve des lemmes 2.2 et 2.3 montre que les jeux de commutation et d'anticommutation sont stables avec module $\delta(\varepsilon) = O(\varepsilon)$.

2.6. Compression de jeu, un survol

L'idée principale introduite par Ji et al. (2020a), et l'ingrédient principal dans la preuve du théorème 2.1, est une procédure de *compression d'un jeu*, qui transforme un jeu compliqué à définir en un nouveau jeu, plus simple à définir, mais dont les bonnes stratégies sont elles-mêmes compliquées, et doivent nécessairement encoder de bonnes stratégies pour le jeu original. En un sens, la compression transfère donc de la complexité du côté de la définition d'un jeu vers celui de ses bonnes stratégies.

Pour rendre cette idée de compression plus précise, il faut expliquer comment la complexité d'un jeu est mesurée. La notion précise semblerait artificielle et nécessite un peu de travail pour être définie. Pour se faire une idée, il vaut peut-être mieux commencer par considérer un modèle (trop simpliste pour que ce qui suit puisse être correct), où la complexité naïve d'un jeu $\mathscr{G} = (\mathscr{X}, \mu, \mathscr{A}, D)$ est donnée par deux entiers $m = \lceil \log|\mathscr{X}| \rceil$, $n = \lceil \log|\mathscr{A}| \rceil$, le logarithme du nombre de questions et du nombre de réponses respectivement. Dans ce modèle trop simple, compresser un jeu,

[7] Il pourrait être préférable de parler de stabilité flexible comme dans le cas des groupes, mais c'est la seule notion de stabilité que l'on considèrera dans ce texte.

cela signifie donc réduire le nombre de questions et le nombre de réponses. Il y a pour cela trois étapes distinctes.

Étape 1 (introspection, ou diminution du nombre de questions) : Il s'agit, étant donné un jeu $\mathscr{G}^{(0)}$ de complexité naïve $(m^{(0)}, n^{(0)})$, de définir un nouveau jeu $\mathscr{G}^{(1)}$ de complexité naïve $(m^{(1)} = \mathrm{polylog}(m^{(0)}), n^{(1)} \leqslant m^{(0)} + n^{(0)})$, et qui vérifie la propriété essentielle

$$\mathrm{val}(\mathscr{G}^{(1)}, d) \geqslant 1 - \delta \implies d \geqslant (1-\varepsilon)e^{m^{(0)}} \text{ et } \mathrm{val}(\mathscr{G}^{(0)}, d) \geqslant 1 - \varepsilon$$

pour tout $\delta > 0$ et $d \in \mathbf{N}$, où $\varepsilon = C\delta$ pour une constante C.

Étape 2 (PCP, ou Diminution du nombre de réponses) : Il s'agit, étant donné un jeu $\mathscr{G}^{(1)}$ de complexité naïve $(m^{(1)}, n^{(1)})$, de définir un nouveau jeu $\mathscr{G}^{(2)}$ de complexité naïve $(m^{(2)}, n^{(2)})$ avec $\max(m^{(2)}, n^{(2)}) = \mathrm{poly}(m^{(1)}, \log n^{(1)}))$, et qui vérifie la propriété essentielle

$$\mathrm{val}(\mathscr{G}^{(2)}, d) \geqslant 1 - \gamma \implies \mathrm{val}(\mathscr{G}^{(1)}, d) \geqslant 1 - \delta$$

pour tout $\gamma > 0$ et $d \in \mathbf{N}$, où $\delta = \mathrm{poly}(m^{(1)}, \log n^{(1)})\gamma^c + o(1)$.

En appliquant successivement ces deux étapes, en partant d'un jeu $\mathscr{G}^{(0)}$ de complexité naïve $(n^{(0)}, m^{(0)}) \leqslant N$, on obtiendrait ainsi un jeu $\mathscr{G}^{(2)}$ de complexité naïve $(n^{(2)}, m^{(2)}) \leqslant \mathrm{poly}(\log N)$, et dont les valeurs satisfont

$$\mathrm{val}(\mathscr{G}^{(2)}, d) \geqslant 1 - \frac{1}{(\log N)^C} \implies d \geqslant \frac{1}{2}2^N \text{ et } \mathrm{val}(\mathscr{G}^{(1)}, d) \geqslant \frac{1}{2}.$$

C'est déjà bien car on a significativement diminué la complexité naïve, mais ce n'est pas suffisant, car le but est d'itérer cette procédure de compression, on voudrait donc avoir une implication du type

$$\mathrm{val}(\mathscr{G}^{(2)}, d) \geqslant \frac{1}{2} \implies d \geqslant \frac{1}{2}n^{(0)} \text{ et } \mathrm{val}(\mathscr{G}^{(0)}, d) \geqslant \frac{1}{2}.$$

C'est résolu par la troisième étape.

Étape 3 (Répétition parallèle) : Étant donné un jeu $\mathscr{G}^{(2)}$ de complexité naïve $(n^{(2)}, m^{(2)})$ et un paramètre $\gamma \in (0,1)$, il existe un nouveau jeu $\mathscr{G}^{(3)}$ de complexité naïve $(m^{(3)} = km^{(2)}, n^{(3)} = kn^{(2)})$ avec $k = \mathrm{poly}(m^{(2)}, \frac{1}{\gamma})$, et qui vérifie la propriété

$$\mathrm{val}(\mathscr{G}^{(3)}, d) \geqslant \frac{1}{2} \implies \mathrm{val}(\mathscr{G}^{(2)}, d) \geqslant 1 - \gamma$$

pour tout $d \in \mathbf{N}$.

De ces trois étapes, seule la troisième est valide pour la notion de complexité naïve. Pour les deux premières, il faudra des notions de complexité plus fines pour que les énoncés deviennent corrects. En particulier, pour la deuxième étape, qui repose sur

des variantes du théorème PCP (pour preuve vérifiable de manière probabiliste), la complexité qui entre en jeu sera une forme de complexité algorithmique, qui mesure le temps nécessaire à une machine de Turing pour calculer $D(x, y, a, b)$.

Les preuves de ces trois étapes sont toutes difficiles. La première est celle qui est la plus innovante. Les deux autres s'inspirent de résultats très importants mais antérieurs en informatique théorique : le théorème PCP de Arora, Lund et al. (1998) et Arora et Safra (1998) et le théorème de répétition parallèle de Raz (1998). Le lecteur pourra se plonger dans les exposés de Chazelle (2003) et Pansu (2013) pour en savoir plus sur ces théorèmes et leur pertinence. Pour un mathématicien ignorant de complexité algorithmique comme moi, l'étape la plus difficile est très nettement la deuxième. Et si j'ai lu sa preuve suffisamment longtemps et attentivement pour me convaincre qu'elle est correcte, je ne pense pas avoir vraiment compris ce qui s'y passe. Ce serait très satisfaisant si on pouvait trouver une autre preuve du théorème 1.2 qui reste du côté de la calculabilité et ne fait pas intervenir de notions de complexité algorithmique.

Dans la suite de ce texte, je présenterai rapidement ces trois étapes, et enfin j'expliquerai un peu plus en détails comment, une fois énoncées avec les notions correctes de complexité, elles permettent de prouver le théorème 2.1.

3. Introspection

Dans la suite de l'exposé, on notera $\mathbf{F}_2$ le corps fini à 2 éléments.

La première étape, celle d'introspection, a pour but de transformer un jeu $\mathscr{G}^{(0)}$ en un jeu $\mathscr{G}^{(1)}$ en réduisant de manière exponentielle le nombre de questions, sans trop augmenter le nombre de réponses.

On ne sait mener à bien cette étape d'introspection que pour des jeux dans la distribution de question $\mu^{(0)}$ est très particulière. C'est un problème ouvert intéressant, posé par Ji et al. (2020a), que d'étendre cette procédure d'introspection à des distributions plus générales.

Initialement, les distributions admissibles étaient ce que les auteurs appelaient conditionnellement linéaires, mais une condition plus générale et un peu plus facile à définir est suffisante : pour un entier N, il existe deux partitions $(E_x)_{x\in\mathscr{X}^{(0)}}$ et $(F_x)_{x\in\mathscr{X}^{(0)}}$ de $\mathbf{F}_2^N$ en sous-espaces affines tels que $\mu^{(0)}(x, y) = 2^{-N}|E_x \cap F_y|$. On dira donc qu'un jeu est de complexité de questions N si la distribution de questions est de cette forme. Le nombre de questions $m^{(0)}$ de $\mathscr{G}^{(0)}$ est alors au plus 2^N ; le paramètre N joue donc le rôle de $m^{(0)}$ dans la section précédente.

De manière un peu plus concrète, en suivant le principe général, le jeu $\mathscr{G}^{(1)}$ aura deux questions centrales « Introspecte-toi ! (1) » et « Introspecte-toi ! (2) », dont la réponse attendue est une paire $(x, a) \in \mathscr{X}^{(0)} \times \mathscr{A}^{(0)}$. Le reste des questions est là pour s'arranger qu'une bonne stratégie pour ce jeu est nécessairement, pour ces questions,

de la forme $q_x^1 \otimes p_a^x$ et $q_x^2 \otimes p_a^x$ pour des partitions de l'unité $(q_x^1)_{x\in\mathscr{X}}$ et $(q_x^2)_{x\in\mathscr{X}}$ vérifiant $\tau(q_x^1 q_y^2) = \mu(x,y)$, et une bonne stratégie p pour $\mathscr{G}^{(0)}$. On peut aussi reformuler ce qui précède dans le langage ludique : en posant la question « Introspecte-toi ! (1) », l'arbitre demande au joueur de générer pour lui-même une paire de questions (x,y) selon la distribution μ, de ne retenir que la première question x, d'y répondre honnêtement a et de lui renvoyer la paire (x,a). Symétriquement pour le deuxième joueur. Une stratégie de la forme $q_x^1 \otimes p_a^x$ et $q_x^2 \otimes p_a^x$ est ce qu'on appelle une stratégie honnête ; toutes les autres questions sont là pour s'arranger pour que les joueurs n'ont pas la possibilité de tricher et n'ont d'autre choix que de suivre une stratégie qui est proche d'une stratégie honnête.

Pour obtenir cela, on utilisera un résultat de stabilité pour les groupes de Heisenberg sur le corps à deux éléments $\mathbf{F}_2$,

$$H_{2N+1} = \left\{ \begin{pmatrix} 1 & a & c \\ 0 & 1_N & b \\ 0 & 0 & 1 \end{pmatrix} \right\} \subset \mathrm{GL}_{N+2}(\mathbf{F}_2).$$

Le groupe H_{2N+1} est une extension centrale de $\mathbf{F}_2^{2N}$ par $\mathbf{F}_2$. Sa théorie des représentations est très singulière. Il a 2^{2N} représentations irréductibles de dimension 1 : celles qui sont triviales sur le centre et proviennent donc de représentations du groupe abélien $\mathbf{F}_2^{2N}$. Il a une seule représentation irréductible qui n'est pas triviale sur le centre. Elle est de dimension 2^N et le centre agit par $\{\mathrm{id}, -\mathrm{id}\}$. Elle peut aussi être décrite par deux représentations unitaires $\sigma^X, \sigma^Z \colon \mathbf{F}_2^N \to \mathscr{U}(2^N)$ qui vérifient

$$\sigma^X(a)\sigma^Z(b) = (-1)^{\sum_i a_i b_i}\sigma^Z(b)\sigma^X(a).$$

Autrement dit, les unitaires $\sigma^X(a)$ et $\sigma^Z(b)$ commutent ou anti-commutent selon la valeur de $\langle a,b\rangle := \sum_i a_i b_i \in \mathbf{F}_2$. Comme la notation l'indique, ces représentations peuvent être réalisées à l'aide des matrices de Pauli

$$\sigma^X(a) = \otimes_{i=1}^n (\sigma^X)^{a_i}, \quad \sigma^Z(b) = \otimes_{i=1}^n (\sigma^Z)^{b_i}.$$

Notons $(\tau_a^X)_{a\in\widehat{\mathbf{F}_2^N}}$ et $(\tau_a^Z)_{a\in\widehat{\mathbf{F}_2^N}}$ les partitions de l'unité correspondant, via la transformée de Fourier, aux représentations σ^X et σ^Z.

En combinant les jeux de commutation ou d'anticommutation, on peut produire un jeu qui prend en compte cette représentation, et de manière économique en termes de questions.

Théorème 3.1. *Pout tout entier N, il existe un jeu $\mathscr{G}_N$ de complexité de questions $O(\log N)$, avec $|\mathscr{A}| = 2^N$. Il a deux questions particulières $X, Z \in \mathscr{X}_N$ vérifiant $\mu_N(X) = \mu_N(Z) = \frac{1}{4}$, et $\mathscr{A}(x) = \mathscr{A}(z) = \mathbf{F}_2^N$. Ce jeu est stable avec module $\varepsilon \mapsto O(\varepsilon)$. De plus, toute stratégie parfaite est sur une algèbre de la forme $(M_{2^N}(\mathbf{C}) \otimes \mathscr{N}, \mathrm{tr} \otimes \tau')$ où $p_a^X = \tau_a^X \otimes 1_{\mathscr{N}}$ et $p_b^Z = \tau_b^Z \otimes 1_{\mathscr{N}}$ pour tout $a, b \in \mathbf{F}_2^N$.*

En particulier, une stratégie de dimension finie et de valeur $\geqslant 1-\varepsilon$ doit être en dimension $\geqslant (1-O(\varepsilon))2^N$ (et même en dimension $O(\varepsilon)$-proche d'un multiple de 2^N).

Toute la force de l'étape d'introspection est contenue dans ce résultat : il y a besoin d'un nombre polynomial de questions, la dimension nécessaire des bonnes stratégies est exponentielle, la stabilité est de module linéaire.

Une autre propriété importante pour l'utilisation de ce résultat pour l'introspection est qu'en restriction à la question X, une bonne stratégie pour $\mathscr{G}$ contient en particulier un générateur de variable aléatoire presque uniforme dans $\mathbf{F}_2^N$ (simplement parce que $\mathrm{tr}(\tau_a^X) = 2^{-N}$ pour tout $a \in \mathbf{F}_2^N$). Il est donc possible pour les joueurs d'utiliser cette source d'aléa d'origine quantique pour engendrer les variables aléatoires de loi $\mu^{(0)}$, et on peut définir un jeu qui force les joueurs à le faire. Pour cela, nous allons voir qu'il suffit d'ajouter un petit nombre (6) de questions à $\mathscr{G}_N$ pour obtenir un nouveau jeu qui permet de mener à bien l'étape d'introspection.

Proposition 3.2. *Étant données deux partitions $\underline{E} = (E_x)_{x\in\mathscr{X}^{(0)}}$ et $\underline{F} = (F_x)_{x\in\mathscr{X}^{(0)}}$ de $\mathbf{F}_2^N$ en sous-espaces affines et un ensemble $\mathscr{A}$, il existe un jeu $\mathscr{G}$ de complexité de questions $O(\log N)$ et avec $\leqslant 2^N|\mathscr{A}|$ réponses qui est stable avec module $\varepsilon \mapsto O(\varepsilon)$. De plus, il a deux questions particulières $I1, I2$ telles que $\mu(I1) = \mu(I2) \geqslant c$ (pour un c indépendant de N), dont les réponses possibles sont $\mathscr{X}\times\mathscr{A}$, et telle que toute stratégie parfaite est, en restriction à $I1$, $I2$, de la forme $(\sum_{v\in E_x}\tau_v^X)\otimes p_a^x$, et $(\sum_{v\in F_x}\tau_v^X)\otimes p_a^x$ pour des partitions de l'unité $(p_a^x)_{a\in\mathscr{A}}$ pour tout $x\in\mathscr{X}^{(0)}$.*

Les partitions $q_x^1 = \sum_{v\in E_x}\tau_v^X$ et $q_x^2 = \sum_{v\in F_x}\tau_v^X$ vérifient bien $\mathrm{tr}(q_x^1 q_y^2) = 2^{-N}|E_x\cap F_y|$. Si on partait d'un jeu $\mathscr{G}^{(0)}$ dont la distribution de questions est donnée par les partitions $\underline{E}$ et $\underline{F}$, en ajoutant au jeu obtenu dans la proposition 3.2 la paire de questions $(I1, I2)$ posée avec probabilité $\geqslant c$, avec fonction de décision $D(I1, I2, (x,a), (y,b)) = D^{(0)}(x,y,a,b)$, on obtient bien un jeu qui vérifie, pour notre notion un peu plus précise de la complexité des questions, la conclusion de la partie introspection de la partie 2.6.

La raison pour laquelle on peut traiter des partitions en sous-espaces affines (mais pas en parties arbitraires) est donnée par la direction « si » (respectivement « seulement si ») du lemme suivant. Dans ce lemme, pour une partie $E \subset \mathbf{F}_2^N$, on note $E_0 = \{a \in \mathbf{F}_2^N \mid a + E = E\}$ le plus grand sous-espace vectoriel tel que E est une union d'espaces affines parallèles à E_0. Par exemple, si E est un sous-espace affine, E_0 est sa partie linéaire. On notera $\cdot^\perp$ l'orthogonal pour la forme $\langle a, b\rangle = \sum_i a_i b_i$.

Lemme 3.3. *Étant données deux parties $E, F \subset \mathbf{F}_2^N$, les projections $\sum_{v\in E}\tau_v^X$ et $\sum_{v\in F}\tau_v^Z = 0$ commutent si et seulement si $E_0^\perp \subset F_0$.*

Pour expliquer comment cette propriété de commutation intervient, donnons la description explicite d'un jeu $\mathscr{G}(\underline{E})$ qui ne dépend que d'une partition, avec une seule

question d'introspection I, et qui vérifie l'analogue de la conclusion de la proposition 3.2 : en restriction à I ses stratégies parfaites sont de la forme $(\sum_{v\in E_x}\tau_v^X)\otimes p_a^{1,x}$ pour une partition de l'unité $(p_a^x)_{a\in\mathscr{A}}$ pour tout $x\in\mathscr{X}^{(0)}$. Une petite modification permet d'obtenir la proposition 3.2.

L'ensemble des questions de $\mathscr{G}(\underline{E})$ est $\mathscr{X}_N\cup\{E,I,L\}$ (pour Échantillonne, Introspecte et Lis). La mesure μ est $\frac{1}{2}(\mu_N+\mu')$ où μ' est la mesure uniforme sur l'ensemble des quatre arêtes de la figure 2. [8]

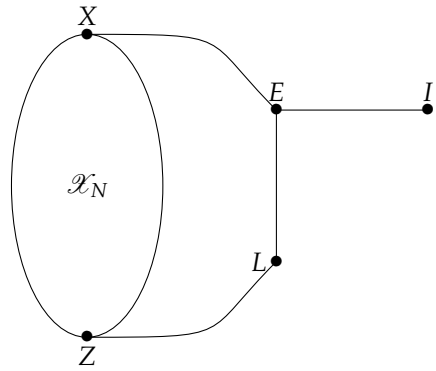


Figure 2 – Les questions du jeu $\mathscr{G}(\underline{E})$

L'ensemble des réponses possibles pour chaque question est

- $\mathscr{A}(x)=\mathscr{A}_N(x)$ si $x\in\mathscr{X}_N$,
- $\mathscr{A}(E)=\{(w,a)\mid w\in\mathbf{F}_2^N, a\in\mathscr{A}\}$,
- $\mathscr{A}(I)=\{(x,b)\mid x\in\mathscr{X}^{(0)}, b\in\mathscr{A}\}$,
- $\mathscr{A}(L)=\{(y,\varphi,c)\mid x\in\mathscr{X}^{(0)},\varphi\in(E_x)_0^*, a\in\mathscr{A}\}$.

La fonction de décision est

- $D=D_N$ en restriction à $\mathscr{G}_N$,
- $D(X,E,w,(v,a))=1_{v=w}$,
- $D(E,I,(v,a),(x,b))=1_{a=b}1_{v\in E_x}$,
- $D(E,L,(v,a),(y,\varphi,c))=1_{a=c}1_{v\in E_y}$,
- $D(L,Z,(y,\varphi,c),w)=1_{\varphi(v)=\langle v,w\rangle\forall v\in(E_y)_0}$.

Voyons comment construire des stratégies parfaites pour le jeu $\mathscr{G}(\underline{E})$. Commençons par une stratégie parfaite q pour $\mathscr{G}_N$; par le théorème 3.1 elle est sur une algèbre de la

[8] Le lecteur attentif remarquera que cette mesure n'est pas de la forme discutée précédemment, mais une petite modification qui n'affecte pas ce qui suit permet de s'y ramener, c'est le contenu de Ji et al., 2020a, §6.3.

forme $M_{2^N}(\mathbf{C}) \otimes \mathscr{N}$ et en restriction à X, Z elle est de la forme $\tau_a^X \otimes 1_{\mathscr{N}}$ et $\tau_a^Z \otimes 1_{\mathscr{N}}$. Alors, pour toute famille $\{(p_a^x)_{a\in\mathscr{A}} \mid x \in \mathscr{X}^{(0)}\}$ de partitions de l'unité dans $(\mathscr{N}, \tau)$, on peut étendre cette stratégie parfaite de $\mathscr{G}_N$ en une stratégie parfaite de $\mathscr{G}(\underline{E})$, en posant

- ▷ $q^E_{w,a} = \tau^X_w \otimes p^x_a$ où $x \in \mathscr{X}^{(0)}$ est caractérisé par $w \in E_x$,
- ▷ $q^I_{x,a} = (\sum_{x\in E_x} \tau^X_w) \otimes p^x_a$,
- ▷ $q^L_{y,\varphi,c} = (\sum_{x\in E_y} \tau^X_w)(\sum_{w|\langle v,w\rangle=\varphi(v)\forall v\in E_{y,0}} \tau^Z_w) \otimes p^x_c$.

Le seul point qui mérite une justification est le fait que q^L est bien constitué de projections, c'est-à-dire que les deux projections $\sum_{x\in E_y} \tau^X_w$ et $\sum_{w|\langle v,w\rangle=\varphi(v)\forall v\in E_{y,0}} \tau^Z_w$ commutent. C'est justifié par le petit résultat d'algèbre linéaire du lemme 3.3. Il est alors immédiat de vérifier que cette stratégie est parfaite pour $\mathscr{G}(\underline{E})$.

Réciproquement, il n'est pas bien dur (on n'a ajouté que 3 questions) de déduire de la stabilité de $\mathscr{G}_N$ la stabilité de $\mathscr{G}(\underline{E})$.

4. PCP

La deuxième étape, celle de vérification probabiliste de preuve, a pour but de transformer un jeu $\mathscr{G}^{(1)}$ en un jeu $\mathscr{G}^{(2)}$ en réduisant le nombre de réponses, sans trop augmenter le nombre de questions.

De manière concrète, certaines des questions du jeu seront « Si la paire de questions posée était (x, y), donne-moi une preuve courte que tu es capable de produire (a, b) tel que $D(x, y, a, b) = 1$. » Les autres questions sont là pour garantir que les joueurs n'ont d'autre choix que de jouer honnêtement. Pour cela, l'argument repose de manière essentielle sur des idées qui sont devenues classiques en informatique théorique : le théorème PCP (pour preuve vérifiable de manière probabiliste), qui de manière informelle affirme qu'un problème de décision NP peut être vérifié de façon probabiliste en ayant accès à un nombre constant de bits de la preuve et en utilisant un nombre logarithmique de bits aléatoires. Ici le problème de décision est « déterminer s'il existe (a, b) tel que $D(x, y, a, b) = 1$ », il est donc naturel que la complexité d'un jeu soit mesurée par la complexité algorithmique de la fonction D plutôt que par le nombre de réponses.

Pour définir de manière plus précise cette notion de complexité, il vaut mieux changer légèrement la définition d'un jeu. Un jeu devient donc la donnée de $(\mathscr{X}, \mu, \mathscr{A}, \mathscr{D})$ où $\mathscr{X}, \mathscr{A}$ sont des ensembles de la forme $\mathbf{F}_2^k, \mathbf{F}_2^\ell$, μ est une distribution de la forme considérée dans la section d'introspection, et $\mathscr{D}$ est une machine de Turing à 4 entrées. Cela donne lieu à un jeu dans le sens précédent si $\mathscr{D}$ a la propriété que $\mathscr{D}$ termine quand elle prend pour argument $(x, y, a, b) \in \mathscr{X} \times \mathscr{X} \times \mathscr{A} \times \mathscr{A}$, et renvoie un élément de $\{0, 1\}$. La complexité de la fonction de décision est mesurée par deux paramètres : la taille de la description de $\mathscr{D}$, et le temps maximum d'exécution de

$\mathscr{D}$ lorsque x, y, a, b varient. Pour prendre en compte l'étape d'introspection où la distribution des questions devient encodée dans la fonction de décision, il faudra aussi mesurer la complexité algorithmique nécessaire pour décrire la mesure μ, mais on ignorera ce point pour simplifier.

Pour mettre en place les idées du théorème PCP classique dans le contexte de jeux et de leurs stratégies quantiques, cette deuxième étape repose sur un autre énoncé particulièrement délicat de stabilité quantitative pour certains jeux très particuliers, obtenu dans Ji et al. (2020b, 2022) pour palier à l'erreur critique qu'ils avaient trouvée dans la preuve de Vidick (2016, 2020). L'énoncé précis est très technique et long à énoncer, il ne sera pas reproduit ici.

Dans Ji et al. (2020a), ce théorème de stabilité était aussi utilisé pour obtenir une forme du théorème 3.1 un peu plus faible mais suffisante. L'amélioration et la simplification présentée dans le théorème 3.1, qui a été obtenue dans de la Salle (2022), repose sur des arguments plus élémentaires de graphes expanseurs et de stabilité pour les groupes.

5. Répétition parallèle

Étant donné un jeu $\mathscr{G} = (\mathscr{X}, \mathscr{A}, \mu, D)$ et un entier $n \geqslant 1$, la répétition en parallèle n fois de $\mathscr{G}$ est le jeu

$$\mathscr{G}^n = (\mathscr{X}^n, \mathscr{A}^n, \mu^{\otimes n}, D^{\otimes n})$$

où

$$D^{\otimes n}\big((x_i)_{i=1}^n, (y_i)_{i=1}^n, (a_i)_{i=1}^n, (b_i)_{i=1}^n\big) = \prod_{i=1}^n D(x_i, y_i, a_i, b_i).$$

Le premier réflexe naïf est de s'attendre à ce que la valeur de $\mathscr{G}^n$ est la puissance n-ième de la valeur de $\mathscr{G}$. C'est faux : il existe un jeu très simple dont la valeur classique (au sens de la sous-section 2.1) et la valeur classique de $\mathscr{G}^2$ sont toutes les deux égales à $\frac{1}{2}$ (Aubrun, 2021, Example 1). Le théorème de répétition parallèle de Raz (1998) affirme que, pour tout jeu de valeur < 1, la valeur classique de $\mathscr{G}^n$ décroît exponentiellement avec n, à un taux qui ne dépend que du nombre de réponses et de la valeur classique. Plus précisément, il affirme que si $\mathscr{G}$ a valeur classique $\leqslant 1 - \varepsilon$, alors $\mathscr{G}^n$ a valeur classique $\leqslant C\exp(-C\varepsilon^{32} n / \log \mathscr{A})$ pour une constante C explicite. Il y a maintenant de nombreuses preuves de cet énoncé, où la constante 32 peut être abaissée à 3. J'apprécie particulièrement la présentation de Aubrun (2021).

La question de savoir si le théorème de répétition parallèle est vrai pour les différentes valeurs quantiques d'un jeu reste un problème ouvert. Une forme a été obtenue par Bavarian, Vidick et Yuen (2022), suffisante pour ce qui suit. Il s'applique à certains types de jeu, appelés des jeux ancrés.

Étant donné un jeu $\mathscr{G} = (\mathscr{X}, \mathscr{A}, \mu, D)$, on peut définir un nouveau jeu $\widetilde{\mathscr{G}} = (\widetilde{\mathscr{X}}, \widetilde{\mathscr{A}}, \widetilde{\mu}, \widetilde{D})$, dit *ancré*, de la manière suivante : $\widetilde{\mathscr{X}} = \mathscr{X} \cup \{\perp\}$, $\widetilde{\mathscr{A}} = \mathscr{A}$,

$$\widetilde{\mu} = \frac{1}{4} \sum_{x,y} \mu(x,y)(\delta_{(x,y)} + \delta_{(\perp,y)} + \delta_{(x,\perp)} + \delta_{(\perp,\perp)})$$

et

$$\widetilde{D}(x,y,a,b) = \begin{cases} D(x,y,a,b) & \text{si } x,y \in \mathscr{X} \\ 1 & \text{si } x = \perp \text{ ou } y = \perp . \end{cases}$$

Autrement dit, on a ajouté une question $\perp$ qui est toujours gagnante, et pour engendrer une paire de questions pour $\widetilde{\mathscr{G}}$, l'arbitre commence par tirer au hasard une paire de questions (x,y) pour $\mathscr{G}$, et indépendamment avec probabilité $\frac{1}{2}$, remplace chaque question par la question gagnante.

Le théorème de répétition parallèle quantique qui suit ne s'applique que pour des jeux qu'on appellera paresseux (en référence aux marches aléatoires paresseuses) si $\mu(x,x) \geqslant \sum_{y \neq x} \frac{1}{2}(\mu(x,y) + \mu(y,x))$ pour tout $x \in \mathscr{X}$. Si $\mathscr{G}$ est un jeu arbitraire, le nouveau jeu ou l'on remplace μ par $\frac{1}{2}(\mu + \sum_{x,y} \frac{1}{2}(\mu(x,y) + \mu(y,x))\delta_{(x,x)})$ est un jeu paresseux.

Théorème 5.1. *Il existe une constante C telle que pour tout jeu paresseux $\mathscr{G} = (\mathscr{X}, \mathscr{A}, \mu, D)$ et tout entier d,*

$$\mathrm{val}((\widetilde{\mathscr{G}})^n, d) \leqslant C \exp\Big(-\frac{(1 - \mathrm{val}(\mathscr{G}, d))^C}{C \log |\mathscr{A}|} n\Big).$$

Démonstration. Ce théorème est le seul pour lequel la notion plus restrictive de stratégie que l'on considère rend les choses un tout petit peu plus difficiles. En effet, on peut définir la valeur asynchrone et de dimension d d'un jeu $\mathscr{G}$, notée $\mathrm{val}_{\mathrm{async}}(\mathscr{G}, d)$, comme le maximum de la quantité (1) où $\mathscr{H} = \mathbf{C}^{d^2}$. Alors le travail très difficile de Bavarian, Vidick et Yuen (2022) démontre exactement ce théorème pour la valeur asynchrone, sans l'hypothèse que le jeu soit paresseux. Le théorème pour la valeur (synchrone) considérée dans ce texte découle du résultat beaucoup plus facile par Vidick (2022a), qui affirme que, si $\mathscr{G}$ est un jeu paresseux, alors

$$\mathrm{val}_{\mathrm{async}}(\mathscr{G}, d) \geqslant 1 - \varepsilon \implies \mathrm{val}(\mathscr{G}, d) \geqslant 1 - K\varepsilon^{\frac{1}{K}}$$

pour tout entier d et tout $\varepsilon \geqslant 0$, où K est une constante universelle. Il est peut-être important de noter que c'est pour cette dernière inégalité que j'ai choisi d'autoriser des stratégies à valeurs $M_k(\mathbf{C})$ pour $k \leqslant d$ (et pas seulement dans $M_d(\mathbf{C})$) dans la définition de $\mathrm{val}(\mathscr{G}, d)$. □

6. Le théorème de compression

Il est maintenant temps d'énoncer une forme précise du théorème de compression, qui peut être obtenu en combinant avec soin les trois étapes qui ont été superficiellement évoquées ci-dessus. Pour son utilisation ultérieure, le théorème de compression sera donné non pas pour des jeux individuels, mais pour des familles de jeux qui sont données de manière uniforme par une machine de Turing.

Une propriété des jeux construits dans l'étape d'introspection est que la distribution de questions ne dépend que de l'entier N, et pas de la distribution des questions $\mu^{(0)}$ du jeu $\mathscr{G}^{(0)}$. En exploitant cette propriété, on peut donc mettre en place un théorème de compression où la suite des distributions de questions est une suite fixée de plus en plus complexe, et les réponses possibles sont également fixées. Autrement dit, on construit une suite explicite $(\mathscr{X}_n, \mu_n, \mathscr{A}_n)$ (des jeux dont il manque la fonction de décision) où la mesure μ_n est donnée comme dans la section 3 par deux partitions de $\mathbf{F}_2^{N_n}$, pour une suite bien choisie d'entiers N_n qui tend vers l'infini. [9]

On considère alors l'ensemble $\mathscr{E}$ des machines de Turing à 2 entrées, qu'on appellera les décideurs. Un décideur $\mathscr{D}$ de $\mathscr{E}$ sera valable si pour tout n, tout $z = (x, y, a, b) \in \mathscr{X}_n \times \mathscr{X}_n \times \mathscr{A}_n \times \mathscr{A}_n$, $\mathscr{D}(n, z)$ termine et renvoie un élément de $\{0, 1\}$. On notera $\mathrm{Temps}_n(\mathscr{D})$ le maximum pour tout z du temps de calcul de $\mathscr{D}(n, z)$. Un décideur valable permet donc de définir une suite de jeux $\mathscr{G}_n(\mathscr{D}) = (\mathscr{X}_n, \mu_n, \mathscr{A}_n, \mathscr{D}(n, \cdot))$.

Théorème 6.1 (Théorème de compression). *Il existe une machine de Turing* $\mathrm{COMPRESS} : \mathscr{E} \to \mathscr{E}$ *de complexité polynomiale et une autre machine de Turing* $f : \mathscr{E} \to \mathbf{N}$ *telle que, pour tout* $\mathscr{D} \in \mathscr{E}$, $\mathscr{D}' = \mathrm{COMPRESS}(\mathscr{D})$ *vérifie les propriétés suivantes :*

- $\mathscr{D}'$ *est valable et de complexité* $\mathrm{Temps}_n(\mathscr{D}') \leqslant \mathrm{poly}(N_n)$ *pour tout* n.
- *si* $\mathscr{D}$ *est valable alors pour tout* $n \geqslant f(\mathscr{D})$ *tel que* $\mathrm{Temps}_n(\mathscr{D}) \leqslant N_n^n$, *on a les deux implications suivantes :*
  1. *si* $\mathscr{G}_n(\mathscr{D})$ *a une stratégie parfaite commutative* [10] *de dimension finie, alors* $\mathscr{G}_{n-1}(\mathscr{D}')$ *aussi.*
  2. *si* $\mathrm{val}(\mathscr{G}_{n-1}(\mathscr{D}'), d) > \frac{1}{2}$, *alors* $d \geqslant N_n$ *et* $\mathrm{val}(\mathscr{G}_n(\mathscr{D}), d) > \frac{1}{2}$.

Expliquons comment le théorème de compression permet de déduire le théorème principal, le théorème 2.1. On définit une nouvelle machine de Turing F à 4 entrées (interprétées comme (R, M, n, z) avec R une machine de Turing à 4 entrées, M une machine de Turing à 0 entrée, n un entier et $z \in \mathscr{X}_n \times \mathscr{X}_n \times \mathscr{A}_n \times \mathscr{A}_n$) de la façon suivante :

[9] Informellement, la compression transforme un jeu de complexité N en un jeu de complexité $\mathrm{polylog} N$. Il n'est donc pas surprenant qu'un choix possible est de prendre pour N_n qui croît presque comme une tour d'exponentielles, mais avec $\log N_{n+1} = o(N_n^\varepsilon)$ pour tout $\varepsilon > 0$. Par exemple, $N_n = a_n^n$ où $a_1 = 1$ et $a_{n+1} = 2^{a_n}$.

[10] Une stratégie p pour un jeu $\mathscr{G} = (\mathscr{X}, \mu, \mathscr{A}, D)$ est dite commutative si pour tout (x, y) dans le support de μ, et tout $a, b \in \mathscr{A}$, $[p_a^x, p_b^y] = 0$

```
1 Exécute M pendant n étapes ;
2 Si l'exécution s'est arrêtée, retourne F(R,M,n,z)=1 ;
3 Sinon, continue ;
4 Définis un décideur D(n',z')=R(R,M,n',z') ;
5 Calcule D'=COMPRESS(D) ;
6 Retourne D'(n,z).
```

Les étapes 4 et 5 sont des instructions de haut niveau, du type *exécute telle machine de Turing dont le code a été donné en argument, ou bien dont le code a été calculé précédemment, avec telle autre entrée* ; une façon de les rendre précises est d'utiliser des machines de Turing universelles, qui en entrée une paire (M, x) où M est une machine de Turing et x une entrée possible de M, retourne le résultat de $M(x)$ si le calcul de $M(x)$ s'arrête, et tourne indéfiniment sinon. Un point important (mais apparemment absent de la littérature sur les machines de Turing) est qu'il est possible de définir une telle machine de Turing universelle de sorte que son temps de calcul en l'entrée (M, x) est au plus polynomial en $|M|, |x|$ et en le temps de calcul de $M(x)$. On déduit de tout cela, du fait que COMPRESS est de complexité polynomiale et du fait que $\mathrm{Temps}_n(\mathscr{D}') \leqslant \mathrm{poly}(N_n)$ dans le théorème 6.1, que le calcul de $F(R, M, n, z)$ termine toujours en temps $\leqslant \mathrm{poly}(n, |R|, |M|, N_n)$ et renvoie toujours 0 ou 1. En particulier, ce temps d'exécution est $\leqslant N_n^n$ pour tout n assez grand, calculable en termes de $|R|$ et $|M|$.

La fonction calculable du théorème 2.1 est alors définie de la manière suivante. Étant donnée une machine de Turing M à 0 entrée :

```
7 Définis un décideur D(M) par D(M)(n,z) = F(F,M,n,z) ;
8 Calcule N >= f(D(M)) tel que Temps_n(D(M))< N_n^n pour tout n>=N ;
9 Retourne le jeu G_N(D(M)).
```

Là encore, l'étape 7 peut être rendue précise en utilisant une machine de Turing universelle. Le fait que l'étape 8 est toujours faisable découle de la discussion qui suit la description du programme F.

Vérifions la conclusion du théorème 2.1. Soit $n_0 \in \mathbf{N}^* \cup \{\infty\}$ le temps d'arrêt de M. L'observation cruciale est que, lorsque dans la ligne 7 on appelle la fonction $F(F, M, n, z)$, pour un $n < n_0$, le décideur $\mathscr{D}$ qui est calculé dans la ligne 4 du code de F est le décideur $\mathscr{D}(M)$ lui-même. Et donc, si on note $\mathscr{D}'(M) = \mathrm{COMPRESS}(\mathscr{D}(M))$, alors $\mathscr{G}_n(\mathscr{D}(M)) = \mathscr{G}_n(\mathscr{D}'(M))$ pour tout $n < n_0$. En appliquant le théorème 6.1 on obtient donc que, si N est l'entier calculé à l'étape 8, on a pour tout entier n tel que $N \leqslant n < n_0$,

(1) si $\mathscr{G}_{n+1}(\mathscr{D}(M))$ a une stratégie parfaite commutative de dimension finie, alors $\mathscr{G}_n(\mathscr{D}(M))$ aussi.

(2) si $\mathrm{val}(\mathscr{G}_n(\mathscr{D}(M)), d) > \frac{1}{2}$, alors $d \geqslant N_{n+1}$ et $\mathrm{val}(\mathscr{G}_{n+1}(\mathscr{D}(M)), d) > \frac{1}{2}$.

Supposons tout d'abord que M s'arrête, c'est-à-dire $n_0 < \infty$. Il s'agit de montrer que $\mathscr{G}_N(D(M))$ a une stratégie parfaite commutative de dimension finie. Alors par définition de la machine de Turing F, pour tout $n \geqslant n_0$, le décideur $\mathscr{D}(M)$ est trivial dans le sens ou $\mathscr{D}(M)(n,z) = 1$ pour tout z. En particulier, le jeu $\mathscr{G}_n(\mathscr{D}(M))$ a une stratégie parfaite commutative de dimension 1 pour tout $n \geqslant n_0$. Si $n_0 \leqslant N$ on a donc fini. Sinon, par le point (1), on obtient que le jeu $\mathscr{G}_{n_0-1}(\mathscr{D}(M))$ a une stratégie parfaite commutative, et donc aussi le jeu $\mathscr{G}_{n_0-2}(D(M))$, etc. Par récurrence on en déduit que c'est le cas du jeu $\mathscr{G}_n(D(M))$ pour tout $n \geqslant N$, et en particulier pour $n = N$.

Supposons maintenant que M ne s'arrête pas, c'est-à-dire $n_0 = \infty$. Supposons par l'absurde que $\mathrm{val}(\mathscr{G}_N(D(M)), <\infty) > \frac{1}{2}$. Il existe un entier d tel que $\mathrm{val}(\mathscr{G}_N(D(M)), d) > \frac{1}{2}$. Par le point (2), on en déduit que $d \geqslant N_{N+1}$ et $\mathrm{val}(\mathscr{G}_{N+1}(D(M)), d) > \frac{1}{2}$. Par récurrence, on en déduit que $d \geqslant N_n$ et $\mathrm{val}(\mathscr{G}_n(D(M)), d) > \frac{1}{2}$ pour tout $n > N$. Comme $\lim_n N_n = \infty$, on obtient une contradiction. Cela conclut donc la preuve du théorème 2.1.

Références

Arora, S., Lund, C. et al. (1998). « Proof verification and the hardness of approximation problems », *J. ACM* **45** (3), p. 501-555.

Arora, S. et Safra, S. (1998). « Probabilistic checking of proofs : a new characterization of NP », *J. ACM* **45** (1), p. 70-122.

Aubrun, G. (2021). « The Parallel repetition theorem », *Notes disponibles sur la page web de l'auteur*.

Bavarian, M., Vidick, T. et Yuen, H. (2022). « Anchored parallel repetition for nonlocal games », *SIAM J. Comput.* **51** (2), p. 214-253.

Bekka, B., Harpe, P. de la et Valette, A. (2008). *Kazhdan's property (T)*. T. 11. New Mathematical Monographs. Cambridge University Press, Cambridge, p. xiv+472.

Ben-Or, M. et al. (1988). « Multi-Prover Interactive Proofs : How to Remove Intractability Assumptions ». In : *Proceedings of the 20th Annual ACM Symposium on Theory of Computing, May 2-4, 1988, Chicago, Illinois, USA*. Sous la dir. de J. Simon. ACM, p. 113-131.

Chazelle, B. (2003). « The PCP theorem [after Arora, Lund, Motwani, Safra, Sudan, Szegedy] », in : Astérisque 290. Séminaire Bourbaki. Vol. 2001/2002, Exp. No. 895, vii, 19-36.

Connes, A. (1976). « Classification of injective factors. Cases II_1, II_∞, III_λ, $\lambda \neq 1$ », *Ann. of Math.* (2) **104** (1), p. 73-115.

De Chiffre, M., Ozawa, N. et Thom, A. (2019). « Operator algebraic approach to inverse and stability theorems for amenable groups », *Mathematika* **65** (1), p. 98-118.

Fritz, T. (2012). « Tsirelson's problem and Kirchberg's conjecture », *Rev. Math. Phys.* **24** (5), p. 1250012, 67.

Gowers, W. T. et Hatami, O. (2017). « Inverse and stability theorems for approximate representations of finite groups », *Mat. Sb.* **208** (12), p. 70-106.

Ji, Z. et al. (2020a). « MIP*=RE ». arXiv.

——— (2020b). « Quantum soundness of the classical low individual degree test ». arXiv.

——— (2022). « Quantum soundness of testing tensor codes », *Discrete Analysis* **17**, 73 pp.

Junge, M. et al. (2011). « Connes embedding problem and Tsirelson's problem », *J. Math. Phys.* **52** (1), p. 012102, 12.

Kirchberg, E. (1993). « On nonsemisplit extensions, tensor products and exactness of group C^*-algebras », *Invent. Math.* **112** (3), p. 449-489.

Natarajan, A. et Wright, J. (2019). « NEEXP is contained in MIP », in : *2019 IEEE 60th Annual Symposium on Foundations of Computer Science*. IEEE Comput. Soc. Press, Los Alamitos, CA, p. 510-518.

Navascués, M., Pironio, S. et Acín, A. (2008). « A convergent hierarchy of semidefinite programs characterizing the set of quantum correlations. » *New J. Phys.* **10**, p. 073013.

Ozawa, N. (2013). « About the Connes embedding conjecture : algebraic approaches », *Jpn. J. Math.* **8** (1), p. 147-183.

Pansu, P. (2013). « Difficulté d'approximation (d'après Khot, Kindler, Mossel, O'Donnell, ...) », in : Astérisque 352. Séminaire Bourbaki. Vol. 2011/2012. Exposés 1043–1058, Exp. No. 1045, vii, 83-120.

Paulsen, V. I. et al. (2016). « Estimating quantum chromatic numbers », *J. Funct. Anal.* **270** (6), p. 2188-2222.

Pisier, G. (2020). *Tensor products of C^*-algebras and operator spaces—the Connes–Kirchberg problem*. T. 96. London Mathematical Society Student Texts. Cambridge University Press, Cambridge, p. x+484.

Raz, R. (1998). « A parallel repetition theorem », *SIAM J. Comput.* **27** (3), p. 763-803.

Salle, M. de la (2022). « Spectral gap and stability for groups and non-local games ». arXiv.

Tsirelson, B. S. (1980). « Quantum generalizations of Bell's inequality », *Lett. Math. Phys.* **4** (2), p. 93-100.

——— (1993). « Some results and problems on quantum Bell-type inequalities », *Hadronic J. Suppl.* **8** (4), p. 329-345.

Vidick, T. (2016). « Three-player entangled XOR games are NP-hard to approximate », *SIAM J. Comput.* **45** (3), p. 1007-1063.

——— (2020). « Erratum : Three-player entangled XOR games are NP-hard to approximate », *SIAM J. Comput.* **49** (6), p. 1423-1427.

——— (2022a). « Almost synchronous quantum correlations », *J. Math. Phys.* **63** (2), Paper No. 022201, 17.

——— (2022b). « MIP*=RE, A negative resolution to Connes' Embedding Problem and Tsirelson's problem », *Proceedings of the ICM 2022*.

Mikael de la Salle

CNRS – Université Claude Bernard Lyon 1
Institut Camille Jordan
43 boulevard du 11 novembre 1918
F-69622 Villeurbanne Cedex

E-mail : delasalle@math.univ-lyon1.fr

Séminaire BOURBAKI Avril 2023
74e année, 2022–2023, no 1204, p. 251 à 284
doi : 10.24033/ast.1213

LE GROUPE DES HOMÉOMORPHISMES DE LA SPHÈRE DE DIMENSION 2 QUI RESPECTENT L'AIRE ET L'ORIENTATION N'EST PAS UN GROUPE SIMPLE.

[*d'après* D. Cristofaro-Gardiner, V. Humilière et S. Seyfaddini]

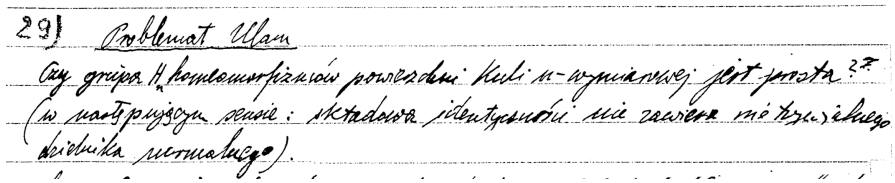
par **Étienne Ghys**

1. Un peu de contexte

Le théorème énoncé dans le titre a des racines anciennes.

1.1. La préhistoire

1.1.1. *Les groupes d'homéomorphismes.* — La question suivante était déjà posée par Ulam (1935) dans le « Księga Szkocka », plus connu sous le nom de « Scottish Book » [(1)].


29) Problemat Ulam
Czy grupa H homeomorfizmów powierzchni kuli n-wymiarowej jest prosta ?
(w sensie następującym : składowa jedności nie zawiera żadnego dzielnika normalnego).

29) Problème d'Ulam
Le groupe H_n de tous les homéomorphismes de la surface de la sphère de dimension n est-il simple ? (dans le sens suivant : la composante de l'identité ne contient pas de sous-groupe distingué non trivial).

Le cas du cercle ($n = 1$) avait été résolu une année auparavant par Schreier et Ulam (1934). Ulam et von Neumann (1947) annoncèrent ensuite une solution pour $n = 2$. Anderson (1958) et Fisher (1960) résolurent le cas $n \leqslant 3$ mais leurs démonstrations se généralisèrent immédiatement en toute dimension une fois que le théorème de Schoenflies généralisé et la conjecture de l'anneau furent établis pour tout n.

(1) Le problème 28, proposé par Mazur, promettait une bouteille de vin à celui qui le résoudrait. Mais aucune récompense n'était prévue pour le problème 29.

1.1.2. *Les difféomorphismes.* — La question bien plus difficile dans le cas des *groupes de difféomorphismes* fut abordée par Epstein (1970), Herman (1971, 1973), Thurston (1974) et Mather (1975) dans une série d'articles impressionnants. Le résultat suivant de Thurston avait été conjecturé par Smale : *la composante neutre du groupe des difféomorphismes de classe C^∞ et à support compact d'une variété connexe est un groupe simple, en toute dimension* [(2)].

1.1.3. *Les difféomorphismes qui préservent le volume ou une forme symplectique.* — Tout naturellement, il s'agissait ensuite d'étudier les difféomorphismes qui préservent une structure additionnelle, comme une forme de volume ou une forme symplectique. Ce sont en effet des exemples emblématiques de groupes de Lie de dimension infinie. Le théorème final fut annoncé [(3)] par Thurston (1973) dans un article « à paraître » qui n'est jamais paru. L'article et le livre de Banyaga (1978, 1997) contiennent en revanche des preuves complètes (voir aussi le complément de Rousseau, 1978).

La situation est alors plus délicate car les groupes correspondants ne sont pas toujours simples.

Soit M une variété différentiable connexe et vol une forme de volume de masse totale finie. Notons $\mathrm{Diff}(M, vol)$ le groupe des difféomorphismes de M (de classe C^∞), à support compact, qui respectent vol. Soit $\mathrm{Diff}_0(M, vol)$ la composante connexe de l'identité et $\widetilde{\mathrm{Diff}}_0(M, vol)$ son revêtement universel. On peut alors définir un homomorphisme, appelé *flux*, de $\widetilde{\mathrm{Diff}}_0(M, vol)$ vers le premier groupe d'homologie $H_1(M, \mathbf{R})$ de la manière suivante. Soit $f_{t\in[0,1])}$ un chemin dans $\mathrm{Diff}(M, vol)$ reliant l'identité f_0 à un difféomorphisme $f = f_1$. Pour chaque point $x \in M$, le chemin $c_x : t \in [0,1] \mapsto f_t(x)$ peut être considéré comme un 1-courant (dont le bord est $f(x) - x$). L'intégrale sur M de c_x par rapport à vol est un 1-cycle dont la classe d'homologie ne dépend que de la classe d'homotopie de f_t à extrémités fixes. Cela définit le flux

$$\Phi : \widetilde{\mathrm{Diff}}_0(M, vol) \to H_1(M, \mathbf{R})$$

qui s'avère être surjectif et qui descend en un homomorphisme

$$\phi : \mathrm{Diff}_0(M, vol) \to H_1(M, \mathbf{R})/\Phi(\pi_1(\mathrm{Diff}(M, vol))).$$

En dimension $n \geqslant 3$, le noyau de ϕ est un groupe simple.

Soit maintenant M une variété symplectique connexe et ω une forme symplectique de volume fini. Notons $\mathrm{Diff}(M, \omega)$ le groupe des difféomorphismes de M (de classe C^∞), à support compact, qui respectent ω. Soit $\mathrm{Diff}_0(M, \omega)$ la composante

(2) Voir Mann (2016) pour une preuve simplifiée et moderne.

(3) McDuff (1980) écrit : « Unfortunately Thurston's proof has remained unpublished. However his results were later generalized by Banyaga to the symplectic case, and one can (with some difficulty) reconstruct Thurston (1973)'s argument ».

connexe de l'identité et $\widetilde{\mathrm{Diff}}_0(M,\omega)$ son revêtement universel. Puisqu'un difféomorphisme symplectique respecte le volume, on dispose de l'homomorphisme flux Φ défini sur $\widetilde{\mathrm{Diff}}_0(M,\omega)$, comme précédemment. Si M est compacte, Banyaga et Thurston démontrent que le noyau du flux est simple. En revanche, lorsque M est non compacte, le noyau $ker(\Phi)$ du flux, n'est pas simple : il est muni d'un homomorphisme $\mathscr{C}al$ à valeurs dans $\mathbf{R}$ dont le noyau est simple. Nous reviendrons plus loin sur la définition de cet invariant introduit par Calabi (1970) [(4)].

1.1.4. *Les difféomorphismes des surfaces qui préservent l'aire.* — En dimension 2 une forme symplectique n'est autre qu'une forme d'aire. La situation est parfaitement comprise dans le cas des surfaces mais pour ne pas alourdir la discussion, limitons-nous ici aux deux cas qui seront au cœur de cet exposé : la sphère et le disque. Puisque ces deux exemples sont simplement connexes, nous n'aurons pas à nous préoccuper du flux. En dimension 2, nous noterons plutôt *aire* une forme de « volume ». Nous considérons *aire* tantôt comme une mesure et tantôt comme une 2-forme différentielle.

Le groupe $\mathrm{Difféo}_0(\mathbf{S}^2, aire)$ *des* ***difféomorphismes*** *de la sphère qui respectent l'aire est simple*.

Pour le disque *fermé* $\mathbf{D}^2$, on note $\mathrm{Difféo}(\mathbf{D}^2, \partial\mathbf{D}^2, aire)$ le groupe des difféomorphismes qui respectent l'aire et qui sont l'identité près du bord. C'est aussi le groupe des difféomorphismes à support compact du disque ouvert qui respectent l'aire.

Le noyau de l'homomorphisme de Calabi $\mathrm{Difféo}(\mathbf{D}^2, \partial\mathbf{D}^2, aire) \to \mathbf{R}$ *est un groupe simple.*

1.1.5. *Les homéomorphismes qui préservent le volume.* — Il fallait étudier également les *homéomorphismes* qui préservent le volume. Cela fut fait dans un article remarquable de Fathi (1980a). Tout d'abord, il généralisa la définition du flux au groupe $\mathrm{Homéo}_0(M, vol)$, composante neutre du groupe des homéomorphismes à support compact qui respectent le volume. Surtout, il démontra qu'**en dimension** $\geqslant 3$ *le noyau du flux est un groupe simple*.

[(4)] C'est l'occasion de rendre hommage à Eugenio Calabi, qui fêtera son centième anniversaire le 11 mai 2023.

1.2. Les homéomorphismes du disque et de la sphère qui préservent l'aire

Le cas des surfaces, et tout particulièrement de la sphère et du disque de dimension 2, a résisté à de nombreux efforts depuis une quarantaine d'années. Les théorèmes de Cristofaro-Gardiner, Humilière et Seyfaddini (2020, 2021) sont une surprise [(5)] :

Le groupe des ***homéomorphismes*** *de la sphère de dimension 2 qui respectent l'aire et l'orientation n'est* ***pas*** *un groupe simple.*

Le groupe des ***homéomorphismes*** *du disque de dimension 2 qui respectent l'aire et qui coïncident avec l'identité près du bord n'est* ***pas*** *un groupe simple.*

La démonstration est un tour de force et fait largement usage de l'homologie de Floer.

La construction explicite d'un sous-groupe distingué n'est pas difficile mais il faut montrer qu'un homéomorphisme très facile à décrire n'est pas dans ce sous-groupe. Hélas, la nature du groupe quotient reste mystérieuse, même si nous en décrirons quelques propriétés.

Ce théorème a été immédiatement précisé et généralisé dans plusieurs prépublications très récentes, avec des méthodes assez différentes. Nous évoquerons plus loin quelques-uns de ces résultats, malheureusement trop superficiellement.

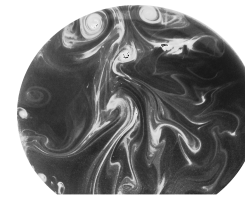
Il est peut-être utile de préciser qu'il s'agit de simplicité au sens algébrique : un groupe est simple s'il ne possède pas de sous-groupe distingué non trivial. Pour un groupe topologique on parle de *simplicité topologique* s'il n'existe pas de sous-groupe distingué *fermé* non trivial. Nous verrons que *le groupe des homéomorphismes de la sphère de dimension 2 qui respectent l'aire et l'orientation* ***est topologiquement simple*** *lorsqu'on le munit de la topologie de la convergence uniforme.*

1.2.1. *Pourquoi s'intéresser aux homéomorphismes ?* — On peut légitimement se demander s'il est utile de dépenser tant d'énergie [(6)] pendant quarante ans pour étudier les groupes d'homéomorphismes qui préservent l'aire alors que la situation des difféomorphismes est connue depuis longtemps.

[(5)] Surprise pour l'auteur de ce texte qui a longtemps tenté de démontrer le contraire. En revanche, ce n'est pas une surprise pour McDuff et Salamon (2017) (problem 42) qui avaient fait la « bonne » conjecture.

[(6)] Nous verrons en effet qu'il faut une énergie infinie.

Une première réponse est de nature physique : il suffit d'observer une rivière pour comprendre que le flot n'est pas de classe C^∞. Une solution classique de l'équation d'Euler pour un fluide parfait incompressible de dimension 2 est donnée par un *tourbillon ponctuel*. Après un temps t un point de coordonnées polaires (r, θ) est transporté au point $(r, \theta + t/r^2)$. Il s'agit d'homéomorphismes non différentiables à l'origine. S'il y a plusieurs tourbillons en interaction, le mouvement global ressemble à celui du lait dans une tasse de café. La dynamique d'un ensemble de k tourbillons ponctuels est un système hamiltonien dans $\mathbf{R}^{2k}$ qui peut être considéré comme une approximation de l'équation d'Euler (Arnold et Khesin, 1998). Beaucoup de topologues rêvent depuis longtemps (à ce jour sans succès) de développer une étude topologique de la dynamique des fluides, qui serait fondée sur la topologie algébrique, sans la moindre dérivée (voir par exemple D. Sullivan, 2011).


Le lait dans le café.

On peut aussi décrire beaucoup de situations très naturelles d'homéomorphismes non lisses qui préservent l'aire. Par exemple, le groupe $\mathrm{SL}(2, \mathbf{Z})$ agit linéairement sur le tore $\mathbf{R}^2/\mathbf{Z}^2$ en préservant l'aire. En passant au quotient par l'involution $x \mapsto -x$, on obtient une action *non lisse* de $\mathrm{PSL}(2, \mathbf{Z})$ sur la sphère. Les homéomorphismes des surfaces qu'on appelle *difféomorphismes pseudo-Anosov* préservent l'aire mais ce ne sont pas des difféomorphismes !

De manière plus fondamentale, il faut rappeler que l'article de Gromov (1987) intitulé « Soft and hard symplectic geometry » a inauguré l'étude de la *topologie symplectique*. Le théorème de rigidité affirme qu'un difféomorphisme qui est la limite uniforme (en topologie C^0) d'une suite de difféomorphismes symplectiques est nécessairement symplectique. Cela conduit à définir un *homéomorphisme symplectique* comme la limite uniforme d'une suite de difféomorphismes symplectiques. Cela signifie-t-il qu'on peut caractériser les homéomorphismes symplectiques en termes uniquement qualitatifs ? Comprendre la structure des homéomorphismes des surfaces qui respectent l'aire est un premier pas vers l'étude des homéomorphismes symplectiques en toute dimension. Il s'agit en quelque sorte de faire passer la mécanique analytique classique du stade de la géométrie différentielle, à la Lagrange, au stade de la topologie.

Il y a bien sûr bien d'autres problèmes actuels de nature algébrique sur les groupes d'homéomorphismes. On en trouvera une présentation très accessible dans Mann (2021).

1.2.2. *Cet exposé.* — Même si les articles de Cristofaro-Gardiner, Humilière et Seyfaddini sont remarquablement écrits, ils sont très longs et parfois techniques. Dans le cadre de cet exposé introductif, qui n'est pas destiné aux spécialistes, je ne peux que présenter le contexte ainsi que les outils utilisés dans la preuve, sans prétendre donner une preuve de ces beaux résultats.

Une première partie sera consacrée à la démonstration du théorème de Fathi qui n'est valable qu'en dimension $\geqslant 3$. Pour apprécier le théorème principal de cet exposé à sa juste mesure, il m'a en effet paru indispensable de comprendre ce qui empêche la preuve de Fathi de se généraliser en dimension 2. Pour une raison inconnue, l'article de Fathi est passé relativement inaperçu pendant plus de trente ans. Il a fallu attendre l'analyse de Le Roux (2010) pour clarifier les obstacles dans le cas des surfaces.

J'exposerai ensuite la stratégie générale de Cristofaro-Gardiner, Humilière et Seyfaddini (2020, 2021).

Pour conclure, je citerai d'autres résultats récents, qui ouvrent un vaste champ de recherches.

Je dédie cet article à la mémoire d'André Haefliger (1929-2023) qui fut tant de fois une source d'inspiration dans ma carrière et qui a souvent contribué à ce séminaire.

2. Le théorème de Fathi

On note Homéo(M) le groupe des homéomorphismes d'une variété connexe M de dimension n. Un indice 0 est utilisé pour la composante neutre. Lorsque M possède un bord non vide, on note Homéo$(M, \partial M)$ le sous-groupe des homéomorphismes qui sont l'identité près du bord. On note Homéo(M, vol) le groupe des homéomorphismes qui préservent (la mesure définie par) un volume vol.

Nous allons nous contenter ici d'un cas particulier du théorème de Fathi : Homéo$_0(\mathbf{S}^n, vol)$ *est simple en dimension* $n \geqslant 3$.

2.1. Quelques outils

2.1.1. *L'astuce d'Alexander (1923).* — *Le groupe des homéomorphismes de la boule unité fermée* $\mathbf{B}^n \subset \mathbf{R}^n$ *qui coïncident avec l'identité sur le bord est connexe par arcs.* En effet un tel homéomorphisme f peut être joint à l'identité par un chemin f^t défini pour $0 < t \leqslant 1$ par $f^t(x) = tf(x/t)$ pour $\|x\| \leqslant t$ et $f^t(x) = x$ pour $\|x\| \geqslant t$.

2.1.2. *Le théorème de Shoenflies généralisé (Brown, 1960) et le théorème de l'anneau (Kirby, 1969; Quinn, 1982)* . — En dimension 2, le théorème de Jordan-Schoenflies garantit que si i_1, i_2 sont deux plongements du cercle $\mathbf{S}^1$ dans la sphère $\mathbf{S}^2$, il existe un homéomorphisme h de la sphère tel que $i_2 = hi_1$. On dispose aujourd'hui de démonstrations élémentaires de ce théorème mais la généralisation en dimension supérieure est bien plus subtile.

Un plongement de la sphère $\mathbf{S}^{n-1}$ dans M est *localement plat* s'il se prolonge en un plongement de $\mathbf{S}^{n-1} \times [-\varepsilon, +\varepsilon]$ dans M. Lorsque $n = 2$ tous les plongements sont localement plats mais ce n'est plus le cas pour $n \geqslant 3$. Le *théorème de Schoenflies généralisé* affirme qu'un plongement localement plat de $\mathbf{S}^{n-1}$ dans $\mathbf{S}^n$ décompose $\mathbf{S}^n$ en deux composantes connexes dont les adhérences sont homéomorphes à des boules fermées $\mathbf{B}^n$.

En général, on appellera *boule* l'image d'un plongement de $\mathbf{B}^n$ dans M. Si la restriction de ce plongement au bord est localement plate, on parle d'une *boule localement plate*. Le théorème de l'anneau affirme que *si B est une boule localement plate contenue dans l'intérieur de $\mathbf{B}^n$, le complémentaire $\mathbf{B}^n \setminus B^\circ$ de son intérieur est homéomorphe à un anneau standard $\{1/2 \leqslant \|x\| \leqslant 1\}$ par un homéomorphisme qui est l'identité sur la sphère unité $\mathbf{S}^{n-1}$, bord de $\mathbf{B}^n$.*

2.1.3. *La fragmentation.* — Le *support* d'un homéomorphisme est l'adhérence de l'ensemble des points qui ne sont pas fixes. Le lemme de fragmentation affirme que *tout homéomorphisme à support compact peut s'écrire comme la composition d'un certain nombre (a priori non borné) d'homéomorphismes dont les supports sont arbitrairement petits* (c'est-à-dire contenus dans l'un des ouverts d'un recouvrement ouvert fixé a priori).

L'énoncé analogue dans le cas des difféomorphismes est également valide et sa preuve n'est pas difficile. Il suffit de le démontrer pour un difféomorphisme f arbitrairement proche de l'identité, temps 1 d'un chemin f^t tel que $f^0 = id$. On peut alors utiliser une partition de l'unité λ_i adaptée à un recouvrement du support par un nombre fini d'ouverts U_i $(i = 1, \ldots, k)$. On définit alors $\zeta_j = \sum_{i \leqslant j} \lambda_i$. Les applications définies par $h_j(x) = f^{\zeta_j(x)}(x)$ sont proches de l'identité et sont donc des difféomorphismes qui sont tels que h_j et h_{j-1} coïncident hors de U_j. La décomposition $f = (h_0^{-1}h_1)(h_1^{-1}h_2)\cdots(h_{k-1}^{-1}h_k)$ est la décomposition cherchée.

Pour les homéomorphismes, la démonstration est plus compliquée et utilise le théorème de l'anneau. On peut en déduire que $\mathrm{Homéo}_0(M, \partial M)$ agit transitivement sur les boules localement plates.

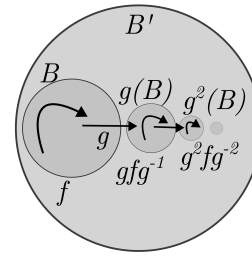
2.1.4. *Le théorème de Oxtoby et Ulam (1941).* — Ce théorème montre que le volume utilisé dans notre discussion des groupes d'homéomorphismes n'est pas nécessairement défini par une forme volume. Considérons deux mesures de probabilité μ_1 et μ_2 sur une variété compacte connexe M qui sont de *bonnes mesures* dans le sens suivant : la mesure du bord est nulle, les mesures des points sont nulles et les mesures des ouverts non vides sont non nulles. Le théorème affirme qu'*il existe un homéomorphisme f de M qui envoie μ_1 sur μ_2.* De plus, on peut supposer que f est dans $\mathrm{Homéo}_0(M)$ et l'identité sur le bord.

2.2. Le théorème d'Anderson

Pour nous préparer à la preuve du théorème de Fathi, voici une esquisse de preuve du théorème d'Anderson : *la composante neutre du groupe des homéomorphismes d'une variété connexe M (compacte sans bord pour fixer les idées) est un groupe simple.*

2.2.1. $\text{Homéo}_0(M)$ *est un groupe parfait.* — Rappelons qu'un groupe est *parfait* si tout élément est un produit de commutateurs [7]. Il est clair qu'un groupe simple (non cyclique) est nécessairement parfait car le sous-groupe engendré par les commutateurs est distingué.

Soit f un homéomorphisme dont le support est contenu dans une boule $B \subset M$ localement plate. Soit B' une autre boule localement plate contenant B dans son intérieur. À conjugaison topologique près, on peut supposer que B' est la boule unité $\mathbf{B}^n$ et que B est une boule euclidienne. Soit g un homéomorphisme à support dans B' tel que les boules $g^k(B)$ soient des boules euclidiennes disjointes deux à deux pour $k \geqslant 0$. Soit h l'homéomorphisme qui coïncide avec $g^k f g^{-k}$ sur $g^k(B)$ et avec l'identité en dehors de la réunion des $g^k(B)$.



Le conjugué ghg^{-1} coïncide avec h partout sauf sur B où il coïncide avec id. Il en résulte que $(ghg^{-1})^{-1}h = f$ et nous avons écrit f comme un commutateur $gh^{-1}g^{-1}h$. Le groupe $\text{Homéo}_0(M)$ est donc parfait puisque nous savons que les homéomorphismes à supports dans des boules (localement plates) engendrent ce groupe.

2.2.2. $\text{Homéo}_0(M)$ *est un groupe simple.* — Passer de la perfection à la simplicité n'est pas toujours facile... mais nous allons utiliser ici une astuce de Thurston.

Soient f_1, f_2 deux homéomorphismes à support dans une boule localement plate B. Soient g_1, g_2 deux homéomorphismes tels que $B, g_1(B), g_2(B)$ soient disjoints. On vérifie alors facilement que le commutateur $[f_1, f_2] = f_1 f_2 f_1^{-1} f_2^{-1}$ est égal au commutateur $[[f_1, g_1], [f_2, g_2]]$. Par conséquent $[f_1, f_2]$ appartient au sous-groupe distingué engendré par g_1. Pour tout $g_1 \neq id$, on peut construire un g_2 qui vérifie les conditions précédentes. Autrement dit, tout sous-groupe distingué non trivial de $\text{Homéo}_0(M)$ contient les commutateurs $[f_1, f_2]$, donc tous les homéomorphismes à support dans une boule localement plate, par perfection, et donc le groupe $\text{Homéo}_0(M)$ tout entier.

2.3. Le théorème de Fathi : le groupe $\text{Homéo}_0(\mathbf{S}^n, vol)$ est simple pour $n \geqslant 3$

La démonstration de la perfection de $\text{Homéo}_0(M)$ reposait sur l'existence d'un homéomorphisme g tel que les images $g^k(B)$ sont disjointes. C'est clairement impossible si g préserve un volume de masse finie.

[7] Postnikov (1985) parle de groupe *caïnien* puisqu'un tel groupe tue Abel.

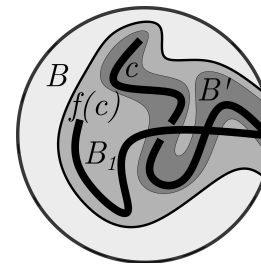
2.3.1. *Fragmentation améliorée.* — *Pour $n \geqslant 3$, tout élément de* $\mathrm{Homéo}_0(\mathbf{B}^n, \partial\mathbf{B}^n, vol)$ *peut être représenté comme la composition de deux éléments dont les supports sont contenus dans des boules de volumes inférieurs à* $\frac{3}{4}vol(\mathbf{B}^n)$.

Nous nous limiterons à une esquisse de preuve.

Convenons de dire qu'une boule localement plate est *bonne* si la mesure *vol* de son bord est nulle. Le groupe $\mathrm{Homéo}_0(M, \partial M, vol)$ agit transitivement sur les bonnes boules dont la mesure *vol* est donnée.

Soit *vol* la mesure de Lebesgue sur $\mathbf{B}^n$ de *masse totale* 1. Soit ν la mesure sur le segment $r\colon \rho \in [3/4, 1] \mapsto (\rho, 0, \cdots, 0) \in \mathbf{B}^n$ obtenue comme image de la mesure de Lebesgue, de masse totale $1/4$. La mesure $\mu = \frac{3}{4}vol + \nu$ vérifie les hypothèses du théorème de Oxtoby–Ulam. Il existe donc un élément de $\mathrm{Homéo}_0(\mathbf{B}^n)$, identité sur le bord, qui envoie μ sur *vol* et donc le rayon r sur un arc c de mesure de Lebesgue $vol(c) = 1/4$ dans $\mathbf{B}^n$. Soit f un élément de $\mathrm{Homéo}_0(\mathbf{B}^n, \partial\mathbf{B}^n, vol)$. Puisque $n \geqslant 3$, le complémentaire de la réunion $c \cup f(c)$ est un ouvert **connexe** dont le volume est compris entre $1/2$ et $3/4$.

Par des arguments purement topologiques, dus à Brown (1962), et valables en toute dimension, dépendant uniquement de la connexité du complémentaire de $c \cup f(c)$, on peut réaliser la situation illustrée sur la figure. La boule $\mathbf{B}^n$ est la réunion de deux parties B et B_1, homéomorphes à une boule, et dont l'intersection, de mesure nulle, est homéomorphe à une boule fermée localement plate de dimension $n-1$. La boule B_1 contient $c \cup f(c)$ dans son intérieur. Les intersections de B et B_1 avec $\mathbf{S}^{n-1} = \partial\mathbf{B}^n$ décomposent $\mathbf{S}^{n-1}$ en deux boules de dimension $n-1$ dont le bord commun est localement plat.

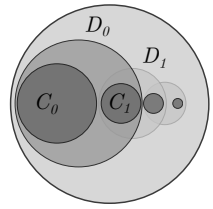


Par ailleurs on peut obtenir cette décomposition de telle sorte que B_1 soit un voisinage arbitrairement petit de $c \cup f(c)$. En particulier, on peut faire en sorte que le volume de B est $> 1/4$. On peut alors trouver une bonne boule $B' \subset B_1$ contenant c qui, de même que B_1, rencontre $\mathbf{S}^{n-1}$ sur une boule de dimension $n-1$. Le volume de B_1 est $> 1/4$. On peut aussi supposer que $B' \cup f(B')$ est contenu dans B_1. Le théorème de Schoenflies généralisé permet alors de garantir l'existence d'un homéomorphisme f_1 préservant *vol* qui coïncide avec f sur B' et qui est l'identité sur B. On a alors décomposé $f = f_1 f_2$ où $f_2 = f_1^{-1} f$. Le support de f_1 est contenu dans B_1, de volume inférieur à $3/4$. Le support de f_2 est contenu dans le complémentaire de $f(B')$ donc de volume inférieur à $3/4$.

2.3.2. $\mathrm{Homéo}_0(B, \partial B, vol)$ *est un groupe parfait.* — Nous avons vu que la preuve d'Anderson utilise un homéomorphisme g tel que des boules $g^k(B)$ soient disjointes deux à deux pour $k \geqslant 0$. C'est bien sûr impossible si g préserve un volume de masse totale finie. Fathi utilise la fragmentation améliorée.

Soit B une bonne boule, identifiée à la boule unité $\mathbf{B}^n$ (munie d'un multiple constant de vol pour que B et $\mathbf{B}^n$ aient la même masse totale). Considérons deux suites de boules euclidiennes $(C_i)_{i\geqslant 0}$ et $(D_i)_{i\geqslant 0}$ à l'intérieur de $\mathbf{B}^n$ vérifiant les propriétés suivantes. L'intérieur de D_i contient C_i et C_{i+1}, les C_i sont disjoints, les D_{2i} sont disjoints, de même que les D_{2i+1}. On suppose que $vol(C_{i+1}) = \frac{3}{4}vol(C_i)$. Considérons un homéomorphisme $f = f_0$ qui respecte le volume et dont le support est dans l'intérieur de C_0. D'après la fragmentation améliorée, on peut décomposer f_0 en un produit de deux homéomorphismes dont les supports sont dans des boules de volumes $< \frac{3}{4}vol(C_0) = vol(C_1)$.

Autrement dit $f_0 = h_{1,0}h_{2,0}$ et le support de $h_{i,0}$ est contenu dans une boule $G_{i,0} \subset C_0$ de volume $< vol(C_1)$ (pour $i = 1, 2$). On peut donc conjuguer chaque $h_{i,0}$ par un homéomorphisme $k_{i,1}$ préservant le volume, de support dans D_0 et envoyant $G_{i,0}$ dans C_1. On obtient des éléments $h_{i,1} = k_{i,1}h_{i,0}k_{i,1}^{-1}$ dont les supports sont contenus C_1. On pose alors $f_1 = h_{1,1}h_{2,1}$. C'est un homéomorphisme dont le support est contenu dans C_1.



Notons que $f_0 f_1^{-1} = (h_{1,0}h_{2,0})(k_{1,1}h_{1,0}k_{1,1}^{-1}k_{2,0}h_{2,1}k_{2,1}^{-1})^{-1}$ est un produit de commutateurs dans le groupe engendré par les $h_{i,0}, k_{i,1}$ car ce mot s'annule lorsqu'on le rend abélien. On pourrait facilement expliciter un tel produit et on peut vérifier qu'il suffit de deux commutateurs (même si ce n'est pas important). Autrement dit $f_0 f_1^{-1} = [a_0, b_0][c_0, d_0]$ où les a_0, b_0, c_0, d_0 sont des mots en les h, k, donc de support dans D_0.

On peut alors faire subir à f_1 le même traitement que nous avons fait subir à f_0. On obtient un homéomorphisme f_2 dont le support est dans C_2. Par récurrence, on construit $f_1, f_2, f_3, \ldots$ dont les supports sont dans $C_1, C_2, C_3 \ldots$ et on a une décomposition $f_k f_{k+1}^{-1} = [a_k, b_k][c_k, d_k]$ pour $k \geqslant 0$. Les supports des a_k, b_k, c_k, d_k sont contenus dans D_k.

Il nous reste à définir les produits infinis (en observant que tous les termes commutent puisque les C_i sont disjoints)

$$g = f_0 f_1^{-1} f_2 f_3^{-1} \cdots \quad \text{et} \quad g' = f_1 f_2^{-1} f_3 f_4^{-1} \cdots$$

Notre f_0 initial est gg'. On peut alors écrire f_0 comme un produit de commutateurs en posant :

$$a = a_0 a_2 a_4 \cdots ; \quad b = b_0 b_2 b_4 \cdots ; \quad c = c_0 c_2 c_4 \cdots ; \quad d = d_0 d_2 d_4 \cdots$$

$$a' = a_1 a_3 a_5 \cdots ; \quad b' = b_1 b_3 b_5 \cdots ; \quad c' = c_1 c_3 c_5 \cdots ; \quad d' = d_1 d_3 d_5 \cdots .$$

Clairement $g = [a, b][c, d]$ et $g' = [a', b'][c', d']$. Nous avons donc écrit f_0 comme un produit de commutateurs.

2.3.3. *Fin de la démonstration du théorème de Fathi : le groupe* $\mathrm{Homéo}_0(\mathbf{S}^n, vol)$ *est simple pour* $n \geqslant 3$. — Nous avons tous les outils. D'après le théorème de l'anneau, l'adhérence du complémentaire d'une boule localement plate dans $\mathbf{S}^n$ est une autre boule localement plate. Un homéomorphisme $\mathrm{Homéo}_0(\mathbf{S}^n, vol)$ proche de l'identité peut-être composé avec un autre pour se ramener au cas où une petite boule (localement plate) est préservée. On en déduit facilement un lemme de fragmentation pour $\mathrm{Homéo}_0(\mathbf{S}^n, vol)$. Pour une variété plus générale que la sphère, il faudrait se placer dans le noyau de l'homomorphisme flux. La perfection de $\mathrm{Homéo}_0(B, \partial B, vol)$ entraîne alors celle de $\mathrm{Homéo}_0(\mathbf{S}^n, vol)$. Avec le même argument de Thurston, la simplicité en résulte comme dans le cas du théorème d'Anderson.

2.4. Pourquoi la démonstration du théorème de Fathi ne s'applique pas en dimension 2

Le théorème principal décrit dans cet exposé consiste en la négation de celui de Fathi en dimension 2 : $\mathrm{Homéo}_0(\mathbf{S}^2, aire)$ *n'est pas un groupe simple.* Il est donc important de comprendre quels sont les arguments dans la preuve de Fathi qui ne sont pas valides en dimension 2. Le seul endroit où on utilise que la dimension est $\geqslant 3$ se situe dans la preuve de la fragmentation améliorée : nous avons utilisé le fait que le complémentaire de la réunion de deux arcs plongés est connexe. Ce n'est pas le cas en dimension 2 lorsque les arcs se rencontrent.

2.4.1. *Fragmentation bornée.* — Un article lumineux de Le Roux (2010) analyse cette question en détail. Il définit la propriété P_ρ pour $0 < \rho < 1$.

P_ρ : Il existe un entier m tel que tout élément de $\mathrm{Homéo}(\mathbf{D}^2, \partial\mathbf{D}^2, aire)$ *à support dans un disque d'aire $< \rho$ est un produit d'au plus m éléments dont les supports sont dans des disques d'aires $< \rho/2$.*

Clairement P_ρ entraîne $P_{\rho'}$ pour $\rho' < \rho$ en itérant plusieurs fois le processus de fragmentation. Le Roux démontre alors que la simplicité de $\mathrm{Homéo}(\mathbf{D}^2, \partial\mathbf{D}^2, aire)$ est équivalente au fait que P_ρ est satisfaite pour au moins une valeur de ρ. Nous savons donc maintenant que P_ρ *n'est satisfaite pour aucun* $0 < \rho < 1$ et que c'est bien cette fragmentation qui est au cœur de la question.

2.4.2. *Invariant de Calabi.* — Soit f un élément de $\mathrm{Difféo}(\mathbf{D}^2, \partial\mathbf{D}^2, aire)$. Soit ζ une 1-forme telle que $d\zeta = aire$. Puisque f préserve l'aire, *i.e.* $f^\star d\zeta = d\zeta$, la forme $f^\star\zeta - \zeta$ est fermée et donc la différentielle d'une unique fonction H nulle près du bord. On définit alors l'*invariant de Calabi*, noté $\mathscr{C}al(f)$, comme l'intégrale de H sur le disque. On vérifie facilement qu'il s'agit d'un homomorphisme surjectif à valeurs dans $\mathbf{R}$ et, comme rappelé plus haut, on sait que le noyau de $\mathscr{C}al$ est un groupe simple (Banyaga, 1997).

Pour montrer que Homéo$(\mathbf{D}^2, \partial\mathbf{D}^2, aire)$ n'est pas simple, il « suffirait » donc de prolonger la définition de $\mathscr{C}al$ aux homéomorphismes. Nous verrons que c'est en effet le cas, mais on ne peut certainement pas prolonger $\mathscr{C}al$ par continuité. Pour s'en convaincre, il suffit de considérer des difféomorphismes « twists » définis en coordonnées polaires par $(r, \theta) \mapsto (r, \theta + u(r))$ où u est défini sur $[0, 1]$ et à valeurs réelles. Si u par exemple est nul en dehors de l'intervalle $[1/2k, 1/k]$ il s'agit bien d'un difféomorphisme qui tend uniformément vers l'identité dans la topologie C^0 lorsque k tend vers l'infini. En revanche, l'invariant de Calabi peut être arbitrairement grand comme le montre un calcul élémentaire que nous verrons plus loin.

2.4.3. *Quasi-morphisme de Calabi.* — Normalisons l'aire de la sphère $aire(\mathbf{S}^2) = 1$. On savait déjà que P_ρ n'est pas satisfaite pour $\rho > 1/2$ grâce à une construction d'un *quasi-morphisme* de Calabi due à Entov, Polterovich et Py (2012). Il s'agit d'une application $\psi\colon \text{Difféo}_0(\mathbf{S}^2, aire) \to \mathbf{R}$ telle que $|\psi(f_1 f_2) - \psi(f_1) - \psi(f_2)|$ est borné indépendamment de f_1, f_2. Par ailleurs, ψ est *homogène*, *i.e.* on a $\psi(f^k) = k\psi(f)$ pour tout entier k. Ce quasi-morphisme est tel que si le support de f est contenu dans un disque D d'aire $< 1/2$, la valeur de $\psi(f)$ est égale à (un multiple constant de) l'invariant $\mathscr{C}al_D(f)$ vu comme un difféomorphisme du disque D. Choisissons un disque D_0 dans $\mathbf{S}^2$ d'aire $\rho > 1/2$ ce qui permet de définir un plongement $i\colon \text{Difféo}(D_0, \partial D_0, aire) \to \text{Difféo}_0(\mathbf{S}^2, aire)$. Il se trouve que la différence $\bar{\psi} = \psi \circ i - \mathscr{C}al_{D_0}$ se prolonge par continuité à Homéo$_0(\mathbf{S}^2, aire)$ et s'annule sur tous les éléments dont le support est contenu dans un disque d'aire $< 1/2$. Si tous les éléments de Homéo$(D_0, \partial D_0, aire)$ pouvaient s'écrire comme le produit d'un nombre *borné* d'éléments à support dans des disques d'aires $< 1/2$, le quasi-morphisme ψ serait borné et donc nul puisqu'il est homogène. Comme ce n'est pas le cas, P_ρ n'est pas satisfaite pour $\rho > 1/2$.

2.4.4. *Une infinité non dénombrable de sous-groupes distingués.* — Le Roux propose une construction de sous-groupes distingués. Si $f \in$ Homéo$(\mathbf{D}^2, \partial\mathbf{D}^2, aire)$ et $0 < \rho < 1$ on note $\|f\|_\rho$ la longueur minimale d'une écriture de f comme produit d'éléments dont les supports sont contenus dans des disques d'aire $< \rho$. Soit $\lambda\colon]0, 1] \to \mathbf{R}_+$ une fonction décroissante. On montre facilement que l'ensemble N_λ des homéomorphismes f tels qu'il existe une constante C_f telle que $\|f\|_\rho \leqslant C_f \lambda(\rho)$ pour ρ assez petit est un sous-groupe distingué.

Le Roux montre que si Homéo$(\mathbf{D}^2, \partial\mathbf{D}^2, aire)$ n'est pas simple — ce que nous savons maintenant être le cas — ces sous-groupes N_λ sont propres. Il démontre même qu'on peut construire de cette manière une infinité non dénombrable de sous-groupes distingués propres. Est-il possible de démontrer ces résultats directement ? par des méthodes « élémentaires » ?

2.5. Aparté : de la perfection à la simplicité

D'ordinaire, la démonstration qu'un groupe n'est pas simple consiste à construire un homomorphisme « naturel » vers un autre groupe tout aussi « naturel ». Qu'on pense par exemple au déterminant dans un groupe linéaire ou à la signature dans le groupe des permutations. Ce n'est pas l'approche dans le cas de Homéo$(\mathbf{D}^2, \partial\mathbf{D}^2, \mathit{aire})$ où l'on se « contente » d'exhiber un sous-groupe distingué sans se préoccuper du quotient. Lorsqu'il s'agit de groupes de transformations, il arrive souvent que la perfection entraîne la simplicité, comme nous venons de le voir. Epstein (1970) a d'ailleurs dégagé des conditions très générales qui sont vérifiées dans notre situation. Nous utilisons ici la contraposée puisque nous savons que les groupes que nous étudions ne sont pas simples : Homéo$(\mathbf{D}^2, \partial\mathbf{D}^2, \mathit{aire})$ *et* Homéo$_0(\mathbf{S}^2, \mathit{aire})$ *ne sont donc pas parfaits.*

Il existe cependant quelques exemples « naturels » de groupes parfaits qui ne sont pas simples pour une raison qui n'est pas évidente. Le groupe des germes de difféomorphismes de classe C^∞ de $\mathbf{R}$ à l'origine qui sont C^∞-tangents à l'identité est un groupe parfait (Sergeraert, 1977). Si f est un tel germe, on peut considérer l'ensemble de ses points fixes $fix(f)$ comme un germe d'ensemble fermé au voisinage de l'origine. L'ensemble des f tels que 0 soit un point de densité de $fix(f)$ au sens de Lebesgue est un sous-groupe distingué. On conçoit que le quotient est compliqué.

McDuff (1981) montre que le groupe des difféomorphismes d'une boule ouverte qui respectent le volume est parfait en dimension $n \geqslant 3$ mais il n'est certainement pas simple puisqu'il contient le sous-groupe distingué des difféomorphismes à support compact (qui est un groupe simple). En dimension 2, le groupe des difféomorphismes du disque ouvert qui respectent l'aire n'est pas parfait, comme nous le verrons plus loin.

3. Le théorème de Thurston-Banyaga-Herman

De même qu'il est important de comprendre pourquoi le théorème de Fathi ne se généralise pas de la dimension $\geqslant 3$ à la dimension 2, il nous faut expliquer pourquoi la preuve de la simplicité de Diff$_0(\mathbf{S}^2, \mathit{aire})$ ne se généralise pas des difféomorphismes aux homéomorphismes.

La démonstration pour les difféomorphismes est fondée sur un théorème important de Herman pour les difféomorphismes du tore. Soit $\alpha \in \mathbf{R}^n$ un vecteur diophantien, c'est-à-dire mal approché par un vecteur rationnel (on pourra consulter Herman (1971) mais la définition importe peu ici). On considère la translation T_α du tore $\mathbf{R}^n/\mathbf{Z}^n$ de vecteur α. Le théorème affirme l'existence d'un voisinage $\mathscr{V}$ de T_α dans Diff$(\mathbf{R}^n/\mathbf{Z}^n)$ et d'une application $s\colon \mathscr{V} \to \mathrm{Diff}(\mathbf{R}^n/\mathbf{Z}^n) \times \mathbf{R}^n/\mathbf{Z}^n$ telle que pour tout $f \in \mathscr{V}$ avec $s(f) = (g, \beta)$, on a :

$$f = \left(gT_\alpha g^{-1}\right) T_{-\beta}.$$

Si f préserve le volume, il n'est pas difficile de vérifier qu'il en est de même pour g. Si le flux de f est nul, on a nécessairement $\alpha = \beta$ et on en conclut que f est un commutateur. Cela permet de montrer que $\mathrm{Diff}_0(\mathbf{R}^n/\mathbf{Z}^n), vol)$ est un groupe simple. On peut ensuite démontrer la simplicité de $\mathrm{Diff}_0(M, vol)$ pour une variété quelconque.

Ce théorème de Herman n'est pas valide pour les homéomorphismes, et on ne peut donc adapter cette preuve, ni pour démontrer le théorème de Fathi en dimension $n \geqslant 3$ (qui est pourtant vrai !) ni pour démontrer la simplicité de $\mathrm{Homéo}_0(\mathbf{S}^2, aire)$ (qui n'est pas vraie !).

En effet, considérons une petite boule B dans le tore $\mathbf{R}^n/\mathbf{Z}^n$ et un point x de l'intérieur de B. Soit y le point de premier retour de x dans B sous l'action des itérés de T_α. On peut alors trouver un homéomorphisme w à support dans B qui préserve le volume et qui envoie y sur x et on pose $f = wT_\alpha$ de sorte que x est un point périodique de f. Clairement le flux de w est nul et celui de f est donc égal à α. Notons que f est arbitrairement proche de T_α (en topologie C^0 qui est la seule disponible dans ce cas). Si le théorème de Herman était vrai dans cette situation topologique, on aurait $\beta = 0$ et f serait conjugué de T_α, ce qui contredit le fait que f possède un point périodique.

Après nous être convaincus que les théorèmes connus de Fathi et Herman ne peuvent pas se généraliser pour montrer la simplicité, nous pouvons maintenant décrire le théorème principal de cet exposé : $\mathrm{Homéo}_0(\mathbf{S}^2, aire)$ n'est pas simple.

4. Le théorème de D. Cristofaro-Gardiner, V. Humilière et S. Seyfaddini

4.1. Quelques outils

4.1.1. *Le théorème de Moser (1965).* — Ce théorème est l'analogue différentiable du théorème de Oxtoby–Ulam pour les mesures. Il affirme que deux formes de volume (de classe C^∞) sur la même variété compacte connexe et de même volume total sont images l'une de l'autre par un difféomorphisme isotope à l'identité. Plus précisément, la donnée d'une métrique riemannienne permet de construire une telle isotopie canonique si bien que le groupe de tous les difféomorphismes d'une variété compacte se rétracte par déformation sur le sous-groupe de ceux qui préservent une forme de volume fixée.

4.1.2. *La topologie de* $\mathrm{Difféo}(\mathbf{S}^2)$ *(Smale, 1959).* — Le groupe $\mathrm{Difféo}_+(\mathbf{S}^2)$ des difféomorphismes qui respectent l'orientation se rétracte par déformation sur le sous-groupe compact maximal $\mathrm{SO}(3)$. On dispose de plusieurs preuves de ce fait, mais la plus rapide est peut-être celle qui consiste à considérer l'espace des structures presque complexes positives sur la sphère, qui est contractile. Pour chaque point de

la sphère, il s'agit en effet de choisir un opérateur J de carré $-id$ dans l'espace tangent, tel que $aire(v, J(v)) > 0$ pour tout vecteur $v \neq 0$. L'espace de ces opérateurs est contractile (il s'identifie à un disque). On observe alors que $\mathrm{Difféo}_+(\mathbf{S}^2)$ opère transitivement sur cet espace de structures presque complexes : c'est une version du théorème classique d'uniformisation. Il en résulte que $\mathrm{Difféo}_+(\mathbf{S}^2)$ se rétracte sur le sous-groupe des difféomorphismes qui respectent une structure complexe, *i.e.* sur les biholomorphismes de $\mathbf{CP}^1$, c'est-à-dire sur $\mathrm{PGL}(2, \mathbf{C})$, et donc sur $\mathrm{SO}(3)$. *On en déduit que* $\mathrm{Difféo}_+(\mathbf{S}^2, aire)$ *se rétracte sur* $\mathrm{SO}(3)$ *et que* $\mathrm{Difféo}(\mathbf{D}^2, \partial\mathbf{D}^2, aire)$ *est contractile.* En particulier $\mathrm{Difféo}_0(\mathbf{S}^2, aire)$ est égal à $\mathrm{Difféo}_+(\mathbf{S}^2, aire)$. Le résultat analogue est valide pour les homéomorphismes (Kneser, 1926).

4.1.3. *Les hamiltoniens.* — Un élément de $\mathrm{Difféo}(\mathbf{S}^2, aire)$ ou de $\mathrm{Difféo}(\mathbf{D}^2, \partial\mathbf{D}^2, aire)$ est donc le temps 1 d'un champ de vecteur $(X_t)_{t\in[0,1]}$ qui dépend du temps et qui respecte l'aire. La 1-forme différentielle définie par $\beta_t(v) = aire(X_t, v)$ est alors fermée. C'est donc une forme exacte dH_t. La fonction H_t qui dépend de $t \in [0, 1]$, et qui est définie sur la sphère ou le disque à une constante additive près, est un *hamiltonien*. Pour cette raison on parle aussi des groupes de *difféomorphismes hamiltoniens*. Dans le cas du disque, on normalise H_t en demandant que sa restriction au bord est nulle. Dans le cas de la sphère, on demande que l'intégrale de H_t soit nulle. Si H_t est un hamiltonien on notera f_H^t l'isotopie qu'il définit.

Bien entendu pour un difféomorphisme hamiltonien f, l'isotopie f_H^t n'est pas unique et il faut fréquemment vérifier que certaines constructions sont indépendantes des choix. Par exemple, dans le cas du disque, on vérifie que l'intégrale $\int_0^1 \int_{\mathbf{D}^2} H_t \, aire \, dt$ ne dépend que de f et on s'assure facilement qu'on retrouve ainsi l'invariant de Calabi de f.

4.1.4. *La métrique de Hofer.* — Pour une fonction H définie sur la sphère ou sur le disque, on note $osc(H)$ la différence entre les valeurs maximales et minimales. La *norme de Hofer* de H_t est définie par

$$\|(H_t)_{t\in[0,1]}\|_{(1,\infty)} = \int_0^1 osc(H_t)\, dt.$$

Si f_1, f_2 sont deux difféomorphismes hamiltoniens, la *distance de Hofer* est la borne intérieure des normes de Hofer des hamiltoniens qui définissent $f_1 f_2^{-1}$. Un résultat difficile de Hofer (1990), Lalonde et McDuff (1995) et Polterovich (2001) affirme que ceci définit en effet une distance sur ces groupes de difféomorphismes, qui possède par ailleurs la propriété intéressante d'être invariante par translations à gauche et à droite. Rappelons que les seuls groupes de Lie connexes qui admettent une métrique bi-invariante sont les produits d'un groupe compact et d'un groupe abélien. Bien sûr, nos groupes de difféomorphismes hamiltoniens sont de dimension infinie et ne sont pas des groupes de Lie.

4.1.5. *Une première tentative infructueuse mais intéressante.* — Müller et Oh (2007) ont défini un sous-groupe distingué de $\mathrm{Homéo}_0(\mathbf{S}^2, aire)$, baptisé groupe des *haméomorphismes*, sans pourtant parvenir à montrer qu'il s'agit d'un sous-groupe propre, même si nous savons maintenant que c'est en effet le cas. Sa définition mérite cependant d'être mentionnée, comme motivation pour la suite. Soit f^t une isotopie dans $\mathrm{Homéo}_0(\mathbf{S}^2, aire)$ joignant l'identité f^0 à un homéomorphisme $f = f^1$. On dit qu'il s'agit d'une *haméotopie* s'il existe une suite de hamiltoniens *lisses* H_t^i ($i \geqslant 1, t \in [0,1]$) tels que d'une part $\|H^i - H^j\|_{(1,\infty)}$ tend vers 0 quand i, j tendent vers l'infini, et d'autre part les flots $f_{H^i}^t$ tendent vers f^t uniformément quand i tend vers l'infini. L'ensemble des homéomorphismes qui sont l'extrémité d'une telle haméotopie est le groupe $\mathrm{Haméo}(\mathbf{S}^2, aire)$ des haméomorphismes.

4.1.6. *Une tentative fructueuse.* — C'est en s'inspirant de diverses tentatives précédentes que D. Cristofaro-Gardiner, V. Humilière et S. Seyfaddini ont construit explicitement un sous-groupe distingué dans $\mathrm{Homéo}_0(\mathbf{S}^2, aire)$ et, surtout, ont montré qu'il s'agit en effet d'un sous-groupe propre.

On dit qu'un élément f de $\mathrm{Homéo}_0(\mathbf{S}^2, aire)$ est d'*énergie finie* [8] s'il est limite uniforme d'une suite f_i d'éléments de $\mathrm{Difféo}_0(\mathbf{S}^2, aire)$ bornée pour la distance de Hofer. Le résultat principal est le suivant.

L'ensemble $\mathrm{FHoméo}_0(\mathbf{S}^2, aire)$ *des homéomorphismes d'énergie finie est un sous-groupe distingué propre.*

Comme pour la tentative précédente, le fait qu'il s'agit d'un sous-groupe distingué est facile. La difficulté est d'exhiber un élément qui n'est pas d'énergie finie.

4.2. La stratégie

Le reste de cette section consiste à esquisser l'idée de la preuve. Les cas du disque et de la sphère suivent des stratégies différentes mais comme nous contentons de décrire le principe général, nous nous concentrons sur le cas du disque.

4.2.1. *Le candidat d'énergie infinie.* — Reprenons un difféomorphisme du disque de la forme $f\colon (r,\theta) \mapsto (r, \theta + u(r))$ où $u\colon [0,1] \to \mathbf{R}$ est nul près de 1, de façon à ce que f soit l'identité près du bord, et constant près de 0 pour que f soit une rotation au voisinage de l'origine. Il s'agit du temps 1 d'un hamiltonien $H\colon \mathbf{D}^2 \to \mathbf{R}$ autonome, qui ne dépend que de r avec $u(r) = (dH/dr)/2\pi r$. L'invariant de Calabi de f est l'intégrale de H sur le disque $\int_0^1 2\pi r H(r) dr$, soit $-2\pi^2 \int_0^1 r^3 u(r) dr$.

[8] Formellement, un hamiltonien a la dimension d'une énergie et la norme de Hofer celle d'une action : une énergie multipliée par un temps. Il serait donc préférable de parler d'homéomorphisme d'action finie.

Une fonction continue u sur l'intervalle $]0,1]$ définit un homéomorphisme f préservant l'aire, quel soit le comportement de u au voisinage de 0. Il est donc tentant de penser qu'un tel homéomorphisme pourrait ne pas être d'énergie finie si u est continue décroissante, nulle près de 1, et telle que l'intégrale $\int_0^1 r^3 u(r)dr$ diverge [(9)]. C'est en effet le cas comme nous allons le voir.

4.2.2. *Les invariants spectraux.* — Les invariants spectraux des difféomorphismes hamiltoniens furent d'abord introduits par Viterbo (1992) puis généralisés par Schwarz (2000) et Oh (2005).

Le cœur de la démonstration consiste à construire des *invariants spectraux* associés à un entier $d \geqslant 1$. Si $H_t\colon \mathbf{D}^2 \to \mathbf{R}$ est un hamiltonien dépendant du temps, nul près du bord, nous allons décrire des invariants $c_d(H_t)$ qui vérifient les propriétés suivantes.

1. $c_d(H_t)$ ne dépend que du difféomorphisme hamiltonien f_H^1 défini par H_t.
2. Continuité : $|c_d(H_t^1) - c_d(H_t^2)| \leqslant d\|H_t^1 - H_t^2\|_{(1,\infty)}$.
3. Monotonie : Si $H_t^1 \leqslant H_t^2$ pour tout t, on a $c_d(H_t^1) \leqslant c_d(H_t^2)$.

4.2.3. *La conjecture de Hutchings.* — Pour un élément f de Difféo$(\mathbf{D}^2, \partial\mathbf{D}^2, aire)$, *Hutchings conjecture que $c_d(f)/d$ converge vers l'invariant de Calabi de f quand d tend vers l'infini.* La question reste ouverte mais c'est en effet le cas lorsque f est un twist monotone associé à une fonction lisse $u\colon [0,1] \to \mathbf{R}$. Même si la preuve de cette dernière assertion n'est pas facile, on ne s'étonnera pas qu'il soit possible de tout calculer pour des difféomorphismes aussi explicites que ces twists.

4.2.4. *La première étape.* — On se fixe donc une fonction u lisse sur $]0,1]$, décroissante, nulle près de 1, et telle que l'intégrale $\int_0^1 r^3 u(r)dr$ diverge. Cela définit un élément f_u de Homéo$(\mathbf{D}^2, \partial\mathbf{D}^2, aire)$ dont il s'agit de montrer qu'il n'est pas dans le groupe FHoméo$(\mathbf{D}^2, \partial\mathbf{D}^2, aire)$ des homéomorphismes d'énergie finie.

La propriété 1 montre que c_d est en fait défini sur Difféo$(\mathbf{D}^2, \partial\mathbf{D}^2, aire)$. La propriété 2 permettra de montrer que *c_d se prolonge par continuité à* Homéo$(\mathbf{D}^2, \partial\mathbf{D}^2, aire)$. Elle permet en outre de montrer que $\limsup_{d\to\infty} c_d(f_H^1)/d < \infty$ pour tout difféomorphisme $f = f_H^1$. On montre que *cela se généralise aux homéomorphismes d'énergie finie :* $\limsup_{d\to\infty} c_d(f)/d < \infty$ *pour* $f \in$ FHoméo$(\mathbf{D}^2, \partial\mathbf{D}^2, aire)$.

[(9)] Pour le tourbillon ponctuel irrotationnel qu'on rencontre en dynamique des fluides on a $u(r) \propto r^{-2}$ si bien que l'intégrale converge même si la dynamique n'est pas lisse à l'origine.

4.2.5. *La deuxième étape.* — On considère une suite de fonctions lisses et décroissantes $(u_k)_{k\geqslant 1}$ définies sur $]0,1]$ et à valeurs réelles telles que :

1. $u_k(r)$ est constant dans un voisinage de 0 de sorte que f_{u_k} est un difféomorphisme,
2. $u_k(r) = u(r)$ pour $r \geqslant 1/k$,
3. $u_{k+1} \geqslant u_k$.

Par monotonie, $c_d(f_u) \geqslant c_d(f_{u_k})$. La conjecture de Hutchings, établie pour les twists, montre que pour tout k la suite $c_d(f_{u_k})/d$ tend vers $\mathscr{C}al(f_{u_k})$ lorsque d tend vers l'infini. Enfin $\mathscr{C}al(f_{u_k})$ tend vers l'infini quand k tend vers l'infini puisque $\int_0^1 r^3 u_k(r)dr$ tend vers l'infini. Il en résulte que $\limsup_{d\to\infty} c_d(f_u)/d = \infty$.

Par conséquent f_u n'est pas d'énergie finie, ce que nous voulions démontrer.

Il nous reste encore à définir les invariants spectraux et à montrer qu'ils vérifient les propriétés utilisées ci-dessus : il s'agit de l'essentiel du travail, que nous ne ferons pas...

4.3. L'homologie de Floer

4.3.1. *Un exemple jouet.* — Avant d'évoquer la construction des invariants c_d, considérons l'exemple classique suivant qui pourra servir de motivation. Soit $h\colon \Lambda \to \mathbf{R}$ une fonction continue définie sur une variété Λ, compacte par exemple, et choisissons une classe d'homologie d dans $H_\star(\Lambda, \mathbf{Z}/2\mathbf{Z})$ (d'un degré quelconque). Notons $\Lambda^{h\leqslant l} = h^{-1}(]-\infty, l])$ et $c_d(h)$ la borne inférieure des l tels que d est dans l'image de l'application naturelle de $H_\star(\Lambda^{h\leqslant l}, \mathbf{Z}/2\mathbf{Z})$ vers $H_\star(\Lambda, \mathbf{Z}/2\mathbf{Z})$. Autrement dit, $c_d(h)$ est le plus petit sous-niveau de h dans lequel on détecte la classe d'homologie d.

Les propriétés suivantes sont évidentes :

1. Si h est lisse, $c_d(h)$ est une valeur critique.
2. Continuité $|c_d(h_1) - c_d(h_2)| \leqslant \sup|h_1 - h_2|$.
3. Monotonie : Si $h_1 \leqslant h_2$ on a $c_d(h_1) \leqslant c_d(h_2)$.

4.3.2. *L'homologie de Morse.* — Avant d'expliquer comment l'homologie de Floer pourra être utilisée, et toujours comme motivation, voici comment on définit classiquement l'homologie de Morse, parfois appelée homologie de Morse–Smale–Thom–Witten (voir par exemple Laudenbach (2004)).

Soit $h\colon \Lambda \to \mathbf{R}$ une fonction de Morse sur une variété compacte. On choisit une métrique riemannienne générique sur Λ, ce qui permet de définir le champ de vecteurs gradient de h. On définit alors un complexe différentiel $(E_k)_{k\geqslant 0}$ d'espaces vectoriels sur $\mathbf{Z}/2\mathbf{Z}$ de la manière suivante. L'espace E_k est librement engendré par les points critiques x de h d'indice k. Le bord ∂x d'un point critique d'indice k est la

somme $\sum_y n_{x,y} y$ où y décrit les points critiques d'indice $k-1$ et $n_{x,y}$ est le nombre (modulo 2) de trajectoires (de l'opposé) du gradient tendant vers x et y lorsque le temps tend vers $\pm\infty$. Il se trouve que $\partial^2 = 0$ et que l'homologie de ce complexe ne dépend ni de la fonction de Morse ni de la métrique riemannienne : on obtient simplement l'homologie singulière usuelle (à coefficients dans $\mathbf{Z}/2\mathbf{Z}$).

4.3.3. *L'action.* — L'exemple jouet utilisait une variété, une fonction, et une classe d'homologie. Dans notre situation, un espace de lacets généralisera la variété et une fonctionnelle d'action jouera le rôle de la fonction. Commençons par simplifier le problème en remplaçant d'abord la sphère par une surface compacte S (munie d'une forme d'aire) de caractéristique d'Euler-Poincaré $\leqslant 0$ de sorte que $\pi_2(S) = 0$.

Considérons un difféomorphisme hamiltonien f_H^1 de S défini par un hamiltonien $H_t\colon S \to \mathbf{R}$ pour $t \in [0,1]$. Soit Λ l'espace des lacets $c\colon \mathbf{S}^1 \to S$ homotopiquement triviaux. Un tel lacet borde une application $\bar{c}\colon \mathbf{D}^2 \to S$. Nous noterons simplement $aire(c)$ l'intégrale $\int_{\mathbf{D}^2} \bar{c}^\star aire$, indépendante du choix de $\bar{c}$, parce que $\pi_2(S) = 0$. L'*action* est définie par

$$\mathscr{A}_H\colon c \in \Lambda \mapsto \int_0^1 H_t(c(t))dt - aire(c) \in \mathbf{R}.$$

Il est facile de vérifier que les points critiques de cette action sont précisément les orbites périodiques de période 1 du flot non autonome f_H^t engendré par H_t.

C'est l'espace Λ qui jouera le rôle de la variété dans l'exemple jouet et $\mathscr{A}_H$ jouera celui de la fonction de Morse.

4.3.4. *Les valeurs critiques.* — L'action $\mathscr{A}_H$ dépend du choix de l'hamiltonien H_t. Le difféomorphisme $f = f_H^1$ ne suffit pas pour la définir. Cependant les valeurs critiques ne dépendent presque pas de l'hamiltonien... Plus précisément, *si H_t et G_t sont deux hamiltoniens définissant le même $f = f_H^1 = f_G^1$, les ensembles des valeurs critiques de $\mathscr{A}_H$ et $\mathscr{A}_G$ coïncident à une translation près.*

La façon la plus simple de s'en convaincre consiste à considérer d'abord le cas où un hamiltonien H_t définit l'identité $f_H^1 = id$. Tous les lacets $t \in [0,1] \mapsto f_H^t(x)$ sont alors des points critiques. Par connexité, l'action $\mathscr{A}_H$ est constante sur tous ces lacets critiques.

Dans le cas général on définit la composition $G\#H$ de deux hamiltoniens H_t et G_t de la manière suivante $(G\#H)_t(x) = G_t(x) + H_t\left((f_G^t)^{-1}(x)\right)$. Le flot hamiltonien $f_{G\#H}^t$ associé à $G\#H$ est la composition $f_G^t f_H^t$. Si les deux hamiltoniens G, H ont le même temps 1, *i.e.* si $f = f_G^1 = f_H^1$, on pose $\bar{G}_t = -H_t(f_G^t(x))$ de sorte que le temps 1 de $\bar{G}\#H$ est l'identité. L'argument précédent, appliqué au cas $f_{\bar{G}\#H}^1 = id$, permet alors de conclure.

4.3.5. *La suspension et sa symplectisation.* — L'objet canoniquement associé à un difféomorphisme f de S est sa *suspension*, *i.e.* la variété Y_f de dimension 3, quotient de $S \times \mathbf{R}$ par l'action de $\mathbf{Z}$ dans laquelle k agit par $(x,t) \mapsto (f^k(x), t+k)$. Cette variété fibre sur le cercle $\mathbf{R}/\mathbf{Z}$ et les fibres sont difféomorphes à S. La donnée d'une isotopie joignant l'identité à f permet d'identifier le fibré Y_f au produit $S \times \mathbf{R}/\mathbf{Z}$. Réciproquement, une trivialisation définit une isotopie.

Le champ de vecteurs $\partial/\partial t$ passe au quotient en un champ R qu'on appelle parfois (à tort) le *champ de Reeb*. Ce champ est bien sûr transverse aux fibres et l'application de premier retour sur une fibre est (conjuguée à) f. La forme d'aire permet de construire une 2-forme fermée ω sur Y_f dont le noyau est précisément R.

La *symplectisation* de f est la variété $\mathbf{R} \times Y_f$, de dimension 4, munie de la forme symplectique $\omega + ds \wedge dt$ où s désigne la première coordonnée. Une *structure presque complexe* J sur $\mathbf{R} \times Y_f$ est *admissible* si

- $J(\partial/\partial s) = R$,
- J est invariant par translations sur le premier facteur,
- $(\omega + ds \wedge dt)(v, Jv) > 0$ pour tout vecteur non nul v.

4.3.6. *L'homologie symplectique de Floer.* — Voici une présentation rapide de l'homologie symplectique de Floer « classique »(Floer, 1989). Voir Laudenbach (2004) ou Audin et Damian (2010) pour une exposition détaillée. On suppose que tous les points fixes de f sont non-dégénérés, c'est-à-dire que 1 n'est pas valeur propre de la différentielle de f en un point fixe. La donnée d'une structure presque complexe admissible J permet de définir les lignes de gradient de l'action et de les interpréter comme des courbes pseudo-holomorphes dans la symplectisation $\mathbf{R} \times Y_f$. Il s'agit de chemins c_s dans Λ qui balaient un cylindre dans $\mathbf{R} \times Y_f$ dont le plan tangent est invariant par J. L'*homologie de Floer* est alors définie à partir d'un complexe de chaînes. Les générateurs sont les points fixes x de f. Soient x, y deux points fixes de f, considérés comme deux orbites fermées dans Y_f. On considère les cylindres pseudo-holomorphes dans $\mathbf{R} \times Y_f$ comme ci-dessus dont les projections dans Y_f convergent vers x et y lorsque s tend vers $\pm\infty$ respectivement. Pour un choix générique de J, l'espace de ces cylindres, quotienté par les translations verticales, est une variété compacte de dimension finie $\mathscr{M}(x,y)$. C'est l'analogue de l'espace des trajectoires de gradient d'une fonction de Morse, connectant deux points critiques. On définit alors le bord ∂x comme la somme $\sum_y n_{x,y} y$ où y décrit les points critiques tels que $\mathscr{M}(x,y)$ se réduit à un nombre fini de cylindres et $n_{x,y}$ est le nombre de ces cylindres (modulo 2).

Il se trouve que cette homologie de Floer est isomorphe à l'homologie usuelle de S.

4.3.7. *La sphère.* — Plaçons-nous maintenant dans le cas qui nous intéresse où S est la sphère $\mathbf{S}^2$. Une courbe fermée de Λ borde encore des disques mais ils ne sont plus uniques à homotopie près et l'aire $aire(c)$ n'est plus définie qu'à l'aire totale de la sphère près (qu'on suppose égale à 1). On ne dispose plus d'une action définie sur Λ à valeurs réelles, mais à valeurs dans le cercle $\mathbf{R}/\mathbf{Z}$.

Il est préférable de définir plutôt Λ comme l'espace des courbes $c\colon \mathbf{R} \to \mathbf{S}^2$ telles que $c(t+1) = f(c(t))$. Une variation infinitésimale de c est une section ξ de c^*TY_f telle que $\xi(t+1) = df(\xi(t))$. Un élément c de Λ définit une courbe $(t, c(t))$ dans $\mathbf{R} \times \mathbf{S}^2$, qui devient une courbe fermée $\tilde{c}$ dans Y_f. C'est alors la différentielle de l'action qui est bien définie comme une 1-forme fermée sur Λ

$$\alpha_c(\xi) = \int_0^1 \omega\left(\frac{d\tilde{c}}{dt}, \xi(t)\right) dt.$$

Les points critiques de α ne sont autres que les courbes c qui ne dépendent pas de t, c'est-à-dire les points fixes de f, ou encore les orbites périodiques de R de période 1.

4.4. L'homologie de Floer périodique

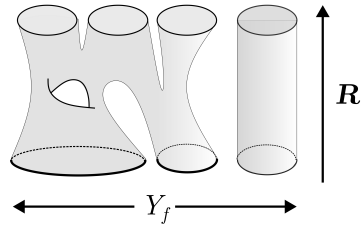
4.4.1. *L'homologie de Floer périodique.* — Nous introduisons maintenant l'homologie de Floer périodique, définie initialement par Hutchings et M. Sullivan (2005), qui dépend d'un entier $d \geqslant 1$ et qui se réduit à l'homologie de Floer classique lorsque $d = 1$. Au lieu de se limiter aux points fixes de f, nous allons utiliser des points périodiques, de période $\leqslant d$. On suppose que tous les points périodiques de f de période $k \leqslant d$ sont non dégénérés : 1 n'est pas valeur propre de la différentielle de f^k au point fixe.

Pour tout entier $d \geqslant 1$, on peut considérer l'hamiltonien H^d obtenu en composant H_t avec lui-même d fois, dont le temps 1 est bien sûr f_H^d. Les points critiques de l'action associée à H^d détectent alors les points périodiques de f_H de période d.

On définit un complexe de chaînes $C(f, d)$ qui est un espace vectoriel sur $\mathbf{Z}/2\mathbf{Z}$. Un générateur de $C(f, d)$ est une *d-multi-orbite*, c'est-à-dire un ensemble fini de la forme $\alpha = \{\alpha_i, m_i\}$ où les α_i sont des entiers $\geqslant 1$ et les m_i sont des orbites périodiques différentes de périodes k_i, tels que $\sum_i \alpha_i k_i = d$. Lorsque m_i est une orbite hyperbolique (*i.e.* la différentielle de f^{k_i} au point fixe est diagonalisable) on demande que $\alpha_i = 1$.

Si $\alpha = \{\alpha_i, m_i\}$ et $\beta = \{\beta_j, n_j\}$ sont deux générateurs, on considère l'ensemble $H_2(Y_f, \alpha, \beta)$ des classes d'homologie relatives Z telles que $\partial Z = \sum_i \alpha_i m_i - \sum_j \beta_j n_j$. Deux tels Z diffèrent d'un élément de $H_2(Y_f; Z) = H_2(\mathbf{S}^2 \times \mathbf{S}^1; \mathbf{Z}) = \mathbf{Z}$. L'étape suivante consiste à définir un indice $I(\alpha, \beta, Z) \in \mathbf{Z}$ purement topologique, dont nous ne donnerons pas la définition précise car elle est technique. Changer Z par une classe de $H_2(Y_f; Z)$ modifie cet indice par une quantité explicite.

Si α et β sont deux générateurs, on considère l'espace des courbes pseudo-holomorphes W dans $\mathbf{R} \times Y_f$ qui sont asymptotes à α quand s tend vers $+\infty$ et à β quand s tend vers $-\infty$ et telles que $I(\alpha, \beta, W) = 1$. Cet espace, modulo les translations en s, contient un nombre fini d'éléments $n_{\alpha,\beta}$ pour un J générique. On pose alors $\partial\alpha = \sum n_{\alpha,\beta}\beta$. Il se trouve que $\partial^2 = 0$ et que l'homologie ainsi définie est indépendante de J : on la note $PFH(Y_f, d)$ (pour Periodic Floer Homology). On note $PFH(Y_f) = \oplus_d PFH(Y_f, d)$.



Lee et Taubes (2012) montrent que cette homologie ne dépend ni de J ni du difféomorphisme f.

4.4.2. *L'homologie de Floer périodique modifiée.* — Nous ne sommes pas encore au bout de nos peines. Il nous faut une version modifiée de $PFH(Y_f)$ qui soit à la fois graduée et munie d'une fonctionnelle d'action.

On commence par fixer une section γ de la fibration $Y_f \to \mathbf{R}/\mathbf{Z}$. C'est une courbe fermée qui engendre $H_1(Y_f)$. Le complexe de chaînes $\widetilde{PFC}(f, d)$ est engendré par les couples (α, Z) où α est une multi-orbite comme plus haut et Z est une classe d'homologie relative dans $H_2(Y_f, \alpha, d\gamma)$. On définit un indice $I(\alpha, Z)$ par une formule analogue à celle qui définissait $I(\alpha, \beta, W)$ et... que nous n'avons pas décrite ! Il est alors facile de vérifier que $I(\alpha, Z) - I(\beta, Z') = I(\alpha, \beta, Z - Z')$.

Si $I(\alpha, Z) - I(\beta, Z') = 1$ on considère les courbes pseudo-holomorphes W dans $\mathbf{R} \times Y_f$ qui sont asymptotes à α et β quand s tend vers $\pm\infty$ et telles que $Z + [W] = Z'$. Comme précédemment, ces courbes, toujours aux translations verticales près, forment un ensemble fini contenant $n_{[(\alpha,Z),(\beta,Z')]}$ éléments ce qui permet de définir $\partial(\alpha, Z)$ comme précédemment. Cela définit $\widetilde{PFH}(f, d)$ muni d'une graduation grâce à l'indice.

Ce groupe d'homologie modifié ne dépend pas du choix de f et on peut le calculer aisément lorsque f est une rotation d'angle irrationnel de $\mathbf{S}^2$ qui n'a comme points périodiques que les deux points fixes. On trouve que $\widetilde{PFH}_\star(f, d)$ est nul si d et $\star$ ne sont pas de même parité et $\mathbf{Z}/2\mathbf{Z}$ s'ils sont de même parité.

Par ailleurs, l'intégrale de ω sur Z définit une action sur $\widetilde{PFC}(f, d)$, ce qui permet de définir une filtration. Pour tout $l \in \mathbf{R}$, on note $\widetilde{PFC}^l(f, d)$ le sous-espace de $\widetilde{PFC}(f, d)$ engendré par les (α, Z) dont l'action est $\leqslant l$. Comme ω est positif sur toute courbe pseudo-holomorphe, l'opérateur ∂ préserve $\widetilde{PFC}^l(f, d)$ et cela définit donc une homologie $\widetilde{PFH}^l(f, d)$.

4.4.3. *Les invariants spectraux.* — Nous pouvons enfin évoquer la définition des invariants spectraux. On souhaite définir $c_d(f)$ comme la borne inférieure des l tels que l'image injective de $\widetilde{PFH}_0^l(f,d) \to \widetilde{PFH}_0(f,d)$ contienne une classe σ donnée.

De nombreuses difficultés techniques subsistent. Par exemple, nous avons choisi arbitrairement une courbe γ qui engendre $H_1(Y_f)$ ou encore il faut choisir σ etc. Beaucoup de conventions n'ont pas été mentionnées.

Les invariants c_d ne sont définis pour l'instant que pour les difféomorphismes dont les points périodiques sont non dégénérés. Il faut les prolonger par continuité à tous les difféomorphismes, mais surtout aux homéomorphismes.

Une fois cette extension réalisée, on peut identifier le disque $\mathbf{D}^2$ à l'hémisphère nord de la sphère $\mathbf{S}^2$ de façon à plonger $\mathrm{Difféo}(\mathbf{D}^2, \partial\mathbf{D}^2, aire)$ dans $\mathrm{Difféo}_0(\mathbf{S}^2, aire)$. On obtient finalement

$$c_d \colon \mathrm{Difféo}(\mathbf{D}^2, \partial\mathbf{D}^2, aire) \to \mathbf{R}.$$

et des invariants spectraux analogues dans le cas de la sphère.

4.4.4. *Les invariants spectraux se prolongent aux homéomorphismes.* — C'est la preuve de l'extension aux homéomorphismes qui est la plus intéressante. On ne peut en présenter ici qu'une lointaine motivation : la *continuité des codes barres dans l'homologie persistante* (Le Roux, Seyfaddini et Viterbo, 2021).

Reprenons notre modèle jouet d'une fonction $h \colon M \to \mathbf{R}$ continue sur une variété fermée M. On peut considérer l'homologie E_l des sous-niveaux $h^{-1}(]-\infty, l])$ et pour $l_1 < l_2$ on dispose d'une application linéaire de E_{l_1} vers E_{l_2}. En termes pédants, on peut considérer la catégorie $\overrightarrow{\mathbf{R}}$ dont les objets sont les nombres réels l et pour laquelle il y a un unique morphisme de l_1 vers l_2 si $l_1 \leqslant l_2$. Les homologies des sous-niveaux définissent donc un foncteur de $\overrightarrow{\mathbf{R}}$ vers la catégorie des espaces vectoriels, par exemple sur $\mathbf{Z}/2\mathbf{Z}$.

Un exemple très simple est donné par un intervalle $I \subset \mathbf{R}$: l'espace vectoriel associé E_l est $\mathbf{Z}/2\mathbf{Z}$ si $l \in I$ et trivial sinon, et les morphismes sont évidents. Il se trouve que, sous une hypothèse très simple de finitude, tout foncteur de $\overrightarrow{\mathbf{R}}$ vers la catégorie des espaces vectoriels de dimension finie s'exprime de manière unique comme une somme directe de foncteurs associés à des intervalles I_k (Crawley-Boevey, 2015). Cette famille d'intervalles est le *code barre* associé à la fonction continue h.

La propriété remarquable (et élémentaire) est que le code barre dépend de manière continue de h. Le théorème de Cohen-Steiner, Edelsbrunner et Harer (2007) affirme que si $\sup |h_1 - h_2| \leqslant C$ les codes barres de h_1 et h_2 sont proches dans le sens précis suivant. On passe d'un code barre à l'autre d'une part en négligeant les intervalles de longueurs $\leqslant C$ et d'autre part en faisant glisser les extrémités des autres d'une distance inférieure à C.

Puisque le code barre détecte les valeurs critiques lorsque h est lisse et qu'il dépend continument de la fonction h, on peut penser que les invariants spectraux se prolongent par continuité aux homéomorphismes. Bien sûr, les invariants spectraux c_d sont définis à partir de valeurs d'une action et d'un sous-niveau dans l'homologie de Floer périodique qui n'est qu'un analogue de notre modèle jouet, et la preuve complète demande beaucoup de pages.

5. Nouveaux résultats

On assiste depuis peu à un florilège de prépublications annonçant des résultats complémentaires. Même s'il s'agit encore d'homologie de Floer, on en utilise d'autres versions. Au lieu d'expliciter un sous-groupe distingué on cherche plutôt à construire un homomorphisme surjectif à valeurs dans un groupe abélien.

5.1. Links lagrangiens

Dans la situation la plus simple de la sphère, on considère des *links lagrangiens* : il s'agit simplement d'une collection $L = \{L_i\}_{1\leqslant i\leqslant d}$ de courbes lisses, fermées, plongées et disjointes, dans la sphère, telle que toutes les composantes connexes du complémentaire de $\cup_i L_i$ ont la même aire. L'exemple le plus simple L_d consiste à choisir les bords de d disques plongés et disjoints dans la sphère, chacun d'aire $1/(d+1)$. L'idée d'utiliser ces links apparaissait déjà dans (Polterovich et Shelukhin, 2021) et (Mak et Smith, 2021).

Rappelons que la *puissance symétrique* $Sym^d(Q)$ d'un espace Q est le quotient de Q^d par l'action du groupe des permutations des coordonnées. Lorsque Q est une surface de Riemann, $Sym^d(Q)$ est une variété non singulière de dimension $2d$. Par exemple, $Sym^d(\mathbf{CP}^1)$ s'identifie naturellement à l'espace projectif complexe $\mathbf{CP}^d$. En effet, étant donné d droites dans $\mathbf{C}^2$, l'espace vectoriel des polynômes homogènes de degré d qui s'annulent précisément sur ces droites (avec multiplicité) est une droite dans un espace de dimension $d+1$.

5.2. De nouveaux invariants spectraux

Si L est un link lagrangien, le produit $\Pi_i L_i$ se plonge dans la puissance symétrique d-ème de la sphère, qui est la variété symplectique $\mathbf{CP}^d$. Un hamiltonien H_t sur la sphère définit un hamiltonien dans les puissances symétriques, en posant $\bar{H}_t(z_1,\dots,z_d) = \sum_i H_t(z_i)$. On dispose donc d'un hamiltonien et d'un tore lagrangien dans $\mathbf{CP}^d$. On peut alors utiliser l'*homologie de Floer lagrangienne* et définir des invariants spectraux $\gamma_d : \mathrm{Diffé o}_0(\mathbf{S}^2, aire) \to \mathbf{R}$ qui ont des propriétés analogues aux invariants précédents (même si le lien entre ces deux constructions n'est pas clair).

5.3. Les autres surfaces

Nous avons déjà signalé que Fathi a construit un homomorphisme surjectif appelé flux

$$\phi\colon \mathrm{Homéo}_0(M, vol) \to H_1(M, \mathbf{R})/\Phi(\pi_1(\mathrm{Homéo}(M, vol)))$$

et qu'il a montré qu'en dimension $\geqslant 3$ le noyau est simple. Nous savons maintenant que ce n'est pas le cas pour la sphère $\mathbf{S}^2$. Le cas de toutes les autres surfaces est traité dans Cristofaro-Gardiner, Humilière, Mak et al. (2022a) : *le noyau du flux n'est jamais simple pour une surface compacte.* Cela utilise aussi des links lagrangiens sur les surfaces.

5.4. Autres sous-groupes distingués

Le sous-groupe distingué $\mathrm{Haméo}(\mathbf{S}^2)$ introduit dans Müller et Oh (2007) *est contenu dans le groupe* $\mathrm{FHoméo}(\mathbf{S}^2)$ *des homéomorphismes d'énergie finie.* Buhovsky (2022) montre que cette inclusion est stricte.

Nous avons vu que la perfection entraîne souvent la simplicité. Par le même genre d'arguments, on montre que tout sous-groupe distingué non trivial de $\mathrm{Homéo}_0(\mathbf{S}^2, aire)$ contient le sous-groupe des commutateurs. Autrement dit, tout quotient non trivial de $\mathrm{Homéo}_0(\mathbf{S}^2, aire)$ est abélien. La nature du quotient (abélien) maximal est encore mystérieuse. Puisque le twist d'énergie infinie que nous avons décrit est évidemment contenu dans un groupe à un paramètre dont tous les éléments (non triviaux) sont d'énergie infinie, donc hors de $\mathrm{FHoméo}(\mathbf{S}^2)$, ce quotient contient des copies de $\mathbf{R}$. Une prépublication récente de Polterovich et Shelukhin (2021) construit explicitement d'énormes groupes abéliens quotients de $\mathrm{Homéo}(\mathbf{S}^2)$.

Dans le même ordre d'idées, on montre que *le groupe des commutateurs de* $\mathrm{Homéo}_0(\mathbf{S}^2, aire)$ *est simple.*

Comment ces sous-groupes distingués se comparent-ils avec ceux définis par les normes de fragmentation, décrits plus haut (Le Roux, 2010) ?

On peut identifier le disque ouvert muni d'une forme d'aire de masse finie au complémentaire du pôle sud dans la sphère. De cette façon, tout difféomorphisme du disque ouvert préservant l'aire peut être vu comme un homéomorphisme de la sphère. On peut en déduire le fait que nous avons déjà signalé : *le groupe des difféomorphismes du disque ouvert qui respectent l'aire n'est pas parfait*, répondant ainsi à une question de McDuff (1981).

Puisque nous savons que $\mathrm{Difféo}_0(\mathbf{S}^2, aire)$ est parfait, que tout sous-groupe distingué non trivial de $\mathrm{Homéo}_0(\mathbf{S}^2, aire)$ contient les commutateurs, et que les difféomorphismes respectant l'aire sont denses dans les homéomorphismes respectant l'aire, on conclut que $\mathrm{Homéo}_0(\mathbf{S}^2, aire)$ *est topologiquement simple*.

5.4.1. *Un groupe de Lie.* — Le fait que le groupe topologique $\mathrm{Homéo}_0(\mathbf{S}^2, aire)$ ne soit pas simple alors que son groupe des commutateurs est dense et simple peut paraître surprenant. Ce phénomène est cependant possible pour un groupe de Lie de dimension finie. Soit $\widetilde{\mathrm{SL}}(2,\mathbf{R})$ le revêtement universel de $\mathrm{SL}(2,\mathbf{R})$, dont le centre Z est isomorphe à $\mathbf{Z}$. Dans le produit $\widetilde{\mathrm{SL}}(2,\mathbf{R}) \times \mathrm{SO}(2)$, considérons le sous-groupe Δ, discret, cyclique et central, engendré par (z,r) où z est un générateur de Z et r est une rotation d'angle irrationnel (*i.e.* d'ordre infini dans $\mathrm{SO}(2)$). Le quotient $G = \widetilde{\mathrm{SL}}(2,\mathbf{R}) \times \mathrm{SO}(2)/\Delta$ est un groupe de Lie qui a a (presque) les propriétés requises. Son groupe dérivé $[G,G]$ est l'image de $\widetilde{\mathrm{SL}}(2,\mathbf{R}) \times \{id\}$ dans G. Ce groupe $[G,G]$ est dense dans G et le quotient $G/[G,G]$ est isomorphe au quotient de $\mathrm{SO}(2)$ par une rotation irrationnelle. D'autre part $[G,G]$ est un groupe de Lie simple (comme groupe de Lie : son seul sous-groupe distingué non trivial est son centre discret).

5.4.2. *Des quasi-plats.* — La géométrie de $\mathrm{Difféo}_0(\mathbf{S}^2, aire)$ équipé de la métrique de Hofer est également mystérieuse mais des progrès considérables ont été obtenus récemment.

Rappelons qu'une application $i\colon (E,d_1) \to (E_2,d_2)$ entre deux espaces métriques est un *plongement quasi-isométrique* s'il existe des constantes $a \geqslant 1, b \geqslant 0$ telles que

$$a^{-1}d_1(x,y) - b \leqslant d_2(i(x),i(y)) \leqslant ad_1(x,y) + b.$$

Polterovich (1998) démontra d'abord que le diamètre de la métrique de Hofer dans $\mathrm{Difféo}_0(\mathbf{S}^2, aire)$ est infini en construisant un plongement quasi-isométrique de $\mathbf{R}$. Puis Cristofaro-Gardiner, Humilière et Seyfaddini (2021) construisirent des plongements quasi-isométriques de $\mathbf{R}^k$ pour des valeurs de k arbitrairement grandes. Le cas des surfaces de genre $\geqslant 2$ avait été traité auparavant par Py (2008). Enfin Polterovich et Shelukhin (2021) montrèrent que *tout espace métrique séparable se plonge quasi-isométriquement dans* $\mathrm{Difféo}_0(\mathbf{S}^2, aire)$ *!*

5.5. L'invariant de Calabi

5.5.1. *Des quasi-morphismes.* — Un plongement d'un disque dans la sphère permet de considérer $\mathrm{Homéo}(\mathbf{D}^2, \partial\mathbf{D}^2, aire)$ comme un sous-groupe de $\mathrm{Homéo}_0(\mathbf{S}^2, aire)$. Cristofaro-Gardiner, Humilière et Seyfaddini (2021) montrent comment utiliser les invariants γ_d pour définir des quasi-morphismes homogènes

$$\bar{\gamma}_d\colon \mathrm{Homéo}(\mathbf{D}^2, \partial\mathbf{D}^2, aire) \to \mathbf{R},$$

dont les défauts tendent vers 0 quand d tend vers l'infini. Précisément

$$|\bar{\gamma}_d(f_1f_2) - \bar{\gamma}_d(f_1) - \bar{\gamma}_d(f_2)| \leqslant \frac{2}{d} \quad \text{pour tout } f_1, f_2 \in \mathrm{Homéo}(\mathbf{D}^2, \partial\mathbf{D}^2, aire)$$

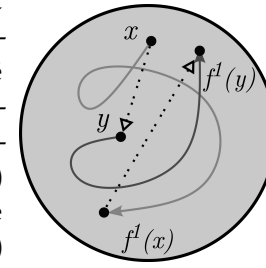
$$\bar{\gamma}_d(f^k) = k\bar{\gamma}_d(f^k) \text{ pour tout } f \in \mathrm{Homéo}(\mathbf{D}^2, \partial\mathbf{D}^2, aire).$$

Ils montrent par ailleurs que *si f est différentiable*, i.e. *dans* Difféo$(\mathbf{D}^2, \partial\mathbf{D}^2, aire)$, *la suite $\bar{\gamma}_d(f)$ tend vers l'invariant $\mathscr{C}al(f)$ quand d tend vers l'infini.* Cette dernière propriété est appelée « loi de Weyl » (pour une raison un peu confuse).

5.5.2. *Rotation de l'argument.* — L'invariant de Calabi est défini *a priori* pour un difféomorphisme mais pas pour un homéomorphisme. Fathi (1980a) demande s'il est possible d'en prolonger la définition aux homéomorphismes. Dans ce but, Fathi (1980b) (non publié) a proposé une définition plus topologique (voir aussi Ghys (2007)). Comme motivation, on peut citer le fait que Gambaudo et Ghys (1997) montrent que deux difféomorphismes conjugués par un homéomorphisme ont le même invariant de Calabi.

Le groupe Homéo$(\mathbf{D}^2, \partial\mathbf{D}^2, aire)$ est contractile.

Choisissons une isotopie f^t connectant $f^0 = id$ et $f^1 = f$. Si x_1, x_2 sont deux points distincts du disque, l'argument du vecteur $f^t(x_1) - f^t(x_2) \in \mathbf{R}^2 \setminus \{(0,0)\}$ tourne d'un certain angle $Angle(f; x_1, x_2)$ quand t varie de 0 à 1 (mesuré par exemple en tours). On s'assure facilement que cette définition est indépendante de l'isotopie choisie et se généralise aux homéomorphismes. La fonction $Angle(f; x_1, x_2)$ est définie hors de la diagonale dans $\mathbf{D}^2 \times \mathbf{D}^2$ mais elle est bornée si f est différentiable. Il se trouve que $\mathscr{C}al(f)$ est égal à (un multiple constant de) l'intégrale (double) de $Angle(f; x_1, x_2)$ sur $\mathbf{D}^2 \times \mathbf{D}^2$.



Pour prolonger l'invariant de Calabi aux homéomorphismes, il suffirait de construire une forme linéaire d'« intégration » définie sur les fonctions continues définies hors de la diagonale dans $\mathbf{D}^2 \times \mathbf{D}^2$, et invariante par l'action diagonale de Homéo$(\mathbf{D}^2, \partial\mathbf{D}^2, aire)$. Une telle forme existe-t-elle ? Dans l'esprit des moyennes à la Hahn-Banach dans les groupes moyennables ?

5.5.3. *La question de Fathi, résolue après quarante ans.* — Cristofaro-Gardiner, Humilière, Mak et al. (2022a) prolongent $\mathscr{C}al$ au groupe des haméomorphismes. Ce n'est pas facile, mais finalement pas très surprenant puisque les haméomorphismes sont définis par des hamiltoniens continus, qu'il « suffit » alors d'intégrer.

Ils démontrent ensuite que *l'invariant de Calabi se prolonge, d'une infinité de manières, à tous les homéomorphismes*, répondant ainsi positivement à la question posée par Fathi il y a plus de quarante ans (Cristofaro-Gardiner, Humilière, Mak et al., 2022b).

Considérons l'espace vectoriel $\mathbf{R}^{\mathbf{N}}$ des suites $(u_k)_{k\geqslant 0}$ de nombres réels et soit $\mathscr{R}$ son quotient par le sous-espace des suites qui tendent vers 0. L'application

$$f \in \text{Homéo}(\mathbf{D}^2, \partial\mathbf{D}^2, aire) \mapsto (\bar{\gamma}_1(f), \bar{\gamma}_2(f), \ldots) \in \mathbf{R}^{\mathbf{N}}$$

définit alors un homomorphisme $f \in$ Homéo$(\mathbf{D}^2, \partial\mathbf{D}^2, aire)$ vers $\mathscr{R}$ (qui s'avère surjectif).

Le lemme de Zorn permet alors d'affirmer l'existence de nombreux homomorphismes $lim : \mathscr{R}$ vers $\mathbf{R}$ qui valent l sur la suite constante $u_k = l$. Les composés $lim \circ \mathscr{C}$ sont tous des prolongements de $\mathscr{C}al$ grâce à la loi de Weyl. Notons que $\mathscr{R}$ est un $\mathbf{R}$-espace vectoriel, isomorphe à $\mathbf{R}$, *comme groupe abélien abstrait*.

Dans certains modèles de la théorie des ensembles, l'axiome du choix est faux et tout homomorphisme entre groupes polonais est nécessairement continu (Rosendal, 2019). Comme nous savons que $\mathscr{C}al$ n'a pas d'extension continue, on comprend qu'il est illusoire de donner un sens « concret » à ces prolongements.

6. Quatre problèmes

6.1. Différentiabilité

Il y a bien sûr beaucoup d'intermédiaires entre les homéomorphismes et les difféomorphismes de classe C^∞. La composante neutre du groupe des difféomorphismes de classe C^r $(0 \leqslant r \leqslant \infty)$ d'une variété (compacte sans bord pour simplifier) de dimension n est un groupe simple si $r \neq n+1$ (Mather, 1975). Le cas $r = n+1$ est totalement mystérieux, même pour $n = 1$. Dans un cas analogue, Mather (1985) montre un exemple surprenant : le groupe des difféomorphismes du cercle qui respectent l'orientation et dont le logarithme de la dérivée est à variation bornée n'est pas un groupe parfait. La situation n'est que partiellement comprise pour les difféomorphismes analytiques réels : Tsuboi (2009) montre que la composante neutre est simple dans un certain nombre d'exemples, dont les sphères, mais le cas général reste ouvert. *Que dire du groupe des difféomorphismes de la sphère* $\mathbf{S}^2$ *de classe* C^r *qui respectent l'aire ?* Un autre cas intéressant est celui des homéomorphismes affines par morceaux du plan, à support compact, qui respectent l'aire. Greenberg (1998) établit un lien avec la K-théorie algébrique.

6.2. Les variétés topologiques symplectiques

On peut définir les *variétés topologiques symplectiques* en demandant que les changements de cartes soient des homéomorphismes symplectiques, limites uniformes de difféomorphismes symplectiques. Sont-elles significativement différentes des variétés symplectiques usuelles ? La caractéristique la plus naïve d'une variété symplectique compacte de dimension $2n$ est de posséder une classe de cohomologie de degré 2 dont la puissance n-ème est non nulle. Est-ce encore vrai dans le cas topologique ? *La sphère de dimension* 4 *est-elle une variété topologique symplectique ?* Cette question fascinante a été posée par Hofer dans les années 80.

Des questions tout à fait analogues se posent dans le cadre des *homéomorphismes de contact* pour lesquels on ne connaît que peu de choses.

6.3. La topologie du classifiant

Dans les années 70, l'étude des groupes de difféomorphismes était avant tout motivée par la théorie des feuilletages. Thurston avait en effet montré un lien étroit entre la topologie du *classifiant de Haefliger des feuilletages* de codimension n et l'homologie du groupe des difféomorphismes de $\mathbf{R}^n$, considéré comme groupe discret : voir par exemple (Sergeraert, 1979) ou (Lawson, 1977). La perfection d'un groupe est équivalente à l'annulation de son premier groupe d'homologie. Les groupes d'homologie d'ordre supérieur ont été largement étudiés, en particulier dans leurs liens avec la cohomologie de Gelfand-Fuchs, mais il reste beaucoup de questions ouvertes (et difficiles). Dans le cadre topologique, Mather (1971) a montré par exemple que le groupe des homéomorphismes de $\mathbf{R}^n$ à support compact est acyclique et cela entraîne que le classifiant correspondant est contractile. *Que dire du classifiant du pseudogroupe des homéomorphismes symplectiques ? ou simplement qui préservent le volume ?* Dans le contexte symplectique différentiable, on dispose de quelques informations sur le classifiant (voir par exemple Kotschick et Morita, 2007 ; McDuff, 1983).

6.4. L'hélicité

L'*hélicité* est un invariant numérique d'un champ de vecteurs non singulier sur $\mathbf{S}^3$ qui préserve le volume, introduit par Arnold sous le nom d'*invariant de Hopf* (Arnold, 1986 ; Arnold et Khesin, 1998). Il mesure le nombre d'enlacement asymptotique moyen de deux trajectoires. Gambaudo et Ghys (1997) ont montré que dans de nombreuses situations, l'hélicité se ramène à l'invariant de Calabi de certaines applications de premier retour sur certaines sections transversales. Maintenant que l'invariant de Calabi a été étendu aux homéomorphismes, on peut espérer répondre positivement à la question suivante, question qui préoccupe l'auteur de ce texte depuis bien trop longtemps. *Peut-on définir naturellement une hélicité* (*peut-être à valeurs dans* $\mathscr{R}$) *pour un flot topologique sans points fixes, qui préserve le volume sur* $\mathbf{S}^3$*, de façon à ce que cette définition prolonge celle d'Arnold et qu'elle soit invariante par conjugaison topologique* (*préservant le volume*) *?*

Références

Alexander, J. W. (1923). « On the deformation of an n-cell », *Proceedings of the National Academy of Sciences of the United States of America* **9** (9), p. 406-407.

Anderson, R. D. (1958). « The algebraic simplicity of certain groups of homeomorphisms », *Amer. J. Math.* **80**, p. 955-963.

Arnold, V. (1986). « The asymptotic Hopf invariant and its applications », *Selecta Math. Soviet.* **5** (4). Selected translations, p. 327-345.

Arnold, V. et Khesin, B. (1998). *Topological methods in hydrodynamics*. T. 125. Applied Mathematical Sciences. Springer-Verlag, New York, p. xvi+374.

Audin, M. et Damian, M. (2010). *Théorie de Morse et homologie de Floer*. Savoirs Actuels. EDP Sciences, Les Ulis ; CNRS Éditions, Paris, p. xii+548.

Banyaga, A. (1978). « Sur la structure du groupe des difféomorphismes qui préservent une forme symplectique », *Comment. Math. Helv.* **53** (2), p. 174-227.

——— (1997). *The structure of classical diffeomorphism groups*. T. 400. Mathematics and its Applications. Kluwer Academic Publishers Group, Dordrecht, p. xii+197.

Brown, M. (1960). « A proof of the generalized Schoenflies theorem », *Bull. Amer. Math. Soc.* **66**, p. 74-76.

——— (1962). « A mapping theorem for untriangulated manifolds ». In : *Topology of 3-manifolds and related topics* (*Proc. The Univ. of Georgia Institute, 1961*). Prentice-Hall, Englewood Cliffs, N.J., p. 92-94.

Buhovsky, L. (2022). *On two remarkable groups of area-preserving homeomorphisms*. eprint : https://arxiv.org/abs/2204.08020.

Calabi, E. (1970). « On the group of automorphisms of a symplectic manifold », in : *Problems in analysis* (*Sympos. in honor of Salomon Bochner, Princeton Univ., Princeton, N.J., 1969*). Princeton Univ. Press, Princeton, N.J., p. 1-26.

Cohen-Steiner, D., Edelsbrunner, H. et Harer, J. (2007). « Stability of persistence diagrams », *Discrete Comput. Geom.* **37** (1), p. 103-120.

Crawley-Boevey, W. (2015). « Decomposition of pointwise finite-dimensional persistence modules », *J. Algebra Appl.* **14** (5), p. 1550066, 8.

Cristofaro-Gardiner, D., Humilière, V., Mak, C. Y. et al. (2022a). « Quantitative Heegaard Floer cohomology and the Calabi invariant », *Forum Math. Pi* **10**, Paper No. e27, 59.

——— (2022b). *Subleading asymptotics of link spectral invariants and homeomorphism groups of surfaces*. eprint : https://arxiv.org/abs/2206.10749.

Cristofaro-Gardiner, D., Humilière, V. et Seyfaddini, S. (2020). « Proof of the simplicity conjecture », *Ann. of Math.*, à paraître.

——— (2021). *PFH spectral invariants on the two-sphere and the large scale geometry of Hofer's metric*. eprint : https://arxiv.org/abs/2102.04404.

Entov, M., Polterovich, L. et Py, P. (2012). « On continuity of quasimorphisms for symplectic maps », in : *Perspectives in analysis, geometry, and topology*. T. 296. Progr. Math. With an appendix by Michael Khanevsky. Birkhäuser/Springer, New York, p. 169-197.

Epstein, D. B. A. (1970). « The simplicity of certain groups of homeomorphisms », *Compositio Math.* **22**, p. 165-173.

Fathi, A. (1980a). « Structure of the group of homeomorphisms preserving a good measure on a compact manifold », *Ann. Sci. École Norm. Sup.* (4) **13** (1), p. 45-93.

——— (1980b). « Transformations et homéomorphismes préservant la mesure, Systèmes dynamiques minimaux ». Thèse, Orsay.

Fisher, G. M. (1960). « On the group of all homeomorphisms of a manifold », *Trans. Amer. Math. Soc.* **97**, p. 193-212.

Floer, A. (1989). « Symplectic fixed points and holomorphic spheres », *Comm. Math. Phys.* **120** (4), p. 575-611.

Gambaudo, J.-M. et Ghys, É. (1997). « Enlacements asymptotiques », *Topology* **36** (6), p. 1355-1379.

Ghys, É. (2007). « Knots and dynamics », in : *International Congress of Mathematicians. Vol. I*. Eur. Math. Soc., Zürich, p. 247-277.

Greenberg, P. (1998). « Area preserving pl homeomorphisms and relations in K_2 », *Ann. Inst. Fourier* (*Grenoble*) **48** (1), p. 133-148.

Gromov, M. (1987). « Soft and hard symplectic geometry ». In : *Proceedings of the International Congress of Mathematicians, Vol. 1, 2* (*Berkeley, Calif., 1986*). Amer. Math. Soc., Providence, RI, p. 81-98.

Herman, M.-R. (1971). « Simplicité du groupe des difféomorphismes de classe C^∞, isotopes à l'identité, du tore de dimension n », *C. R. Acad. Sci. Paris Sér. A* **273**, p. 232-234.

——— (1973). « Sur le groupe des difféomorphismes du tore », *Ann. Inst. Fourier* (*Grenoble*) **23** (2), p. 75-86.

Hofer, H. (1990). « On the topological properties of symplectic maps », *Proc. Roy. Soc. Edinburgh Sect. A* **115** (1-2), p. 25-38.

Hutchings, M. et Sullivan, M. (2005). « The periodic Floer homology of a Dehn twist », *Algebr. Geom. Topol.* **5**, p. 301-354.

Kirby, R. C. (1969). « Stable homeomorphisms and the annulus conjecture », *Ann. of Math.* (2) **89**, p. 575-582.

Kneser, H. (1926). « Die Deformationssätze der einfach zusammenhängenden Flächen », *Math. Z.* **25** (1), p. 362-372.

Kotschick, D. et Morita, S. (2007). « Characteristic classes of foliated surface bundles with area-preserving holonomy », *J. Differential Geom.* **75** (2), p. 273-302.

Lalonde, F. et McDuff, D. (1995). « Hofer's L^∞-geometry : energy and stability of Hamiltonian flows. I, II », *Invent. Math.* **122** (1), p. 1-33, 35-69.

Laudenbach, F. (2004). « Symplectic geometry and Floer homology », in : *Symplectic geometry and Floer homology. A survey of the Floer homology for manifolds with contact type boundary or symplectic homology*. T. 7. Ensaios Mat. Soc. Brasil. Mat., Rio de Janeiro, p. 1-50.

Lawson Jr., H. B. (1977). *The quantitative theory of foliations*. Conference Board of the Mathematical Sciences Regional Conference Series in Mathematics, No. 27. Expository lectures from the CBMS Regional Conference held at Washington

University, St. Louis, Mo., January 6–10, 1975. American Mathematical Society, Providence, R.I., p. v+65.

Le Roux, F. (2010). « Simplicity of Homeo($\mathbb{D}^2, \partial\mathbb{D}^2$, Area) and fragmentation of symplectic diffeomorphisms », *J. Symplectic Geom.* **8** (1), p. 73-93.

Le Roux, F., Seyfaddini, S. et Viterbo, C. (2021). « Barcodes and area-preserving homeomorphisms », *Geom. Topol.* **25** (6), p. 2713-2825.

Lee, Y.-J. et Taubes, C. H. (2012). « Periodic Floer homology and Seiberg-Witten-Floer cohomology », *J. Symplectic Geom.* **10** (1), p. 81-164.

Mak, C. Y. et Smith, I. (2021). « Non-displaceable Lagrangian links in four-manifolds », *Geom. Funct. Anal.* **31** (2), p. 438-481.

Mann, K. (2016). « A short proof that $\mathrm{Diff}_c(M)$ is perfect », *New York J. Math.* **22**, p. 49-55.

——— (2021). « The structure of homeomorphism and diffeomorphism groups », *Notices Amer. Math. Soc.* **68** (4), p. 482-492.

Mather, J. N. (1971). « The vanishing of the homology of certain groups of homeomorphisms », *Topology* **10**, p. 297-298.

——— (1975). « Commutators of diffeomorphisms. II », *Comment. Math. Helv.* **50**, p. 33-40.

——— (1985). « Commutators of diffeomorphisms. III. A group which is not perfect », *Comment. Math. Helv.* **60** (1), p. 122-124.

McDuff, D. (1980). « On the group of volume-preserving diffeomorphisms of $\mathbf{R}^n$ », *Trans. Amer. Math. Soc.* **261** (1), p. 103-113.

——— (1981). « On groups of volume-preserving diffeomorphisms and foliations with transverse volume form », *Proc. London Math. Soc.* (3) **43** (2), p. 295-320.

——— (1983). « Local homology of groups of volume-preserving diffeomorphisms. III », *Ann. Sci. École Norm. Sup.* (4) **16** (4), 529-540 (1984).

McDuff, D. et Salamon, D. (2017). *Introduction to symplectic topology*. 3rd edition. Oxford Graduate Texts in Mathematics. Oxford University Press, Oxford, p. xi+623.

Moser, J. (1965). « On the volume elements on a manifold », *Trans. Amer. Math. Soc.* **120**, p. 286-294.

Müller, S. et Oh, Y.-G. (2007). « The group of Hamiltonian homeomorphisms and C^0-symplectic topology », *J. Symplectic Geom.* **5** (2), p. 167-219.

Oh, Y.-G. (2005). « Construction of spectral invariants of Hamiltonian paths on closed symplectic manifolds », in : *The breadth of symplectic and Poisson geometry*. T. 232. Progr. Math. Birkhäuser Boston, Boston, MA, p. 525-570.

Oxtoby, J. C. et Ulam, S. (1941). « Measure-preserving homeomorphisms and metrical transitivity », *Ann. of Math.* (2) **42**, p. 874-920.

Polterovich, L. (1998). « Hofer's diameter and Lagrangian intersections », *Internat. Math. Res. Notices* (4), p. 217-223.

——— (2001). *The geometry of the group of symplectic diffeomorphisms*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, p. xii+132.

Polterovich, L. et Shelukhin, E. (2021). *Lagrangian configurations and Hamiltonian maps*. eprint : https://arxiv.org/abs/2102.06118.

Postnikov, M. (1985). *Leçons de géométrie*. Traduit du Russe par Djilali Embarek. Groupes et algèbres de Lie. "Mir", Moscou, p. 375.

Py, P. (2008). « Quelques plats pour la métrique de Hofer », *J. Reine Angew. Math.* **620**, p. 185-193.

Quinn, F. (1982). « Ends of maps. III. Dimensions 4 and 5 », *J. Differential Geometry* **17** (3), p. 503-521.

Rosendal, C. (2019). « Continuity of universally measurable homomorphisms », *Forum Math. Pi* **7**, e5, 20.

Rousseau, G. (1978). « Difféomorphismes d'une variété symplectique non compacte », *Comment. Math. Helv.* **53** (4), p. 622-633.

Schreier, O. et Ulam, S. (1934). « Eine Bemerkung über die Gruppe der topolischen Abbildungen der Kreislinie auf sich selbst », *Studia Math* **5**, p. 155-159.

Schwarz, M. (2000). « On the action spectrum for closed symplectically aspherical manifolds », *Pacific J. Math.* **193** (2), p. 419-461.

Sergeraert, F. (1977). « Feuilletages et difféomorphismes infiniment tangents à l'identité », *Invent. Math.* **39** (3), p. 253-275.

——— (1979). « $B\Gamma$ [d'après John N. Mather et William Thurston] », in : *Séminaire Bourbaki, 30e année (1977/78)*. T. 710. Lecture Notes in Math. Springer, Berlin, Exp. No. 524, p. 300-315.

Smale, S. (1959). « Diffeomorphisms of the 2-sphere », *Proc. Amer. Math. Soc.* **10**, p. 621-626.

Sullivan, D. (2011). « Algebra, topology and algebraic topology of 3D ideal fluids », in : *Low-dimensional and symplectic topology*. T. 82. Proc. Sympos. Pure Math. Amer. Math. Soc., Providence, RI, p. 1-7.

Thurston, W. (1973). « On the structure of volume preserving diffeomorphisms ». Non publié.

——— (1974). « Foliations and groups of diffeomorphisms », *Bull. Amer. Math. Soc.* **80**, p. 304-307.

Tsuboi, T. (2009). « On the group of real analytic diffeomorphisms », *Ann. Sci. Éc. Norm. Supér. (4)* **42** (4), p. 601-651.

Ulam, S. (1935). *The Scottish Book*. 2nd edition. Mathematics from the Scottish Café with selected problems from the new Scottish Book, Including selected papers presented at the Scottish Book Conference held at North Texas University, Denton, TX, May 1979. Birkhäuser/Springer, Cham, p. xvii+322.

Ulam, S. et von Neumann, J. (1947). « On the group of homeomorphisms of the surface of a sphere (abstract) », *Bull. AMS* **53**, p. 506.

VITERBO, C. (1992). « Symplectic topology as the geometry of generating functions », *Math. Ann.* **292** (4), p. 685-710.

Étienne Ghys

UMPA, ENS Lyon,
46 Allée d'Italie, 69364 Lyon
et
Académie des sciences,
23 quai de Conti, 75006 Paris

E-mail : etienne.ghys@ens-lyon.fr

Séminaire BOURBAKI Avril 2023
74e année, 2022–2023, no 1205, p. 285 à 363
doi: 10.24033/ast.1214

POINTWISE CONVERGENCE FOR THE SCHRÖDINGER EQUATION
[*after* Xiumin Du and Ruixiang Zhang]

by **Jonathan Hickman**

1. Introduction: the Carleson problem

1.1. Solutions to the Schrödinger equation

Suitably normalised, the free Schrödinger equation on $\mathbb{R}^n$ is the second order partial differential equation

$$iu_t - \Delta_x u = 0. \tag{1}$$

Here u is a complex-valued function of the space-time variables $(x,t) \in \mathbb{R}^n \times \mathbb{R}$, whilst u_t and $\Delta_x u$ denote the first order time derivative and spatial Laplacian, respectively. We are interested in the Cauchy problem for this equation, whereby we specify an initial datum f and wish to solve

$$\begin{cases} iu_t - \Delta_x u = 0, \\ u(x,0) = f(x) \end{cases} \qquad (x,t) \in \mathbb{R}^n \times \mathbb{R}. \tag{2}$$

Depending on our hypotheses on f, what it means for u to be a 'solution' to the equation (2) varies. Here we consider two examples:

Classical solution. If f is sufficiently regular, then elementary Fourier transform methods show that (2) has a unique solution in the classical sense.[(1)] For instance, if we assume $f \in \mathscr{S}(\mathbb{R}^n)$, the Schwartz space, then the unique solution is given by

$$u(x,t) := e^{it\Delta} f(x)$$

where $e^{it\Delta}$ is the *Schrödinger propagator*

$$e^{it\Delta} f(x) := \frac{1}{(2\pi)^n} \int_{\widehat{\mathbb{R}}^n} e^{i(x\cdot\xi + t|\xi|^2)} \hat{f}(\xi)\, \mathrm{d}\xi. \tag{3}$$

[(1)] In particular, the derivatives u_t and $\Delta_x u$ are all well-defined in the usual sense from calculus, and the identities in (2) hold pointwise.

Note that the regularity – or smoothness – of the initial datum f is crucial to these observations. Indeed, the smoothness of f directly translates into the decay of the Fourier transform $\hat{f}(\xi)$ as $|\xi| \to \infty$. This decay ensures the integral in (3) is well-defined and also allows one to pass the derivatives inside the integral in order to verify (1).

L^2 solution. Now suppose $f \in L^2(\mathbb{R}^n)$, without any additional regularity assumptions. In this case, Plancherel's theorem allows us to define the Fourier transform $\hat{f}$ as a function in $L^2(\widehat{\mathbb{R}}^n)$, but in general we cannot conclude that $\hat{f}$ is integrable. Consequently, the integral formula (3) is not well-defined in the classical sense.

To circumvent these issues, we further appeal to the L^2 theory of Fourier transform. Note that the propagator $e^{it\Delta}$ introduced above can be interpreted as a linear operator on $\mathscr{S}(\mathbb{R}^n)$ which, given an initial datum f, outputs the solution at time t. Using Plancherel's theorem, we can extend $e^{it\Delta}$ to a Fourier multiplier operator acting on the whole of $L^2(\mathbb{R}^n)$. In particular, we define

$$e^{it\Delta}f := \mathscr{F}^{-1}\big(e^{it|\cdot|^2} \cdot \mathscr{F}f\big) \qquad \text{for } f \in L^2(\mathbb{R}^n),$$

where here $\mathscr{F}$ denotes the Fourier transform acting on $L^2(\mathbb{R}^n)$. Furthermore, this operator is an isometry of the L^2 space, in the sense that

$$\|e^{it\Delta}f\|_{L^2(\mathbb{R}^n)} = \|f\|_{L^2(\mathbb{R}^n)} \qquad \text{for all } f \in L^2(\mathbb{R}^n) \text{ and all } t \in \mathbb{R}; \tag{4}$$

this identity is typically referred to as *conservation of energy*.

As before, we may define

$$u(x,t) := e^{it\Delta}f(x),$$

but in general this is no longer a classical solution to the Schrödinger equation: for instance, for a fixed time t, the best we can say about $u(\,\cdot\,,t)$ is that it belongs to $L^2(\mathbb{R}^n)$ and so the Laplacian $\Delta_x u$ is not defined in the classical sense. However, we can interpret u as a solution to (1) in the sense of distributions. Indeed, using (4) it is not difficult to show u defines a distribution in $\mathscr{S}'(\mathbb{R}^{n+1})$ and so $\partial_t u$ and $\Delta_x u$ can be understood in the distributional sense. Furthermore, a simple Fourier analytic argument shows $\langle i\partial_t u - \Delta_x u, \phi\rangle = 0$ for all test functions $\phi \in \mathscr{S}(\mathbb{R}^{n+1})$.

1.2. The Carleson problem

Once a solution u to (2) has been constructed, it is natural to investigate the behaviour of u and how it relates to the initial datum f. There is a huge variety of different questions one can ask in this direction. Here we are interested in the classical *Carleson problem*, which aims to understand whether the initial datum can be recovered as a pointwise limit of the solution.

First consider the case where $f \in \mathscr{S}(\mathbb{R}^n)$, so that the solution $u(x,t) := e^{it\Delta}f(x)$ is classically defined. By definition, we know the solution u satisfies $u(x,0) = f(x)$ and is differentiable, and therefore continuous, with respect to t. In particular,

$$\lim_{t\to 0_+} e^{it\Delta}f(x) = f(x) \qquad \text{for all } x \in \mathbb{R}^n. \tag{5}$$

The Carleson problem asks to what extent this elementary limit identity continues to hold when we consider more general L^2 solutions to the Schrödinger equation.

Since an L^2 function is only defined almost everywhere, in order to make sense of the problem for general initial data in $L^2(\mathbb{R}^n)$ it is necessary to weaken the requirement that convergence holds for *all* $x \in \mathbb{R}^n$ in (5) to *almost all* $x \in \mathbb{R}^n$. That is, given $f \in L^2(\mathbb{R}^n)$ we wish to determine whether

$$\lim_{t\to 0_+} e^{it\Delta}f(x) = f(x) \qquad \text{for almost every } x \in \mathbb{R}^n. \tag{6}$$

Nevertheless, it is still unclear how to precisely interpret the above limit, since for every time slice t (belonging to the *continuum* $[0,1]$, say) we have a choice of representation for $e^{it\Delta}f$. We shall gloss over these technicalities for now and return to them in §3.2 below.

It is not difficult to show that the limit holds in the L^2-sense: that is, given $f \in L^2(\mathbb{R}^n)$ we have

$$\lim_{t\to 0_+} \|e^{it\Delta}f - f\|_{L^2(\mathbb{R}^n)} = 0. \tag{7}$$

Indeed, this can be easily verified for $f \in \mathscr{S}(\mathbb{R}^n)$ using the integral formula (3) for the propagator and the dominated convergence theorem. One can then pass to general $f \in L^2(\mathbb{R}^n)$ via density, using the conservation of energy identity (4).

On the other hand, there are examples of $f \in L^2(\mathbb{R}^n)$ for which (6) in fact **fails** (see §1.3 below). Thus, we are interested in determining an additional hypothesis on f under which the above norm convergence (7) can be 'upgraded' to almost everywhere convergence. Contrasting the situation for $f \in \mathscr{S}(\mathbb{R}^n)$ with that for general $f \in L^2(\mathbb{R}^n)$, it is natural that the additional hypothesis should enforce some degree of regularity on the initial datum.

The above considerations lead us to consider the Sobolev spaces $H^s(\mathbb{R}^n)$. Roughly speaking, $H^s(\mathbb{R}^n)$ consists of all $f \in L^2(\mathbb{R}^n)$ with derivatives up to order s lying also in $L^2(\mathbb{R}^n)$. More precisely,

$$H^s(\mathbb{R}^n) := \big\{ f \in L^2(\mathbb{R}^n) : (1-\Delta_x)^{s/2} f \in L^2(\mathbb{R}^n) \big\}, \qquad s \geqslant 0,$$

where $(1-\Delta)^{s/2}$ denotes the fractional differential operator, defined in terms of the Fourier transform $\mathscr{F}$ now acting on the space of distributions $\mathscr{S}'(\mathbb{R}^n)$ by

$$(1-\Delta_x)^{s/2} f := \mathscr{F}^{-1}\big((1+|\cdot|^2)^{s/2} \cdot \mathscr{F} f\big).$$

In particular, given $f \in L^2(\mathbb{R}^n)$, we can always make sense of the fractional derivative $(1-\Delta_x)^{s/2} f$ as a distribution, and $f \in H^s(\mathbb{R}^n)$ if this distribution coincides with an L^2 function. It is clear from the definitions that

$$H^0(\mathbb{R}^n) = L^2(\mathbb{R}^n) \quad \text{and} \quad H^{s_1}(\mathbb{R}^n) \supseteq H^{s_2}(\mathbb{R}^n) \quad \text{for } 0 \leqslant s_1 \leqslant s_2.$$

Sobolev spaces provide a natural framework in which to formalise the Carleson problem.

Problem 1.1 (Carleson, 1980). *Determine the values of $s \geqslant 0$ such that*

$$\text{if } f \in H^s(\mathbb{R}^n), \quad \text{then} \quad \lim_{t\to 0} e^{it\Delta} f(x) = f(x) \quad \text{for almost every } x \in \mathbb{R}^n. \tag{8}$$

That is, we wish to determine the minimal degree of regularity (measured in terms of the Sobolev space index s) for which almost everywhere convergence is guaranteed to hold.

Aside from its intrinsic appeal, Problem 1.1 is intimately related to important questions regarding the distribution of the solution $e^{it\Delta} f(x)$ in space-time. Pointwise convergence is typically proved via analysis of the *Schrödinger maximal operator*, an object of interest in its own right. The maximal operator can in turn be studied using *fractal energy estimates* for the Schrödinger solutions. We introduce these concepts in §3.2 and §3.5 below. Through these connections, progress on Problem 1.1 has led to new developments on a surprising array of different problems, such as the Falconer distance problem (see, for instance, Du and Zhang, 2019; Guth, Iosevich, et al., 2020) and the Fourier restriction conjecture (see Wang and Wu, 2022).

1.3. A resolution of the Carleson problem: introducing the key results

Problem 1.1 has a rich history, paralleling many important developments in harmonic analysis over the last 40 years. We do not intend to give a complete survey of the relevant literature, but focus on definitive results and recent highlights.

Whilst the $n = 1$ case of Problem 1.1 was fully understood by the early 1980s through the works of Carleson (1980) and Dahlberg and Kenig (1982), in higher dimensions the situation is much more nuanced. Nevertheless, a recent series of dramatic developments brought about an almost complete resolution.

Necessary conditions. — Problem 1.1 splits into two parts: finding necessary conditions for the index s for (8) to hold and finding sufficient conditions. Both parts are difficult. The recent spate of activity on the Carleson problem was initiated by the surprising discovery of a new necessary condition on s.

Theorem 1.2 (Bourgain, 2016). *For all $s < \frac{n}{2(n+1)}$, there exists some $f \in H^s(\mathbb{R}^n)$ such that* (6) *fails.*

Theorem 1.2 relies on the construction of an explicit[2] initial datum f; the proof is intricate, involving number theoretic considerations. Prior to Bourgain (2016), weaker necessary conditions were established in Bourgain (2013b), Dahlberg and Kenig (1982), and Lucà and Rogers (2017).

We shall not discuss the proof of Theorem 1.2 here, but instead refer the reader to the detailed exposition in Pierce (2020). An alternative argument, based on ergodic arguments rather than number theory, can also be found in Lucà and Rogers (2019).

Sufficient conditions. — We now turn to positive results, which form the focus of this article. In the wake of Bourgain's counterexample, there was a flurry of activity on the Carleson problem. In a major advance, the $n = 2$ case was completely settled through work of Du, Guth, and Li (2017). The higher dimensional case later followed in a landmark paper of Du and Zhang (2019).

Theorem 1.3 (Du and Zhang, 2019[3]). *If $f \in H^s(\mathbb{R}^n)$ for some $s > \frac{n}{2(n+1)}$, then*

$$\lim_{t \to 0_+} e^{it\Delta} f(x) = f(x) \qquad \text{holds for almost every } x \in \mathbb{R}^n.$$

Remark 1.4. As previously noted, there are measure-theoretic technicalities regarding the meaning of the above statement, since the limit is taken over a continuum. We address this in §3.2 below.

Together, Theorem 1.2 and Theorem 1.3 give an almost complete[4] answer to the Carleson problem and constitute a major milestone in harmonic analysis and PDE. Furthermore, Theorem 1.3 is in fact a special case of a significantly more general result proved in Du and Zhang (2019), which has a variety of additional applications: see §3.5 below.

The proof of Theorem 1.3 builds on many important developments in harmonic analysis and previous works on the Carleson problem in particular. For the purpose of this article, we shall roughly divide the recent history of the problem into two epochs.

[2] Strictly speaking, the proof of Theorem 1.2 proceeds by constructing a counterexample to the $H^s(\mathbb{R}^n) \to L^1(\mathbb{R}^n)$ boundedness of the *Schrödinger maximal operator*. This in turn implies the existence of a counterexample to (8) through a variant of Stein's maximal principle. See §3.2.

[3] The $n = 1$ and $n = 2$ cases of Theorem 1.3 were established earlier in Carleson (1980) and Du, Guth, and Li (2017), respectively.

[4] That is, except for the question of behaviour at the endpoint exponent $s = n/(2(n+1))$.

1. **Multilinear theory / broad-narrow analysis**. A key development in modern harmonic analysis was the introduction of multilinear Strichartz estimates for the Schrödinger equation in Bennett, Carbery, and Tao (2006) (see Theorem 5.5 below). These estimates were applied to the study of (linear) oscillatory integral operators by Bourgain and Guth (2011), where an important mechanism was introduced for estimating linear operators via their multilinear counterparts. The technique of Bourgain and Guth (2011) has since become known as *broad-narrow analysis*; the relevant ideas are discussed in detail in §6 below. The multilinear technology was applied to the study of Problem 1.1 in Bourgain (2013b), where Theorem 1.3 was shown to hold for the more restrictive range $s > \frac{2n-1}{4n}$ (this result was previously established for $n = 2$ in Lee (2006) using bilinear methods).

2. **Refined Strichartz estimates.** A landmark paper of Du, Guth, and Li (2017) established the $n = 2$ case of Theorem 1.3. Their work relied on important advances in harmonic analysis such as the ℓ^2 decoupling theorem of Bourgain and Demeter (2015) (see Theorem 6.5 below) and polynomial partitioning techniques introduced in Guth (2016) and Guth and Katz (2015). Moreover, Du, Guth, and Li (2017) introduced *refined Strichartz estimates*, which were later developed to attack the Carleson problem in higher dimensions in Du, Guth, Li, and Zhang (2018) and have been found to have an array of additional applications (see, for instance, Guth, Iosevich, et al., 2020; Wang and Wu, 2022).

The argument of Du and Zhang (2019) incorporates many of the tools and ideas mentioned above: in particular, the broad-narrow analysis of Bourgain and Guth (2011); the multilinear Strichartz inequalities of Bennett, Carbery, and Tao (2006) and the decoupling estimates of Bourgain and Demeter (2015). We shall discuss these ingredients in detail in §6 below. On the other hand, the methods of Du and Zhang (2019) are in many respects quite different from those used in Du, Guth, and Li (2017) to settle the $n = 2$ case. Here no polynomial partitioning is used and the refined Strichartz estimates are not necessary to the argument.[(5)] Nevertheless, the main novel ingredient in Du and Zhang (2019) is an ingenious induction-on-scale method which has its roots in the proof of the refined Strichartz estimates from Du, Guth, and Li (2017) and Du, Guth, Li, and Zhang (2018). We shall discuss these techniques in §7 below.

[(5)] In Du and Zhang (2019) multilinear refined Strichartz estimates from Du, Guth, Li, and Zhang (2018) are applied but, as noted in the paper, the more elementary multilinear Strichartz estimates of Bennett, Carbery, and Tao (2006) suffice for the proof of Theorem 1.3.

1.4. About this article

What follows is an exposition of the proof of Theorem 1.3, following the methods of Du and Zhang (2019). As described above, the proof combines sophisticated modern machinery from harmonic analysis and, in particular, the multilinear Strichartz estimates of Bennett, Carbery, and Tao (2006) and the ℓ^2 decoupling theory of Bourgain and Demeter (2015). We shall introduce these two key ingredients in Theorem 5.5 and Theorem 6.5 below, but we shall not provide proofs. The rest of the article is self-contained. Equally important to the argument are a variety of elementary guiding principles, rooted in Fourier analysis, which govern the behaviour of solutions to the Schrödinger equation. We shall spend some time in §4 discussing these principles, and as such this article could serve as an accessible introduction to this highly active area of harmonic analysis and PDE.

1.5. Acknowledgement

The author wishes to thank Marco Vitturi, Bernat Ramis Vich and an anonymous referee for many tremendously helpful comments which improved the exposition. He also wishes to thank Anthony Carbery and Andreas Seeger for some interesting conversations concerning aspects of the theory.

2. Notation

Throughout this article, we work either in the *space-time domain* $\mathbb{R}^{n+1}$ or *spatial domain* $\mathbb{R}^n$. The latter is endowed with the product metric

$$|z-\bar{z}| = \max\{|x-\bar{x}|, |t-\bar{t}|\} \qquad \text{for } z=(x,t), \bar{z}=(\bar{x},\bar{t}) \in \mathbb{R}^{n+1}, \tag{9}$$

where the norms $|\cdot|$ appearing on the right-hand side are the usual Euclidean norms on $\mathbb{R}^d$ for $d=n$ and $d=1$, respectively. The *spatial domain*, on the other hand, is endowed with the usual Euclidean metric. We write $B^{n+1}(z,R)$ for the space-time ball centred at $z \in \mathbb{R}^{n+1}$ of radius R, defined with respect to (9), and $B^n(x,R)$ for the usual Euclidean ball centred at $x \in \mathbb{R}$ of radius R. This gives rise to a slight ambiguity in the notation, but the meaning should always be clear from the context. In some cases we will write B_R^{n+1} or B_R^n for $B^{n+1}(0,R)$ and $B^n(0,R)$, respectively.

Given functions $f \in L^1(\mathbb{R}^n)$ and $g \in L^1(\widehat{\mathbb{R}}^n)$, we define the Fourier transform of f and the inverse Fourier transform of g by the integral formulæ

$$\hat{f}(\xi) := \int_{\mathbb{R}^n} e^{-ix\cdot\xi} f(x)\,\mathrm{d}x \quad \text{and} \quad \check{g}(x) := \frac{1}{(2\pi)^n}\int_{\widehat{\mathbb{R}}^n} e^{ix\cdot\xi} g(\xi)\,\mathrm{d}\xi.$$

Note that we distinguish between the spatial domain $\mathbb{R}^n$ and the *frequency domain* $\widehat{\mathbb{R}}^n$; this is simply for notational purposes and $\widehat{\mathbb{R}}^n$ can be thought of as a copy of $\mathbb{R}^n$.

Given a set $E \subseteq \mathbb{R}^d$, we let $\chi_E \colon \mathbb{R}^d \to \mathbb{R}$ denote its characteristic function, so that $\chi_E(x) = 1$ if $x \in E$ and $\chi_E(x) = 0$ otherwise. If E is Lebesgue measurable, we let $|E|$ denote its Lebesgue measure.

An *r-cube* (or simply a *cube*) $Q \subset \mathbb{R}^n$ is a set of the form

$$Q := x + [-r/2, r/2]^n \qquad \text{for some } x \in \mathbb{R}^n \text{ and } r > 0;$$

in this case x is referred to as the *centre* of the cube and r the *side-length*. Note that, for the purposes of this article, all cubes have faces parallel to the coordinate axes. We say $Q \subseteq \mathbb{R}^{n+1}$ is a *lattice r-cube* for some $r > 0$ if it is an r-cube centred at a point on the integer lattice $r\mathbb{Z}^{n+1}$. In the case $r = 1$, we will also refer to Q as a lattice unit-cube. Given a cube Q and $M > 0$, we let $M \cdot Q$ denote the cube concentric to Q but side-length scaled by a factor of M.

Given a list of objects L and real numbers $A, B \geqslant 0$, we write $A \lesssim_L B$ or $B \gtrsim_L A$ to indicate $A \leqslant C_L B$ for some constant C_L which depends only items in the list L and perhaps other admissible parameters such as the dimension n or Lebesgue exponents p. We write $A \sim_L B$ to indicate $A \lesssim_L B$ and $B \lesssim_L A$.

3. Standard reductions and reformulations

In this section we perform a series of arguments to reduce Theorem 1.3 to the *fractal energy estimate* stated in Theorem 3.7 below. Whilst these arguments are standard (use of maximal estimates, linearisation, discretisation), the final form of the fractal energy estimate in Theorem 3.7 is an interesting aspect of the approach.

3.1. Elementary symmetries

For fixed time $t \in \mathbb{R}$, the operator $e^{it\Delta}$ is a Fourier multiplier, and therefore automatically enjoys a number of special symmetries.[6] In particular, $e^{it\Delta}$ commutes with any other multiplier and, therefore, with (spatial) translations. We call this property *spatial translation invariance*. With respect to dilations, the operators satisfy

$$e^{it\Delta}(\delta_R f) = \delta_R(e^{iR^2 t\Delta} f) \qquad \text{where} \qquad \delta_R f(x) := f(Rx),$$

and so are dilation invariant up to a scaling of the temporal parameter. Finally, the operators $e^{it\Delta}$ form a semigroup, which leads to temporal translation invariance properties, at least at the level of L^2 norms.

[6] We will discuss additional special symmetries particular to $e^{it\Delta}$ in §4.2 below.

3.2. Schrödinger maximal estimates

The first step is to apply a standard argument to reduce the pointwise convergence problem to proving an *a priori* bound for a maximal operator.

Theorem 3.1 (Du and Zhang, 2019). *For all $s > \frac{n}{2(n+1)}$, we have a maximal estimate*

$$\| \sup_{0\leqslant t\leqslant 1} |e^{it\Delta} f| \|_{L^2(B^n(0,1))} \lesssim \|f\|_{H^s(\mathbb{R}^n)} \qquad \text{for all } f \in \mathscr{S}(\mathbb{R}^n). \tag{10}$$

Here, given $f \in H^s(\mathbb{R}^n)$, the *H^s-norm* is defined by

$$\|f\|_{H^s(\mathbb{R}^n)} := \|(1-\Delta_x)^{s/2} f\|_{L^2(\mathbb{R}^n)}.$$

Note that we only ask for (10) to hold for $f \in \mathscr{S}(\mathbb{R}^n)$; in this case the solution $u(x,t) := e^{it\Delta} f$ is continuous in (x,t) and so the expression appearing on the left-hand side of (10) is certainly well-defined. We refer to the operation $f \mapsto \sup_{0\leqslant t\leqslant 1} |e^{it\Delta} f|$ as the *Schrödinger maximal operator*.

By a standard argument, Theorem 3.1 implies Theorem 1.3. For completeness, the details of this implication are given below. We remark that the choice of L^2 space here is likely sub-optimal and (10) may hold with the left-hand L^2-norm replaced with an L^p-norm for larger p. Indeed, for $n = 1$ it was shown in Kenig, Ponce, and Vega (1991) that one may take $p = 4$ and for $n = 2$ it was shown in Du, Guth, Li, and Zhang (2018) that one may take $p = 3$. These exponents are sharp. In general, an obvious necessary condition is $p \leqslant 2 \cdot \frac{n+1}{n}$ which arises through Sobolev embedding. Indeed, since we know $\lim_{t\to 0} e^{it\Delta} f$ for $f \in \mathscr{S}(\mathbb{R}^n)$, we can only hope to have an $H^s(\mathbb{R}^n) \to L^p(\mathbb{R}^n)$ maximal estimate if $H^s(\mathbb{R}^n)$ embeds into $L^p(\mathbb{R}^n)$. Surprisingly, a more restrictive necessary condition on p was found in Du, Kim, et al. (2020) by adapting Bourgain's counterexample from Bourgain (2016).

Theorem 3.1 has the following consequence, which formalises the meaning of the limit identity in Theorem 1.3 (*c.f.* Remark 1.4 and the discussion following (6)).

Corollary 3.2. *Let $s > \frac{n}{2(n+1)}$ and $f \in H^s(\mathbb{R}^n)$. Then for each $0 \leqslant t \leqslant 1$ there is a choice of representative of the L^2-function $e^{it\Delta} f$ such that for almost every $x \in \mathbb{R}^n$ the map $t \mapsto e^{it\Delta} f(x)$ is continuous.*

Proof. By translation invariance, it suffices to consider only $x \in B^n(0,1)$. Fixing f as in the statement of the corollary, for each $k \in \mathbb{N}$ we can find some $f_k \in \mathscr{S}(\mathbb{R}^n)$ such that $\|f - f_k\|_{H^s(\mathbb{R}^n)} \leqslant 2^{-2k}$. We claim that the set

$$E := \big\{ x \in B^n(0,1) : \limsup_{k\to\infty} 2^k \sup_{0\leqslant t\leqslant 1} |e^{it\Delta}(f_k - f_{k-1})(x)| > 1 \big\}$$

is of Lebesgue measure zero. Temporarily assuming this is true, we see that the sequence $e^{it\Delta} f_k(x)$ of functions is Cauchy over $(x,t) \in B^n(0,1) \setminus E \times [0,1]$, uniformly

in t. The sequence therefore converges pointwise in x and uniformly in t on this set to some limit function $L(x,t)$, which is continuous in t. On the other hand, for fixed $0 \leqslant t \leqslant 1$, by the conservation of energy identity we have

$$\|e^{it\Delta}f_k - e^{it\Delta}f\|_{L^2(\mathbb{R}^n)} = \|f_k - f\|_{L^2(\mathbb{R}^n)} \leqslant \|f_k - f\|_{H^s(\mathbb{R}^n)} \leqslant 2^{-2k},$$

and so $e^{it\Delta}f_k$ converges to $e^{it\Delta}f$ in the $L^2(\mathbb{R}^n)$ sense. Since convergence in L^2 implies almost everywhere convergence along a subsequence, it follows that for each $0 \leqslant t \leqslant 1$, the function $L(x,t)$ must agree with $e^{it\Delta}f(x)$ for almost every $x \in B^n(0,1)$. This is precisely the desired conclusion.

It remains to show E is of Lebesgue measure zero, which is achieved using a Borel–Cantelli argument. Note that

$$E \subseteq \bigcap_{K\in\mathbb{N}} \bigcup_{k\geqslant K} E_k$$

where

$$E_k := \big\{x \in B^n(0,1) : \sup_{0\leqslant t\leqslant 1} |e^{it\Delta}(f_k - f_{k-1})(x)| > 2^{-k}\big\}.$$

By Tchebyshev's inequality and the maximal estimate,

$$|E| \leqslant \sum_{k\geqslant K} |E_k| \leqslant \sum_{k\geqslant K} 2^{2k} \big\| \sup_{0\leqslant t\leqslant 1} |e^{it\Delta}(f_k - f_{k-1})| \big\|_{L^2(\mathbb{R}^n)}^2 \lesssim_s \sum_{k\geqslant K} 2^{-2k} \lesssim 2^{-2K}$$

for all $K \in \mathbb{N}$. Thus, $|E| = 0$, as required.

□

For $s > \frac{n}{2(n+1)}$ and $f \in H^s(\mathbb{R}^n)$, we can use Corollary 3.2 to clarify the meaning of our limits. We choose representatives of the functions $e^{it\Delta}f$ for $0 \leqslant t \leqslant 1$ to guarantee the conclusion of Corollary 3.2. Since $e^{it\Delta}f(x)|_{t=0} = f(x)$ for almost every $x \in \mathbb{R}^n$, with this choice of representatives we do indeed have

$$\lim_{t\to 0_+} e^{it\Delta}f(x) = f(x) \qquad \text{for almost every } x \in \mathbb{R}^n.$$

Thus, Corollary 3.2 can be interpreted as the precise formulation of Theorem 1.3.

3.3. Littlewood–Paley decomposition

The next step is to recast the $H^s(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ maximal bound in Theorem 3.1 as an $L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ estimate for frequency localised data. We are easily able to do this since Theorem 3.1 involves an open range of s.

For $R \geqslant 1$, let $\chi_{A(R)}$ denotes the characteristic function of the frequency annulus

$$A(R) := \{\xi \in \widehat{\mathbb{R}}^n : R/2 \leqslant |\xi| < R\}$$

and let $\chi_{A(R)}(D)$ denote the Fourier multiplier operator defined by

$$\chi_{A(R)}(D)f := \mathscr{F}^{-1}(\chi_{A(R)} \cdot \mathscr{F}f) \qquad \text{for all } f \in L^2(\mathbb{R}^n).$$

Thus, $\chi_{A(R)}(D)$ corresponds to a frequency projection to $A(R)$, with a rough cutoff. As a simple consequence of Plancherel's theorem,

$$\|f\|_{H^s(\mathbb{R}^n)} \sim \Big(\sum_{k=1}^{\infty} 2^{2ks}\|\chi_{A(2^k)}(D)f\|^2_{L^2(\mathbb{R}^n)}\Big)^{1/2} + \|f\|_{L^2(\mathbb{R}^n)}; \tag{11}$$

this is known as the *Littlewood–Paley characterisation* of the H^s-norm.

In view of the characterisation (11), to prove Theorem 3.1 it suffices to show that for all $\varepsilon > 0$ and all $R \geqslant 1$, the inequality

$$\| \sup_{0\leqslant t\leqslant 1} |e^{it\Delta}f|\|_{L^2(B^n(0,1))} \lesssim_\varepsilon R^{n/(2(n+1))+\varepsilon}\|f\|_{L^2(\mathbb{R}^n)} \tag{12}$$

holds for all $f \in L^2(\mathbb{R}^n)$ with $\operatorname{supp}\hat{f} \subseteq A(R)$. By scaling and exploiting certain pseudo-local properties of the propagator, the problem is further reduced to the following proposition.

Proposition 3.3. *For all $\varepsilon > 0$ and all $R \geqslant 1$ the inequality*

$$\| \sup_{0\leqslant t\leqslant R} |e^{it\Delta}f|\|_{L^2(B^n(0,R))} \lesssim_\varepsilon R^{n/(2(n+1))+\varepsilon}\|f\|_{L^2(\mathbb{R}^n)} \tag{13}$$

holds whenever $f \in L^2(\mathbb{R}^n)$ satisfies $\operatorname{supp}\hat{f} \subseteq A(1)$.

Scaling alone shows that (12) is equivalent to a variant of (13) in which the supremum is taken over the longer time interval $0 < t < R^2$. To pass to the shorter time interval, we use an argument of Lee (2006), based on pseudo-local properties of the operator; we describe the details in §4.4.

3.4. Linearising the maximal operator

To prove Proposition 3.3, we linearise the maximal operator using the Kolmogorov–Seliverstov–Plessner method. In particular, it suffices to show, for all $\varepsilon > 0$, $R \geqslant 1$ and all measurable functions $\mathbf{t}\colon B^n(0,R) \to (0,R)$, the inequality

$$\Big(\int_{B^n(0,R)} |e^{i\mathbf{t}(x)\Delta}f(x)|^2\,\mathrm{d}x\Big)^{1/2} \lesssim_\varepsilon R^{n/(2(n+1))+\varepsilon}\|f\|_{L^2(\mathbb{R}^n)} \tag{14}$$

holds for all $f \in L^2(\mathbb{R}^n)$ with $\operatorname{supp}\hat{f} \subseteq B^n(0,1)$.

Let U denote the operator defined initially on the Schwartz class by

$$Uf(x,t) := \frac{1}{(2\pi)^n}\int_{B^n(0,1)} e^{i(\langle x,\xi\rangle + t|\xi|^2)}\hat{f}(\xi)\,\mathrm{d}\xi.$$

Thus, $Uf(\,\cdot\,,t)$ corresponds to the composition of $e^{it\Delta}f$ with a rough frequency projection to the unit ball. We think of the estimate (14) as bounding Uf over the space-time graph $\Gamma_{\mathbf{t}} := \{(x,\mathbf{t}(x)) : x \in B^n(0,R)\}$ of the measurable function $\mathbf{t}$. This perspective turns out to be useful and we shall formulate all our key estimates over the space-time domain.

Since Uf is localised in frequency, we do not expect the values $|Uf(z)|$ to vary greatly at small scales. This is a manifestation of the *uncertainty principle*, which is discussed in detail in §4.1 below. These observations allow us to discretise our setup.

Definition 3.4. Let $\mathscr{Q}$ be a family of lattice unit cubes in $\mathbb{R}^{n+1}$.

i) We let $Z_{\mathscr{Q}}$ denote the union $\bigcup_{Q\in\mathscr{Q}} Q$.

ii) We say $\mathscr{Q}$ satisfies the *vertical line test* if for almost every $x \in \mathbb{R}^n$, at most one cube from Q intersects the line $\{(x,t) : t \in \mathbb{R}\}$.

A lattice unit cube Q should be thought of as a 'discretised point' and $Z_{\mathscr{Q}}$ for a family $\mathscr{Q}$ satisfying the vertical line test as a 'discretised graph'. With these definitions, Proposition 3.3 is a consequence of the following bound.

Proposition 3.5 (Du and Zhang, 2019). *For all $\varepsilon > 0$ and all $R \geqslant 1$, the inequality*

$$\|Uf\|_{L^2(Z_{\mathscr{Q}})} \lesssim_\varepsilon R^{n/(2(n+1))+\varepsilon}\|f\|_{L^2(\mathbb{R}^n)}$$

holds whenever $f \in L^2(\mathbb{R}^n)$ and $\mathscr{Q}$ is a collection of unit lattice cubes lying in $B^{n+1}(0,R)$ which satisfy the vertical line test.

We will show in detail why the discretised bound in Proposition 3.5 implies the linearised maximal estimate (14) in §4.1 below.

3.5. Fractal energy estimates

Proposition 3.5 is in fact a special case of a significantly more general theorem proved in Du and Zhang (2019). Here we describe the general framework, which we adopt for the remainder of the exposition.

Definition 3.6. Let $M > 0$ and $\mathscr{Q}$ be a family of lattice M-cubes in $\mathbb{R}^{n+1}$.

i) For $1 \leqslant \alpha \leqslant n+1$ and a space-time ball $B \subseteq \mathbb{R}^{n+1}$, we define

$$\Delta_\alpha(\mathscr{Q},B) := \frac{\#\{Q \in \mathscr{Q} : Q \subset B\}}{\mathrm{rad}(B)^\alpha},$$

where $\mathrm{rad}(B)$ denotes the radius of B.

ii) Furthermore, let

$$\Delta_\alpha(\mathscr{Q}) := \sup_B \Delta_\alpha(\mathscr{Q}, B)$$

where the supremum is taken over all space-time balls $B = B^{n+1}(z, r) \subseteq \mathbb{R}^{n+1}$.

If $1 \leqslant M \leqslant R$ and $\mathscr{Q}$ is a non-empty collection of lattice M-cubes contained in $B^{n+1}(0, R)$, then it follows from the definition that

$$M^{-\alpha} \lesssim \Delta_\alpha(\mathscr{Q}) \qquad \text{and} \qquad \#\mathscr{Q} \lesssim \Delta_\alpha(\mathscr{Q}) R^\alpha.$$

We may now (finally) state the main result of Du and Zhang (2019).[7]

Theorem 3.7 (Du and Zhang, 2019). *For all $\varepsilon > 0$ and all $R \geqslant 1$, $1 \leqslant \alpha \leqslant n+1$, the inequality*

$$\|Uf\|_{L^2(Z_{\mathscr{Q}})} \lesssim_\varepsilon \Delta_\alpha(\mathscr{Q})^{1/(n+1)} R^{\alpha/(2(n+1))+\varepsilon} \|f\|_{L^2(\mathbb{R}^n)} \tag{15}$$

holds whenever $f \in L^2(\mathbb{R}^n)$ and $\mathscr{Q}$ is a family of lattice unit cubes in $B^{n+1}(0, R)$.

If $\mathscr{Q}$ satisfies the vertical line test, then it is not difficult to see $\Delta_n(\mathscr{Q}) \lesssim 1$. Consequently, Theorem 3.7 implies Proposition 3.5 and therefore, by the preceding reductions, also pointwise convergence result in Theorem 1.3.

We refer to the estimate (15) as a *fractal energy estimate*. This terminology is partly motivated by the conservation of energy identity (4). Indeed, from (4) we have

$$\|Uf\|_{L^2(Z_{\mathscr{Q}})} \leqslant \Big(\int_{-R}^{R} \|e^{it\Delta} f\|_{L^2(\mathbb{R}^n)}^2 \, \mathrm{d}t \Big)^{1/2} \lesssim R^{1/2} \|f\|_{L^2(\mathbb{R}^n)}, \tag{16}$$

which directly implies the $\alpha = n+1$ case of Theorem 3.7. For general exponents $1 \leqslant \alpha \leqslant n+1$, it is useful to think of a family of cubes $\mathscr{Q}$ satisfying $\Delta_\alpha(\mathscr{Q}) \lesssim 1$ as a discretised version of an α-dimensional 'fractal' set.

There are two main advantages in working with the general framework of fractal energy estimates:

1) Theorem 3.7 has an array of additional applications beyond the pointwise convergence problem for the Schrödinger maximal function. These include estimates for the dimension of the divergence set in the Carleson problem and partial results towards the Falconer distance conjecture. We refer to Du and Zhang (2019, §2) for further details.

2) The form of the estimate in (15) is useful when it comes to the proof. In particular, the arguments involve an induction scheme and the inclusion of the $\Delta_\alpha(\mathscr{Q})$ factor allows greater leverage from the induction hypothesis.[8]

[7] In fact, Du and Zhang (2019) prove a further strengthening of Theorem 3.7 involving an additional parameter λ: see Du and Zhang (2019, Theorem 1.6). However, this strengthened result is unnecessary for typical applications and so here we stick to the simpler statement in Theorem 3.7.

[8] The induction argument is not applied to the statement in Theorem 3.7 itself, but a variant described in Proposition 7.2.

For the remainder of this article we shall discuss the proof of Theorem 3.7. We start with some basic background in harmonic analysis and dispersive PDEs in §4, before moving to more advanced topics in §§5–7.

4. Basic tools from harmonic analysis

4.1. The uncertainty principle

We begin with a discussion of the uncertainty principle for the Fourier transform. This is a set of heuristics which roughly state:

> If $\hat{F}$ is localised at scale R^{-1}, then $|F|$ should be locally constant at scale R.

The following lemma provides a rigorous interpretation of this principle at the unit scale.

Lemma 4.1 (Locally constant property). *There exists a continuous function $\eta\colon \mathbb{R}^d \to [0,\infty)$ satisfying the following:*

i) If $F \in \mathscr{S}(\mathbb{R}^d)$ satisfies $\operatorname{supp} \hat{F} \subseteq Q_0 := [-1/2,1/2]^d$, then

$$|F(z)| \leqslant \sum_{Q\in\mathscr{Q}_{\text{all}}} a_Q \chi_Q(z) \lesssim |F| * \eta(z) \qquad \text{for all } z \in \mathbb{R}^d \tag{17}$$

where here $\mathscr{Q}_{\text{all}}$ is the collection of all lattice unit cubes in $\mathbb{R}^d$ and

$$a_Q := \sup_{z\in Q} |F(z)| \qquad \text{for all } Q \in \mathscr{Q}_{\text{all}}.$$

ii) The function η is L^1-normalised and rapidly decaying away from Q_0 in the sense that

$$\eta(z) \lesssim_N (1+2|z|_\infty)^{-N} \qquad \text{for all } N \in \mathbb{N}.$$

Proof. Fix $\eta_0 \in \mathscr{S}(\mathbb{R}^d)$ satisfying $\hat{\eta}_0(\xi) = 1$ for all $\xi \in [-1,1]^d$. Thus, if $F \in \mathscr{S}(\mathbb{R}^d)$ satisfies the hypothesis of part i), we have the reproducing formula $F = F * \eta_0$. In particular, given $Q \in \mathscr{Q}_{\text{all}}$ and $z_Q \in Q$ chosen to satisfy $|F(z_Q)| = a_Q$, it follows that

$$a_Q = |F(z_Q)| \leqslant \int_{\mathbb{R}^d} |F(y)||\eta_0(z_Q - y)|\,\mathrm{d}y.$$

Now define $\eta\colon \mathbb{R}^d \to [0,\infty)$ by

$$\eta(z) := \sup_{|w-z|_\infty \leqslant 1} |\eta_0(w)|.$$

The rapid decay of the Schwartz function η_0 then ensures η is rapidly decaying away from $[-1,1]^d$. It therefore only remains to show that (17) holds.

If $z \in Q$ is an arbitrary element, then

$$|(z_Q - y) - (z - y)|_\infty \leqslant 1 \qquad \text{for all } y \in \mathbb{R}^d.$$

Consequently, $|\eta_0(z_Q - y)| \leqslant \eta(z - y)$ for all $y \in \mathbb{R}^d$, and so

$$a_Q \leqslant |F| * \eta(z) \qquad \text{for all } z \in Q,$$

which immediately implies the second inequality in (17). On the other hand, the first inequality in (17) is a trivial consequence of the definitions. □

We may use the rigorous formulations of the uncertainty principle introduced above to justify the discretisation procedure described in §3.4.

Proof (*Proposition 3.5 ⇒ Proposition 3.3*). Assume Proposition 3.5 holds. Recall that it suffices to show the linearised maximal estimate (14).

Let $\tilde{\chi} \in \mathscr{S}(\mathbb{R}^{n+1})$ satisfy $|\tilde{\chi}(z)| \gtrsim 1$ for all $|z| \leqslant 2$ and $\operatorname{supp} \mathscr{F}\tilde{\chi} \subseteq B^{n+1}(0,1)$. Defining

$$F(z) := Uf(z) \cdot \tilde{\chi}(R^{-1}z),$$

then it follows that $|Uf(z)| \lesssim |F(z)|$ for all $z \in B^{n+1}(0,2R)$ and $\operatorname{supp} \hat{F} \subseteq B^{n+1}(0,2)$. The proof will follow from the resulting locally constant properties of the function F.

Fix $\varepsilon > 0$ and a measurable function $\mathbf{t}\colon B^n(0,R) \to (0,1]$ and let $\mathfrak{q}_R$ denote the collection of lattice unit cubes in $\mathbb{R}^n$ which intersect $B^n(0,R)$. For each $q \in \mathfrak{q}_R$ there exists some choice of $x_q \in q$ such that

$$\sup_{x \in q} |e^{i\mathbf{t}(x)\Delta} f(x)| \leqslant 2|e^{i\mathbf{t}(x_q)\Delta} f(x_q)|.$$

If we define $z_q := (x_q, \mathbf{t}(x_q))$ for each $q \in \mathfrak{q}_R$, then it follows that

$$\Big(\int_{B^n(0,R)} |e^{i\mathbf{t}(x)\Delta} f(x)|^2 \, \mathrm{d}x \Big)^{1/2} \lesssim \Big(\sum_{q \in \mathfrak{q}_R} |F(z_q)|^2 \Big)^{1/2}.$$

For each $q \in \mathfrak{q}_R$ let $I_q \subseteq \mathbb{R}$ denote a choice of lattice unit interval containing $\mathbf{t}(x_q)$ and define $Q_q := q \times I_q$. By Lemma 4.1, we have

$$|F(z_q)| \lesssim \| |F| * \eta \|_{L^2(Q_q)}$$

where η is rapidly decaying away from the unit cube in $\mathbb{R}^{n+1}$ centred at the origin. In particular, if we let $\delta := \varepsilon/(2n)$ and define the enlarged cube $Q_q^{(\delta)} := R^\delta \cdot Q_q$, then

$$|F(z_q)| \lesssim_{N,\varepsilon} \|F\|_{L^2(Q_q^{(\delta)})} + R^{-N} \|F\|_{L^2(\mathbb{R}^{n+1})} \qquad \text{for all } N \in \mathbb{N}_0.$$

Thus, if we define the family of space-time unit cubes $\mathscr{Q}^{(\delta)} := \{Q_q^{(\delta)} : q \in \mathfrak{q}_R\}$, then

$$\Big(\int_{B^n(0,R)} |e^{i\mathbf{t}(x)\Delta} f(x)|^2 \, \mathrm{d}x \Big)^{1/2} \lesssim \|Uf\|_{L^2(Z_{\mathscr{Q}^{(\delta)}})} + R^{-100n} \|Uf\|_{L^2(w_{B_R^{n+1}})}, \tag{18}$$

for $w_{B_R^{n+1}}$ a weight adapted to B_R^{n+1}. The rapidly decaying error term is easily bounded using the conservation of energy identity. In particular, it follows from the rapid decay of the weight and translation invariance properties of the operator that

$$\|Uf\|_{L^2(w_{B_R^{n+1}})} \lesssim R^{1/2} \|f\|_{L^2(\mathbb{R}^n)}. \tag{19}$$

On the other hand, the set $Z_{\mathscr{Q}^{(\delta)}}$ can be covered by $O(R^{(n+1)\delta})$ sets of the form $Z_{\mathscr{Q}}$ where $\mathscr{Q}$ is a family of unit lattice cubes satisfying the hypotheses of Proposition 3.5 (and, in particular, the vertical line test). For any such $\mathscr{Q}$, we may apply Proposition 3.5 to deduce that

$$\|Uf\|_{L^2(Z_{\mathscr{Q}})} \lesssim_\varepsilon R^{n/(2(n+1))+\delta} \|f\|_{L^2(\mathbb{R}^n)}.$$

Summing together these contributions, we obtain an estimate for $\|Uf\|_{L^2(Z_{\mathscr{Q}^{(\delta)}})}$. This can be combined with (18) and (19) to deduce the desired bound (14). □

We now discuss some further manifestations of the uncertainty principle which are of use in later arguments.

By the basic scaling properties of the Fourier transform, Lemma 4.1 implies a generalisation of itself. We define a *parallelepiped* to be set $\pi \subseteq \mathbb{R}^d$ given by the image of $Q_0 := [-1/2, 1/2]^d$ under an affine transformation. In particular, $\pi = A(Q_0) + a$ for some $A \in \mathrm{GL}(d, \mathbb{R})$ and $a \in \mathbb{R}^d$. Given such a parallelepiped, we define the *dual parallelepiped* $\pi^* := A^{-\top}(Q_0)$, where $A^{-\top}$ is the inverse transpose of A, and

$$\mathscr{P}_{\mathrm{all}}(\pi) := \{A^{-\top} Q : Q \in \mathscr{Q}_{\mathrm{all}}\}.$$

Thus, $\mathscr{P}_{\mathrm{all}}(\pi)$ is a family of translates of π^* which tile the whole of $\mathbb{R}^d$.

As a direct consequence of Lemma 4.1 and an obvious change of variables, we obtain the following.

Corollary 4.2. *Let $\pi \subseteq \widehat{\mathbb{R}}^d$ be a parallelepiped. There exists a function $\eta_\pi \colon \mathbb{R}^n \to [0, \infty)$ satisfying the following:*

i) If $F \in \mathscr{S}(\mathbb{R}^d)$ satisfies $\operatorname{supp} \hat{F} \subseteq \pi$, then

$$|F(z)| \leqslant \sum_{P \in \mathscr{P}_{\mathrm{all}}(\pi)} a_P \chi_P \lesssim |F| * \eta_\pi(z) \qquad \text{for all } z \in \mathbb{R}^d \tag{20}$$

where $\mathscr{P}_{\mathrm{all}}(\pi)$ is as defined above and

$$a_P := \sup_{z \in P} |F(z)| \qquad \text{for all } P \in \mathscr{P}_{\mathrm{all}}(\pi).$$

ii) The function η_π is L^1-normalised and rapidly decaying away from π^* in the sense that

$$\eta(z) \lesssim_N |\pi^*|^{-1}(1+|z|_{\pi^*})^{-N} \qquad \text{for all } N \in \mathbb{N}.$$

Here $|\cdot|_{\pi^*}$ is the norm given by the Minkowski function of π^*.

Corollary 4.2 realises the uncertainty principle by allowing us to pass from F to its discretisation over $\mathscr{P}_{\mathrm{all}}(\pi)$ and back again. However, the appearance of the mollifier η_π on the right-hand side of (20) is a source of minor annoyance. It can be removed at the level of $L^q(\mathbb{R}^d)$ norm estimates: for instance, combining (17) with Young's inequality we deduce that

$$\|F\|_{L^q(\mathbb{R}^d)} \leqslant \Big(\sum_{P \in \mathscr{P}_{\mathrm{all}}(\pi)} |a_P|^q |P| \Big)^{1/q} \lesssim \|F\|_{L^q(\mathbb{R}^d)}$$

for all $1 \leqslant q < \infty$. Generalising this argument, we arrive at the following fundamental estimate.

Lemma 4.3 (Bernstein inequality). *Let $1 \leqslant p \leqslant q \leqslant \infty$ and suppose $F \in \mathscr{S}(\mathbb{R}^d)$ satisfies* $\operatorname{supp} \hat{F} \subseteq \pi$ *for some parallelepiped $\pi \subseteq \widehat{\mathbb{R}}^d$. Then*

$$\|F\|_{L^q(\mathbb{R}^d)} \lesssim |\pi|^{1/p-1/q} \|F\|_{L^p(\mathbb{R}^d)}.$$

The moral here is that, by the locally constant property, F behaves like a discrete function and, consequently, the L^p norms of F satisfy a nesting property like the ℓ^p norms of a sequence.

Proof (of Lemma 4.3). We assume $q < \infty$; the same proof goes through for $q = \infty$ *mutatis mutandis*. Fix $F \in \mathscr{S}(\mathbb{R}^d)$ satisfying the Fourier support hypothesis and apply Corollary 4.2 to bound

$$|F| \leqslant \sum_{P \in \mathscr{P}_{\mathrm{all}}(\pi)} a_P \chi_P.$$

Since $|P| = |\pi|^{-1}$ for any $P \in \mathscr{P}_{\mathrm{all}}(\pi)$, by the nesting of ℓ^p norms,

$$\|F\|_{L^q(\mathbb{R}^d)} \leqslant \Big(\sum_{P \in \mathscr{P}_{\mathrm{all}}(\pi)} |a_P|^q |P| \Big)^{1/q} \leqslant |\pi|^{1/p-1/q} \Big(\sum_{P \in \mathscr{P}_{\mathrm{all}}(\pi)} |a_P|^p |P| \Big)^{1/p}.$$

Applying Corollary 4.2, we have

$$\Big(\sum_{P \in \mathscr{P}_{\mathrm{all}}(\pi)} |a_P|^p |P| \Big)^{1/p} = \Big\| \sum_{P \in \mathscr{P}_{\mathrm{all}}(\pi)} |a_P| \chi_P \Big\|_{L^p(\mathbb{R}^d)} \lesssim \| |F| * \eta_\pi \|_{L^p(\mathbb{R}^d)}.$$

The desired result now follows by Young's convolution inequality. □

The Bernstein inequality can be localised in space, provided this localisation occurs at a scale which is coarse enough to respect the uncertainty principle. Before we state this local version, we introduce the following definition which plays a somewhat technical rôle in our arguments.

Definition 4.4. Let $S_0 \subseteq \mathbb{R}^d$ be a symmetric convex body[(9)] and $S := S_0 + z_0$ some translate of S. We say a function $w_S \colon \mathbb{R}^d \to [0,\infty)$ is a (*rapidly decaying*) *weight adapted to* S if w_S is continuous and satisfies

$$w_S(z) \lesssim_N (1 + |z - z_0|_{S_0})^{-N} \qquad \text{for all } N \in \mathbb{N},$$

where here $|\cdot|_{S_0}$ is the norm given by the Minkowski function of S_0.

Such weight functions are used to account for 'Schwartz tails errors' which arise when attempting to localise a function simultaneously in the physical and the frequency domain.

Corollary 4.5 (Local Bernstein inequality). *Let $1 \leqslant p \leqslant q \leqslant \infty$, and $\pi \subseteq \widehat{\mathbb{R}}^d$ be a parallelepiped. For every $P \in \mathscr{P}_{\mathrm{all}}(\pi)$ there exists a rapidly decaying weight w_P adapted to P such that the following holds. If $F \in \mathscr{S}(\mathbb{R}^d)$ satisfies* $\operatorname{supp} \hat{F} \subseteq \pi$, *then*

$$\|F\|_{L^q(P)} \lesssim |\pi|^{1/p-1/q} \|F\|_{L^p(w_P)}.$$

Proof. We may assume without loss of generality that $p < \infty$. It is a simple exercise to show that there exists some $\beta_P \in \mathscr{S}(\mathbb{R}^d)$ with $\operatorname{supp} \hat{\beta}_P \subseteq \pi$ satisfying

$$1 \lesssim |\beta_P(z)| \quad \text{for all } z \in P$$

and such that $|\beta_P(z)|$ is a rapidly decaying weight adapted to P. The function $G := F \cdot \beta_P$ has Fourier support in the Minkowski sum $\pi + \pi$ and satisfies $|F(z)| \lesssim |G(z)|$ for all $z \in P$. Applying Bernstein's inequality to G, the desired result follows with $w_P := |\beta_P|^p$. □

4.2. Parabolic geometry

Given $f \in \mathscr{S}(\mathbb{R}^n)$ with $\operatorname{supp} \hat{f} \subseteq B^n(0,1)$, we recall the integral formula for the solution

$$Uf(x,t) = \frac{1}{(2\pi)^n} \int_{B^n(0,1)} e^{i\phi(x,t;\xi)} \hat{f}(\xi)\, \mathrm{d}\xi,$$

where the *phase function* ϕ is given by

$$\phi(x,t;\xi) := \langle x, \xi \rangle + t|\xi|^2.$$

(9) In practice, we will only consider simple examples such as euclidean balls, cubes and cartesian products of these sets.

The phase ϕ can be interpreted as the inner product of the space-time vector $(x,t) \in \mathbb{R}^{n+1}$ with a point lying on the bounded piece of the paraboloid

$$\Sigma_0 := \big\{\Sigma(\xi) : \xi \in B(0,1)\big\} \subset \Sigma := \big\{\Sigma(\xi) : \xi \in \widehat{\mathbb{R}}^n\big\} \tag{21}$$

where $\Sigma(\xi) := (\xi, |\xi|^2)$. It follows that the (distributional) spatio-temporal Fourier transform of Uf is supported on Σ_0.

The geometry of Σ underpins our entire analysis of the propagator $e^{it\Delta}$. As a first example of this, we observe certain symmetries of the paraboloid, which translate into symmetries of the propagator.

Any ball $\theta \subset \widehat{\mathbb{R}}^n$ of radius r corresponds to a *cap*, or *r-cap*, on the paraboloid, given by

$$\Sigma_\theta := \big\{\Sigma(\xi) : \xi \in \theta\big\}.$$

In view of this, we shall often refer to balls $\theta \subset \widehat{\mathbb{R}}^n$ themselves as 'caps'. Note that the bounded piece of the paraboloid Σ_0 featured in (21) corresponds to the cap formed over the unit ball centred at the origin. We observe an important self-similarity property of the paraboloid, relating every cap to Σ_0.

Lemma 4.6 (Parabolic rescaling: geometric version). *Given a cap Σ_θ, corresponding to a ball $\theta \subseteq \widehat{\mathbb{R}}^n$, there exists an affine transformation A_θ of the ambient space $\widehat{\mathbb{R}}^{n+1}$ which restricts to a bijection from Σ_0 to Σ_θ.*

Remark 4.7. By inverting and composing affine transformations, Lemma 4.6 further implies that given any two caps $\Sigma_{\theta_1}, \Sigma_{\theta_2}$, there exists an affine transformation $A_{\theta_2 \to \theta_1}$ of the ambient space $\widehat{\mathbb{R}}^{n+1}$ which restricts to a bijection from Σ_{θ_2} to Σ_{θ_1}.

Before giving the (simple) proof of Lemma 4.6, we introduce some notation and, in fact, give an explicit formula for A_θ. Given a ball $\theta \subset \widehat{\mathbb{R}}^n$, let ξ_θ denote its centre and $\mathrm{rad}(\theta)$ its radius. We let M_θ be the shear transformation and D_θ the anisotropic scaling on $\widehat{\mathbb{R}}^{n+1}$ defined by

$$M_\theta := \begin{pmatrix} I_n & 0 \\ 2\xi_\theta^\top & 1 \end{pmatrix} \qquad \text{and} \qquad D_\theta := \begin{pmatrix} \mathrm{rad}(\theta) I_n & 0 \\ 0 & \mathrm{rad}(\theta)^2 \end{pmatrix};$$

here I_n denotes the $n \times n$ identity matrix. It will be shown in the proof below that the affine transformation A_θ in the statement of Lemma 4.6 can be taken to be

$$A_\theta \colon \zeta \mapsto (\mathscr{L}_\theta)^\top \zeta + \Sigma(\xi_\theta) \qquad \text{where} \qquad \mathscr{L}_\theta := D_\theta \circ M_\theta^\top, \tag{22}$$

where here $\top$ is used to denote the matrix transpose.

Proof (of Lemma 4.6). Let $\theta \subseteq \widehat{\mathbb{R}}^n$ be a ball, so that the map

$$\eta \mapsto \xi_\theta + r_\theta \eta \quad \text{for} \quad \eta \in \widehat{\mathbb{R}}^n$$

restricts to a bijection from $B(0,1)$ to θ. Here $r_\theta := \mathrm{rad}(\theta)$. Moreover, if we fix $\xi \in \theta$ and write $\xi = \xi_\theta + r_\theta \eta$, then a simple computation shows that

$$\Sigma(\xi) = \Sigma(\xi_\theta) + M_\theta \circ D_\theta \Sigma(\eta) \tag{23}$$

where M_θ and D_θ are as defined above. Indeed, (23) follows directly from the expansion of the inner product

$$|\xi|^2 = |\xi_\theta + r_\theta \eta|^2 = |\xi_\theta|^2 + 2r_\theta \langle \xi_\theta, \eta \rangle + r_\theta^2 |\eta|^2,$$

which may also be interpreted as a Taylor series expansion. Thus, the map A_θ defined in (22) satisfies the desired property. □

We now relate the scaling property of the paraboloid to the solution operator U via the formula $\phi(z;\xi) := \langle z, \Sigma(\xi) \rangle$ for $z = (x,t) \in \mathbb{R}^{n+1}$.

Corollary 4.8 (Parabolic rescaling). *Let $\theta \subset \widehat{\mathbb{R}}^n$ be ball and suppose $f \in L^2(\mathbb{R}^n)$ satisfies* $\operatorname{supp} \hat{f} \subseteq \theta \cap B^n(0,1)$. *Then*

$$|Uf(z)| = \mathrm{rad}(\theta)^{n/2} |U\tilde{f} \circ \mathscr{L}_\theta(z)| \tag{24}$$

for some function $\tilde{f} \in L^2(\mathbb{R}^n)$ satisfying

$$\|\tilde{f}\|_{L^2(\mathbb{R}^n)} = \|f\|_{L^2(\mathbb{R}^n)} \quad \textit{and} \quad \operatorname{supp} \mathscr{F}(\tilde{f}) \subseteq B^n(0,1). \tag{25}$$

Here the linear rescaling $\mathscr{L}_\theta$ is as defined in (22).

Proof (*of Corollary 4.8*). Let ξ_θ denote the centre of θ and $r_\theta := \mathrm{rad}(\theta)$. We simply define $\tilde{f}$ via the Fourier transform by

$$\mathscr{F}(\tilde{f})(\eta) := r_\theta^{n/2} \hat{f}(\xi_\theta + r_\theta \eta),$$

so that (25) immediately holds. On the other hand, we apply a change of variables to write

$$Uf(z) = r_\theta^{n/2} \int_{\widehat{\mathbb{R}}^n} e^{i\phi(z;\xi_\theta + r_\theta \eta)} \mathscr{F}(\tilde{f})(\eta) \, \mathrm{d}\eta.$$

The remaining property (24) now follows from the identity

$$\phi(z;\xi_\theta + r_\theta \eta) = \langle z, \Sigma(\xi_\theta + r_\theta \eta) \rangle = \langle z, \Sigma(\xi_\theta) \rangle + \langle \mathscr{L}_\theta z, \Sigma(\eta) \rangle,$$

which is itself a simple consequence of (23) and the definition of $\mathscr{L}_\theta$. □

4.3. Wave packets

We now analyse solutions to the Schrödinger equation for a class of very simple, well-behaved initial data. We shall see that the behaviour of our solutions is closely related to the geometry of the paraboloid Σ.

Example 4.9 (Unit-scale localised datum). Fix $\psi \in \mathscr{S}(\mathbb{R}^n)$ with non-negative Fourier transform satisfying

$$\operatorname{supp} \hat{\psi} \subseteq B^n(0,4) \quad \text{and} \quad \hat{\psi}(\xi) = 1 \quad \text{for all } \xi \in B^n(0,2).$$

By a simple integration-by-parts argument, the propagator

$$e^{it\Delta}\psi(x) = \frac{1}{(2\pi)^n} \int_{\widehat{\mathbb{R}}^n} e^{i\langle x,\xi\rangle} \left(e^{it|\xi|^2} \hat{\psi}(\xi) \right) \mathrm{d}\xi$$

satisfies

$$|e^{it\Delta}\psi(x)| \lesssim_N (1+|x|)^{-N} \qquad \text{for all } N \in \mathbb{N}_0 \text{ whenever } |t| \leqslant 1. \tag{26}$$

Indeed, the point here is that for $|t| \leqslant 1$, the functions $e^{it|\cdot|^2}\hat{\psi}$ form a uniformly bounded family in the Schwartz class.

Example 4.10 (Wave packets). We now rescale and translate the unit-scale localised datum ψ to create a rich family of examples. Let $\rho \geqslant 1$ and $\theta_T \subset B^n(0,1)$ be a ball of radius $\rho^{-1/2}$ centred at $\xi_T \in \widehat{\mathbb{R}}^n$. Consider the L^2-normalised function ψ_{θ_T} given by[(10)]

$$\hat{\psi}_{\theta_T}(\xi) := \rho^{n/4} \hat{\psi}\big(\rho^{1/2}(\xi - \xi_T)\big). \tag{27}$$

In addition, for $x_T \in \mathbb{R}^n$ we consider the translated datum

$$\psi_T(x) := \psi_{\theta_T}(x - x_T). \tag{28}$$

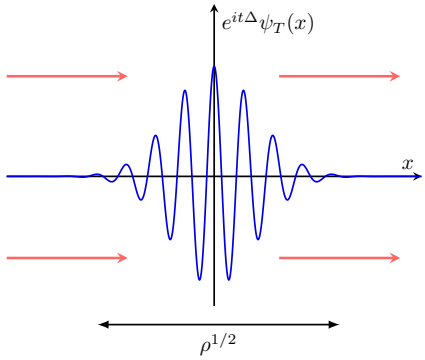
Thus, ψ_T corresponds to a modulated bump function spatially localised in a ball of radius $\rho^{1/2}$ centred at x_T and has frequency support lying in $4 \cdot \theta_T$.

It follows from parabolic rescaling in the form of Corollary 4.8, translation invariance and (26) that

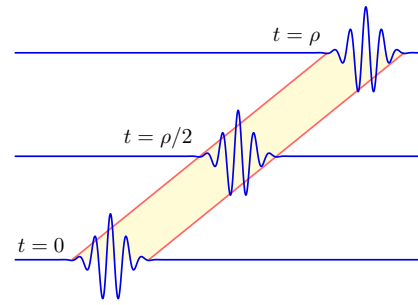
$$|e^{it\Delta}\psi_T(x)| \lesssim_N (1 + \rho^{-1/2}|x - x_T + 2t\xi_T|)^{-N} \qquad \text{for all } N \in \mathbb{N}_0 \text{ whenever } |t| \leqslant \rho.$$

Thus, during the time interval $|t| \leqslant \rho$, the solution at time t is concentrated in the spatial ball $B(x_T - 2t\xi_T, \rho^{1/2})$, in the sense that $e^{it\Delta}\psi_T$ rapidly decays away from this set. We illustrate this phenomenon in 1 spatial dimension in Figure 1a. This solution is an example of what is known as a *wave packet*.

[(10)] The reason for the apparently superfluous subscript T will be made clear below.



(a) A wave packet at time t. The wave packet is concentrated in an interval in the physical space and, as time evolves, travels with fixed velocity $v(T) = -2\xi_T$ proportional to the frequency ξ_T.



(b) A space-time tube T. The tube describes the spatial localistion of the solution $U\psi_T$ at each time slice. Three such time slices are illustrated in the figure.

Figure 1: Two perspectives on wave packets.

We highlight the basic properties of the wave packets from Example 4.10:

- ▷ The initial datum ψ_T (and the solution $e^{it\Delta}\psi_T$ at each fixed time) is frequency supported in a ball centred at ξ_T of radius $\sim \rho^{-1/2}$;
- ▷ For each time t satisfying $|t| \leqslant \rho$ the solution is concentrated in a spatial ball of radius $\rho^{1/2}$;
- ▷ The spatial ball travels from the initial position x_T at $t = 0$ along a linear trajectory with velocity $-2\xi_T$. In particular, the velocity is determined by the frequency.

This illustrates the fundamental *dispersion relation* between the velocity v of a Schrödinger wave and its frequency ξ, summed up by the formula

$$v = -2\xi.$$

The fact that waves of different frequency travel with different velocities (and therefore disperse at large time scales) is the defining characteristic of *dispersive PDE*.

Remark 4.11. For our wave packets ψ_T, we have localised the frequency at scale $\rho^{-1/2}$ so that the distinct frequency modes making up the wave $e^{it\Delta}\psi_T$ travel at the same velocity up to an error of $O(\rho^{-1/2})$. Consequently, the wave packet remains 'stable' for the time scale $|t| \leqslant \rho$. Beyond this time scale, the difference in velocities between the distinct modes causes the wave packet to disperse.

In view of what follows, it is useful to adopt a new perspective and visualise the wave packet $e^{it\Delta}\psi_T(x)$ as a function on the spatio-temporal domain $\mathbb{R}^{n+1}$.

Definition 4.12. For $\rho \geqslant 1$, a *space-time ρ-tube*, or simply a *ρ-tube*, is a set $T \subset \mathbb{R}^{n+1}$ of the form

$$T := \big\{(x,t) \in \mathbb{R}^{n+1} : |x - x(T) - tv(T)| \leqslant \rho^{1/2} \quad \text{and} \quad |t| \leqslant \rho\big\}$$

for some $x(T), v(T) \in \mathbb{R}^n$. In this case, we say T has initial position $x(T)$, velocity $v(T)$, duration ρ and spatial radius $\rho^{1/2}$.

As a function of (x,t), we see that the wave packet[(11)] $U\psi_T(x,t)$ is concentrated on the space-time tube T centred at x_T with velocity $v(T) = -2\xi_T$, duration ρ and spatial radius $\rho^{1/2}$; see Figure 1b. In particular, if for $0 < \delta < 1$ we define the slightly enlarged space-time tube

$$T^{(\delta)} := \big\{(x,t) \in \mathbb{R}^{n+1} : |x - x(T) - tv(T)| \leqslant \rho^{1/2+\delta} \quad \text{and} \quad |t| \leqslant \rho\big\}, \tag{29}$$

then

$$|U\psi_T(x,t)| \lesssim_{\delta,N} \rho^{-N} \qquad \text{if } (x,t) \in B^{n+1}(0,\rho) \setminus T^{(\delta)}. \tag{30}$$

We update our notation to accommodate our change in perspective. If $T \subset \mathbb{R}^{n+1}$ is a space-time ρ-tube as in Definition 4.12, then we define $\theta_T \subset \widehat{\mathbb{R}}^n$ to be the ball centred at $\xi_T := -v(T)/2$ of side-length $\rho^{-1/2}$. Moreover, we let ψ_T denote the function (28) from Example 4.10.

Note that a ρ-tube T is centred around its *core line*

$$\ell_T := \{(x,t) \in \mathbb{R}^{n+1} : x = x_T - 2t\xi_T\}.$$

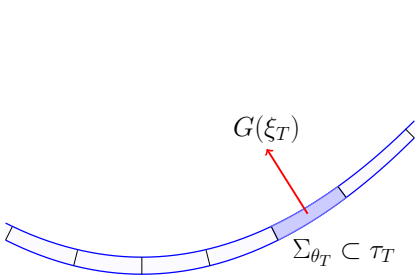
This has direction $G(\xi_T)$ where

$$G(\xi) := \frac{1}{\sqrt{1+4|\xi|^2}} \begin{pmatrix} -2\xi \\ 1 \end{pmatrix}; \tag{31}$$

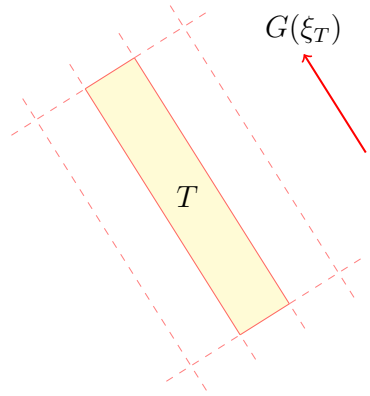
we will also refer to $G(\xi_T)$ as the *direction* of T. Note that G is precisely the Gauss map associated to the paraboloid Σ. Thus, the direction of the tube T corresponds to the normal vector to the paraboloid at the centre of the cap Σ_T associated to θ_T.

Remark 4.13. These observations are consistent with the uncertainty principle and should be compared with the statement of Corollary 4.2. Indeed, $e^{it\Delta}\psi_T(x)$ has spatio-temporal Fourier support in Σ_{θ_T}. This cap is itself contained in a rectangle τ_T centred at ξ_T with n sides of length $O(\rho^{-1/2})$ lying in the tangent directions of Σ at $\Sigma(\xi_T)$ and a remaining side of length $O(\rho^{-1})$ lying in the normal direction. The (essential) spatial-temporal support T therefore corresponds to a translate of the dual τ_T^* of the frequency support τ_T. See Figure 2.

[(11)] By the hypothesis $\theta \subseteq B^n(0,1)$, we have $e^{it\Delta}\psi_T(x) = U\psi_T(x,t)$.



(a) Each $\rho^{-1/2}$-cap θ_T is contained in a rectangle τ_T of dimension approximately $\rho^{-1/2} \times \cdots \times \rho^{-1/2} \times \rho^{-1}$, with short side in the direction of $G(\xi_T)$.



(b) The space-time ρ-tube T associated to the cap θ_T. This tube has dimensions $\rho^{1/2} \times \cdots \times \rho^{1/2} \times \rho$, with long side in the direction of $G(\xi_T)$.

Figure 2: The wave packets decomposition respects the uncertainty principle.

4.4. Wave packet decomposition

The examples considered of the previous section appear rather specialised. Nevertheless, using a Fourier series decomposition, *any* initial datum $f \in L^2(\mathbb{R}^n)$ can be expressed as a superposition

$$f = \sum_{T \in \mathbf{T}} a_T \psi_T$$

where $\mathbf{T}$ is a (possibly infinite) collection of space-time tubes; the ψ_T are basic initial data introduced above and $(a_T)_{T \in \mathbf{T}}$ is a sequence of complex coefficients. Consequently,

$$e^{it\Delta} f = \sum_{T \in \mathbf{T}} a_T e^{it\Delta} \psi_T$$

and so any solution can be realised as a superposition of wave packets. This observation is referred to as the *wave packet decomposition*. It is essentially a consequence of the uncertainty principle and is closely related to the local constancy property described in Corollary 4.2.

Definition and basic properties. — Turning to the precise details, we first introduce some notation.

Definition 4.14. For $\rho \geqslant 1$, let $\mathbf{T}[\rho]$ denote the collection of all space-time ρ-tubes as in Definition 4.12 where

▷ The initial position $x(T)$ is free to vary over the lattice $\rho^{1/2}\mathbb{Z}^n$;

▷ The velocity $v(T)$ is free to vary over the lattice $c_n\rho^{-1/2}\mathbb{Z}^n$.

Here $c_n := 1/2n^{1/2}$ is a fixed constant which plays a minor technical rôle.

The precise form of the wave packet decomposition is given by the following lemma.

Lemma 4.15 (Wave packet decomposition). *Given $f \in L^2(\mathbb{R}^n)$ with $\operatorname{supp} \hat{f} \subseteq B^n(0,1/2)$ and $\rho \geqslant 10$ we may write*

$$f = \sum_{T \in \mathbf{T}[\rho]} f_T \tag{32}$$

where the functions $f_T \in \mathscr{S}(\mathbb{R}^n)$ satisfy the following:

i) ***Dispersion relation.*** *Let $T \in \mathbf{T}[\rho]$ and suppose f_T is not identically 0. Then*

$$|v(T) + 2\xi| \leqslant 2\rho^{-1/2} \qquad \text{for some } \xi \in \operatorname{supp} \hat{f}.$$

ii) ***Orthogonality.*** *For any collection $\mathbf{W} \subseteq \mathbf{T}[\rho]$, we have*

$$\Big\| \sum_{T \in \mathbf{W}} f_T \Big\|^2_{L^2(\mathbb{R}^n)} \lesssim \sum_{T \in \mathbf{W}} \|f_T\|^2_{L^2(\mathbb{R}^n)} \lesssim \|f\|^2_{L^2(\mathbb{R}^n)}. \tag{33}$$

iii) ***Spatio-temporal localisation.*** *If for $T \in \mathbf{T}[\rho]$ and $0 < \delta < 1$ we define the slightly enlarged space-time tube $T^{(\delta)}$ as in (29), then*

$$|Uf_T(x,t)| \lesssim_{\delta,N} \rho^{-N} \|f\|_{L^2(\mathbb{R}^n)} \qquad \text{if } (x,t) \in B^{n+1}(0,\rho) \setminus T^{(\delta)}. \tag{34}$$

We refer to (32) as the wave packet decomposition of f at scale ρ.

The idea is to first decompose f as a sum of pieces f_θ, each frequency supported in some cap θ. The individual f_θ are then further decomposed using Fourier series. Before giving the details, we fix some notation. For $\rho \geqslant 1$ and let

$$\Theta[\rho] := \big\{ B(\xi_T, \rho^{-1/2}) : \xi_T \in c_n\rho^{-1/2}\mathbb{Z}^n \cap B^n(0,2) \big\}$$

denote a covering of $B^n(0,2)$ by $\rho^{-1/2}$-caps. As at the end of §4.3, we associate to each $T \in \mathbf{T}[\rho]$ a cap $\theta_T \in \Theta[\rho]$ with centre ξ_T satisfying the dispersion relation $v(T) = -2\xi_T$.

Fix $\phi \in \mathscr{S}(\mathbb{R}^n)$ satisfying[(12)]

$$\operatorname{supp} \hat{\phi} \subseteq [-c_n, c_n]^n \quad \text{and} \quad \sum_{k \in \mathbb{Z}^n} \hat{\phi}(\xi - c_n k) = 1 \quad \text{for all } \xi \in \widehat{\mathbb{R}}^n.$$

[(12)] To construct such a function, choose $\beta \in C_c^\infty(\mathbb{R}^n)$ such that $\operatorname{supp} \beta \subseteq Q_0 := [-1/2, 1/2]^n$ and $\int_{\mathbb{R}} \beta = 1$. Since $\sum_{k \in \mathbb{Z}^n} \chi_{Q_0} * \beta(\,\cdot\, - k) \equiv 1$, we may take $\hat{\phi}$ to be a suitable scaling of $\chi_{Q_0} * \beta$.

Furthermore, let $\psi \in \mathscr{S}(\mathbb{R}^n)$ be as in Example 4.9, so that

$$\hat{\phi}(\xi) = \hat{\phi}(\xi)\hat{\psi}(\xi) \qquad \text{for all } \xi \in \widehat{\mathbb{R}}^n. \tag{35}$$

Given a cap $\theta_T \in \Theta[\rho]$ with centre ξ_T, define $\phi_{\theta_T} \in \mathscr{S}(\mathbb{R}^n)$ by

$$\hat{\phi}_{\theta_T}(\xi) := \hat{\phi}(\rho^{1/2}(\xi - \xi_T)).$$

The family of functions $\hat{\phi}_{\theta_T}$ then forms a partition of unity subordinate to the covering $\Theta[\rho]$ of $B^n(0,2)$. Consequently,

$$f = \sum_{\theta_T \in \Theta[\rho]} f_{\theta_T} \qquad \text{where} \qquad f_{\theta_T} := \phi_{\theta_T} * f. \tag{36}$$

With these definitions, we turn to the proof of the wave packet decomposition.

Proof (*of Lemma 4.15*). For f_{θ_T} as defined in (36), it follows that $\hat{f}_{\theta_T}$ is compactly supported in a cube of side-length $\rho^{-1/2}$. Consequently, we can expand $\hat{f}_{\theta_T}$ as a (suitably scaled) Fourier series, giving

$$\hat{f}_{\theta_T}(\xi) = \sum_{k \in \rho^{1/2}\mathbb{Z}^n} \rho^{n/2} e^{-i\langle \xi, k\rangle} f_{\theta_T}(k) = \sum_{k \in \rho^{1/2}\mathbb{Z}^n} \rho^{n/4} e^{-i\langle \xi, k\rangle} f_{\theta_T}(k) \hat{\psi}_{\theta_T}(\xi). \tag{37}$$

Here the convergence can be taken in the L^2 sense and ψ_θ is as defined in (27). For the second step in (37) we have used (35). Note, by Plancherel's theorem,

$$\sum_{k \in \rho^{1/2}\mathbb{Z}^n} |f_{\theta_T}(k)|^2 = \rho^{-n/2} \|f_\theta\|_{L^2(\mathbb{R}^n)}^2.$$

In light of the above, we may write

$$f = \sum_{T \in \mathbf{T}[\rho]} f_T \quad \text{where} \quad f_T(x) := \rho^{n/4} f_{\theta_T}(x(T)) \hat{\psi}_T(x)$$

for ψ_T as defined in (28). It remains to prove the functions f_T have the desired properties.

i) **Dispersion relation.** From the definitions, $\operatorname{supp} \hat{\phi}_{\theta_T}$ lies in the ball θ_T of radius $\rho^{-1/2}$. If f_T is not identically zero, then f_{θ_T} is not identically zero. In this case, θ_T must intersect the frequency support of f, which immediately implies the desired property.

ii) **Orthogonality.** For each $\theta \in \Theta[\rho]$, let $\mathbf{T}_\theta[\rho]$ denote the collection of all tubes $T \in \mathbf{T}[\rho]$ associated to θ and $\mathbf{W}_\theta := \mathbf{W} \cap \mathbf{T}_\theta[\rho]$. Since the functions $\sum_{T \in \mathbf{W}_\theta} f_T$ have finitely overlapping Fourier support,

$$\Big\| \sum_{T \in \mathbf{W}} f_T \Big\|_{L^2(\mathbb{R}^n)}^2 = \Big\| \sum_{\theta \in \Theta[\rho]} \sum_{T \in \mathbf{W}_\theta} f_T \Big\|_{L^2(\mathbb{R}^n)}^2 \lesssim \sum_{\theta \in \Theta[\rho]} \Big\| \sum_{T \in \mathbf{W}_\theta} f_T \Big\|_{L^2(\mathbb{R}^n)}^2, \tag{38}$$

where in the final step we use Plancherel's theorem and the Cauchy–Schwarz inequality.

Fix $\theta \in \Theta[\rho]$ and note that $\operatorname{supp} \hat{\psi}_\theta$ lies in the cube $\tilde{\theta}$ concentric to θ with side-length $8\rho^{-1/2}$. In particular, by Plancherel's theorem (and periodicity),

$$\Big\| \sum_{T\in\mathbf{W}_\theta} f_T \Big\|^2_{L^2(\mathbb{R}^n)} \lesssim \rho^n \Big\| \sum_{T\in\mathbf{W}_\theta} f_\theta(x(T)) e^{-i\langle \xi, x(T)\rangle} \Big\|^2_{L^2(\tilde{\theta})} \sim \rho^{n/2} \sum_{T\in\mathbf{W}_\theta} |f_\theta(x(T))|^2. \quad (39)$$

However, from the definition of f_T and the choice of normalisation of ψ_T, it follows that

$$\rho^{n/2} |f_\theta(x(T))|^2 \sim \|f_T\|^2_{L^2(\mathbb{R}^n)} \qquad \text{for any } T \in \mathbf{T}_\theta[\rho]. \quad (40)$$

Combining (38), (39) and (40), we deduce

$$\Big\| \sum_{T\in\mathbf{W}} f_T \Big\|^2_{L^2(\mathbb{R}^n)} \lesssim \sum_{T\in\mathbf{W}} \|f_T\|^2_{L^2(\mathbb{R}^n)},$$

which is the first of the desired inequalities.

On the other hand, in view of (37) and (40), we have

$$\sum_{T\in\mathbf{W}} \|f_T\|^2_{L^2(\mathbb{R}^n)} \lesssim \rho^{n/2} \sum_{\theta\in\Theta[\rho]} \sum_{T\in\mathbf{W}_\theta} |f_\theta(x(T))|^2 \lesssim \sum_{\theta\in\Theta[\rho]} \|f_\theta\|^2_{L^2(\mathbb{R}^n)} \lesssim \|f\|_{L^2(\mathbb{R}^n)},$$

where the last inequality follows since the f_θ has finitely-overlapping frequency support.

iii) **Spatio-temporal locality.** It follows from (30) that

$$|Uf_T(x,t)| \lesssim_N \rho^{-N} |f_{\theta_T}(x(T))| \qquad \text{if } (x,t) \in B^{n+1}(0,\rho) \setminus T^{(\delta)}. \quad (41)$$

However, since f_{θ_T} has Fourier support in a small ball, it follows from Bernstein's inequality and the orthogonality properties of the wave packets that

$$|f_{\theta_T}(x(T))| \leqslant \|f_{\theta_T}\|_{L^\infty(\mathbb{R}^n)} \lesssim \|f_{\theta_T}\|_{L^2(\mathbb{R}^n)} \lesssim \|f\|_{L^2(\mathbb{R}^n)}. \quad (42)$$

Combining (42) and (41) gives the desired bound. □

We can use Lemma 4.15 to address a point left open from §3. The following argument is essentially taken from Lee (2006).

Proof (*Proposition 3.3 ⇒ Theorem 3.1*). Assume Proposition 3.3 holds. For all $\varepsilon > 0$ and $R \geqslant 1$, it suffices to show

$$\Big\| \sup_{0<t<R^2} |e^{it\Delta} f| \Big\|_{L^2(B^n(0,R))} \lesssim_\varepsilon R^{n/(2(n+1))+\varepsilon} \|f\|_{L^2(\mathbb{R}^n)} \quad (43)$$

for all $f \in L^2(\mathbb{R}^n)$ with $\operatorname{supp} \hat{f} \subseteq A(1)$. Indeed, as discussed in §3.3, the desired result then follows from a simple scaling argument and the Littlewood–Paley characterisation of $H^s(\mathbb{R}^n)$.

Fix $\varepsilon > 0$, $R \geqslant 1$ and $f \in L^2(\mathbb{R}^n)$ with $\operatorname{supp} \hat{f} \subseteq A(1)$. Let $\mathscr{I}_R$ be the collection of all lattice R-intervals which intersect the long time interval $[0, R^2]$. Trivially, we have

$$\begin{aligned} \| \sup_{0<t\leqslant R^2} |e^{it\Delta} f| \|_{L^2(B^n(0,R))} &= \| \sup_{I\in\mathscr{I}_R} \sup_{t\in I} |e^{it\Delta} f| \|_{L^2(B^n(0,R))} \\ &\leqslant \Big(\sum_{I\in\mathscr{I}_R} \| \sup_{t\in I} |e^{it\Delta} f| \|^2_{L^2(B^n(0,R))} \Big)^{1/2}. \end{aligned} \tag{44}$$

On the other hand, by Proposition 3.3 and the invariance of the estimates under temporal translation,

$$\| \sup_{t\in I} |e^{it\Delta} g| \|_{L^2(B^n(0,R))} \lesssim_\varepsilon R^{n/(2(n+1))+\varepsilon/2} \|g\|_{L^2(\mathbb{R}^n)} \qquad \text{for all } I \in \mathscr{I}_R \tag{45}$$

whenever $g \in L^2(\mathbb{R}^n)$ with $\operatorname{supp} \hat{g} \subseteq A(1)$.

In order to sum the estimates from (45) we observe a certain orthogonality between the localised maximal operators associated to distinct time intervals I. We first discuss this orthogonality at a heuristic level. As above, let $\mathbf{T}[R]$ denote the collection of all space-time R-tubes. Since $\operatorname{supp} \hat{f} \subseteq A(1)$, given $T \in \mathbf{T}[R]$ the wave packet $e^{it\Delta} f_T$ has speed $|v(T)| \sim 1$. This means that the wave $e^{it\Delta} f_T$ spends roughly R units of time in the spatial ball $B^n(0, R)$. Hence, for each $T \in \mathbf{T}[R]$ there is essentially a unique time interval $I \in \mathscr{I}_R$ for which $\sup_{t\in I} |e^{it\Delta} f_T|$ is non-negligible. Thus, each wave packet contributes only to a single term in the sum on the right-hand side of (44); since the functions $(f_T)_{T\in\mathbf{T}[R]}$ are themselves orthogonal, this leads to the desired orthogonality between the maximal operators.

We now turn to the formal details of the proof. Fix $\delta := \varepsilon/2$ and for each $I \in \mathscr{I}_R$, let $Q_I := B^n(0, R) \times I$ and $Q_I^{(\delta)} := R^\delta \cdot Q_I$. Define

$$\mathbf{T}_I[R] := \big\{ T \in \mathbf{T}[R] : T \cap Q_I^{(\delta)} \neq \emptyset \text{ and } 1/4 \leqslant |v(T)| \leqslant 2 \big\} \tag{46}$$

so that, by frequency support hypotheses on f and the spatio-temporal locality of the wave packets,

$$\| \sup_{t\in I} |e^{it\Delta} f| \|_{L^2(B^n(0,R))} \lesssim_\varepsilon \| \sup_{t\in I} |e^{it\Delta} f_I| \|_{L^2(B^n(0,R))} + R^{-100n} \|f\|_{L^2(\mathbb{R}^n)} \tag{47}$$

where

$$f_I := \sum_{T\in\mathbf{T}_I[R]} f_T.$$

The key observation is that the sets $\mathbf{T}_I[R]$ are essentially disjoint: more precisely,

$$\max_{T\in\mathbf{T}[R]} \#\{ I \in \mathscr{I}_R : T \in \mathbf{T}_I[R] \} \lesssim R^\delta. \tag{48}$$

Indeed, once we have (48), we may combine (44), (45) and (47) to deduce that

$$\begin{aligned}\| \sup_{0<t\leqslant R^2} |e^{it\Delta}f|\|_{L^2(B^n(0,R))} &\lesssim_\varepsilon R^{n/(2(n+1))+\varepsilon/2}\Big(\sum_{I\in\mathscr{I}_R}\|f_I\|^2_{L^2(\mathbb{R}^n)}\Big)^{1/2} + R^{-100n}\|f\|_{L^2(\mathbb{R}^n)}\\ &\lesssim_\varepsilon R^{n/(2(n+1))+\varepsilon}\Big(\sum_{T\in\mathbf{T}[R]}\|f_T\|^2_{L^2(\mathbb{R}^n)}\Big)^{1/2} + R^{-100n}\|f\|_{L^2(\mathbb{R}^n)}\end{aligned}$$

where the second step follows from orthogonality, interchanging the order of summation and (48). The desired estimate (43) now follows by the orthogonality property of the wave packets.

It remains to show (48). Fix $T \in \mathbf{T}[R]$ with initial position $x(T)$ and velocity $v(T)$. Suppose $T \in \mathbf{T}_{I_1}[R] \cap \mathbf{T}_{I_2}[R]$ for a pair of time intervals $I_1, I_2 \in \mathscr{I}_R$. It suffices to show $\operatorname{dist}(I_1, I_2) \leqslant 30R^{1+\delta}$. Suppose this is not the case and let $\bar{t}_j$ denote the centre of I_j for $j = 1, 2$. By the definition of the sets $\mathbf{T}_{I_j}[R]$, there exist $(x_j, t_j) \in \mathbb{R}^{n+1}$ such that

$$|x_j - x(T) - t_j v(T)| \leqslant R^{1/2}, \quad |x_j| \leqslant R^{1+\delta} \quad \text{and} \quad |t_j - \bar{t}_j| \leqslant R^{1+\delta}$$

for $j = 1, 2$. By the triangle inequality,

$$|\bar{t}_1 - \bar{t}_2||v(T)| \leqslant |x_1 - x_2| + \sum_{j=1}^{2} |x_j - x(T) - t_j v(T)| + |t_j - \bar{t}_j| \leqslant 6R^{1+\delta}.$$

Since, by hypothesis, $|\bar{t}_1 - \bar{t}_2| \geqslant 30R^{1+\delta}$, it follows that $|v(T)| \leqslant 1/10$. In view of the definition of $\mathbf{T}_{I_j}[R]$ from (46), this is a contradiction. □

4.5. Pseudo-local properties

Throughout the following, we fix a large spatio-temporal scale $R \geqslant 1$. This scale plays a similar rôle to R in the statement of Theorem 3.7 and, in particular, we shall consider estimates for Uf which are localised in space-time to $B^{n+1}(0, R)$. For this reason, it is useful to perform a wave packet decomposition at scale R, so that the corresponding space-time tubes have duration R.

We shall often work with an additional, much smaller, intermediate scale $1 \leqslant K \leqslant R^{1/2}$; this should be thought of as reasonably large, but still vastly smaller than R. For instance, in applications we typically either take K to be a large dimensional constant or $K = R^\delta$ for some very small $\delta > 0$.

Let $\mathscr{T}_{K^{-1}}$ be a finitely-overlapping covering of $B^n(0,1)$ by $(2K)^{-1}$-caps with centres lying in $B^n(0,2)$. Fix $\tau \in \mathscr{T}_{K^{-1}}$ with $\tau \subseteq B^n(0,1)$ and suppose $f_\tau \in L^2(\mathbb{R}^n)$ satisfies $\operatorname{supp} \hat{f}_\tau \subseteq \tau$.[13] By applying a wave packet decomposition to f_τ at scale R, we have

$$f_\tau = \sum_{T\in\mathbf{T}_\tau[R]} f_T \qquad \text{where} \qquad \mathbf{T}_\tau[R] := \big\{T \in \mathbf{T}[R] : |v(T) + 2\xi_\tau| \leqslant 4K^{-1}\big\};$$

here the dispersion relation property from Lemma 4.15 is used to restrict the velocities.

[13] The caps are chosen to have radius $(2K)^{-1}$ for technical reasons; one should think of f_τ as supported 'well within' a cap of radius K^{-1}.

Each $T \in \mathbf{T}_\tau[R]$ is aligned in the direction $G(\xi_\tau)$ up to an angular difference of $O(K^{-1})$. Consequently, we can group the wave packets together according to a partition of space-time into disjoint strips S of dimension $R/K \times \cdots \times R/K \times R$: see Figure 3. In view of this, the solution operator U essentially acts independently between distinct strips.

To describe these properties more precisely, let $\mathbf{S}_\tau[R]$ denote the collection of all strips

$$S = \big\{(x,t) \in \mathbb{R}^{n+1} : |x - x(S) - tv(S)| \leqslant R/K \quad \text{and} \quad |t| \leqslant R\big\}$$

where $x(S) \in \mathbb{R}^n$ is a choice of lattice point in $R/K \cdot \mathbb{Z}^n$ and $v(S) := -2\xi_\tau$. Form a partition $\mathbf{T}_S[R]$ of $\mathbf{T}_\tau[R]$, parameterised by the the strips $S \in \mathbf{S}_\tau[R]$ such that

$$\mathbf{T}_S[R] \subseteq \big\{T \in \mathbf{T}_\tau[R] : T \cap S \neq \emptyset\big\}.$$

Then we may write

$$f_\tau = \sum_{S \in \mathbf{S}_\tau[R]} f_S \qquad \text{where} \qquad f_S = \sum_{T \in \mathbf{T}_S[R]} f_T \quad \text{for all } S \in \mathbf{S}_\tau[R]. \tag{49}$$

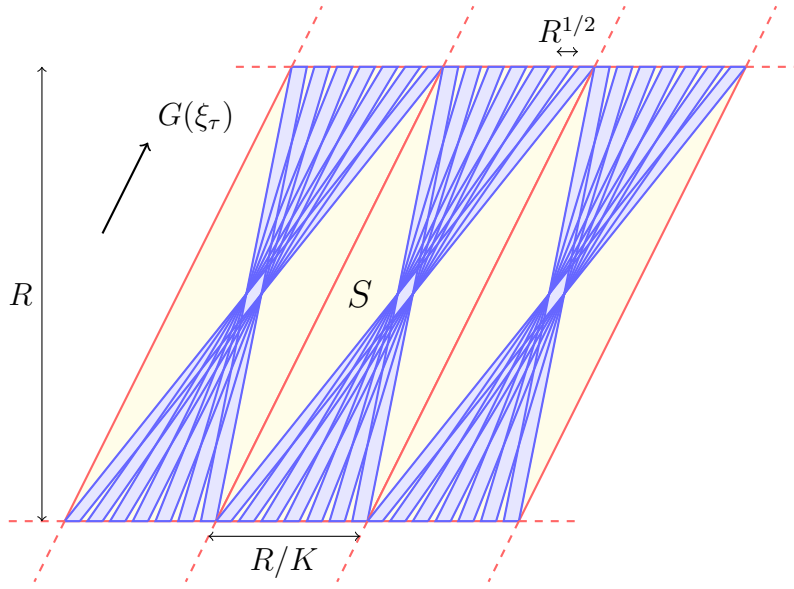


Figure 3: A schematic demonstrating the pseudolocal property. The tubes T arising from the wave packet decomposition of f_τ have directions lying in a K^{-1}-neighbourhood of $G(\xi_\tau)$. Consequently, they can (essentially) be partitioned into families which lie in disjoint strips S.

Morally, if $S_1, S_2 \in \mathbf{S}_\tau[R]$ are distinct strips, then the waves Uf_{S_1} and Uf_{S_2} concentrate on S_1 and S_2, respectively, and therefore do not interact. To make this statement precise, we define the slightly enlarged strip

$$\bar{S} = \big\{(x,t) \in \mathbb{R}^{n+1} : |x - x(S) - tv(S)| \leqslant 20R/K \quad \text{and} \quad |t| \leqslant R\big\}.$$

Then we have the following lemma.

Lemma 4.16. *Suppose $0 < \delta < 1$ and $1 \leqslant K \leqslant R^{1/2-\delta}$. With the above definitions, for all $S \in \mathbf{S}_\tau[R]$ we have*

$$|Uf_S(z)| \lesssim_{\delta,N} |Uf_S(z)|\chi_{\bar{S}}(z) + R^{-N}\|f_\tau\|_{L^2(\mathbb{R}^n)} \qquad \text{for all } z \in B^{n+1}(0,R) \text{ and } N \in \mathbb{N}.$$

Proof. Let $S \in \mathbf{S}_\tau[R]$ and $T \in \mathbf{T}_S[R]$. By the localisation of the wave packets from (34), it suffices to show $T^{(\delta)} \subseteq \bar{S}$, where $T^{(\delta)}$ is the enlargement of T defined in (29).

Let $z = (x,t) \in B^{n+1}(0,R) \setminus \bar{S}$, so that

$$|x - x(S) - tv(S)| > 20R/K.$$

Since $T \in \mathbf{T}_S[R]$, it follows that $|v(T) - v(S)| \leqslant 4K^{-1}$ and there exists some $(x',t') \in \mathbb{R}^{n+1}$ satisfying

$$|x' - x(T) - t'v(T)| \leqslant R^{1/2}, \quad |x' - x(S) - t'v(S)| \leqslant R/K \quad \text{and} \quad |t'| \leqslant R.$$

Therefore, by the triangle inequality, $|x(T) - x(S)| \leqslant 6R/K$. Consequently,

$$|x - x(T) - tv(T)| \geqslant R/K > R^{1/2+\delta}$$

and so $(x,t) \notin T^{(\delta)}$, as required. □

Remark 4.17. Note that the time zero slice $B_S := S \cap \{(x,0) : x \in \mathbb{R}^n\}$ is an R/K-ball. Another interpretation of the preceding observations is that $S \in \mathbf{S}_\tau[R]$ essentially corresponds to the domain of influence for the initial datum f_τ localised to B_S, over the time interval $[-R,R]$.

For $S \in \mathbf{S}_\tau[R]$ define the mapping

$$\mathscr{A}_S \colon (x,t) \mapsto (K^{-1}(x - x(S) - tv(S)), K^{-2}t).$$

Thus, $\mathscr{A}_S$ maps the strip S to the ball $B^{n+1}(0, R/K^2)$; see Figure 4. Furthermore, we may write $\mathscr{A}_S(z) = \mathscr{L}_\tau(z) - z(S)$ where $\mathscr{L}_\tau$ is the scaling map associated to the cap τ as defined in (22) and $z(S) := (x(S), 0)$. In light of these observations, we obtain the following L^p variant of the parabolic rescaling from Corollary 4.8.

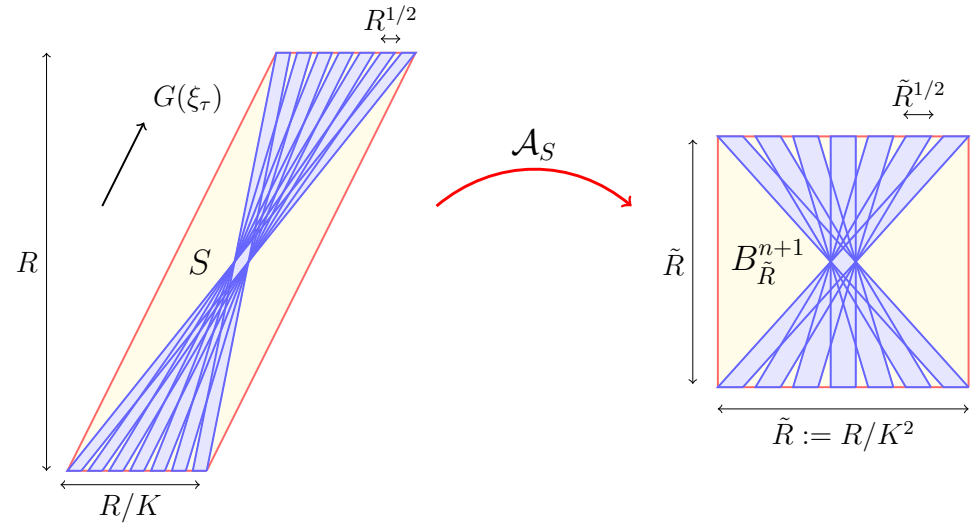


Figure 4: Rescaling in Lemma 4.18 in the (physical) space-time domain. The strip S is mapped to $B^{n+1}(0,\tilde{R})$ under the affine map $\mathscr{A}_S$.

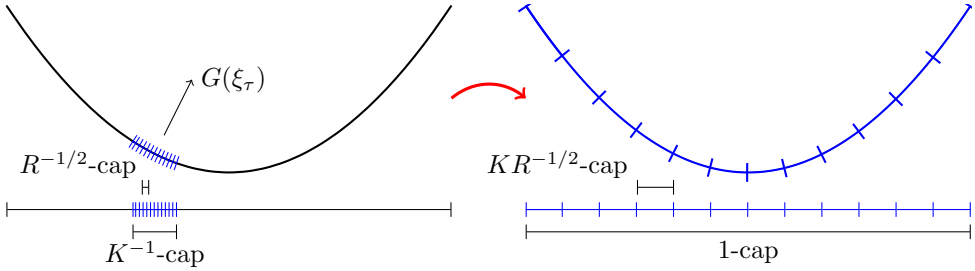


Figure 5: Rescaling in Lemma 4.18 in the frequency domain. The K^{-1}-cap τ is mapped to the whole paraboloid. Note the relationship between the directions of the space-time tubes in Figure 4 and the underlying frequencies.

Lemma 4.18. *Let $1 \leqslant p \leqslant \infty$. With the above setup,*

$$\|Uf_S\|_{L^p(\bar{S})} \leqslant K^{(n+2)/p-n/2}\|U\tilde{f}_S\|_{L^p(B^{n+1}(0,\tilde{R}))}$$

for $\tilde{R} = 20R/K^2$ and some function $\tilde{f}_S \in L^2(\mathbb{R}^n)$ satisfying

$$\|\tilde{f}_S\|_{L^2(\mathbb{R}^n)} = \|f_S\|_{L^2(\mathbb{R}^n)} \quad \textit{and} \quad \operatorname{supp}\mathscr{F}(\tilde{f}_S) \subseteq B^n(0,1).$$

Proof. This follows by applying parabolic rescaling in the form of Corollary 4.8, exploiting the relationship between $\mathscr{L}_\tau$ and $\mathscr{A}_S$ and changing the spatio-temporal variables. The scaling is represented in the physical and frequency domains in Figure 4 and Figure 5, respectively.[(14)] □

4.6. Dyadic pigeonholing and reverse Hölder

We shall make extensive use of pigeonholing arguments in the proof of Theorem 3.7 in §7. Although such arguments are entirely elementary, they are nevertheless surprisingly useful and it is worth discussing the *dyadic pigeonholing* method in particular.

Definition 4.19. We say $\mathscr{B} \subseteq (0,\infty)$ is *dyadically constant* if there exists some $j \in \mathbb{Z}$ such that $\mathscr{B} \subseteq [2^j, 2^{j+1}]$.

Remark 4.20. Typically, $\mathscr{B} = \{H(a) : a \in \mathscr{A}\}$ is a sequence of positive numbers indexed over some finite set $\mathscr{A}$; in such cases we shall often write

$$H(a) \qquad \text{is dyadically constant over } a \in \mathscr{A}$$

to mean $\mathscr{B}$ is dyadically constant.

Let $M > 0$, $R \geqslant 1$. By taking logarithms, we see that there are only $O(\log R)$ values of $j \in \mathbb{Z}$ such that $2^j \in [MR^{-1}, MR]$. It follows that any set $\mathscr{B} \subseteq [MR^{-1}, MR]$ can be written as a union

$$\mathscr{B} = \bigcup_{j=1}^{J} \mathscr{B}_j \tag{50}$$

where each $\mathscr{B}_j$ is dyadically constant and $J \lesssim \log R$. Applying the pigeonhole principle to the sets $\mathscr{B}_j$ arising from this decomposition, we deduce the following elementary (but surprisingly powerful) lemma.

Lemma 4.21 (Dyadic pigeonholing). *Let $M > 0$, $R \geqslant 1$ and $\mathscr{B} \subseteq [MR^{-1}, MR]$ be a finite set.*

i) There exists some $\mathscr{B}' \subseteq \mathscr{B}$ which is dyadically constant and satisfies

$$\#\mathscr{B}' \gtrsim (\log R)^{-1} \#\mathscr{B}.$$

ii) More generally, given $F\colon \mathscr{B} \to (0,\infty)$, there exists some $\mathscr{B}_F \subseteq \mathscr{B}$ which is dyadically constant and satisfies

$$\sum_{a\in\mathscr{B}} F(a) \lesssim \log R \sum_{a\in\mathscr{B}_F} F(a).$$

[(14)] There are some minor technical complications in the proof of Lemma 4.18 owing to slightly larger frequency support of f_S (compared with f_τ) and the use of the enlarged strip $\bar{S}$ rather than S. This accounts for the factor of 20 in the definition of $\tilde{R}$.

A close relative of Lemma 4.21 is the following reverse form of Hölder's inequality.

Lemma 4.22 (Reverse Hölder). *Let $M > 0$, $R \geqslant 1$ and $H\colon \mathscr{A} \to [MR^{-1}, MR]$ be a function defined on a finite set $\mathscr{A}$. Then we may write $\mathscr{A} = \bigcup_{j=1}^{J} \mathscr{A}_j$ where*

$$\Big(\sum_{a\in\mathscr{A}_j} H(a)^p \Big)^{1/p} \lesssim \log R\, [\#\mathscr{A}_j]^{-(1/p-1/q)} \Big(\sum_{a\in\mathscr{A}_j} H(a)^q \Big)^{1/q}$$

holds for all $1 \leqslant p \leqslant q < \infty$ and all $1 \leqslant j \leqslant J$ and $J \lesssim \log R$.

Proof. Let $\mathscr{B} := \{H(a) : a \in \mathscr{A}\}$ and decompose this set into $O(\log R)$ disjoint dyadically constant pieces $\mathscr{B}_j$, $1 \leqslant j \leqslant J$, as in (50). The result then follows by taking $\mathscr{A}_j := H^{-1}(\mathscr{B}_j)$ for $1 \leqslant j \leqslant J$. □

In view of Lemma 4.22, the pigeonhole principle is useful for 'real interpolation' arguments; that is, when one wishes to reconcile distinct estimates involving different L^p norms. We shall see multiple instances of this later in §7.

5. Tools from mutlilinear harmonic analysis

5.1. Linear Strichartz estimates

Recall that our goal is to prove Theorem 3.7, which bounds the solution Uf over the union $Z_{\mathscr{Q}}$ of a family of space-time cubes. Here we consider a simpler class of estimates, which bound the solution over the whole space-time domain.

Theorem 5.1 (Strichartz estimate). *Let $2 \cdot \frac{n+2}{n} \leqslant q \leqslant \infty$. The inequality*

$$\|Uf\|_{L^q(\mathbb{R}^{n+1})} \lesssim \|f\|_{L^2(\mathbb{R}^n)} \tag{51}$$

holds for all $f \in L^2(\mathbb{R}^n)$.

This is a reinterpretation of the Stein–Tomas Fourier restriction theorem, dating back to Stein (1986), Strichartz (1977), and Tomas (1975).

Remark 5.2. Theorem 5.1 is in fact a special case of the more general *Strichartz estimates for Schrödinger equation*, which involve mixed norms in the spatio-temporal variables.

Theorem 5.1 is clearly related to the theory of fractal energy estimates. Indeed, if $\mathscr{Q}$ is a family of lattice unit cubes in $B^{n+1}(0,R)$ and $Z_{\mathscr{Q}}$ denotes their union, then

$$\|Uf\|_{L^2(Z_{\mathscr{Q}})} \leqslant |Z_{\mathscr{Q}}|^{1/(n+2)} \|Uf\|_{L^{2\cdot\frac{n+2}{n}}(\mathbb{R}^{n+1})}.$$

Applying the simple bound $|Z_{\mathscr{Q}}| = \#\mathscr{Q} \leqslant \Delta_\alpha(\mathscr{Q})R^\alpha$ and Theorem 5.1, we deduce that

$$\|Uf\|_{L^2(Z_{\mathscr{Q}})} \lesssim \Delta_\alpha(\mathscr{Q})^{1/(n+2)} R^{\alpha/(n+2)} \|f\|_{L^2(\mathbb{R}^n)}.$$

Unfortunately, this is not strong enough to imply Theorem 3.7, since the R-exponent $\alpha/(n+2)$ is always larger than the required value $\alpha/(2(n+1))$.

Remark 5.3. The range $q \geqslant 2 \cdot \frac{n+2}{n}$ in Theorem 5.1 is sharp, so there is no hope of directly improving the Strichartz estimate to give the desired fractal energy estimate via the above argument. To see this, we fix a spatial scale $R \geqslant 1$ and a wave packet ψ_T as in §4.3 with $\rho = R$. Complementing the rapid decay property (30), it is not difficult to show

$$|U\psi_T(x,t)| \gtrsim R^{-n/4} \qquad \text{for all } (x,t) \in \tfrac{1}{100} \cdot T.$$

In particular, for any $1 \leqslant q \leqslant \infty$ we have

$$\|U\psi_T\|_{L^q(B^{n+1}(0,R))} \gtrsim R^{-n/4+(n+2)/(2q)} \qquad \text{and} \qquad \|\psi_T\|_{L^2(\mathbb{R}^n)} \lesssim 1.$$

Since these inequalities hold for all $R \geqslant 1$, it follows that (51) fails whenever $q < 2 \cdot \frac{n+2}{n}$.

Theorem 5.1 is not used in the proof of Theorem 3.7; the linear Strichartz estimate is included here to motivate the multilinear Strichartz estimates introduced in the following section. For this reason, the full proof of Theorem 5.1 is omitted.[15]

5.2. Multilinear Strichartz estimates

Suppose we have initial data ψ_{T_1} and ψ_{T_2} as in Example 4.10 which oscillate at well-separated frequencies ξ_{T_1} and ξ_{T_2}, respectively. Since the waves $U\psi_{T_1}$ and $U\psi_{T_2}$ have distinct frequencies, we expect destructive interference between them; indeed, a rigorous manifestation of this is the orthogonality property (33). Furthermore, by the dispersion relation, the velocities $v(T_1)$ and $v(T_2)$ are also well-separated. Thus, $U\psi_{T_1}$ and $U\psi_{T_2}$ only interact for a short time interval. By these considerations, we expect that the product $U\psi_{T_1} U\psi_{T_2}$ is small, since it measures the interaction between the waves.

These heuristics naturally lead to the study of multilinear Stichartz estimates for the Schrödinger equation. This multilinear theory is a central ingredient in the proof of Theorem 3.7. To introduce the main results, we first discuss the basic example of the bilinear estimate for $n = 1$.

As in §4.5, let $\mathscr{T}_{K^{-1}}$ be a finitely-overlapping covering of $B^n(0,1)$ by K^{-1}-caps with centres lying in $B^n(0,2)$. Given $\tau \in \mathscr{T}_{K^{-1}}$, we let $G(\tau) \subseteq S^n$ denote the image of τ under the Gauss map G introduced in (31). If f_τ is Fourier supported on τ, then $G(\tau)$ is essentially the set of directions of the wavepackets in the wavepacket decomposition of f_τ.

[15] See §6, however, where we do prove slightly weakened version of the $n = 1$ case.

We consider well-separated pairs of caps (τ_1, τ_2), corresponding to wavepackets with distinct frequencies. By the dispersion relation, we may equivalently consider pairs of caps (τ_1, τ_2) corresponding to transverse tubes (T_1, T_2). In particular, given $\xi_1, \xi_2 \in \widehat{\mathbb{R}}^n$, let $|G(\xi_1) \wedge G(\xi_2)|$ denote the absolute value of the determinant of the 2×2 matrix with jth column $G(\xi_j)$. We then define

$$\mathscr{T}^{\text{trans}}_{K^{-1}} := \big\{(\tau_1, \tau_2) \in (\mathscr{T}_{K^{-1}})^2 : |G(\tau_1) \wedge G(\tau_2)| \geqslant K^{-1}\big\}, \tag{52}$$

where

$$|G(\tau_1) \wedge G(\tau_2)| := \inf\big\{|G(\xi_1) \wedge G(\xi_2)| : \xi_j \in \tau_j \text{ for } j = 1, 2\big\}.$$

Let $(\tau_1, \tau_2) \in \mathscr{T}^{\text{trans}}_{K^{-1}}$ and suppose $f_1, f_2 \in L^2(\mathbb{R})$ satisfy $\operatorname{supp} \hat{f}_j \subseteq \tau_j$ for $j = 1, 2$. As in the above discussion, the functions Uf_j oscillate at distinct frequencies and their constituent wave packets travel at distinct velocities. In view of this, we again expect the two waves to interact weakly.

A rigorous manifestation of the above principle is the bilinear identity

$$\int_{\mathbb{R}^{1+1}} |Uf_1(z) Uf_2(z)|^2 \, \mathrm{d}z = 2\pi^2 \int_{\mathbb{R}^2} \frac{|\hat{f}_1(\xi_1)|^2 |\hat{f}_2(\xi_2)|^2}{|\xi_1 - \xi_2|} \, \mathrm{d}\xi_1 \mathrm{d}\xi_2, \tag{53}$$

which holds whenever the f_j satisfy the above hypothesis. This identity dates back to foundational work of Fefferman (1970). The proof is simple: we write

$$Uf_1(z) Uf_2(z) = \int_{\mathbb{R}^2} e^{ix(\xi_1+\xi_2)+it(\xi_1^2+\xi_2^2)} \hat{f}_1(\xi_1) \hat{f}_2(\xi_2) \, \mathrm{d}\xi_1 \mathrm{d}\xi_2$$

and perform the change of variables $\eta_1 = \xi_1 + \xi_2$, $\eta_2 = \xi_1^2 + \xi_2^2$ to deduce

$$Uf_1(z) Uf_2(z) = \check{F}(z) \qquad \text{where} \qquad F(\eta) := (2\pi)^2 \cdot \frac{\hat{f}_1 \circ \xi_1(\eta) \hat{f}_2 \circ \xi_2(\eta)}{2|\xi_1(\eta) - \xi_2(\eta)|}.$$

The desired identity (53) now follows by an application of Plancherel's theorem and changing back the variables.

As an immediate consequence of (53) and interpolation with trivial L^∞ bounds, we deduce the following *bilinear* Strichartz inequality.

Proposition 5.4 (1d Bilinear Strichartz). *Let $4 \leqslant p \leqslant \infty$ and $(\tau_1, \tau_2) \in \mathscr{T}^{\text{trans}}_{K^{-1}}$. The inequality*

$$\Big\| \prod_{j=1}^{2} |Uf_j|^{1/2} \Big\|_{L^p(\mathbb{R}^2)} \lesssim K^{1/p} \prod_{j=1}^{2} \|f_j\|^{1/2}_{L^2(\mathbb{R})} \tag{54}$$

holds whenever $f_j \in L^2(\mathbb{R})$ satisfies $\operatorname{supp} \hat{f}_j \subseteq \tau_j$ for $j = 1, 2$.

In applications, we typically take $K = O(1)$ to be bounded (or to depend subpolynomially on a scale parameter R) so that the additional $K^{1/p}$ factor is admissible. A key advantage of the bilinear setup of Proposition 5.4, as opposed to the $n = 1$ case of Theorem 5.1, is that the estimate (54) is valid all the way down to the exponent $p = 4$. By contrast, the linear inequality is only true down to $p = 6$. The wider range of estimates in the bilinear setup reflects the principle that transverse wave packets interact weakly. In particular, the 'critical behaviour' in Theorem 5.1 arises from non-transverse interactions.

Establishing satisfactory higher dimensional analogues of Proposition 5.4 is a deep problem. For our purposes, we are interested in the following sharp $(n+1)$-linear Strichartz estimate in $\mathbb{R}^n$, which is a celebrated theorem of Bennett, Carbery, and Tao (2006). Generalising the definition of the set of transverse pairs form (52), we consider the set of transverse $(n+1)$-tuples

$$\mathscr{T}^{\mathrm{trans}}_{K^{-1}} := \big\{(\tau_1, \dots, \tau_{n+1}) \in (\mathscr{T}_{K^{-1}})^{n+1} : |G(\tau_1) \wedge \cdots \wedge G(\tau_{n+1})| \geqslant K^{-n}\big\}.$$

Here $|G(\tau_1) \wedge \cdots \wedge G(\tau_{n+1})|$ denotes the infimum of $|G(\xi_1) \wedge \cdots \wedge G(\xi_{n+1})|$ over all $\xi_j \in \tau_j$, $1 \leqslant j \leqslant n+1$. The main result then reads as follows.

Theorem 5.5 (Bennett, Carbery, and Tao, 2006). *Let $p_n := 2 \cdot \frac{n+1}{n}$ and $p_n \leqslant p \leqslant \infty$ and suppose $(\tau_1, \dots, \tau_{n+1}) \in \mathscr{T}^{\mathrm{trans}}_{K^{-1}}$. For all $\varepsilon > 0$ and all $R \geqslant 1$, the inequality*

$$\big\| \prod_{j=1}^{n+1} |Uf_j|^{1/(n+1)} \big\|_{L^p(B_R^{n+1})} \lesssim_\varepsilon K^E R^\varepsilon \prod_{j=1}^{n+1} \|f_j\|^{1/(n+1)}_{L^2(\mathbb{R}^n)} \tag{55}$$

holds whenever $f_j \in L^2(\mathbb{R}^n)$ satisfies $\operatorname{supp} \hat{f}_j \subseteq \tau_j$ for $1 \leqslant j \leqslant n+1$. Here E is a dimensional constant.

The technique used to prove the $n = 1$ case in Proposition 5.4 does not generalise to higher dimensions. The proof of Theorem 5.5 lies beyond the scope of this exposition. A short argument, using ideas of a similar flavour to those explored in this article, is given in Guth (2015).[16] We also refer the reader to Bennett (2014) for an accessible introduction to this topic.

As a direct consequence of Theorem 5.5, we have the following multilinear variant of the fractal energy estimate from Theorem 3.1.

Corollary 5.6 (Multilinear fractal energy estimate). *For all $\varepsilon > 0$ and all $R \geqslant 1$, $1 \leqslant \alpha \leqslant n+1$ the inequality*

$$\big\| \prod_{j=1}^{n+1} |Uf_j|^{1/(n+1)} \big\|_{L^2(Z_{\mathscr{Q}})} \lesssim_\varepsilon K^E \Delta_\alpha(\mathscr{Q})^{1/(2(n+1))} R^{\alpha/(2(n+1))+\varepsilon} \prod_{j=1}^{n+1} \|f_j\|^{1/(n+1)}_{L^2(\mathbb{R}^n)} \tag{56}$$

[16] More precisely, Guth (2015) establishes the (non-endpoint) *multilinear Kakeya inequality*, which is equivalent to Theorem 5.5 by an argument described in Bennett, Carbery, and Tao (2006). Alternatively, one can apply the argument of Guth (2015) to directly prove the multilinear Strichartz inequality by incorporating wave packet decomposition techniques.

holds whenever $f_j \in L^2(\mathbb{R}^n)$ satisfy the hypotheses of Theorem 5.5 and $\mathscr{Q}$ is a family of unit lattice cubes in $B^{n+1}(0,R)$.

Proof. This is a repeat of the simple argument discussed in the linear setting in §5.1. We apply Hölder's inequality and Theorem 5.5 to deduce that

$$\begin{aligned}\big\| \prod_{j=1}^{n+1} |Uf_j|^{1/(n+1)} \big\|_{L^2(Z_{\mathscr{Q}})} &\leqslant |Z_{\mathscr{Q}}|^{1/(2(n+1))} \big\| \prod_{j=1}^{n+1} |Uf_j|^{1/(n+1)} \big\|_{L^{p_n}(B_R^{n+1})} \\ &\lesssim_\varepsilon K^E \Delta_\alpha(\mathscr{Q})^{1/(2(n+1))} R^{\alpha/(2(n+1))+\varepsilon} \prod_{j=1}^{n+1} \|f_j\|_{L^2(\mathbb{R}^n)}^{1/(n+1)}.\end{aligned}$$

Here, in the second step, we use the bound $|Z_{\mathscr{Q}}| = \#\mathscr{Q} \leqslant \Delta_\alpha(\mathscr{Q}) R^\alpha$, which is a direct consequence of the definitions. □

If we assume, say, $K = O_\varepsilon(1)$, then (56) is a multilinear analogue of the desired fractal energy estimate (15) from Theorem 3.7 (and, in fact, (56) has a better dependence on $\Delta_\alpha(\mathscr{Q})$). Note that transversality plays a crucial rôle in the multilinear Strichartz estimate underpinning these observations. Indeed, if we were free to drop the transversality hypothesis in Theorem 5.5 and take $f = f_1 = \cdots = f_{n+1}$, then we would arrive at the linear estimate

$$\|Uf\|_{L^{p_n}(B(0,R))} \lesssim_\varepsilon R^\varepsilon \|f\|_{L^2(\mathbb{R}^n)}; \tag{57}$$

however, taking f to be a wave packet as defined in Example 4.10, it is easy to see that (57) fails (see Remark 5.3). Thus, the key difficulty in proving Theorem 3.7 is to control the non-transversal interactions.

5.3. Multilinear Bernstein inequality

In this subsection we address some slightly technical multilinear extensions of the results discussed in §4.1 which will be of use in later arguments.

Lemma 5.7 (Multilinear Bernstein inequality). *Let $1 \leqslant p \leqslant q \leqslant \infty$ and suppose $F_j \in \mathscr{S}(\mathbb{R}^d)$ satisfy* $\operatorname{supp} \hat{F}_j \subseteq Q_0 := [-1/2, 1/2]^d$ *for $1 \leqslant j \leqslant d$. Then*

$$\big\| \prod_{j=1}^{d} |F_j|^{1/d} \big\|_{L^q(\mathbb{R}^d)} \lesssim \big\| \prod_{j=1}^{d} |F_j|^{1/d} \big\|_{L^p(\mathbb{R}^d)}.$$

Remark 5.8. By applying an affine scaling, Lemma 5.7 immediately implies a generalisation of itself for functions Fourier supported in some fixed parallelepiped π. Since, for our purposes, we only require the result at unit scale, we omit the details.

There is, in fact, nothing particularly multilinear *per se* about Lemma 5.7: it is a direct consequence of linear Bernstein inequalities. Indeed, under the hypothesis of Lemma 5.7, if we define $F := F_1 \cdots F_d$, then we may equivalently express (5.7) as

$$\|F\|_{L^{q/d}(\mathbb{R}^d)} \lesssim \|F\|_{L^{p/d}(\mathbb{R}^d)}.$$

Suppose for $p/d \geqslant 1$. Since F has Fourier support in $Q_0 + \cdots + Q_0$, the d-fold Minkowski sum, in this case Lemma 5.7 is a direct consequence of the linear Bernstein inequality applied to F. To deal with the remaining case $p/d < 1$, we establish a variant of Lemma 4.1.

Lemma 5.9. *Let $0 < s \leqslant 1$ and $M \geqslant 1$. There exists a function $\eta_M \colon \mathbb{R}^d \to [0,\infty)$ satisfying the following:*

i) If $F \in \mathscr{S}(\mathbb{R}^d)$ satisfies $\operatorname{supp} \hat{F} \subseteq Q_0 := [-1/2, 1/2]^d$, then

$$|F(z)|^s \leqslant \sum_{Q \in \mathscr{Q}_{M,\mathrm{all}}} |a_Q|^s \chi_Q(z) \lesssim M^d |F|^s * \eta_M(z) \qquad \text{for all } z \in \mathbb{R}^d$$

where $\mathscr{Q}_{M,\mathrm{all}}$ is the collection of all lattice M-cubes and

$$a_Q := \sup_{z \in Q} |F(z)| \qquad \text{for all } Q \in \mathscr{Q}_{M,\mathrm{all}}.$$

ii) The function η_M is L^1-normalised and rapidly decaying away from $[-M/2, M/2]^d$ in the sense that

$$\eta_M(z) \lesssim_{N,s} M^{-d}(1 + 2M^{-1}|z|_\infty)^{-N} \qquad \text{for all } N \in \mathbb{N}.$$

Once Lemma 5.9 is proved, Lemma 5.7 follows easily by adapting the argument used to prove the linear Bernstein inequality in Lemma 4.3. Note that we only need the case $M = 1$ of Lemma 5.9 for this purpose; it is useful, however, to have the result for general M in view of later applications.

Proof (of Lemma 5.9). Fix $\eta_0 \in \mathscr{S}(\mathbb{R}^d)$ satisfying $\operatorname{supp} \hat{\eta}_0 \subseteq [-2,2]^d$ and $\hat{\eta}_0(\xi) = 1$ for all $\xi \in [-1,1]^d$. Let $F \in \mathscr{S}(\mathbb{R}^d)$ satisfy the hypothesis of part i). Given $Q \in \mathscr{Q}_{M,\mathrm{all}}$ and $z_Q \in Q$ such that $|F(z_Q)| = a_Q$, we have

$$a_Q = |F(z_Q)| \leqslant \int_{\mathbb{R}^d} |F(w)||\eta_0(z_Q - w)| \, \mathrm{d}w. \tag{58}$$

We interpret the right-hand side as the norm $\|G_R\|_{L^1(\mathbb{R}^d)}$, where

$$G_R(w) := F(w)\eta_0(z_Q - w).$$

It is not difficult to see $\hat{G}_R$ is supported in $[-10,10]^d$. We claim that, by Bernstein's inequality,

$$\|G_R\|_{L^1(\mathbb{R}^d)} \lesssim \|G_R\|_{L^s(\mathbb{R}^d)}. \tag{59}$$

The issue here is that $0 < s \leqslant 1$ and so we cannot appeal to Lemma 4.3 directly. However, applying Bernstein's inequality with exponents $p = 1$ and $q = \infty$ gives

$$\|G_R\|_{L^1(\mathbb{R}^d)} \leqslant \|F\|_{L^\infty(\mathbb{R}^d)}^{1-s} \|F\|_{L^s(\mathbb{R}^d)}^{s} \lesssim \|F\|_{L^1(\mathbb{R}^d)}^{1-s} \|F\|_{L^s(\mathbb{R}^d)}^{s},$$

which then rearranges to produce the desired bound (59).

In light of the above, we can upgrade (58) to

$$|a_Q|^s \lesssim \int_{\mathbb{R}^d} |F(w)|^s |\eta_0(z_Q - w)|^s \, \mathrm{d}w.$$

If we now define $\eta_M \colon \mathbb{R}^d \to [0,\infty)$ by

$$\eta_M(z) := M^{-d} \sup_{|w-z|_\infty \leqslant M} |\eta_0(w)|^s,$$

then the desired result follows as in the proof of Lemma 4.1. □

We may also spatially localise the multlinear Bernstein inequality, as in Corollary 4.5.

Corollary 5.10. *Under the hypotheses of Lemma 5.7, if $Q \subseteq \mathbb{R}^d$ is any cube of side-length at least 1, then*

$$\big\| \prod_{j=1}^d |F_j|^{1/d} \big\|_{L^q(Q)} \lesssim \big\| \prod_{j=1}^d |F_j|^{1/d} \big\|_{L^p(w_Q)},$$

where $w_Q \colon \mathbb{R}^d \to [0,\infty)$ is a weight adapted to Q (see Definition 4.4).

Proof. This follows from Lemma 5.7 via the same argument used to prove Corollary 4.5. □

6. Broad-narrow analysis

6.1. Motivation

In the previous section we derived a multilinear variant of the fractal energy estimate, Corollary 5.6, as a direct consequence of the multilinear Strichartz estimates of Bennett, Carbery, and Tao (2006). The problem now is to obtain *bona fide* linear estimates from their multilinear counterparts.

On the face of it, passing from multilinear to linear estimates appears challenging. Indeed, the proof of Corollary 5.6 crucially exploits the transversality hypothesis, required in order to invoke Theorem 5.5. Consequently, the methods of §5 are ill equipped to deal with interactions between resonant wave packets.

In this section we describe an ingenious method induced in Bourgain and Guth (2011) which does in fact allow passage from multilinear to linear estimates. This is now commonly referred to as the *Bourgain–Guth method* or *broad-narrow analysis* and forms the backbone of many recent advances in harmonic analysis.

Broad-narrow analysis was originally introduced to study the famous Fourier restriction conjecture, but was later adapted in Bourgain (2013b) to make progress on the Carleson problem. It relies on decomposition, scaling and induction-on-scale techniques which have their roots in earlier works of Tao, Vargas, and Vega (1998) and Wolff (2001), amongst others.

6.2. Broad-narrow analysis for $n = 1$: an illustration

To begin, we illustrate the core ideas behind broad-narrow analysis in a very simple setting. In particular, we use the bilinear estimate from Proposition 5.4 to prove the following (slightly weaker) variant of the $n = 1$ case of Theorem 5.1.

Proposition 6.1. *For all $\varepsilon > 0$ and all $R \geqslant 1$, the inequality*

$$\|Uf\|_{L^6(B^{1+1}(0,R))} \lesssim_\varepsilon R^\varepsilon \|f\|_{L^2(\mathbb{R})}$$

holds whenever $f \in L^2(\mathbb{R})$.

The following proof of Proposition 6.1 is included for illustrative purposes only: we do not require the result later in the discussion. The proof does, however, provide a simple and effective introduction to broad-narrow methods.

The first step is to relate the linear operator Uf to bilinear $Uf_1 Uf_2$ expressions involving functions f_1, f_2 satisfying the transversality hypothesis of Proposition 5.4. For $n = 1$, this is easily achieved via an elementary pointwise inequality.

Let $f \in L^2(\mathbb{R})$ and assume, without loss of generality, that $\operatorname{supp} \hat{f} \subseteq B^1(0, 1/2)$. We decompose

$$f = \sum_{\tau \in \mathscr{T}_{K^{-1}}} f_\tau,$$

similarly to the discussion in §4.4, so that each $f_\tau \in L^2(\mathbb{R})$ satisfies $\operatorname{supp} \hat{f}_\tau \subseteq \tau$.

Lemma 6.2 (Broad-narrow decomposition, $n = 1$). *For all $z = (x, t) \in \mathbb{R}^{1+1}$, we have*

$$|Uf(z)| \lesssim \max_{\tau \in \mathscr{T}_{K^{-1}}} |Uf_\tau(z)| + K \max_{(\tau_1, \tau_2) \in \mathscr{T}^{\mathrm{trans}}_{K^{-1}}} \prod_{j=1}^{2} |Uf_{\tau_j}(z)|^{1/2}. \tag{60}$$

Here $\mathscr{T}^{\mathrm{trans}}_{K^{-1}}$ denotes the collection of all K^{-1}-transverse pairs of caps in $\mathscr{T}_{K^{-1}}$, as defined in (52). The first term on the right-hand side of (60) is referred to as the *narrow* term, whilst the second is referred to as the *broad* term.

Proof (of Lemma 6.2). Given $z = (x,t) \in \mathbb{R}^{1+1}$, we let $\tau_z \in \mathscr{T}_{K^{-1}}$ be a choice of cap satisfying

$$|Uf_{\tau_z}(z)| = \max_{\tau \in \mathscr{T}_{K^{-1}}} |Uf_\tau(z)|.$$

We define the set of *narrow* and *broad* caps (for Uf at z) by

$$\mathscr{N}_z := \big\{\tau \in \mathscr{T}_{K^{-1}} : |G(\tau) \wedge G(\tau_z)| < K^{-1}\big\} \quad \text{and} \quad \mathscr{B}_z := \mathscr{T}_{K^{-1}} \setminus \mathscr{N}_z,$$

respectively. Note that $\mathscr{N}_z$ simply consists of the cap τ_z and some of its neighbours.

The decomposition of $\mathscr{T}_{K^{-1}}$ into broad and narrow caps induces a decomposition of the operator

$$|Uf(z)| = \Big|\sum_{\tau \in \mathscr{T}_{K^{-1}}} Uf_\tau(z)\Big| \leqslant \Big|\sum_{\tau \in \mathscr{N}_z} Uf_\tau(z)\Big| + \sum_{\tau \in \mathscr{B}_z} |Uf_\tau(z)|. \tag{61}$$

For the term involving narrow caps, we simply note that $\#\mathscr{N}_z \lesssim 1$ and so

$$\Big|\sum_{\tau \in \mathscr{N}_z} Uf_\tau(z)\Big| \lesssim \max_{\tau \in \mathscr{T}_{K^{-1}}} |Uf_\tau(z)|. \tag{62}$$

On the other hand, clearly $(\tau, \tau_z) \in \mathscr{T}^{\mathrm{trans}}_{K^{-1}}$ for all $\tau \in \mathscr{B}_z$ and so

$$\begin{aligned}\sum_{\tau \in \mathscr{B}_z} |Uf_\tau(z)| &\leqslant \sum_{\tau \in \mathscr{B}_z} |Uf_\tau(z)|^{1/2} |Uf_{\tau_z}(z)|^{1/2} \\ &\lesssim K \max_{(\tau_1,\tau_2) \in \mathscr{T}^{\mathrm{trans}}_{K^{-1}}} \prod_{j=1}^{2} |Uf_{\tau_j}(z)|^{1/2}.\end{aligned} \tag{63}$$

Plugging (62) and (63) into (61), we deduce the desired estimate. □

As an immediate consequence of the pointwise bound in Lemma 6.2, we deduce the following norm bound.

Corollary 6.3 (L^q broad-narrow decomposition, $n = 1$). *For all $1 \leqslant q \leqslant \infty$ and all $R \geqslant 1$, we have*

$$\|Uf\|_{L^q(B_R^{1+1})} \lesssim \Big(\sum_{\tau \in \mathscr{T}_{K^{-1}}} \|Uf_\tau\|^q_{L^q(B_R^{1+1})}\Big)^{1/q} + K^3 \max_{(\tau_1,\tau_2) \in \mathscr{T}^{\mathrm{trans}}_{K^{-1}}} \Big\| \prod_{j=1}^{2} |Uf_{\tau_j}|^{1/2} \Big\|_{L^q(B_R^{1+1})}. \tag{64}$$

For $q = \infty$ the ℓ^q expression is understood as a maximum.

Proof. The $q = \infty$ case is immediate. Fixing $1 \leqslant q < \infty$, dominate the maxima in (60) by the corresponding ℓ^q sum and take the L^q-norm in z of both sides of the resulting expression. The desired result then follows by interchanging ℓ^q and L^q norms. □

To motivate what follows, we pause to consider the terms appearing on the right-hand side of (64).

- ▷ The broad term involves a bilinear expression and, for appropriate q, can be estimated using the bilinear Strichartz estimate from Proposition 5.4.
- ▷ To estimate the narrow term, the key observation is that the expressions $\|Uf_\tau\|_{L^q(B_R^{1+1})}$ appearing on the right-hand side are of the same form as the expression $\|Uf\|_{L^q(B_R^{1+1})}$ appearing on the left-hand side. This is a symmetry (or self-similarity) of the inequality. Moreover, the functions f_τ appearing on the right are frequency localised versions of the function f on the left; in this sense the f_τ are simpler objects than f. These considerations naturally lead to an inductive argument.

Proof (*of Proposition 6.1*). Fix $\varepsilon > 0$. We let $\mathbf{C}_\varepsilon \geqslant 1$ and $K \geqslant 2$ denote fixed constants, depending only on the admissible parameter ε, which are chosen large enough to satisfy the forthcoming requirements of the proof.

We argue by induction on the scale parameter R. As a simple consequence of the Cauchy–Schwarz inequality and Plancherel's theorem, we have

$$\|Uf\|_{L^6(B^{1+1}(0,R))} \lesssim R^{1/3}\|Uf\|_{L^\infty(B^{1+1}(0,R))} \lesssim R^{1/3}\|\hat{f}\|_{L^2(\mathbb{R})} = R^{1/3}\|f\|_{L^2(\mathbb{R})}.$$

Thus, Proposition 6.1 holds trivially for small scales, say $R \leqslant 100$; this serves as a base case for the induction.

Fix $R \geqslant 100$ and $f \in L^2(\mathbb{R})$ and assume, without loss of generality, that $\operatorname{supp} \hat{f} \subseteq B^1(0, 1/2)$. We further assume the following holds.

Induction hypothesis. *For all* $1 \leqslant \tilde{R} \leqslant R/2$, *the inequality*

$$\|Ug\|_{L^6(B^{1+1}(0,\tilde{R}))} \leqslant \mathbf{C}_\varepsilon \tilde{R}^\varepsilon \|g\|_{L^2(\mathbb{R})}$$

holds whenever $g \in L^2(\mathbb{R})$.

For $2 \leqslant q < \infty$ we apply the broad-narrow decomposition from Corollary 6.3 to deduce that[17]

$$\|Uf\|_{L^q(B_R^{1+1})} \lesssim \Big(\sum_{\tau \in \mathscr{T}_{K^{-1}}} \|Uf_\tau\|^2_{L^q(B_R^{1+1})} \Big)^{1/2} + K^3 \max_{(\tau_1,\tau_2) \in \mathscr{T}^{\mathrm{trans}}_{K^{-1}}} \Big\| \prod_{j=1}^{2} |Uf_{\tau_j}|^{1/2} \Big\|_{L^q(B_R^{1+1})}. \tag{65}$$

Here we have weakened the ℓ^q sum to an ℓ^2 sum in the narrow term, using the nesting of ℓ^q norms.

[17] Later we fix $q = 6$, but it is useful to keep the parameter q free for the time being to see how the numerology works out in general.

The broad term. We begin by estimating the broad contribution: that is, the second term on the right-hand side of (65). This is achieved by direct application of the bilinear Strichartz estimate from Proposition 5.4. In particular, by (54), for $4 \leqslant q < \infty$ we have

$$\|Uf\|_{L^q(B_R^{1+1})} \lesssim \Big(\sum_{\tau \in \mathscr{T}_{K^{-1}}} \|Uf_\tau\|^2_{L^q(B_R^{1+1})} \Big)^{1/2} + K^4 \|f\|_{L^2(\mathbb{R})}. \tag{66}$$

The narrow term. It remains to estimate the narrow contribution, corresponding to the first term of the right-hand side of (65). This is achieved via a combination of parabolic rescaling and appeal to the induction hypothesis.

For each $\tau \in \mathscr{T}_{K^{-1}}$, as in (49) we decompose[18]

$$f_\tau = \sum_{S \in \mathbf{S}_\tau[R]} f_S \qquad \text{where} \qquad f_S := \sum_{T \in \mathbf{T}_S[R]} f_T. \tag{67}$$

By Lemma 4.16, we have

$$\|Uf_\tau\|^q_{L^q(B_R^{1+1})} \lesssim \sum_{S \in \mathbf{S}_\tau[R]} \|Uf_S\|^q_{L^q(\bar{S})} + R^{-100nq} \|f\|^q_{L^2(\mathbb{R})}. \tag{68}$$

We now invoke parabolic rescaling in the form of Lemma 4.18. In particular, for each $S \in \mathbf{S}_\tau[R]$ there exists a function $\tilde{f}_S \in L^2(\mathbb{R})$ which is Fourier supported in $B^1(0,1)$ and satisfies

$$\|Uf_S\|_{L^q(\bar{S})} \leqslant K^{3/q-1/2} \|U\tilde{f}_S\|_{L^q(B^{1+1}(0,\tilde{R}))} \qquad \text{and} \qquad \|\tilde{f}_S\|_{L^2(\mathbb{R})} = \|f_S\|_{L^2(\mathbb{R})}, \tag{69}$$

where $\tilde{R} := 20R/K^2$. Provided K is chosen sufficiently large, $\tilde{R} \leqslant R/2$, and so we may apply the induction hypothesis to conclude that

$$\|U\tilde{f}_S\|_{L^q(B^{1+1}(0,\tilde{R}))} \leqslant \mathbf{C}_\varepsilon \tilde{R}^\varepsilon \|\tilde{f}_S\|_{L^2(\mathbb{R})} \lesssim \mathbf{C}_\varepsilon K^{-2\varepsilon} R^\varepsilon \|f_S\|_{L^2(\mathbb{R})}. \tag{70}$$

Combining (68), (69) and (70), we therefore deduce that

$$\Big(\sum_{\tau \in \mathscr{T}_{K^{-1}}} \|Uf_\tau\|^2_{L^q(B_R^{1+1})} \Big)^{1/2} \lesssim \mathbf{C}_\varepsilon K^{3/q-1/2-2\varepsilon} R^\varepsilon \Big(\sum_{S \in \mathbf{S}[R]} \|f_S\|^2_{L^2(\mathbb{R})} \Big)^{1/2} + R^{-10n} \|f\|_{L^2(\mathbb{R})}, \tag{71}$$

where $\mathbf{S}[R]$ denotes the union of all the sets $\mathbf{S}_\tau[R]$ over all $\tau \in \mathscr{T}_{K^{-1}}$. The families of wave packets $\mathbf{T}_S[R]$ appearing in the definition (67) are essentially disjoint as S varies over $\mathbf{S}[R]$. Thus, by the orthogonality properties of the wave packets,

$$\Big(\sum_{S \in \mathbf{S}[R]} \|f_S\|^2_{L^2(\mathbb{R})} \Big)^{1/2} \lesssim \|f\|_{L^2(\mathbb{R})}.$$

[18] The decomposition according to the strips S is somewhat overkill for the purposes of this argument, but is an important feature when adapting these methods to prove Theorem 3.7 in §7. An alternative approach to the current proof is to rescale Uf_τ directly using Corollary 4.8.

Applying this bound to (71), we deduce that

$$\Big(\sum_{\tau \in \mathscr{T}_{K^{-1}}} \|Uf_\tau\|^2_{L^q(B_R^{1+1})} \Big)^{1/2} \lesssim \mathbf{C}_\varepsilon K^{3/q-1/2-2\varepsilon} R^\varepsilon \|f\|_{L^2(\mathbb{R})}, \tag{72}$$

which is the final estimate for the narrow term.

To conclude the proof, we combine (66) and (72); thus, for $4 \leqslant q \leqslant \infty$, we have

$$\|Uf\|_{L^q(B_R^{1+1})} \leqslant C\big(\mathbf{C}_\varepsilon K^{3/q-1/2} K^{-2\varepsilon} R^\varepsilon + K^4\big) \|f\|_{L^2(\mathbb{R})},$$

where C is suitable a choice of absolute constant, which accounts for all the factors arising from the implicit constants in the above argument. Note that the exponent $3/q - 1/2$ is non-positive precisely when $q \geqslant 6$. In particular, specialising to the case $q = 6$, we have

$$\|Uf\|_{L^6(B_R^{1+1})} \leqslant C\big(\mathbf{C}_\varepsilon K^{-2\varepsilon} + K^4\big) R^\varepsilon \|f\|_{L^2(\mathbb{R})}. \tag{73}$$

The estimate (73) involves two free parameters (both of which must be chosen independently of f and R):

- ▷ The constant $\mathbf{C}_\varepsilon$ appearing in the induction hypothesis;
- ▷ The intermediate scale K.

We fine tune these parameters to ensure that the induction closes. We first choose $\mathbf{C}_\varepsilon$ in terms of K so as to satisfy $\mathbf{C}_\varepsilon = 2CK^4$. Thus, (72) becomes

$$\|Uf\|_{L^6(B_R^{1+1})} \leqslant \Big(C\mathbf{C}_\varepsilon K^{-2\varepsilon} R^\varepsilon + \frac{\mathbf{C}_\varepsilon}{2} \Big) \|f\|_{L^2(\mathbb{R})}.$$

Finally, we choose K, depending only on ε, so that $CK^{-2\varepsilon} \leqslant 1/2$. With this choice, we conclude

$$\|Uf\|_{L^6(B_R^{1+1})} \leqslant \mathbf{C}_\varepsilon R^\varepsilon \|f\|_{L^2(\mathbb{R})},$$

which closes the induction and completes the proof. □

The simple argument presented above is particular to the $n = 1$ case. This is due to the innocuous nature of the narrow term when $n = 1$; in higher dimensions we shall see that the narrow term is significantly more complex.

6.3. What is going on here?

The induction argument used in the proof of Proposition 6.1 is very neat, but perhaps obscures the mechanics of what is happening in the proof. Indeed, arguments like this can sometimes seem a little magical. To get a better sense of what is going on, it is helpful to 'unpack' the induction argument and think of it as a recursive process.

This tends to be messier, but can give a better sense of how the argument works. Here we give an informal sketch of this recursive process, reinterpreting the proof from the previous subsection.

As the process progresses, we pass through a decreasing chain of spatial scales

$$R \to R/K^2 \to R/K^4 \to \cdots \to R/K^{2N}, \tag{74}$$

where N corresponds to the total number of steps in the recursion. At the terminal step (corresponding to the base case in the proof of Proposition 6.1), we have reached the unit scale, and so we have, say, $R/K^{2N} \leqslant 100$. From this, we see that the total number of steps is roughly

$$N \sim \frac{\log R}{\log K}.$$

At each step we use Corollary 6.3 to split the norm into two parts: the broad term, which is analysed directly, and the narrow term which we continue to decompose. Thus, at each step we gain one additional piece in the decomposition so that throughout the whole process we split the norm into $N+1$ pieces. In particular, we have:

- ▷ A broad term for each step of the process;
- ▷ A narrow term for the terminal step only.

Since $N+1 \lesssim \log R \lesssim R^{\varepsilon/2}$, it suffices to show that the contribution from each piece of this decomposition is $O_\varepsilon(R^{\varepsilon/2}\|f\|_{L^2(\mathbb{R})})$.

The remaining narrow term from the terminal step is controlled using the trivial energy estimate, corresponding to the analysis of the base case in the inductive setup. On the other hand, each of the broad terms is estimated using the bilinear Strichartz inequality from Proposition 5.4. By applying parabolic rescaling, we can always separate the pairs of caps appearing in the broad term to ensure they are K^{-1}-transverse. This means that the constant arising from the bilinear estimate is uniform over all steps of the recursion.[19]

There is one final subtlety. When carrying out the broad-narrow decomposition at each step, we incur some additional absolute constant $C_\circ$ in our inequality, say $C_\circ = 10$. On their own, these constants are harmless, but as we iterate the procedure they accumulate as powers $C_\circ^k$. At the terminal stage, we will have gained an additional factor of $C_\circ^N$: since N is logarithmic in R, this could be catastrophic. The parameter K is used to deal with this issue. In particular, by choosing K sufficiently large, depending only on ε, we can ensure

$$C_\circ^N = R^{C \log C_\circ / \log K} \leqslant R^{\varepsilon/2},$$

[19] We did not apply parabolic rescaling to the broad term directly in the inductive proof of Proposition 6.1. However, by applying parabolic rescaling to the narrow term at step k of the process, we effectively rescale the broad term at step $k+1$ (which is formed from decomposing the step k narrow term). Thus, the above is indeed an accurate representation of the proof of Proposition 6.1.

which is admissible for our purposes. In short, the larger K, the larger the jumps between scales in (74) and, consequently, the fewer the number of steps N in the recursive process. By choosing K sufficiently large, we can favourably control the constant $C_\circ^N$ which arises through the recursive procedure.

6.4. Pointwise broad-narrow decomposition in higher dimensions

In the remainder of this section we explore extensions of the simple ideas introduced in §6.2 to higher spatial dimensions. This provides the framework for the proof of the fractal energy estimate (Theorem 3.7) in the next section.[20]

For our purposes, the correct implementation of the broad-narrow decomposition for $n \geqslant 2$ turns out to be highly non-trivial and relies on the deep *decoupling theory* of Bourgain and Demeter (2015). Our first step is to generalise the simple pointwise broad-narrow decomposition from Lemma 6.2 to higher dimensions. This will in fact prove a misstep, and we shall go back and refine our estimates in the proof of Lemma 6.7 below. Nevertheless, we consider the pointwise decomposition to gain an initial understanding of the problem.

Recall in the proof of Lemma 6.2 we defined a collection of *narrow caps*, which were caps with normals $G(\tau)$ aligned along some fixed 1-dimensional subspace. More generally, given a d-dimensional linear subspace $V \subseteq \mathbb{R}^{n+1}$, we define

$$\mathscr{T}_{K^{-1}}(V) := \big\{\tau \in \mathscr{T}_{K^{-1}} : |\sin \angle(G(\tau), V)| \leqslant C_n K^{-1}\big\}. \tag{75}$$

Here $\angle(G(\tau), V)$ denotes the infimum of the angles $\angle(G(\xi), V)$ over all $\xi \in \tau$ and $C_n \geqslant 1$ is a dimensional constant, chosen large enough to satisfy the forthcoming requirements of the proof. With this definition, the general from of Lemma 6.2 reads thus.

Lemma 6.4 (Pointwise broad-narrow decomposition). *For all $z = (x,t) \in \mathbb{R}^{n+1}$, we have*

$$|Uf(z)| \lesssim \max_{V \in \mathrm{Gr}(n,\mathbb{R}^{n+1})} \Big| \sum_{\tau \in \mathscr{T}_{K^{-1}}(V)} Uf_\tau(z) \Big| + K^n \max_{(\tau_1,\dots,\tau_{n+1}) \in \mathscr{T}_{K^{-1}}^{\mathrm{trans}}} \prod_{j=1}^{n+1} |Uf_{\tau_j}(z)|^{1/(n+1)}. \tag{76}$$

Here the left-hand maximum is taken over all n-dimensional linear subspaces in $\mathbb{R}^{n+1}$.

[20] In particular, the goal is **not** to extend the proof of Proposition 6.1 (which is included for illustrative purposes only) to higher dimensions. Whilst broad-narrow analysis can be used to prove Strichartz estimates for $n \geqslant 2$, this approach is cumbersome in the extreme compared with, say, the original proof of Theorem 5.1 from Tomas (1975). On the other hand, the tools developed here are effective when it comes to studying Theorem 3.7 in general dimensions.

We immediate see that Lemma 6.4 implies Lemma 6.2 in the $n = 1$ case. Comparing (76) with (60), a significant additional complication in higher dimensions is the form of the narrow term. This involves the function

$$f_V := \sum_{\tau \in \mathscr{T}_{K^{-1}}(V)} f_\tau,$$

which is localised to a whole family of caps (aligned along a strip), rather than a single cap. For $n \geqslant 2$, the analysis of this term involves highly non-trivial tools from decoupling theory, described in §6.5 below.

We remark that the precise form of Lemma 6.4 is not used in our subsequent analysis; instead, we will rely on an L^q variant introduced in Lemma 6.7 below.[21] Nevertheless, Lemma 6.4 provides a useful conceptual stepping stone.

Proof (*of Lemma 6.4*). The proof is an elaboration of the argument used to establish the $n = 1$ case in Lemma 6.2. The first step is to identify a codimension 1 subspace $V_z \in \mathrm{Gr}(n, \mathbb{R}^{n+1})$ which 'captures' as many of the large $|Uf_\tau(z)|$ as possible. More precisely, define the *broad part* of the operator

$$U_{\mathrm{Br}} f(z) := \min_{V \in \mathrm{Gr}(n, \mathbb{R}^{n+1})} \max_{\tau \notin \mathscr{T}_{K^{-1}}(V)} |Uf_\tau(z)| \tag{77}$$

and suppose $V_z \in \mathrm{Gr}(n, \mathbb{R}^{n+1})$ realises the minimum in (77). Then V_z has the desired property, in the sense that the largest value of $|Uf_\tau(z)|$ for $\tau \notin \mathscr{T}_{K^{-1}}(V)$ is minimised.

For slightly technical reasons, we also define

$$\mathscr{T}_{K^{-1}}(V, z) := \big\{\tau \in \mathscr{T}_{K^{-1}}(V) : |Uf_\tau(z)| \geqslant U_{\mathrm{Br}} f(z)\big\}$$

and further choose V_z so that, of all spaces realising the minimum in (77), the space V_z also maximises $\#\mathscr{T}_{K^{-1}}(V, z)$. Once again, we can think of this condition as ensuring V_z captures as many large $|Uf_\tau(z)|$ as possible.

From the definition, the caps $\tau \in \mathscr{T}_{K^{-1}}(V_z, z)$ (or, more precisely, their normals) are aligned around the n-dimensional subspace V_z. However, they do not align around around any lower dimensional subspace.

Claim. There does not exist a subspace $W \subseteq \mathbb{R}^{n+1}$ of dimension $n - 1$ such that $\mathscr{T}_{K^{-1}}(V_z, z) \subseteq \mathscr{T}_{K^{-1}}(W)$.

The idea is that if all the caps in $\mathscr{T}_{K^{-1}}(V_z, z)$ are 'captured' by a subspace W of dimension $n - 1$, then we have the freedom to extend W to a subspace V_z' of dimension n which captures even more large caps. But this contradicts the maximality of V_z. We postpone the precise details of this argument until the end of the proof.

[21] In contrast with the $n = 1$ case, in higher dimensions our formulation of the L^q broad-narrow decomposition does not directly follow from the pointwise decomposition.

Assuming the claim, it is a simple matter to conclude the proof of Lemma 6.4. In analogy with the proof of Lemma 6.2, we define the collections of *narrow* and *broad* caps (for Uf at z) by

$$\mathscr{N}_z := \mathscr{T}_{K^{-1}}(V_z) \qquad \text{and} \qquad \mathscr{B}_z := \mathscr{T}_{K^{-1}} \setminus \mathscr{T}_{K^{-1}}(V_z),$$

respectively. By the triangle inequality and the defining properties of V_z, we have

$$\begin{aligned} |Uf(z)| &\leqslant \Big| \sum_{\tau \in \mathscr{N}_z} Uf_\tau(z) \Big| + \sum_{\tau \in \mathscr{B}_z} |Uf_\tau(z)| \\ &\lesssim \max_{V \in \mathrm{Gr}(n, \mathbb{R}^{n+1})} \Big| \sum_{\tau \in \mathscr{T}_{K^{-1}}(V)} Uf_\tau(z) \Big| + K^n U_{\mathrm{Br}} f(z). \end{aligned} \tag{78}$$

In view of the claim, there exists an n-tuple of caps $(\tau_{z,1}, \dots, \tau_{z,n}) \in \mathscr{T}_{K^{-1}}(V_z, z)^n$ which is $K^{-(n+1)}$-transverse in the sense that

$$\big| G(\tau_{z,1}) \wedge \cdots \wedge G(\tau_{z,n}) \big| \geqslant K^{-(n-1)}. \tag{79}$$

To see this, simply take $(\tau_{z,1}, \dots, \tau_{z,n})$ to be a tuple which maximises the left-hand wedge product. If (79) fails for this choice, then

$$\big| G(\tau_{z,1}) \wedge \cdots \wedge G(\tau_{z,n-1}) \wedge G(\tau) \big| < K^{-(n-1)} \qquad \text{for all } \tau \in \mathscr{T}_{K^{-1}}(V_z, z).$$

Define $W := \operatorname{span}\{G(\xi_{\tau_{z,1}}), \dots, G(\xi_{\tau_{z,n-1}})\}$, where $\xi_{\tau_{z,j}}$ denotes the centre of $\tau_{z,j}$ for $1 \leqslant j \leqslant n-1$. If the constant $C_n \geqslant 1$ in (75) is chosen sufficiently large, depending only on n, then W is an $(n-1)$-dimensional subspace satisfying $\mathscr{T}_{K^{-1}}(V_z, z) \subseteq \mathscr{T}_{K^{-1}}(W)$, contradicting the claim. Hence (79) must hold.

By the definition of the broad functional, there exists a cap $\tau_{z,n+1} \notin \mathscr{T}_{K^{-1}}(V_z)$ such that

$$|Uf_{\tau_{z,n+1}}(z)| = U_{\mathrm{Br}} f(z).$$

It follows that the $(n+1)$-tuple $(\tau_{z,1}, \dots, \tau_{z,n+1})$ is K^{-n}-transverse and satisfies

$$U_{\mathrm{Br}} f(z) \leqslant \prod_{j=1}^{n+1} |Uf_{\tau_{z,j}}(z)|^{1/(n+1)} \leqslant \max_{(\tau_1, \dots, \tau_{n+1}) \in \mathscr{T}_{K^{-1}}^{\mathrm{trans}}} \prod_{j=1}^{n+1} |Uf_{\tau_j}(z)|^{1/(n+1)}. \tag{80}$$

The desired result now follows by combining (80) and (78).

It remains to prove the claim. Following the proof sketch, we argue by contradiction, assuming such a subspace W exists. Let $\tau^* := \tau_{z,n+1}$ be as above, so that $\tau^* \notin \mathscr{T}_{K^{-1}}(V_z)$ realises the maximum in the definition (77) for $V = V_z$. Define V_z' to be a subspace of dimension n which contains W and $G(\xi_{\tau^*})$, where ξ_{τ^*} is the centre of τ^*.

First note that V_z' realises the minimum in (77). Indeed, from the hypothesis on W and the definition of V_z' we have $\mathscr{T}_{K^{-1}}(V_z) \subseteq \mathscr{T}_{K^{-1}}(V_z')$ and so

$$U_{\mathrm{Br}}f(x) = \max_{\tau \notin \mathscr{T}_{K^{-1}}(V_z)} |Uf_\tau(z)| \geqslant \max_{\tau \notin \mathscr{T}_{K^{-1}}(V_z')} |Uf_\tau(z)| \geqslant U_{\mathrm{Br}}f(x).$$

In light of the maximality of V_z, it follows that $\#\mathscr{T}_{K^{-1}}(V_z', z) \leqslant \#\mathscr{T}_{K^{-1}}(V_z, z)$. However, it is clear from the definitions that

$$\mathscr{T}_{K^{-1}}(V_z, z) \subseteq \mathscr{T}_{K^{-1}}(V_z', z), \qquad \text{whilst} \quad \tau^* \in \mathscr{T}_{K^{-1}}(V_z', z) \quad \text{and} \quad \tau^* \notin \mathscr{T}_{K^{-1}}(V_z, z).$$

Thus, $\#\mathscr{T}_{K^{-1}}(V_z, z) < \#\mathscr{T}_{K^{-1}}(V_z', z)$, which is a contradiction. □

As in the $n = 1$ case discussed in §6.2, the inequality (76) is designed to access the multilinear theory from §5. In particular, the multilinear Strichartz estimates can be used to control the 'broad' contribution coming from the second term on the right-hand side of (76). What remains is to devise a method to analyse the 'narrow' contribution, corresponding to the first term on the right-hand side of (76).

In the proof of 1-dimensional Strichartz estimate (Proposition 6.1), we used a combination of parabolic rescaling and induction-on-scale to estimate the narrow contribution. This direct approach is particular to the 1-dimensional setting and significant complications arise when trying to extending these ideas to higher dimensions.

To understand the difficulties, recall that the narrow term involves the operator U applied to functions of the form

$$f_V := \sum_{\tau \in \mathscr{T}_{K^{-1}}(V)} f_\tau \qquad \text{for } V \in \mathrm{Gr}(d, \mathbb{R}^{n+1}).$$

When $n = d = 1$, the function f_V essentially corresponds to some f_τ.[(22)] In this case, Uf_V is Fourier supported on a cap on the parabola and we can exploit the scaling structure of the parabola in the guise of Lemma 4.18. However, for higher dimensions, the best we can say is that the function Uf_V is supported on a parabolic strip. There is no viable scaling which maps a strip to the whole paraboloid.

6.5. Analysis of the narrow term: decoupling

To analyse the narrow term in (76), we appeal to *decoupling estimates* and the following celebrated theorem of Bourgain and Demeter. To introduce the result, let $\mathscr{Q}_{K^2}$ denote the collection of space-time lattice K^2-cubes which intersect $B^{n+1}(0, R)$.

[(22)] More precisely, f_V is a sum of f_τ over a family of $O(1)$ adjacent caps, but this distinction is unimportant.

Theorem 6.5 (Bourgain and Demeter, 2015). *Let $1 \leqslant d \leqslant n+1$ and $2 \leqslant q \leqslant 2 \cdot \frac{d+1}{d-1}$. For all $\varepsilon > 0$, $K \geqslant 1$ we have*

$$\|Uf_V\|_{L^q(Q)} \lesssim_\varepsilon K^\varepsilon \Big(\sum_{\tau \in \mathscr{T}_{K^{-1}}(V)} \|Uf_\tau\|^2_{L^q(w_Q)} \Big)^{1/2} \tag{81}$$

whenever $Q \in \mathscr{Q}_{K^2}$ and $V \subseteq \mathbb{R}^{n+1}$ be a linear subspace of dimension d.

The inequality (81) is understood to hold for all $f \in L^2(\mathbb{R}^n)$ with f_V and f_τ as defined above. The weight w_Q is as in Definition 4.4. A crucial feature of Theorem 6.5 is that the implied constant does not depend on the scale parameter K.

Remark 6.6. Theorem 6.5 is not explicitly stated in Bourgain and Demeter (2015), but it can be easily be deduced as a consequence of Theorem 1.1 in Bourgain and Demeter (2015): see the proof of Lemma 9.5 in Guth (2018).

Theorem 6.5 provides an effective comparison between the function f_V and its constituent parts f_τ for $\tau \in \mathscr{T}_{K^{-1}}(V)$. Note that the right-hand side of (81) involves the norms $\|Uf_\tau\|_{L^q(w_Q)}$, which are amenable to parabolic rescaling. Thus, the Bourgain–Demeter theorem allows us to access in higher dimensions the parabolic rescaling and induction-on-scale arguments which are more readily available in the 1-dimensional case.

We will not present a proof of Theorem 6.5; indeed, the argument of Bourgain and Demeter (2015) is lengthy and complex, incorporating all the techniques we have so far encountered (wave packet analysis, parabolic rescaling, multilinear Strichartz estimates, broad-narrow analysis, and so on). We will, however, make some elementary remarks to contextualise the result.

We first note that there are elementary ways to compare f_V with the f_τ. One example is the triangle inequality, which can be combined with Cauchy–Schwarz to give

$$\|Uf_V\|_{L^q(Q)} \leqslant \sum_{\tau \in \mathscr{T}_{K^{-1}}(V)} \|Uf_\tau\|_{L^q(Q)} \lesssim K^{(d-1)/2} \Big(\sum_{\tau \in \mathscr{T}_{K^{-1}}(V)} \|Uf_\tau\|^2_{L^q(Q)} \Big)^{1/2} \tag{82}$$

for all $1 \leqslant q \leqslant \infty$. However, (for large K) the $K^{(d-1)/2}$ factor on the right-hand side of (82) is much larger than corresponding factor on the right-hand side of (81). The weakness in the triangle inequality is that it does not take into account cancellation between the terms Uf_τ in the sum defining $Uf_V = \sum_{\tau \in \mathscr{T}_{K^{-1}}(V)} Uf_\tau$. Indeed, the Uf_τ oscillate with distinct frequencies (that is, they have disjoint Fourier support) and so it is natural to expect significant cancellation.

For $q = 2$, we can use Plancherel's theorem to exploit the disjoint frequency support and arrive at the substantially stronger estimate

$$\|Uf_V\|_{L^2(Q)} \leqslant \Big(\sum_{\tau \in \mathscr{T}_{K^{-1}}(V)} \|Uf_\tau\|^2_{L^2(w_Q)} \Big)^{1/2}.$$

This can then be interpolated[(23)] with a trivial Cauchy–Schwarz estimate at $q = \infty$ (using the fact that $\#\mathscr{T}_{K^{-1}}(V) \lesssim K^{d-1}$) to give

$$\|Uf_V\|_{L^q(Q)} \lesssim K^{(d-1)(1/2-1/q)} \Big(\sum_{\tau \in \mathscr{T}_{K^{-1}}(V)} \|Uf_\tau\|^2_{L^q(w_Q)} \Big)^{1/2} \tag{83}$$

for $2 \leqslant q \leqslant \infty$. Although (83) improves over (82) by taking into account orthogonality properties, it still falls far short of the decoupling estimate in (81). To improve over, (83) it is necessary to not only use the basic disjointness of the Fourier support, but also the specific parabolic geometry. Thus, all the techniques we have encountered in §4 and §5 are relevant to the proof of Theorem 6.5.

6.6. L^p broad-narrow decomposition in higher dimensions

We wish to apply the decoupling estimate from Theorem 6.5 to bound the narrow term arising from our broad-narrow decomposition. However, the form of the pointwise broad-narrow decomposition from Lemma 6.4 is not suited to this. The main problem is that the maximum over $V \in \mathrm{Gr}(n, \mathbb{R}^{n+1})$ is taken pointwise, and so the maximising subspace V_z can vary from point to point. In order to apply Theorem 6.5, we need to fix a *single* subspace V over a whole K^2-cube.

To get around this issue, we use the locally constant properties of the solution operator U, dictated by the uncertainty principle. In particular, since Uf (or Uf_τ, or Uf_V) is frequency localised at unit scale, $|Uf|$ (or $|Uf_\tau|$, or $|Uf_V|$) should be locally constant at unit scale. Thus, we should be able to fix a single maximising subspace V over any given unit cube.

Unit scale cubes are still too small to apply Theorem 6.5, which requires cubes of side-length at least K^2. Nevertheless, we can adapt the above argument to work at the K^2 spatial scale. The idea is that $|Uf|$ will still be 'locally constant up to K^2 factors' over K^2-cubes.[(24)] We then proceed as before, working with a fixed maximising subspace V over any given K^2-cube, but including additional K^2 factors due to the lack of true local constancy at this scale. It is vital, however, that matters are arranged so

[(23)] The decoupling inequalities are not norm inequalities for linear operators in the usual sense, so one has to be somewhat careful when it comes to interpolation. Nevertheless, it is possible to appeal to classical interpolation results by suitably interpreting the estimates.

[(24)] Indeed, on a heuristic level, this is just a consequence of the local constancy property at unit scale. To rigorously implement this principle, we shall apply Lemma 5.9 with $M = K^2$.

that these additional powers of K^2 appear only in the broad term. Indeed, as we saw in the proof of Proposition 6.1, we must carefully control the K power in the narrow term in order for the decomposition to be effective.

Rigorous implementation of these ideas leads to the following bound.

Lemma 6.7 (L^q broad-narrow decomposition). *For all $1 \leqslant q \leqslant \infty$ and $Q \in \mathscr{Q}_{K^2}$, we have*

$$\|Uf\|_{L^q(Q)} \lesssim \max_{V \in \mathrm{Gr}(n,\mathbb{R}^{n+1})} \|Uf_V\|_{L^q(Q)} + K^E \max_{(\tau_1,\dots,\tau_{n+1}) \in \mathscr{T}^{\mathrm{trans}}_{K^{-1}}} \Big\| \prod_{j=1}^{n+1} |Uf_{\tau_j}|^{1/(n+1)} \Big\|_{L^{q,*}(Q)}. \tag{84}$$

Here $E = E_n$ is a dimensional constant.

Here the $L^{q,*}$ expression is a 'fuzzy' variant of the usual L^q norm, which plays a largely unimportant technical rôle. In particular, let η_{K^2} be a continuous, L^1-normalised function which is concentrated in $[-K^2/2, K^2/2]^{n+1}$, as in the statement of Lemma 5.9. For

$$\vec{\eta}_{K^2}(\vec{w}) := \eta_{K^2}(w_1) \cdots \eta_{K^2}(w_{n+1}), \qquad \vec{w} = (w_1, \dots, w_{n+1}) \in (\mathbb{R}^{n+1})^{n+1},$$

and any choice of measurable set $S \subseteq \mathbb{R}^{n+1}$, we then define

$$\Big\| \prod_{j=1}^{n+1} |Uf_{\tau_j}|^{1/(n+1)} \Big\|_{L^{q,*}(S)} := \int_{(\mathbb{R}^{n+1})^{n+1}} \Big\| \prod_{j=1}^{n+1} |Uf_{\tau_j}(\,\cdot - w_j)|^{1/(n+1)} \Big\|_{L^q(S)} \vec{\eta}_{K^2}(\vec{w})\, \mathrm{d}\vec{w}.$$

Such expressions arise due to the appearance of the mollifier in the rigorous formulation of the locally constant property from Lemma 4.1 and Lemma 5.9. In view of the translation invariance, multilinear estimates such as (55) automatically imply 'fuzzy' variants, with the left-hand L^p norm replaced with the corresponding $L^{p,*}$ norm.

Before presenting the proof of Lemma 6.7, we discuss an immediate consequence. By combining the L^q broad-narrow decomposition with the decoupling inequality from Theorem 6.5, we arrive at the following broad-narrow decomposition.

Proposition 6.8. *For all $2 \leqslant q \leqslant q_n := 2 \cdot \frac{n+1}{n-1}$, $Q \in \mathscr{Q}_{K^2}$ and all $\varepsilon > 0$, we have*

$$\begin{aligned} \|Uf\|_{L^q(Q)} \lesssim_\varepsilon K^\varepsilon \Big(\sum_{\tau \in \mathscr{T}_{K^{-1}}} \|Uf_\tau\|^2_{L^q(w_Q)} \Big)^{1/2} \\ + K^E \max_{(\tau_1,\dots,\tau_{n+1}) \in \mathscr{T}^{\mathrm{trans}}_{K^{-1}}} \Big\| \prod_{j=1}^{n+1} |Uf_{\tau_j}|^{1/(n+1)} \Big\|_{L^{q,*}(Q)}. \end{aligned} \tag{85}$$

Here $E = E_n$ is a dimensional constant.

Remark 6.9. By examining the proof below, it is possible to take $E := 4n^2$ in both Lemma 6.7 and Proposition 6.8.

Proposition 6.8 plays a central rôle in the proof of Theorem 3.7. The narrow term is now of a similar form to that of the $n = 1$ case in (65), and therefore amenable to parabolic rescaling and induction arguments. A crucial feature of Proposition 6.8 is the large range of exponents $2 \leqslant q \leqslant 2 \cdot \frac{n+1}{n-1}$ for which the estimate holds. One could attempt to dispense with the broad-narrow decomposition entirely and apply Theorem 6.5 directly with $d = n + 1$. This leads to the bound

$$\|Uf\|_{L^q(Q)} \lesssim_\varepsilon K^\varepsilon \Big(\sum_{\tau \in \mathscr{T}_{K^{-1}}} \|Uf_\tau\|_{L^q(w_Q)}^2 \Big)^{1/2} \qquad \text{for } 2 \leqslant q \leqslant 2 \cdot \tfrac{n+2}{n}.$$

However, in a similar spirit to the observations of §5.1, the more restrictive range $2 \leqslant q \leqslant 2 \cdot \frac{n+2}{n}$ is insufficient for the purpose of proving Theorem 3.7.

Remark 6.10. Proposition 6.8 essentially appears in Bourgain and Guth (2011) for the restricted range $2 \leqslant q \leqslant 2 \cdot \frac{n}{n+1}$. More precisely, the arguments of Bourgain and Guth (2011) can be used to show that if the decoupling estimate (81) is valid for some $q \geqslant 2$ (and $d = n$), then (85) holds for the same q. It also follows from the methods of Bourgain and Guth (2011) that the decoupling estimate holds for $2 \leqslant q \leqslant 2 \cdot \frac{n}{n+1}$ (see also Bourgain (2013a), where the connection with decoupling is more explicit). The full range $2 \leqslant q \leqslant q_n$ for decoupling followed later in Bourgain and Demeter (2015).

Proof (of Lemma 6.7). Given a cube $Q \in \mathscr{Q}_{K^2}$, we define the *L^q broad functional*

$$\|Uf\|_{\mathrm{BL}^q(Q)} := \min_{V \in \mathrm{Gr}(n, \mathbb{R}^{n+1})} \max_{\tau \notin \mathscr{T}_{K^{-1}}(V)} \|Uf_\tau\|_{L^q(Q)} \tag{86}$$

and let

$$\mathscr{T}_{K^{-1}}(V, Q) := \big\{ \tau \in \mathscr{T}_{K^{-1}}(V) : \|Uf_\tau\|_{L^q(Q)} \geqslant \|Uf\|_{\mathrm{BL}^q(Q)} \big\};$$

these definitions are the natural L^q analogues of the pointwise definitions appearing in the proof of Lemma 6.4.

From all spaces realising the minimum in (86), choose $V_Q \in \mathrm{Gr}(n, \mathbb{R}^{n+1})$ so as to also maximise $\#\mathscr{T}_{K^{-1}}(V, Q)$. Define the collections of *narrow* and *broad* caps (for Uf over Q) by

$$\mathscr{N}_Q := \mathscr{T}_{K^{-1}}(V_Q) \qquad \text{and} \qquad \mathscr{B}_Q := \mathscr{T}_{K^{-1}} \setminus \mathscr{T}_{K^{-1}}(V_Q).$$

Arguing as in the proof of Lemma 6.4, we then have

$$\begin{aligned} \|Uf\|_{L^q(Q)} &\leqslant \Big\| \sum_{\tau \in \mathscr{N}_Q} Uf_\tau \Big\|_{L^q(Q)} + \sum_{\tau \in \mathscr{B}_Q} \|Uf_\tau\|_{L^q(Q)} \\ &\lesssim \max_{V \in \mathrm{Gr}(n, \mathbb{R}^{n+1})} \|Uf_V\|_{L^q(Q)} + K^n \|Uf\|_{\mathrm{BL}^q(Q)}. \end{aligned} \tag{87}$$

Furthermore,

$$\|Uf\|_{\mathrm{BL}^q(Q)} \leqslant \max_{(\tau_1,\dots,\tau_{n+1})\in \mathscr{T}_{K^{-1}}^{\mathrm{trans}}} \prod_{j=1}^{n+1} \|Uf_{\tau_j}\|_{L^q(Q)}^{1/(n+1)}. \tag{88}$$

Comparing the 'broad' term on the right-hand side of (88) with the corresponding term in (84), we can see that the order of the geometric mean and L^q norms are interchanged. We can in fact bound the right-hand side of (84) by the right-hand side of (88) simply by Hölder's inequality, but this estimate goes in the wrong direction. Thus, the problem is to prove a reverse Hölder inequality, which is achieved using the locally constant properties.

Let $\tilde{\chi} \in \mathscr{S}(\mathbb{R}^{n+1})$ satisfy $|\tilde{\chi}(z)| \gtrsim 1$ for all $|z| \leqslant 2$ and $\operatorname{supp} \mathscr{F}\tilde{\chi} \subseteq B^{n+1}(0,1)$. Fix caps $\tau_j \in \mathscr{T}_{K^{-1}}$ for $1 \leqslant j \leqslant n+1$ and define the functions

$$F_j(z) := Uf_{\tau_j}(z) \cdot \tilde{\chi}(R^{-1}z), \quad \text{for } 1 \leqslant j \leqslant n+1.$$

Since each cube $Q \in \mathscr{Q}_{K^2}$ satisfies $Q \subseteq B^{n+1}(0,2R)$, it follows that

$$\prod_{j=1}^{n+1} \|Uf_{\tau_j}\|_{L^q(Q)}^{1/(n+1)} \lesssim \prod_{j=1}^{n+1} \|F_j\|_{L^q(Q)}^{1/(n+1)} \leqslant \Big[\prod_{j=1}^{n+1} |a_{j,Q}|^{1/(n+1)}\Big] |Q|^{1/q},$$

where $a_{j,Q}$ denotes the supremum of $|F_j(z)|$ over all $z \in Q$. Applying Lemma 5.9 to each function F_j with exponent $s := 1/(n+1)$ and $M := K^2$, we deduce that

$$\begin{aligned}\prod_{j=1}^{n+1} \|Uf_{\tau_j}\|_{L^q(Q)}^{1/(n+1)} &\lesssim K^{2(n+1)} \Big\| \prod_{j=1}^{n+1} |F_j|^{1/(n+1)} * \eta_{K^2} \Big\|_{L^q(Q)} \\ &\lesssim K^{2(n+1)} \Big\| \prod_{j=1}^{n+1} |Uf_{\tau_j}|^{1/(n+1)} \Big\|_{L^{q,*}(Q)}, \end{aligned} \tag{89}$$

where the second step follows by Minkowski's inequality. Note that the additional factor of $K^{2(n+1)}$ arises since the F_j are only Fourier localised to scale 1 (rather than scale K^{-2}) and therefore only enjoy local constancy at unit scale. Combining (89) with (87) and (88), we deduce the desired bound. □

7. Proof of the fractal energy estimate

7.1. Recap and a final reduction

We now combine all the tools introduced in the previous sections to prove the fractal energy estimate. For convenience, here we reproduce the statement.

Theorem 7.1 (Du and Zhang, 2019, *c.f.* Theorem 3.7). *For all $\varepsilon > 0$ and all $R \geqslant 1$, $1 \leqslant \alpha \leqslant n+1$, the inequality*

$$\|Uf\|_{L^2(Z_{\mathscr{Q}})} \lesssim_\varepsilon \Delta_\alpha(\mathscr{Q})^{1/(n+1)} R^{\alpha/(2(n+1))+\varepsilon} \|f\|_{L^2(\mathbb{R}^n)} \tag{90}$$

holds whenever $f \in L^2(\mathbb{R}^n)$ and $\mathscr{Q}$ is a family of lattice unit cubes in $B^{n+1}(0,R)$.

From §3, we know that Theorem 7.1 implies the sharp L^2 maximal estimate in Theorem 3.1 and therefore also the pointwise convergence result for the Schrödinger equation in Theorem 1.3.

The proof of Theorem 7.1 hinges on the broad-narrow decomposition and induction-on-scale methods. Prior to Du and Zhang (2019), these techniques were applied to bound the Schrödinger maximal function in Bourgain (2013b). The latter work establishes $H^s \to L^2$ bounds for the maximal function in the range $s > \frac{2n-1}{4n}$, which is more restrictive than the (essentially sharp) condition $s > \frac{n}{2(n+1)}$ from Du and Zhang (2019). One of the main advantages of the approach of Du and Zhang (2019) is the novel form of the inductive statement, which allows a great deal of information to be translated between scales.

Rather than attempting to prove (90) directly, we work with an auxiliary $L^2 \to L^{q_n}$ estimate, where $q_n := 2 \cdot \frac{n+1}{n-1}$. This is the exponent featured in the broad-narrow decomposition from Proposition 6.8 (which arises from the application of the lower-dimensional decoupling inequality from Theorem 6.5).

Proposition 7.2. *Let $q_n := 2 \cdot \frac{n+1}{n-1}$. For all $0 < \varepsilon < 1$ and $R \geqslant 1$, defining $\delta := \varepsilon/100n^2$ and $K := R^\delta$, the following holds. Suppose $f \in L^2(\mathbb{R}^n)$ and $\mathscr{Q}$ is a family of lattice K^2-cubes contained in $B^{n+1}(0,R)$ such that*[25]

$$\|Uf\|_{L^{q_n}(Q)} \qquad \text{are dyadically constant over } Q \in \mathscr{Q}. \tag{91}$$

Then for all $1 \leqslant \alpha \leqslant n+1$, we have

$$\|Uf\|_{L^{q_n}(Z_{\mathscr{Q}})} \lesssim_\varepsilon \left[\frac{\Delta_\alpha(\mathscr{Q})}{\#\mathscr{Q}}\right]^{1/(n+1)} R^{\alpha/(2(n+1))+\varepsilon} \|f\|_{L^2(\mathbb{R}^n)}. \tag{92}$$

It is not difficult to show that Proposition 7.2 implies Theorem 7.1. This relies on a pigeonholing argument, used to pass to the dyadically constant setup, which we isolate in Lemma 7.3 below. Before giving the details of this reduction, we indicate why the formulation of Proposition 7.2 is useful. Two major ingredients in our analysis are the multilinear Strichartz estimates from Theorem 5.5 and the decoupling inequality used in the broad-narrow decomposition in Proposition 6.8. The critical exponent for the former is $p_n := 2 \cdot \frac{n+1}{n}$ and for the latter is q_n. Roughly speaking, the setup in Proposition 7.2 allows us to apply both these estimates at their respective critical

[25] For the definition of dyadically constant, see §4.6.

exponents. The proposition itself is stated for the exponent q_n, whilst the dyadic constancy hypothesis (91) allows one to efficiently pass to the exponent p_n using reverse Hölder-type arguments (see §7.4 below). Thus, the pigeonholing can be thought of roughly as a weak form of real interpolation, which allows us to reconcile two very different estimates at distinct Lebesgue exponents.

Lemma 7.3. *Let $2 \leqslant q \leqslant \infty$ and $1 \leqslant M \leqslant R$. Suppose $f \in L^2(\mathbb{R}^n)$ and $\mathscr{Q}_M$ is a collection of lattice M-cubes contained in $B^{n+1}(0,R)$. Then there exists a subcollection $\mathscr{Q}'_M \subseteq \mathscr{Q}_M$ such that*

$$\|Uf\|_{L^q(Q)} \qquad \text{are dyadically constant over } Q \in \mathscr{Q}'_M. \tag{93}$$

and

$$\|Uf\|_{L^2(Z_{\mathscr{Q}_M})} \lesssim (\log R)^{1/2} \|Uf\|_{L^2(Z_{\mathscr{Q}'_M})} + R^{-100n} \|f\|_{L^2(\mathbb{R}^n)}. \tag{94}$$

Proof. Begin by bounding

$$\|Uf\|_{L^2(Z_{\mathscr{Q}_M})} \leqslant \Big(\sum_{Q \in \mathscr{Q}_{M,0}} \|Uf\|^2_{L^2(Q)} \Big)^{1/2} + \Big(\sum_{Q \in \mathscr{Q}_{M,1}} \|Uf\|^2_{L^2(Q)} \Big)^{1/2} \tag{95}$$

where $\mathscr{Q}_{M,0}$ and $\mathscr{Q}_{M,1}$ are defined by

$$\mathscr{Q}_{M,0} := \big\{ Q \in \mathscr{Q}_M : \|Uf\|_{L^q(Q)} < R^{-200n} \|f\|_{L^2(\mathbb{R}^n)} \big\}, \quad \mathscr{Q}_{M,1} := \mathscr{Q}_M \setminus \mathscr{Q}_{M,0}.$$

The net contribution to (95) arising from the cubes $Q \in \mathscr{Q}_{M,0}$ is negligible, and can be bounded by the second term on the right-hand side of (94).

By the definition of $\mathscr{Q}_{M,0}$ and elementary estimates,

$$R^{-200n} \|f\|_{L^2(\mathbb{R}^n)} \leqslant \|Uf\|_{L^q(Q)} \lesssim R^{(n+1)/q} \|f\|_{L^2(\mathbb{R}^n)} \qquad \text{for all } Q \in \mathscr{Q}_{M,1}.$$

Consequently, we may apply dyadic pigeonholing in the form of Lemma 4.21 ii) to find $\mathscr{Q}'_M \subseteq \mathscr{Q}_M$ satisfying (93) and such that

$$\Big(\sum_{Q \in \mathscr{Q}_{M,1}} \|Uf\|^2_{L^2(Q)} \Big)^{1/2} \lesssim (\log R)^{1/2} \Big(\sum_{Q \in \mathscr{Q}'_M} \|Uf\|^2_{L^2(Q)} \Big)^{1/2},$$

which combines with (95) and our earlier observation to give the desired bound. □

Proof (Proposition 7.2 ⇒ Theorem 7.1). Let $\mathscr{Q}$ be a family of lattice unit cubes in $B^{n+1}(0,R)$ and $\mathscr{Q}_{K^2}$ the collection of lattice K^2-cubes which intersect $Z_{\mathscr{Q}}$. Here $K := R^{\delta}$ is as in the statement of Proposition 7.2.

By Lemma 7.3, there exists a subcollection $\mathscr{Q}'_{K^2} \subseteq \mathscr{Q}_{K^2}$ such that

$$\|Uf\|_{L^{q_n}(Q)} \qquad \text{are dyadically constant over } Q \in \mathscr{Q}'_{K^2}$$

and

$$\|Uf\|_{L^2(Z_{\mathscr{Q}})} \leqslant \|Uf\|_{L^2(Z_{\mathscr{Q}_{K^2}})} \lesssim (\log R)^{1/2}\|Uf\|_{L^2(Z_{\mathscr{Q}'_{K^2}})} + R^{-100n}\|f\|_{L^2(\mathbb{R}^n)}. \quad (96)$$

We can therefore apply Proposition 7.2 to the family $\mathscr{Q}'_{K^2}$ to deduce that

$$\|Uf\|_{L^{q_n}(Z_{\mathscr{Q}'_{K^2}})} \lesssim_\varepsilon \Big[\frac{\Delta_\alpha(\mathscr{Q}'_{K^2})}{\#\mathscr{Q}'_{K^2}}\Big]^{1/(n+1)} R^{\alpha/(2(n+1))+\varepsilon/2}\|f\|_{L^2(\mathbb{R}^n)}. \quad (97)$$

We now apply Hölder's inequality and combine (96) and (97) to deduce that

$$\begin{aligned}\|Uf\|_{L^2(Z_{\mathscr{Q}})} &\lesssim (\log R)^{1/2}|Z_{\mathscr{Q}'_{K^2}}|^{1/(n+1)}\|Uf\|_{L^{q_n}(Z_{\mathscr{Q}'_{K^2}})} + R^{-100n}\|f\|_{L^2(\mathbb{R}^n)}\\ &\lesssim_\varepsilon R^{2\delta}(\log R)^{1/2}\Delta_\alpha(\mathscr{Q}'_{K^2})^{1/(n+1)}R^{\alpha/(2(n+1))+\varepsilon/2}\|f\|_{L^2(\mathbb{R}^n)},\end{aligned}$$

where we have used the elementary bound

$$|Z_{\mathscr{Q}'_{K^2}}| \leqslant K^{2(n+1)}[\#\mathscr{Q}'_{K^2}] = R^{2(n+1)\delta}[\#\mathscr{Q}'_{K^2}]$$

Note that the factor $R^{2\delta}(\log R)^{1/2}$ can be bounded by $R^{\varepsilon/2}$.

Finally, since each $Q_{K^2} \in \mathscr{Q}_{K^2}$ has the property that $2\cdot Q_{K^2}$ contains at least one cube from $\mathscr{Q}$, it easily follows that $\Delta_\alpha(\mathscr{Q}'_{K^2}) \lesssim \Delta_\alpha(\mathscr{Q})$. Consequently,

$$\|Uf\|_{L^2(Z_{\mathscr{Q}})} \lesssim_\varepsilon \Delta_\alpha(\mathscr{Q})^{1/(n+1)}R^{\alpha/(2(n+1))+\varepsilon}\|f\|_{L^2(\mathbb{R}^n)},$$

which is precisely the desired bound. □

7.2. Induction-on-scale

The proof of Proposition 7.2 follows by induction on the scale parameter R. Here we introduce the underlying induction scheme.

Fix $\varepsilon > 0$, $1 \leqslant \alpha \leqslant n+1$ and set $\delta := \varepsilon/100n^2$. We shall say a parameter is *admissible* if it depends only on the dimension n and ε. Recall from (16) that the basic energy estimate

$$\|Uf\|_{L^2(Z_{\mathscr{Q}})} \lesssim \Delta_\alpha(\mathscr{Q})^{1/(n+1)}R^{1/2}\|f\|_{L^2(\mathbb{R}^n)}$$

always holds. Thus, Proposition 7.2 is trivial for small scales. In particular, let $R_0 \geqslant 1$ denote a fixed scale, depending only on admissible parameters n and ε. For $\mathbf{C}_\varepsilon \geqslant 1$ a suitable choice of admissible constant,

$$\|Uf\|_{L^{q_n}(Z_{\mathscr{Q}})} \leqslant \mathbf{C}_\varepsilon\Big[\frac{\Delta_\alpha(\mathscr{Q})}{\#\mathscr{Q}}\Big]^{1/(n+1)}R^{\alpha/(2(n+1))+\varepsilon}\|f\|_{L^2(\mathbb{R}^n)}, \quad (98)$$

holds under the hypotheses of Proposition 7.2 whenever $1 \leqslant R \leqslant R_0$. This serves as the base case for our induction.

Henceforth we assume R_0 and $\mathbf{C}_\varepsilon$ are fixed admissible constants, chosen so that the above base case holds and large enough to satisfy the forthcoming requirements of the proof. Let $R \geqslant R_0$ and define $K := R^\delta$. We shall work under the following induction hypothesis.

Induction hypothesis. *Let $1 \leqslant \tilde{R} \leqslant R/2$ and define $\tilde{K} := \tilde{R}^\delta$. The inequality*

$$\|Ug\|_{L^{q_n}(Z_{\tilde{\mathscr{Q}}})} \leqslant \mathbf{C}_\varepsilon \Big[\frac{\Delta_\alpha(\tilde{\mathscr{Q}})}{\#\tilde{\mathscr{Q}}}\Big]^{1/(n+1)} \tilde{R}^{\alpha/(2(n+1))+\varepsilon} \|g\|_{L^2(\mathbb{R}^n)}.$$

holds whenever $g \in L^2(\mathbb{R}^n)$ and $\tilde{\mathscr{Q}}$ is a non-empty family of lattice $\tilde{K}^2$-cubes contained in $B^{n+1}(0,R)$ such that

$$\|Ug\|_{L^{q_n}(\tilde{Q})} \qquad \text{are dyadically constant over } \tilde{Q} \in \tilde{\mathscr{Q}}.$$

We now fix $f \in L^2(\mathbb{R}^n)$ and for $R \geqslant R_0$ as above let $\mathscr{Q}$ be a family of K^2-cubes satisfying the hypotheses of Proposition 7.2. The goal is to prove (98). Without loss of generality, we may assume $\operatorname{supp} \hat{f} \subseteq B^n(0,1/2)$.

Remark 7.4. The high-level structure of the proof is similar to the induction-on-scale argument used earlier to prove the 1-dimensional Strichartz estimate (Proposition 6.1). There are, however, some notable differences:

- ▷ The intermediate scale K introduced in the statement of Proposition 7.2 plays the same rôle as the K parameter in the proof of Proposition 6.1. Here, however, $K = R^\delta$ depends on the inadmissible parameter R whereas in the proof of Proposition 6.1 the parameter K was chosen independently of R. This choice of K allows us to compensate for small losses in the narrow term arising from the K^ε factor in the decoupling inequality (Theorem 6.5).
- ▷ To close the induction, later in the argument we must ensure that K is sufficiently large. Rather than fix a specific value of K, as in the proof of Proposition 6.1, here we fix a lower bound for $K = R^\delta$ by introducing the parameter R_0.

7.3. Broad / narrow dichotomy

The next step is to apply the broad-narrow decomposition described in §6.6. This will allow us to bring the powerful multilinear Strichartz estimates into play.

By Proposition 6.8, there exists admissible constants $C_\varepsilon^{\mathrm{bn}}$, $E \geqslant 1$ such that

$$\begin{aligned}\|Uf\|_{L^{q_n}(Q)} \leqslant C_\varepsilon^{\mathrm{bn}} K^\varepsilon \Big(\sum_{\tau \in \mathscr{T}_{K^{-1}}} \|Uf_\tau\|^2_{L^{q_n}(w_Q)} \Big)^{1/2} \\ + C_\varepsilon^{\mathrm{bn}} K^E \max_{\tau \in \mathscr{T}^{\mathrm{trans}}_{K^{-1}}} \Big\| \prod_{j=1}^{n+1} |Uf_{\tau_j}|^{1/(n+1)} \Big\|_{L^{q_n,*}(Q)}. \end{aligned} \tag{99}$$

holds for all $Q \in \mathscr{Q}$. Here the maximum is taken over all transverse $(n+1)$-tuples $\tau = (\tau_1, \dots, \tau_{n+1}) \in \mathscr{T}_{K^{-1}}^{\mathrm{trans}}$. By Remark 6.9 we may (and shall) take $E := 4n^2$.

We say a cube $Q \in \mathscr{Q}$ is *narrow* if

$$\|Uf\|_{L^{q_n}(Q)} \leqslant 2C_\varepsilon^{\mathrm{bn}} K^\varepsilon \Big(\sum_{\tau \in \mathscr{T}_{K^{-1}}} \|Uf_\tau\|_{L^{q_n}(w_Q)}^2 \Big)^{1/2};$$

otherwise, we say Q is *broad*. As a consequence of (99), the inequality

$$\|Uf\|_{L^{q_n}(Q)} \leqslant 2C_\varepsilon^{\mathrm{bn}} K^E \max_{\tau \in \mathscr{T}_{K^{-1}}^{\mathrm{trans}}} \Big\| \prod_{j=1}^{n+1} |Uf_{\tau_j}|^{1/(n+1)} \Big\|_{L^{q_n,*}(Q)} \tag{100}$$

holds whenever $Q \in \mathscr{Q}$ is broad.

We denote by $\mathscr{Q}_{\mathrm{broad}}$ and $\mathscr{Q}_{\mathrm{narrow}}$ the collections of broad and narrow cubes, respectively. The proof will split into two subcases, depending on whether the majority of the cubes are broad or narrow.

7.4. Broad-dominant case

Suppose the majority of the cubes $Q \in \mathscr{Q}$ are broad: that is,

$$\#\mathscr{Q}_{\mathrm{broad}} \geqslant \#\mathscr{Q}/2. \tag{101}$$

We refer to this as the *broad-dominant case*. Here we bound our operator by direct appeal to the multilinear Strichartz estimate from Theorem 5.5. This should come as no surprise, since we have already seen in Corollary 5.6 that Theorem 5.5 implies multilinear fractal energy estimates. Due to the form of the desired estimate in (92), we do not appeal directly to Corollary 5.6, but the underlying idea is nevertheless the same.

Let $\lambda \geqslant 1$ and $\mathscr{Q}' \subseteq \mathscr{Q}_{\mathrm{broad}}$ be any collection of broad cubes which satisfies $\#\mathscr{Q}' \geqslant \lambda^{-1}\#\mathscr{Q}$. Then, by the dyadic constancy hypothesis (91), we may bound

$$\|Uf\|_{L^{q_n}(Z_{\mathscr{Q}})} \lesssim \lambda^{1/q_n} \|Uf\|_{L^{q_n}(Z_{\mathscr{Q}'})} \leqslant \lambda \Big(\sum_{Q \in \mathscr{Q}'} \|Uf\|_{L^{q_n}(Q)}^{q_n} \Big)^{1/q_n}.$$

Combining this with (100), there exists an assignment of a transverse $(n+1)$-tuple of caps $\tau_Q = (\tau_{Q,1}, \dots, \tau_{Q,n+1}) \in \mathscr{T}_{K^{-1}}^{\mathrm{trans}}$ to each $Q \in \mathscr{Q}'$ such that

$$\|Uf\|_{L^{q_n}(Z_{\mathscr{Q}})} \lesssim \lambda K^E \Big(\sum_{Q \in \mathscr{Q}'} \Big\| \prod_{j=1}^{n+1} |Uf_{\tau_{Q,j}}|^{1/(n+1)} \Big\|_{L^{q_n,*}(Q)}^{q_n} \Big)^{1/q_n}.$$

Let $p_n := 2 \cdot \frac{n+1}{n}$ denote the critical exponent in the multilinear Strichartz estimate (Theorem 5.5). By the local multilinear Bernstein inequality from Corollary 5.10, we deduce that[(26)]

$$\|Uf\|_{L^{q_n}(Z_{\mathscr{Q}})} \lesssim \lambda K^E \Big(\sum_{Q \in \mathscr{Q}'} \Big\| \prod_{j=1}^{n+1} |Uf_{\tau_{Q,j}}|^{1/(n+1)} \Big\|_{L^{p_n,*}(w_Q)}^{q_n} \Big)^{1/q_n}. \tag{102}$$

One way to estimate the ℓ^{q_n} sum on the right-hand side of (102) is to simply use the nesting of ℓ^p-norms to deduce that

$$\Big(\sum_{Q \in \mathscr{Q}'} \Big\| \prod_{j=1}^{n+1} |Uf_{\tau_{Q,j}}|^{1/(n+1)} \Big\|_{L^{p_n,*}(w_Q)}^{q_n} \Big)^{1/q_n} \lesssim K^E \max_{\tau \in \mathscr{T}_{K^{-1}}^{\mathrm{trans}}} \Big\| \prod_{j=1}^{n+1} |Uf_{\tau_j}|^{1/(n+1)} \Big\|_{L^{p_n,*}(w_{B_R})}. \tag{103}$$

Here we have interchanged the ℓ^{p_n}-norm and maximum in τ at the expense of an additional factor of K^E. The multilinear expression can now be bounded using Theorem 5.5. However, the resulting estimates are insufficient for our purpose and we shall instead use the flexibility to choose $\mathscr{Q}'$ to improve (103). In particular, we choose $\mathscr{Q}'$ via a pigeonholing argument, along the lines of the reverse Hölder inequality from Lemma 4.22.

Pigeonholing will repeatedly feature in the forthcoming arguments; it is convenient to introduce asymptotic notation to suppress the resulting logarithmic factors.

Notation. *Let A, B be non-negative real numbers. We write $A \lessapprox B$ or $B \gtrapprox A$ if for all $\eta > 0$ there exists a constant $C_{\varepsilon,\eta} \geqslant 1$, depending only on η and the admissible parameters n and ε, such that $A \leqslant C_{\varepsilon,\eta} R^{\eta} B$.*

Let $\mathscr{Q}_{\mathrm{tiny}}$ denote the collection of all cubes $Q \in \mathscr{Q}_{\mathrm{broad}}$ such that

$$\Big\| \prod_{j=1}^{n+1} |Uf_{\tau_{Q,j}}|^{1/(n+1)} \Big\|_{L^{p_n,*}(w_Q)} \leqslant R^{-100n} \|f\|_{L^2(\mathbb{R}^n)}.$$

If $\#\mathscr{Q}_{\mathrm{tiny}} \geqslant \#\mathscr{Q}_{\mathrm{broad}}/2$, then we may apply (102) with $\mathscr{Q}' := \mathscr{Q}_{\mathrm{tiny}}$ to obtain a very favourable estimate. Thus, we assume $\#\mathscr{Q}_{\mathrm{tiny}} < \#\mathscr{Q}_{\mathrm{broad}}/2$.

Note that

$$R^{-100n} \|f\|_{L^2(\mathbb{R}^n)} < \Big\| \prod_{j=1}^{n+1} |Uf_{\tau_{Q,j}}|^{1/(n+1)} \Big\|_{L^{p_n,*}(w_Q)} \lesssim R \|f\|_{L^2(\mathbb{R}^n)}$$

for all $Q \in \mathscr{Q}_{\mathrm{broad}} \setminus \mathscr{Q}_{\mathrm{tiny}}$, where the upper bound is a trivial consequence of the Cauchy–Schwarz inequality and Plancherel's theorem. Thus, by dyadic pigeonholing, there exists some $\mathscr{Q}' \subseteq \mathscr{Q}_{\mathrm{broad}}$ satisfying $\#\mathscr{Q}' \gtrsim \#\mathscr{Q}$ such that

$$\Big\| \prod_{j=1}^{n+1} |Uf_{\tau_{Q,j}}|^{1/(n+1)} \Big\|_{L^{p_n,*}(w_Q)} \qquad \text{are dyadically constant over } Q \in \mathscr{Q}'.$$

[(26)] The weighted $L^{p_n,*}$-norms are defined in the obvious manner; we omit the details.

For this choice of set $\mathscr{Q}'$, we can upgrade (103) to the reverse Hölder inequality

$$\Big(\sum_{Q\in\mathscr{Q}'} \Big\| \prod_{j=1}^{n+1} |Uf_{\tau_{Q,j}}|^{1/(n+1)} \Big\|_{L^{p_n,*}(w_Q)}^{q_n} \Big)^{1/q_n} \\ \lessapprox K^E [\#\mathscr{Q}]^{-1/(2(n+1))} \max_{\tau\in\mathscr{T}_{K^{-1}}^{\mathrm{trans}}} \Big\| \prod_{j=1}^{n+1} |Uf_{\tau_j}|^{1/(n+1)} \Big\|_{L^{p_n,*}(w_{B_R})}. \tag{104}$$

Combining (102) (for the choice of $\mathscr{Q}'$ above) and (104), we deduce that

$$\|Uf\|_{L^{q_n}(Z_{\mathscr{Q}})} \lessapprox K^{2E} [\#\mathscr{Q}]^{-1/(2(n+1))} \max_{\tau\in\mathscr{T}_{K^{-1}}^{\mathrm{trans}}} \Big\| \prod_{j=1}^{n+1} |Uf_{\tau_j}|^{1/(n+1)} \Big\|_{L^{p_n,*}(w_{B_R})}.$$

Now applying the multilinear Strichartz estimate,[(27)]

$$\|Uf\|_{L^{q_n}(Z_{\mathscr{Q}})} \lessapprox K^{3E} [\#\mathscr{Q}]^{-1/(n+1)} [\#\mathscr{Q}]^{1/(2(n+1))} R^{\varepsilon/2} \|f\|_{L^2(\mathbb{R}^n)}.$$

Recall that $\#\mathscr{Q} \leqslant \Delta_\alpha(\mathscr{Q}) R^\alpha$, and therefore

$$\|Uf\|_{L^{q_n}(Z_{\mathscr{Q}})} \lessapprox K^{3E} [\#\mathscr{Q}]^{-1/(n+1)} \Delta_\alpha(\mathscr{Q})^{1/(2(n+1))} R^{\alpha/(2(n+1))+\varepsilon/2} \|f\|_{L^2(\mathbb{R}^n)}.$$

Finally, since $\Delta_\alpha(\mathscr{Q}) \gtrsim K^{-2\alpha}$ and $K = R^\delta$ where $\delta \leqslant \varepsilon/8E$, we conclude that

$$\|Uf\|_{L^{q_n}(Z_{\mathscr{Q}})} \lesssim_\varepsilon \Big[\frac{\Delta_\alpha(\mathscr{Q})}{\#\mathscr{Q}} \Big]^{1/(n+1)} R^{\alpha/(2(n+1))+\varepsilon} \|f\|_{L^2(\mathbb{R}^n)}. \tag{105}$$

This provides a favourable estimate in the broad-dominant case.

7.5. Narrow-dominant case: introduction

Now suppose (101) fails, so that

$$\#\mathscr{Q}_{\mathrm{narrow}} \geqslant \#\mathscr{Q}/2$$

We refer to this as the *narrow-dominant case*. Here we estimate our operator using a combination of parabolic rescaling and appeal to the induction hypothesis.

The analysis of the narrow-dominant case is a major innovation of Du and Zhang (2019). Indeed, as remarked earlier, Bourgain (2013b) introduced the broad/narrow dichotomy to the study of the Schrödinger maximal operator. Many aspects of the arguments of Bourgain (2013b) and Du and Zhang (2019) are similar (the use of

[(27)] Here we are estimating a weighted norm $L^{p_n}(w_{B_R})$ rather than a $L^{p_n}(B_R)$-norm with sharp cut-off as in Theorem 5.5. However, Theorem 5.5 automatically extends to the weighted case, using the translation invariance of the estimate and rapid decay of the weight.

induction-on-scale,[28] broad/narrow dichotomy, decoupling-type estimates, multilinear Strichartz); however, the novel form of the key estimate (92) of Du and Zhang (2019) allows information to be efficiently passed between scales and, consequently, leads to a much tighter bound in the narrow-dominant case.

We shall discuss the narrow-dominant case at length, and attempt to develop both a heuristic and technically detailed understanding of the argument. In particular, the following subsections are structured as follows:

- ▷ In §7.6 we describe the initial steps of the analysis of the narrow-dominant case, setting the scene for the main argument.
- ▷ In §7.7 we provide a non-technical overview of the main argument. For this overview, we make a number of simplifying assumptions.
- ▷ In §7.8 we provide a rigorous description of main argument. The goal here is primarily to remove the simplifying assumptions used in the previous subsection.

7.6. Narrow-dominant case: initial steps

The first few steps of the argument mirror those of the inductive proof of the 1-dimensional Strichartz estimate (Proposition 6.1). For $\tau \in \mathscr{T}_{K^{-1}}$ recall that, since f_τ has frequency localisation to the cube τ, the wave Uf_τ behaves pseudo-locally. In particular, we may write

$$f_\tau = \sum_{S \in \mathbf{S}_\tau[R]} f_S$$

as in (49) where, by Lemma 4.16, the pointwise inequality

$$|Uf_\tau(z)| \lesssim \sum_{S \in \mathbf{S}_\tau[R]} |Uf_S(z)| \chi_{\bar{S}}(z) + R^{-10n} \|f\|_{L^2(\mathbb{R}^n)}$$

holds for all $z \in B^{n+1}(0,R)$.

Fix a narrow cube $Q \in \mathscr{Q}_{\text{narrow}}$. Since the sets $\bar{S}$ have bounded overlap,

$$\|Uf_\tau\|_{L^{q_n}(w_Q)} \lesssim \Big(\sum_{S \in \mathbf{S}_\tau[R]} \|Uf_S\|^{q_n}_{L^{q_n}(w_Q)} \Big)^{1/q_n} + R^{-5n} \|f\|_{L^2(\mathbb{R}^n)}. \tag{106}$$

Let $\mathbf{S}$ denote the (disjoint) union of the sets $\mathbf{S}_\tau[R]$ over all $\tau \in \mathscr{T}_{K^{-1}}$. By (106), the definition of $\mathscr{Q}_{\text{narrow}}$ and the nesting of the ℓ^p spaces,

$$\begin{aligned} \|Uf\|_{L^{q_n}(Q)} &\lesssim_\varepsilon K^\varepsilon \Big(\sum_{\tau \in \mathscr{T}_{K^{-1}}} \|Uf_\tau\|^2_{L^{q_n}(w_Q)} \Big)^{1/2} \\ &\lesssim_\varepsilon K^\varepsilon \Big(\sum_{S \in \mathbf{S}} \|Uf_S\|^2_{L^{q_n}(w_Q)} \Big)^{1/2} + R^{-5n} \|f\|_{L^2(\mathbb{R}^n)}. \end{aligned} \tag{107}$$

[28] The argument of Bourgain (2013b) is presented as a recursive process rather than an induction, but this is tantamount to the same thing.

This situation looks very similar to the analysis of the narrow case in the proof of Proposition 6.1. The key difference, however, is that we must now keep track of the localisation to the various cubes $Q \in \mathscr{Q}$.

7.7. Analysis of the narrow case: non-technical overview

We now give a non-technical overview of the remainder of the proof. Any outstanding technical details are discussed in the following subsection.

Since the weight w_Q is rapidly decaying away from Q, the only strips S which significantly contribute to the sum in (107) are those belonging to

$$\mathbf{S}(Q) := \{S \in \mathbf{S} : S \cap Q \neq \emptyset\}.$$

Thus, we essentially have[29]

$$\|Uf\|_{L^{q_n}(Q)} \lesssim_\varepsilon K^\varepsilon \Big(\sum_{S \in \mathbf{S}(Q)} \|Uf_S\|_{L^{q_n}(Q)}^2 \Big)^{1/2}.$$

We remark that, for any fixed $\tau \in \mathscr{T}_{K^{-1}}$, the collection $\mathbf{S}(Q)$ contains only $O(1)$ strips lying in $\mathbf{S}_\tau[R]$. We may therefore think of $\mathbf{S}(Q)$ as a collection of strips passing through Q which are oriented in K^{-1}-separated directions.

We wish to sum the contributions over all $Q \in \mathscr{Q}$. To do this effectively, we apply Hölder's inequality to convert the ℓ^2 expression into an ℓ^q expression:

$$\|Uf\|_{L^{q_n}(Q)} \lesssim_\varepsilon K^\varepsilon [\#\mathbf{S}(Q)]^{1/(n+1)} \Big(\sum_{S \in \mathbf{S}} \|Uf_S\|_{L^{q_n}(Q)}^{q_n} \Big)^{1/q_n}.$$

We may now take the ℓ^q sum over all $Q \in \mathscr{Q}_{\text{narrow}}$ to deduce that

$$\|Uf\|_{L^{q_n}(Z_{\mathscr{Q}})} \lesssim_\varepsilon K^\varepsilon [\max_{Q \in \mathscr{Q}} \#\mathbf{S}(Q)]^{1/(n+1)} \Big(\sum_{S \in \mathbf{S}} \|Uf_S\|_{L^{q_n}(Z_{\mathscr{Q}})}^{q_n} \Big)^{1/q_n}. \tag{108}$$

Parabolic rescaling and the induction hypothesis. — At this stage, we wish to apply parabolic rescaling and the induction hypothesis to bound each of the terms $\|Uf_S\|_{L^q(Z_{\mathscr{Q}})}$. This is exactly as in the inductive proof of Proposition 6.1. However, one major complication in the present setup is that our induction hypothesis involves some underlying family of $\tilde{K}^2$-cubes $\tilde{\mathscr{Q}}$. We must therefore prepare the ground so that, after rescaling, a suitable family of $\tilde{K}^2$-cubes arises.

Fix $S \in \mathbf{S}$, so that

$$S = \big\{(x,t) \in \mathbb{R}^{n+1} : |x - x(S) - tv(S)| \leqslant R/K \text{ and } |t| \leqslant R\big\}$$

[29] For the sake of this discussion, we will ignore rapidly decaying error terms in the estimates and assume we have sharp cutoffs rather than weighted L^q norms. We address such technical points in the following subsection.

for some choice of initial position $x(S) \in B^n(0,R)$ and velocity $v(S) \in B^n(0,1)$. As in §4.5, let $\mathscr{A}_S\colon \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ denote the affine transformation

$$\mathscr{A}_S\colon (x,t) \mapsto \big(K^{-1}(x - x(S) - tv(S)), K^{-2}t\big)$$

which maps bijectively between S and $B^{n+1}(0, R/K^2)$. Define

$$\tilde{R} := 20R/K^2 \quad \text{and} \quad \tilde{K} := \tilde{R}^{\delta}. \tag{109}$$

Let $\mathscr{P}(S)$ denote a cover of the strip S by parallelepipeds aligned parallel to S and of dimensions $K\tilde{K}^2 \times \cdots \times K\tilde{K}^2 \times K^2\tilde{K}^2$. In particular, $\mathscr{P}(S)$ consists of the sets

$$P := \big\{(x,t) \in \mathbb{R}^{n+1} : |x - x(P) - tv(S)|_\infty \leqslant K\tilde{K}^2/2 \text{ and } |t - t(P)| \leqslant K^2\tilde{K}^2/2\big\}, \tag{110}$$

where the centres $z(P) = (x(P), t(P))$ vary over the lattice points $K\tilde{K}^2\mathbb{Z}^n \times K^2\tilde{K}^2\mathbb{Z}$. The dimensions of these sets are chosen so that the transformation $\mathscr{A}_S$ maps each parallelepiped $P \in \mathscr{P}(S)$ to a lattice $\tilde{K}^2$-cube

$$\mathscr{A}_S(P) = \big\{(\tilde{x},\tilde{t}) \in \mathbb{R}^{n+1} : |\tilde{x} - K^{-1}(x(S) - x(P))|_\infty \leqslant \tilde{K}^2/2 \text{ and } |\tilde{t} - K^{-2}t(P)| \leqslant \tilde{K}^2/2\big\};$$

see Figure 6.

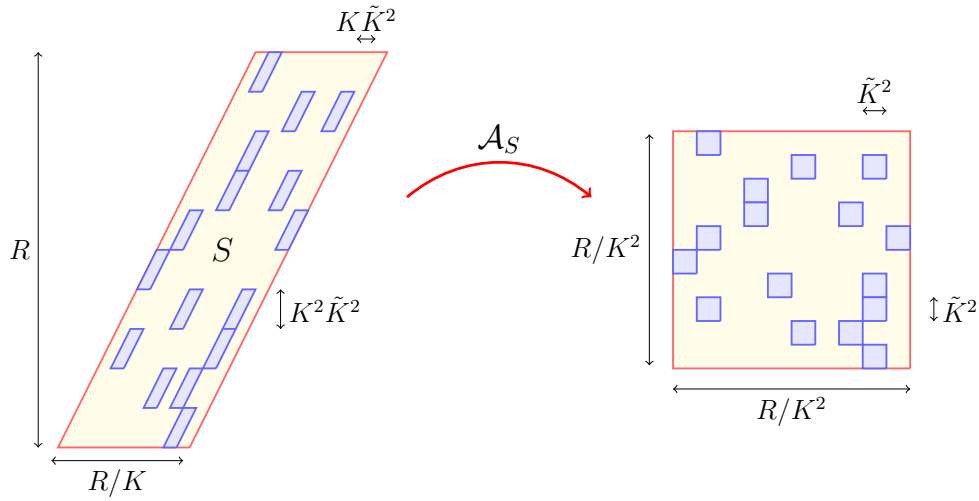


Figure 6: The map $\mathscr{A}_S$ sends S to $B^{n+1}(0, R/K^2)$ and each parallelepiped P to a lattice $\tilde{K}^2$-cube. Note that the left and right-hand sides are drawn at different scales (the right-hand ball is in fact much smaller than the strip).

Since we are forming our L^q-norms over the set $Z_{\mathscr{Q}}$, it suffices to only consider the subcollection $\mathscr{P}(S;\mathscr{Q})$ of all parallelepipeds $P \in \mathscr{P}(S)$ which intersect some

cube $Q \in \mathscr{Q}$. By the preceding discussion,

$$\tilde{\mathscr{Q}}(S) := \big\{ \mathscr{A}_S(P) : P \in \mathscr{P}(S;\mathscr{Q}) \big\} \tag{111}$$

is a collection of lattice $\tilde{K}^2$-cubes in $B^{n+1}(0,\tilde{R})$.

Since the set $S \cap Z_{\mathscr{Q}}$ is contained in

$$Z_{\mathscr{P}(S;\mathscr{Q})} := \bigcup_{P \in \mathscr{P}(S;\mathscr{Q})} P,$$

and by Lemma 4.16, the function Uf_S is essentially supported on S, we may estimate

$$\|Uf_S\|_{L^{q_n}(Z_{\mathscr{Q}})} \lesssim \|Uf_S\|_{L^{q_n}(Z_{\mathscr{P}(S;\mathscr{Q})})}. \tag{112}$$

which combines with (108) to give

$$\|Uf\|_{L^{q_n}(Z_{\mathscr{Q}})} \lesssim_\varepsilon K^\varepsilon [\max_{Q \in \mathscr{Q}} \#\mathbf{S}(Q)]^{1/(n+1)} \Big(\sum_{S \in \mathbf{S}} \|Uf_S\|^{q_n}_{L^{q_n}(Z_{\mathscr{P}(S;\mathscr{Q})})} \Big)^{1/q_n}. \tag{113}$$

In order to apply the induction hypothesis later in the argument, it is useful to make the following assumption.

Simplifying Assumption P1. For each $S \in \mathbf{S}$, we assume that

$$\|Uf_S\|_{L^{q_n}(P)} \qquad \text{are dyadically constant over } P \in \mathscr{P}(S;\mathscr{Q}). \tag{114}$$

For the purpose of this non-technical discussion, we shall make a number of such simplifying assumptions. Later, in §7.7, we give a rigorous justification of these assumptions using pigeonholing.

We now rescale the norm on the right-hand side of (112) using Lemma 4.18. In particular,

$$\|Uf_S\|_{L^{q_n}(Z_{\mathscr{P}(S;\mathscr{Q})})} \leqslant K^{(n+2)/q_n - n/2} \|U\tilde{f}_S\|_{L^{q_n}(Z_{\tilde{\mathscr{Q}}(S)})} = K^{-1/(n+1)} \|U\tilde{f}_S\|_{L^{q_n}(Z_{\tilde{\mathscr{Q}}(S)})}$$

for some function $\tilde{f}_S \in L^2(\mathbb{R}^n)$ satisfying

$$\|\tilde{f}_S\|_{L^2(\mathbb{R}^n)} = \|f_S\|_{L^2(\mathbb{R}^n)} \quad \text{and} \quad \operatorname{supp} \mathscr{F}(\tilde{f}_S) \subseteq B^n(0,1).$$

Each localised norm $\|U\tilde{f}_S\|_{L^{q_n}(P)}$ rescales to some $K^{-1/(n+1)}\|U\tilde{f}_S\|_{L^{q_n}(\tilde{Q})}$. Thus,

$$\|U\tilde{f}_S\|_{L^{q_n}(\tilde{Q})} \qquad \text{are dyadically constant over } \tilde{Q} \in \tilde{\mathscr{Q}}(S),$$

as a consequence of (114). Therefore, we may apply the induction hypothesis to conclude that

$$\|Uf_S\|_{L^{q_n}(Z_{\mathscr{P}(S;\mathscr{Q})})} \lesssim \mathbf{C}_\varepsilon K^{-1/(n+1)} \Big[\frac{\Delta_\alpha(\tilde{\mathscr{Q}}(S))}{\#\tilde{\mathscr{Q}}(S)} \Big]^{1/(n+1)} \tilde{R}^{\alpha/(2(n+1))+\varepsilon} \|f\|_{L^2(\mathbb{R}^n)}.$$

Since $\tilde{R} = 20R/K^2$, this becomes

$$\|Uf_S\|_{L^{q_n}(Z_{\mathscr{P}(S;\mathscr{Q})})} \lesssim \mathbf{C}_\varepsilon K^{-2\varepsilon}\Big[K^{-1-\alpha}\frac{\Delta_\alpha(\tilde{\mathscr{Q}}(S))}{\#\tilde{\mathscr{Q}}(S)}\Big]^{1/(n+1)} R^{\alpha/(2(n+1))+\varepsilon}\|f\|_{L^2(\mathbb{R}^n)}. \tag{115}$$

The decay factor of $K^{-2\varepsilon}$ arising from the application of the induction hypothesis is crucial for closing the argument.

Summing the estimates. — The next step is to sum the localised estimate (115) over all choices of strip $S \in \mathbf{S}$. Substituting the estimate (115) into (113), we deduce that

$$\|Uf\|_{L^{q_n}(Z_{\mathscr{Q}})} \lesssim \mathbf{C}_\varepsilon K^{-\varepsilon} M_K(\mathscr{Q})^{1/(n+1)} R^{\alpha/(2(n+1))+\varepsilon}[\#\mathbf{S}]^{1/(n+1)}\Big(\sum_{S\in\mathbf{S}}\|f_S\|_{L^2(\mathbb{R}^n)}^{q_n}\Big)^{1/q_n} \tag{116}$$

where

$$M_K(\mathscr{Q}) := K^{-1-\alpha}\Big[\max_{Q\in\mathscr{Q}}\frac{\#\mathbf{S}(Q)}{\#\mathbf{S}}\Big]\Big[\max_{S\in\mathbf{S}}\frac{\Delta_\alpha(\tilde{\mathscr{Q}}(S))}{\#\tilde{\mathscr{Q}}(S)}\Big].$$

One way to estimate the ℓ^q sum on the right-hand side of (116) is to simply use the nesting of ℓ^q-norms and orthogonality between the wave packets:

$$\Big(\sum_{S\in\mathbf{S}}\|f_S\|_{L^2(\mathbb{R}^n)}^{q_n}\Big)^{1/q_n} \leqslant \Big(\sum_{S\in\mathbf{S}}\|f_S\|_{L^2(\mathbb{R}^n)}^{2}\Big)^{1/2} \lesssim \|f_S\|_{L^2(\mathbb{R}^{n+1})}. \tag{117}$$

However, we shall use a more nuanced bound, improving over the above. The idea is to use a reverse Hölder inequality, similar to that used in §7.4. Ideally,

$$\Big(\sum_{S\in\mathbf{S}}\|f_S\|_{L^2(\mathbb{R}^n)}^{q_n}\Big)^{1/q_n} \lesssim [\#\mathbf{S}]^{-1/(n+1)}\Big(\sum_{S\in\mathbf{S}}\|f_S\|_{L^2(\mathbb{R}^n)}^{2}\Big)^{1/2} \lesssim [\#\mathbf{S}]^{-1/(n+1)}\|f\|_{L^2(\mathbb{R}^n)}. \tag{118}$$

which gains an additional factor of $[\#\mathbf{S}]^{-1/(n+1)}$ over (117). This leads us to our next simplifying assumption.

Simplifying Assumption S2.

$$\|f_S\|_{L^2(\mathbb{R})} \qquad \text{are dyadically constant over } S \in \mathbf{S}. \tag{119}$$

Under this assumption, the desired reverse Hölder inequality in (117) holds.

Combining (116) and (118), we obtain

$$\|Uf\|_{L^{q_n}(Z_{\mathscr{Q}})} \lesssim \mathbf{C}_\varepsilon K^{-\varepsilon} M_K(\mathscr{Q})^{1/(n+1)} R^{\alpha/(2(n+1))+\varepsilon}\|f\|_{L^2(\mathbb{R}^n)}. \tag{120}$$

It remains to estimate the $M_K(\mathscr{Q})$ factor.

Closing the induction. — In order to conclude the argument, we shall show

$$M_K(\mathscr{Q}) \lesssim \frac{\Delta_\alpha(\mathscr{Q})}{\#\mathscr{Q}}. \tag{121}$$

Indeed, once we have this bound, we can plug it into (120) and then choose $\mathbf{C}_\varepsilon$ and R_0 appropriately to close the induction and complete the proof.

The definition of $M_K(\mathscr{Q})$ involves two factors:

$$\Big[\max_{Q\in\mathscr{Q}} \frac{\#\mathbf{S}(Q)}{\#\mathbf{S}}\Big] \qquad \text{and} \qquad K^{-1-\alpha}\Big[\max_{S\in\mathbf{S}} \frac{\Delta_\alpha(\tilde{\mathscr{Q}}(S))}{\#\tilde{\mathscr{Q}}(S)}\Big]. \tag{122}$$

We shall split the proof of (121) into 3 steps: the first two steps shall treat the first factor in (122) and the remaining step treats the remaining.

1. Multiplicity bounds. Recall from the definitions that $\#\mathbf{S}(Q)/\#\mathbf{S}$ is the proportion of all strips $S \in \mathbf{S}$ which pass through Q. Hence, we refer to this quantity as the *multiplicity* of Q. On the other hand, if we define

$$\mathscr{Q}(S) := \{Q \in \mathscr{Q} : S \cap Q \neq \emptyset\}, \qquad S \in \mathbf{S},$$

then $\#\mathscr{Q}(S)/\#\mathscr{Q}$ is the proportion of all cubes $Q \in \mathscr{Q}$ which lie in a fixed strip S. We refer to this quantity as the *multiplicity* of S. Ideally, we would like to show

$$\max_{Q\in\mathscr{Q}} \frac{\#\mathbf{S}(Q)}{\#\mathbf{S}} \lesssim \max_{S\in\mathbf{S}} \frac{\#\mathscr{Q}(S)}{\#\mathscr{Q}}; \tag{123}$$

in other words, if there exists a high multiplicity cube, then there must exist a high multiplicity strip.

Example 7.5. Without further hypotheses, it is easy to see (123) may fail. For instance, suppose there are M strips in $\mathbf{S}$, there are $M+1$ cubes in $\mathscr{Q}$. We may arrange things so that:

- ▷ There exists precisely one cube $Q_0 \in \mathscr{Q}$ which lies in every strip: $\#\mathbf{S}(Q_0) = M$;
- ▷ Every other cube in $\mathscr{Q}$ lies in precisely one strip: $\#\mathbf{S}(Q) = 1$ for all $Q \in \mathscr{Q} \setminus \{Q_0\}$;
- ▷ Every strip contains 2 cubes: $\#\mathscr{Q}(S) = 2$ for all $S \in \mathbf{S}$.

Then

$$\max_{Q\in\mathscr{Q}} \frac{\#\mathbf{S}(Q)}{\#\mathbf{S}} = \frac{\#\mathbf{S}(Q_0)}{\#\mathbf{S}} = 1 \qquad \text{and} \qquad \max_{S\in\mathbf{S}} \frac{\#\mathscr{Q}(S)}{\#\mathscr{Q}} = \frac{2}{M+1}.$$

Thus, if $M \gg 1$, then this violates (123). The idea here, however, is that the violation arises from the single 'outlier' cube Q_0; if we throw away such outliers, and focus only on 'typical' cubes, then we can hope for (123) to hold.

By double-counting we always have a trivial inequality

$$\min_{Q\in\mathscr{Q}} \frac{\#\mathbf{S}(Q)}{\#\mathbf{S}} \leqslant \max_{S\in\mathbf{S}} \frac{\#\mathscr{Q}(S)}{\#\mathscr{Q}}. \tag{124}$$

Indeed, (124) immediately follows once we note

$$\sum_{Q\in\mathscr{Q}} \#\mathbf{S}(Q) = \#\{(S,Q) : S\in\mathbf{S}, Q\in\mathscr{Q}(S)\} = \sum_{S\in\mathbf{S}} \#\mathscr{Q}(S).$$

To exploit the bound (124), we make a further assumption on the multiplicities of the cubes.

Simplifying Assumption Q.

$$\#\mathbf{S}(Q) \qquad \text{are dyadically constant over } Q\in\mathscr{Q}. \tag{125}$$

This assumption ensures there are no 'outlier' cubes as in Example 7.5. In particular, using (125), we may upgrade (124) to the desired bound (123). This helps us to control the first factor in (122).

2. From strips to parallelepipeds. The next step is to relate $\#\mathscr{Q}(S)$ to $\#\tilde{\mathscr{Q}}(S)$. We begin by recalling some of the definitions. For each strip $S\in\mathbf{S}$ there is an associated family of parallelepipeds $\mathscr{P}(S;\mathscr{Q})$. Furthermore, the family of cubes $\tilde{\mathscr{Q}}(S)$ is obtained by transforming the $P\in\mathscr{P}(S;\mathscr{Q})$ under $\mathscr{A}_S$ (see (111)). In particular,

$$\#\mathscr{P}(S;\mathscr{Q}) = \#\tilde{\mathscr{Q}}(S). \tag{126}$$

Thus, our task here is to relate $\#\mathscr{Q}(S)$ to $\#\mathscr{P}(S;\mathscr{Q})$.

Let $\mathscr{P}(\mathscr{Q})$ denote the union of the $\mathscr{P}(S;\mathscr{Q})$ over all $S\in\mathbf{S}$ and let

$$\mathscr{Q}(P) := \{Q\in\mathscr{Q} : P\cap Q\neq\emptyset\} \qquad \text{for all } P\in\mathscr{P}(\mathscr{Q}).$$

It follows from the definitions that

$$\mathscr{Q}(S) = \bigcup_{P\in\mathscr{P}(S;\mathscr{Q})} \mathscr{Q}(P).$$

We can therefore compare the multiplicity of a strip to the multiplicities of the parallelepipeds via the simple inequality

$$\#\mathscr{Q}(S) \leqslant \sum_{P\in\mathscr{P}(S;\mathscr{Q})} \#\mathscr{Q}(P) \leqslant [\max_{P\in\mathscr{P}(S;\mathscr{Q})} \#\mathscr{Q}(P)][\#\mathscr{P}(S;\mathscr{Q})].$$

Thus, in view of (123) and (126), we have

$$\max_{Q\in\mathscr{Q}} \frac{\#\mathbf{S}(Q)}{\#\mathbf{S}} \leqslant [\max_{P\in\mathscr{P}(\mathscr{Q})} \#\mathscr{Q}(P)][\max_{S\in\mathbf{S}} \#\tilde{\mathscr{Q}}(S)].$$

The maxima on the right-hand side of the above inequality are awkward to bound. However, the situation is improved if we introduce the following additional assumptions.

Simplifying Assumption P2.

$$\#\mathscr{Q}(P) \qquad \text{are dyadically constant over all } P \in \mathscr{P}(\mathscr{Q}). \tag{127}$$

Simplifying Assumption S2.

$$\#\mathscr{P}(S;\mathscr{Q}) \qquad \text{are dyadically constant over } S \in \mathbf{S}.$$

These are (thankfully!) our final simplifying assumptions. By (127) and (119), it follows that

$$\Big[\max_{Q\in\mathscr{Q}} \frac{\#\mathbf{S}(Q)}{\#\mathbf{S}}\Big] \lesssim \frac{1}{\#\mathscr{Q}}[\min_{P\in\mathscr{P}(\mathscr{Q})} \#\mathscr{Q}(P)][\min_{S\in\mathbf{S}} \#\tilde{\mathscr{Q}}(S)]. \tag{128}$$

3. Comparing densities. The final step is to relate the densities $\Delta_\alpha(\tilde{\mathscr{Q}}(S))$ and $\Delta_\alpha(\mathscr{Q})$. This is achieved via the following simple lemma.

Lemma 7.6. *With the above setup, for any $S \in \mathbf{S}$ we have*

$$[\min_{P\in\mathscr{P}(\mathscr{Q})} \#\mathscr{Q}(P)]\Delta_\alpha(\tilde{\mathscr{Q}}(S)) \lesssim K^{1+\alpha}\Delta_\alpha(\mathscr{Q}). \tag{129}$$

Proof. Let $\tilde{B} \subseteq B^{n-1}(0,\tilde{R})$ be a ball of radius $r := \mathrm{rad}(\tilde{B}) \geqslant \tilde{K}^2$. Note that

$$\#\{\tilde{Q} \in \tilde{\mathscr{Q}}(S) : \tilde{Q} \subseteq \tilde{B}\} = \#\{P \in \mathscr{P}(S;\mathscr{Q}) : P \subseteq \mathscr{A}_S^{-1}(\tilde{B})\},$$

where we have used the definition of the family of cubes $\tilde{\mathscr{Q}}(S)$ from (111). Letting $T := \mathscr{A}_S^{-1}(\tilde{B})$, it follows that

$$[\min_{P\in\mathscr{P}(\mathscr{Q})} \#\mathscr{Q}(P)][\#\{\tilde{Q} \in \tilde{\mathscr{Q}}(S) : \tilde{Q} \subseteq \tilde{B}\}] \leqslant \#\{Q \in \mathscr{Q} : Q \subseteq T\}.$$

Since T is a strip of dimension $Kr \times \cdots \times Kr \times K^2 r$, it can be covered by $O(K)$ balls $B \subseteq B^{n+1}(0,R)$ of radius Kr. Furthermore, this cover can be chosen such that if $Q \in \mathscr{Q}$ satisfies $Q \subseteq T$, then $Q \subseteq B$ for some choice of ball in the cover. Thus,

$$\#\{Q \in \mathscr{Q} : Q \subseteq T\} \lesssim K[\max_{\substack{B\subseteq B^{n+1}(0,R)\\ \mathrm{rad}(B)=Kr}} \#\{Q \in \mathscr{Q} : Q \subseteq B\}] \leqslant K^{1+\alpha} r^\alpha \Delta_\alpha(\mathscr{Q}). \tag{130}$$

Combining (7.7) and (130), we deduce that

$$[\min_{P\in\mathscr{P}(\mathscr{Q})} \#\mathscr{Q}(P)]\Delta_\alpha(\tilde{\mathscr{Q}};\tilde{B}) \lesssim K^{1+\alpha}\Delta_\alpha(\mathscr{Q})$$

and the desired result follows by taking a supremum over all choices of $\tilde{B}$. □

Combining (128) and (129), we therefore have

$$\Big[\max_{Q\in\mathscr{Q}}\frac{\#\mathbf{S}(Q)}{\#\mathbf{S}}\Big]\Big[\max_{S\in\mathbf{S}}\frac{\Delta_\alpha(\tilde{\mathscr{Q}}(S))}{\#\tilde{\mathscr{Q}}(S)}\Big]\lesssim K^{1+\alpha}\frac{\Delta_\alpha(\mathscr{Q})}{\#\mathscr{Q}}.$$

Recalling the definition of $M_k(\mathscr{Q})$, this immediately implies the desired bound (121). Consequently, we have the narrow estimate

$$\|Uf\|_{L^{q_n}(Z_{\mathscr{Q}})}\lesssim \mathbf{C}_\varepsilon K^{-\varepsilon}\Big[\frac{\Delta_\alpha(\mathscr{Q})}{\#\mathscr{Q}}\Big]^{1/(n+1)}R^{\alpha/(2(n+1))+\varepsilon}\|f\|_{L^2(\mathbb{R}^n)},$$

at least under the Simplifying Assumptions P1, P2, S1, S2 and Q introduced above.

7.8. Analysis of the narrow case: technical details

We now reexamine the argument from §7.7, including technical details. The main additional ingredient is a sequence of pigeonholing steps used to rigorously justify the various simplifying assumptions used above.

Define the scales $\tilde{R}$ and $\tilde{K}$ as in (109). For each $S\in\mathbf{S}$, we decompose each enlarged strip $\bar{S}$ into parallel parallelepipeds $\mathscr{P}(S)$ as in (110). In particular, each $P\in\mathscr{P}(S)$ is aligned parallel to S and has dimensions $K\tilde{K}^2\times\cdots\times K\tilde{K}^2\times K^2\tilde{K}^2$. By Lemma 4.16, we have the pointwise bound

$$|Uf_S(z)|\lesssim\sum_{P\in\mathscr{P}(S)}|Uf_S(z)|\chi_P(z)+R^{-100n}\|f\|_{L^2(\mathbb{R}^n)}. \tag{131}$$

We let $\mathscr{P}_{\text{all}}$ denote the disjoint union of the sets $\mathscr{P}(S)$ over all $S\in\mathbf{S}$. Thus, given a parallelepiped $P\in\mathscr{P}_{\text{all}}$, there exists a unique element $S\in\mathbf{S}$ such that $P\in\mathscr{P}(S)$, which we denote by $S(P)$.

Pigeonholing. — We first pigeonhole in the parallelepipeds $P\in\mathscr{P}_{\text{all}}$. To this end, define

$$\mathscr{P}_{\text{tiny}}:=\big\{P\in\mathscr{P}:\|Uf_{S(P)}\|_{L^{q_n}(P)}\leqslant R^{-100n}\|f\|_{L^2(\mathbb{R}^n)}\big\}.$$

Parallelepipeds $P\in\mathscr{P}_{\text{tiny}}$ are negligible for our purpose: for a formal interpretation of this see (133) below. On the other hand, if $P\in\mathscr{P}_{\text{all}}\setminus\mathscr{P}_{\text{tiny}}$, then a trivial estimate using the Cauchy–Schwarz inequality and Plancherel's theorem shows that

$$R^{-100n}\|f\|_{L^2(\mathbb{R}^n)}\leqslant\|Uf_{S(P)}\|_{L^{q_n}(P)}\lesssim|P|^{1/q}\|f\|_{L^2(\mathbb{R}^n)}\lesssim R\|f\|_{L^2(\mathbb{R}^n)}.$$

This will allow us to dyadically pigeonhole in $\|Uf_{S(P)}\|_{L^{q_n}(P)}$. We shall also pigeonhole in the cardinality of the sets

$$\mathscr{Q}(P):=\big\{Q\in\mathscr{Q}:P\cap Q^{(\delta)}\neq\emptyset\big\},\quad\text{where}\quad Q^{(\delta)}:=K^\delta\cdot Q\quad\text{for } Q\in\mathscr{Q};$$

that is, the number of K^2-cubes $Q\in\mathscr{Q}$ which lie in the vicinity of a given P.

Pigeonholing the parallelepipeds. The family of parallelepipeds $\mathscr{P}_{\text{all}} \setminus \mathscr{P}_{\text{tiny}}$ can be written as a disjoint union $\mathscr{P}_{\text{all}} \setminus \mathscr{P}_{\text{tiny}} = \mathscr{P}_1 \cup \cdots \cup \mathscr{P}_I$ where

$$\|Uf_{S(P)}\|_{L^{q_n}(P)} \qquad \text{are dyadically constant over } P \in \mathscr{P}_i \tag{132}$$

and

$$\#\mathscr{Q}(P) \qquad \text{are dyadically constant over } P \in \mathscr{P}_i$$

for each $1 \leqslant i \leqslant I$ and $I \lessapprox 1$. This corresponds to Simplifying Assumptions P1 and P2 in §7.7.

We now turn to pigeonholing the strips $S \in \mathbf{S}$. To this end, first define

$$\mathbf{S}_{\text{tiny}} := \big\{S \in \mathbf{S} : \|f_S\|_{L^2(\mathbb{R}^n)} \leqslant R^{-100n}\|f\|_{L^2(\mathbb{R}^n)}\big\}.$$

The strips $S \in \mathbf{S}_{\text{tiny}}$ are negligible for our purpose. More precisely, let $\mathscr{Q}_{\text{tiny}}$ denote the set of cubes $Q \in \mathscr{Q}$ such that

$$\Big(\sum_{S\in\mathbf{S}} \|Uf_S\|^2_{L^{q_n}(w_Q)}\Big)^{1/2} \leqslant 2\Big(\sum_{S\in\mathbf{S}_{\text{tiny}}} \|Uf_S\|^2_{L^{q_n}(w_Q)}\Big)^{1/2}.$$

If $\#\mathscr{Q}_{\text{tiny}} \geqslant \#\mathscr{Q}/2$, then (107) and the dyadically constant hypothesis (91) immediately yield a very favourable estimate. Henceforth, we assume $\mathscr{Q}_{\text{main}} := \mathscr{Q} \setminus \mathscr{Q}_{\text{tiny}}$ satisfies $\#\mathscr{Q}_{\text{main}} \geqslant \#\mathscr{Q}/2$.

If $S \in \mathbf{S} \setminus \mathbf{S}_{\text{tiny}}$, then the orthogonality properties of the wave packets (see Lemma 4.15) imply that

$$R^{-100n}\|f\|_{L^2(\mathbb{R}^n)} \leqslant \|f_S\|_{L^2(\mathbb{R}^n)} \lesssim \|f\|_{L^2(\mathbb{R}^n)}.$$

This will allow us to dyadically pigeonhole in $\|f_S\|_{L^2(\mathbb{R}^n)}$. We shall also pigeonhole in the cardinalities of the sets

$$\mathscr{P}_i(S) := \mathscr{P}(S) \cap \mathscr{P}_i \qquad \text{for } S \in \mathbf{S} \text{ and } 1 \leqslant i \leqslant I.$$

Pigeonholing the strips. For each $1 \leqslant i \leqslant I$, the family of strips $\mathbf{S} \setminus \mathbf{S}_{\text{tiny}}$ can be written as a disjoint union $\mathbf{S} \setminus \mathbf{S}_{\text{tiny}} = \mathbf{S}_{i,1} \cup \cdots \cup \mathbf{S}_{i,J_i}$ where

$$\|f_S\|_{L^2(\mathbb{R}^n)} \qquad \text{are dyadically constant over } S \in \mathbf{S}_{i,j}$$

and

$$\#\mathscr{P}_i(S) \qquad \text{are dyadically constant over } S \in \mathbf{S}_{i,j}$$

and $J_i \lessapprox 1$ for all $1 \leqslant i \leqslant I$, $1 \leqslant j \leqslant J_i$. This corresponds to Simplifying Assumptions S1 and S2 in §7.7.

Finally, we pigeonhole in the cubes $Q \in \mathscr{Q}$. This step is a little more involved and we require a number of preliminary observations.

Given $Q \in \mathscr{Q}$ and any set $\mathscr{P} \subseteq \mathscr{P}_{\text{all}}$, define

$$w_Q|_{\mathscr{P}} := w_Q \cdot \sum_{P \in \mathscr{P}} \chi_P.$$

Fix $Q \in \mathscr{Q}_{\text{main}}$. By the pointwise bound (131) and nesting of ℓ^q norms, we have

$$\|Uf_S\|_{L^{q_n}(w_Q)} \lesssim \Big(\sum_{i=1}^{I} \|Uf_S\|^2_{L^{q_n}(w_Q|_{\mathscr{P}_i(S)})} \Big)^{1/2} + R^{-10n} \|f\|_{L^2(\mathbb{R}^n)}. \tag{133}$$

Here the contribution from the parallelepipeds $P \in \mathscr{P}_{\text{tiny}}$ is absorbed into the rapidly decaying term. Taking the ℓ^2-norm over $S \in \mathbf{S}$, we have

$$\Big(\sum_{S \in \mathbf{S}} \|Uf_S\|^2_{L^{q_n}(w_Q)} \Big)^{1/2} \lesssim \Big(\sum_{S \in \mathbf{S} \setminus \mathbf{S}_{\text{tiny}}} \sum_{i=1}^{I} \|Uf_S\|^2_{L^{q_n}(w_Q|_{\mathscr{P}_i(S)})} \Big)^{1/2} + R^{-5n} \|f\|_{L^2(\mathbb{R}^n)};$$

note that, since $Q \in \mathscr{Q}_{\text{main}}$, the contribution from the strips $S \in \mathbf{S}_{\text{tiny}}$ is negligible. By reordering the right-hand sum in the above expression, we have

$$\sum_{S \in \mathbf{S} \setminus \mathbf{S}_{\text{tiny}}} \sum_{i=1}^{I} \|Uf_S\|^2_{L^{q_n}(w_Q|_{\mathscr{P}_i(S)})} = \sum_{i=1}^{I} \sum_{j=1}^{J_i} \sum_{S \in \mathbf{S}_{i,j}} \|Uf_S\|^2_{L^{q_n}(w_Q|_{\mathscr{P}_i(S)})}.$$

Finally, we combine the above observations with (107) to deduce that

$$\|Uf\|_{L^{q_n}(Q)} \lesssim_\varepsilon K^\varepsilon \Big(\sum_{i=1}^{I} \sum_{j=1}^{J_i} \sum_{S \in \mathbf{S}_{i,j}} \|Uf_S\|^2_{L^{q_n}(w_Q|_{\mathscr{P}_i(S)})} \Big)^{1/2} + R^{-5n} \|f\|_{L^2(\mathbb{R}^n)}. \tag{134}$$

We now turn to pigeonholing the cubes $Q \in \mathscr{Q}$, which involves a two-step process.

Pigeonholing the cubes. Let $Q \in \mathscr{Q}_{\text{main}}$. Applying pigeonholing to (134), there exists some index pair (i_Q, j_Q) such that $\mathbf{S}_Q := \mathbf{S}_{i_Q, j_Q}$ and $\mathscr{P}_Q(S) := \mathscr{P}_{i_Q}(S)$ satisfy

$$\|Uf\|_{L^{q_n}(Q)} \lessapprox K^\varepsilon \Big(\sum_{S \in \mathbf{S}_Q} \|Uf_S\|^2_{L^{q_n}(w_Q|_{\mathscr{P}_Q(S)})} \Big)^{1/2} + R^{-5n} \|f\|_{L^2(\mathbb{R}^n)}.$$

Furthermore, by pigeonholing, there exists some refinement $\mathscr{Q}'_0 \subseteq \mathscr{Q}_{\text{main}}$ and index (i_0, j_0) such that $\mathbf{S}' := \mathbf{S}_{i_0, j_0}$ and $\mathscr{P}' := \mathscr{P}_{i_0}$ satisfy

$$\mathbf{S}_Q = \mathbf{S}' \quad \text{for all } Q \in \mathscr{Q}'_0 \quad \text{and} \quad \#\mathscr{Q}'_0 \gtrsim \#\mathscr{Q}$$

Given $Q \in \mathscr{Q}$, define

$$\mathbf{S}'(Q) := \big\{ S \in \mathbf{S}' : \bar{S} \cap Q^{(\delta)} \neq \emptyset \big\}$$

where, as above, $Q^{(\delta)} := K^{\delta} \cdot Q$ for $Q \in \mathscr{Q}$. By a second application of the pigeonhole principle, we can find a further refinement $\mathscr{Q}' \subseteq \mathscr{Q}'_0$ such that

$$\#\mathbf{S}'(Q) \quad \text{are dyadically constant over } Q \in \mathscr{Q}'. \tag{135}$$

This corresponds to Simplifying Assumption Q in §7.7.

If $Q \in \mathscr{Q}'$, then it follows from the above construction and definitions that

$$\|Uf\|_{L^{q_n}(Q)} \lessapprox K^{\varepsilon}\Big(\sum_{S\in\mathbf{S}'(Q)} \|Uf_S\|^2_{L^{q_n}(w_Q|_{\mathscr{P}'(S)})}\Big)^{1/2} + R^{-5n}\|f\|_{L^2(\mathbb{R}^n)}.$$

In order to sum in Q, we apply Hölder's inequality to pass from an ℓ^2 to an ℓ^q norm. Furthermore, since the weight w_Q is rapidly decaying away from Q, we may pass from $w_Q|_{\mathscr{P}'(S)}$ to the weight $w_Q|_{\mathscr{P}'(S;\mathscr{Q}')}$ supported on the parallelepipeds

$$\mathscr{P}'(S;\mathscr{Q}') := \big\{P \in \mathscr{P}'(S) : P \cap Q^{(\delta)} \neq \emptyset \text{ for some } Q \in \mathscr{Q}'\big\}.$$

In particular,

$$\|Uf\|_{L^{q_n}(Q)} \lessapprox K^{\varepsilon}[\#\mathbf{S}'(Q)]^{1/(n+1)}\Big(\sum_{S\in\mathbf{S}'(Q)} \|Uf_S\|^{q_n}_{L^{q_n}(w_Q|_{\mathscr{P}'(S;\mathscr{Q}')})}\Big)^{1/q_n} + R^{-5n}\|f\|_{L^2(\mathbb{R}^n)}.$$

Taking q_n-powers and summing over all the cubes in the refined collection,

$$\|Uf\|_{L^{q_n}(Z_{\mathscr{Q}})} \lessapprox K^{\varepsilon}[\min_{Q\in\mathscr{Q}'} \#\mathbf{S}'(Q)]^{1/(n+1)}\Big(\sum_{S\in\mathbf{S}'} \|Uf_S\|^{q_n}_{L^{q_n}(Z_{\mathscr{P}'(S;\mathscr{Q}')})}\Big)^{1/q_n} + \|f\|_{L^2(\mathbb{R}^n)} \tag{136}$$

where $Z_{\mathscr{P}'(S;\mathscr{Q}')}$ denotes the union of the $P \in \mathscr{P}'(S;\mathscr{Q}')$. Here we have used the dyadically constant hypothesis (91) and (135).

Parabolic rescaling and the induction hypothesis. — Fix $S \in \mathbf{S}'$. We applying parabolic rescaling from Lemma 4.18 to deduce that

$$\|Uf_S\|_{L^{q_n}(Z_{\mathscr{P}'(S;\mathscr{Q}')})} \leqslant K^{-1/(n+1)}\|U\tilde{f}_S\|_{L^{q_n}(Z_{\tilde{\mathscr{Q}}(S)})}$$

where $\tilde{f}_S \in L^2(\mathbb{R}^n)$ satisfies

$$\|\tilde{f}_S\|_{L^2(\mathbb{R}^n)} = \|f_S\|_{L^2(\mathbb{R}^n)} \quad \text{and} \quad \operatorname{supp}\mathscr{F}(\tilde{f}_S) \subseteq B^n(0,1).$$

Here $\tilde{\mathscr{Q}}(S)$ is a collection of lattice $\tilde{K}^2$-cubes contained in $B^{n+1}(0,\tilde{R})$. Furthermore, for each $\tilde{Q} \in \tilde{\mathscr{Q}}(S)$ there exists some $P \in \mathscr{P}'(S;\mathscr{Q}')$ such that

$$\|Uf_S\|_{L^{q_n}(P)} = K^{-1/(n+1)}\|U\tilde{f}_S\|_{L^{q_n}(\tilde{Q})};$$

it therefore follows from (132) that

$$\|U\tilde{f}_S\|_{L^{q_n}(\tilde{Q})} \qquad \text{are dyadically constant over } \tilde{Q} \in \tilde{\mathscr{Q}}(S).$$

In light of the above, we may apply the induction hypothesis to conclude that

$$\|Uf_S\|_{L^{q_n}(Z_{\mathscr{P}'(S;\mathscr{Q}')})} \lesssim \mathbf{C}_\varepsilon K^{-2\varepsilon}\Big[K^{-1-\alpha}\frac{\Delta_\alpha(\tilde{\mathscr{Q}}(S))}{\#\tilde{\mathscr{Q}}(S)}\Big]^{1/(n+1)} R^{\alpha/(2(n+1))+\varepsilon}\|f_S\|_{L^2(\mathbb{R}^n)}. \tag{137}$$

This is, of course, the analogue of the estimate (115) from the non-technical argument in §7.7. The key difference here is that the family of cubes $\tilde{\mathscr{Q}}(S)$ is formed from the refined set of parallelepipeds $\mathscr{P}'(S;\mathscr{Q}')$; this will allow us to (rigorously) exploit the various dyadically constancy properties.

Summing the estimates. — As in §7.7, the next step is to sum the localised estimate (137) over all choices of strip $S \in \mathbf{S}'$. Substituting (137) into (136), we deduce that

$$\|Uf\|_{L^{q_n}(Z_{\mathscr{Q}})} \lessapprox \mathbf{C}_\varepsilon K^{-\varepsilon} M'_K(\mathscr{Q}')^{1/(n+1)} R^{\alpha/(2(n+1))+\varepsilon}[\#\mathbf{S}']^{1/(n+1)}\Big(\sum_{S\in\mathbf{S}'}\|f_S\|^{q_n}_{L^2(\mathbb{R}^n)}\Big)^{1/q_n}, \tag{138}$$

where

$$M'_K(\mathscr{Q}') := K^{-1-\alpha}\Big[\max_{Q\in\mathscr{Q}'}\frac{\#\mathbf{S}'(Q)}{\#\mathbf{S}'}\Big]\Big[\max_{S\in\mathbf{S}'}\frac{\Delta_\alpha(\tilde{\mathscr{Q}}(S))}{\#\tilde{\mathscr{Q}}(S)}\Big]. \tag{139}$$

is defined in a similar manner to the corresponding quantity in §7.7.

To estimate the ℓ^q sum on the right-hand side of (138), we use a reverse Hölder inequality as in (118). Indeed, reverse Hölder can now be applied rigorously without further assumptions, since the norms $\|f_S\|_{L^2(\mathbb{R}^n)}$ for $S \in \mathbf{S}'$ are dyadically constant by construction. Consequently,

$$\|Uf\|_{L^{q_n}(Z_{\mathscr{Q}})} \lesssim \mathbf{C}_\varepsilon K^{-\varepsilon} M'_K(\mathscr{Q}')^{1/(n+1)} R^{\alpha/(2(n+1))+\varepsilon}\|f\|_{L^2(\mathbb{R}^n)}.$$

Multiplicity bounds. — The final step of the narrow analysis is to control the constant $M'_K(\mathscr{Q}')$ and, in particular, show that

$$M'_K(\mathscr{Q}') \lessapprox K^{O(\delta)}\frac{\Delta_\alpha(\mathscr{Q})}{\#\mathscr{Q}}. \tag{140}$$

Recall from our initial pigeonholing:

$\#\mathbf{S}'(Q)$	are dyadically constant over $Q \in \mathscr{Q}'$;
$\#\mathscr{P}'(S;\mathscr{Q}')$	are dyadically constant over $S \in \mathbf{S}'$;
$\#\mathscr{Q}'(P)$	are dyadically constant over $P \in \mathscr{P}'$.

Arguing exactly as in §7.7, we therefore deduce that

$$\Big[\max_{Q\in\mathscr{Q}'}\frac{\#\mathbf{S}'(Q)}{\#\mathbf{S}'}\Big]\Big[\max_{S\in\mathbf{S}'}\frac{\Delta_\alpha(\tilde{\mathscr{Q}}(S))}{\#\tilde{\mathscr{Q}}(S)}\Big]\lesssim\frac{1}{\#\mathscr{Q}'}\big[\min_{P\in\mathscr{P}'(\mathscr{Q}')}\#\mathscr{Q}'(P)\big]\big[\max_{S\in\mathbf{S}'}\Delta_\alpha(\tilde{\mathscr{Q}}(S))\big]. \quad (141)$$

As before, we can use Lemma 7.6 to compare the densities, giving

$$\big[\min_{P\in\mathscr{P}'(\mathscr{Q}')}\#\mathscr{Q}'(P)\big]\big[\max_{S\in\mathbf{S}'}\Delta_\alpha(\tilde{\mathscr{Q}}(S))\big]\lesssim K^{O(\delta)}K^{1+\alpha}\Delta_\alpha(\mathscr{Q}'). \quad (142)$$

Note that here we lose a factor of $K^{O(\delta)}$ due to the fact that the set of parallelepipeds $\mathscr{P}(S;\mathscr{Q})$ is defined with the enlarged cubes $Q^{(\delta)}$.

Combining (141) and (142), we have

$$\Big[\max_{Q\in\mathscr{Q}'}\frac{\#\mathbf{S}'(Q)}{\#\mathbf{S}'}\Big]\Big[\max_{S\in\mathbf{S}'}\frac{\Delta_\alpha(\tilde{\mathscr{Q}}(S))}{\#\tilde{\mathscr{Q}}(S)}\Big]\lessapprox K^{O(\delta)}K^{1+\alpha}\frac{\Delta_\alpha(\mathscr{Q})}{\#\mathscr{Q}};$$

here we have also used the fact that $\#\mathscr{Q}'\gtrsim\#\mathscr{Q}$ and $\Delta_\alpha(\mathscr{Q}')\leqslant\Delta_\alpha(\mathscr{Q})$. Recalling the definition of $M_K'(\mathscr{Q}')$ from (139), this immediately implies the desired bound (140). Consequently, we have the narrow estimate

$$\|Uf\|_{L^{q_n}(Z_{\mathscr{Q}})}\lesssim_\varepsilon \mathbf{C}_\varepsilon K^{-\varepsilon/2}\Big[\frac{\Delta_\alpha(\mathscr{Q})}{\#\mathscr{Q}}\Big]^{1/(n+1)}R^{\alpha/(2(n+1))+\varepsilon}\|f\|_{L^2(\mathbb{R}^n)}, \quad (143)$$

which essentially agrees with the bound derived in §7.7.

7.9. Concluding the argument

To conclude the proof, it remains to collect together the estimates proved above and show that they can be used to close the induction. Recall:

- ▷ In §7.4, we showed that (105) holds in the broad-dominant case;
- ▷ In §§7.6-7.8, we showed that (143) holds in the narrow-dominant case.

Combining (105) and (143), we see that there exists a constant $C_\varepsilon\geqslant 1$ such that the bound

$$\|Uf\|_{L^{q_n}(Z_{\mathscr{Q}})}\leqslant C_\varepsilon\big(\mathbf{C}_\varepsilon K^{-\varepsilon/2}+1\big)K^{-\varepsilon/2}\Big[\frac{\Delta_\alpha(\mathscr{Q})}{\#\mathscr{Q}}\Big]^{1/(n+1)}R^{\alpha/(2(n+1))+\varepsilon}\|f\|_{L^2(\mathbb{R}^n)}. \quad (144)$$

holds in either case.

The estimate (144) involves two free parameters:

- ▷ We are free to choose the cutoff R_0 for the base case;
- ▷ We are free to choose the constant $\mathbf{C}_\varepsilon$ appearing in the induction hypothesis, provided our choice is independent of R.[30]

At this point, we fine tune these parameters to ensure that the induction closes. Recalling that $K = R^\delta$, we choose R_0 large enough so that $C_\varepsilon K^{-\varepsilon/2} \leqslant 1/2$ whenever $R \geqslant R_0$ and take $\mathbf{C}_\varepsilon = 2C_\varepsilon$. With these parameters, (144) implies

$$\|Uf\|_{L^{q_n}(Z_{\mathscr{Q}})} \leqslant \mathbf{C}_\varepsilon \Big[\frac{\Delta_\alpha(\mathscr{Q})}{\#\mathscr{Q}}\Big]^{1/(n+1)} R^{\alpha/(2(n+1))+\varepsilon} \|f\|_{L^2(\mathbb{R}^n)},$$

which is precisely what is required to close the induction.

References

Bennett, J. (2014). "Aspects of multilinear harmonic analysis related to transversality", in: *Harmonic analysis and partial differential equations*. Vol. 612. Contemp. Math. Amer. Math. Soc., Providence, RI, pp. 1–28.

Bennett, J., Carbery, A., and Tao, T. (2006). "On the multilinear restriction and Kakeya conjectures", *Acta Math.* **196** (2), pp. 261–302.

Bourgain, J. (2013a). "Moment inequalities for trigonometric polynomials with spectrum in curved hypersurfaces", *Israel J. Math.* **193** (1), pp. 441–458.

——— (2013b). "On the Schrödinger maximal function in higher dimension", *Tr. Mat. Inst. Steklova* **280** (Ortogonal'nye Ryady, Teoriya Priblizhenii i Smezhnye Voprosy), pp. 53–66.

——— (2016). "A note on the Schrödinger maximal function", *J. Anal. Math.* **130**, pp. 393–396.

Bourgain, J. and Demeter, C. (2015). "The proof of the l^2 decoupling conjecture", *Ann. of Math.* (2) **182** (1), pp. 351–389.

Bourgain, J. and Guth, L. (2011). "Bounds on oscillatory integral operators based on multilinear estimates", *Geom. Funct. Anal.* **21** (6), pp. 1239–1295.

Carleson, L. (1980). "Some analytic problems related to statistical mechanics". In: *Euclidean harmonic analysis* (*Proc. Sem., Univ. Maryland, College Park, Md., 1979*). Vol. 779. Lecture Notes in Math. Springer, Berlin, pp. 5–45.

Dahlberg, B. E. J. and Kenig, C. E. (1982). "A note on the almost everywhere behavior of solutions to the Schrödinger equation", in: *Harmonic analysis* (*Minneapolis, Minn., 1981*). Vol. 908. Lecture Notes in Math. Springer, Berlin-New York, pp. 205–209.

[30] On a minor technical note, $\mathbf{C}_\varepsilon$ should be chosen large depending on R_0 to ensure the base case holds. In this sense, the two parameters are not entirely independent of one another.

Du, X., Guth, L., and Li, X. (2017). "A sharp Schrödinger maximal estimate in $\mathbb{R}^2$", *Ann. of Math.* (2) **186** (2), pp. 607–640.

Du, X., Guth, L., Li, X., and Zhang, R. (2018). "Pointwise convergence of Schrödinger solutions and multilinear refined Strichartz estimates", *Forum Math. Sigma* **6**, Paper No. e14, 18.

Du, X., Kim, J., et al. (2020). "Lower bounds for estimates of the Schrödinger maximal function", *Math. Res. Lett.* **27** (3), pp. 687–692.

Du, X. and Zhang, R. (2019). "Sharp L^2 estimates of the Schrödinger maximal function in higher dimensions", *Ann. of Math.* (2) **189** (3), pp. 837–861.

Fefferman, C. (1970). "Inequalities for strongly singular convolution operators", *Acta Math.* **124**, pp. 9–36.

Guth, L. (2015). "A short proof of the multilinear Kakeya inequality", *Math. Proc. Cambridge Philos. Soc.* **158** (1), pp. 147–153.

——— (2016). "A restriction estimate using polynomial partitioning", *J. Amer. Math. Soc.* **29** (2), pp. 371–413.

——— (2018). "Restriction estimates using polynomial partitioning II", *Acta Math.* **221** (1), pp. 81–142.

Guth, L., Iosevich, A., et al. (2020). "On Falconer's distance set problem in the plane", *Invent. math.* **219** (3), pp. 779–830.

Guth, L. and Katz, N. H. (2015). "On the Erdős distinct distances problem in the plane", *Ann. of Math.* (2) **181** (1), pp. 155–190.

Kenig, C. E., Ponce, G., and Vega, L. (1991). "Oscillatory integrals and regularity of dispersive equations", *Indiana Univ. Math. J.* **40** (1), pp. 33–69.

Lee, S. (2006). "On pointwise convergence of the solutions to Schrödinger equations in $\mathbb{R}^2$", *Int. Math. Res. Not.*, Art. ID 32597, 21.

Lucà, R. and Rogers, K. M. (2017). "Coherence on fractals versus pointwise convergence for the Schrödinger equation", *Comm. Math. Phys.* **351** (1), pp. 341–359.

——— (2019). "A note on pointwise convergence for the Schrödinger equation", *Math. Proc. Cambridge Philos. Soc.* **166** (2), pp. 209–218.

Pierce, L. B. (2020). "On Bourgain's counterexample for the Schrödinger maximal function", *Q. J. Math.* **71** (4), pp. 1309–1344.

Stein, E. M. (1986). "Oscillatory integrals in Fourier analysis", in: *Beijing lectures in harmonic analysis* (*Beijing, 1984*). Vol. 112. Ann. of Math. Stud. Princeton Univ. Press, Princeton, NJ, pp. 307–355.

Strichartz, R. S. (1977). "Restrictions of Fourier transforms to quadratic surfaces and decay of solutions of wave equations", *Duke Math. J.* **44** (3), pp. 705–714.

Tao, T., Vargas, A., and Vega, L. (1998). "A bilinear approach to the restriction and Kakeya conjectures", *J. Amer. Math. Soc.* **11** (4), pp. 967–1000.

Tomas, P. A. (1975). "A restriction theorem for the Fourier transform", *Bull. Amer. Math. Soc.* **81**, pp. 477–478.

Wang, H. and Wu, S. (2022). "An improved restriction estimate in $\mathbb{R}^3$". Preprint: arXiv:2210.03878.

Wolff, T. (2001). "A sharp bilinear cone restriction estimate", *Ann. of Math.* (2) **153** (3), pp. 661–698.

Jonathan Hickman

School of Mathematics,
James Clerk Maxwell Building,
The King's Buildings,
Peter Guthrie Tait Road,
Edinburgh,
EH9 3FD, UK.

E-mail:
jonathan.hickman@ed.ac.uk

Séminaire BOURBAKI Avril 2023
74e année, 2022–2023, no 1206, p. 365 à 382
doi: 10.24033/ast.1215

EXPONENTIAL GROWTH RATES IN HYPERBOLIC GROUPS
[*after* Koji Fujiwara and Zlil Sela]

by **Clara Löh**

A classical result of Jørgensen and Thurston shows that the set of volumes of finite volume complete hyperbolic 3-manifolds is a well-ordered subset of the real numbers of order type ω^ω; moreover, each volume can only be attained by finitely many isometry types of hyperbolic 3-manifolds.

Fujiwara and Sela (2020) established a group-theoretic companion of this result: If Γ is a non-elementary hyperbolic group, then the set of exponential growth rates of Γ is well-ordered, the order type is at least ω^ω, and each growth rate can only be attained by finitely many finite generating sets (up to automorphisms).

In this talk, we outline this work of Fujiwara and Sela and discuss related results.

1. Main results

Geometric group theory provides a rich interaction between the Riemannian geometry of manifolds and the large-scale geometry of finitely generated groups. This bond is particularly strong in the presence of negative curvature and explains a variety of rigidity phenomena. The group-theoretic analogues of closed hyperbolic manifolds are hyperbolic groups; more generally, the group-theoretic analogues of finite volume complete hyperbolic manifolds are relatively hyperbolic groups.

The volume growth behaviour of Riemannian balls in the universal covering of a compact Riemannian manifold is the same as the growth behaviour of balls in Cayley graphs of the fundamental group. By definition, the exponential growth rates of finitely generated groups measure the exponential expansion rate of balls in Cayley graphs and thus are entropy-like invariants. While there is no direct connection between the volume of a hyperbolic manifold M and the exponential growth rates of $\pi_1(M)$, the results of Fujiwara and Sela (2020) show that certain sets of such values share fundamental structural similarities.

To state these results, for a finitely generated group Γ, we write $\mathrm{Exp}(\Gamma) \subset \mathbb{R}$ for the (countable) set of all exponential growth rates $e(\Gamma, S)$ with respect to finite generating

sets S of Γ. The automorphism group $\mathrm{Aut}(\Gamma)$ acts on the set $\mathrm{FG}(\Gamma)$ of all finite generating sets of Γ and $\mathrm{e}(\Gamma, f(S)) = \mathrm{e}(\Gamma, S)$ holds for all $S \in \mathrm{FG}(\Gamma)$ and all $f \in \mathrm{Aut}(\Gamma)$. More details on terminology and notation can be found in Appendix A.

Theorem 1.1 (well-orderedness; Fujiwara and Sela, 2020, Theorem 2.2)**.** *If Γ is a hyperbolic group, then $\mathrm{Exp}(\Gamma)$ is well-ordered (with respect to the standard order on $\mathbb{R}$).*

Theorem 1.2 (finite ambiguity; Fujiwara and Sela, 2020, Theorem 3.1)**.** *The set $\{S \in \mathrm{FG}(\Gamma) \mid \mathrm{e}(\Gamma, S) = r\} / \mathrm{Aut}(\Gamma)$ is finite for every non-elementary hyperbolic group Γ and every $r \in \mathbb{R}$.*

Theorem 1.3 (growth ordinals; Fujiwara and Sela, 2020, Proposition 4.3)**.** *Let Γ be a non-elementary hyperbolic group. Then the ordinal number $\mathrm{ord}_{\mathrm{Exp}}(\Gamma)$ associated with $\mathrm{Exp}(\Gamma)$ satisfies $\mathrm{ord}_{\mathrm{Exp}}(\Gamma) \geqslant \omega^\omega$.*

Moreover, Fujiwara and Sela (2020, Proposition 4.3) show that $\mathrm{ord}_{\mathrm{Exp}}(\Gamma) = \omega^\omega$ if epi-limit groups over Γ have a Krull dimension. In analogy with the case of hyperbolic 3-manifolds, they conjecture that $\mathrm{ord}_{\mathrm{Exp}}(\Gamma) = \omega^\omega$ holds for all non-elementary hyperbolic groups Γ (Fujiwara and Sela, 2020, Section 4) .

Example 1.4. If F is a finitely generated free group of rank at least 2, then limit groups over F have a Krull dimension (Louder, 2012). Hence, Theorems 1.1–1.3 show that $\mathrm{ord}_{\mathrm{Exp}}(F) = \omega^\omega$ and each value in $\mathrm{Exp}(F)$ is realised by only finitely many generating sets (up to automorphisms of F).

The key idea for the proofs of Theorems 1.1–1.3 is inspired by the proofs by Thurston and Jørgensen for the set of volumes of hyperbolic 3-manifolds and model theory: One passes from sequences of generating sets (of bounded size) of the given hyperbolic group Γ to a limit group over Γ with an associated finite generating set; *i.e.*, limit groups play the role of cusped manifolds. The main challenge is then to compute the exponential growth rate of this limiting object in terms of the exponential growth rates appearing in the original sequence.

Overview

Basics on hyperbolic groups, exponential growth rates, and well-ordered countable sets are recalled in Appendix A. We briefly explain the manifold context of the results above in Section 2, with a focus on hyperbolic and simplicial volume. Section 3 gives a proof outline of the main results. Finally, in Section 4, we mention applications and extensions of the main results.

2. Context: Volumes of manifolds and hyperbolicity

The results of Fujiwara and Sela (2020) are analogues of the behaviour of volumes of finite volume complete hyperbolic 3-manifolds. We recall this background in Section 2.1. The situation for simplicial volume is discussed in Section 2.2. In addition, we mention right-computability as a further structural property of "volume" sets (Section 2.3).

2.1. Hyperbolic volume

The structure and volumes of hyperbolic 3-manifolds was analysed in the breakthrough work of Jørgensen and Thurston.

Theorem 2.1 (volumes of hyperbolic 3-manifolds; Thurston, 1979, Chapter 6). *The set*

$$\{\mathrm{vol}(M) \mid M \text{ is a finite volume complete hyperbolic 3-manifold}\}$$

is well-ordered (with respect to the standard order on $\mathbb{R}$*) and the associated ordinal is* ω^ω*. Moreover, every value arises only from finitely many isometry classes of finite volume hyperbolic 3-manifolds.*

We briefly summarise the main steps of the proof (Gromov, 1981); the key is to study the convergence of sequences of hyperbolic manifolds and to understand the role of hyperbolic manifolds with cusps as limits of such sequences:

1. Every sequence $(M_n)_{n\in\mathbb{N}}$ of complete hyperbolic 3-manifolds with uniformly bounded volume contains a subsequence that converges in a strong geometric sense to a finite volume complete hyperbolic 3-manifold M and $\lim_{n\to\infty} \mathrm{vol}(M_n) = \mathrm{vol}(M)$. Furthermore, for "non-trivial" such sequences, one can show that $\mathrm{vol}(M) > \mathrm{vol}(M_n)$ holds for all members M_n of the subsequence.

   This can be used to show that the set of hyperbolic volumes is well-ordered and that every value can only be obtained in finitely many ways.

2. Every finite volume complete hyperbolic 3-manifold with $k \in \mathbb{N}$ cusps can be obtained for each $p \in \{0, \dots, k\}$ as the limit of a sequence of finite volume complete hyperbolic 3-manifolds with exactly p cusps.

   This can be used to show that the volume ordinal is at least ω^k. Constructing hyperbolic 3-manifolds with arbitrarily large numbers of cusps thus shows that the volume ordinal is at least ω^ω. In combination with the first part, one can derive that the volume ordinal equals ω^ω.

In contrast, in higher dimensions, the set of volumes of finite volume complete hyperbolic manifolds leads to the ordinal ω. This follows from Wang's finiteness theorem and the unboundedness of hyperbolic volumes.

Theorem 2.2 (Wang's finiteness theorem; Wang, 1972). *Let $n \in \mathbb{N}_{\geqslant 4}$ and $v \in \mathbb{R}_{\geqslant 0}$. Then there exist only finitely many isometry classes of finite volume complete hyperbolic n-manifolds M with* $\mathrm{vol}(M) \leqslant v$.

2.2. Simplicial volume

Simplicial volume is a homotopy invariant of closed manifolds. For several geometrically relevant classes of Riemannian manifolds, the simplicial volume encodes topological rigidity properties of the Riemannian volume.

Definition 2.3 (simplicial volume; Gromov, 1982). The *simplicial volume* of an oriented closed connected manifold M is the ℓ^1-semi-norm of its (singular) $\mathbb{R}$-fundamental class:

$$\|M\| := \|[M]_{\mathbb{R}}\|_1 := \inf\Big\{ \sum_{j=1}^{k} |a_j| \;\Big|\; \sum_{j=1}^{k} a_j \cdot \sigma_j \text{ is a singular } \mathbb{R}\text{-fundamental cycle of } M \Big\}$$

For genuine hyperbolic manifolds, the simplicial volume leads to the same ordering and finiteness behaviour as the hyperbolic volume (Section 2.1):

Example 2.4 (hyperbolic manifolds). If M is an oriented closed connected hyperbolic manifold of dimension n, then

$$\|M\| = \frac{\mathrm{vol}(M)}{v_n},$$

where $v_n \in \mathbb{R}_{>0}$ is the hyperbolic volume of ideal regular geodesic n-simplices in hyperbolic n-space (Benedetti and Petronio, 1992; Thurston, 1979). A similar relationship also holds in the complete finite volume case (Thurston, 1979; Fujiwara and Manning, 2011, Appendix A). In particular, this proportionality can be used to prove mapping degree estimates in terms of the hyperbolic volume for continuous maps between hyperbolic manifolds (Gromov, 1982).

Passing to the setting of fixed hyperbolic fundamental groups, we obtain:

Example 2.5 (hyperbolic fundamental group). Let Γ be a finitely presented group and let $n \in \mathbb{N}$. Then the set

$$\mathrm{SV}_\Gamma(n) := \{\|M\| \mid M \text{ is an oriented closed connected } n\text{-manifold with } \pi_1(M) \cong \Gamma\}$$

is a subset of $\{\|\alpha\|_1 \mid \alpha \in H_n(\Gamma;\mathbb{R}) \text{ is integral}\}$, where a class in $H_n(\Gamma;\mathbb{R})$ is *integral* if it lies in the image of the change of coefficients map $H_n(\Gamma;\mathbb{Z}) \to H_n(\Gamma;\mathbb{R})$ (Löh, 2023, Section 3.1).

If Γ is hyperbolic and $n \geqslant 2$, then $\|\cdot\|_1$ is a norm on $H_n(\Gamma;\mathbb{R})$ (by the results of Mineyev (2001) on bounded cohomology and the duality principle). In particular: The set $\mathrm{SV}_\Gamma(n) \subset \mathbb{R}$ is well-ordered and for $n \geqslant 4$ the ordinal associated with $\mathrm{SV}_\Gamma(n)$ is

- either 0 (if $H_n(\Gamma;\mathbb{R}) \cong 0$);

▷ or ω (if $H_n(\Gamma;\mathbb{R}) \ncong 0$): In this case, normed Thom realisation shows that indeed infinitely many different values are realised (Löh, 2023, Section 3.1).

For $n \geqslant 4$, finite ambiguity breaks down in this general topological setting: If M is an oriented closed connected n-manifold, then for each $k \in \mathbb{N}$, the manifold M and the iterated connected sums $M_k := M \# (S^2 \times S^{n-2})^{\#k}$ have the same simplicial volume (Gromov, 1982) and isomorphic fundamental groups. However, the manifolds $M_0, M_1, \dots$ all have different homotopy types (as can be seen from the homology in degree 2).

2.3. Right-computability

In the previous discussion, we focussed on the order structure of volumes and exponential growth rates. Many real-valued invariants in geometric group theory and geometric topology also carry another, complementary, structure: They tend to have an intrinsic limit on their computational complexity. In particular, such a limit gives additional constraints on the possible sets of values.

Definition 2.6 (right-computable)**.** A real number α is *right-computable* if the set $\{x \in \mathbb{Q} \mid x > \alpha\}$ is recursively enumerable.

For example, simplicial volumes of oriented closed connected manifolds are right-computable real numbers (Heuer and Löh, 2023). On the group-theoretic side, right-computability naturally arises for stable commutator length of recursively presented groups (Heuer, 2019) or L^2-Betti numbers of groups with controlled word problem (Löh and Uschold, 2022). Concerning exponential growth rates, we have the following:

Proposition 2.7 (right-computability of exponential growth rates)**.** *There exists a Turing machine that*

▷ *given a finite presentation $\langle S \mid R\rangle$ and a finite set S' of words over $S \sqcup S^{-1}$,*

▷ *does*

- not *terminate if S' does not represent a generating set of the group Γ described by $\langle S \mid R\rangle$;*
- *terminate and return an enumeration of $\{x \in \mathbb{Q} \mid x > \mathrm{e}(\Gamma, S')\}$ if S' represents a generating set of Γ.*

Corollary 2.8. *Let Γ be a finitely presented group.*

1. *For every $S \in \mathrm{FG}(\Gamma)$, the real number $\mathrm{e}(\Gamma, S)$ is right-computable.*
2. *For every $r \in \mathbb{Q}$, the truncated set $\{S \in \mathrm{FG}(\Gamma) \mid \mathrm{e}(\Gamma, S) < r\}$ is recursively enumerable.*

Proofs of these observations are provided in Appendix B. In particular, such results could be used to give a crude a priori upper bound for $\mathrm{ord}_{\mathrm{Exp}}(\Gamma)$ by a "large" countable ordinal for all finitely presented groups Γ with well-ordered set $\mathrm{Exp}(\Gamma)$.

3. Proof technique

We outline the proofs of Theorem 1.1–1.3 by Fujiwara and Sela (2020). These proofs roughly follow the blueprint of the case of hyperbolic 3-manifolds (Section 2.1), where limit groups will play the role of cusped manifolds:

- ▷ A compactness phenomenon turns convergence of exponential growth rates into convergence of actions (of subsequences) to the action of a limit group.
- ▷ The main challenge is then to compute the exponential growth rates of these limit groups as the limit of the given exponential growth rates.

Before going into these arguments, we first recall basic notions on limit groups.

3.1. Limit groups

Limit groups are groups that arise as "limits" – in various senses – of groups. These groups are convenient tools in the model theory of groups and in geometric group theory (Groves and Wilton, 2018; Kharlampovich and Myasnikov, 1998a,b; Sela, 2006). In analogy with 3-manifolds, limit groups over hyperbolic groups admit a JSJ-decomposition (Sela, 2009; Weidmann and Reinfeldt, 2019, Section 4). Limit groups over a given group Γ are the finitely generated subgroups of non-principal ultraproducts of Γ. More explicitly:

Definition 3.1 (limit group). Let Γ be a group.

- ▷ A *stable homomorphism* from a group Λ to Γ is a sequence $(f_n \colon \Lambda \to \Gamma)_{n\in\mathbb{N}}$ of homomorphisms with the property

$$\forall_{x\in\Lambda} \quad \exists_{N\in\mathbb{N}} \quad (\forall_{n\in\mathbb{N}_{\geqslant N}} \quad f_n(x) = 1) \vee (\forall_{n\in\mathbb{N}_{\geqslant N}} \quad f_n(x) \neq 1).$$

  The *stable kernel* of a stable homomorphism $f_* \colon \Lambda \to \Gamma$ is defined as

$$\ker f_* := \left\{ x \in \Lambda \;\middle|\; \exists_{N\in\mathbb{N}} \quad \forall_{n\in\mathbb{N}_{\geqslant N}} \quad f_n(x) = 1 \right\} \subset \Lambda.$$

- ▷ A *limit group over* Γ is a group of the form $\Lambda/\ker f_*$, where Λ is a finitely generated group and $f_* \colon \Lambda \to \Gamma$ is a stable homomorphism. The canonical projection $f \colon \Lambda \to \Lambda/\ker f_*$ is the *limit homomorphism* and we say that f_* *converges to* f. A limit group over Γ is an *epi-limit group over* Γ if the f_n can be chosen to be epimorphisms.
- ▷ A *limit group* is a limit group over a finitely generated free group.

Example 3.2. If Γ is a group, then every finitely generated subgroup of Γ can be viewed as a limit group over Γ (via the inclusion homomorphisms). In particular, Γ is an epi-limit group over Γ if Γ is finitely generated.

3.2. Limits and their exponential growth rate

Theorem 3.3 (compactness; Fujiwara and Sela, 2020, proof of Theorem 2.2). *Let Γ be a hyperbolic group, let (X, d) be a Cayley graph of Γ, and let $(S_n)_{n\in\mathbb{N}}$ be a sequence of finite generating sets of Γ such that the sequence $(\mathrm{e}(\Gamma, S_n))_{n\in\mathbb{N}}$ is bounded. Then there exists a subsequence (again denoted by $(S_n)_{n\in\mathbb{N}}$) with the following properties:*

- *All S_n have the same size. Let S be a set of this cardinality and let F be the free group freely generated by S.*
- *There exist epimorphisms $f_n\colon F \to \Gamma$ for all $n \in \mathbb{N}$ such that $f_n(S)$ is conjugate to S_n. Moreover, f_* is a stable homomorphism $F \to \Gamma$. Let L denote the associated limit group.*
- *The sequence*
$$\left(F \curvearrowright_{f_n} \left(X, \frac{1}{\max_{s\in S} d(1, f_n(s))} \cdot d\right)\right)_{n\in\mathbb{N}}$$
of actions (induced by $f_n\colon F \to \Gamma$ and the translation action of Γ on X) converges in the F-Gromov–Hausdorff distance to a faithful action of L on a real tree.

Proof. The boundedness of the exponential growth rates allows us to fix the size of the generating set because the exponential growth rate grows at least linearly in the size of the generating set by an estimate of Arzhantseva and Lysenok (2006).

One can then apply the Bestvina–Paulin method after conjugating and rescaling appropriately (Fujiwara and Sela, 2020, proof of Theorem 2.2; Weidmann and Reinfeldt, 2019, Section 2). □

Theorem 3.4 (limits of exponential growth rates; Fujiwara and Sela, 2020, Proposition 2.3). *Let Γ be a hyperbolic group, let Λ be a finitely generated group with finite generating set S, and let $(f_n\colon \Lambda \to \Gamma)_{n\in\mathbb{N}}$ be a stable homomorphism consisting of epimorphisms that converges to a limit group $f\colon \Lambda \to L$ over Γ with a faithful action on a real tree. Then $\mathrm{e}(\Gamma, f_n(S)) \leqslant \mathrm{e}(L, f(S))$ holds for all large enough $n \in \mathbb{N}$ and*
$$\lim_{n\to\infty} \mathrm{e}(\Gamma, f_n(S)) = \mathrm{e}(L, f(S))$$

Sketch of proof. By precomposition, without loss of generality we may assume that Λ is free and that S a free generating set.

We first explain why "$\leqslant$" holds (provided the limit exists): Because Γ is hyperbolic, for all large enough $n \in \mathbb{N}$, there exists a homomorphism $h_n\colon L \to \Gamma$ with $f_n = h_n \circ f$ (Weidmann and Reinfeldt, 2019, Lemma 6.5, Corollary 7.13):

$$\begin{array}{ccc} (\Lambda, S) & & \\ {\scriptstyle f}\downarrow & \searrow^{f_n} & \\ (L, f(S)) & \xrightarrow[h_n]{} & (\Gamma, f_n(S)) \end{array}$$

Therefore, monotonicity of the exponential growth rates (Remark A.5) implies that $\mathrm{e}(\Gamma, f_n(S)) \leqslant \mathrm{e}(L, f(S))$ for all large enough $n \in \mathbb{N}$.

The hard work lies in proving convergence and "$\geqslant$": Given $\varepsilon \in \mathbb{R}_{>0}$, the goal is to show that for all large enough $n \in \mathbb{N}$, we have

$$\log \mathrm{e}(\Gamma, f_n(S)) \geqslant \log \mathrm{e}(L, f(S)) - \varepsilon.$$

Matters would be simple if, given $N \in \mathbb{N}$, the multiplication-projection map

$$\begin{aligned} B_N(L, f(S))^q &\to B_{q\cdot N}(\Gamma, f_n(S)) \\ (w_1, \dots, w_q) &\mapsto h_n(w_1 \cdot \dots \cdot w_q) \end{aligned}$$

were injective for all large enough $n \in \mathbb{N}$ and all $q \in \mathbb{N}$. However, this will not happen in general. Using the faithful limit action of L on a real tree, Fujiwara and Sela (2020, proof of Proposition 2.3) find enough freeness inside L to show through delicate estimates that there exists a $b \in \mathbb{N}$, a four-element subset $U \subset B_b(L, f(S))$, and a constant $C \in \mathbb{R}_{>0}$ with:

> For all $q \in \mathbb{N}$, there is a map $\varphi_q \colon L^q \to L$ of the form
>
> $$(w_1, \dots, w_q) \mapsto w_1 \cdot u_1 \cdot \dots \cdot w_q \cdot u_q,$$
>
> where the "separators" $u_1, \dots, u_q \in U$ may depend on $w_1, \dots, w_q$ and satisfy a "small cancellation condition" that ensures the following: Given $N \in \mathbb{N}$, for all large enough $n \in \mathbb{N}$ and all $q \in \mathbb{N}$, the map $h_n \circ \varphi_q \colon L^q \to \Gamma$ is injective on at least a subset $A_{N,n,q}$ of size $(1/C \cdot \beta_N(L, f(S)))^q$ of $B_N(L, f(S))^q$. In particular,
>
> $$\beta_{q\cdot(N+b)}(\Gamma, f_n(S)) \geqslant \#A_{N,n,q} \geqslant \Big(\frac{1}{C} \cdot \beta_N(L, f(S))\Big)^q.$$
>
> More specifically, this works for all $n \in \mathbb{N}$ that are large enough so that h_n is injective on $B_{2\cdot N}(L, f(S))$; such n exist in view of the convergence of actions.

Given $\varepsilon \in \mathbb{R}_{>0}$, we choose $N \in \mathbb{N}$ large enough to have

$$\frac{1}{N+b} \cdot (\log \beta_N(L, f(S)) - \log C) \geqslant \frac{1}{N} \cdot \log \beta_N(L, f(S)) - \varepsilon.$$

Then, we obtain for all large enough $n \in \mathbb{N}$ that

$$\begin{aligned} \log \mathrm{e}(\Gamma, f_n(S)) &= \lim_{q\to\infty} \frac{1}{q \cdot (N+b)} \cdot \log \beta_{q\cdot(N+b)}(\Gamma, f_n(S)) \\ &\geqslant \frac{1}{N+b} \cdot \log \frac{\beta_N(L, f(S))}{C} \\ &\geqslant \frac{1}{N} \cdot \log \beta_N(L, f(S)) - \varepsilon \\ &\geqslant \log \mathrm{e}(L, f(S)) - \varepsilon, \end{aligned}$$

as desired. □

Remark 3.5. The proof of Theorem 3.4 is mainly based on properties of the limit action on the real tree. In fact, the theorem also holds under the following weaker assumptions on Γ (Fujiwara, 2021, Proposition 3.2): The group Γ is finitely generated, equationally Noetherian, and admits a non-elementary isometric action on a hyperbolic graph X that satisfies a uniform weak proper discontinuity condition (Fujiwara, 2021, Definition 2.1) and that admits a constant N such that for every $S \in \mathrm{FG}(\Gamma)$, the set S^N contains an element that acts hyperbolically on X.

3.3. Well-orderedness

Sketch of proof of Theorem 1.1. If the given hyperbolic group Γ is virtually cyclic, then $\mathrm{Exp}(\Gamma) = \{1\}$, which clearly is well-ordered.

In the following, we consider the case when Γ is non-elementary hyperbolic. We assume for a contradiction that there exists a sequence $(S_n)_{n\in\mathbb{N}}$ of finite generating sets of Γ such that $(\mathrm{e}(\Gamma, S_n))_{n\in\mathbb{N}}$ is strictly monotonically decreasing. In particular, the sequence $(\mathrm{e}(\Gamma, S_n))_{n\in\mathbb{N}}$ is bounded. In view of the compactness theorem (Theorem 3.3) and the invariance of the exponential growth rates under conjugation, we may assume without loss of generality that there exists a free group F with free generating set S and epimorphisms $(f_n \colon F \to \Gamma)_{n\in\mathbb{N}}$ with $f_n(S) = S_n$ and such that f_* converges to a limit group $f \colon F \to L$ over Γ with a faithful action on a real tree.

We therefore obtain from Theorem 3.4 that

$$\begin{aligned}\lim_{n\to\infty} \mathrm{e}(\Gamma, S_n) &= \lim_{n\to\infty} \mathrm{e}(\Gamma, f_n(S)) \\ &= \mathrm{e}(L, f(S)) && \text{(Theorem 3.4)} \\ &\geqslant \mathrm{e}(\Gamma, S_N) > \mathrm{e}(\Gamma, S_{N+1}) > \dots, && \text{(for all } N \gg 0\text{; Theorem 3.4)}\end{aligned}$$

which is impossible. This contradiction shows that no such strictly decreasing sequence exists and hence $\mathrm{Exp}(\Gamma)$ is well-ordered. □

3.4. Finite ambiguity

Sketch of proof of Theorem 1.2. Let $r \in \mathbb{R}_{>1}$ and let us assume for a contradiction that there exists a sequence $(S_n)_{n\in\mathbb{N}}$ of finite generating sets that all represent different $\mathrm{Aut}(\Gamma)$-orbits and that satisfy $\mathrm{e}(\Gamma, S_n) = r$ for all $n \in \mathbb{N}$.

Proceeding as before, by the compactness theorem (Theorem 3.3), we may assume without loss of generality that there exists a free group F with free generating set S and epimorphism $(f_n \colon F \to \Gamma)_{n\in\mathbb{N}}$ with $f_n(S) = S_n$ and such that f_* converges to a limit group $f \colon F \to L$ over Γ with a faithful action on a real tree. Hence, Theorem 3.4 shows that

$$\mathrm{e}(L, f(S)) = \lim_{n\to\infty} \mathrm{e}(\Gamma, f_n(S)) = r.$$

On the other hand, by passing to a subsequence, we may furthermore assume that for all $n \in \mathbb{N}$, there exists a homomorphism $h_n \colon L \to \Gamma$ with $f_n = h_n \circ f$ (as in the proof of Theorem 3.4), that at most one of the epimorphisms h_n is an isomorphism (because the S_n lie in different $\mathrm{Aut}(\Gamma)$-orbits), and that the kernels of the h_n contain no torsion. Then a careful refinement of the proof of Theorem 3.4 shows that the strict inequality

$$\mathrm{e}(L, f(S)) > \mathrm{e}(\Gamma, f_n(S)) = r$$

holds for all $n \in \mathbb{N}$ (Fujiwara and Sela, 2020, Proposition 3.2). This contradicts the previous computation that $\mathrm{e}(L, f(S)) = r$. □

3.5. Growth ordinals

Sketch of proof of Theorem 1.3. It suffices to show that $\mathrm{ord}_{\mathrm{Exp}}(\Gamma) \geqslant \omega^m$ for every $m \in \mathbb{N}$. Let $m \in \mathbb{N}$. We consider the sequence

$$L_1 := \Gamma * F_m \to L_2 := \Gamma * F_{m-1} \to \cdots \to L_m := \Gamma * \mathbb{Z} \to L_{m+1} := \Gamma$$

of epimorphisms, where F_j is a free group of rank j and where the epimorphisms successively kill free generators and keep the Γ-factor intact. It helps to think of j as the number of cusps.

Let us first focus on a single step: If Λ is a non-elementary hyperbolic group, then there exists a stable homomorphism $(f_n \colon \Lambda * \mathbb{Z} \to \Lambda)_{n \in \mathbb{N}}$ consisting of epimorphisms that converges to $\Lambda * \mathbb{Z}$. Let S be a finite generating set of Λ and let $\widetilde{S} \subset \Lambda * \mathbb{Z}$ be a generating set of $\Lambda * \mathbb{Z}$, e.g., obtained by adding a free generator of $\mathbb{Z}$. By passing to subsequences of f_*, one can achieve the following strict monotonicity:

- ▷ The sequence $(\mathrm{e}(\Lambda, f_n(\widetilde{S})))_{n \in \mathbb{N}}$ is increasing and converges to $\mathrm{e}(\Gamma * \mathbb{Z}, \widetilde{S})$; this uses Theorem 3.3 and Theorem 3.4, as before.
- ▷ The values in the sequence are all different; this uses a finite ambiguity theorem for finitely generated subgroups of limit groups over Λ (Fujiwara and Sela, 2020, Theorem 5.8).

For notational simplicity, we now restrict to the case $m = 2$. We choose a finite generating set S of Γ and take the extended finite generating set $\widetilde{S}$ of $L_1 = \Gamma * F_2$. Applying the single step to $L_1 \to L_2$ leads to a stable homomorphism f_*^1 with strict monotonicity. Let f_*^2 be a stable homomorphism for $L_2 \to L_3$. For each $n_1 \in \mathbb{N}$, we apply the single step to $L_2 \to L_3$ and the generating set $f_{n_1}(\widetilde{S})$ to select a subsequence of f_*^2 with strict monotonicity. By composing with f_{n_1}, we obtain a sequence $f_{n_1,*}$ from L_1 to $L_3 = \Gamma$ such that $(\mathrm{e}(\Gamma, f_{n_1,n}(\widetilde{S})))_{n \in \mathbb{N}}$ is strictly increasing and converges to $\mathrm{e}(L_2, f_{n_1}(\widetilde{S}))$. By varying n_1, we thus see that $\mathrm{ord}_{\mathrm{Exp}}(\Gamma) \geqslant \omega^2$.

For higher values of m, one iterates these considerations appropriately. □

To prove $\mathrm{ord}_{\mathrm{Exp}}(\Gamma) \leqslant \omega^\omega$ under additional hypotheses, Fujiwara and Sela (2020, proof of Theorem 4.2) construct proper epimorphism chains of limit groups over Γ from convergent sequences of convergent sequences of etc... of exponential growth rates of Γ; the Krull dimension property then gives control on the maximal lengths of such chains, whence on the maximal powers of ω that appear below a given threshold.

4. Applications and extensions

The well-orderedness of exponential growth rates (Theorem 1.1) in particular contains the fact that all non-elementary hyperbolic groups have uniform exponential growth.

4.1. Hyperbolic groups are Hopfian

A group Γ is *Hopfian* if every self-epimorphism $\Gamma \to \Gamma$ is an automorphism. This property has applications in the context of degrees of self-maps of closed manifolds. Hyperbolic groups are known to be Hopfian (Sela, 1999; Weidmann and Reinfeldt, 2019). Using that the exponential growth rates of hyperbolic groups are well-ordered, Fujiwara and Sela (2020, Corollary 2.9) complete an approach to proving that hyperbolic groups are Hopfian outlined by de la Harpe (2002); this is not an independent alternative proof because the current proof of Theorem 1.1 uses the very results on limit groups that go into the previous proofs that hyperbolic groups are Hopfian.

Corollary 4.1. *Every hyperbolic group is Hopfian.*

Proof. Elementary hyperbolic groups are Hopfian because they are virtually cyclic (whence finitely generated and residually finite).

Let Γ be a non-elementary hyperbolic group and let $f\colon \Gamma \to \Gamma$ be an epimorphism. Because $\mathrm{Exp}(\Gamma)$ is well-ordered (Theorem 1.1), there exists a finite generating set S of Γ with $\mathrm{e}(\Gamma) = \mathrm{e}(\Gamma, S)$. *Assume* for a contradiction that the kernel of f is non-trivial. Then, Arzhantseva and Lysenok (2002) show that there is a *strict* monotonicity

$$\mathrm{e}(\Gamma, S) > \mathrm{e}(\Gamma, f(S)).$$

However, this contradicts the minimality property of S. Thus, f is an automorphism. □

All finitely generated residually finite groups are Hopfian. While fundamental groups of closed hyperbolic manifolds are residually finite and hyperbolic, it is a long-standing open problem whether all hyperbolic groups are residually finite.

4.2. Generalisations

The methods discussed in Section 3 by Fujiwara and Sela (2020) extend to cover also the following generalisations:

- If Γ is a hyperbolic group, then the set

$$\{\mathrm{e}(H,S) \mid H < \Gamma \text{ finitely generated and non-elementary, } S \in \mathrm{FG}(H)\}$$

  is well-ordered (Fujiwara and Sela, 2020, Theorem 5.1). This can be viewed as an addition to the Tits alternative for hyperbolic groups.

- Moreover, in this subgroup setting, there is a corresponding finite ambiguity statement for non-elementary hyperbolic groups (Fujiwara and Sela, 2020, Theorem 5.3).

As a consequence, they also obtain analogous results for limit groups over non-elementary hyperbolic groups (Fujiwara and Sela, 2020, Corollary 5.6–5.10). Furthermore, the approach is robust enough to admit an extension to the case of sub-semigroups (Fujiwara and Sela, 2020, Section 6).

Fujiwara (2021) adapted the method to obtain well-orderedness of exponential growth rates sets for other classes of groups, including certain groups acting acylindrically on hyperbolic spaces, rank-1 lattices, fundamental groups of strictly negatively curved Riemannian manifolds, and certain relatively hyperbolic groups. These results can, for instance, be applied to certain subgroups of right-angled Artin groups (Kerr, 2021, Corollary 1.0.11).

A. Terminology

For the sake of completeness, we recall the basic terminology appearing in the main results (Section 1).

A.1. Hyperbolic groups

Finitely generated groups are hyperbolic if their Cayley graphs are "negatively curved" in the sense that geodesic triangles in are uniformly slim (Bridson and Haefliger, 1999; Gromov, 1987):

Definition A.1 (hyperbolic group). A finitely generated group Γ is *hyperbolic* if the Cayley graph of Γ with respect to one (whence every (Bridson and Haefliger, 1999, Theorem III.H.1.9)) finite generating set is a hyperbolic metric space. A hyperbolic group is *non-elementary* if it is not virtually cyclic.

Example A.2 (hyperbolic groups). Fundamental groups of closed smooth manifolds that admit a Riemannian metric of negative sectional curvature are hyperbolic in view of the Švarc–Milnor lemma and the fact that CAT(κ)-spaces with $\kappa < 0$ are hyperbolic metric spaces. In particular, this includes the fundamental groups of closed hyperbolic manifolds. Such fundamental groups are virtually cyclic if and only if the dimension is at most 1.

Finitely generated free groups are hyperbolic. The class of hyperbolic groups is closed under quasi-isometries (Bridson and Haefliger, 1999, Theorem III.H.1.9) and under certain amalgamations (Bestvina and Feighn, 1996).

The group $\mathbb{Z}^2$ is *not* hyperbolic. More generally, all finitely generated groups that contain a subgroup isomorphic to $\mathbb{Z}^2$ are *not* hyperbolic (Bridson and Haefliger, 1999, Corollary III.Γ.3.10). In general, subgroups of hyperbolic groups need *not* be hyperbolic.

A.2. Exponential growth rates of groups

The exponential growth rate of groups measures the exponential expansion rate of the size of balls in Cayley graphs (de la Harpe, 2000, Chapter VII.B):

Remark A.3. Let Γ be a finitely generated group and let $S \subset \Gamma$ be a finite generating set of Γ. We write $\beta_n(\Gamma, S)$ for the number of elements in the n-ball $B_n(\Gamma, S)$ of the Cayley graph of Γ with respect to S. Then $\beta_{n+m}(\Gamma, S) \leqslant \beta_n(\Gamma, S) \cdot \beta_m(\Gamma, S)$ for all $n, m \in \mathbb{N}$. Therefore, the Fekete lemma shows that the limit of $(\beta_n(\Gamma, S)^{1/n})_{n\in\mathbb{N}}$ exists and that

$$\lim_{n\to\infty} \beta_n(\Gamma, S)^{1/n} = \inf_{n\in\mathbb{N}_{>0}} \beta_n(\Gamma, S)^{1/n}.$$

Definition A.4 (exponential growth rate). Let Γ be a finitely generated group.

- ▷ Let $S \subset \Gamma$ be a finite generating set of Γ. The *exponential growth rate of* Γ *with respect to* S is defined as
$$\mathrm{e}(\Gamma, S) := \lim_{n\to\infty} \beta_n(\Gamma, S)^{1/n}.$$
- ▷ We write $\mathrm{Exp}(\Gamma) := \{\mathrm{e}(\Gamma, S) \mid S \in \mathrm{FG}(\Gamma)\}$ for the (countable) set of all exponential growth rates of Γ, where $\mathrm{FG}(\Gamma)$ denotes the set of finite generating sets of Γ.
- ▷ The *exponential growth rate of* Γ is the infimum
$$\mathrm{e}(\Gamma) := \inf \mathrm{Exp}(\Gamma).$$
- ▷ The group Γ has *exponential growth* if there exists an $S \in \mathrm{FG}(\Gamma)$ with $\mathrm{e}(\Gamma, S) > 1$. The group Γ has *uniform exponential growth* if $\mathrm{e}(\Gamma) > 1$.

Remark A.5 (monotonicity of exponential growth rates). Let Γ and Λ be finitely generated groups.

1. If $f\colon \Gamma \to \Lambda$ is an epimorphism and $S \in \mathrm{FG}(\Gamma)$, then $\mathrm{e}(\Gamma, S) \geqslant \mathrm{e}(\Lambda, f(S))$.
2. If Λ is a subgroup of Γ and Λ has exponential growth, then also Γ has exponential growth.

Example A.6 (exponential growth). Finitely generated free groups have exponential growth if and only if they are of rank at least 2. More precisely (DE LA HARPE, 2000, Proposition VII.13): If F is a free group and $S \in \mathrm{FG}(F)$, then

$$\mathrm{e}(F, S) \geqslant 2 \cdot \mathrm{rk}(F) - 1.$$

Because non-elementary hyperbolic groups [uniformly] contain free groups of rank 2, monotonicity shows that they have [uniform] exponential growth (KOUBI, 1998).

There exist finitely generated groups that have exponential growth but do *not* have uniform exponential growth (WILSON, 2004). In particular, for such groups Γ, the set $\mathrm{Exp}(\Gamma)$ is *not* well-ordered.

Exponential growth rates seem to be fragile under quasi-isometries: It is an open problem to determine whether uniform exponential growth is stable under quasi-isometries.

A.3. Well-ordered countable sets and ordinals

Well-orderings are orderings that allow for induction principles. Moreover, well-orderings admit an arithmetic, the ordinal arithmetic.

Definition A.7 (well-ordered sets, ordinals). An ordered set $(A, <)$ is *well-ordered* if every non-empty subset of A contains a $<$-minimal element. An *ordinal* is an isomorphism class of well-ordered ordered sets. An ordinal is *countable* if the underlying set is countable.

In the context of Section 1, the following ordinals are important (Figure 1):

Example A.8 (ω^ω). The natural numbers $\mathbb{N}$ are well-ordered with respect to the standard order. The corresponding ordinal is denoted ω. For $k \in \mathbb{N}$, we write ω^k for the ordinal represented by $\mathbb{N}^k$ with the lexicographic order. Equipping the finite support functions $\mathbb{N} \to \mathbb{N}$ with the lexicographic order leads to a well-ordered set; its ordinal number is denoted by ω^ω. The ordinal ω^ω can alternatively also be described as $\sup_{k\in\mathbb{N}} \omega^k$.

Example A.9. The subsets $A := \{1 - 1/n \mid n \in \mathbb{N}_{>0}\}$ and $B := \bigcup_{n\in\mathbb{N}}(n + A)$ of $\mathbb{R}$ are well-ordered with respect to the standard order on $\mathbb{R}$. The set A represents the ordinal ω and B represents the ordinal ω^2. The subset $\mathbb{Q}_{\geqslant 0} \subset \mathbb{R}$ is *not* well-ordered.

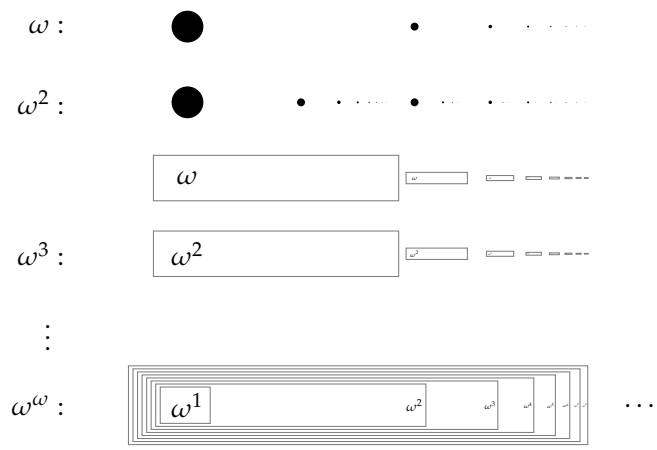


Figure 1: The ordinals ω, ω^2, ..., ω^ω, schematically

B. Right-computability of exponential growth rates

We provide proofs for the right-computability claims in Section 2.3.

Proof of Proposition 2.7. Let $F(S)$ be the set of reduced words over $S \sqcup S^{-1}$. In particular, $F(S)$ is a free group, freely generated by S, with respect to the composition given by concatenation and reduction. It is well known that we can Turing-enumerate all finite subsets of $F(S)$ that represent generating sets of Γ under the canonical projection $F(S) \to \langle S \mid R \rangle = \Gamma$ (by Turing-enumerating the normal closure of R in $F(S)$).

Therefore it suffices to show that there exists a Turing machine that given a finite generating set $S' \subset F(S)$ of Γ enumerates the set $A(S') := \{x \in \mathbb{Q} \mid x > \mathrm{e}(\Gamma, S')\}$. By the Fekete lemma (Remark A.3), for all generating sets S', we have

$$A(S') = \left\{x \in \mathbb{Q} \;\middle|\; \exists_{n \in \mathbb{N}_{>0}} \quad x^n > \beta_n(\Gamma, S')\right\}.$$

The numbers $\beta_n(\Gamma, S')$ are not necessarily computable in terms of n and S' (as the word problem might not be solvable in Γ), but recursively enumerating the normal closure of R in $F(S)$ shows that there exists a Turing machine that given a finite generating set $S' \subset F(S)$ of Γ enumerates $\{(n, m) \mid n, m \in \mathbb{N},\ m \geqslant \beta_n(\Gamma, S')\}$; hence, there is also a Turing machine for $A(\,\cdot\,)$. □

Proof of Corollary 2.8. The first part is a direct consequence of Proposition 2.7.

For the second part, let $r \in \mathbb{Q}$. For all $S \in \mathrm{FG}(\Gamma)$, we have

$$\mathrm{e}(\Gamma, S) < r \iff \exists_{x \in \mathbb{Q}} \quad (r > x \wedge x > \mathrm{e}(\Gamma, S)).$$

Let $\langle S \mid R \rangle$ be a finite presentation of Γ. Using a Turing machine as provided by Proposition 2.7, we can thus construct a Turing machine that enumerates all finite sets S' of words over $S \sqcup S^{-1}$ such that S' represents a generating set of Γ and such that $e(\Gamma, S') < r$. □

For simplicity, we restricted the discussion to finitely presented groups. Similar arguments also apply to finitely generated recursively presented groups. Conversely, one might wonder whether every right-computable real number $\geqslant 1$ can be realised as the exponential growth rate of some finite generating set of some finitely/recursively presented group.

Acknowledgements

This work was supported by the CRC 1085 *Higher Invariants* (Universität Regensburg, funded by the DFG).

References

Arzhantseva, G. N. and Lysenok, I. G. (2002). "Growth tightness for word hyperbolic groups", *Math. Z.* **241** (3), pp. 597–611.

——— (2006). "A lower bound on the growth of word hyperbolic groups", *J. London Math. Soc.* **73** (1), pp. 109–125.

Benedetti, R. and Petronio, C. (1992). *Lectures on hyperbolic geometry*. Universitext. Springer.

Bestvina, M. and Feighn, M. (1996). "Addendum and correction to: "A combination theorem for negatively curved groups" [J. Differential Geom., 35(1), 85–101, 1992]", *J. Differential Geom.* **43** (4), pp. 783–788.

Bridson, M. R. and Haefliger, A. (1999). *Metric spaces of non-positive curvature*. Vol. 319. Grundlehren der mathematischen Wissenschaften. Springer.

de la Harpe, P. (2000). *Topics in geometric group theory*. Chicago Lectures in Mathematics. University of Chicago Press.

——— (2002). "Uniform growth in groups of exponential growth". In: *Proceedings of the Conference on Geometric and Combinatorial Group Theory, Part II (Haifa, 2000)*. Vol. 95, pp. 1–17.

Fujiwara, K. (2021). "The rates of growth in an acylindrically hyperbolic group". preprint, arXiv:2103.01430.

Fujiwara, K. and Manning, J. F. (2011). "Simplicial volume and fillings of hyperbolic manifolds", *Algebr. Geom. Topol.* **11** (4), pp. 2237–2264.

Fujiwara, K. and Sela, Z. (2020). "The rates of growth in a hyperbolic group". preprint, arXiv:2002.10278.

Gromov, M. (1981). "Hyperbolic manifolds (according to Thurston and Jørgensen)", in: *Bourbaki Seminar, Vol. 1979/80*. Vol. 842. Lecture Notes in Math. Springer, pp. 40–53.

——— (1982). "Volume and bounded cohomology", *Inst. Hautes Études Sci. Publ. Math.* **56**, 5–99 (1983).

——— (1987). "Hyperbolic groups", in: *Essays in group theory*. Vol. 8. Math. Sci. Res. Inst. Publ. Springer, pp. 75–263.

Groves, D. and Wilton, H. (2018). "The structure of limit groups over hyperbolic groups", *Israel J. Math.* **226** (1), pp. 119–176.

Heuer, N. (2019). "The full spectrum of scl on recursively presented groups". preprint, arXiv:1909.01309.

Heuer, N. and Löh, C. (2023). "Transcendental simplicial volumes", *Annales de l'Institut Fourier*. To appear.

Kerr, A. (2021). "Product set growth in mapping class groups". preprint, arXiv:2103.12643.

Kharlampovich, O. and Myasnikov, A. (1998a). "Irreducible affine varieties over a free group. I. Irreducibility of quadratic equations and Nullstellensatz", *J. Algebra* **200** (2), pp. 472–516.

——— (1998b). "Irreducible affine varieties over a free group. II. Systems in triangular quasi-quadratic form and description of residually free groups", *J. Algebra* **200** (2), pp. 517–570.

Koubi, M. (1998). "Croissance uniforme dans les groupes hyperboliques", *Ann. Inst. Fourier* (*Grenoble*) **48** (5), pp. 1441–1453.

Löh, C. (2023). "The spectrum of simplicial volume with fixed fundamental group", *Geom. Dedicata* **217** (2), Paper No. 16.

Löh, C. and Uschold, M. (2022). "L^2-Betti numbers and computability of reals". preprint, arXiv:2202.03159, to appear in *Computability*, DOI 10.3233/COM-220416.

Louder, L. (2012). "Krull dimension for limit groups", *Geom. Topol.* **16** (1), pp. 219–299.

Mineyev, I. (2001). "Straightening and bounded cohomology of hyperbolic groups", *Geom. Funct. Anal.* **11** (4), pp. 807–839.

Sela, Z. (1999). "Endomorphisms of hyperbolic groups. I. The Hopf property", *Topology* **38** (2), pp. 301–321.

——— (2006). "Diophantine geometry over groups. VI. The elementary theory of a free group", *Geom. Funct. Anal.* **16** (3), pp. 707–730.

——— (2009). "Diophantine geometry over groups. VII. The elementary theory of a hyperbolic group", *Proc. Lond. Math. Soc.* **99** (1), pp. 217–273.

Thurston, W. P. (1979). *The geometry and topology of* 3*-manifolds*. mimeographed notes. Princeton.

Wang, H. C. (1972). "Topics on totally discontinuous groups", in: *Symmetric spaces (Short Courses, Washington Univ., St. Louis, Mo., 1969–1970)*. Pure and Appl. Math., Vol. 8. Dekker, pp. 459–487.

Weidmann, R. and Reinfeldt, C. (2019). "Makanin–Razborov diagrams for hyperbolic groups", *Ann. Math. Blaise Pascal* **26** (2), pp. 119–208.

Wilson, J. S. (2004). "On exponential growth and uniformly exponential growth for groups", *Invent. Math.* **155** (2), pp. 287–303.

Clara Löh

Fakultät für Mathematik
Universität Regensburg
93040 Regensburg
Germany

E-mail: clara.loeh@ur.de

Séminaire BOURBAKI Avril 2023
74e année, 2022–2023, no 1207, p. 383 à 416
doi: 10.24033/ast.1216

STRONG FORCING AXIOMS AND THE CONTINUUM PROBLEM
[*after* Asperó's and Schindler's proof that MM^{++} implies Woodin's Axiom $(*)$]

by **Matteo Viale**

Introduction

This note addresses the continuum problem, taking advantage of the breakthrough mentioned in the subtitle, and relating it to many recent advances occurring in set theory.[(1)] We try to the best of our possibilities to make our presentation self-contained and accessible to a general mathematical audience.[(2)]

Let us start by stating Asperó's and Schindler's result:

Theorem 0.1 (Asperó and Schindler, 2021). *Assume* $\mathbf{MM}^{++}$ *holds. Then Woodin's axiom* $(*)$ *holds as well.*

We will address the following three questions:

- ▷ What is the axiom $\mathbf{MM}^{++}$?
- ▷ What is Woodin's axiom $(*)$?
- ▷ What is the bearing of Asperó's and Schindler's result on the continuum problem, and why their result is regarderd as a major breakthrough in the set theoretic community?

We give rightaway a spoiler of the type of answers we sketch for the above questions.

We have two major approaches to produce witnesses x of certain mathematical properties $P(x)$.

(1)The author acknowledges support from the project: *PRIN 2017-2017NWTM8R Mathematical Logic: models, sets, computability* and from GNSAGA.

(2)Surveys on the topic complementing this note are (among an ample list) Bagaria, 2005; Koellner, 2010; Venturi and Viale, 2023b; Woodin, 2001a,b.

A topological approach is exemplified by Baire's category theorem: given a compact Hausdorff topological space X one can find a "generic" point $x \in X$ satisfying a certain topological property $P(x)$ by showing that $P(x)$ can fail only on a "small" (more precisely meager) set of points of X.

An algebraic approach is exemplified by the construction of algebraic numbers: one takes a set of Diophantine equations $P(\vec{x})$ which are not jointly inconsistent, and builds abstractly a formal solution in the ring $\mathbb{Q}(\vec{x})/_{P(\vec{x})}$.

Duality theorems connect the algebraic point of view to the geometric one, for example Hilbert's Nullstellensatz relates solutions of irreducible sets of Diophantine equations to generic points of algebraic varieties.

We will outline that Woodin's axiom $(*)$ provides an "algebraic approach" to the construction of set theoretic witnesses for "elementary" set theoretic properties, $\mathbf{MM}^{++}$ a "geometric approach", and Asperó's and Schindler's result connects these two perspectives.

We plan to do this while gently introducing the reader to the fundamental concepts of set theory.

The note is structured as follows:

- ▷ 1 is a brief review of the basic results of set theory with a focus on its historical development and on the topological complexity of sets of reals witnessing the failure of the continuum hypothesis.
- ▷ In 2 we quote some of Gödel's thoughts on the continuum problem and on the ontology of mathematical entities.
- ▷ 3 gives a brief overview of (the use in mathematics of) large cardinal axioms.
- ▷ In 4 we introduce forcing axioms with a focus on their topological presentations, while giving a precise formulation of the axiom $\mathbf{MM}^{++}$. We also list some of the major undecidable problems which get a solution assuming this axiom, among which the continuum problem.
- ▷ 5 is a small interlude giving some insights on the forcing method, while relating it to the notions of sheaf and of Grothendieck topos.
- ▷ 6 revolves about the notion of algebraic closure. In particular we outline how Robinson's notion of model companionship gives the means to transfer the concept of "algebraic closure" developed for rings to a variety of other mathematical theories.
- ▷ 7 discusses what is the right language in which set theory should be axiomatized in order to unfold its "algebraic closure" properties.

- ▷ 8 relates Woodin's generic absoluteness results for second order number theory to properties of algebraic closure for the initial fragment of the universe of sets given by $H_{\aleph_1}$.
- ▷ 9 brings to light why Woodin's axiom $(*)$ can be regarded as an axiom of "algebraic closure" for the larger initial fragment of set theory given by $H_{\aleph_2}$. Putting everything together we conclude by showing why Asperó's and Schindler's result establish a natural correspondence between the geometric approach and the algebraic approach to forcing axioms.

I thank Alberto Albano, David Asperó, Vivina Barutello, Raphaël Carroy, Ralf Schindler for many helpful comments on the previous drafts of this manuscript. Many thanks to Nicolas Bourbaki for the invitation and the precious editorial support in the preparation and revision of this work.

1. Basics of set theory

Set theory deals with the properties of sets (the "manageable" mathematical objects) and classes (the "not so manageable" entities).[(3)]

1.1. Axioms

The axioms of set theory can be split in three types (as is the case for many other mathematical theories):

- ▷ **Universal axioms** which establish properties valid for all sets;
- ▷ **Existence axioms** which establish the existence of certain sets;
- ▷ **Construction principles** which allow for the construction of new sets from ones which are already known to exist.

We present the axiomatization of set theory by Morse–Kelley MK with sets and classes. Its axioms are distributed in the three categories as follows:

Universal axioms

- ▷ **Extensionality:** Two classes (or sets) are equal if they have exactly the same elements.

[(3)] We refer the reader to Jech, 2003; Kunen, 1980; Monk, 1969 for a systematic treatment of the topic. The reader familiar with set theory can skim through or just skip this section.

- ▷ **Comprehension (a):** Every class (or set) is a subset of V, the (proper) class whose elements are exactly the sets.
  (a **proper class** is a class which is not a set, a **set** is a class which belongs to V).
- ▷ **Foundation:** There is no infinite sequence $\langle x_n : n \in \mathbb{N} \rangle$ of classes such that $x_{n+1} \in x_n$ for all n.

Existence axioms

- ▷ **Infinity:** $\emptyset$ and $\mathbb{N}$ are sets.

Weak construction principles

- ▷ **Union, Pair, Product:** If X, Y are sets, so are $X \cup Y$, $\{X, Y\}$, $X \times Y$.
- ▷ **Separation:** If P is a class and X is a set, $P \cap X$ is a set.

Strong construction principles

- ▷ **Comprehension (b):** For every property $\psi(x)$, $P_\psi = \{a \in V : \psi(a)\}$ is a class.
- ▷ **Replacement:** If F is a class function and $X \subseteq \mathrm{dom}(F)$ is a set, the pointwise image $F[X]$ of X under F is a set.
- ▷ **Powerset:** If X is a set, so is the class $\mathscr{P}(X) = \{Y : Y \subseteq X\}$.
- ▷ **Global Choice:** For all classes $C = \{X_i : i \in I\}$ of non-empty sets X_i, $\prod_{i \in I} X_i$ (the family of functions F with domain I and such that $F(i) \in X_i$ for all $i \in I$) is non-empty.

Some comments:

- ▷ By Foundation V cannot be a set else $\langle x_n : n \in \mathbb{N} \rangle$ with each x_n constantly assigned to V defines a decreasing $\in$-chain.[(4)]
- ▷ Many of the objects of interest in mathematics are proper classes, for example the family of groups, or the family of topological spaces. More generally for a given (first order) theory T, the family of stuctures which satisfy the axioms of T is a proper class (and exists in view of Comprehension (b)). There are delicate ontological issues related to the notion of proper class, but they are foreign to almost all domains of mathematics, with the notable exceptions of category theory and set theory.

[(4)] *V is not a set* can also be proved without Foundation. Set theorists need foundation in order to infer that the notion of well-foundedness is an elementary set theoretic property (more precisely it is a provably Δ_1-property).

▷ It is convenient for natural numbers to distinguish their ordinal type (which confronts them according to which of these numbers "comes first") from their cardinal type (which assigns to each natural number n the family of sets which have exactly n elements). When dealing with arbitrary sets, their ordinal type may not be defined, while the cardinal type always is. Von Neumann devised a simple trick to represent the finite ordinal types. One can inductively define the natural number n as the set $\{0, \dots, n-1\}$ (*i.e.* $0 = \emptyset, 1 = \{\emptyset\}, 2 = \{\emptyset, \{\emptyset\}\}$, ...).[(5)]

▷ Set theoretic construction principles are of two sorts: the simple (or weak) ones are for example those bringing from sets X, Y to sets $X \cup Y$, $\{X, Y\}$, $X \times Y$, or from set X and class P to the set $P \cap X$; the strong ones are the power-set axiom, the replacement axiom, and the axiom of choice. Let us discuss briefly the role of such axioms in the development of routine mathematics.

Weak construction principles The integers and rationals can be constructed from the naturals using only weak construction principles: $\mathbb{Z}$ can be seen as the subset of $\mathbb{N} \times \{0, 1\}$ which assigns the positive integers to the ordered pairs with second coordinate 0 and the negative ones to those pairs with second coordinate 1 (paying attention to the double counting of 0 as $(0, 0)$ and $(0, 1)$); $\mathbb{Q}$ can be seen as the subset of $\mathbb{Z} \times (\mathbb{N} \setminus \{0\})$ given by ordered pairs which are coprime.

Powerset axiom In order to build the reals from the rationals, one needs this axiom: $\mathbb{R}$ is the subset of $\mathscr{P}(\mathbb{Q})$ given by Dedekind cuts.

Replacement axiom An adequate development of set theory requires it: consider the function F on $\mathbb{N}$ given by $F(0) = \mathbb{N}$, $F(n+1) = \mathscr{P}(F(n))$. Without replacement it cannot be proved that F (or even the image of F) is a set, it might only be a proper class.

Choice Choice also has a special status in ordinary mathematics, and many mathematicians feel uneasy about it. However Choice is unavoidable: it is essential in the proofs of the Hahn–Banach theorem, of the existence of a base for infinite-dimensional vector spaces, or of the existence of a maximal ideal on a ring,...Even the equivalence of sequential continuity and topological continuity for real valued functions requires it: if $f : \mathbb{R} \to \mathbb{R}$ is not continuous at x, there is $\varepsilon > 0$ such that for each n one can find x_n so that $|x_n - x| < 1/n$ and $|f(x_n) - f(x)| > \varepsilon$. The sequence $(x_n)_n$ is (and in most cases can only be) defined appealing to (countable) Choice.

[(5)] The transfinite ordinal types (or the Von Neumann ordinals) are those (possibly infinite) sets α which are linearly ordered by $\in$ and are transitive (*i.e.* such that when $x \in y \in \alpha$, we have that $x \in \alpha$ as well). The proper class of Von Neumann ordinals is linearly well-ordered by $\in$. One can check that the natural numbers are the finite Von Neumann ordinals and that $\mathbb{N}$ (the set of finite Von Neumann ordinals) is the first infinite Von Neumann ordinal.

1.2. Cardinal arithmetic

Let us now develop arithmetic on cardinal types.

Given sets X, Y

- ▷ The cardinal $|X|$ is the (proper) class $\{Y : \exists f : X \to Y \text{ bijection}\}$;
- ▷ $|X| \leqslant |Y|$ if and only if *there is* $f : X \to Y$ *injection* if and only if *there is* $g : Y \to X$ *surjection*;[6]
- ▷ $|X| < |Y|$ iff $|X| \leqslant |Y|$ and $|X| \neq |Y|$;
- ▷ $|X| + |Y| = |(X \times \{0\}) \cup (Y \times \{1\})|$ (the size of the disjoint union of X and Y);
- ▷ $|X| \cdot |Y| = |X \times Y|$ (the size of the product of X and Y);
- ▷ $|X|^{|Y|} = |X^Y|$ where

$$X^Y = \{f : \ f \text{ is a function with domain } Y \text{ and range } X\}.$$

Note that each cardinality class on a set X is a proper class if it contains a non-empty element. The equivalence class of $\emptyset$ is $\{\emptyset\}$. It can also be shown that C is a proper class if and only if there is a surjection of C onto V, and X is a set if and only if there is no surjection of X onto V.

The finite sets are those in the equivalence class of some $n \in \mathbb{N}$. For finite sets X, Y of size respectively m, n:

- ▷ $|X| = |Y|$ if and only if $n = m$ if and only if $|X| \leqslant |Y|$ and $|Y| \leqslant |X|$;
- ▷ $|X| + |Y|$ is (the equivalence class of) $m + n$;
- ▷ $|X| \cdot |Y|$ is (the equivalence class of) $m \cdot n$;
- ▷ $|X|^{|Y|}$ is (the equivalence class of) m^n;

i.e. the map $n \mapsto |n|$ defines an embedding of the stucture $(\mathbb{N}, <, \cdot, +, n \mapsto 2^n, 0, 1)$ in the family of cardinals endowed by the above operations.

Basic questions were then asked for the infinite cardinals:

1. Assume $|X| \leqslant |Y|$ and $|Y| \leqslant |X|$ holds for infinite sets X, Y, do we have also in this case that $|X| = |Y|$?
2. What is the structure of the order on infinite cardinals given by $<$?
3. Do the arithmetic of cardinals obey the associative, commutative, distributive laws which hold for the naturals?

[6] The last equivalence gives an alternative definition of the axiom of choice.

4. What is the computation table for the arithmetic operations of sum and product on cardinals, when one of the factors is an infinite cardinal?

5. What about the computation table for the exponential map $|X| \mapsto 2^{|X|}$ when X is infinite?

Soon most of these questions had a clear cut answer:

1. For arbitrary sets X, Y we have that $|X| \leqslant |Y|$ and $|Y| \leqslant |X|$ if and only if $|X| = |Y|$ (Cantor 1887, Bernstein 1897, Dedekind 1898).[(7)]

2. $\leqslant$ is a well-order on cardinals (Zermelo+...$\sim$ 1904), *i.e.* it is a linear order on cardinals such that for every class $C \neq \emptyset$ there is min $\{|X| : X \in C\}$.[(8)]

3. The arithmetic operations on cardinals endow this family of a structure of ordered commutative semiring with exponentiation; the natural numbers form an initial segment of this semiring (Cantor - published in 1895, but most likely known earlier).

4. $|X| + |Y| = |X| \cdot |Y| = \max\{|X|, |Y|\}$ if at least one among X, Y is infinite and both are non-empty (Special cases by Cantor, before 1895, general proofs by Harward 1905, Jourdain 1908, Hausdorff 1914...).

5. For all sets X, $|X| < |\mathscr{P}(X)| = |2|^{|X|}$ (Cantor 1891).[(9)]

We denote cardinals by greek letters κ, λ, θ.

- ▷ $\aleph_0$ is the cardinality of $\mathbb{N}$ the least infinite cardinal;
- ▷ for κ a cardinal κ^+ is the cardinality of the least cardinal above κ (which exists by 2 and 5);
- ▷ $\aleph_1 = \aleph_0^+$, $\aleph_2 = \aleph_1^+$.

[(7)] This is non trivial: there is a topological embedding of $[0;1]$ into $(0;1)$ and conversely; there is no continuous bijection of $[0;1]$ onto $(0;1)$; however one can find a Borel bijection.

[(8)] Furthermore in each class $|X|$ there is a least Von Neumann ordinal κ which is called the cardinal of X and is the canonical representative of $|X|$.

[(9)] $x \mapsto \{x\}$ witnesses $|X| \leqslant |\mathscr{P}(X)|$; if $g : \mathscr{P}(X) \to X$, $Y_g = \{y \in Y : y \notin g(y)\}$ witnesses that g is not a surjection.

Regarding the map $(|X|, |Y|) \mapsto |X|^{|Y|}$, the computation of its table can be done in terms of that of the exponential map (Bukovský, 1965).

1.3. The continuum problem

The so called continuum problem (Cantor 1878) remained unsettled:

What is the value of $2^{|\mathbb{N}|}$?

The continuum hypothesis can be equivalently phrased:

- $2^{\aleph_0} = \aleph_1$;
- For all $X \subseteq \mathbb{R}$ either X is countable or X has size continuum.

The equivalence of the two follows once one observes that $|\mathbb{R}| = |\mathscr{P}(\mathbb{N})|$.

1.4. Progresses on the definable version of the continuum problem

For decades there were scant progresses on the solution of this problem. On the positive side one can note proofs that "simply definable" subsets of $\mathbb{R}$ cannot witness the negation of the continuum hypothesis (the so called "continuum hypothesis for definable sets").

The *analytic* (or Σ^1_1) subsets of $\mathbb{R}^n$ are the projections on the first n-coordinates of a Borel subset of $\mathbb{R}^{n+k}$ for some natural number k. The *coanalytic* (or Π^1_1) subsets of $\mathbb{R}^n$ are the complements of the analytic subsets.

- No *closed* subset of $\mathbb{R}$ is a counterexample to CH (Cantor 1883).
- No *Borel* subset of $\mathbb{R}$ is a counterexample to CH (Alexandroff 1916, Hausdorff 1917).
- No *analytic* subset of $\mathbb{R}$ is a counterexample to CH (Suslin+Alexandroff 1917).
- *Coanalytic* subsets of $\mathbb{R}$ have either size $\aleph_0, \aleph_1, 2^{\aleph_0}$ (Luzin–Sierpinski 1917$\approx$).

This was as far as the first rounds of mathematicians addressing Cantor's problem could get.

With the introduction of large cardinals further major progresses were obtained on the definable version of CH.

The *projective* subsets of $\mathbb{R}^n$ are those subsets of $\mathbb{R}^n$ which are Σ^1_m (or Π^1_m) for some m, where $X \subseteq \mathbb{R}^n$ is Σ^1_{m+1} if it is the projection of a Π^1_m-subset of $\mathbb{R}^{n+k}$ and is a Π^1_{m+1}-subset if it is the complement of a Σ^1_{m+1}-set.

Projective sets define a natural family of subsets of $\mathbb{R}^n$ which ought to be topologically simple.

Another natural family has been isolated by Feng, Magidor, and Woodin (1992):

Recall that for a given topological space (Y, τ), a set X is *nowhere dense* in Y if its complement contains an open dense subset of Y, it is *meager* if it is a countable union of nowhere dense subsets, it has the *Baire property* if it has meager symmetric difference with an open set.

Definition 1.1 (Feng, Magidor, and Woodin, 1992). $X \subseteq \mathbb{R}^k$ is *universally Baire* if for all continuous maps $f : Y \to \mathbb{R}^k$ with Y compact Hausdorff, $f^{-1}[X]$ has the Baire property in Y.

Analytic sets are universally Baire, and this family forms a σ-algebra.

To appreciate the strength of this property, consider $2^{\mathbb{N}}$ when 2 is given the discrete topology and $2^{\mathbb{N}}$ has the product topology.

The map $\theta : f \mapsto \sum_{i=0}^{\infty} f(i)/3^{i+1}$ for $f : \mathbb{N} \to \{0, 2\}$ defines a topological embedding of $2^{\mathbb{N}}$ into $[0; 1]$ whose image is meager and has Lebesgue measure 0.

Now take a subset P of $2^{\mathbb{N}}$ which does not have the Baire property in $2^{\mathbb{N}}$.

Seen as a subset of $[0; 1]$, $\theta[P]$ is meager, hence it has the Baire property in $[0; 1]$. On the other hand $P = \theta^{-1}[\theta[P]]$ does not have the Baire property in $2^{\mathbb{N}}$. Hence $\theta[P]$ is not universally Baire even if it has the Baire property.

A similar argument based on measure, shows that the non-measurable sets produced by Vitali are not universally Baire. More generally it can be shown that assuming large cardinals a subset of the reals is "non-pathological" if and only if it is universally Baire. The theorem below makes this a sound mathematical assertion.

Theorem 1.2. *Assume there is a proper class of Woodin cardinals. Then:*

- ▷ *No universally Baire subset of* $\mathbb{R}$ *is a counterexample to* CH (*Davis, 1964; Feng, Magidor, and Woodin, 1992; Martin and Steel, 1989*).
- ▷ *Borel sets, analytic sets, projective sets,...are all universally Baire* (*Feng, Magidor, and Woodin, 1992*).
- ▷ *The universally Baire subsets of* $\mathbb{R}$ *are determined* (*Martin and Steel, 1989*).
- ▷ *The determined subsets of the reals are Lebesgue measurable, have the perfect set property,...*(*various authors one for each relevant regularity property*).

1.5. Independence of CH

CH is independent of the axioms of set theory:

- ▷ There is a model of the axioms of MK where CH holds (Gödel, 1947).
- ▷ There is a model of the axioms of MK where CH fails (Cohen, 1963).
- ▷ In the model of the axioms of MK where CH fails produced by Cohen, this failure can be witnessed by a Σ^1_2-set.

2. Gödel's program

Gödel (1947) wrote an influential survey on the continuum problem which has been a source of inspiration for the work of many logicians to come. In my opinion the best summary of its content can be given by quoting a few excerpts from the paper:

On the independence of the continuum problem (p. 520) *Only someone who (like the intuitionist) denies that the concepts and axioms of classical set theory have any meaning (or any well-defined meaning) could be satisfied with such a solution, not someone who believes them to describe some well-determined reality. For in this reality Cantor's conjecture must be either true or false, and its undecidability from the axioms as known today can only mean that these axioms do not contain a complete description of this reality;*

On Large Cardinals (p. 520) *For first of all the axioms of set theory by no means form a system closed in itself, but, quite on the contrary, the very concept of set on which they are based suggests their extension by new axioms which assert the existence of still further iterations of the operation "set of". These axioms can also be formulated as propositions asserting the existence of very great cardinal numbers or (which is the same) of sets having these cardinal numbers. The simplest of these strong "axioms of infinity" assert the existence of inaccessible numbers (and of numbers inaccessible in the stronger sense)* $> \aleph_0$.

On success as a criterion to detect new axioms (p. 521) *There might exist axioms so abundant in their verifiable consequences, shedding so much light upon a whole discipline, and furnishing such powerful methods for solving given problems (and even solving them, as far as that is possible, in a constructivistic way) that quite irrespective of their intrinsic necessity they would have to be assumed at least in the same sense as any well established physical theory.*

3. Large cardinals

Consider the universe of sets construed from the emptyset using all other axioms with the exception of *Infinity*; it can be shown that one ends up with a "baby" universe $H_{\aleph_0}$ characterized by the following property: if $X \in H_{\aleph_0}$, X is finite, all the elements Y of X are finite, all the elements Z of some element Y of X are finite,...More precisely

Definition 3.1. Given a set X,

- $\cup^0 X = X$, $\cup^{n+1} X = \cup(\cup^n X)$,
- $\mathrm{trcl}(X) = \bigcup_{n\in\mathbb{N}}(\cup^n X)$ is the *transitive closure* of X.

X is *hereditarily finite* if $\mathrm{trcl}(X)$ is finite.

$H_{\aleph_0}$ is the collection of hereditarily finite sets.

One can check that $\mathscr{P}\left(H_{\aleph_0}\right)$ (or -more precisely- the structure $\left(\mathscr{P}\left(H_{\aleph_0}\right), H_{\aleph_0}, \in\right)$) is a model of all axioms of Morse–Kelley set theory with the exception of the infinity axiom asserting that $\mathbb{N}$ is a set. Indeed in this model $H_{\aleph_0}$ is the class of all sets and $\mathbb{N} \subseteq H_{\aleph_0}$ is a "proper class".

In particular the axiom of Infinity is a way to assert that not all sets can be described using the construction principles and starting from the emptyset. $\mathbb{N}$ is an example of such a set.

Large cardinal axioms posit the existence of sets which cannot be described from $\mathbb{N}$ using the construction principles encoded in the axioms of MK. We already mentioned from Gödel the axiom stating the existence of inaccessible cardinals, *i.e.* cardinals behaving like $|\mathbb{N}|$, but of larger size: κ is inaccessible if and only if $\mathscr{P}(H_\kappa)$ is a model of Morse–Kelley set theory whose universe of sets is H_κ,[(10)] where:

Definition 3.2. Given a cardinal λ, a set X is *hereditarily of size less than* λ if $\mathrm{trcl}(X)$ has size less than λ.

H_λ is the set of all sets which are hereditarily of size less than λ.

The original proof by Wiles of Fermat's last theorem uses the notion of Grothendieck universe; when correctly formalized in set theory, the existence of a Grothendieck universe is equivalent to the existence of an inaccessible cardinal. Grothendieck's theory of universes finds its natural formulation in (it is actually equivalent to) set theory enriched with the axiom stating the existence of a proper class of inaccessible cardinals (McLarty, 2010).

Wiles' proof provides evidence grounded on Gödel's criterion of success for the adoption of large cardinal axioms. There are plenty of (less celebrated) such cases. For example:

Definition 3.3 (Vopenka's principle VP). For every *proper class* of **directed graphs with no loops**, there are two members of the class with a homomorphism between them.

This is an equivalent formulation of VP by Adámek and Rosický (1994). VP is a tool which category theorists employ successfully, see for example Bagaria et al., 2015; Casacuberta, Scevenels, and Smith, 2005; Rosický and Tholen, 2003.

VP entails the existence of all large cardinal axioms one may require in any of the results presented elsewhere in this note; for example if VP holds there is a proper class of Woodin cardinals. Henceforth assuming VP no counterexample to CH can be universally Baire, the projective sets of reals are universally Baire, the universally Baire sets are determined.

Later on we will delve more on the effects of large cardinals on second order number theory and on the family of topologically simple sets of reals.

[(10)] Equivalently κ is inaccessible if it is regular and $(H_\kappa, \in) \models \mathsf{ZFC}$.

4. Forcing axioms

Loosely speaking forcing axioms try to encapsulate the idea that the powerset of some set X is "as thick as possible". Forcing axioms for X can be divided in two categories:

Axioms of topological maximality They can be reformulated as strong forms of Baire's category theorem, are inspired by the notion of generic point, include MM^{++} among their instantiations.

Axioms of algebraic maximality They assert the closure of $\mathscr{P}(X)$ under a variety of set theoretic operations, are inspired by the notion of algebraic closure, include Woodin's axiom $(*)$ among their instationations.

We remark the following:

- ▷ MM^{++} and $(*)$ are forcing axioms for $X = \aleph_1$.
- ▷ Baire's category theorem is a "topological" forcing axiom for $X = \mathbb{N} = \aleph_0$.
- ▷ Large cardinals entail "algebraic" forcing axioms for $X = \mathbb{N} = \aleph_0$.

4.1. Topological maximality and Martin's maximum

Recall Baire's category theorem:

> Let (X, τ) be a compact Hausdorff space and $\{D_n : n \in \mathbb{N}\}$ be a countable family of dense open subsets of X. Then $\bigcap_n D_n$ is dense in X.

Let us parametrize the conclusion in all cardinals κ rather than just $\aleph_0$:

Definition 4.1. Let κ be an infinite cardinal and (X, τ) a topological space.

$\mathsf{FA}_\kappa(X, \tau)$ holds if $\bigcap_{i\in\kappa} D_i$ is *dense* in X for all $\{D_i : i \in \kappa\}$ family of dense open subsets of X.

For $\kappa > \aleph_0$ not all (X, τ) compact Hausdorff satisfy $\mathsf{FA}_\kappa(X, \tau)$. For example:

> Let Y be an *uncountable set* and (X, τ) be the Stone-Čech compactification of the product space $Y^{\mathbb{N}}$ where Y is endowed with the discrete topology. Then $\mathsf{FA}_{\aleph_1}(X, \tau)$ fails.

It can also be shown that $\mathsf{FA}_{\aleph_1}(X, \tau)$ holds for certain compact Hausdorff spaces (X, τ).[(11)]

[(11)] Actually an equivalent formulation of the axiom of choice states that $\mathsf{FA}_\kappa(X, \tau)$ holds for any compact Hausdorff space (X, τ) such that τ admits a base P of non-empty sets with the property that $(P, \subseteq)$ is a $< \kappa$-closed forcing (see for example Viale, 2017).

Abraham isolated a necessary condition on (X, τ) so that $\mathsf{FA}_{\aleph_1}(X, \tau)$ is not inconsistent.

Proposition 4.2 (Abraham). *Assume (X, τ) is a compact Hausdorff space which is not*[12] SSP. *Then $\mathsf{FA}_{\aleph_1}(X, \tau)$ fails.*

Foreman, Magidor, and Shelah (1988) showed that it can also be a sufficient condition:

Definition 4.3 (Foreman, Magidor, and Shelah, 1988). Martin's maximum $\mathsf{MM} \equiv \mathsf{FA}_{\aleph_1}(X, \tau)$ *holds for all compact Hausdorff spaces (X, τ) which are* SSP.

Theorem 4.4 (Foreman, Magidor, and Shelah, 1988). *Assume there exists a supercompact cardinals. Then there is a model of* MK *where* MM *holds.*

Furthermore one can strenghten the Theorem by asking the preservation of any sufficiently strong large cardinal assumption; for example if there is a model of Vopenka's principle in which there is a supercompact cardinal, then there is (a possibly different) model of Vopenka's principle and MM.

In particular the theorem establishes that (if one is eager to accept large cardinal axioms) there is a model of set theory enriched with any reasonable large cardinal axiom such that

> $\mathsf{FA}_{\aleph_1}(X, \tau)$ *holds for all compact Hausdorff spaces (X, τ) for which it is not impossible.*

4.2. What are the mathematical consequences of forcing axioms?

We collect here some of the major applications of forcing axioms in set theory and in other domains of mathematics.

Assume Martin's maximum holds. Then:

- ▷ CH is false and the continuum is the second uncountable cardinal, *i.e.* $2^{\aleph_0} = \aleph_2$ (Foreman, Magidor, and Shelah, 1988).
- ▷ Whitehead's conjecture on free groups is false, (*i.e.* there are uncountable Whitehead groups which are not free) (Shelah, 1974).
- ▷ Kaplansky's conjecture on Banach algebras holds (*i.e.* every algebra homomorphism from the Banach algebra $C(X)$ -where X is compact Hausdorff- into any other Banach algebra is necessarily continuous) (Woodin and Solovay, unpublished).[13]

[12] See Def. 9.1; however now we do not need to know what SSP means for now.

[13] A complete proof from published sources can be obtained combining results in Dales and Woodin, 1987, Chapter 3 with Todorčević, 1989, Thm. 7.7.

- ▷ There are five uncountable linear orders such that any uncountable linear order contains an isomorphic copy of one of them (Moore, 2006)
- ▷ All automorphisms of the Calkin algebra are inner (Farah, 2011).
- ▷ ...

All these conclusions are independent of MK+*Vopenka's principle* (or any other large cardinal assumption).

4.3. MM^{++}

MM^{++} is a natural technical strengthening of MM. For the sake of completeness we state one of its possible definitions; we caution the reader that it will make sense only for those familiar with the forcing method and with the basic theory of $\mathscr{P}(\aleph_1)$. Those not willing to delve into set theoretic technicalities may skip to the next section. This axiom also appears in Foreman, Magidor, and Shelah (1988), as well as all the results of this section, unless otherwise specified.

Definition 4.5. Given a complete boolean algebra B, $\dot{S} \in V^{\mathsf{B}}$ is a B-name for a stationary subset of $\aleph_1$, if $[\![\dot{S}$ is a stationary subset of $\check{\aleph}_1]\!]_{\mathsf{B}} = 1_{\mathsf{B}}$.

Given G ultrafilter on B, we let $\dot{S}_G = \{\alpha < \aleph_1 : [\![\check{\alpha} \in \dot{G}]\!]_{\mathsf{B}} \in G\}$.

Given C club subset of $\aleph_1$, we let $D_{C,\dot{S}} = \{G \in \mathsf{St}(\mathsf{B}) : \dot{S}_G \cap C \neq \emptyset\}$.

Note that $D_{C,\dot{S}}$ is dense open in $\mathsf{St}(\mathsf{B})$ for any C club subset of $\aleph_1$.

Recall that (X,τ) is a compact extremally disconnected Hausdorff space if and only if it is the Stone space of its algebra of regular open sets.

Proposition 4.6. *For a compact Hausdorff space (X,τ), $\mathsf{FA}_{\aleph_1}(X,\tau)$ holds if and only if so does $\mathsf{FA}_{\aleph_1}(\mathsf{St}(\mathsf{B}),\tau_{\mathsf{B}})$; where B is the algebra of regular open subsets of X and τ_{B} is the compact Hausdorff topology on the Stone space $\mathsf{St}(\mathsf{B})$ of ultrafilters on B induced by the Stone duality applied to B.*

Definition 4.7. Given a compact extremally disconnected Hausdorff space (X,τ), we let B be the complete boolean algebra given by its regular open sets.

$\mathsf{FA}^{++}_{\aleph_1}(X,\tau)$ holds if for any family $\{D_\alpha : \alpha < \aleph_1\}$ of dense open subsets of X, and any family $\{\dot{S}_\alpha : \alpha < \aleph_1\}$ of B-names for stationary subsets of $\aleph_1$, there is an ultrafilter $G \in X = \mathsf{St}(\mathsf{B})$ such that

$$G \in D_\alpha \text{ for all } \alpha < \aleph_1,$$

and

$$G \in D_{C,\dot{S}_\alpha} \text{ for all } \alpha < \aleph_1 \text{ and } C \text{ club subset of } \aleph_1.$$

Note that $2^{\aleph_1} > \aleph_1$ and there are $2^{\aleph_1}$-many clubs C on the cardinal $\aleph_1$, each one producing a different set $D_{C,\dot{S}_\alpha}$ for each $\alpha < \aleph_1$. Therefore $\mathsf{FA}^{++}_{\aleph_1}(X,\tau)$ requires the existence of points on X meeting families of dense open sets of size much larger than $\aleph_1$.

Definition 4.8. MM^{++} holds if $\mathsf{FA}^{++}_{\aleph_1}(X,\tau)$ holds for all compact extremally disconnected Hausdorff spaces which are SSP.

It is known (for example by a combination of Asperó and Schindler, 2021; Larson, 2008) that MM^{++} is an axiom strictly stronger than MM. However the standard proof of the consistency of Martin's maximum produces a model of MM^{++}, and we have the following:

Theorem 4.9. *Assume there is a supercompact cardinal. Then there is a model of* MK *with a where* MM^{++} *holds.*[(14)]

5. Forcing

We give in this section some basic information on the forcing method and its relation with the notion of Grothendieck Topos. The content of this section is not needed in the subsequent parts of this note.

We recall that Stone duality identifies complete boolean algebras with compact Hausdorff extremally disconnected spaces (a space is extremally disconnected if the closure of an open set is open).

Below we organized a text sketching on its left-side column the forcing procedure according to a set theorist, and highlighting on the right-side column the corresponding steps viewed with the lenses of a category theorist. The reader can skim through the left-side text, then through the right side text, and finally compare the paragraphs of the two texts with same alignment.

[(14)] The same considerations on the consistency of MM with large cardinal axioms apply to MM^{++}.

SET THEORIST	CATEGORY THEORIST
Given the complete boolean algebra B,	Given the compact extremally disconnected Hausdorff space $\mathrm{St}(\mathsf{B})$,
one forms	
the boolean valued model V^{B} by Cohen/Scott–Solovay–Vopenka;	the topos $\mathsf{Sh}(\mathrm{St}(\mathsf{B}), \mathsf{CompHaus})$, given by topological sheaves on $\mathrm{St}(\mathsf{B})$ with target a compact Hausdorff space;
one chooses	
a V-generic ultrafilter G on B,	a generic point of $\mathrm{St}(\mathsf{B})$ belonging to all dense open subset of $\mathrm{St}(\mathsf{B})$,
and one obtains	
the MK-model $V[G]$ generic extension of V by G.	the topos $\mathsf{Sh}(\mathrm{St}(\mathsf{B}), \mathsf{CompHaus})/_G$ of stalks at G of the sheaves in $\mathsf{Sh}(\mathrm{St}(\mathsf{B}), \mathsf{CompHaus})$.
The properties of	
$V[G]$ depend mainly on B	$\mathsf{Sh}(\mathrm{St}(\mathsf{B}), \mathsf{CompHaus})/_G$ depend mainly on $\mathrm{St}(\mathsf{B})$
and minimally on G.[15]	

We give some glimpses on how forcing assigns truth values to logical properties in a structure of the form $\mathscr{F}(X)/_G$ when $\mathscr{F}$ is a sheaf on a compact extremally disconnected Hausdorff space X, and G is a point of this space. The machinery of forcing generalizes to the topos of such sheaves this procedure.

Consider the space $L^{\infty+}(\mathbb{R})$ of measurable functions $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ which take value ∞ on a null set. For example $1/x$ is such a function and all essentially bounded measurable functions are also in $L^{\infty+}(\mathbb{R})$. Note that with pointwise multiplication and sum $L^{\infty+}(\mathbb{R})/_{\approx}$ is a commutative ring (where $\approx$ identifies functions overlapping an a conull set, and $f + g(x) = f \cdot g(x) = 0$ if either one of $f(x), g(x)$ is ∞).

[15] Note that generic points do not exist for atomless boolean algebras, however it can be of help for our intuition to work under the assumption that such points exist and that -on average- the points of $\mathrm{St}(\mathsf{B})$ behave like generic points. A strictly formal approach to the topic can eliminate the use of generic points in the forcing analysis of the properties of V^{B} (or equivalently of $\mathsf{Sh}(\mathrm{St}(\mathsf{B}), \mathsf{CompHaus})$).

Let MALG be the complete boolean algebra given by Lebesgue measurable sets modulo null sets. Given a universally Baire (hence Lebesgue measurable) relation $R \subseteq \mathbb{R}^n$ we can lift it to a boolean relation

$$R^{\mathsf{MALG}} : L^{\infty+}(\mathbb{R})^n \to \mathsf{MALG}$$

by setting

$$R^{\mathsf{MALG}}(f_1, \dots, f_n) = [\{x \in \mathbb{R} : R(f_1(x), \dots, f_n(x))\}],$$

where $[A]$ is the equivalence class in MALG of the measurable set $A \subseteq \mathbb{R}$. For example

$$\sin(x) <^{\mathsf{MALG}} \cos(x) = [\bigcup_{n \in \mathbb{Z}} ((2n-1) \cdot \pi + \frac{\pi}{4}; 2n \cdot \pi + \frac{\pi}{4})]$$

gets a MALG-value which is neither true nor false.

A key observation is that whenever $R, S \subseteq \mathbb{R}^n$ are universally Baire (Lebesgue measurable), so are $R \cap S, R \cup S, \mathbb{R}^n \setminus S$. Furthermore (assuming the existence of a proper class of Woodin cardinals) so is also $\pi_j[R]$ where π_j is the projection on the j-th coordinate of R.

This gives that

$$\begin{aligned}(R \cap S)^{\mathsf{MALG}}(f_1, \dots, f_n) &\\ &= [\{x \in \mathbb{R} : R(f_1(x), \dots, f_n(x))\} \cap \{x \in \mathbb{R} : S(f_1(x), \dots, f_n(x))\}] \\ &= R^{\mathsf{MALG}}(f_1, \dots, f_n) \wedge S^{\mathsf{MALG}}(f_1, \dots, f_n)\end{aligned}$$

Similarly it can be shown (appealing again to universal Baireness) that:

$$(\mathbb{R}^n \setminus S)^{\mathsf{MALG}}(f_1, \dots, f_n) = \neg_{\mathsf{MALG}} S^{\mathsf{MALG}}(f_1, \dots, f_n),$$

$$(R \cup S)^{\mathsf{MALG}}(f_1, \dots, f_n) = R^{\mathsf{MALG}}(f_1, \dots, f_n) \vee S^{\mathsf{MALG}}(f_1, \dots, f_n),$$

$$(\pi_n[R])^{\mathsf{MALG}}(f_1, \dots, f_{n-1}) = \bigvee_{\mathsf{MALG}} \left\{ R^{\mathsf{MALG}}(f_1, \dots, f_{n-1}, g) : g \in L^{\infty+}(\mathbb{R}) \right\}.$$

Now when G is an ultrafilter on MALG, we can pass to the quotient structure $L^{\infty+}(\mathbb{R})/_G$ given by the classes

$$[f]_G = \{h : [f = h] \in G\}$$

and relations

$$R/_G([f_1]_G, \dots, [f_n]_G)$$

holding when $R^{\mathsf{MALG}}(f_1, \dots, f_n) \in G$.

One can check (among many things) that the axioms of ordered field holds in $L^{\infty+}(\mathbb{R})/_G$, as G makes a coherent selection of truth values. For example given $f, g, h \in L^{\infty+}(\mathbb{R})$ it cannot be the case that the three boolean values $(f <^{\mathsf{MALG}} g)$, $(g <^{\mathsf{MALG}} h)$, $(h <^{\mathsf{MALG}} f)$ are all simultaneously in G. Similar arguments yield that the relation $<^{\mathsf{MALG}}/_G$ defines a dense linear order without endpoints on $L^{\infty+}(\mathbb{R})/_G$.

Letting $f \approx g$ iff $[\{x : f(x) = g(x)\}] = 1_{\mathsf{MALG}}$, $L^{\infty+}(\mathbb{R})/_\approx$ is the space of global sections with regard to the sheafification according to the dense Grothendieck topology of the presheaf $\mathscr{F}$ which to any $[A] \in \mathsf{MALG}$ assigns the family of essentially bounded measurable functions defined on A modulo null sets. $L^{\infty+}(\mathbb{R})/_G$ is then the stalk at G of this space of global sections (see Pierobon and Viale, 2020 for details on this example).

Cohen's forcing method applied to MALG devises a procedure which generalizes the above simultaneously to the whole topos $\mathsf{Sh}(\mathsf{St}(\mathsf{MALG}), \mathsf{CompHaus})$ (which in forcing is described as the structure V^{MALG}) producing a MALG-valued model of set theory. The procedure is modular and can be applied with input any complete boolean algebra B (or any extremally disconnected compact Hausdorff space), and links the logical properties of the B-valued structure V^{B} to the combinatorial properties of B. By suitably choosing B, Cohen was able to produce a B-valued model of set theory where CH gets value 0_{B}, or (equivalently) such that CH is false in $V^{\mathsf{B}}/_G$.

6. Algebraic maximality and model companionship

It is now time to make a detour in model theory and investigate the notion of algebraic closure and its possible transfers/generalizations to other mathematical theories.[16]

6.1. Algebraic closure of structures

Consider the signature $\{+, \cdot, 0, 1\}$. The table below shows on the left column the algebraic theory under consideration, in the central column the first order axiomatization in the above signature, in the right column a standard model of the axioms.

The table describes some features of the passages from $\mathbb{N}$ to $\mathbb{Z}$, from $\mathbb{Z}$ to $\mathbb{Q}$, from $\mathbb{Q}$ to $\mathbb{C}$. On the semantic level these passages came along with the introduction of new operations, and the construction of structures closed under these operations (additive inverses, multiplicative inverses, closure under solutions of polynomial equations), examples of which are listed in the left-side column. On the syntactic side the cell detected by the intersection of the first row with the central column lists the universal axioms common to all theories; the closure under the new operations of the structures on the left column is reflected by the satisfaction of corresponding Π_2-axioms, as listed in the corresponding cells of the central column.

[16] Standard model theory textbooks are Chang and Keisler, 1990; Hodges, 1997; Marker, 2002; Tent and Ziegler, 2012.

Structures	Axioms	Example
Commutative semirings with no zero divisors	$\forall x, y\,(x \cdot y = y \cdot x)$ $\forall x, y, z\,[(x \cdot y) \cdot z = x \cdot (y \cdot z)]$ $\forall x\,(x \cdot 1 = x \wedge 1 \cdot x = x)$ $\forall x, y\,(x + y = y + x)$ $\forall x, y, z\,[(x + y) + z = x + (y + z)]$ $\forall y\,(x + 0 = x \wedge 0 + x = x)$ $\forall x, y, z\,[(x + y) \cdot z = (x \cdot y) + (x \cdot z)]$ $\forall x, y\,[x \cdot y = 0 \to (x = 0 \vee y = 0)]$	$\mathbb{N}$
Integral domains	$\forall x \exists y\,(x + y = 0)$	$\mathbb{Z}$
Fields	$\forall x\,[x \neq 0 \to \exists y\,(x \cdot y = 1)]$	$\mathbb{Q}$
Algebraically closed fields	for all $n \geqslant 1$ $\forall x_0 \ldots x_n \exists y \sum x_i \cdot y^i = 0$	$\mathbb{C}$

Roughly the more closed-off a semiring with no zero-divisors is, the more Π_2-axioms it satisfies.

Robinson came up with model theoretic notions giving an abstract description of the closed-off structures for an arbitrary universal theory.

6.2. Existentially closed structures and model companionship

▷ A vocabulary τ is a list of predicates, constants and function symbols (we are accustomed to the vocabulary $\{+, \cdot, 0, 1, <, =\}$ which we can use to write down diophantine (in)equations, such as $x + 2y < xz + 5w$). Associated to a vocabulary τ we naturally have a notion of atomic τ-formula (which for $\{+, \cdot, 0, 1, <, =\}$ are exactly the diophantine (in)equations).

▷ A τ-formula $\phi(x_1, \ldots, x_n)$ is *quantifier free* if it is a boolean combination of *atomic* formulae (*i.e.* obtained by atomic formulae using conjunction $\wedge$, disjunction $\vee$, negation $\neg$, implication $\to$); for example $(x + y = 1 + 1 + z) \wedge \neg(x \cdot y < z \cdot z)$ is quantifer free for $\tau = \{\cdot, +, 0, 1, <\}$.

▷ A τ-formula $\psi(x_1, \ldots, x_n)$ is a Σ_1-formula if it is of the form

$$\exists x_0, \ldots, x_k \, \phi(x_0, \ldots, x_k, x_{k+1}, \ldots, x_n)$$

with $\phi(x_0, \ldots, x_k, x_{k+1}, \ldots, x_n)$ quantifier free; for example

$$\exists x, y \, [(x + y = 1 + 1 + z) \wedge \neg(x \cdot y < z \cdot z)]$$

is Σ_1 for $\tau = \{\cdot, +, 0, 1, <\}$).

Example 6.1. In the vocabulary $\{+, \cdot, 0, 1\}$, the atomic formulae are *diophantine equations* and the *quantifier free formulae* with parameters in a ring $\mathscr{M}$ define the *constructible sets* (in the sense of algebraic geometry) of $\mathscr{M}$. Below a standard example of a quantifier free formula (with parameters in $\mathscr{M}$):

$$\bigvee_{j=1}^{l} \left[\bigwedge_{i=1}^{k_j} p_{ij}(a_1^{ij}, \ldots, a_{m_{ij}}^{ij}, x_1, \ldots, x_n) = 0 \wedge \bigwedge_{d=1}^{m_j} \neg q_{dj}(b_1^{dj}, \ldots, b_{k_{dj}}^{dj}, x_1, \ldots, x_n) = 0 \right]$$

with a_k^{ij}, b_k^{dj} elements of $\mathscr{M}$ and $p_{ij}(y_1, \ldots, y_{m_{ij}}, x_1, \ldots, x_n)$, $q_{dj}(z_1, \ldots, z_{k_{dj}}, x_1, \ldots, x_n)$ polynomials with coefficients in $\mathbb{N}$ (of degree 1 in the y_l, z_h-s).

Definition 6.2. Given a vocabulary τ and τ-structures[17] $\mathscr{M} \sqsubseteq \mathscr{N}$, $\mathscr{M} \prec_1 \mathscr{N}$ if every Σ_1-formula with parameters in $\mathscr{M}$ and true in $\mathscr{N}$ is true also in $\mathscr{M}$.

For example $\langle \mathbb{C}, +, \cdot, 0, 1 \rangle \prec_1 \langle \mathbb{C}[X], +, \cdot, 0, 1 \rangle$, but $\langle \mathbb{Z}, +, \cdot, 0, 1 \rangle \not\prec_1 \langle \mathbb{C}, +, \cdot, 0, 1 \rangle$.

Definition 6.3. Given a τ-theory S, a τ-structure $\mathscr{M}$ is S-ec if:

▷ there is a model of S $\mathscr{N} \sqsupseteq \mathscr{M}$,

▷ $\mathscr{M} \prec_1 \mathscr{N}$ for any $\mathscr{N} \sqsupseteq \mathscr{M}$ which models S.

Example 6.4. For S the $\{+, \cdot, 0, 1\}$-theory of *integral domains* the *algebraically closed fields* are exactly the S-ec models.

Definition 6.5. Given a τ-theory S, a τ-theory T is the *model companion* of S if TFAE for any τ-structure $\mathscr{M}$:

▷ $\mathscr{M}$ is a model of T,

▷ $\mathscr{M}$ is S-ec.

Example 6.6. The $\{+, \cdot, 0, 1\}$-theory of *integral domains* has the $\{+, \cdot, 0, 1\}$-theory of *algebraically closed fields* as its model companion.

[17] $\sqsubseteq$ denotes the substructure relation among the τ-structures

Definition 6.7. A τ-theory T is *model complete* if it is its own model companion, *i.e.* if $\mathscr{M} \prec_1 \mathscr{N}$ whenever[18] $\mathscr{M} \sqsubseteq \mathscr{N}$ are models of T.

Example 6.8. The $\{+, \cdot, 0, 1\}$-theory of *algebraically closed fields* is *model complete*.

In particular model companionship and model completeness describe a notion of algebraic closure which makes sense for an arbitrary mathematical theory, when axiomatized in a certain signature τ.

Note that the signature plays a crucial role in model companionship results; for example the theory of algebraically closed fields is the model companion of the theory of integral domains in signature $\{+, \cdot, 0, 1\}$.[19] It is not anymore so in the signature $\{+, \cdot, ^{-1}, 0, 1\}$ where to interpret $^{-1}$ in an integral domain we use the axiom:

$$\forall x \, [\exists y \, (x \cdot y = 1 \wedge x \cdot x^{-1} = 1) \vee (\neg \exists y \, (x \cdot y = 1) \wedge x^{-1} = 0)].$$

With this axiom we still get that

$$\mathscr{M} = \langle \mathbb{C}, +, \cdot, ^{-1}, 0, 1 \rangle \sqsubseteq \mathscr{N} = \langle \mathbb{C}[X], +, \cdot, ^{-1}, 0, 1 \rangle,$$

but not that $\mathscr{M} \prec_1 \mathscr{N} : \exists x \, (x \neq 0 \wedge x^{-1} = 0)$ is true in $\mathscr{N}$ (as witnessed by X), but false in $\mathscr{M}$.

In particular $\langle \mathbb{C}, +, \cdot, ^{-1}, 0, 1 \rangle$ is not anymore T-ec for T the theory of integral domains in signature $\{+, \cdot, ^{-1}, 0, 1\}$. Note however that the theory of algebraically closed fields is still model complete[20] even when formalized in signature $\{+, \cdot, ^{-1}, 0, 1\}$.

7. Algebraic closure for set theory

We now want to use the notion of model companionship to describe what is "algebraic closure" for set theory. This will allow to formulate a reasonably accessible definition of Woodin's axiom $(*)$.

7.1. The right vocabulary for set theory

The standard axiomatization of set theory in textbooks is done in vocabulary $\{\in\}$, eventually with extra symbol $\subseteq$. However this is not giving an efficient formalization. For example in the $\{\in\}$-vocabulary the notion of ordered pair is formalized by means of *Kuratowski's trick:* $\langle y, z \rangle$ is (coded by) the set $\{\{y\}, \{y, z\}\}$.

[18] $\mathscr{M} \prec \mathscr{N}$ holds if all formulae with parameters in $\mathscr{M}$ get the same truth value in the two structures (hence $\prec_1$ is weaker than $\prec$). Model completeness can equivalently be defined replacing $\prec_1$ by $\prec$.

[19] It is also the model companion of the theory of fields in this same signature.

[20] It is also still the model companion of the theory of fields in $\{+, \cdot, ^{-1}, 0, 1\}$.

The standard $\in$-formula expressing $x = \langle y, z\rangle$ is

$$\exists t \exists u\ [\forall w\,(w \in x \leftrightarrow w = t \vee w = u) \wedge \forall v\,(v \in t \leftrightarrow v = y) \wedge \forall v\,(v \in u \leftrightarrow v = y \vee v = z)].$$

We do not regard the concept of ordered pair as a complicated one. It should then be formalizable by a simple formula. Accordingly many other basic set theoretic concepts such as that of being an n-ary relation, a function, the emptyset, the set of natural numbers,...should all be formalizable by simple formulae.

The lightface Δ_0-properties isolate the simplest set theoretic properties and include the following classes:

- $\{R \in V : R \text{ is an } n\text{-ary relation}\}$,
- $\{f \in V : f \text{ is a function}\}$,
- $\{\langle a, b\rangle : a \subseteq b \text{ are sets}\}$,
- ...

Definition 7.1. (Kunen, 1980, Def. IV.3.5)

An $\{\in\}$-formula ϕ is a Δ_0-formula if all its quantified variables are bounded to range in a set (e.g. $y \subseteq z \equiv \forall x\,(x \in y \to x \in z) \equiv \forall x \in y\,(x \in z)$ is a Δ_0-formula).

A lightface Δ_0-property is a class of sets whose extension is given by a Δ_0-formula.

It is thus natural to expand the vocabulary of set theory as follows:

Definition 7.2.

$\in_{\Delta_0}$ has the following list of symbols:

- constants for $\emptyset, \mathbb{N}$,
- relation symbols R_ϕ for any Δ_0-formula $\phi(x_1, \ldots, x_n)$,
- function symbols G_i for the list of ten Gödel operations given in Jech, 2003, Def. 13.6.

The base $\in_{\Delta_0}$-theory for sets interprets the new symbols according to their expected meaning, for example it must include the following axioms:

- $\forall \vec{x}\,[R_\phi(\vec{x}) \leftrightarrow \phi(\vec{x})]$ whenever ϕ is a Δ_0-formula
- $\forall \vec{x}, y\,(R_{\phi_i}(\vec{x}, y) \leftrightarrow G_i(\vec{x}) = y)$ if $\phi_i(\vec{x}, y)$ is the Δ_0-formula describing the graph of the Gödel operation G_i.
- $\forall x\,(x \notin \emptyset)$.
- $\forall x\,(x \in \mathbb{N} \leftrightarrow x \text{ is a finite Von Neumann ordinal})$.

Note that $(x$ is a finite Von Neumann ordinal$)$ is (formalizable by) a Δ_0-formula.

We now give an explicit first order axiomatization of MK in signature $\in_{\Delta_0} \cup \{\mathsf{Set}, V\}$ where Set is a unary predicate symbol for the property of being a set and V is a constant which denotes the universe of all sets. We stipulate for the sake of readability that smallcase letters indicate sets, uppercase letters indicate classes. To the above list of $\in_{\Delta_0}$-axioms[21] we add:

Universal axioms

Extensionality $\forall X, Y\,[(X \subseteq Y \wedge Y \subseteq X) \leftrightarrow X = Y]$.

Comprehension (a) $\forall X\,(\mathsf{Set}(X) \leftrightarrow X \in V) \wedge \forall X\,(X \subseteq V)$.

Foundation

$$\forall F\,[(F \text{ is a function } \wedge \operatorname{dom}(F) = \mathbb{N}) \rightarrow \exists n \in \mathbb{N}\, F(n+1) \notin F(n)].$$

Existence Axioms

Emptyset $\mathsf{Set}(\emptyset)$,

Infinity $\mathsf{Set}(\mathbb{N})$.

Basic construction principles

Gödel operations For each $i = 1, \ldots, 10$

$$\forall x_0, \ldots, x_{k_i}[(\bigwedge_{i=0}^{k}(x_i \in V) \rightarrow \exists! z\,(z \in V \wedge R_{\phi_i}(\vec{x}, z))],$$

where ϕ_i is the Δ_0-formula whose extension is the graph of the Gödel operation G_i.

Separation $\forall P, x\,[\mathsf{Set}(x) \rightarrow \mathsf{Set}(P \cap x)]$.

Strong construction principles

Comprehension (b) For every $\in_{\Delta_0}$-formula $\psi(x_0, \ldots, x_n, \vec{Y})$

$$\forall \vec{Y}\, \exists Z\, \forall x\, [x \in Z \leftrightarrow (x \in V \wedge \exists x_0, \ldots, x_n (x = \langle x_0, \ldots, x_n \rangle \wedge \psi(x_0, \ldots, x_n, \vec{Y})))].$$

Replacement

$$\forall F, x\,[(F \text{ is a function } \wedge \mathsf{Set}(x) \wedge (x \subseteq \operatorname{dom}(F))) \rightarrow \mathsf{Set}(F[x])].$$

[21] To be picky, the above list of axioms should be reformulated so that quantifiers apply only to variables ranging over sets.

Powerset

$$\forall x\,[\mathsf{Set}(x) \to \exists! y\,[\mathsf{Set}(y) \wedge (\forall z\,(z \in y \leftrightarrow z \subseteq x))]].$$

Global Choice

$$\begin{array}{l}\forall F[\\ \quad F \text{ is a function} \wedge \forall x\,(x \in \mathrm{dom}(F) \to F(x) \neq \emptyset)\\ \to\\ \quad \exists G\,(G \text{ is a function} \wedge \mathrm{dom}(G) = \mathrm{dom}(F) \wedge \forall x\,(x \in \mathrm{dom}(G) \to G(x) \in F(x))\\ \quad]\end{array}$$

An $\in_{\Delta_0}$ model of MK is a "sorted" structure $(\mathscr{C}, V, \in_{\Delta_0})$, where $\mathscr{C}$ is the family of all classes, and V is an element of $\mathscr{C}$ whose extension is the subfamily of $\mathscr{C}$ given by sets. One can also check that in this case $(V, \in_{\Delta_0})$ is a model of the ZFC-axioms (the classical presentation of set theory which avoids the mention of proper classes).

With a certain degree of approximation, it is customary to denote a model $(\mathscr{C}, V, \in_{\Delta_0})$ by its second component V, as $\mathscr{C}$ could be recovered as the family of subclasses of V. However this is not entirely correct as there can be $(\mathscr{C}, V, \in_{\Delta_0})$, $(\mathscr{D}, V, \in_{\Delta_0})$ both models of MK with $\mathscr{C} \neq \mathscr{D}$. When confusion on this issue may arise we will be more explicit. In general we stick to the convention to denote a model of (some of the axioms of) MK by $(\mathscr{C}, V, \in_{\Delta_0})$ and when we write just $(V, \in_{\Delta_0})$ we are considering the substructure whose elements are sets and not proper classes.

7.2. The H_κs

A finite set may not be simple, for example to understand the singleton $\{\mathbb{R}\}$ we need to know $\mathbb{R}$. We want to stratify sets according to the cardinalities required to generate them. Recall (Def. 3.2 and Def. 3.1) that a set X is *hereditarily of size less than* (*at most*) κ if $\mathrm{trcl}(X)$ has size less than (at most) κ.

- $\{\mathbb{R}\}$ is not hereditarily countable (*i.e.* $\{\mathbb{R}\} \notin H_{\aleph_1}$);
- Any subset A of $\mathbb{N}$ is hereditarily countable (*i.e.* $A \in H_{\aleph_1}$);
- $\mathbb{Q}$ and $\mathbb{Z}$ as defined in any textbook are hereditarily countable (*i.e.* in $H_{\aleph_1}$);
- $\mathbb{R}$ and $\mathscr{P}(\mathbb{N})$ are subsets of $H_{\aleph_1}$ (but not elements!);
- $\mathscr{P}(\mathbb{N})$ is definable by the atomic $\in_{\Delta_0}$-formula $(x \subseteq \mathbb{N})$ in the structure $\langle H_{\aleph_1}, \in_{\Delta_0}\rangle$;
- similarly for $\mathbb{R}$ or for (a representative of the homeomorphism class of) any Polish space.

- ▷ $\mathscr{P}(\aleph_1)$ is definable by the atomic $\in_{\Delta_0}$-formula $(x \subseteq \aleph_1)$ in parameter $\aleph_1$ (the first uncountable ordinal) in the structure $\langle H_{\aleph_2}, \in_{\Delta_0}\rangle$,
- ▷ NS, the non-stationary ideal on $\aleph_1$, is Σ_1-definable in parameter $\aleph_1$ in the same structure (see Section 9 for a definition of NS).

$$H_{\aleph_0} \subseteq H_{\aleph_1} \subseteq H_{\aleph_2} \subseteq \cdots \subseteq H_{\kappa^+} \subseteq \ldots$$
$$V = \bigcup \{H_\lambda : \lambda \text{ an infinite cardinal}\}$$

$(\mathscr{P}(H_{\aleph_i}), H_{\aleph_i}, \in_{\Delta_0})$ for $i = 1, 2$ are models of all axioms of MK with the exception of powerset. The role of V in these models is played by $H_{\aleph_i}$.[22]

$(\mathscr{P}(H_\kappa), H_\kappa, \in_{\Delta_0})$ models MK if and only if κ is inaccessible.

7.3. Existentially closed structures for set theory

Theorem 7.3 (Levy absoluteness). *Let κ be an infinite cardinal.*
Then

$$\langle H_{\kappa^+}, \in_{\Delta_0}, A : A \subseteq \mathscr{P}(\kappa)\rangle \prec_1 \langle V, \in_{\Delta_0}, A : A \subseteq \mathscr{P}(\kappa)\rangle$$

We argue below that $H_{\aleph_1}$ and $H_{\aleph_2}$ provide natural candidates for existentially closed structures for set theory. The choice of whether to focus on $H_{\aleph_1}$ or $H_{\aleph_2}$ depends on the signature in which one formalizes set theory.

8. Algebraic maximality for $\mathscr{P}(\mathbb{N})$ and generic absoluteness

We connect here the notion of algebraic closure given by existentially closed model, to generic absoluteness results for second order arithmetic.

Theorem 8.1 (Shoenfield, 1961). *Let $V[G]$ be a forcing extension of V. Then*[23]

$$\langle H_{\aleph_1}, \in_{\Delta_0}\rangle \prec_1 \langle H_{\aleph_1}^{V[G]}, \in_{\Delta_0}\rangle.$$

A key observation is that $\mathscr{P}(\mathbb{N})$ can be identified with Cantor's space $2^{\mathbb{N}}$ via characteristic functions. With this identification the universally Baire subsets[24] of $2^{\mathbb{N}}$ describe a particularly nice σ-algebra contained in $\mathscr{P}(\mathscr{P}(\mathbb{N}))$.

[22] Similarly it can be shown that for $i = 1, 2$, $(H_{\aleph_i}, \in_{\Delta_0})$ are model of all axioms of ZFC with the exception of Powerset.

[23] Note that $\langle H_{\aleph_1}^{V[G]}, \in_{\Delta_0}\rangle \prec_1 \langle V[G], \in_{\Delta_0}\rangle$ follows from Levy Absoluteness applied in $V[G]$.

[24] Def. 1.1 was given only on $\mathbb{R}^k$ but makes sense for any uncountable Polish space X (a second countable topological space whose topology can be induced by a complete metric) which is locally compact. Actually the families of universally Baire subsets of uncountable locally compact Polish spaces X and Y can be identified modulo any Borel isomorphism existing between the two spaces (Feng, Magidor, and Woodin, 1992; Kechris, 1995).

UB^V denotes the family of universally Baire subsets of $2^{\mathbb{N}}$ existing in V. Given $V[G]$ forcing extension of V, each universally Baire set A in V has a canonical extension $A^{V[G]}$ to a universally Baire set of $V[G]$ (Feng, Magidor, and Woodin, 1992).

The following is an elaboration which rephrases in model theoretic terminology deep results of Feng, Magidor, Woodin, Steel, Martin (see Venturi and Viale, 2023a):

Theorem 8.2 (Woodin 1985 + Steel, Martin, 1989 + Feng, Magidor, Woodin, 1992 + V., Venturi, 2019). *Assume there is a proper class of Woodin's cardinals. Then the theory of the structure*

$$\langle H^V_{\aleph_1}, \in_{\Delta_0}, A^V : A \in \mathsf{UB}^V \rangle$$

is model complete and is the model companion of the theory of

$$\langle V[H], \in_{\Delta_0}, A^{V[H]} : A \in \mathsf{UB}^V \rangle$$

whenever $V[H]$ is some generic extension of V.

8.1. Algebraic maximality for $\mathscr{P}(\mathbb{N})$: a summary

The table below summarizes the effects of large cardinals on the algebraic closure properties of set theory relative to $\mathscr{P}(\mathbb{N})$ (or -better- to $H_{\aleph_1}$):

Theory	Degree of algebraic closure
MK	$\langle H_{\aleph_1}, \in_{\Delta_0}, A : A \in \mathsf{UB}^V \rangle$ is Σ_1-elementary in $\langle V[G], \in_{\Delta_0}, A^{V[G]} : A \in \mathsf{UB}^V \rangle$ for all generic extension $V[G]$ of V
MK+ large cardinals	$\langle H_{\aleph_1}, \in_{\Delta_0}, A : A \in \mathsf{UB}^V \rangle$ has a **model complete theory** which is the **model companion** of the theory of $\langle V[G], \in_{\Delta_0}, A^{V[G]} : A \in \mathsf{UB}^V \rangle$ for all generic extension $V[G]$ of V

9. Algebraic maximality for $\mathscr{P}(\aleph_1)$ and axiom $(*)$

The notions of club and stationarity are given for the sake of completeness. One could take the extension of these concepts as a blackbox and still get some useful insights on Woodin's axiom $(*)$.

- C is a club subset[(25)] of $\aleph_1$ if $\cup C = \aleph_1$ and for all $\beta \notin C$ there is $\alpha \in \beta$ such that $(\alpha, \beta] \cap C$ is empty (where $(\alpha, \beta]$ is given by those ordinals γ with $\alpha \in \gamma$ and $\gamma \in \beta$ or $\gamma = \beta$).

[(25)] We here represent $\aleph_1$ by the least Von Neumann ordinal in the equivalence class of $\aleph_1$.

- $S \subseteq \aleph_1$ is stationary if for all C club subset of $\aleph_1$ $S \cap C$ is non-empty.
- $\mathsf{NS} \subseteq \mathscr{P}(\aleph_1)$ is the ideal of non-stationary subsets of $\aleph_1$ (*i.e.* subsets disjoint from some club).

Definition 9.1. Let B be a complete boolean algebra. B is SSP if whenever $V[G]$ is a forcing extension of V by B

$$\langle H_{\aleph_2}, \in_{\Delta_0}, \mathsf{NS}^V \rangle \sqsubseteq \langle V[G], \in_{\Delta_0}, \mathsf{NS}^{V[G]} \rangle.$$

Definition 9.2 (Bagaria, 2000)**. Bounded Martin's maximum** BMM holds if whenever B is an SSP complete boolean algebra and $V[G]$ is a forcing extension of V by B

$$\langle H_{\aleph_2}, \in_{\Delta_0} \rangle \prec_1 \langle H^{V[G]}_{\aleph_2}, \in_{\Delta_0} \rangle.$$

Compare it with Shoenfield's theorem and Levy absoluteness.

Theorem 9.3 (Bagaria, 2000)**.** MM *implies* BMM.

Definition 9.4 (Woodin, 1999)**.** BMM^{++} holds if whenever B is an SSP complete boolean algebra and $V[G]$ is a forcing extension of V by B

$$\langle H_{\aleph_2}, \in_{\Delta_0}, \mathsf{NS}^V \rangle \prec_1 \langle H^{V[G]}_{\aleph_2}, \in_{\Delta_0}, \mathsf{NS}^{V[G]} \rangle.$$

Theorem 9.5 (Woodin, 1999)**.** MM^{++} *implies* BMM^{++}.

9.1. Applications of BMM

The consequences of bounded forcing axioms are almost the same of those of MM although in some cases new proofs had to be found.

Assume BMM. Then:

- $2^{\aleph_0} = \aleph_2 = \aleph_1^+$ (Todorčević, 2002).
- Whitehead's conjecture on free groups is false.
- Kaplansky's conjecture on Banach algebras holds.

On the other hand Moore's result on the existence of a five element basis for uncountable linear orders and Farah's result establishing that MM implies all automorphism of the Calkin algebra are inner are not known to follow from BMM^{++}.

9.2. Woodin's axiom $(*)$

We let UB^V denote the family of universally Baire subsets of $\mathscr{P}(\mathbb{N})$ existing in V.

Remark 9.6. Let B be an SSP complete boolean algebra and $V[G]$ be a forcing extension of V by B. Then

$$\langle H_{\aleph_2}, \in_{\Delta_0}, \mathsf{NS}, A : A \in \mathsf{UB}^V\rangle \sqsubseteq \langle H^{V[G]}_{\aleph_2}, \in_{\Delta_0}, \mathsf{NS}^{V[G]}, A^{V[G]} : A \in \mathsf{UB}^V\rangle.$$

It is not clear to me whether the following definition is due to Woodin or Shelah or Goldstern:

Definition 9.7 (Woodin, Shelah, Goldstern?)**.** $\mathsf{UB\text{-}BMM}^{++}$ holds if whenever B is an SSP complete boolean algebra and $V[G]$ is a forcing extension of V by B

$$\langle H_{\aleph_2}, \in_{\Delta_0}, \mathsf{NS}, A : A \in \mathsf{UB}^V\rangle \prec_1 \langle H^{V[G]}_{\aleph_2}, \in_{\Delta_0}, \mathsf{NS}^{V[G]}, A^{V[G]} : A \in \mathsf{UB}^V\rangle.$$

By variations of the results in Woodin (1999), one gets:

Theorem 9.8 (Woodin)**.** MM^{++} *implies* $\mathsf{UB\text{-}BMM}^{++}$.

$(*)_{\mathsf{UB}}$ is a natural strengthening of Woodin's axiom $(*)$.

Theorem 9.9 (Asperó and Schindler, 2021)**.** *Assume there is a proper class of Woodin cardinals. Then* $(*)_{\mathsf{UB}}$ *if and only if* $\mathsf{UB\text{-}BMM}^{++}$.

If one is interested in Woodin's axiom $(*)$, here is an equivalent reformulation of it:

Theorem 9.10 (Asperó and Schindler, 2021)**.** *Assume there is a proper class of Woodin cardinals. Then Woodin's axiom* $(*)$ *holds if and only if whenever* B *is an* SSP *complete boolean algebra and* $V[G]$ *is a forcing extension of* V *by* B

$$\langle H_{\aleph_2}, \in_{\Delta_0}, \mathsf{NS}, A : A \text{ is in } \mathscr{P}(\mathbb{R})^{L(\mathbb{R})^V}\rangle$$

is Σ_1*-elementary in*

$$\langle H^{V[G]}_{\aleph_2}, \in_{\Delta_0}, \mathsf{NS}^{V[G]}, A^{V[G]} : A \text{ is in } \mathscr{P}(\mathbb{R})^{L(\mathbb{R})^V}\rangle.$$

The original formulation of $(*)$ can be found in Woodin (1999). $(*)$ follows from $(*)_{\mathsf{UB}}$ once one notes that any set of reals definable in $L(\mathbb{R})$ is universally Baire (assuming the existence of a proper class of Woodin cardinals).

9.3. Woodin's axiom $(*)$ and model completeness for the theory of $H_{\aleph_2}$

Recall that ψ is a Π_2-sentence if it is of the form $\forall \vec{x}\, \exists \vec{y} \phi(\vec{x}, \vec{y})$ with $\phi(\vec{x}, \vec{y})$ quantifier free.

In signature $\in_{\Delta_0} \neg\mathsf{CH}$ can be formalized by the Π_2-sentence in parameter $\aleph_1$ (the first uncountable ordinal/cardinal):

$$\forall f\, [(\underbrace{f \text{ is a function}}_{\Delta_0(f)} \wedge \underbrace{\mathrm{dom}(f) = \aleph_1}_{\Delta_0(f,\aleph_1)}) \to \exists r\, (\underbrace{r \subseteq \mathbb{N}}_{\Delta_0(r,\mathbb{N})} \wedge \underbrace{r \notin \mathrm{ran}(f)}_{\Delta_0(r,f)})]$$

Recall that NS is saturated if the boolean algebra $\mathscr{P}(\aleph_1)/_{\mathsf{NS}}$ has only partitions of size at most $\aleph_1$.

We quote the following facts on the non-stationary ideal:

- ▷ Assume NS is saturated. Then it is precipitous.
- ▷ Assume MM. Then NS is saturated (Foreman, Magidor, and Shelah, 1988).
- ▷ NS is precipitous is consistent with CH.

Theorem 9.11 (Woodin, 1999). *Assume there is a proper class of supercompact cardinals,* Sealing[(26)], and NS is precipitous. *TFAE:*

- ▷ $(*)_{\mathsf{UB}}$ (*or* UB-BMM^{++}).
- ▷ *For any* Π_2*-sentence*[(27)] ψ *for* $\in_{\Delta_0} \cup \{\aleph_1, \mathsf{NS}\} \cup \left\{A : A \in \mathsf{UB}^V\right\}$

$$\langle H_{\aleph_2}, \in_{\Delta_0}, \aleph_1, \mathsf{NS}, A :\ A \in \mathsf{UB}^V \rangle \models \psi$$

if and only if

$$\psi \textbf{ is true in } H_{\aleph_2}^{V[G]} \textbf{ for some forcing extension } V[G] \textbf{ of } V.$$

Sealing can be removed if one replaces UB^V with $\mathscr{P}(\mathbb{R})^{L(\mathrm{Ord}^{\mathbb{N}})}$ in the formulation of UB-BMM^{++} and in all relevant spots (note that the sets of reals definable in $L(\mathrm{Ord}^{\mathbb{N}})$ are universally Baire assuming the existence of class many supercompact cardinals). The large cardinal assumptions are far stronger than needed.

Theorem 9.12 (Viale, 2022). *Assume there is a proper class of supercompact cardinals,* Sealing, and NS is precipitous. *TFAE:*

- ▷ $(*)_{\mathsf{UB}}$ (*or* UB-BMM^{++}).

(26) See Def. 9.13.

(27) Among which $\neg\mathsf{CH}$ and a strong form of $2^{\aleph_0} = \aleph_2$.

▷ *The theory T of the structure*

$$\mathscr{M} = \langle H_{\aleph_2}, \in_{\Delta_0}, \aleph_1, \mathsf{NS}, A : A \in \mathsf{UB}^V \rangle$$

is the ***model companion*** *of the theory S of the structure*

$$\langle V, \in_{\Delta_0}, \aleph_1, \mathsf{NS}, A : A \in \mathsf{UB}^V \rangle.$$

(*i.e. T is* ***model complete***)

▷ *Letting $S_{\forall\vee\exists}$ be the boolean combination of existential sentences which are in S, and ψ be a Π_2-sentence,*

$$\mathscr{M} \textbf{ models } \psi \textbf{ if and only } \psi + S_{\forall\vee\exists} \textbf{ is consistent}.$$

▷ *For any Π_2-sentences ψ*

$$\langle H_{\aleph_2}, \in_{\Delta_0}, \mathsf{NS}, A : A \in \mathsf{UB}^V \rangle \models \psi$$

if and only if

$$\psi \textbf{ is true in } H^{V[G]}_{\aleph_2} \textbf{ for some forcing extension } V[G] \textbf{ of } V.$$

Sealing can be removed if one replaces UB^V with $\mathscr{P}(R)^{L(\mathrm{Ord}^{\mathbb{N}})}$ in the formulation of UB-BMM^{++} and in the relevant spots.

9.4. Sealing

For the sake of completeness a form of Sealing sufficient to prove both theorems is the following:

Given $(\mathscr{D}, W, \in_{\Delta_0})$ transitive model of MK, let N^W be the set $\mathscr{P}\left(H_{\aleph_1}\right)^{L(\mathsf{UB})^W}$, where $L(\mathsf{UB})^W$ is the smallest transitive model of ZF containing UB^W.

Definition 9.13 (Woodin). **Sealing** holds in a model $(\mathscr{C}, V, \in_{\Delta_0})$ of MK if:

▷ there are class many Woodin cardinals;

▷ the theory T of $(N^V, H^V_{\aleph_1}, \in_{\Delta_0})$ is model complete;

▷ $(N^{V[G]}, H^{V[G]}_{\aleph_1}, \in_{\Delta_0})$ models T whenever $V[G]$ is a generic extension of V.

It is not known whether Sealing follows rightaway from large cardinals; however Woodin has established a strong form of consistency for sealing relative to large cardinals:

Theorem 9.14 (Woodin). *Assume V models κ is supercompact and there is a proper class of Woodin cardinals. Let $V[H]$ be a generic extension of V where κ is countable. Then sealing holds in $V[H]$.*

In particular assuming large cardinals, Sealing can be forced to be true, and once it is true in a forcing extension there is no way to force it to become false in a further forcing extension. See Larson, 2004, Section 3.4 for details.

9.5. Algebraic maximality for $\mathscr{P}(\aleph_1)$: a summary

The table below summarizes the effects of large cardinals and forcing axioms on the algebraic closure properties of set theory relative to $\mathscr{P}(\aleph_1)$ (or -better- to $H_{\aleph_2}$):

Theory	**Degree of algebraic closure**
MK	$\langle H^V_{\aleph_2}, \in_{\Delta_0}, \aleph_1^V, \mathsf{NS}, A^V : A \in \mathsf{UB}^V\rangle$ is a *substructure* of $\langle V[G], \in_{\Delta_0}, \aleph_1^{V[G]}, \mathsf{NS}^{V[G]}, A^{V[G]} : A \in \mathsf{UB}^V\rangle$ for all generic extension $V[G]$ of V by an SSP-forcing
MK+ forcing axioms	$\langle H^V_{\aleph_2}, \in_{\Delta_0}, \aleph_1^V, \mathsf{NS}^V, A^V : A \in \mathsf{UB}^V\rangle$ is a Σ_1-*substructure* of $\langle V[G], \in_{\Delta_0}, \aleph_1^{V[G]}, \mathsf{NS}^{V[G]}, A^{V[G]} : A \in \mathsf{UB}^V\rangle$ for all generic extension $V[G]$ of V by an SSP-forcing
MK+ large cardinal axioms	for all generic extension $V[G]$ of V the theories of $\langle V[G], \in_{\Delta_0}, \aleph_1^{V[G]}, \mathsf{NS}^{V[G]}, A^{V[G]} : A \in \mathsf{UB}^V\rangle$ have the same **model companion** theory
MK+ large cardinals + forcing axioms	for all generic extension $V[G]$ of V the theories of $\langle V[G], \in_{\Delta_0}, \aleph_1^{V[G]}, \mathsf{NS}^{V[G]}, A^{V[G]} : A \in \mathsf{UB}^V\rangle$ have as **model companion** the theory of $\langle H^V_{\aleph_2}, \in_{\Delta_0}, \aleph_1^V, \mathsf{NS}^V, A^V : A \in \mathsf{UB}^V\rangle$

References

Adámek, J. and Rosický, J. (1994). *Locally presentable and accessible categories*. Vol. 189. London Mathematical Society Lecture Note Series. Cambridge University Press, Cambridge, pp. xiv+316.

Asperó, D. and Schindler, R. (2021). "Martin's Maximum^{++} implies Woodin's axiom $(*)$", *Ann. of Math.* (2) **193** (3), pp. 793–835.

Bagaria, J. (2000). "Bounded forcing axioms as principles of generic absoluteness", *Arch. Math. Logic* **39** (6), pp. 393–401.

——— (2005). "Natural axioms of set theory and the continuum problem", in: *Proceedings of the 12th International Congress of Logic, Methodology, and Philosophy of Science*. King's College London Publications, pp. 43–64.

Bagaria, J. et al. (2015). "Definable orthogonality classes in accessible categories are small", *J. Eur. Math. Soc. (JEMS)* **17** (3), pp. 549–589.

Bukovský, L. (1965). "The continuum problem and powers of alephs", *Comment. Math. Univ. Carolinae* **6**, pp. 181–197.

Casacuberta, C., Scevenels, D., and Smith, J. H. (2005). "Implications of large-cardinal principles in homotopical localization", *Adv. Math.* **197** (1), pp. 120–139.

Chang, C.-C. and Keisler, H. J. (1990). *Model theory*. Third. Vol. 73. Studies in Logic and the Foundations of Mathematics. North-Holland Publishing Co., Amsterdam, pp. xvi+650.

Cohen, P. J. (1963). "The independence of the continuum hypothesis", *Proc. Nat. Acad. Sci. U.S.A.* **50**, pp. 1143–1148.

Dales, H. G. and Woodin, H. (1987). *An introduction to independence for analysts*. Vol. 115. London Mathematical Society Lecture Note Series. Cambridge University Press, Cambridge, pp. xiv+241.

Davis, M. D. (1964). "Infinite games of perfect information", in: *Advances in Game Theory*. Princeton Univ. Press, Princeton, N.J., pp. 85–101.

Farah, I. (2011). "All automorphisms of the Calkin algebra are inner", *Ann. of Math.* (2) **173** (2), pp. 619–661.

Feng, Q., Magidor, M., and Woodin, H. (1992). "Universally Baire sets of reals", in: *Set theory of the continuum (Berkeley, CA, 1989)*. Vol. 26. Math. Sci. Res. Inst. Publ. Springer, New York, pp. 203–242.

Foreman, M., Magidor, M., and Shelah, S. (1988). "Martin's maximum, saturated ideals, and nonregular ultrafilters. I", *Ann. of Math.* (2) **127** (1), pp. 1–47.

Gödel, K. (1947). "What is Cantor's continuum problem?", *Amer. Math. Monthly* **54**, pp. 515–525.

Hodges, W. (1997). *A shorter model theory*. Cambridge University Press, Cambridge, pp. x+310.

Jech, T. (2003). *Set theory*. Springer Monographs in Mathematics. The third millennium edition, revised and expanded. Berlin: Springer, pp. xiv+769.

Kechris, A. S. (1995). *Classical descriptive set theory*. Vol. 156. Graduate Texts in Mathematics. Springer-Verlag, New York, pp. xviii+402.

Koellner, P. (2010). "On the question of absolute undecidability", in: *Kurt Gödel: essays for his centennial*. Vol. 33. Lect. Notes Log. Assoc. Symbol. Logic, La Jolla, CA, pp. 189–225.

Kunen, K. (1980). *Set theory*. Vol. 102. Studies in Logic and the Foundations of Mathematics. An introduction to independence proofs. Amsterdam: North-Holland, pp. xvi+313.

Larson, P. B. (2004). *The stationary tower*. Vol. 32. University Lecture Series. Notes on a course by W. Hugh Woodin. Providence, RI: American Mathematical Society, pp. x+132.

——— (2008). "Martin's maximum and definability in $H(\aleph_2)$", *Ann. Pure Appl. Logic* **156** (1), pp. 110–122.

Marker, D. (2002). *Model theory: An introduction*. Vol. 217. Graduate Texts in Mathematics. Springer-Verlag, New York, pp. viii+342.

Martin, D. A. and Steel, J. R. (1989). "A proof of projective determinacy", *J. Amer. Math. Soc.* **2** (1), pp. 71–125.

McLarty, C. (2010). "What Does It Take to Prove Fermat's Last Theorem? Grothendieck and the Logic of Number Theory", *Bulletin of Symbolic Logic* **16** (3), pp. 359–377.

Monk, J. D. (1969). *Introduction to set theory*. McGraw-Hill Book Co., New York-London-Sydney, pp. ix+193.

Moore, J. T. (2006). "A five element basis for the uncountable linear orders", *Ann. of Math. (2)* **163** (2), pp. 669–688.

Pierobon, M. and Viale, M. (2020). *Boolean valued models, presheaves, and étalé spaces*. arXiv: 2006.14852.

Rosický, J. and Tholen, W. (2003). "Left-determined model categories and universal homotopy theories", *Trans. Amer. Math. Soc.* **355** (9), pp. 3611–3623.

Shelah, S. (1974). "Infinite abelian groups, Whitehead problem and some constructions", *Israel J. Math.* **18**, pp. 243–256.

Tent, K. and Ziegler, M. (2012). *A course in model theory*. Cambridge University Press.

Todorčević, S. (1989). *Partition problems in topology*. Vol. 84. Contemporary Mathematics. American Mathematical Society, Providence, RI, pp. xii+116.

——— (2002). "Generic absoluteness and the continuum", *Math. Res. Lett.* **9** (4), pp. 465–471.

Venturi, G. and Viale, M. (2023a). "Second order arithmetic as the model companion of set theory", *Arch. Math. Logic* **62** (1-2), pp. 29–53.

Venturi, G. and Viale, M. (2023b). "What model companionship can say about the continuum problem", *The Review of Symbolic Logic*, pp. 1–40. eprint: 2204.13756.

Viale, M. (2017). "Useful axioms", *IfCoLog Journal of Logics and their Applications* **4** (10), pp. 3427–3462.

——— (2022). *Absolute model companionship, forcibility, and the continuum problem*. arXiv: 2109.02285.

Woodin, H. (1999). *The axiom of determinacy, forcing axioms, and the nonstationary ideal*. Vol. 1. de Gruyter Series in Logic and its Applications. Berlin: Walter de Gruyter & Co., pp. vi+934.

——— (2001a). "The continuum hypothesis. I", *Notices Amer. Math. Soc.* **48** (7), pp. 567–576.

——— (2001b). "The continuum hypothesis. II", *Notices Amer. Math. Soc.* **48** (7), pp. 681–690.

Matteo Viale

Dipartimento di Matematica
Università di Torino
Via Carlo Alberto 10 - 10125
Torino - Italie

E-mail: matteo.viale@unito.it

Séminaire BOURBAKI Juin 2023
74e année, 2022–2023, no 1208, p. 417 à 452
doi : 10.24033/ast.1217

NON-UNICITÉ DES SOLUTIONS DU SYSTÈME DE NAVIER–STOKES AVEC TERME SOURCE
[*d'après* **Dallas Albritton, Elia Brué et Maria Colombo**]

par **Anne-Laure Dalibard**

1. Introduction

Le système de Navier–Stokes décrit l'évolution des fluides visqueux incompressibles. Il traduit la conservation locale de la quantité de mouvement, et s'écrit, en dimension d et lorsque le fluide est soumis à une force extérieure,

$$\begin{aligned} \partial_t u + (u\cdot\nabla)u + \nabla p - \nu\Delta u = f \qquad & t>0,\ x\in\mathbb{R}^d, \\ \operatorname{div} u = 0 \qquad & t>0, x\in\mathbb{R}^d, \\ u(t=0,x) = u_0(x) \qquad & \forall x\in\mathbb{R}^d, \end{aligned} \tag{1}$$

où $u\colon [0,+\infty[\times\mathbb{R}^d\to\mathbb{R}^d$ désigne le champ de vitesse du fluide, $p\colon [0,+\infty[\times\mathbb{R}^d\to\mathbb{R}$ le champ de pression, et $f\colon [0,+\infty[\times\mathbb{R}^d\to\mathbb{R}^d$ la force appliquée au fluide. Les opérateurs différentiels ∇, div et Δ agissent ici sur la variable spatiale $x\in\mathbb{R}^d$. La fonction $u_0\colon \mathbb{R}^d\to\mathbb{R}^d$ est la donnée initiale du champ de vitesse, et le paramètre $\nu>0$ est la viscosité du fluide. Lorsque $\nu=0$, c'est-à-dire lorsque les forces de viscosité sont absentes du fluide, le système porte le nom d'équation d'Euler, et ses propriétés mathématiques sont différentes. En effet, le terme de dissipation $-\nu\Delta u$ a un effet régularisant qui joue un rôle fondamental dans les théories d'existence et d'unicité des solutions. La force extérieure f et la donnée initiale u_0 sont des données du problème, tandis que (u,p) est l'inconnue. Au moins formellement, si la solution u est régulière et suffisamment décroissante à l'infini, on peut déterminer p en prenant la divergence de la première équation, *i.e.*

$$-\Delta p = \operatorname{div}\left((u\cdot\nabla)u\right).$$

Ce texte est consacré à un problème difficile, resté longtemps ouvert, et résolu récemment par Albritton, Brué et Colombo (2022) en s'appuyant sur les travaux de Vishik (2018a,b) (revisités par Albritton, Brué, Colombo et al., 2021) : *la non-unicité des solutions faibles (dites de Leray–Hopf) en dimension trois.*

Solutions de Leray : existence globale et non-unicité en dimension trois

Les solutions de Leray s'appuient sur l'observation suivante : si (u,p) est une solution régulière et décroissante à l'infini du système de Navier–Stokes (1), alors, en faisant le produit scalaire de (1) avec u et en intégrant par parties en espace, on obtient

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}^d}|u(t,x)|^2\,dx+\int_{\mathbb{R}^d}(u\cdot\nabla)u(t,x)\cdot u(t,x)\,dx\\ -\int_{\mathbb{R}^d}\operatorname{div}u(t,x)\;p(t,x)\,dx+\nu\int_{\mathbb{R}^d}|\nabla u(t,x)|^2\,dx\\ =\int_{\mathbb{R}^d}f(t,x)\cdot u(t,x)\,dx.\end{aligned}$$

On peut réécrire le terme d'advection $(u\cdot\nabla)u\cdot u$ comme $u\cdot\nabla(|u|^2/2)$. Une intégration par parties et la condition de divergence nulle mènent à

$$\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}^d}|u(t,x)|^2\,dx+\nu\int_{\mathbb{R}^d}|\nabla u(t,x)|^2\,dx=\int_{\mathbb{R}^d}f(t,x)\cdot u(t,x)\,dx. \tag{2}$$

Soit $T>0$ quelconque. Supposons que $f\in L^1([0,T],L^2(\mathbb{R}^d)^d)$. L'intégration en temps de (2) et l'utilisation du lemme de Grönwall mènent à l'inégalité d'énergie

$$\begin{aligned}&\|u\|_{L^\infty([0,T],L^2(\mathbb{R}^d)^d)}+\nu^{1/2}\|\nabla u\|_{L^2([0,T]\times\mathbb{R}^d,\mathscr{M}_d(\mathbb{R}))}\\ &\leqslant C\left(\|u_0\|_{L^2(\mathbb{R}^d)^d}+\|f\|_{L^1([0,T],L^2(\mathbb{R}^d)^d)}\right),\end{aligned} \tag{3}$$

où C est une constante universelle, indépendante de T et de ν. Il apparaît que l'espace $L^\infty_{\mathrm{loc}}(\mathbb{R}_+,L^2(\mathbb{R}^d)^d)\cap L^2_{\mathrm{loc}}(\mathbb{R}_+,H^1(\mathbb{R}^d)^d)$ est un espace fonctionnel naturel pour chercher des solutions ; on l'appellera « espace d'énergie » dans la suite de ce texte. La notion de solution faible introduite par Leray (1934) et généralisée ensuite par Hopf (1950) au cas de domaines bornés, s'appuie sur l'analyse précédente. On adopte les notations suivantes : si $a,b\in\mathbb{R}^d$, on note $a\otimes b$ la matrice de $\mathscr{M}_d(\mathbb{R})$ définie par $(a\otimes b)_{i,j}=a_ib_j$ pour $1\leqslant i,j\leqslant d$. Pour $M=(m_{ij})_{1\leqslant i,j\leqslant d}$, $N=(n_{i,j})_{1\leqslant i,j\leqslant d}\in\mathscr{M}_d(\mathbb{R})$, on note $M:N$ le produit scalaire canonique entre M et N, c'est-à-dire $M:N=\sum_{1\leqslant i,j\leqslant d}m_{ij}n_{ij}$.

Définition 1.1 (Solutions de Leray de (1)). Soit $T>0$ quelconque.
Soit $u_0\in L^2(\mathbb{R}^d)^d$ telle que $\operatorname{div}u_0=0$, et $f\in L^1([0,T],L^2(\mathbb{R}^d)^d)$. Soit $u\in L^\infty([0,T],L^2(\mathbb{R}^d)^d)\cap L^2(([0,T],H^1(\mathbb{R}^d)^d)$.

On dit que u est une solution de Leray de (1) sur l'intervalle $[0,T]$ si et seulement si, pour tout $\varphi\in C^1([0,T],H^1(\mathbb{R}^d)^d)$ tel que $\operatorname{div}\varphi=0$ et $\varphi(T,x)=0$ pour tout

$x \in \mathbb{R}^d$, on a

$$\begin{aligned} & -\int_0^T \int_{\mathbb{R}^d} u(s,x) \cdot \partial_s \varphi(s,x)\, ds\, dx - \int_{\mathbb{R}^d} u_0(x) \cdot \varphi(0,x)\, dx \\ = \quad & -\nu \int_0^T \int_{\mathbb{R}^d} \nabla u(s,x) : \nabla \varphi(s,x)\, ds\, dx \\ & + \int_0^T \int_{\mathbb{R}^d} u(s,x) \otimes u(s,x) : \nabla \varphi(s,x)\, ds\, dx + \int_0^T \int_{\mathbb{R}^d} f(s,x) \cdot \varphi(s,x)\, ds\, dx. \end{aligned}$$

Remarque 1.2. Cette définition peut être étendue au cas où $f \in L^1([0,T], L^2(\mathbb{R}^d)^d) + L^2([0,T], H^{-1}(\mathbb{R}^d)^d)$, où $H^{-1}(\mathbb{R}^d)$ est le dual de $H^1(\mathbb{R}^d)$. L'inégalité d'énergie (3) est alors légèrement modifiée, mais l'espace d'énergie reste le même.

Jean Leray (1934) a démontré l'existence globale de solutions faibles du système de Navier–Stokes (1) en dimension trois ; la preuve s'étend aisément au cas de la dimension deux :

Théorème 1.3. *On suppose que $d \in \{2,3\}$. Soit $u_0 \in L^2(\mathbb{R}^d)^d$ telle que* $\operatorname{div} u_0 = 0$, *$f \in L^1_{\mathrm{loc}}(\mathbb{R}_+, L^2(\mathbb{R}^d)^d)$. Alors pour tout $T > 0$, il existe une solution de Leray $u \in L^\infty([0,T], L^2(\mathbb{R}^d)^d) \cap L^2([0,T], H^1(\mathbb{R}^d)^d)$ du système de Navier–Stokes* (1) *sur l'intervalle $[0,T]$.*

Lions et Prodi (1959) ont démontré en dimension deux l'unicité des solutions de Leray. Cette propriété est liée à une particularité d'invariance par changement d'échelle de l'espace d'énergie $L^\infty(\mathbb{R}_+, L^2(\mathbb{R}^d)^d) \cap L^2(\mathbb{R}_+, H^1(\mathbb{R}^d)^d)$ lorsque $d = 2$, sur laquelle nous reviendrons dans la prochaine section (voir (8)). En dimension $d = 3$, la question de l'unicité des solutions faibles était restée ouverte depuis les travaux de Leray. Lorsque la donnée initiale est régulière — dans un sens que l'on précisera ultérieurement — on peut construire une unique solution locale en temps, par exemple par une méthode de point fixe. Une telle solution est appelée « solution forte ». De façon remarquable, on a alors un *principe d'unicité « fort-faible »* : si une solution forte existe, alors toutes les solutions de Leray issues de la même donnée initiale coïncident avec la solution forte. Dans ce cas, l'unicité des solutions de Leray est donc acquise sur le temps d'existence de la solution forte. Par conséquent, pour des données initiales et des termes sources réguliers, la question de l'unicité est reliée à celle de la formation de singularités en temps fini pour les solutions de (1), qui est un autre problème ouvert majeur de l'analyse mathématique des équations de la mécanique des fluides. Cependant, on considèrera dans ce texte des termes sources avec peu de régularité (typiquement, avec la régularité requise dans le théorème 1.3 ou dans la Remarque 1.2), pour lesquels il n'existe pas de solution forte [(1)]. La question de l'unicité ne peut donc se réduire au principe d'unicité « fort-faible ». Les travaux

[(1)] En effet, s'il en existait une, le terme source associé serait alors régulier...

de Albritton, Brué et Colombo (2022) que nous décrirons ici apportent précisément une réponse (négative) à la question de l'unicité en dimension trois :

Théorème 1.4 (Non-unicité des solutions de Leray avec terme source). *Il existe $T > 0$ et $f \in L^1([0,T], L^2(\mathbb{R}^3)^3)$ tels que le système de Navier–Stokes (1) admette deux solutions de Leray distinctes sur $[0,T] \times \mathbb{R}^3$ avec la donnée initiale $u_0 = 0$ et le terme source f.*

Remarque 1.5. L'unicité des solutions de Leray en dimension 3 en l'absence de terme source (*i.e.* pour $f = 0$) ou dans un domaine à bords demeurent des problèmes ouverts. La non-unicité des solutions de Navier–Stokes avec $f = 0$ dans l'espace $C([0,T], H^\beta(\mathbb{T}^3)^3)$ avec $\beta > 0$ a été obtenue par Buckmaster et Vicol (2019) par des méthodes d'intégration convexe, complètement différentes de celles présentées ici. Cependant les solutions ainsi construites sont loin d'avoir la régularité des solutions de Leray : Buckmaster et Vicol (2019) montrent que la vorticité $\omega = \nabla \wedge u$ des solutions obtenues par intégration convexe appartient à l'espace $C([0,T], L^1(\mathbb{T}^3)^3)$, mais elle n'est pas *a priori* de carré intégrable.

Schéma de la preuve

La méthode de preuve repose sur une stratégie voisine de celle de Guillod et Šverák (2017) et Jia et Šverák (2014, 2015), qui sera décrite dans la prochaine partie. L'idée fondamentale est de tirer parti de l'invariance par changement d'échelle du système (1) (voir (8) ci-dessous), de façon à transformer la question de la non-unicité des solutions en un problème d'instabilité spectrale. En effet, soit u une solution de Leray de (1), associée à un terme source f. En s'appuyant sur l'invariance par changement d'échelle du système de Navier–Stokes (1) (voir (8) dans la section 2), on introduit les variables auto-similaires $\tau = \ln t$, $\xi = x/\sqrt{t}$, et les fonctions $\mathscr{U}$, $\mathscr{F}$, $\mathscr{P}$ définies par

$$u(t,x) = \frac{1}{\sqrt{t}}\mathscr{U}(\tau,\xi), \quad f(t,x) = \frac{1}{t^{3/2}}\mathscr{F}(\tau,\xi), \quad p = \frac{1}{t}\mathscr{P}(\tau,\xi). \tag{4}$$

Dans ces nouvelles variables, le système (1) devient

$$\begin{aligned} \partial_\tau \mathscr{U} - \frac{1}{2}\left(1 + \xi\cdot\nabla_\xi\right)\mathscr{U} - \nu\Delta_\xi \mathscr{U} + (\mathscr{U}\cdot\nabla_\xi)\mathscr{U} + \nabla_\xi \mathscr{P} = \mathscr{F}, \\ \operatorname{div}\mathscr{U} = 0. \end{aligned} \tag{5}$$

La donnée initiale en $t = 0$ devient une donnée en $\tau = -\infty$. Pour $\overline{\mathscr{U}} \in H^1(\mathbb{R}^3)^3$, on note $\mathscr{H}_{\overline{\mathscr{U}}}$ l'opérateur linéarisé autour de $\overline{\mathscr{U}}$, *i.e.*

$$\mathscr{H}_{\overline{\mathscr{U}}} : \mathscr{U} \mapsto -\frac{1}{2}\left(1 + \xi\cdot\nabla_\xi\right)\mathscr{U} - \nu\Delta_\xi\mathscr{U} + \nabla_\xi\mathscr{P} + (\overline{\mathscr{U}}\cdot\nabla_\xi)\mathscr{U} + (\mathscr{U}\cdot\nabla_\xi)\overline{\mathscr{U}}, \tag{6}$$

où le gradient de pression $\nabla_\xi\mathscr{P}$ assure que $\mathscr{H}_{\overline{\mathscr{U}}}(\mathscr{U})$ est à divergence nulle. Le domaine de $\mathscr{H}_{\overline{\mathscr{U}}}$ est $\mathscr{D}(\mathscr{H}_{\overline{\mathscr{U}}}) = \{\mathscr{U} \in H^2(\mathbb{R}^3)^3,\ \operatorname{div}\mathscr{U} = 0,\ \xi\cdot\nabla_\xi\mathscr{U} \in L^2(\mathbb{R}^3)^3\}$.

Supposons qu'il existe un champ de vecteur $\overline{\mathscr{U}} \in H^2(\mathbb{R}^3)^3$ à divergence nulle possédant la propriété suivante :

(P) $\mathscr{H}_{\overline{\mathscr{U}}}$ admet une valeur propre λ de partie réelle strictement négative, relative à une fonction propre $\mathscr{V}_\lambda \in \mathscr{D}(\mathscr{H}_{\overline{\mathscr{U}}})$.

Considérons le terme source associé à $\overline{\mathscr{U}}$, c'est-à-dire

$$\overline{\mathscr{F}} := -\frac{1}{2}\left(1 + \xi \cdot \nabla_\xi\right)\overline{\mathscr{U}} - \nu\Delta_\xi\overline{\mathscr{U}} + (\overline{\mathscr{U}} \cdot \nabla_\xi)\overline{\mathscr{U}}.$$

Alors par construction, $\overline{\mathscr{U}}$ est une solution stationnaire de (5) pour le terme source $\overline{\mathscr{F}}$ et avec une pression nulle, tandis que $\overline{\mathscr{U}} + \Re(e^{-\lambda\tau}\mathscr{V}_\lambda)$ est une solution de (5) avec un terme source $\overline{\mathscr{F}} + O(e^{-2\Re(\lambda)\tau})$. On remarque que le terme de reste $O(e^{-2\Re(\lambda)\tau})$ est négligeable quand $\tau \to -\infty$ (ce qui correspond à l'asymptotique $t \to 0$). On s'attend donc à pouvoir construire une solution $\mathscr{U}_1$ de (5) pour le terme source $\overline{\mathscr{F}}$ sur un intervalle $]-\infty, \tau_0[$ avec $\tau_0 \in \mathbb{R}$, avec $\mathscr{U}_1$ de la forme

$$\mathscr{U}_1 := \overline{\mathscr{U}} + \Re(e^{-\lambda\tau}\mathscr{V}_\lambda) + \mathscr{W},$$

où $\mathscr{W}$ est un correcteur non linéaire, vérifiant $\mathscr{W} = O(e^{-2\Re(\lambda)\tau})$. En posant

$$u_1(t,x) = \frac{1}{\sqrt{t}}\,\mathscr{U}_1\left(\ln t, \frac{x}{\sqrt{t}}\right), \quad u_2(t,x) = \frac{1}{\sqrt{t}}\,\overline{\mathscr{U}}\left(\frac{x}{\sqrt{t}}\right), \tag{7}$$

on vérifie que u_1 et u_2 sont deux solutions distinctes de (1) sur l'intervalle $]0, e^{\tau_0}[$ avec la même donnée initiale $u_0 = 0$ et le même terme source

$$f(t,x) = \frac{1}{t^{3/2}}\,\overline{\mathscr{F}}\left(\frac{x}{\sqrt{t}}\right).$$

On remarque en particulier que

$$\begin{aligned}
\|u_2(t)\|_{L^2(\mathbb{R}^3)^3} &= t^{1/4}\|\overline{\mathscr{U}}\|_{L^2(\mathbb{R}^3)^3} \to 0 \text{ quand } t \to 0,\\
\|u_1(t)\|_{L^2(\mathbb{R}^3)^3} &\leqslant t^{1/4}\|\overline{\mathscr{U}}\|_{L^2(\mathbb{R}^3)^3} + O(t^{1/4-\Re(\lambda)}) \to 0 \text{ quand } t \to 0,\\
\text{et } \|f(t)\|_{L^2(\mathbb{R}^3)^3} &= t^{-3/4}\|\overline{\mathscr{F}}\|_{L^2(\mathbb{R}^3)^3},
\end{aligned}$$

de sorte que $f \in L^1([0,T], L^2(\mathbb{R}^3)^3)$. Le résultat de non-unicité du théorème 1.4 s'ensuit.

On constate que le point central de la preuve est l'identification d'un profil $\overline{\mathscr{U}}$ possédant la propriété d'instabilité spectrale **(P)** mentionnée plus haut. La majeure partie de ce manuscrit sera donc consacrée à cette question. Dans la seconde partie, on rappelle quelques généralités sur le système de Navier–Stokes (invariance d'échelle, solutions fortes, unicité fort-faible), et on esquisse les principaux points de la stratégie de Guillod et Šverák (2017) et Jia et Šverák (2014, 2015), en la comparant à

celle exposée ici. La troisième partie est consacrée à la preuve d'un résultat de Vishik (2018a,b), reprise dans Albritton, Brué, Colombo et al. (2021), portant sur l'instabilité spectrale de flots tourbillonnaires particuliers pour le système d'Euler 2d. La preuve du théorème 1.4 à proprement parler est donnée dans la quatrième et dernière section, et repose sur plusieurs arguments perturbatifs. Le tourbillon bidimensionnel de Vishik est d'abord relevé en un champ de vitesse tridimensionnel supporté dans un anneau. On montre que ce champ de vitesse est instable pour le système d'Euler 3d à condition que le rayon de l'anneau soit suffisamment grand, en s'appuyant sur la proximité entre le système d'Euler 2d et le système d'Euler 3d axisymétrique, lorsque la distance à l'axe de symétrie est grande. Enfin, on établit que ce champ de vitesse est instable pour le système de Navier–Stokes 3d lorsque que son amplitude est suffisamment grande, en traitant perturbativement le terme visqueux et le terme de transport $\mathscr{U} + \xi \cdot \nabla_\xi \mathscr{U}$. Ceci conclut la preuve de la propriété (**P**).

Remarque 1.6. Le terme de dissipation visqueuse $-\nu\Delta u$ joue un rôle stabilisant dans l'équation de Navier–Stokes (1), qui permet par exemple de montrer l'existence globale de solutions fortes à données petites. Pour en comprendre le mécanisme, on pourra s'inspirer de l'analogie avec l'équation différentielle

$$X'(t) + \alpha X(t) = X(t)^2, \quad X(0) = X_0,$$

avec $X\colon \mathbb{R}_+ \to \mathbb{R}$, $\alpha > 0$ et $X_0 \in \mathbb{R}$. On constate que si $|X_0| < \alpha$, cette équation admet une solution globale, qui vérifie en outre $X(t) = O(e^{-\alpha t})$. En revanche, si $X_0 > \alpha$, le terme de dissipation αX n'est pas suffisant pour éviter l'explosion.

On comprend, sur ce modèle jouet, qu'il convient de bien analyser les tailles relatives du terme non linéaire $(\mathscr{U} \cdot \nabla)\mathscr{U}$ (ou de sa version linéarisée $(\overline{\mathscr{U}} \cdot \nabla)\mathscr{U} + (\mathscr{U} \cdot \nabla)\overline{\mathscr{U}}$) et du terme de dissipation $-\nu\Delta\mathscr{U}$ au sein de l'opérateur $\mathscr{H}_{\overline{\mathscr{U}}}$. Lorsque $\overline{\mathscr{U}}$ est petit (disons dans $W^{1,\infty}$), on s'attend à ce que le terme de dissipation visqueuse l'emporte. Cela correspond au régime $|X_0| < \alpha$ pour le modèle jouet ci-dessus. En revanche, lorsque $\overline{\mathscr{U}}$ devient grand, le terme de dissipation visqueuse devient négligeable et l'advection domine. On bascule dans l'analogue du régime $|X_0| > \alpha$ pour le modèle jouet. Le système est alors proche, dans un certain sens, du système d'Euler. Bien évidemment, la notion de petitesse doit être quantifiée convenablement et dépend de ν. Ces considérations sont centrales dans la notion de nombre de Reynolds critique, que nous introduirons dans la prochaine section.

On pourrait se demander s'il est possible de construire un mode instable pour le système de Navier–Stokes 2d, en partant du tourbillon instable de Vishik pour le système d'Euler 2d et en traitant le terme de viscosité perturbativement. Bien que ce résultat ne figure pas dans la littérature, il est probable que les arguments présentés ici pour la preuve de la propriété (**P**) se généralisent au cas bidimensionnel. Cependant, contrairement au cas tridimensionnel, cette instabilité spectrale dans les variables auto-similaires n'aboutit pas à un résultat de non-unicité des solutions de

Leray (ce qui est heureux, puisque l'unicité des solutions de Leray en deux dimensions est un résultat bien connu, voir Lions et Prodi (1959)). En effet, les analogues des solutions u_1, u_2 pour $d = 2$ n'appartiennent pas à l'espace d'énergie, puisqu'on a dans ce cas $\|\nabla u_2(t)\|_{L^2(\mathbb{R}^2)^2} = t^{-1/2}\|\nabla \overline{U}\|_{L^2(\mathbb{R}^2)^2} \notin L^2([0,T])$.

Nous reviendrons sur ces questions dans la remarque 3.12.

Quelques éléments d'analyse spectrale : outils généraux et notations

La pierre angulaire du théorème 1.4 est le résultat d'instabilité spectrale **(P)**. Par ailleurs, comme expliqué plus haut, l'existence d'une valeur propre de partie réelle strictement négative pour $\mathscr{H}_{\overline{\mathscr{U}}}$ est obtenue par des arguments perturbatifs. On rappelle donc dans ce paragraphe quelques définitions et résultats qui seront utiles dans le reste de la preuve.

Soit H un espace de Hilbert, et $\mathbf{L}\colon D(\mathbf{L}) \subset H \to H$ un opérateur linéaire fermé, défini sur un domaine dense. On note $\rho(\mathbf{L})$ l'ensemble résolvant défini par

$$\rho(\mathbf{L}) := \{\lambda \in \mathbb{C},\ \lambda - \mathbf{L} \text{ est une bijection}\},$$

et $\sigma(\mathbf{L}) := \mathbb{C} \setminus \rho(\mathbf{L})$ le spectre. On définit le spectre essentiel $\sigma_{\text{ess}}(\mathbf{L})$ comme l'ensemble des $\lambda \in \sigma(\mathbf{L})$ tels que $\lambda - \mathbf{L}$ n'est pas un opérateur de Fredholm d'indice zéro. (On rappelle que d'autres conventions sont possibles pour définir le spectre essentiel.) Sur $\rho(\mathbf{L})$, on définit la résolvante par $R(\lambda, \mathbf{L}) := (\lambda - \mathbf{L})^{-1}$. On rappelle le résultat suivant, qui sera fréquemment utilisé dans ce texte : si $\mathbf{L}\colon D(\mathbf{L}) \subset H \to H$ est un opérateur linéaire fermé, et $\mathbf{K}\colon D(\mathbf{L}) \to H$ un opérateur compact, alors $\sigma_{\text{ess}}(\mathbf{L} + \mathbf{K}) = \sigma_{\text{ess}}(\mathbf{L})$. En effet, si $\lambda \in \mathbb{C} \setminus \sigma_{\text{ess}}(L)$, alors $\lambda - \mathbf{L}$ est un opérateur de Fredholm d'indice zéro. On en déduit (voir Kato, 2013, Chapitre IV, Théorème 5.26) que $\lambda - \mathbf{L} - \mathbf{K}$ est également un opérateur de Fredholm d'indice zéro, et par conséquent $\sigma_{\text{ess}}(\mathbf{L} + \mathbf{K}) \subset \sigma_{\text{ess}}(\mathbf{L})$. Par symétrie, on en déduit l'égalité des deux spectres essentiels.

On utilisera fréquemment les deux résultats suivants :

Lemme 1.7. *Soit $\mathbf{L}\colon D(\mathbf{L}) \subset H \to H$ un opérateur linéaire fermé, et soit $\mathbf{K}\colon D(\mathbf{L}) \to H$ un opérateur compact. On suppose qu'il existe $\mu \in \mathbb{R}$ tel que $\sigma_{ess}(\mathbf{L}) \subset \{\Re(\lambda) \geqslant \mu\}$, et que $\mathbf{L} + \mathbf{K}$ admet une valeur propre $\lambda_0 \in \mathbb{C}$ telle que $\Re(\lambda_0) < \mu$. On suppose enfin que si $\lambda \in \mathbb{C}$ est tel que $|\Re(\lambda)| \gg 1$ et $\Re(\lambda) < 0$, alors $\lambda \in \rho(\mathbf{L} + \mathbf{K})$.*

Alors λ_0 est un élément isolé de $\sigma(\mathbf{L} + \mathbf{K})$.

Démonstration. Tout d'abord, puisque $\mathbf{K}$ est compact, $\sigma_{\text{ess}}(\mathbf{L}+\mathbf{K}) = \sigma_{\text{ess}}(\mathbf{L}) \subset \{\Re(\lambda) \geqslant \mu\}$. De plus, dans chaque composante connexe U de $\mathbb{C} \setminus \sigma_{\text{ess}}(\mathbf{L} + \mathbf{K})$, on a l'alternative suivante :

- ▷ soit $U \subset \sigma(\mathbf{L} + \mathbf{K})$;
- ▷ soit $U \cap \sigma(\mathbf{L} + \mathbf{K})$ ne comporte que des points isolés.

Soit U_0 la composante connexe de $\mathbb{C} \setminus \sigma_{\rm ess}(\mathbf{L}+\mathbf{K})$ contenant λ_0. Alors $\{\Re(\lambda) < \mu\} \subset U_0$, et donc par hypothèse $U_0 \cap \rho(\mathbf{L}+\mathbf{K}) \neq \emptyset$. Le résultat du lemme découle alors de l'alternative ci-dessus. □

Lemme 1.8. *Soit $\mathbf{M}_\infty$ un opérateur linéaire fermé, $\mathbf{K}_\infty : D(\mathbf{M}_\infty) \to H$ un opérateur compact, et soit $(\mathbf{M}_n)_{n\in\mathbb{N}}$, (resp. $(\mathbf{K}_n)_{n\in\mathbb{N}}$) une suite d'opérateurs linéaires fermés (resp. compacts) définis sur $D(\mathbf{M}_\infty)$. On suppose que les hypothèses suivantes sont vérifiées :*

- ▷ *Il existe $\mu \in \mathbb{R}$ et une suite réelle $(\mu_n)_{n\in\mathbb{N}}$ telle que $\lim_{n\to\infty} \mu_n = \mu$, tels que*
$$\sigma(\mathbf{M}_n) \subset \{\Re(\lambda) \geqslant \mu_n\} \quad \forall n \in \mathbb{N}, \quad \sigma(\mathbf{M}_\infty) \subset \{\Re(\lambda) \geqslant \mu\};$$
- ▷ *Pour tout $f \in H$, $R(\lambda, \mathbf{M}_n)f \to R(\lambda, \mathbf{M}_\infty)f$ uniformément sur tout compact de $\rho(\mathbf{M}_\infty) \cap \{\Re(\lambda) < \mu\}$;*
- ▷ *$\mathbf{K}_n \to \mathbf{K}_\infty$ pour la norme d'opérateur.*

Soit λ_∞ une valeur propre isolée de $\mathbf{M}_\infty + \mathbf{K}_\infty$ telle que $\Re(\lambda_\infty) < \mu$, et soit $V \subset \{\Re(\lambda) < \mu\}$ un voisinage de λ_∞ tel que $V \cap \sigma(\mathbf{A}_\infty) = \{\lambda_\infty\}$.

Alors pour n suffisamment grand, $\mathbf{M}_n + \mathbf{K}_n$ admet une valeur propre dans V.

Démonstration. Dans toute la preuve, on pose $\mathbf{A}_n = \mathbf{M}_n + \mathbf{K}_n$, $\mathbf{A}_\infty = \mathbf{M}_\infty + \mathbf{K}_\infty$. On commence par observer que pour tout ensemble compact $C \subset \rho(\mathbf{A}_\infty) \cap \{\Re(\lambda) < \mu\}$, pour tout $f \in H$, $R(\lambda, \mathbf{A}_n)f \to R(\lambda, \mathbf{A}_\infty)f$ uniformément sur C. Pour cela on écrit tout d'abord, pour $\lambda \in C$, et pour n suffisamment grand (de sorte que $\Re(\lambda) < \mu_n$),

$$\lambda - \mathbf{M}_n - \mathbf{K}_\infty = (\lambda - \mathbf{M}_n)(I - R(\lambda, \mathbf{M}_n)\mathbf{K}_\infty).$$

Puisque $\mathbf{K}_\infty$ est un opérateur compact, on peut approcher $\mathbf{K}_\infty$ par une suite d'opérateurs de rang fini. On en déduit que $R(\lambda, \mathbf{M}_n)\mathbf{K}_\infty \to R(\lambda, \mathbf{M}_\infty)\mathbf{K}_\infty$, uniformément sur C et pour la norme d'opérateur. Par conséquent, $R(\lambda, \mathbf{M}_n + \mathbf{K}_\infty)f \to R(\lambda, \mathbf{A}_\infty)f$ pour tout $f \in H$, uniformément sur C. Enfin, en écrivant

$$\lambda - \mathbf{A}_n = (\lambda - \mathbf{M}_n - \mathbf{K}_\infty)(I - R(\lambda, \mathbf{M}_n + \mathbf{K}_\infty)(\mathbf{K}_\infty - \mathbf{K}_n)),$$

on obtient la convergence annoncée.

Soit Γ un contour fermé encerclant λ_∞, tel que $\Gamma \subset V \cap \rho(\mathbf{A}_\infty)$, orienté dans le sens trigonométrique. On définit alors les projecteurs de Riesz

$$\Pr_\Gamma(\mathbf{A}_\infty) = \frac{1}{2\pi i}\int_\Gamma R(\lambda, \mathbf{A}_\infty)d\lambda,$$
$$\Pr_\Gamma(\mathbf{A}_n) = \frac{1}{2\pi i}\int_\Gamma R(\lambda, \mathbf{A}_n)d\lambda.$$

L'opérateur $\Pr_\Gamma(\mathbf{A}_\infty)$ est le projecteur spectral sur $\ker(\lambda_\infty - \mathbf{A}_\infty)$, et est donc non trivial par hypothèse. D'après ce qui précède, $\Pr_\Gamma(\mathbf{A}_n)f \to \Pr_\Gamma(\mathbf{A}_\infty)f$ pour tout f dans H. On en déduit que $\Pr_\Gamma(\mathbf{A}_n)$ est non trivial pour n suffisamment grand. Comme $\sigma_{\rm ess}(\mathbf{A}_n) = \sigma_{\rm ess}(\mathbf{M}_n) \subset \{\Re(\lambda) \geqslant \mu_n\}$, V contient une valeur propre de $\mathbf{A}_n$. □

Notations générales

Les opérateurs seront notés en gras. Les lettres calligraphiques ($\mathscr{U}$, $\mathscr{G}$, etc.) désignent des objets tridimensionnels (champ de vitesse, opérateur d'Euler linéarisé, etc.), tandis que les lettres droites ($\overline{U}$, $\boldsymbol{G}$) désignent des objets bidimensionnels (à l'exception du tourbillon Ω, qui sera noté de la même façon en deux et en trois dimensions). L'opérateur d'Euler 2d (resp. 3d) linéarisé autour d'un champ $\overline{U}$ (resp. $\overline{\mathscr{U}}$) sera noté $\boldsymbol{G}_{\overline{U}}$ (resp. $\mathscr{G}_{\overline{\mathscr{U}}}$). L'opérateur de Navier–Stokes tridimensionnel linéarisé autour d'un champ $\overline{\mathscr{U}}$ dans les variables autosimilaires sera noté $\mathscr{H}_{\overline{\mathscr{U}}}$. Il sera souvent utile de distinguer les opérateurs linéarisés en formulation vitesse et en formulation vorticité (voir la partie suivante), et on notera les opérateurs correspondants $\boldsymbol{G}^{\mathrm{vel}}$, $\boldsymbol{G}^{\mathrm{vor}}$, etc.

2. Généralités sur le système de Navier–Stokes

Invariance par changement d'échelle et solutions fortes

L'équation de Navier–Stokes possède une invariance par changement d'échelle, qui joue un rôle crucial pour la construction de solutions fortes et des solutions auto-similaires (4). Plus précisément, soit u une solution des équations de Navier–Stokes (1) pour le terme source f et la donnée initiale u_0. On observe alors que pour tout $\lambda > 0$, la fonction u_λ définie par

$$u_\lambda(t,x) = \lambda u(\lambda^2 t, \lambda x) \tag{8}$$

est également solution de (1) pour le terme source $\lambda^3 f(\lambda^2 t, \lambda x)$ et la donnée initiale $\lambda u_0(\lambda x)$. En particulier, une solution stationnaire et invariante par le changement d'échelle est nécessairement homogène de degré -1. On remarque par ailleurs que cette invariance a guidé la recherche de solutions auto-similaires de la forme (4). Dès lors, il est intéressant d'analyser quelles sont les normes invariantes par ce changement d'échelle. En effet, la construction de solutions fortes repose souvent sur un argument de point fixe (de type Cauchy-Lipschitz). Par analogie avec les équations différentielles ordinaires (voir la remarque 1.6), on peut espérer démontrer des résultats de deux types : (i) existence locale et unicité de solutions fortes, pour des données initiales et des termes sources de tailles arbitraires ; (ii) existence globale et unicité de solutions fortes, pour des données initiales et des termes sources suffisamment petits. Si l'espace fonctionnel considéré n'est *pas* invariant par changement d'échelle, alors quitte à modifier artificiellement les données par un changement d'échelle, l'hypothèse de petitesse du (ii) est systématiquement vérifiée, ce qui semble peu raisonnable. Les espaces considérés pour les solutions fortes possèdent donc des normes invariantes par le changement d'échelle (8).

Il semble naturel de comparer la régularité des espaces dont les normes sont invariantes par changement d'échelle à celle de l'espace d'énergie de l'équation. Si la régularité de l'espace d'énergie est strictement supérieure à celle des espaces invariants par changement d'échelle, alors la situation est très favorable. En effet, la borne sur l'énergie fournit automatiquement un contrôle de la solution dans l'espace dans lequel on réalise le point fixe. La solution peut donc être définie pour tout temps. On parle en ce cas d'équation *sous-critique*. Au contraire, si la régularité de l'espace d'énergie est strictement inférieure à celle des espaces invariants par changement d'échelle, le contrôle de l'énergie ne donne aucune information sur une éventuelle explosion de la norme associée à l'argument de point fixe. On parle alors d'équation *sur-critique*. Lorsque la norme associée à l'espace d'énergie est invariante par le changement d'échelle, l'équation est dite *critique*.

Notons $\|\cdot\|_{\dot{H}^s(\mathbb{R}^d)}$, pour $s \in \mathbb{R}$ la norme de Sobolev homogène

$$\|u\|_{\dot{H}^s(\mathbb{R}^d)} = \left(\int_{\mathbb{R}^d} |\xi|^{2s} |\hat{u}(\xi)|^2 \, d\xi \right)^{1/2},$$

la fonction $\hat{u}$ étant la transformée de Fourier de u dans $L^2(\mathbb{R}^d)$. On renvoie à Bahouri, Chemin et Danchin (2011) pour une étude des espaces $\dot{H}^s(\mathbb{R}^d)$ et de leurs propriétés. En dimension deux, on vérifie que l'espace d'énergie $L^\infty(\mathbb{R}_+, L^2(\mathbb{R}^2)^2) \cap L^2(\mathbb{R}_+, \dot{H}^1(\mathbb{R}^2)^2)$ est invariant par (8). Dans ce cas, l'équation de Navier–Stokes est donc critique. Une conséquence fondamentale de cette invariance réside dans l'unicité des solutions de Leray en dimension deux. En revanche, en dimension trois, on vérifie que

$$\|u_\lambda\|_{L^\infty(\mathbb{R}_+, L^2(\mathbb{R}^3)^3)} + \|u_\lambda\|_{L^2(\mathbb{R}_+, \dot{H}^1(\mathbb{R}^3)^3)} = \lambda^{-1} \left(\|u\|_{L^\infty(\mathbb{R}_+, L^2(\mathbb{R}^3)^3)} + \|u\|_{L^2(\mathbb{R}_+, \dot{H}^1(\mathbb{R}^3)^3)} \right),$$

tandis que

$$\|u_\lambda\|_{L^\infty(\mathbb{R}_+, \dot{H}^{1/2}(\mathbb{R}^3)^3)} + \|u_\lambda\|_{L^2(\mathbb{R}_+, \dot{H}^{3/2}(\mathbb{R}^3)^3)} = \|u\|_{L^\infty(\mathbb{R}_+, \dot{H}^{1/2}(\mathbb{R}^3)^3)} + \|u\|_{L^2(\mathbb{R}_+, \dot{H}^{3/2}(\mathbb{R}^3)^3)}.$$

L'équation de Navier–Stokes en dimension trois est donc sur-critique. La situation est ainsi radicalement différente de la dimension deux, ce qui ouvre la porte à des résultats de non-unicité dans l'espace d'énergie comme le théorème 1.4.

Depuis les années 1960, différents auteurs ont montré les résultats (i) et/ou (ii) dans des espaces invariants d'échelle, dont nous donnons ici quelques exemples :

- ▷ Fujita et Kato, 1964 : on considère une donnée initiale u_0 dans l'espace $\dot{H}^{1/2}(\mathbb{R}^3)^3$, à divergence nulle, et un terme source f dans l'espace $L^2_{\text{loc}}(\mathbb{R}_+, \dot{H}^{-1/2}(\mathbb{R}^3)^3)$. L'espace des solutions est donné par l'image de $\dot{H}^{1/2}(\mathbb{R}^3)^3$ par le flot de la chaleur : autrement dit, on cherche des solutions dans $L^\infty_{\text{loc}}(\mathbb{R}_+, \dot{H}^{1/2}(\mathbb{R}^3)^3) \cap L^2_{\text{loc}}(\mathbb{R}_+, \dot{H}^{3/2}(\mathbb{R}^3)^3)$. Fujita et Kato (1964) ont démontré les points (i) et (ii) dans cet espace, ainsi qu'une propriété d'unicité fort-faible : si une telle solution forte existe sur un intervalle $[0, T]$ pour la donnée initiale $u_0 \in \dot{H}^{1/2}(\mathbb{R}^3)^3$, alors toutes les solutions de Leray pour la même donnée initiale u_0 coïncident avec elle sur l'intervalle $[0, T]$.

▷ Kato, 1984 : on considère une donnée initiale $u_0 \in L^3(\mathbb{R}^3)^3$, à divergence nulle, et un terme source $f = 0$ (pour simplifier). On remarquera que $\dot{H}^{1/2}(\mathbb{R}^3) \hookrightarrow L^3(\mathbb{R}^3)$. De nouveau, on cherche des solutions dans l'image de $L^3(\mathbb{R}^3)^3$ par le flot de la chaleur, c'est-à-dire telles que

$$u \in C(\mathbb{R}_+, L^3(\mathbb{R}^3)^3), \quad \sup_{t>0} t^{\alpha/2}\|u(t)\|_{L^p(\mathbb{R}^3)^3} < +\infty,$$
$$u \in L^r_{\mathrm{loc}}(\mathbb{R}_+, L^q(\mathbb{R}^3)^3) \quad \text{et} \quad \sup_{t>0} t^{1/2}\|\nabla u\|_{L^3(\mathbb{R}^3)^9} < +\infty,$$

où les exposants α, p, q et r sont tels que $\alpha = 1 - 3/p$ et $3 < p \leqslant 6$ d'une part, et $2/r = 1 - 3/q$ avec $3 < q < 9$ d'autre part. Kato (1984) a montré les propriétés (i) et (ii) dans cet espace ainsi qu'une propriété d'unicité fort-faible.

La propriété (ii) a ensuite été généralisée par Cannone (1997) et Planchon (1996) à des données initiales de taille arbitraire dans $L^3(\mathbb{R}^3)^3$, mais petites dans l'espace de Besov homogène $\dot{B}^{-1+\frac{3}{p}}_{p,\infty}$ avec $3 < p < \infty$. On renvoie à Bahouri, Chemin et Danchin (2011) pour une définition des espaces de Besov en Fourier, utilisant la décomposition de Littlewood–Paley, mais pour les besoins de ce texte, on pourra utiliser la caractérisation suivante (voir Cannone (1997, Lemme 1.1)) : pour $1 < p < \infty$, $\alpha > 0$,

$$\|u\|_{\dot{B}^{-\alpha}_{p,\infty}(\mathbb{R}^d)} = \sup_{t \geqslant 0} t^{\alpha/2}\|e^{t\Delta}u\|_{L^p(\mathbb{R}^d)}.$$

Mentionnons que $L^3(\mathbb{R}^3) \hookrightarrow \dot{B}^{-1+\frac{3}{p}}_{p,\infty}$, mais cette inclusion est stricte : en particulier, l'espace $\dot{B}^{-1+\frac{3}{p}}_{p,\infty}$ contient des fonctions homogènes de degré -1 non identiquement nulles.

▷ Koch et Tataru, 2001 : on considère une donnée initiale à divergence nulle, appartenant à l'espace BMO^{-1}, c'est-à-dire telle que

$$\|u_0\|_{BMO^{-1}} := \left(\sup_{x\in\mathbb{R}^3, R>0} |B(x,R)|^{-1} \int_0^{R^2} \int_{B(x,R)} |(e^{t\Delta}u_0)(y)|^2 \, dy \, dt \right)^{1/2} < +\infty.$$

De nouveau, Koch et Tataru (2001) ont démontré la propriété (ii) dans l'espace image de BMO^{-1} par le flot de la chaleur et pour un terme source $f = 0$; autrement dit, la solution appartient à l'espace X défini par

$$\|u\|_X = \sup_{t>0} t^{1/2}\|u(t)\|_{L^\infty(\mathbb{R}^3)^3} + \left(\sup_{x,R} |B(x,R)|^{-1} \int_{B(x,R)} |u(t,y)|^2 \, dy \, dt \right)^{1/2}. \quad (9)$$

On peut vérifier que BMO^{-1} contient des fonctions homogènes de degré -1, et que par ailleurs la solution u_2 donnée par (7) appartient à l'espace X.

Une conséquence immédiate des résultats d'unicité fort-faible mentionnés ci-dessus est la suivante : s'il existe deux solutions de Leray distinctes issues de la même donnée initiale, alors ces deux solutions appartiennent nécessairement aux complémentaires des espaces dans lesquels on a un principe d'unicité fort-faible. Avec les notations (7), on observe par exemple que

$$\|u_2(t)\|_{\dot{H}^{3/2}(\mathbb{R}^3)^3} = \frac{1}{\sqrt{t}}\|\overline{\mathscr{U}}\|_{\dot{H}^{3/2}(\mathbb{R}^3)^3},$$

de sorte que $u_2 \notin L^2_{\rm loc}(\mathbb{R}_+, \dot{H}^{3/2}(\mathbb{R}^3))$. Notons que cette propriété est précisément liée au fait que l'espace $L^2_{\rm loc}(\mathbb{R}_+, \dot{H}^{3/2}(\mathbb{R}^3))$ est invariant par changement d'échelle, tandis que la solution u_2 est définie à partir du changement d'échelle (8). En revanche, puisque la fonction u_2 donnée par (7) appartient à l'espace X de Koch et Tataru (2001), défini par (9), le théorème 1.4 entraîne immédiatement qu'il n'y pas d'unicité fort-faible dans l'espace X, du moins en présence d'un terme source. Par ailleurs, la fonction u_1 donnée par (7) appartient également à une version localisée en temps de l'espace X, à savoir

$$\|u\|_{X_T} = \sup_{t\in]0,T[} t^{1/2}\|u(t)\|_{L^\infty(\mathbb{R}^3)^3} + \left(\sup_{\substack{x\in\mathbb{R}^3,\\ R<\sqrt{T}}} |B(x,R)|^{-1}\int_0^{R^2}\int_{B(x,R)}|u(t,y)|^2\,dy\,dt\right)^{1/2}. \quad (10)$$

D'après le théorème 1.4, il n'y a pas unicité dans X_T des solutions du système de Navier–Stokes avec terme source, de sorte que l'hypothèse de petitesse de Koch et Tataru (2001, Theorème 3) semble nécessaire (on rappelle que les solutions u_1 et u_2 sont, par construction, grandes dans X_T). On reviendra sur ces questions dans le cas de la dimension deux dans la remarque 3.12.

Un enjeu crucial, tant du point de vue des mathématiques fondamentales que des applications physiques, est de comprendre si une donnée initiale régulière (disons, dans $\dot{H}^{1/2}(\mathbb{R}^3)^3$) et un terme source régulier (dans $L^2(\mathbb{R}_+, \dot{H}^{-1/2}(\mathbb{R}^3)^3)$) peuvent donner naissance à une solution singulière en un temps fini T^*, *i.e.* vérifiant $\lim_{t\to T^*, t<T^*}\|u(t)\|_{\dot{H}^{1/2}(\mathbb{R}^3)^3} = +\infty$. Cette question est au cœur d'un des Millenium Problems de la fondation Clay, et des questions de développement de la turbulence en physique. Elle est néanmoins distincte du problème qui nous occupe ici, puisque le terme source considéré n'appartient pas à $L^2_{\rm loc}(\mathbb{R}_+, \dot{H}^{-1/2}(\mathbb{R}^3)^3)$. De surcroît, les solutions u_1 et u_2 sont régulières sur $[\delta, T]\times\mathbb{R}^3$ pour $\delta > 0$: la singularité (et donc la non-unicité) se produit quand $t \to 0^+$. Au moment de l'écriture de ce texte, les scénarios potentiels d'explosion demeurent très ouverts. Rappelons toutefois que des résultats d'explosion sont connus pour d'autres systèmes fluides, comme le système d'Euler 3d (voir Elgindi, 2021) ou le système de Navier–Stokes compressible barotrope (voir Merle et al., 2022a,b, ainsi que l'exposé Bourbaki de Perelman, 2022).

Vocabulaire et notations

Les fonctions axisymétriques (*i.e.* invariantes par rotation autour de l'axe z) jouent un rôle particulier dans la preuve, le flot $\overline{\mathscr{U}}$ étant lui-même axisymétrique. On introduit donc quelques éléments de vocabulaire spécifique à ces fonctions, ainsi que quelques éléments de notation. On note $(r,\theta,z) \in \mathbb{R}_+ \times [0,2\pi[\times \mathbb{R}$ les trois variables des coordonnées cylindriques, et e_r, e_θ, e_z les vecteurs associés. Si $\overline{\mathscr{U}} : \mathbb{R}^3 \to \mathbb{R}^3$, on écrit $\overline{\mathscr{U}} = \overline{\mathscr{U}}_r e_r + \overline{\mathscr{U}}_\theta e_\theta + \overline{\mathscr{U}}_z e_z$. Une fonction *axisymétrique* est une fonction indépendante de θ dans ce jeu de coordonnées. Une fonction *axisymétrique sans swirl* (*i.e.* sans composante suivant e_θ) est une fonction $\overline{\mathscr{U}} : \mathbb{R}^3 \to \mathbb{R}^3$ indépendante de θ et telle que $\overline{\mathscr{U}}_\theta = 0$ pour tout $(r,z) \in \mathbb{R}_+ \times \mathbb{R}$. Une fonction *axisymétrique pure swirl* est une fonction axisymétrique telle que $\overline{\mathscr{U}}_r = \overline{\mathscr{U}}_z = 0$ pour tout $(r,z) \in \mathbb{R}_+ \times \mathbb{R}$. Si $\mathscr{U} : \mathbb{R}^3 \to \mathbb{R}^3$ est une fonction C^1 axisymétrique sans swirl, alors son rotationnel $\Omega = \nabla \wedge \mathscr{U}$ est une fonction axisymétrique pure swirl, et on a $\Omega = (\partial_z \mathscr{U}_r - \partial_r \mathscr{U}_z) e_\theta$.

Soit $U \in W^{1,1} \cap W^{1,\infty}(\mathbb{R}^2, \mathbb{R}^2)$ un champ de vecteurs à divergence nulle, et soit $\Omega = \operatorname{curl} U = \partial_1 U_2 - \partial_2 U_1$. Alors $U = \mathrm{BS}[\Omega] := K_{\mathrm{BS}} * \Omega$, où K_{BS} est le noyau de Biot et Savart, défini par

$$K_{\mathrm{BS}}(x) := \frac{1}{2\pi} \frac{x^\perp}{|x|^2}, \quad \text{où} \quad x^\perp = \begin{pmatrix} -x_2 \\ x_1 \end{pmatrix} \quad \forall x \in \mathbb{R}^2. \tag{11}$$

Autrement dit, $\mathrm{BS}[\Omega] = \nabla^\perp \psi$, où $\Delta \psi = \Omega$. La fonction $\psi : \mathbb{R}^2 \to \mathbb{R}$ est appelée *fonction courant*. Ainsi, si $U : \mathbb{R}_+ \times \mathbb{R}^2 \to \mathbb{R}^2$ est une solution du système de Navier–Stokes (1) en dimension deux, sa vorticité Ω vérifie (au moins formellement) l'équation

$$\partial_t \Omega + U \cdot \nabla \Omega - \nu \Delta \Omega = 0, \quad U = \mathrm{BS}[\Omega].$$

On effectue ici les calculs pour le système de Navier–Stokes, mais on remarquera aisément que ceux-ci sont identiques pour le système d'Euler en prenant $\nu = 0$. L'opérateur de Navier–Stokes 2d linéarisé autour d'un flot $\overline{U}$ s'écrit donc, en formulation vorticité,

$$\Omega \mapsto \overline{U} \cdot \nabla \Omega + \mathrm{BS}[\Omega] \cdot \nabla \overline{\Omega} - \nu \Delta \Omega, \tag{12}$$

où $\overline{\Omega} = \operatorname{curl} \overline{U}$.

Considérons maintenant une solution $\mathscr{U}$ du système de Navier–Stokes tridimensionnel (1), et supposons que $\mathscr{U}$ est axisymétrique sans swirl. On pose $\Omega = \nabla \wedge \mathscr{U} = \Omega_\theta e_\theta$, où $\Omega_\theta = \partial_z \mathscr{U}_r - \partial_r \mathscr{U}_z$. On vérifie (formellement) que Ω_θ est solution de l'équation

$$\partial_t \Omega_\theta + \mathscr{U} \cdot \nabla \Omega_\theta - \frac{1}{r} \mathscr{U}_r \Omega_\theta - \nu \left(\partial_r^2 + \partial_z^2 \right) \Omega_\theta - \nu \partial_r \left(\frac{1}{r} \Omega_\theta \right) = 0.$$

De plus, $\mathscr{U} = \mathrm{BS}_{\mathrm{ass}}[\Omega_\theta]$, où l'opérateur $\mathrm{BS}_{\mathrm{ass}}$ (opérateur de Biot et Savart axisymétrique sans swirl) est défini par

$$\begin{aligned} \mathrm{BS}_{\mathrm{ass}}[\Omega_\theta] &= -\partial_z \psi e_r + \left(\partial_r \psi + \frac{1}{r}\psi\right) e_z, \\ \left(\partial_r^2 + \frac{1}{r}\partial_r - \frac{1}{r^2} + \partial_z^2\right)\psi &= -\Omega_\theta. \end{aligned} \tag{13}$$

Ainsi, l'opérateur de Navier–Stokes linéarisé autour d'un profil $\overline{\mathscr{U}}$ axisymétrique sans swirl s'écrit, en formulation vitesse,

$$\mathscr{U} \mapsto \overline{\mathscr{U}} \cdot \nabla \mathscr{U} + \mathscr{U} \cdot \nabla \overline{\mathscr{U}} - \nu \Delta \mathscr{U} + \nabla \mathscr{P},$$

où le champ de pression $\mathscr{P}$ assure que l'image de $\mathscr{U}$ par l'opérateur précédent est à divergence nulle. En formulation vorticité, après restriction aux vitesses axisymétriques sans swirl, l'opérateur devient

$$\Omega \mapsto \overline{\mathscr{U}} \cdot \nabla \Omega + \mathrm{BS}_{\mathrm{ass}}[\Omega] \cdot \nabla \overline{\Omega} - \frac{1}{r}\overline{\mathscr{U}}_r \Omega - \frac{1}{r}\mathrm{BS}_{\mathrm{ass}}[\Omega] \cdot e_r \overline{\Omega} - \nu\left(\partial_r^2 + \partial_z^2\right)\Omega - \nu \partial_r\left(\frac{1}{r}\Omega\right). \tag{14}$$

On observe (toujours formellement) que pour r très grand, les opérateurs donnés par (12) et (14) sont proches. Ce fait remarquable sera justifié rigoureusement dans la quatrième partie, et permettra de transformer l'instabilité identifiée par Vishik pour le flot d'Euler 2d en une instabilité pour le flot d'Euler 3d.

Panorama des travaux de Jia, Guillod et Šverák

Revenons à présent sur les travaux de Guillod et Šverák (2017) et Jia et Šverák (2014, 2015), qui ont initié partiellement la stratégie décrite dans ce texte. Dans ce paragraphe, on prend $\nu = 1$, ce qui est la convention choisie dans les articles sus-mentionnés. Comme expliqué dans l'introduction, l'invariance d'échelle (8) est au cœur de ces articles. Tout d'abord, Jia et Šverák (2014) ont démontré que si $u_0 \in C^\infty(\mathbb{R}^3 \setminus \{0\})$ est une donnée initiale invariante par le changement d'échelle (et donc homogène de degré -1), alors le système de Navier–Stokes (1) avec $f = 0$ admet une solution globale, invariante par le changement d'échelle (8) (et donc auto-similaire), qui est régulière dans $]0, +\infty[\times \mathbb{R}^3$. Pour cela, Jia et Šverák (2014) considèrent la donnée initiale ηu_0, avec $\eta \in [0,1]$. Pour η petit, le résultat de Koch et Tataru (2001) assure qu'il existe une unique solution ; celle-ci est nécessairement invariante par changement d'échelle, et il s'agit donc d'une solution auto-similaire. On écrit cette dernière sous la forme $u(t,x) = t^{-1/2}\overline{\mathscr{U}}_\eta(x/\sqrt{t})$, et on montre que $\overline{\mathscr{U}}_\eta(x) - u_0(x) = O(|x|^{-3})$ quand $|x| \to \infty$. L'existence de solutions pour $\eta \in [0,1]$ est ensuite obtenue grâce à la théorie du degré de Leray–Schauder.

La méthodologie proposée ensuite dans Jia et Šverák (2015) est la suivante : on considère la donnée initiale ηu_0, avec $\eta > 0$, ainsi qu'une solution auto-similaire associée, notée $\overline{\mathscr{U}}_\eta$, qui vérifie le système

$$\begin{aligned} -\Delta_\xi \overline{\mathscr{U}}_\eta - \frac{1}{2}\left(1+\xi\cdot\nabla_\xi\right)\overline{\mathscr{U}}_\eta + \overline{\mathscr{U}}_\eta\cdot\nabla_\xi\overline{\mathscr{U}}_\eta + \nabla_\xi P_\eta = 0,\\ \operatorname{div}_\xi \overline{\mathscr{U}}_\eta = 0,\\ \overline{\mathscr{U}}_\eta(x) - \eta u_0(x) = O(|x|^{-3}) \quad \text{quand } |x|\to\infty. \end{aligned}$$

On cherche ensuite une solution de (1) avec $f = 0$ de la forme

$$u(t,x) = \frac{1}{\sqrt{t}}\overline{\mathscr{U}}_\eta\left(\frac{x}{\sqrt{t}}\right) + \frac{1}{\sqrt{t}}\mathscr{V}\left(\ln t, \frac{x}{\sqrt{t}}\right),$$

où $\mathscr{V}$ est une solution de

$$\partial_\tau\mathscr{V} + \mathscr{H}_{\overline{\mathscr{U}}_\eta}\mathscr{V} + \mathscr{V}\cdot\nabla\mathscr{V} = 0.$$

On rappelle que l'opérateur $\mathscr{H}_{\overline{\mathscr{U}}_\eta}$ est défini dans (6). Comme expliqué dans l'introduction, on étudie le spectre de l'opérateur $\mathscr{H}_{\overline{\mathscr{U}}_\eta}$ dans $L^2(\mathbb{R}^3)^3$, en cherchant à identifier des champs de vitesse u_0 et des valeurs de η pour lesquels $\mathscr{H}_{\overline{\mathscr{U}}_\eta}$ possède une valeur propre de partie réelle strictement négative. Pour $\eta = 0$, le champ de vitesse $\overline{\mathscr{U}}_0$ est identiquement nul, et d'après Gallay et Wayne (2002a,b), lorsque $\nu = 1$ [(2)],

$$\sigma(\mathscr{H}_{\overline{\mathscr{U}}_0}) = \sigma(\mathscr{H}_0) = \left\{\lambda\in\mathbb{C},\ \Re(\lambda)\geqslant\frac{1}{4}\right\}\cup\left\{1+\frac{n}{2},\ n\in\mathbb{N}^*\right\} = \left\{\lambda\in\mathbb{C},\ \Re(\lambda)\geqslant\frac{1}{4}\right\},$$

où le demi-plan $\{\Re(\lambda)\geqslant 1/4\}$ correspond au spectre continu, tandis que les nombres $1+n/2$ sont des valeurs propres de $\mathscr{H}_0$. Pour $\nu = 1$ et $\overline{\mathscr{U}} \in L^\infty\cap H^2(\mathbb{R}^3)$ à divergence nulle et suffisamment décroissant en $+\infty$, d'après Jia et Šverák (2015),

$$\sigma(\mathscr{H}_{\overline{\mathscr{U}}}) \subset \left\{\lambda\in\mathbb{C},\ \Re(\lambda)\geqslant\frac{1}{4}\right\}\cup S,$$

où l'ensemble $S\cap\{\Re(\lambda) < 1/4\}$ ne comporte que des valeurs propres isolées. Pour $\eta \ll 1$, $\sigma(\mathscr{H}_{\overline{\mathscr{U}}_\eta}) \subset \{\lambda\in\mathbb{C}, \Re(\lambda) > 0\}$: le profil $\overline{\mathscr{U}}_\eta$ est stable et on a unicité de la solution auto-similaire, ce que l'on savait déjà grâce aux travaux de Koch et Tataru (2001). Jia et Šverák (2015) proposent deux scénarios potentiels de croisement de $\sigma(\mathscr{H}_{\overline{\mathscr{U}}_\eta})$ (et plus spécifiquement, du spectre discret de $\mathscr{H}_{\overline{\mathscr{U}}_\eta}$) avec l'axe $i\mathbb{R}$ pour $\eta = \eta_0 > 0$, et démontrent un résultat de non-unicité pour chacun de ces

[(2)] La formule donnée par Gallay et Wayne (2002a,b) est en fait plus générale, et donne le spectre de l'opérateur $\mathscr{H}_0$ dans l'espace $\{\mathscr{V}\in L^2(\mathbb{R}^3)^3,\ |\cdot|^m\mathscr{V}\in L^2(\mathbb{R}^3)^3,\ \operatorname{div}\mathscr{V} = 0\}$ avec $m\in\mathbb{N}$. La partie continue du spectre de $\mathscr{H}_0$ est alors $\{\lambda\in\mathbb{C},\ \Re(\lambda)\geqslant\frac{1}{4}+\frac{m}{2}\}$.

scénarios. Cependant les solutions ainsi obtenues sont faiblement décroissantes à l'infini, et n'appartiennent donc pas à l'espace d'énergie. Dans un second temps, il faut donc perturber et tronquer les solutions ainsi obtenues ; sous des hypothèses de non-dégénérescence supplémentaires, on obtient la non-unicité des solutions de Leray.

L'étape suivante est donc de déterminer si l'un des deux scénarios d'apparition d'une valeur propre instable pour $\mathscr{H}_{\overline{\mathscr{U}}_\eta}$ se produit effectivement. Guillod et Šverák (2017) ont mis en évidence numériquement l'un des deux scénarios, pour un champ de vitesse u_0 explicite [(3)]. Les auteurs concluent à la non-unicité des solutions de Leray, sous réserve que les observations numériques puissent être justifiées rigoureusement.

Néanmoins, obtenir une preuve rigoureuse des observations numériques de Guillod et Šverák (2017) est un défi mathématique conséquent. L'une des difficultés réside dans le fait que le profil $\overline{\mathscr{U}}_\eta$ n'est pas explicite, ce qui complique l'analyse spectrale de l'opérateur $\mathscr{H}_{\overline{\mathscr{U}}_\eta}$. La stratégie de Vishik (2018a,b) pour montrer la non-unicité des équations d'Euler 2d, puis de Albritton, Brué et Colombo (2022) pour celle du système de Navier–Stokes 3d repose sur la même idée générale : identifier des profils de vitesse instables dans les variables auto-similaires. Cependant, le fait d'autoriser un terme source dans l'équation (1) permet une latitude beaucoup plus grande dans le choix du profil de vitesse, puisque ce dernier n'est plus contraint d'être une solution particulière des équations. De fait, le profil identifié par Vishik (2018a,b) est quasiment explicite, et sa forme précise joue un rôle important dans le résultat d'instabilité (voir la preuve de la proposition 3.1 dans la prochaine partie). Le profil $\overline{\mathscr{U}}$, qui est obtenu en perturbant le profil de Vishik (et en le plongeant dans un cadre axisymétrique), est donc lui aussi explicite à l'ordre principal. Dans une moindre mesure, la présence du terme source dans le résultat de Albritton, Brué et Colombo (2022) facilite les raisonnements perturbatifs successifs dans la preuve du théorème 1.4 : chaque modification du profil de vitesse engendre un nouveau terme d'erreur, qui s'ajoutera au terme source $\overline{\mathscr{F}}$. Enfin, le fait de travailler avec un profil de vitesse invariant par le changement d'échelle dans Guillod et Šverák (2017) et Jia et Šverák (2015) nécessite de tronquer les solutions obtenues (qui ne décroissent que comme $|x|^{-1}$ en $+\infty$, et n'appartiennent donc pas à l'espace d'énergie), ce qui est une source de complications techniques substantielles.

Terminons cette partie par une comparaison entre les résultats de Albritton, Brué et Colombo (2022), Guillod et Šverák (2017) et Jia et Šverák (2015), et la notion de nombre de Reynolds critique en physique. Par définition, le nombre de Reynolds Re est un nombre sans dimension, évaluant dans un écoulement donné le rapport entre le terme d'advection $u \cdot \nabla u$ dans (1) et la dissipation visqueuse $\nu \Delta u$. Si L^* (resp. U^*) est une longueur typique (resp. une vitesse typique) de l'écoulement, le nombre de

(3) On prend $u_0(r,\theta,z) = \exp(-4(z/r)^2)(r^2+z^2)^{-1/2} e_\theta$ en coordonnées cylindriques, de sorte que u_0 est axisymétrique pure swirl et symétrique par rapport au plan $z = 0$.

Reynolds est défini par

$$\mathrm{Re} := \frac{L^* U^*}{\nu}.$$

Ainsi, après adimensionnement des équations, le système de Navier–Stokes (1) devient

$$\begin{aligned} \partial_t \tilde{u} + (\tilde{u} \cdot \widetilde{\nabla})\tilde{u} + \widetilde{\nabla}\tilde{p} - \frac{1}{\mathrm{Re}}\Delta\tilde{u} &= \tilde{f}, \\ \widetilde{\mathrm{div}}\,\tilde{u} &= 0, \end{aligned}$$

où le $\tilde{}$ indique que l'on travaille avec des variables adimensionnées (*i.e.* $u(t,x) = U^*\tilde{u}(t/T^*, x/L^*)$, etc.). Soit $\tilde{u}_I$ un profil stationnaire instable pour l'équation d'Euler (au sens où l'opérateur $\mathscr{G}_{\tilde{u}_I} \colon \tilde{u} \mapsto (\tilde{u}_I \cdot \nabla)\tilde{u} + (\tilde{u} \cdot \nabla)\tilde{u}_I + \nabla\tilde{p}$ admet une valeur propre de partie réelle négative). Pour $\mathrm{Re} \ll 1$, le spectre de l'opérateur $\mathscr{G}_{\tilde{u}_I} - \mathrm{Re}^{-1}\Delta$ est inclus dans $\{\lambda \in \mathbb{C}, \Re(\lambda) > 0\}$ (la diffusion l'emporte sur l'advection). En revanche, pour $\mathrm{Re} \gg 1$, on s'attend à ce que la diffusion puisse être traitée perturbativement, et donc à ce que $\mathscr{G}_{\tilde{u}_I} - \mathrm{Re}^{-1}\Delta$ ait une valeur propre de partie réelle strictement négative. Le nombre de Reynolds critique Re_c est précisément défini comme étant la première valeur de Re pour laquelle le spectre de $\mathscr{G}_{\tilde{u}_I} - \mathrm{Re}^{-1}\Delta$ (ou autrement dit, de $\mathscr{G}_{\mathrm{Re}\tilde{u}_I} - \Delta$) intersecte l'axe $i\mathbb{R}$:

$$\mathrm{Re}_c := \inf\{\mathrm{Re} > 0,\ \sigma(\mathscr{G}_{\mathrm{Re}\tilde{u}_I} - \Delta) \cap \{\lambda \in \mathbb{C},\ \Re(\lambda) \leqslant 0\} \neq \emptyset\}.$$

Évidemment, la valeur du nombre de Reynolds critique dépend du profil $\tilde{u}_I$. On voit ici immédiatement la similarité avec l'approche de Guillod et Šverák (2017) et Jia et Šverák (2015) : le nombre de Reynolds critique est précisément le nombre η_0. La preuve de Albritton, Brué et Colombo (2022) utilise également ce concept, dans une certaine mesure. On considère l'opérateur $\mathscr{H}_{\beta\overline{\mathscr{U}}}$, où $\overline{\mathscr{U}}$ est un profil instable pour Euler 3d. On observe que

$$\mathscr{H}_{\beta\overline{\mathscr{U}}} = \mathscr{G}_{\beta\overline{\mathscr{U}}} - \frac{1}{2}(1 + \xi \cdot \nabla_\xi) - \Delta_\xi.$$

Pour β suffisamment grand, on montre que cet opérateur a une valeur propre de partie réelle strictement négative, ce qui signifie (si on oublie momentanément le terme de transport) que $\beta > \mathrm{Re}_c$. Autrement dit, les travaux d'Albritton, Brué et Colombo (2022) se situent dans le régime $\mathrm{Re} \gg 1$.

3. Instabilité spectrale pour Euler 2d (d'après les travaux de Vishik)

Cette partie reprend les travaux de Vishik (2018a,b), en s'appuyant sur la présentation faite dans Albritton, Brué, Colombo et al. (2021). Comme expliqué dans la

partie précédente, le point de départ est de montrer qu'il existe un flot bidimensionnel $\overline{U}$ à divergence nulle tel que l'opérateur

$$\boldsymbol{G}_{\overline{U}}^{\mathrm{vel}} : U \in H \mapsto \overline{U} \cdot \nabla U + U \cdot \nabla \overline{U} + \nabla P, \quad H := \{U \in \dot{H}^1(\mathbb{R}^2)^2,\ \mathrm{div}\, U = 0\},$$

admette une valeur propre de partie réelle strictement négative. Le champ de pression P est choisi de sorte que le champ de vecteur $\overline{U} \cdot \nabla U + U \cdot \nabla \overline{U} + \nabla P$ soit à divergence nulle. Pour cela, on commence par écrire le problème aux valeurs propres en formulation vorticité : si $U \in H$, on peut écrire $U = \mathrm{BS}[\Omega]$, où $\Omega = \mathrm{curl}\ U = -\partial_2 U_1 + \partial_1 U_2$ est le tourbillon, et BS est le noyau de Biot et Savart défini dans (11).

S'il existe $\lambda \in \mathbb{C}$ et $U \in H$ tels que $\boldsymbol{G}_{\overline{U}}^{\mathrm{vel}} U = \lambda U$, alors, en posant $\Omega = \mathrm{curl}\ U$, $\overline{\Omega} = \mathrm{curl}\ \overline{U}$, on obtient

$$\overline{U} \cdot \nabla \Omega + U \cdot \nabla \overline{\Omega} = \overline{U} \cdot \nabla \Omega + \mathrm{BS}[\Omega] \cdot \nabla \overline{\Omega} = \lambda \Omega. \tag{15}$$

Réciproquement, si Ω vérifie l'équation aux valeurs propres (15), alors en posant $U = \mathrm{BS}[\Omega]$, on observe que

$$\mathrm{curl}\ \left(\overline{U} \cdot \nabla U + U \cdot \nabla \overline{U} - \lambda U\right) = 0.$$

Il existe donc formellement un champ de pression P tel que $\overline{U} \cdot \nabla U + U \cdot \nabla \overline{U} + \nabla P = \lambda U$, et les valeurs propres de $\boldsymbol{G}_{\overline{U}}^{\mathrm{vor}}$ et $\boldsymbol{G}_{\overline{U}}^{\mathrm{vel}}$ sont donc les mêmes.

Ainsi, on est amené à conduire une analyse spectrale de l'opérateur

$$\boldsymbol{G}_{\overline{U}}^{\mathrm{vor}} : \Omega \in H^1(\mathbb{R}^2) \mapsto \overline{U} \cdot \nabla \Omega + \mathrm{BS}[\Omega] \cdot \nabla \overline{\Omega},$$

pour des flots $\overline{U}$ bien choisis. En particulier, dans toute la suite, on prendra $\overline{\Omega}$ de la forme

$$\overline{\Omega}(x) = g(|x|) = g(r), \tag{16}$$

avec $g \in C_b^\infty(\mathbb{R}_+)$ et $x = (r\cos\theta, r\sin\theta)$, de sorte que $\overline{U} = \zeta(|x|)x^\perp = r\zeta(r)e_\theta$, où

$$\zeta(r) = \frac{1}{r^2} \int_0^r \rho g(\rho)\, d\rho.$$

Notons que cette analyse est voisine de celle des instabilités des flots de cisaillement du type $(\overline{U}_1(y), 0)$ pour le système d'Euler 2d posé dans l'espace entier, ou dans une bande du type $\mathbb{R} \times (0,1)$ (voir par exemple Grenier, 2000 ou Drazin et Reid, 2004). Comme les coefficients de l'opérateur $\boldsymbol{G}_{\overline{U}}^{\mathrm{vor}}$ ne dépendent pas de la variable angulaire θ, il est naturel d'analyser l'action de l'opérateur $\boldsymbol{G}_{\overline{U}}^{\mathrm{vor}}$ sur un mode de Fourier de fréquence k. Plus précisément, on peut décomposer $L^2(\mathbb{R}^2)$ comme

$$L^2(\mathbb{R}^2) = \bigoplus_{k \in \mathbb{Z}} E_k, \quad E_k := \{f(r)e^{ik\theta},\ f \in L^2(\mathbb{R}_+, r\, dr)\},$$

et on remarque que pour tout $k \in \mathbb{Z}$, $\boldsymbol{G}_{\overline{U}}^{\mathrm{vor}}(E_k \cap H^1(\mathbb{R}^2)) \subset E_k$.

Une première idée de Vishik est donc de restreindre l'analyse spectrale de l'opérateur $\boldsymbol{G}^{\mathrm{vor}}_{\overline{U}}$ à l'ensemble E_k, *i.e.* à un ensemble de fonctions invariantes par des rotations d'angle $2\pi/k$, avec $k \in \mathbb{N}$, $k \geqslant 2$. Le résultat fondamental de cette partie est le suivant :

Proposition 3.1. *Il existe $g \in C^\infty_b(\mathbb{R}_+)$, décroissant comme $r^{-\alpha}$ en $+\infty$ avec $0 < \alpha < 1$, et $k \in \mathbb{Z}$, $k \geqslant 2$, tels que l'opérateur $\boldsymbol{G}^{\mathrm{vor}}_{\overline{U}}|_{E_k}$ admet une valeur propre de partie réelle strictement négative. De plus, cette valeur propre est isolée dans $\sigma(\boldsymbol{G}^{\mathrm{vor}}_{\overline{U}}|_{E_k})$.*

Par conséquent, pour ce choix de g, l'opérateur $\boldsymbol{G}^{\mathrm{vel}}_{\overline{U}}$ admet une valeur propre de partie réelle strictement négative.

En tronquant la fonction ζ et en utilisant un argument de perturbation, on peut ensuite se ramener au cas où g est à support compact :

Corollaire 3.2. *Il existe $R > 0$ et une fonction $g_R \in C^\infty_c(\mathbb{R}_+)$ telle que* Supp $g_R \subset [0, R]$, Supp $\zeta_R \subset [0, R]$, *et telle que l'énoncé de la proposition 3.1 reste vrai en remplaçant g par g_R.*

Le reste de cette partie est donc dédié à l'analyse spectrale de l'opérateur $\boldsymbol{G}^{\mathrm{vor}}_{\overline{U}}$, restreint à l'espace E_k. Pour alléger les notations, on omet dans la suite l'indice $\overline{U}$. La preuve est organisée comme suit :

- ▷ On commence par décomposer $\boldsymbol{G}^{\mathrm{vor}}|_{E_k}$ en une partie antisymétrique (égale au terme de transport $\overline{U} \cdot \nabla\Omega$) et une partie compacte (égale au terme $\mathrm{BS}[\Omega] \cdot \nabla\overline{\Omega}$). Cette structure sera cruciale dans cette partie et la suivante pour analyser les propriétés spectrales de l'opérateur linéarisé.
- ▷ On montre ensuite l'existence de modes instables, que l'on construit par perturbations des modes neutres. Cette démarche est assez classique dans l'analyse spectrale des équations fluides. La subtilité réside ici dans le fait que les modes instables trouvés de cette façon ne correspondent pas nécessairement à des modes de Fourier k entiers, mais à des paramètres $k \in \mathbb{R}$, qui ne sont donc pas des modes de Fourier.
- ▷ Il faut donc ensuite montrer qu'il existe un mode instable correspondant à un entier $k \geqslant 2$. Cette construction repose sur un bel argument de connexité (du type théorème des valeurs intermédiaires), et sur un choix judicieux de la fonction g.

Décomposition de l'opérateur $G^{\rm vor}$

On commence par étudier la restriction de l'opérateur $\boldsymbol{G}^{\rm vor}$ à l'espace E_k pour $k \in \mathbb{Z}$.

Lemme 3.3. *Soit $k \in \mathbb{Z}$, $k \geqslant 2$, et soit $\Omega \in E_k$. On écrit $\Omega(x) = f(r)e^{ik\theta}$ avec $f \in L^2(\mathbb{R}_+, r\,dr)$. On suppose que $\overline{\Omega}$ est donné par (16) avec $g \in C_b^\infty(\mathbb{R}_+)$ tel que $|g(r)| \leqslant Cr^{-\alpha}$, $|g'(r)| \leqslant Cr^{-\alpha-1}$ pour $r \geqslant 1$.*

Alors

$$\boldsymbol{G}^{\rm vor}\Omega(r,\theta) = ike^{ik\theta}\left[\zeta(r)f(r) - \frac{1}{r}\psi(r)g'(r)\right],$$

où

$$\psi(r) = -\frac{1}{2k}r^k\int_r^\infty f(s)s^{1-k}\,ds - \frac{1}{2k}r^{-k}\int_0^r f(s)s^{1+k}\,ds \in L^2(\mathbb{R}_+, dr/r^3). \qquad (17)$$

De plus, l'opérateur

$$\mathbf{S}: \begin{array}{ccc} L^2(\mathbb{R}_+, r\,dr) & \to & L^2(\mathbb{R}_+, r\,dr) \\ f & \mapsto & \zeta f \end{array}$$

est auto-adjoint et borné, tandis que pour tout $k \in [1, +\infty[$ (pas nécessairement entier), l'opérateur

$$\mathbf{K}_k: \begin{array}{ccc} L^2(\mathbb{R}_+, r\,dr) & \to & L^2(\mathbb{R}_+, r\,dr) \\ f & \mapsto & \left(r \mapsto -\frac{1}{r}\psi(r)g'(r)\right) \end{array}$$

est compact.

Éléments de preuve. Pour $\Omega = f(r)e^{ik\theta} \in E_k$, on pose $U = \mathrm{BS}[\Omega] = \nabla^\perp\Delta^{-1}\Omega$. On introduit donc la (partie radiale de la) fonction courant $\psi = \psi(r)$, telle que

$$\Delta(\psi(r)e^{ik\theta}) = f(r)e^{ik\theta}.$$

On obtient alors

$$\psi'' + \frac{1}{r}\psi' - \frac{k^2}{r^2}\psi = f,$$

ce qui s'écrit encore

$$\frac{d^2}{dr^2}\left(\frac{\psi(r)}{r}\right) + \frac{3}{r}\frac{d}{dr}\left(\frac{\psi(r)}{r}\right) - \frac{k^2-1}{r^2}\frac{\psi(r)}{r} = \frac{f(r)}{r}.$$

On montre aisément que pour tout $f \in L^2(\mathbb{R}_+, rdr)$, cette équation admet une unique solution ψ telle que $\psi/r \in L^2(dr/r)$ et $d(\psi(r)/r)/dr \in L^2(rdr)$, qui est donnée par (17). Le reste de la preuve est laissé au lecteur ou à la lectrice, et découle des bornes sur g et ζ, ainsi que de la compacité de l'opérateur $f \in L^2(\mathbb{R}_+, rdr) \mapsto \psi/r \in L^2(a,b)$ pour tout $0 < a < b < +\infty$. □

Remarque 3.4.

- ▷ On déduit du Lemme précédent que $\sigma_{\text{ess}}(\mathbf{S}+\mathbf{K}_k) = \sigma_{\text{ess}}(\mathbf{S}) \subset \mathbb{R}$. Le Lemme 1.7 assure que si $\lambda \in \sigma(\mathbf{S}+\mathbf{K}_k)$ est tel que $\Re(\lambda) < 0$, alors λ est une valeur propre isolée de $\mathbf{S}+\mathbf{K}_k$.
- ▷ Supposons qu'il existe une valeur propre $\lambda_k \in \mathbb{C}\setminus\{0\}$ de $\mathbf{S}+\mathbf{K}_k$, et soit $f_k \in L^2(rdr)$ une fonction propre non triviale associée. Soit $\psi_k \in L^2(dr/r^3)$ la fonction courant donnée par (17). On a alors

$$(\zeta(r)-\lambda_k)f_k(r) = \frac{\psi_k(r)}{r}g'(r).$$

La définition de ζ assure que $\lim_{r\to\infty}\zeta(r) = 0$. Par ailleurs, $\psi_k(r)/r \in L^\infty(\mathbb{R})$. On en déduit qu'il existe une constante C_k telle que pour $r \geqslant 1$, $|f_k(r)| \leqslant C_k r^{-\alpha-1}$. En insérant cette estimation dans (17), on obtient, pour $r \geqslant 1$, $|\psi_k(r)| \lesssim r^{1-\alpha} + r^{-k}$. Par conséquent, $|f_k(r)| \lesssim r^{-2\alpha-1} + r^{-k-\alpha-2}$. En itérant ce raisonnement, on arrive finalement à $|\psi_k(r)| \lesssim r^{-k}$ pour $r \geqslant 1$, et donc $\psi_k \in L^2(dr/r)$.

Chercher une valeur propre λ_k de partie réelle strictement négative de $\boldsymbol{G}^{\text{vor}}$, relative à une fonction propre dans E_k avec $k \in \mathbb{Z}$, $k \geqslant 2$, revient donc à chercher un couple $(z,\psi) \in \mathbb{C}\times L^2(\mathbb{R}_+, dr/r)$ avec $\Im(z) > 0$ ($z = \lambda_k/ik$) et $\psi \neq 0$ tels que

$$-\psi''(r) - \frac{1}{r}\psi'(r) + \frac{k^2}{r^2}\psi(r) + \frac{g'(r)}{r(\zeta(r)-z)}\psi(r) = 0. \tag{18}$$

En faisant le changement de variables $r = e^s$, $s \in \mathbb{R}$, et $\psi(r) = \varphi(s)$, on obtient l'équation de Rayleigh

$$(\Xi(s)-z)\left(-\varphi''(s) + k^2\varphi(s)\right) + A(s)\varphi(s) = 0, \tag{19}$$

où

$$\Xi(s) = \zeta(r) = \int_{-\infty}^{s} e^{-2(s-\tau)}g(e^\tau)\,d\tau,$$

$$A(s) = \frac{d}{ds}g(e^s) = \Xi''(s) + 2\Xi'(s).$$

L'idée est à présent de choisir convenablement la fonction A (ou, de façon équivalente, en posant $\Xi(+\infty) = g(+\infty) = 0$, la fonction Ξ ou la fonction g) de sorte qu'il existe $k \in \mathbb{Z}$, $k \geqslant 2$, tel que l'équation (19) admette une solution non triviale (z,φ) avec $z \in \mathbb{C}$, $\Im(z) > 0$ et $\varphi \in L^2 \cap H^2_{\text{loc}}(\mathbb{R},\mathbb{C})$.

Remarque 3.5. Une différence cruciale avec les travaux antérieurs sur la stabilité des flots de cisaillement pour Euler 2d (voir Grenier, 2000 ou la section 22 de l'ouvrage de référence de Drazin et Reid, 2004) réside dans le fait qu'on cherche ici des solutions de l'équation de Rayleigh avec $k \in \mathbb{Z}$ (et non $k \in \mathbb{R}$).

Construction de modes instables de l'équation de Rayleigh (19) pour $k \in \mathbb{R}$

On commence par analyser l'équation (19) lorsque $z \in \mathbb{R}$:

Lemme 3.6 (Modes neutres). *Soit $N \in \mathbb{N}$ quelconque. Il existe une fonction $A \in C_b^\infty(\mathbb{R})$ telle que $A(0) = A(1/2) = 0$ et $k_0 \geqslant N$, $k_{1/2} > 0$, $k_0, k_{1/2} \in \mathbb{R}$ tels que l'équation (19) admette une solution $\varphi \in H^1(\mathbb{R})$ non triviale pour $(k,z) \in \{(k_0, \Xi(0)), (k_{1/2}, \Xi(1/2))\}$.*

Réciproquement, soit $A \in C_b^\infty(\mathbb{R})$ telle que $A(0) = A(1/2) = 0$, et $A(s) \neq 0$ pour $s \in \mathbb{R} \setminus \{0, 1/2\}$. Supposons qu'il existe (k, z, φ) vérifiant (19) avec $k > 1$, $z \in \mathbb{R}$ et $\varphi \in L^2 \cap H^2_{\mathrm{loc}}(\mathbb{R})$, $\varphi \neq 0$. Alors $z \in \{\Xi(0), \Xi(1/2)\}$.

On cherchera la fonction A dans la classe suivante :

Définition 3.7. La classe $\mathscr{C}$ est l'ensemble des fonctions $A \in C_b^\infty(\mathbb{R})$ telles que

1. Il existe $c_0, M_0 > 0$ tels que $A(s) = -8c_0 e^{2s}$ pour tout $s \leqslant -M_0$; de façon équivalente, Ξ admet une limite finie en $-\infty$, et $\Xi(s) = \Xi(-\infty) - c_0 e^{2s}$ si $s \leqslant -M_0$;
2. Il existe $\alpha > 0$ tel que $A(s) = -\alpha e^{-\alpha s}$ pour $s \geqslant \ln 2$; de façon équivalente, il existe $c_1 \in \mathbb{R}$ tel que $\Xi(s) = c_1 e^{-2s} + \frac{1}{2-\alpha} e^{-\alpha s}$ si $s \geqslant \ln 2$;
3. $A(0) = A(1/2) = 0$ et $A < 0$ sur $]-\infty, 0[\cup]1/2, +\infty[$, $A > 0$ sur $]0, 1/2[$, avec $A'(1/2) < 0 < A'(0)$;
4. $\Xi'(s) < 0$ pour tout $s \in \mathbb{R}$.

Remarque 3.8. Les modes neutres correspondent aux points d'annulation de A, et donc de g'. Pour cette raison, ils sont fréquemment appelés « points critiques » dans la littérature sur les instabilités hydrodynamiques. Par ailleurs, $g(r) = r\zeta'(r) + 2\zeta(r)$. Ainsi, un changement de signe de A correspond à un changement de signe de $r \mapsto r\zeta''(r) + 3\zeta'(r)$. La condition 3. de la définition 3.7 peut donc être interprétée comme une version du critère de Rayleigh (changement de signe de la dérivée de la vorticité, ou de la dérivée seconde de la vitesse de cisaillement) dans ce contexte axisymétrique sans swirl (voir Drazin et Reid, 2004).

Remarque 3.9. Ces hypothèses impliquent les propriétés suivantes sur la fonction $g \in C^\infty(\mathbb{R}_+)$:

1. $g(r) = g(0) + \frac{g''(0)}{2} r^2$ dans un voisinage de $r = 0$, avec $g''(0) = -8c_0 < 0$;
2. $g(r) = r^{-\alpha}$ pour $r \geqslant 2$;
3. $g'(1) = g'(e^{1/2}) = 0$, et $g' < 0$ sur $]0, 1[\cup]e^{1/2}, +\infty[$;
4. $\zeta' < 0$ sur $]0, +\infty[$.

Ébauche de la preuve du lemme 3.6. Soit $a \in \mathbb{R}$ tel que $A(a) = 0$. En posant $z_a = \Xi(a)$, et en considérant l'opérateur auto-adjoint

$$\mathbf{L}_{A,a} := -\frac{d^2}{ds^2} + \frac{A(s)}{\Xi(s) - \Xi(a)} \tag{20}$$

on observe que l'équation de Rayleigh (19) devient

$$\mathbf{L}_{A,a}\varphi = -k^2\varphi.$$

Autrement dit, avec le choix $z_a = \Xi(a)$, (19) devient un problème aux valeurs propres pour l'opérateur $\mathbf{L}_{A,a}$. Si on choisit $A \in \mathscr{C}$ (voir la définition 3.7), alors on observe que les fonctions $s \mapsto A(s)/(\Xi(s) - \Xi(a))$ sont bornées et changent de signe en un unique point : $s = 1/2$ lorsque $a = 0$, et $s = 0$ lorsque $a = 1/2$. De plus, d'après les propriétés 3. et 4. de la définition 3.7,

$$\frac{A(s)}{\Xi(s) - \Xi(0)} < \frac{A(s)}{\Xi(s) - \Xi(1/2)} \quad \forall s \in \mathbb{R}. \tag{21}$$

Notons $-\lambda_{A,a} = \inf \sigma(\mathbf{L}_{A,a})$. Puisque

$$-\lambda_{A,a} = \inf_{\substack{\varphi \in H^1(\mathbb{R}), \\ \|\varphi\|_{L^2} = 1}} \int_{\mathbb{R}} \left(|\varphi'(s)|^2 + \frac{A(s)}{\Xi(s) - \Xi(a)} \varphi(s)^2 \right) ds \tag{22}$$

on déduit de (21) que $-\lambda_{A,0} < -\lambda_{A,1/2} < 0$. De plus, l'opérateur $\mathbf{L}_{A,a}$ étant auto-adjoint, $-\lambda_{A,a}$ est une valeur propre de $\mathbf{L}_{A,a}$. D'après le théorème de Sturm–Liouville la fonction propre associée $\varphi_{A,a}$ ne s'annule pas sur $\mathbb{R}$. On prend donc $k_a = \sqrt{\lambda_{A,a}}$, de sorte que $0 < k_{1/2} < k_0$.

Par ailleurs, en prenant $a = 0$ dans (22) et $\varphi(s) = \mathbf{1}_{|s| \leqslant 1/2} \sqrt{2} \cos(\pi s)$, on en déduit la majoration

$$-\lambda_{A,0} \leqslant \pi^2 + 2 \int_{-1/2}^{1/2} \frac{A(s)}{\Xi(s) - \Xi(0)} \cos^2(\pi s)\, ds.$$

On choisit alors la fonction A de sorte que $A(s) = Bs$ et $\Xi'(s) \in [-2, -1/4]$ sur l'intervalle $[-\sqrt{B^{-1}}, 0]$, avec B une grande constante à choisir ultérieurement. On vérifie par un calcul que dans ce cas

$$-\lambda_{A,0} \leqslant \pi^2 - \int_{-\sqrt{B^{-1}}}^{0} B \cos^2(\pi s)\, ds \leqslant \pi^2 - \frac{\sqrt{B}}{2},$$

ce qui conclut la première partie du lemme pourvu que $\pi^2 - \sqrt{B}/2 \leqslant -N^2$. La seconde partie est laissée au lecteur ou à la lectrice : on commence par montrer que $z \in \Xi(\mathbb{R})$, puis que $z = \Xi(a)$ avec $a \in A^{-1}(\{0\})$. □

On cherche ensuite des modes instables (*i.e.* tels que $\Im(z) > 0$) de l'équation de Rayleigh (19) au voisinage des modes neutres. Plus précisément, en prenant $a \in \{0, 1/2\}$ et avec les notations précédentes, on cherche (φ, k, z) sous la forme

$$\begin{aligned} \varphi &= \varphi_{A,a} + \varepsilon\varphi_1 + \varepsilon^2\varphi_2 + \cdots, \\ k &= \sqrt{\lambda_{A,a}} + \varepsilon k_1 + \varepsilon^2 k_2 + \cdots, \\ z &= \Xi(a) + \varepsilon z_1 + \varepsilon^2 z_2 + \cdots, \end{aligned}$$

avec $\varepsilon \ll 1$. L'équation vérifiée par φ_1 est alors, à l'ordre principal en ε,

$$-z_1(-\varphi''_{A,a}+\lambda_{A,a}\varphi_{A,a}) + 2(\Xi-\Xi(a))\sqrt{\lambda_{A,a}}k_1\varphi_{A,a} + (\Xi-\Xi(a))(\mathbf{L}_{A,a}\varphi_1 + \lambda_{A,a}\varphi_1) = 0.$$

L'idée est alors de diviser par $\Xi - \Xi(a)$ et d'utiliser l'équation vérifiée par $\varphi_{A,a}$. Cependant, comme $\Xi - \Xi(a)$ s'annule, cette division n'est pas possible sans précaution préalable. On divise donc les deux membres de l'équation par $\Xi - \Xi(a) \pm i\delta$ avec $\delta > 0$ et on fait tendre δ vers zéro. On note $\Xi - \Xi(a) \pm i0$ la limite ainsi obtenue, et on obtient

$$(\mathbf{L}_{A,a} + \lambda_{A,a})\varphi_1 = -2\sqrt{\lambda_{A,a}}k_1\varphi_{A,a} - z_1(\Xi - \Xi(a) \pm i0)^{-1}\frac{A}{\Xi - \Xi(a)}\varphi_{A,a}.$$

D'après l'alternative de Fredholm, l'équation ci-dessus admet une solution si et seulement si

$$\int_{\mathbb{R}} \left(-2\sqrt{\lambda_{A,a}}k_1\varphi_{A,a} - z_1(\Xi - \Xi(a) \pm i0)^{-1}\frac{A}{\Xi - \Xi(a)}\varphi_{A,a} \right) \varphi_{A,a} = 0.$$

En utilisant des arguments d'analyse complexe (formule de Plemelj), on obtient alors une formule pour z_1, et en particulier

$$\mathrm{sgn}\Im(z_1) = \mp \mathrm{sgn}\frac{k_1 A'(a)}{\Xi'(a)^2}.$$

On en déduit qu'un choix convenable du signe de la régularisation complexe conduit à un mode instable. Une version plus rigoureuse de l'heuristique ci-dessus mène au résultat suivant :

Lemme 3.10 (Modes instables au voisinage des modes neutres). *Soit $A \in \mathscr{C}$, et soit $a \in \{0, 1/2\}$. On pose $\eta_0 = -1$, $\eta_{1/2} = 1$, et on suppose que $\lambda_{A,a} > 1$. Il existe des constantes $h_0 > 0$ et $c_{A,a} \in \mathbb{C}$ avec $\Im(c_{A,a}) > 0$, telles que pour tout $h \in]0, h_0[$, l'équation de Rayleigh (19) a une solution non triviale $\varphi \in H^2_{\mathrm{loc}} \cap L^2(\mathbb{R})$ avec $k = \sqrt{\lambda_{A,a}} + \eta_a h$ et $z = \Xi(a) + c_{A,a}h + o(h)$.*

Modes instables avec $k \in \mathbb{Z}$

À ce stade, le lemme 3.10 assure de l'existence de modes instables de l'équation de Rayleigh. Cependant, ces modes ne correspondent pas nécessairement à des valeurs entières de k, puisqu'ils se situent au voisinage de $\sqrt{\lambda_{A,a}}$ (qui n'est pas entier *a priori*). Ils ne coïncident donc pas avec des modes propres de $\boldsymbol{G}^{\mathrm{vor}}|_{E_k}$. La construction d'un mode instable avec $k \in \mathbb{Z}$ repose sur les deux arguments suivants :

▷ Tout d'abord, la courbe $z = Z(k)$ correspondant aux modes instables, dont les contours aux voisinages de $\sqrt{\lambda_{A,0}}$ et de $\sqrt{\lambda_{A,1/2}}$ sont esquissés au lemme 3.10, peut être prolongée à tout l'intervalle $\left]\max(\sqrt{\lambda_{A,1/2}}, 1), \sqrt{\lambda_{A,0}}\right[$. Plus précisément, on a le résultat de connexité suivant :

Lemme 3.11. *Soit $A \in \mathscr{C}$. On suppose que $\lambda_{A,0} > 1$. Alors pour tout $k \in \left]\max(\sqrt{\lambda_{A,1/2}}, 1), \sqrt{\lambda_{A,0}}\right[$, il existe $z \in \mathbb{C}$ avec $\Im(z) > 0$ tel que l'équation (19) admette une solution non triviale.*

Éléments de preuve. On commence par une observation préliminaire, qui découle de la seconde partie du Lemme 3.6 et des propriétés des solutions [(4)] de l'équation (19), et dont la preuve est laissée au lecteur ou à la lectrice. Pour tout compact C de $\left]\max(\sqrt{\lambda_{A,1/2}}, 1), \sqrt{\lambda_{A,0}}\right[$, il existe $R_C, \delta_C > 0$ tel que pour tout triplet $(k, z_k, \varphi_k) \in C \times \mathbb{C} \times L^2 \cap H^2_{\mathrm{loc}}(\mathbb{R})$ solution de l'équation de Rayleigh (19) avec φ_k non identiquement nulle, on a $|z_k| \leqslant R_C$ et $|\Im(z_k)| > \delta_C$.

Soit G l'ensemble des $k \in \left]\max(\sqrt{\lambda_{A,1/2}}, 1), \sqrt{\lambda_{A,0}}\right[$ tels qu'il existe une solution non triviale de (19) avec $\Im(z) > 0$. D'après le Lemme 3.10, G contient un voisinage à gauche de $\sqrt{\lambda_{A,0}}$. Montrons que G est un ouvert-fermé de $\left]\max(\sqrt{\lambda_{A,1/2}}, 1), \sqrt{\lambda_{A,0}}\right[$. Pour montrer le caractère fermé de G, on prend une suite $(k_j)_{j\in\mathbb{N}}$ d'éléments de G, convergeant vers $k_\infty \in \left]\max(\sqrt{\lambda_{A,1/2}}, 1), \sqrt{\lambda_{A,0}}\right[$. On pose $C = \{k_j,\ j \in \mathbb{N} \cup \{+\infty\}\}$. L'ensemble C est compact ; en utilisant l'observation préliminaire et en passant à la limite dans l'équation de Rayleigh (19) écrite pour chaque k_j, on en déduit que $k_\infty \in G$.

On montre ensuite que G^c est fermé. Pour ce faire, on considère une suite $(m_j)_{j\in\mathbb{N}}$ d'éléments de G^c qui converge vers $m_\infty \in \left]\max(\sqrt{\lambda_{A,1/2}}, 1), \sqrt{\lambda_{A,0}}\right[$. Pour tout $j \in \mathbb{N}$, par définition de G, les valeurs propres de $\mathbf{S} + \mathbf{K}_{m_j}$ sont réelles, et on a donc $\sigma(\mathbf{S} + \mathbf{K}_{m_j}) \subset \mathbb{R} \cup \sigma_{\mathrm{ess}}(\mathbf{S} + \mathbf{K}_{m_j}) \subset \mathbb{R}$. On pose à présent $C = \{m_\infty\}$ (qui est évidemment compact), et on note R_C, δ_C les constantes associées par l'observation préliminaire. Il suffit donc de montrer que $\sigma(\mathbf{S} + \mathbf{K}_{m_\infty}) \cap \{|z| \leqslant R_C, \Im(z) > \delta_C\} = \emptyset$. Pour cela, on note Γ le contour dans le plan complexe de la région $\{|z| \leqslant R_C, \Im(z) > \delta_C\}$, et on considère les

[(4)] En particulier, les solutions de (19) se comportent comme $C_\pm \exp(\mp|k|s)$ avec $C_\pm \in \mathbb{C}$ quand $s \to \pm\infty$.

projecteurs de Riesz

$$\mathbf{P}_m := \frac{1}{2\pi i}\int_\Gamma R(z, \mathbf{S}+\mathbf{K}_m)\, dz.$$

Pour tout $j \in \mathbb{N}$, $\mathbf{P}_{m_j} = 0$ par hypothèse. De plus,

$$\begin{aligned} R(z, \mathbf{S}+\mathbf{K}_{m_j}) &= (I + R(z, \mathbf{S}+\mathbf{K}_{m_\infty})(\mathbf{K}_{m_\infty} - \mathbf{K}_{m_j}))^{-1} R(z, \mathbf{S}+\mathbf{K}_{m_\infty}) \\ &\to R(z, \mathbf{S}+\mathbf{K}_{m_\infty}), \end{aligned}$$

uniformément pour $z \in \Gamma$. On en déduit que $\mathbf{P}_{m_\infty} = 0$, et donc $m_\infty \notin G$. □

▷ Il reste donc à construire une fonction particulière $A \in \mathscr{C}$ telle que l'intervalle $\left]\sqrt{\lambda_{A,1/2}}, \sqrt{\lambda_{A,0}}\right[$ contienne un entier supérieur ou égal à 2. Pour cela, on considère une première fonction $A_0 \in \mathscr{C}$. Soit $k \in \mathbb{Z}$ tel que $k > \max(\sqrt{\lambda_{A_0,0}}, 1)$. D'après le lemme 3.6, il existe une seconde fonction $A_1 \in \mathscr{C}$ telle que $\inf(\sigma(\mathbf{L}_{A_1,0})) = -\lambda_{A_1,0} < -k^2$. On pose $A_\eta = (1-\eta)A_0 + \eta A_1$ pour $\eta \in [0,1]$. Il existe η_0 tel que $\lambda_{A_{\eta_0},0} = k^2$, et on peut choisir η_0 maximal parmi les η vérifiant cette propriété. Remarquons que pour cet η_0, $\lambda_{A_{\eta_0},1/2} < k^2$. Alors pour h suffisamment petit, $\lambda_{A_{\eta_0+h},0} > k^2$ et $\lambda_{A_{\eta_0+h},1/2} < k^2$. La fonction $A = A_{\eta_0+h}$ convient donc.

Ceci conclut la preuve de la proposition 3.1. Pour la suite, il sera utile d'avoir un résultat d'instabilité spectrale pour des tourbillons à support compact. Nous concluons donc cette partie par quelques éléments de preuve pour le corollaire 3.2.

Troncature du tourbillon instable

Soit $\overline{U}$ le champ de vecteurs bidimensionnel obtenu dans la proposition 3.1, et soit λ la valeur propre de partie réelle strictement négative associée. Soit $\phi \in C_c^\infty(\mathbb{R})$ telle que $\phi \equiv 1$ dans un voisinage de zéro. Pour tout $R > 0$, $x \in \mathbb{R}^2$, on pose $\phi_R := \phi(\cdot/R)$, et

$$\overline{U}^R(x) = \zeta(|x|)\phi_R(|x|)x^\perp, \quad \overline{\Omega}^R := \operatorname{curl} \overline{U}^R.$$

On note également $\mathbf{L}_R = \boldsymbol{G}^{\mathrm{vor}}_{\overline{U}^R}$ l'opérateur linéarisé autour de $\overline{U}^R$.

On invoque alors le lemme 1.8 pour montrer que pour R suffisamment grand, en posant $r = -\Re(\lambda)/2$, on a $\sigma(\mathbf{L}_R) \cap B(\lambda, r) \neq \emptyset$. Par ailleurs, comme $\overline{U}^R$ est à divergence nulle,

$$\sigma_{\mathrm{ess}}\left(\overline{U}^R \cdot \nabla\right) \subset i\mathbb{R}.$$

Comme l'opérateur $U \mapsto U \cdot \nabla \overline{U}^R$ est une perturbation compacte de $\overline{U}^R \cdot \nabla$, on en déduit que $\sigma_{\mathrm{ess}}(\mathbf{L}_R) \subset i\mathbb{R}$. Par conséquent, pour R suffisamment grand, $B(\lambda, r)$ contient

une valeur propre de $\mathbf{L}_R$. Le lemme 1.7 entraîne de surcroît que cette valeur propre est isolée. On obtient alors le corollaire 3.2 en prenant $g_R : r \mapsto g(r)\phi_R(r) + r\zeta(r)\phi_R'(r)$.

Remarque 3.12. Dans les travaux de Albritton, Brué, Colombo et al. (2021) et Vishik (2018a,b), cette instabilité linéaire est exploitée pour démontrer la non-unicité des solutions de l'équation d'Euler 2d avec un terme source $f \in L^1_{\mathrm{loc}}(\mathbb{R}_+, L^1 \cap L^p(\mathbb{R}^2))$ avec $p > 2$. La démarche est très similaire à celle qui a été présentée dans l'introduction : on effectue un changement de variables autosimilaire, de sorte que l'instant initial $t = 0$ est envoyé en $\tau = -\infty$. On construit ensuite deux branches de solutions en tirant parti du mode instable. Une différence (importante !) entre les preuves de non-unicité pour Navier–Stokes 3d ou pour Euler 2d réside dans le choix des variables auto-similaires. Dans le cas de l'équation d'Euler 2d, on prend

$$v(t,x) = \frac{1}{t^{1-\frac{1}{\alpha}}} V\left(\ln t, \frac{x}{t^{\frac{1}{\alpha}}}\right),$$

avec $\alpha \in (0,1)$. On peut légitimement s'interroger sur l'ajout d'un terme de viscosité dans cette stratégie. Comme rappelé ci-dessus, les solutions de Leray du système de Navier–Stokes 2d sont uniques, et on s'attend donc à ce que la méthode échoue. De fait, on peut vérifier qu'avec le choix de variables auto-similaires ci-dessus, le terme de diffusion est dominant quand $t \to 0^+$, et ne peut donc être traité perturbativement. Si on utilise le même choix de variables auto-similaires que (4) lorsque $d = 2$, le terme source n'est pas dans $L^2(H^{-1}) + L^1(L^2)$, et les solutions u_1 et u_2 ne sont pas dans l'espace d'énergie. Albritton et Colombo (2023) ont établi que le théorème de non-unicité pour le système d'Euler 2d s'étend au système de Navier–Stokes avec laplacien fractionnaire $(-\Delta)^{\beta/2}$, pourvu que $\beta < 2$. Le cas $\beta = 2$ est donc critique.

Pour conclure, mentionnons une dernière piste dans cette direction. Bien que ce résultat ne figure pas dans la littérature, il semble raisonnable de penser que les arguments perturbatifs que l'on présentera dans la prochaine partie fonctionnent en dimension deux, et que par conséquent l'opérateur $\boldsymbol{G}_{\beta\overline{U}^R} - \Delta$ admet une valeur propre de partie réelle strictement négative pour β suffisamment grand. Si on utilise le choix de variables auto-similaires (4) pour $d = 2$, les solutions ainsi obtenues n'appartiennent pas à l'espace d'énergie, mais en revanche on peut vérifier que $u_i \in L^2([0,T], \dot{H}^s(\mathbb{R}^2)^2)$ pour tout $s < 1$, et que u_1, u_2 appartiennent également à l'espace X_T de Koch–Tataru défini par (10). Sous réserve que la preuve présentée ici s'exporte en deux dimensions, on en déduit donc que les solutions du système de Navier–Stokes avec terme source ne sont pas uniques dans $L^2([0,T], \dot{H}^s(\mathbb{R}^2)^2)$ ni dans X_T.

4. Instabilité linéaire et non linéaire pour Navier–Stokes 3d

Cette partie est dédiée à la preuve du théorème 1.4. Le principe est d'identifier un profil de vitesse $\overline{\mathscr{U}}$ vérifiant la propriété (**P**). Pour cela, on utilise plusieurs arguments perturbatifs, d'abord pour passer de deux à trois dimensions, puis pour passer d'Euler 3d à Navier–Stokes 3d.

4.1. D'Euler 2d à Euler 3d

La première idée est de transformer, à l'aide d'un relèvement ad hoc, l'instabilité obtenue pour le système d'Euler bidimensionnel en une instabilité pour le système d'Euler tridimensionnel. Pour cela, on crée un anneau instable tridimensionnel à partir du tourbillon instable 2d obtenu dans le corollaire 3.2. Plus précisément, si on note (r,θ,z) les coordonnées cylindriques en dimension trois, et $\overline{U}^R$ le champ de vitesse donné par le corollaire 3.2, on pose, pour $\ell \gg 1$ à fixer ultérieurement,

$$\widetilde{\mathscr{U}_\ell}(r,\theta,z) := \overline{U}_1^R(r-\ell,z)e_r + \overline{U}_2^R(r-\ell,z)e_z.$$

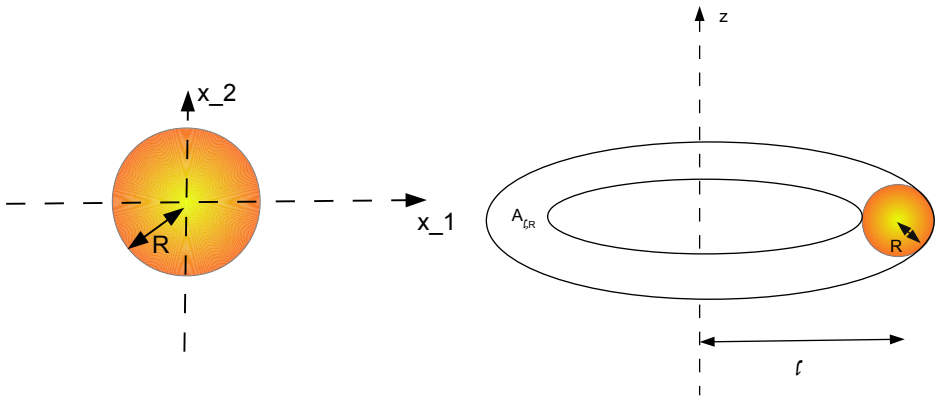


Figure 1 – À gauche, le tourbillon bidimensionnel de Vishik. À droite, son relèvement tridimensionnel et axisymétrique dans un anneau $A_{\ell,R}$.

Par construction, le champ $\widetilde{\mathscr{U}_\ell}$ est axisymétrique sans swirl et $\mathrm{Supp}\ \widetilde{\mathscr{U}_\ell} \subset \{|r-\ell|^2+z^2 \leqslant R^2\} =: A_{\ell,R}$. En particulier, si on choisit $\ell \geqslant 2R$, $\widetilde{\mathscr{U}_\ell}$ est supporté dans l'ensemble $r \in [\ell-R, \ell+R] \subset [\ell/2, 3\ell/2]$.

Notons qu'un tel $\widetilde{\mathscr{U}_\ell}$ n'est *pas* à divergence nulle : en effet,

$$\mathrm{div}\ \widetilde{\mathscr{U}_\ell} = \frac{1}{r}\overline{U}_1^R(r-\ell,z),$$

et donc $|\mathrm{div}\,\widetilde{\mathscr{U}_\ell}| \lesssim \ell^{-1}\mathbf{1}_{A_{\ell,R}}$. On commence donc par relever la divergence et construire un correcteur $\mathscr{V}_\ell \in C_c^\infty(\mathbb{R}^2)^2$ tel que $\mathrm{Supp}\,\mathscr{V}_\ell \subset A_{\ell,R}$, $\|\mathscr{V}_\ell\|_{C^k} \lesssim \ell^{-1}$ pour tout $k \in \mathbb{N}$ et $\mathrm{div}\,(\widetilde{\mathscr{U}_\ell} + \mathscr{V}_\ell) = 0$. On pose $\overline{\mathscr{U}}_\ell := \widetilde{\mathscr{U}_\ell} + \mathscr{V}_\ell$ et $\Omega_\ell := \partial_z \overline{\mathscr{U}}_{\ell,r} - \partial_r \overline{\mathscr{U}}_{\ell,z}$.

On considère ensuite l'opérateur d'Euler tridimensionnel linéarisé autour du profil $\overline{\mathscr{U}}_\ell$, et restreint aux fonctions axisymétriques sans swirl, c'est-à-dire de la forme $\mathscr{V} = \mathscr{V}_r(r,z)e_r + \mathscr{V}_z(r,z)e_z$. Pour de telles fonctions, le tourbillon $\Omega = \nabla \wedge \mathscr{V}$ est axisymétrique *pure swirl*, et s'écrit $\Omega = (\partial_z \mathscr{V}_r - \partial_r \mathscr{V}_z)e_\theta$. On peut donc écrire l'opérateur linéarisé autour de $\overline{\mathscr{U}}_\ell$ en formulation vorticité (voir (14)), que l'on note $\mathscr{G}^{\mathrm{vor}}_{\overline{\mathscr{U}}_\ell}$. On rappelle que

$$\mathscr{G}^{\mathrm{vor}}_{\overline{\mathscr{U}}_\ell} : \Omega \mapsto \overline{\mathscr{U}}_\ell \cdot \nabla\Omega + \mathrm{BS}_{\mathrm{ass}}[\Omega] \cdot \nabla\Omega_\ell - \frac{1}{r}\overline{\mathscr{U}}_{\ell,r}\Omega - \frac{1}{r}\Omega_\ell \mathrm{BS}_{\mathrm{ass}}[\Omega] \cdot e_r,$$

où l'opérateur $\mathrm{BS}_{\mathrm{ass}}$ est défini dans (13). On observe immédiatement que si Ω est une fonction propre de $\mathscr{G}^{\mathrm{vor}}_{\overline{\mathscr{U}}_\ell}$, alors $\mathrm{Supp}\,\Omega \subset A_{\ell,r}$. Heuristiquement, l'opérateur $\mathscr{G}^{\mathrm{vor}}_{\overline{\mathscr{U}}_\ell}$ est proche de $\boldsymbol{G}^{\mathrm{vor}}_{\tau_\ell \overline{U}^R}$ pour $\ell \gg 1$, où τ_ℓ est l'opérateur de translation $\tau_\ell f(r,z) = f(r-\ell,z)$. Ainsi, afin de se ramener à un opérateur proche de $\boldsymbol{G}^{\mathrm{vor}}_{\overline{U}^R}$, on compose avec la translation $\tau_{-\ell}$. On définit donc l'opérateur $\mathscr{L}_\ell$ par

$$\begin{aligned}\mathscr{L}_\ell\Omega := \tau_{-\ell}\,\overline{\mathscr{U}}_\ell \cdot \nabla\Omega + \mathrm{BS}_\ell[\Omega] \cdot \nabla\tau_{-\ell}\Omega_\ell \\ - \frac{1}{r+\ell}\tau_{-\ell}\,\overline{\mathscr{U}}_{\ell,r}\Omega - \frac{1}{r+\ell}\tau_{-\ell}\Omega_\ell \mathrm{BS}_\ell[\Omega] \cdot e_r,\end{aligned}$$

avec $\mathrm{BS}_\ell[\Omega] = \tau_{-\ell}\mathrm{BS}_{\mathrm{ass}}[\tau_\ell\Omega]$ et $D(\mathscr{L}_\ell) = \{\Omega \in L^2((r+\ell)_+ dr\,dz), \nabla\Omega \in L^2((r+\ell)_+ dr\,dz)\}$. Pour exploiter la ressemblance entre $\mathscr{L}_\ell$ et $\boldsymbol{G}^{\mathrm{vor}}_{\overline{U}^R}$, on analyse le spectre de $\mathscr{L}_\ell$ non dans $L^2((r+\ell)_+ dr\,dz)$, mais dans $L^2(\gamma\,dr\,dz)$, où $\gamma \in C^\infty(\mathbb{R})$ est un poids régulier tel que $\gamma \equiv 1$ sur B_R, et $\gamma(r) = (1+r^2)^N$ pour $|r| \geqslant 2$, avec N assez grand.

L'opérateur $\mathscr{L}_\ell$ peut être décomposé en une somme de la forme

$$\mathscr{L}_\ell = \mathscr{M}_\ell + \mathscr{K}_\ell + \mathscr{S}_\ell,$$

où les opérateurs $\mathscr{M}_\ell, \mathscr{K}_\ell, \mathscr{S}_\ell$ vérifient les propriétés suivantes :

- ▷ $\mathscr{M}_\ell$ est antisymétrique pour le produit scalaire dans $L^2(\gamma)$;
- ▷ $\mathscr{K}_\ell$ est compact ;
- ▷ $R(\lambda, \mathscr{M}_\ell + \mathscr{K}_\ell)\Omega \to R(\lambda, \boldsymbol{G}^{\mathrm{vor}}_{\overline{U}^R})\Omega$ localement uniformément sur $\rho(\boldsymbol{G}^{\mathrm{vor}}_{\overline{U}^R})$, pour tout $\Omega \in L^2(\gamma)$;
- ▷ $\mathscr{S}_\ell \to 0$ quand $\ell \to \infty$.

Plus précisément, on adopte les définitions suivantes :

▷ Terme principal (*main term*) $\mathscr{M}_\ell$: on pose

$$\mathscr{M}_\ell\Omega = \overline{U}_1^R\partial_r\Omega + \overline{U}_2^R\partial_z\Omega + \tau_{-\ell}\mathscr{V}_\ell\cdot\nabla\Omega + \frac{1}{2}(\partial_r\tau_{-\ell}\mathscr{V}_{\ell,r} + \partial_z\tau_{-\ell}\mathscr{V}_{\ell,z})\Omega$$

Le troisième terme assure l'antisymétrie de $\mathscr{M}_\ell$ sur $L^2(\gamma\mathbf{1}_{r+\ell>0})$, et est petit lorsque $\ell \gg 1$. On rappelle que Supp $\overline{U}^R \subset B_R$ et Supp $\tau_{-\ell}\mathscr{V}_\ell \subset B_R$. On peut montrer, avec des arguments similaires à ceux du lemme 4.2 plus bas, que pour tout $\Omega \in L^2(\gamma)$, pour tout compact C de $\{\Re(\lambda) < 0\}$, $R(\lambda,\mathscr{M}_\ell)(\mathbf{1}_{r+\ell>0}\Omega) \to R(\lambda,\overline{U}^R\cdot\nabla)\Omega$ uniformément pour $\lambda\in C$.

▷ Terme compact $\mathscr{K}_\ell$: on pose

$$\mathscr{K}_\ell\Omega = \mathrm{BS}_\ell[\Omega]\cdot\nabla\tau_{-\ell}\Omega_\ell - \frac{1}{r+\ell}\tau_{-\ell}\Omega_\ell\mathrm{BS}_\ell[\Omega]\cdot e_r.$$

Comme Supp $\Omega_\ell \subset A_{\ell,R}$, Supp $\tau_{-\ell}\Omega_\ell \subset B_R$. La compacité de l'opérateur $\mathscr{K}_\ell$ découle de cette observation ainsi que de l'effet régularisant de l'opérateur de Biot et Savart. On montre de surcroît que $\mathscr{K}_\ell\mathbf{1}_{r+\ell>0} \to \mathrm{BS}[\,\cdot\,]\cdot\nabla\overline{\Omega}^R$ pour la norme d'opérateur dans $L^2(\gamma)$.

▷ Terme perturbatif (*small stretching term*) $\mathscr{S}_\ell$: on pose

$$\mathscr{S}_\ell\Omega = -\frac{1}{r+\ell}\tau_{-\ell}\overline{\mathscr{U}}_{\ell,r}\,\Omega - \frac{1}{2}(\partial_r\tau_{-\ell}\mathscr{V}_{\ell,r} + \partial_z\tau_{-\ell}\mathscr{V}_{\ell,z})\Omega.$$

Ici encore, Supp $\tau_{-\ell}\overline{\mathscr{U}}_{\ell,r} \subset B_R$, Supp $\tau_{-\ell}\mathscr{V}_\ell \subset B_R$. On en déduit que

$$\|\mathscr{S}_\ell\|_{L^2(\gamma\mathbf{1}_{r>-\ell})\to L^2(\gamma)} \to 0 \quad \text{quand } \ell\to\infty.$$

Ces trois propriétés, combinées avec les lemmes 1.7 et 1.8, entraînent que $\mathscr{L}_\ell$ admet une valeur propre isolée de partie réelle strictement négative :

Proposition 4.1. *Soit $R>0$ tel que l'opérateur $\boldsymbol{G}^{\mathrm{vor}}_{\overline{U}^R}$ admette une valeur propre λ^R de partie réelle strictement négative.*

Pour tout $\varepsilon \in]0,-\Re(\lambda^R)[$, il existe $\ell_0 > 2R$ tel que pour tout $\ell>\ell_0$, l'opérateur $\mathscr{L}_\ell$ admette une valeur propre λ_ℓ telle que $|\lambda_\ell-\lambda^R|<\varepsilon$, de sorte que $\Re(\lambda_\ell)<0$.

Par conséquent, λ_ℓ est également valeur propre de l'opérateur $\mathscr{G}^{\mathrm{vor}}_{\overline{\mathscr{U}}_\ell}$, et la fonction propre associée est supportée dans $A_{\ell,R}$. De plus λ_ℓ est un élément isolé de $\sigma(\mathscr{G}^{\mathrm{vor}}_{\overline{\mathscr{U}}_\ell})$.

4.2. D'Euler 3d à Navier–Stokes 3d

Soient $R,\ell>0$ tels que l'opérateur $\mathscr{G}^{\mathrm{vor}}_{\overline{\mathscr{U}}_\ell}$ admette une valeur propre λ_ℓ de partie réelle strictement négative, donnée par la proposition 4.1. Bien évidemment, pour

tout $\beta > 0$, $\beta\lambda_\ell$ est valeur propre de l'opérateur $\mathscr{G}^{\mathrm{vor}}_{\beta\overline{\mathscr{U}}_\ell}$. Dans la suite, on prend R et ℓ fixés, et on omet donc la dépendance en ℓ dans $\overline{\mathscr{U}}_\ell$.

On considère à présent l'opérateur de Navier–Stokes tridimensionnel dans les variables auto-similaires, linéarisé autour du profil $\beta\overline{\mathscr{U}}$ avec $\beta > 0$, restreint aux fonctions axisymétriques sans swirl, et écrit en formulation vorticité. Pour simplifier, on prend $\nu = 1$ dans toute cette partie. Cet opérateur s'écrit

$$\mathscr{H}^{\mathrm{vor}}_{\beta\overline{\mathscr{U}}} : \Omega \mapsto -\left(1 + \frac{\xi}{2}\cdot\nabla_\xi\right)\Omega - \Delta\Omega + \beta\mathscr{G}^{\mathrm{vor}}_{\overline{\mathscr{U}}}\Omega,$$

et son domaine est $D(\mathscr{H}^{\mathrm{vor}}_{\beta\overline{\mathscr{U}}}) = \{\Omega \in H^2(\mathbb{R}^3),\ \Omega = \Omega(r,z),\ \xi\cdot\nabla_\xi\Omega \in L^2(\mathbb{R}^3)\}$.

On veut raisonner perturbativement. On écrit donc

$$\mathscr{H}^{\mathrm{vor}}_{\beta\overline{\mathscr{U}}} = \beta\left(\mathscr{G}^{\mathrm{vor}}_{\overline{\mathscr{U}}} - \frac{1}{\beta} - \frac{\xi}{2\beta}\cdot\nabla_\xi - \frac{1}{\beta}\Delta\right).$$

L'idée est ensuite d'utiliser le lemme 1.8 et la proposition 4.1 pour montrer que $\mathscr{H}^{\mathrm{vor}}_{\beta\overline{\mathscr{U}}}$ possède une valeur propre de partie réelle strictement négative pour β suffisamment grand. Pour cela, on utilise de nouveau une décomposition du type

$$\mathscr{G}^{\mathrm{vor}}_{\overline{\mathscr{U}}} = \mathscr{M} + \mathscr{S} + \mathscr{K},$$

où $\mathscr{M} = \overline{\mathscr{U}}\cdot\nabla$ est antisymétrique, $\mathscr{K} : \Omega \mapsto \mathrm{BS}_{\mathrm{ass}}[\Omega]\cdot(\nabla\Omega_\ell - r^{-1}\Omega_\ell e_r)$ est compact et $\mathscr{S} : \Omega \mapsto -r^{-1}\overline{\mathscr{U}}\,\Omega$ est petit pour $\ell \gg 1$. Il convient de souligner que les opérateurs $\mathscr{M}, \mathscr{S}, \mathscr{K}$ sont différents des opérateurs $\mathscr{M}_\ell, \mathscr{S}_\ell, \mathscr{K}_\ell$ du paragraphe précédent, en raison des compositions avec les translations $\tau_{\pm\ell}$ qui interviennent dans la définition de l'opérateur $\mathscr{L}_\ell$. Posons

$$a := \inf_{\lambda\in\sigma(\mathscr{G}^{\mathrm{vor}}_{\overline{\mathscr{U}}})} \Re(\lambda) < 0.$$

On peut montrer que $\mu := \|\mathscr{S}\| \lesssim \ell^{-2/3}$, où $\|\mathscr{S}\|$ est la norme d'opérateur de $\mathscr{S}$, de sorte que l'on peut choisir ℓ tel que $\|\mathscr{S}\| \leqslant |a|/2$. On isole la partie antisymétrique du terme de transport en posant (rappelons que l'on travaille ici en dimension trois)

$$\mathscr{T} : \Omega \mapsto -\frac{3}{4}\Omega - \frac{\xi}{2}\cdot\nabla_\xi\Omega.$$

On a alors le résultat suivant :

Lemme 4.2. *Pour tout $\beta > 0$, et pour tout λ tel que $\Re(\lambda) < -\mu$,*

$$\|R(\lambda, \beta^{-1}\mathscr{T} - \beta^{-1}\Delta + \mathscr{M} + \mathscr{S})\| \leqslant |\Re(\lambda) + \mu|^{-1}, \tag{23}$$

et pour tout $\Omega \in L^2(rdrdz)$, localement uniformément en λ,

$$\lim_{\beta\to\infty} R(\lambda, \beta^{-1}\mathscr{T} - \beta^{-1}\Delta + \mathscr{M} + \mathscr{S})\Omega = R(\lambda, \mathscr{M} + \mathscr{S})\Omega \quad \textit{dans } L^2(\mathbb{R}^3).$$

Démonstration. Soit $\Omega_0 \in L^2(rdrdz)$. On considère l'équation d'advection-diffusion

$$\begin{aligned} \partial_t \Omega^\beta + \left(\beta^{-1}\mathscr{T} - \beta^{-1}\Delta + \mathscr{M} + \mathscr{S}\right)\Omega^\beta = 0, \\ \Omega^\beta|_{t=0} = \Omega_0. \end{aligned} \tag{24}$$

Comme les opérateurs $\mathscr{T}$ et $\mathscr{M}$ sont antisymétriques, on a l'estimation d'énergie

$$\frac{1}{2}\frac{d}{dt}\|\Omega^\beta\|_{L^2}^2 + \beta^{-1}\|\nabla\Omega^\beta\|_{L^2}^2 \leqslant \|\mathscr{S}\|\|\Omega^\beta\|_{L^2} \leqslant \mu\|\Omega^\beta\|_{L^2},$$

et donc $\|\Omega^\beta(s)\|_{L^2} \leqslant e^{\mu t}\|\Omega_0\|_{L^2}$. En écrivant $R(\lambda, -\beta^{-1}\mathscr{T} + \beta^{-1}\Delta - \mathscr{M} - \mathscr{S})$ comme une transformée de Laplace, on a

$$R(\lambda, -\beta^{-1}\mathscr{T} + \beta^{-1}\Delta - \mathscr{M} - \mathscr{S})\Omega_0 = \int_0^\infty \exp(-s\lambda)\Omega^\beta(s)\,ds \quad \forall \lambda \in \mathbb{C},\ \Re(\lambda) > \mu,$$

et l'inégalité (23) s'ensuit.

Par ailleurs, Ω^β est borné uniformément en β dans $L^\infty_{\mathrm{loc}}(\mathbb{R}_+, L^2(\mathbb{R}^3))$. On peut donc en extraire une sous-suite qui converge faiblement lorsque $\beta \to 0$. La limite Ω vérifie l'équation de transport

$$\begin{aligned} \partial_t \Omega + (\mathscr{M} + \mathscr{S})\,\Omega = 0, \\ \Omega|_{t=0} = \Omega_0, \end{aligned} \tag{25}$$

et est donc unique. On montre aisément que si $\Omega_0 \in H^1(\mathbb{R}^3, \rho(\xi)d\xi)$, où ρ est un poids régulier à croissance algébrique, alors $\Omega \in L^\infty_{\mathrm{loc}}(\mathbb{R}_+, H^1(\mathbb{R}^3, \rho(\xi)d\xi))$. On en déduit que Ω est une solution de (24) avec un reste d'ordre β^{-1}, puis que, pour tout $\tau > 0$,

$$\|\Omega^\beta - \Omega\|_{L^\infty([0,\tau], L^2(\mathbb{R}^3))} \leqslant C(\tau, \Omega_0)\beta^{-1} \quad \forall \Omega_0 \in H^1(\mathbb{R}^3, \rho(\xi)d\xi).$$

On écrit alors, pour $\lambda \in \mathbb{C}$, $\Re(\lambda) > \mu$ et $\Omega_0 \in H^1(\mathbb{R}^3, \rho(\xi)d\xi)$,

$$\begin{aligned} &\left[R(\lambda, -\beta^{-1}\mathscr{T} + \beta^{-1}\Delta - \mathscr{M} - \mathscr{S}) - R(\lambda, -\mathscr{M} - \mathscr{S})\right]\Omega_0 \\ &= \int_0^\infty \exp(-s\lambda)\left(\Omega^\beta(s) - \Omega(s)\right)\,ds. \end{aligned}$$

On en déduit que pour tout $\tau > 0$, pour tout $\Omega_0 \in H^1(\mathbb{R}^3, \rho(\xi)d\xi)$, pour $\lambda \in \mathbb{C}$, $\Re(\lambda) > \mu$,

$$\begin{aligned} &\left\|\left[R(\lambda, -\beta^{-1}\mathscr{T} + \beta^{-1}\Delta - \mathscr{M} - \mathscr{S}) - R(\lambda, -\mathscr{M} - \mathscr{S})\right]\Omega_0\right\|_{L^2} \\ \leqslant & C(\tau, \Omega_0)\frac{1}{\beta\mu} + 2\int_\tau^\infty \exp(s(\mu - \Re(\lambda))\|\Omega_0\|_{L^2}\,ds \\ \leqslant & C(\tau, \Omega_0)\frac{1}{\beta\mu} + 2\frac{\exp(\tau(\mu - \Re(\lambda))}{\Re(\lambda) - \mu}\|\Omega_0\|_{L^2}. \end{aligned}$$

En choisissant d'abord τ grand pour que le second terme soit petit, puis β grand pour que le premier terme le soit également, on obtient la convergence annoncée pour $\Omega_0 \in H^1(\mathbb{R}^3, \rho(\xi)d\xi)$, puis pour $\Omega \in L^2(r\,dr\,dz)$ par densité. □

On en déduit immédiatement que

$$\sigma(\beta^{-1}\mathscr{T} - \beta^{-1}\Delta + \mathscr{M} + \mathscr{S}) \subset \{\Re(\lambda) \geqslant -\mu\},$$

et donc $\sigma_{\text{ess}}(\beta^{-1}\mathscr{T} - \beta^{-1}\Delta + \mathscr{M} + \mathscr{S} + \mathscr{K}) \subset \{\Re(\lambda) \geqslant -\mu\}$. On rappelle que $\mathscr{G}^{\text{vor}}_{\overline{\mathscr{U}}} = \mathscr{M} + \mathscr{S} + \mathscr{K}$. Soit λ une valeur propre de partie réelle strictement négative de $\mathscr{G}^{\text{vor}}_{\overline{\mathscr{U}}}$. On peut montrer que l'on peut choisir λ tel que $\Re(\lambda) = a$. D'après le lemme 1.8, pour β suffisamment grand, l'opérateur $\beta^{-1}\mathscr{T} - \beta^{-1}\Delta + \mathscr{G}^{\text{vor}}_{\overline{\mathscr{U}}}$ possède une valeur propre dans $B(\lambda, |a|/4)$. Par conséquent, si $\beta > |a|^{-1}$, l'opérateur $\mathscr{H}^{\text{vor}}_{\beta\overline{\mathscr{U}}}$ possède une valeur propre dans $B(\beta\lambda, \beta|a|/2)$. On obtient finalement le résultat suivant :

Proposition 4.3. *Soit $\overline{\mathscr{U}}$ le profil construit dans la proposition 4.1.*

Il existe $\beta_0 > 0$ tel que si $\beta \geqslant \beta_0$, l'opérateur $\mathscr{H}^{\text{vor}}_{\beta\overline{\mathscr{U}}}$ admet une valeur propre de partie réelle strictement négative.

Par conséquent, l'opérateur $\mathscr{H}^{\text{vel}}_{\beta\overline{\mathscr{U}}}$ admet également une valeur propre de partie réelle strictement négative.

4.3. Du linéaire au non linéaire

On considère dans cette dernière partie l'équation (5), avec le terme source

$$\overline{\mathscr{F}} = -\frac{\beta}{2}(1 + \xi \cdot \nabla_\xi)\overline{\mathscr{U}} - \beta\nu\Delta_\xi\overline{\mathscr{U}} + \beta^2(\overline{\mathscr{U}} \cdot \nabla_\xi)\overline{\mathscr{U}}, \quad \beta > 0.$$

Par définition de $\overline{\mathscr{F}}$, la fonction $\beta\overline{\mathscr{U}}$ est une solution de (5). Le but est de construire une autre solution de l'équation (5) de la forme

$$\mathscr{U} = \beta\overline{\mathscr{U}} + \mathscr{U}_{\text{lin}} + \mathscr{U}_{\text{per}},$$

où $\mathscr{U}_{\text{lin}}$ est un mode propre instable de $\mathscr{H}^{\text{vel}}_{\beta\overline{\mathscr{U}}}$ (avec β suffisamment grand, voir la proposition 4.3), *i.e.* $\mathscr{U}_{\text{lin}} = \Re(\exp(-\lambda\tau)\mathscr{V}_\lambda)$, où λ est une valeur propre de partie réelle strictement négative de $\mathscr{H}^{\text{vel}}_{\beta\overline{\mathscr{U}}}$, et $\mathscr{V}_\lambda$ la fonction propre associée. On construit alors $\mathscr{U}_{\text{per}}$ à l'aide d'un théorème de point fixe. Plus précisément, on pose $a = \inf\{\Re(\mu), \mu \in \sigma(\mathscr{H}^{\text{vel}}_{\beta\overline{\mathscr{U}}})\}$. On sait déjà que $a < 0$ d'après ce qui précède, et on peut montrer qu'il existe une valeur propre μ^* telle que $a = \Re(\mu^*)$. On prend donc $\mathscr{U}_{\text{lin}} = \Re(\exp(-\mu^*\tau)\mathscr{V}_{\mu^*})$. On introduit l'espace de Banach

$$X_T = \{\mathscr{U} \in C((-\infty, T), H^3(\mathbb{R}^3)),\ \sup_{\tau<T} e^{-3a\tau/2}\|\mathscr{U}(\tau)\|_{H^3} < +\infty\},$$

et on cherche $\mathscr{U}_{\rm per}$ comme un point fixe de la fonctionnelle

$$\mathscr{I}(\mathscr{U}) := -\int_{-\infty}^{\tau} e^{-(\tau-s)\mathscr{H}^{\rm vel}_{\beta\overline{\mathscr{U}}}}\,\mathbb{P}\left[(\mathscr{U}\cdot\nabla\mathscr{U}) + (\mathscr{U}_{\rm lin}\cdot\nabla\mathscr{U}) + (\mathscr{U}\cdot\nabla\mathscr{U}_{\rm lin})\right]\,ds$$
$$-\int_{-\infty}^{\tau} e^{-(\tau-s)\mathscr{H}^{\rm vel}_{\beta\overline{\mathscr{U}}}}\,\mathbb{P}((\mathscr{U}_{\rm lin}\cdot\nabla)\mathscr{U}_{\rm lin})\,ds,$$

où $\mathbb{P}$ est le projecteur de Leray sur les champs de vecteurs à divergence nulle. Le choix de l'espace fonctionnel $H^3(\mathbb{R}^3)$ dans la définition de X_T est guidé par l'injection de Sobolev $H^s(\mathbb{R}^3) \subset W^{1,\infty}(\mathbb{R}^3)$ pour $s > 5/2$ (et on choisit de surcroît ici s entier). On montre alors, en utilisant les propriétés de $\mathscr{H}^{\rm vel}_{\beta\overline{\mathscr{U}}}$, qu'il existe $T \in \mathbb{R}$ tel que la boule $B_{X_T} := \{\mathscr{U} \in X_T,\ \|\mathscr{U}\|_{X_T} < 1\}$ est stable par $\mathscr{I}$, et tel que $\mathscr{I}|_{B_{X_T}}$ est une contraction. L'existence et l'unicité de $\mathscr{U}_{\rm per} \in X_T$ s'ensuivent. On définit ensuite deux solutions de Leray u_1 et u_2 de (1) par (7), et on vérifie que u_1 et u_2 sont deux solutions distinctes de (1) pour la même donnée initiale $u_0 = 0$, ce qui conclut la preuve du théorème 1.4.

Remerciements. — Je remercie vivement Nicolas Bourbaki, Isabelle Gallagher, Thierry Gallay et Charlotte Perrin pour leur relecture attentive, leurs conseils et leurs explications, qui auront permis d'améliorer la qualité de ce texte. Ce travail a bénéficié du soutien l'Institut Universitaire de France et de la bourse ANR-18-CE40-0027 de l'Agence Nationale de la Recherche (projet SingFlows).

Références

Albritton, D., Brué, E. et Colombo, M. (2022). « Non-uniqueness of Leray solutions of the forced Navier–Stokes equations », *Annals of Mathematics* **196** (1), p. 415-455.

Albritton, D., Brué, E., Colombo, M. et al. (2021). « Instability and nonuniqueness for the 2d Euler equations in vorticity form, after M. Vishik », arXiv : 2112.04943.

Albritton, D. et Colombo, M. (2023). « Non-uniqueness of Leray solutions to the hypodissipative Navier–Stokes equations in two dimensions », *Communications in Mathematical Physics*, p. 1-18.

Bahouri, H., Chemin, J.-Y. et Danchin, R. (2011). *Fourier analysis and nonlinear partial differential equations*. T. 343. Springer.

Buckmaster, T. et Vicol, V. (2019). « Nonuniqueness of weak solutions to the Navier–Stokes equation », *Annals of Mathematics* **189** (1), p. 101-144.

Cannone, M. (1997). « A generalization of a theorem by Kato on Navier–Stokes equations », *Revista matemática iberoamericana* **13** (3), p. 515-541.

Drazin, P. G. et Reid, W. H. (2004). *Hydrodynamic stability*. Cambridge university press.

Elgindi, T. M. (2021). « Finite-time Singularity Formation for $C^{1,\alpha}$ Solutions to the Incompressible Euler Equations on $\mathbb{R}^3$ », *Annals of Mathematics* **194** (3), p. 647-727.

Fujita, H. et Kato, T. (1964). « On the Navier–Stokes initial value problem. I », *Archive for Rational Mechanics and Analysis* **16** (4), p. 269-315.

Gallay, T. et Wayne, E. (2002a). « Invariant manifolds and the long-time asymptotics of the Navier–Stokes and vorticity equations on $\mathbb{R}^2$ », *Archive for Rational Mechanics and Analysis* **163**, p. 209-258.

——— (2002b). « Long-time asymptotics of the Navier–Stokes and vorticity equations on $\mathbb{R}^3$ », *Philosophical Transactions of the Royal Society of London. Series A : Mathematical, Physical and Engineering Sciences* **360** (1799), p. 2155-2188.

Grenier, E. (2000). « On the nonlinear instability of Euler and Prandtl equations », *Communications on Pure and Applied Mathematics : A Journal Issued by the Courant Institute of Mathematical Sciences* **53** (9), p. 1067-1091.

Guillod, J. et Šverák, V. (2017). « Numerical investigations of non-uniqueness for the Navier–Stokes initial value problem in borderline spaces », arXiv : 1704.00560.

Hopf, E. (1950). « Über die Anfangswertaufgabe für die hydrodynamischen Grundgleichungen. Erhard Schmidt zu seinem 75. Geburtstag gewidmet », *Mathematische Nachrichten* **4** (1-6), p. 213-231.

Jia, H. et Šverák, V. (2014). « Local-in-space estimates near initial time for weak solutions of the Navier–Stokes equations and forward self-similar solutions », *Inventiones mathematicae* **196**, p. 233-265.

——— (2015). « Are the incompressible 3d Navier–Stokes equations locally ill-posed in the natural energy space ? », *Journal of Functional Analysis* **268** (12), p. 3734-3766.

Kato, T. (1984). « Strong L^p solutions of the Navier–Stokes equation in $\mathbb{R}^m$, with applications to weak solutions », *Mathematische Zeitschrift* **187**, p. 471-480.

——— (2013). *Perturbation theory for linear operators*. T. 132. Springer Science & Business Media.

Koch, H. et Tataru, D. (2001). « Well-posedness for the Navier–Stokes equations », *Advances in Mathematics* **157** (1), p. 22-35.

Leray, J. (1934). « Sur le mouvement d'un liquide visqueux emplissant l'espace », *Acta Mathematica* **63**, p. 193-248.

Lions, J.-L. et Prodi, G. (1959). « Un théorème d'existence et unicité dans les équations de Navier- Stokes en dimension 2 », *C. R. Acad. Sci., Paris* **248**, p. 3519-3521.

Merle, F. et al. (2022a). « On the implosion of a compressible fluid I : Smooth self-similar inviscid profiles », *Annals of Mathematics* **196** (2), p. 567-778.

——— (2022b). « On the implosion of a compressible fluid II : Singularity formation », *Annals of Mathematics* **196** (2), p. 779-889.

Perelman, G. (2022). « Finite time blow-up for the compressible fluids and for the energy supercritical defocusing nonlinear Schrödinger equation (after Frank Merle, Pierre Raphaël, Igor Rodnianski and Jérémie Szeftel) », in : *Séminaire Bourbaki. Volume 2021/2022, Exposés 1181–1196*. Paris : Société Mathématique de France (SMF), p. 403-432.

Planchon, F. (1996). « Global strong solutions in Sobolev or Lebesgue spaces to the incompressible Navier–Stokes equations in $\mathbb{R}^3$ », *Annales de l'Institut Henri Poincaré C, Analyse non linéaire* **13** (3), p. 319-336.

Vishik, M. (2018a). « Instability and non-uniqueness in the Cauchy problem for the Euler equations of an ideal incompressible fluid. Part I », arXiv : 1805.09426.

——— (2018b). « Instability and non-uniqueness in the Cauchy problem for the Euler equations of an ideal incompressible fluid. Part II », arXiv : 1805.09440.

Anne-Laure Dalibard

Sorbonne Université
Laboratoire Jacques-Louis Lions
F-75005 Paris
&
École Normale Supérieure
DMA
F-75005 Paris

E-mail : anne-laure.dalibard@sorbonne-universite.fr

Séminaire BOURBAKI Juin 2023
74e année, 2022–2023, no 1209, p. 453 à 480
doi : 10.24033/ast.1218

CATÉGORIES TENSORIELLES SYMÉTRIQUES EN CARACTÉRISTIQUE POSITIVE
[*d'après* Kevin Coulembier, Pavel Etingof, Victor Ostrik...]

par **Daniel Juteau**

> [...] J'ai fini par comprendre comment la notion de motif fournissait la clef d'une compréhension de ce mystère — comment, par le seul fait de la présence d'une catégorie (ici celle des motifs « lisses » sur un schéma de base donné, par exemple les motifs sur un corps de base donné), ayant des structures internes similaires à celles qu'on trouve sur la catégorie des représentations linéaires d'un pro-groupe algébrique sur un corps $\Bbbk$ (le charme de la notion de pro-groupe algébrique m'ayant été révélé précédemment par Serre également), on arrive à reconstituer bel et bien un tel progroupe (dès qu'on dispose d'un « foncteur fibre » convenable), et à interpréter la catégorie « abstraite » comme la catégorie de ses représentations linéaires.
>
> Alexander Grothendieck, *Récoltes et Semailles* (II.B.IV.1).

Introduction

Considérons $\mathbf{G}$ un groupe algébrique (voire pro-algébrique) sur un corps $\Bbbk$ (que nous supposerons algébriquement clos pour simplifier), ainsi que la catégorie $\mathscr{C} = \mathrm{Rep}_{\Bbbk} \mathbf{G}$ des représentations linéaires de $\mathbf{G}$ sur des $\Bbbk$-espaces vectoriels de dimension finie. Quelles sont donc ces structures internes de $\mathscr{C}$ mentionnées par Grothendieck ?

(1) $\mathscr{C}$ est $\Bbbk$-linéaire, abélienne. [(1)]

(2) Tous les objets de $\mathscr{C}$ sont de longueur finie et tous les espaces de morphismes sont de dimension finie.

[(1)] Pour une référence générale sur les catégories, *cf.* MacLane, 1971.

(3) $\mathscr{C}$ est **monoïdale** : elle est munie d'un bifoncteur ($\Bbbk$-bilinéaire) $\mathscr{C} \times \mathscr{C} \to \mathscr{C}$, d'un isomorphisme de foncteurs

$$\alpha \colon (- \otimes -) \otimes - \to - \otimes (- \otimes -),$$

dit contrainte d'associativité, vérifiant une condition de cohérence (axiome du pentagone), et d'un objet unité $\mathbf{1}$ tel que $\mathbf{1} \otimes \mathbf{1} \simeq \mathbf{1}$ et que les foncteurs $\mathbf{1} \otimes -$ et $- \otimes \mathbf{1}$ sont des équivalences.

(4) $\mathscr{C}$ est **tressée** : on a des isomorphismes fonctoriels $c_{X,Y} \colon X \otimes Y \to Y \otimes X$, compatibles avec α (axiomes de l'hexagone).

(5) $\mathscr{C}$ est **symétrique** : le tressage c vérifie $c_{Y,X} \circ c_{X,Y} = \mathrm{id}_{X \otimes Y}$.

(6) $\mathscr{C}$ est **rigide** : tout objet X est dualisable, *i.e.* il admet un dual [(2)] X^* muni de morphismes $\mathrm{ev}_X \colon X^* \otimes X \to \mathbf{1}$ et $\mathrm{coev}_X \colon \mathbf{1} \to X \otimes X^*$, tels que les composés

$$(\mathrm{id}_X \otimes \mathrm{ev}_X) \circ (\mathrm{coev}_X \otimes \mathrm{id}_X) \colon X \longrightarrow X \otimes X^* \otimes X \to X,$$
$$(\mathrm{ev}_X \otimes \mathrm{id}_{X^*}) \circ (\mathrm{id}_{X^*} \otimes \mathrm{coev}_X) \colon X^* \longrightarrow X^* \otimes X \otimes X^* \to X^*$$

sont l'identité de X et de X^* respectivement.

(7) $\mathscr{C}$ admet un **foncteur fibre** $F \colon \mathscr{C} \to \mathrm{Vec}_\Bbbk$, à savoir le foncteur d'oubli, à valeurs dans la catégorie des $\Bbbk$-espaces vectoriels de dimension finie (muni du produit tensoriel usuel, avec les contraintes d'associativité et de commutativité évidentes). Cela veut dire que c'est un foncteur exact, monoïdal (préservant l'unité et muni d'un isomorphisme fonctoriel $J \colon F(X) \otimes F(Y) \xrightarrow{\sim} F(X \otimes Y)$ compatible aux contraintes d'associativité et aux unités) et symétrique (on demande que J soit aussi compatible aux contraintes de commutativité). Il résulte de la rigidité qu'un tel foncteur est automatiquement fidèle.

Une **catégorie tensorielle symétrique** (sur $\Bbbk$) est une catégorie monoïdale symétrique rigide, abélienne et $\Bbbk$-linéaire (avec un produit tensoriel $\Bbbk$-bilinéaire).

Si de plus on a $\mathrm{End}(\mathbf{1}) = \Bbbk$ et que tous les objets sont de longueur finie, on dira que c'est une catégorie **prétannakienne**.

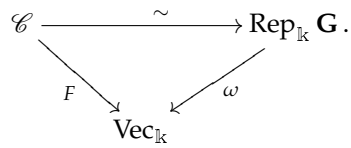
Une catégorie prétannakienne munie d'un foncteur fibre à valeurs dans $\mathrm{Vec}_\Bbbk$ est dite **tannakienne**. [(3)] Un résultat fondamental du formalisme est que toute catégorie tannakienne est en fait une catégorie de représentations pour un schéma en groupes affine.

[(2)] Notons que la définition de dual ci-dessus est celle d'un dual à gauche (qui signifie que le foncteur $X^* \otimes -$ est adjoint à gauche du foncteur $X \otimes -$) ; en fait, la rigidité demande des duaux à gauche et à droite, mais dans le cas tressé (voire symétrique), un dual à gauche est aussi naturellement un dual à droite en utilisant la contrainte de commutativité.

[(3)] Il faudrait préciser : tannakienne neutre. La notion de catégorie tannakienne est plus générale (le foncteur fibre pouvant être à valeurs dans Mod_A, pour A une $\Bbbk$-algèbre), mais en fait, dans la suite, la plupart du temps nous supposerons $\Bbbk$ algébriquement clos et on ne s'intéressera qu'au cas neutre ; par conséquent, nous dirons simplement catégorie tannakienne pour cette notion. Voir (Deligne et Milne, 1982, §2 et 3) pour une discussion détaillée des foncteurs fibres.

Théorème 0.1 (Deligne et Milne, 1982 ; Saavedra Rivano, 1972). *Soit $\mathscr{C}$ une catégorie tannakienne sur $\Bbbk$, avec foncteur fibre $F\colon \mathscr{C} \to \mathrm{Vec}_\Bbbk$. Alors le foncteur qui à une $\Bbbk$-algèbre R associe le groupe des automorphismes tensoriels symétriques de $F_R\colon \mathscr{C} \to \mathrm{Mod}_R$, $X \mapsto F(X) \otimes_\Bbbk R$ est représentable par un $\Bbbk$-schéma en groupes affine $\mathbf{G} := \underline{\mathrm{Aut}}^{\otimes}(F)$.*

De plus, le foncteur fibre F se factorise par une équivalence (tensorielle symétrique) suivie du foncteur d'oubli :



La terminologie « foncteur fibre » s'explique par le cas où $\mathscr{C}$ est la catégorie des $\Bbbk$-systèmes locaux sur un espace topologique X raisonnable : pour chaque point $x \in X$, on a un « foncteur fibre au point x », sur lequel le groupe fondamental $\pi_1(X, x)$ agit. On a alors une équivalence de catégories monoïdales symétriques entre les systèmes locaux de rang fini et les représentations de $\pi_1(X, x)$. Attention, si on applique la reconstruction tannakienne dans le cadre linéaire comme ci-dessus, on ne retrouve pas $\pi_1(X, x)$, mais seulement sa complétion proalgébrique (un schéma en groupes affine est toujours limite projective de groupes algébriques). Un précurseur était la théorie de Galois–Grothendieck du groupe fondamental, où on regarde plutôt les revêtements finis et des actions de $\pi_1(X, x)$ sur des ensembles finis (ou plutôt des actions continues de son complété profini). On voit que dans ce cas le groupe fondamental est bien défini à un automorphisme intérieur près. L'objet plus canonique serait le groupoïde fondamental ; les isomorphismes entre deux foncteurs fibres $\pi(X, x)$ et $\pi(X, y)$ sont décrits par le torseur $\pi_1(X, x, y)$ des chemins à homotopie près de x à y.

Le « rêve des motifs » de Grothendieck avait plusieurs sources et motivations : notamment la géométrie énumérative, la recherche d'un invariant cohomologique universel à coefficients dans $\mathbb{Q}$ factorisant toutes les cohomologies de Weil, et la théorie de Galois (André, 2004). Grothendieck ne s'est pas laissé abattre par l'argument bien connu de Serre montrant qu'on ne peut pas avoir d'invariant cohomologique prenant ses valeurs dans $\mathrm{Vec}_\mathbb{Q}$ (en considérant les endomorphismes de courbes elliptiques super-singulières en caractéristique $p > 0$) : il a imaginé qu'il puisse exister une catégorie abélienne $\mathbb{Q}$-linéaire, avec les structures appropriées, qui fasse l'affaire ; notamment, il faut pouvoir exprimer la propriété de Künneth, donc il faut avoir un produit tensoriel. Les différents foncteurs de réalisation des motifs donnant les différentes cohomologies classiques sont des foncteurs fibres. Cette approche donne lieu à un groupe de Galois motivique, vaste généralisation de la théorie de Galois en dimension supérieure.

En plus de fournir le cadre conceptuel pour la théorie des motifs, les catégories tensorielles symétriques sont intéressantes à bien des égards. Chacune de ces catégories donne un cadre où on peut faire de l'algèbre commutative et de la géométrie algébrique : ainsi, on peut donner un sens à la notion de schéma en groupes affine dans une telle catégorie ; par exemple, si on remplace la catégorie $\mathrm{Vec}_{\Bbbk}$ par celle des super-espaces vectoriels $\mathrm{sVec}_{\Bbbk}$, on obtient les notions de super-groupe algébrique, de super-algèbre de Lie, etc. On peut aussi dire, comme Etingof, que dans un premier temps, on peut faire de la théorie des représentations sans groupes (groupes quantiques, etc), puis même sans espaces vectoriels. Vu le théorème de reconstruction tannakienne, on peut voir les catégories tensorielles symétriques comme des généralisations des groupes (pro)algébriques.

Un problème naturel est de tenter de classifier ces catégories tensorielles symétriques. On a vu que beaucoup d'entre elles peuvent être décrites comme les représentations d'un groupe. Comment caractériser ces catégories tannakiennes, parmi toutes les autres ? En caractéristique 0, une réponse élégante a été donnée par Deligne (1990), puis dans le cas super : Deligne (2002, 2011). Des exemples non (super-)tannakiens ont aussi été étudiés (Deligne, 2007 ; Harman et Snowden, 2022 ; Knop, 2007).

En caractéristique $p > 0$, des exemples de catégories non super-tannakiennes, et pourtant à croissance modérée (*cf.* §1.3), étaient connus (Gelfand et Kazhdan, 1992 ; Georgiev et Mathieu, 1994) : il s'agit des catégories Ver_p. Mais pendant longtemps, il n'était pas clair comment on pourrait obtenir un analogue du critère de Deligne en caractéristique p. Ces dernières années, d'énormes progrès ont été accomplis. Il est naturel de chercher à caractériser les catégories Ver_p-tannakiennes (admettant un « foncteur fibre » vers Ver_p). Sans surprise, un mot clé est Frobenius. On peut définir plusieurs foncteurs tâchant d'imiter le twist de Frobenius des représentations des groupes réductifs en caractéristique p. Il s'avère que la notion de catégorie Frobenius-exacte (pour laquelle un de ces foncteurs de Frobenius, peu importe lequel, est exact) est cruciale et apparaît dans les critères dégagés par Coulembier, Etingof et Ostrik (2023b) pour caractériser les catégories Ver_p-tannakiennes. Mais l'histoire ne s'arrête pas là : Benson et Etingof (2019) et Benson, Etingof et Ostrik (2023) dévoilent une richesse foisonnante insoupçonnée parmi ces catégories tensorielles symétriques en caractéristique p, et il y a bien sûr encore beaucoup de questions ouvertes.

Dans la section 1, nous commencerons de manière très naïve, en approchant la notion de dimension de façon intrinsèque à partir de la structure monoïdale symétrique, car cela donne beaucoup de conditions pour pouvoir être une catégorie (super-)tannakienne. Au fur et à mesure, nous verrons des exemples et contre-exemples. Puis dans la section 2, nous rappellerons les résultats connus en caractéristique 0. Dans la section 3, nous expliquerons la procédure de semi-simplification, qui permet de définir la catégorie Ver_p. Dans la section 4, nous parlerons des différents

foncteurs de Frobenius et d'un analogue (partiel) au théorème de Deligne en caractéristique p, dans le cas des catégories Frobenius-exactes. Dans la section 5, nous décrirons les nouvelles catégories tensorielles symétriques découvertes récemment, qui ne sont ni semi-simples ni Frobenius-exactes, et mentionnerons une conjecture qui nous dit que cette liste pourrait suffire pour décrire, via le formalisme tannakien, toutes les catégories tensorielles symétriques à croissance modérée en caractéristique p. Enfin, dans la section 6, nous verrons des applications aux représentations modulaires des groupes finis.

1. Dimensions

Comment pouvons-nous décrire la dimension d'un espace vectoriel V dans $\mathrm{Vec}_{\Bbbk}$, uniquement en termes de la structure tensorielle symétrique de cette catégorie ? En réalité, plusieurs moyens sont à notre disposition et cela donnera lieu à plusieurs notions de « dimension » dans une catégorie tensorielle symétrique $\mathscr{C}$ quelconque. Ces notions seront préservées par les foncteurs tensoriels symétriques ; nous obtiendrons donc des conditions nécessaires sur $\mathscr{C}$ pour qu'elle soit tannakienne.

1.1. La trace de l'identité

Première idée : la dimension, c'est la trace de l'identité ! Dans $\mathrm{Vec}_{\Bbbk}$, la coévaluation $\mathrm{coev}_V \colon \Bbbk \to V \otimes V^* \simeq \mathrm{End}_{\Bbbk}(V)$ envoie 1 sur id_V et l'évaluation $\mathrm{ev}_V \colon \mathrm{End}_{\Bbbk}(V) \simeq V^* \otimes V \to \Bbbk$ s'interprète comme la trace. Cela nous suggère comment définir la trace d'un endomorphisme dans une catégorie tensorielle symétrique générale.

Définition 1.1. On définit la trace d'un endomorphisme $f \in \mathrm{End}_{\mathscr{C}}(X)$ par

$$\mathrm{Tr}\, f = \mathrm{ev}_X \circ c_{X,X^*} \circ (f \otimes \mathrm{id}_{X^*}) \circ \mathrm{coev}_X \in \mathrm{End}(\mathbf{1}) = \Bbbk,$$

et en particulier la dimension de X par

$$\dim(X) := \mathrm{Tr}(\mathrm{id}_X) \in \Bbbk.$$

Voilà un bel invariant, conservé par tout foncteur tensoriel symétrique ! Ah oui... Un bémol : il est à valeurs dans $\Bbbk$. Donc c'est très bien si $\Bbbk$ est de caractéristique 0, mais s'il est de caractéristique p on n'a accès qu'à la dimension modulo p. En tout cas, si $\dim(X)$ n'est pas l'image d'un entier naturel dans $\Bbbk$, c'est très clairement une obstruction à l'existence d'un foncteur fibre.

Voici un exemple basique qui ne vérifie pas cette condition, si $\Bbbk$ est un corps de caractéristique 0 : la catégorie $\mathrm{sVec}_{\Bbbk}$ des $\Bbbk$-espaces vectoriels $\mathbb{Z}/2$-gradués $V = V_0 \oplus V_1$, avec composante paire V_0 et composante impaire V_1 de dimension finie, munie du

produit tensoriel évident avec la contrainte d'associativité habituelle, mais avec la règle de Koszul comme contrainte de commutativité : pour des éléments homogènes x et y de degrés $|x|$ et $|y|$,

$$c_{X,Y}(x \otimes y) = (-1)^{|x||y|} y \otimes x.$$

On vérifie facilement que $\dim(V) = \dim_{\Bbbk} V_0 - \dim_{\Bbbk} V_1 \in \mathbb{Z}$ peut prendre des valeurs aussi bien positives que négatives. La catégorie $\mathrm{sVec}_{\Bbbk}$ n'est donc pas tannakienne. Remarquons qu'on a un foncteur $\Bbbk$-linéaire, monoïdal, exact et fidèle $F\colon \mathrm{sVec}_{\Bbbk} \to \mathrm{Vec}_{\Bbbk}$, qu'on pourrait noter $V \mapsto |V|$, mais bien sûr il n'est pas symétrique. On notera $\dim_{\Bbbk} V = \dim_{\Bbbk} |V| = \dim_{\Bbbk} V_0 + \dim_{\Bbbk} V_1$. Si on note $\overline{\mathbf{1}}$ la droite impaire, tout objet de $\mathrm{sVec}_{\Bbbk}$ est donc de la forme $\mathbf{1}^m \oplus \overline{\mathbf{1}}^n$; un tel objet est dit de dimension $m|n$.

Les objets de $\mathrm{sVec}_{\Bbbk}$ sont appelés super-espaces vectoriels (de dimension finie). Remarquons qu'en prenant un objet de type algèbre de Lie, algèbre de Hopf, etc, dans $\mathrm{sVec}_{\Bbbk}$ (ce sont des notions qui ont un sens dans n'importe quelle catégorie tensorielle symétrique), on obtient en appliquant le foncteur $F\colon V \mapsto |V|$ les notions correspondantes en termes classiques : super-algèbre de Lie, super-algèbre de Hopf, etc. On peut aussi faire de la super-géométrie algébrique : la catégorie des super-schémas affines est la catégorie opposée de celle des algèbres commutatives dans $\mathrm{Ind}\,\mathrm{sVec}_{\Bbbk}$, cette dernière étant la catégorie des ind-super-espaces vectoriels (super-espaces vectoriels de dimension quelconque). Quant aux super-schémas en groupes affines, on peut les voir comme des objets en groupes dans les super-schémas affines, ou comme des algèbres de Hopf dans $\mathrm{Ind}\,\mathrm{sVec}_{\Bbbk}$.

1.2. La longueur

Cherchons maintenant des invariants à valeurs dans $\mathbb{N}$, voire dans $\mathbb{R}^+$, ce qui nous permettra de mieux contrôler la taille de nos objets, y compris en caractéristique p. En fait, il faut aussi admettre la valeur ∞ (qui sera sans doute le signe d'une situation pas très tannakienne...).

La longueur $\ell(X)$ d'un objet X de $\mathscr{C}$ est la plus grande longueur possible d'une chaîne d'inclusions strictes entre sous-objets, ou ∞ si cette longueur n'est pas bornée. Comme $\mathscr{C}$ est abélienne, pour un objet de longueur finie, une chaîne de longueur maximale est une suite de composition, c'est-à-dire une filtration dont tous les sous-quotients sont simples ; et on a le théorème de Jordan–Hölder : toutes les suites de composition sont de même longueur et les objets simples (à isomorphisme près) apparaissant comme sous-quotients, ainsi que leurs multiplicités, ne dépendent pas du choix de la suite de composition.

Si $F\colon \mathscr{C} \to \mathscr{D}$ est un foncteur exact et fidèle (par exemple un foncteur fibre), il ne peut que faire augmenter la longueur : si X se dévisse en les objets simples S_i, par

exactitude $F(X)$ a une filtration de quotients successifs les $F(S_i)$, qui par fidélité sont non nuls et donc de longueur au moins égale à 1. On a donc

$$\ell(X) \leqslant \ell(F(X)). \tag{1}$$

La rigidité de $\mathscr{C}$ implique que $X \otimes Y \neq 0$ dès que $X \neq 0$ et $Y \neq 0$. En effet, les norphismes de coévaluation $\mathrm{coev}_X \colon \mathbf{1} \to X \otimes X^*$ et $\mathrm{coev}_Y \colon \mathbf{1} \to Y \otimes Y^*$ sont alors non nuls, donc des monomorphismes car $\mathbf{1}$ est simple (Deligne et Milne, 1982, Prop. 1.17). D'autre part, le foncteur $X \otimes -$ est exact, puisqu'il a des adjoints à gauche et à droite. Donc $\mathbf{1} \hookrightarrow X \otimes X^* \hookrightarrow X \otimes X^* \otimes Y \otimes Y^* \simeq X \otimes Y \otimes Y^* \otimes X^*$, ce qui implique $X \otimes Y \neq 0$. On en déduit que

$$\ell(X \otimes Y) \geqslant \ell(X)\ell(Y). \tag{2}$$

En effet, si X et Y ont des filtrations avec n et m sous-quotients non nuls respectivement, alors la remarque précédente nous donne une filtration sur $X \otimes Y$ ayant nm sous-quotients non nuls. En particulier, si X est de longueur infinie et $Y \neq 0$, alors $X \otimes Y$ est aussi de longueur infinie.

Dans $\mathrm{Vec}_{\Bbbk}$, le seul objet simple est $\Bbbk$ et la longueur est égale à la dimension. En vue de (1), si $\mathscr{C}$ contient un objet de longueur infinie, c'est clairement une obstruction à l'existence d'un foncteur fibre. Reprenons l'exemple de Deligne (1990, §2.19). Prenons $\Bbbk$ de caractéristique 0 contenant t transcendant sur $\mathbb{Q}$. La catégorie (GL_t) est librement engendrée par un objet X_t de dimension t : les objets sont les $X_t^{\otimes a} \otimes (X_t^*)^{\otimes b}$ et on peut représenter les morphismes par des diagrammes, avec pour générateurs ev_X et coev_X (« caps and cups » orientés) et pour unique relation $\dim(X_t) = t$, c'est-à-dire qu'un cercle peut être remplacé par le scalaire t. L'objet X_t est « l'objet universel de dimension t » : si X est un autre objet de dimension t dans une catégorie tensorielle symétrique $\mathscr{T}$ sur $\Bbbk$, alors il existe un foncteur tensoriel symétrique exact de (GL_t) vers $\mathscr{T}$ envoyant X_t sur X, unique à automorphisme tensoriel symétrique près (induit par un automorphisme de X). On a donc une suite de foncteurs tensoriels symétriques exacts

$$(\mathrm{GL}_t) \longrightarrow (\mathrm{GL}_{t-1}) \longrightarrow (\mathrm{GL}_{t-2}) \longrightarrow \cdots \longrightarrow (\mathrm{GL}_{t-n}) \longrightarrow \cdots$$

$$X_t \longmapsto \mathbf{1} \oplus X_{t-1} \longmapsto \mathbf{1}^2 \oplus X_{t-2} \longmapsto \cdots \longmapsto \mathbf{1}^n \oplus X_{t-n} \longmapsto \cdots$$

L'algèbre d'endomorphismes de $X_t = \mathbf{1}^n \oplus X_{t-n}$ dans (GL_{t-n}) est $M_n(k) \times k$.

Soit $\mathscr{C}$ la limite inductive des catégories (GL_{t-n}). C'est encore une catégorie tensorielle symétrique. On peut la voir comme librement engendrée par un objet X_t de dimension t muni de décompositions $X_t = \mathbf{1} \oplus X_{t-1}$, $X_{t-1} = \mathbf{1}^2 \oplus X_{t-2}$, etc. Dans cet exemple, X_t n'est pas de longueur finie, et $\mathrm{End}_{\mathscr{C}}(X_t) = M_\infty(\Bbbk) \times \Bbbk$ est de dimension infinie. La catégorie $\mathscr{C}$ n'est donc clairement pas tannakienne.

1.3. La dimension de croissance

Jusqu'ici, nous avons considéré X de façon isolée. Mais bien sûr, il est intéressant de considérer ses puissances tensorielles $X^{\otimes n}$, par exemple on peut se demander comment cet objet « croît » avec n. Posons

$$d_n(X) := \ell(X^{\otimes n}).$$

On dit que X est à **croissance modérée** s'il existe une constante réelle C telle que $d_n(X) \leqslant C^n$ et que $\mathscr{C}$ est à croissance modérée si c'est le cas pour tous ses objets. Remarquons que l'existence d'un foncteur fibre $F\colon \mathscr{C} \to \mathrm{Vec}_{\Bbbk}$, voire d'un superfoncteur fibre $F\colon \mathscr{C} \to \mathrm{sVec}_{\Bbbk}$, implique que $\mathscr{C}$ est à croissance modérée : d'après (1), on a $d_n(X) = \ell(X^{\otimes n}) \leqslant \ell(F(X^{\otimes n})) = (\dim_{\Bbbk} F(X))^n$.

On peut donc se dire qu'on obtiendra une bonne notion de dimension en étudiant plus précisément le taux de croissance de $d_n(X)$. D'après (2), on a

$$d_{n+m}(X) \geqslant d_n(X) d_m(X)$$

Si $d_n(X) = \infty$ pour un certain n, alors on a encore $d_m(X) = \infty$ pour tout $m \geqslant n$. Supposons au contraire que les $d_n(X)$ soient finis. Alors la suite $(1/d_n(X))$ est sous-multiplicative et le lemme de Fekete (Benson, 2020a, Lemme 1.6.3) implique que la limite suivante existe :

$$\mathrm{gd}(X) := \lim_{n\to\infty} d_n(X)^{1/n} = \sup_n d_n(X)^{1/n} \in \mathbb{R}^+ \cup \{+\infty\}. \tag{3}$$

On l'appelle la dimension de croissance de X. On a $\mathrm{gd}(X) \geqslant 1$ dès que $X \neq 0$. Un objet est à croissance modérée si et seulement si $\mathrm{gd}(X) < \infty$. Pour V dans $\mathrm{Vec}_{\Bbbk}$ ou $\mathrm{sVec}_{\Bbbk}$, on a $\mathrm{gd}(V) = \dim_{\Bbbk} V$.

1.4. La dimension alternée

Une autre idée pour caractériser la dimension d'un espace vectoriel : c'est la dernière puissance alternée non nulle. Il serait quand même temps d'utiliser le fait que comme $\mathscr{C}$ est une catégorie tensorielle symétrique, le groupe $\mathfrak{S}_n$ agit sur $X^{\otimes n}$... Soit

$$a_n := \sum_{\sigma\in\mathfrak{S}_n} \mathrm{sgn}(\sigma)\sigma \in \Bbbk\mathfrak{S}_n \tag{4}$$

l'élément antisymétrisant. Si $\Bbbk$ est de caractéristique 0 ou $> n$, on dispose de l'idempotent central $\frac{1}{n!}a_n$, qui agit sur $X^{\otimes n}$ comme la projection sur la composante sgn-isotypique, et son image donne la seule définition possible d'une puissance extérieure.

Mais pour $\Bbbk$ de caractéristique $p > 0$, on peut envisager trois versions de la n-ième puissance extérieure : les anti-invariants Λ^n (plus grand sous-objet sur lequel $\mathfrak{S}_n$

agit par le caractère sgn), l'image A_n de a_n, ou les co-anti-invariants $\wedge^n$ (plus grand quotient sur lequel $\mathfrak{S}_n$ agit par le caractère sgn). On a des morphismes canoniques

$$\wedge^n X \twoheadrightarrow A^n X \hookrightarrow \Lambda^n X.$$

Pour $\mathscr{C} = \mathrm{Vec}_{\Bbbk}$, ce sont des isomorphismes ; mais ce n'est pas vrai en général, par exemple pour $\mathrm{sVec}_{\Bbbk}$ si $p > 2$. Coulembier, Etingof et Ostrik (2023b) préfèrent utiliser A^n. Ils posent :

$$\mathrm{ad}(X) := \sup \{n \in \mathbb{N} \mid A^n X \neq 0\} \in \mathbb{N} \cup \{\infty\}.$$

Notons qu'en caractéristique 0, la dimension alternée ad permet de faire une distinction entre $\mathrm{Vec}_{\Bbbk}$ et $\mathrm{sVec}_{\Bbbk}$: pour un super-espace vectoriel purement impair V, on a $|\Lambda^n V| = \mathrm{Sym}^n |V|$, donc $\mathrm{ad}(V) = \infty$. Par contre, on peut annuler un tel V avec une puissance symétrique. Plus généralement, un super-espace vectoriel pourra être annulé par un foncteur de Schur.

1.5. La dimension de Schur

Supposons temporairement que $\Bbbk$ soit de caractéristique 0. Dans ce cas, pour n dans $\mathbb{N}$, l'algèbre de groupe $\Bbbk\mathfrak{S}_n$ est semi-simple. On rappelle que les représentations irréductibles de $\Bbbk\mathfrak{S}_n$ sont paramétrées par les partitions λ de n (James et Kerber, 1981). On les notera V_λ. En particulier, $V_{(n)} = \Bbbk$ est la représentation triviale et $V_{(1^n)} = \mathrm{sgn}$ est la représentation signe. Pour chaque λ, on a un foncteur de Schur $\mathbb{S}_\lambda : \mathscr{C} \to \mathscr{C}$ (voir Deligne, 2002, §1 pour plus de détails) :

$$\mathbb{S}_\lambda X := (V_\lambda^* \otimes X^{\otimes n})^{\mathfrak{S}_n}$$

permettant de décomposer $X^{\otimes n}$ en composantes V_λ-isotypiques :

$$\bigoplus_\lambda V_\lambda \otimes \mathbb{S}_\lambda(X) \xrightarrow{\sim} X^{\otimes n}. \tag{5}$$

Parmi les facteurs directs, on retrouve

$$\mathbb{S}_{(n)} X = \mathrm{Sym}^n X \quad \text{et} \quad \mathbb{S}_{(1^n)} X = \Lambda^n X.$$

Un espace vectoriel V dans $\mathrm{Vec}_{\Bbbk}$, de dimension d, est annulé par le foncteur de Schur $\mathbb{S}_\lambda$ si et seulement si λ a plus de d lignes. Donc si $n \leqslant d$, tous les $\mathbb{S}_\lambda V$ sont non nuls ; pour $n > d$, au moins un $\mathbb{S}_\lambda$ annule V. Par exemple, pour $n = d+1$, le foncteur $\mathbb{S}_{(1^{d+1})} = \Lambda^{d+1}$ annule V. Autrement dit, le morphisme canonique, $\Bbbk\mathfrak{S}_n \simeq \bigoplus_\lambda V_\lambda^* \otimes V_\lambda \to \mathrm{End}_{\Bbbk}(V^{\otimes n})$, est injectif pour $n \leqslant d$ et ne l'est plus pour $n > d$.

On a supposé $\Bbbk$ de caractéristique 0 pour disposer des foncteurs de Schur, mais la dernière reformulation a un sens en toute généralité. On pose

$$\mathrm{sd}(X) := \sup\{n \in \mathbb{N} \mid \mathrm{can}\colon \Bbbk\mathfrak{S}_n \to \mathrm{End}_{\mathscr{C}}(X^{\otimes n}) \text{ est injectif}\} \in \mathbb{N} \cup \{\infty\},$$

où can désigne l'action canonique de $\Bbbk\mathfrak{S}_n$ sur $X^{\otimes n}$, déduite de la structure de catégorie tensorielle symétrique.

Si $\mathrm{sd}(X) = \infty$, alors

$$\dim_{\Bbbk} \mathrm{Hom}_{\mathscr{C}}(\mathbf{1}, (X^* \otimes X)^{\otimes n}) = \dim_{\Bbbk} \mathrm{End}_{\mathscr{C}}(X^{\otimes n}) \geqslant \dim_{\Bbbk} \Bbbk\mathfrak{S}_n = n!,$$

donc le socle de $(X^* \otimes X)^{\otimes n}$ (*i.e.* son plus grans sous-objet semi-simple) contient au moins $n!$ copies de $\mathbf{1}$, la longueur $\ell((X^* \otimes X)^{\otimes n}) \geqslant n!$ croît plus vite que toute exponentielle, et $\mathscr{C}$ n'est pas à croissance modérée.

Pour $\Bbbk$ de caractéristique zéro, en utilisant (5), on pourrait en déduire directement que X lui-même n'est pas à croissance modérée, car alors

$$\ell(X^{\otimes n}) \geqslant \sum_{\lambda} \dim(V_\lambda) \geqslant \Big(\sum_{\lambda} \dim(V_\lambda)^2\Big)^{1/2} = (n!)^{1/2}.$$

1.6. Relations entre les différentes notions de dimension

Les différentes notions de dimension que nous avons vues sont reliées de plusieurs manières.

Proposition 1.2 (Coulembier, Etingof et Ostrik, 2023b, Prop. 4.7). *Pour X dans $\mathscr{C}$, on a*

(i) $\mathrm{sd}(X) \leqslant \mathrm{ad}(X)$;

(ii) $\ell(X) \leqslant \mathrm{sd}(X \otimes X^*)$;

(iii) $\mathrm{gd}(X \otimes X^*) < \infty \Longrightarrow \mathrm{sd}(X) < \infty$;

(iv) $\ell(X) \leqslant \mathrm{ad}(X)$;

(v) $\mathrm{gd}(X) \leqslant \mathrm{ad}(X)$;

(vi) *Si de plus $\Bbbk$ est de caractéristique $p > 0$, alors* $\mathrm{sd}(X) < \infty \Longrightarrow \mathrm{ad}(X) < \infty$.

Il se passe quelque chose de différent en caractéristique p : notamment, la dimension alternée ne permettra plus de distinguer les catégories tannakiennes des catégories super-tannakiennes !

Corollaire 1.3. *Pour une catégorie tensorielle symétrique $\mathscr{C}$ sur $\Bbbk$ de caractéristique $p > 0$, les conditions suivantes sont équivalentes :*

(i) $\mathrm{gd}(X) < \infty$ *pour tout X dans $\mathscr{C}$;*

(ii) $\mathrm{sd}(X) < \infty$ *pour tout X dans $\mathscr{C}$;*

(iii) $\mathrm{ad}(X) < \infty$ *pour tout X dans $\mathscr{C}$.*

2. Caractérisations internes en caractéristique zéro

Dans toute cette section, $\Bbbk$ est supposé de caractéristique zéro.

2.1. Catégories tannakiennes

Théorème 2.1 (DELIGNE, 1990). *Soit $\mathscr{C}$ une catégorie prétannakienne* [4] *sur $\Bbbk$ supposé de caractéristique nulle. Alors les conditions suivantes sont équivalentes :*

(i) *$\mathscr{C}$ est tannakienne ;*

(ii) *Pour tout X objet de $\mathscr{C}$, on a $\dim(X) \in \mathbb{N}$;*

(iii) *Pour tout X objet de $\mathscr{C}$, il existe un $n \in \mathbb{N}$ tel que $\Lambda^n X = 0$.*

En caractéristique zéro, il n'y a pas lieu de distinguer entre différents types de puissances alternées. On note donc $\Lambda^n X$ l'image du projecteur antisymétrisant

$$\frac{1}{n!} \sum_{\sigma \in \mathfrak{S}_n} \operatorname{sgn}(\sigma)\sigma$$

agissant sur $X^{\otimes n}$. On montre d'abord que la dimension de $\Lambda^n X$ est donnée par la même formule que pour les espaces vectoriels.

Lemme 2.2. *Pour X objet de $\mathscr{C}$, on a*

$$\dim(\Lambda^n X) = \binom{\dim(X)}{n} = \frac{\dim(X)(\dim(X)-1)\cdots(\dim(X)-n+1)}{n!}.$$

Démonstration. On commence par montrer que la trace sur $X^{\otimes n}$ d'une permutation $\sigma \in \mathfrak{S}_n$ ayant k orbites [5] est $(\dim(X))^k$ (on peut s'en convaincre facilement par la diagrammatique des catégories tensorielles symétriques : prendre la trace revient à fermer le diagramme de σ ; chaque orbite donne une boucle et donc un facteur $\dim(X)$). Il en résulte qu'il existe un polynôme universel $P_n \in \mathbb{Q}[T]$ (valable pour toute catégorie prétannakienne $\mathscr{C}$), tel que pour tout X dans $\mathscr{C}$, on ait $\dim(\Lambda^n X) = P_n(\dim(X))$, à savoir

$$P_n(T) = \frac{1}{n!} \sum_{\sigma \in \mathfrak{S}_n} \operatorname{sgn}(\sigma) T^{c(\sigma)},$$

où $c(\sigma)$ est le nombre d'orbites de σ. En considérant $\mathscr{C} = \mathrm{Vec}_{\Bbbk}$, on connaît une infinité de valeurs du polynôme P_n, ce qui suffit à le déterminer. □

[4] Nous avons inclus dans la définition de catégorie prétannakienne le fait que les objets doivent être de longueur finie, mais il n'est pas nécessaire de le supposer dans cet énoncé, cela découle du reste.

[5] On dit « orbites » plutôt que « cycles » pour qu'il soit bien clair qu'on compte aussi les points fixes (orbites de cardinal 1), qui ne sont pas des cycles dans la définition standard.

Lemme 2.3. *On suppose toujours $\Bbbk$ de caractéristique zéro. Si tout objet X de $\mathscr{C}$ vérifie $\dim(X) \in \mathbb{N}$, alors tout objet de dimension nulle est nul.*

Démonstration. Soit $X \neq 0$ dans $\mathscr{C}$. Alors $\mathrm{id}_X \neq 0$, donc par adjonction $0 \neq \mathrm{coev}_X : \mathbf{1} \to X \otimes X^*$ et c'est nécessairement un monomorphisme puisque $\mathbf{1}$ est simple. Par hypothèse,

$$\dim \operatorname{coker} \mathrm{coev}_X = \dim(X)\dim(X^*) - 1 \geqslant 0,$$

ce qui implique que $\dim(X) \geqslant 1$. □

Début de la preuve du théorème. Montrons d'abord que (ii) implique (iii). Soit X un objet de $\mathscr{C}$. Par hypothèse, sa dimension est un entier $n \in \mathbb{N}$. Il en résulte que $\dim \Lambda^{n+1} X = 0$, d'où $\Lambda^{n+1} X = 0$ par le lemme précédent.

Réciproquement, si un objet X vérifie $\Lambda^n X = 0$, on en déduit que $\dim \Lambda^n X = \binom{\dim(X)}{n} = 0$, ce qui implique que $\dim(X)$ est un entier naturel compris entre 0 et $n-1$. Donc (iii) implique (ii).

Il est clair qu'une catégorie tannakienne vérifie (ii) et (iii). La preuve de la réciproque est beaucoup plus longue ! □

2.2. Catégories super-tannakiennes

Soit $\mathbf{G}$ un super-schéma en groupes affine sur $\Bbbk$. Cela revient à se donner une algèbre de Hopf commutative $\mathscr{O}(G)$ dans $\mathrm{sVec}_\Bbbk$. Étant donné un élément $z \in \mathbf{G}(\Bbbk)$ d'ordre divisant 2 agissant par l'automorphisme de parité sur $\mathscr{O}(G)$ (c'est-à-dire par $+1$ sur la partie paire et par -1 sur la partie impaire), on considère la catégorie $\mathrm{Rep}_\Bbbk(\mathbf{G}, z)$ des super-représentations (*i.e.* comodules sur l'algèbre $\mathscr{O}(G)$, dans $\mathrm{sVec}_\Bbbk$) de dimension finie sur lesquelles z agit par l'automorphisme de parité.

Théorème 2.4 (Coulembier, 2020 ; Deligne, 2002, 2011). *On suppose $\Bbbk$ de caractéristique 0. Soit $\mathscr{C}$ une catégorie prétannakienne sur $\Bbbk$. Alors les conditions suivantes sont équivalentes :*

(i) *$\mathscr{C}$ est super-tannakienne ;*

(ii) *tout X objet de $\mathscr{C}$ est à croissance modérée ;*

(iii) *pour tout X objet de $\mathscr{C}$, il existe un $n \in \mathbb{N}$ et une partition λ de n tels que le foncteur de Schur $\mathbb{S}_\lambda$ annule X.*

3. Une catégorie modérée, mais pas super !

On peut maintenant se poser la question : en caractéristique p, existe-t-il des catégories tensorielles symétriques qui ne soient pas super-tannakiennes, bien qu'elles soient à croissance modérée ? La réponse est oui ! Pour les décrire, nous avons besoin d'introduire le concept de semi-simplification.

3.1. Semi-simplification d'une catégorie tensorielle symétrique

La notion de semi-simplification d'une catégorie tensorielle sphérique est introduite par Barrett et Westbury (1999) et développée dans Etingof et Ostrik (2022), où notamment de nombreux exemples importants sont calculés. En fait, en géométrie algébrique, on trouve déjà un exemple de cette notion dans la preuve de Jannsen (1992) de la semi-simplicité des motifs numériques effectifs pour les variétés projectives lisses.

Un idéal tensoriel dans une catégorie $\Bbbk$-linéaire monoïdale $\mathscr{C}$ est une collection $\mathscr{N}$ de sous-espaces $\mathscr{N}(X,Y) \subset \mathrm{Hom}_{\mathscr{C}}(X,Y)$ (pour X et Y objets de $\mathscr{C}$), qui est stable par composition et produit tensoriel avec des morphismes arbitraires de $\mathscr{C}$. On peut alors définir une catégorie monoïdale quotient $\mathscr{C}/\mathscr{N}$ qui a les mêmes objets que $\mathscr{C}$ et pour espaces de morphismes $\mathrm{Hom}_{\mathscr{C}/\mathscr{N}}(X,Y) = \mathrm{Hom}_{\mathscr{C}}(X,Y)/\mathscr{N}(X,Y)$. Si $\mathscr{C}$ est rigide, tressée ou symétrique, il en sera de même pour $\mathscr{C}/\mathscr{N}$.

Définition 3.1. Soit $\mathscr{C}$ une catégorie tensorielle symétrique. Un morphisme $f\colon X \to Y$ dans $\mathscr{C}$ est *négligeable* si pour tout $g\colon Y \to X$, on a $\mathrm{Tr}(fg) = 0$ (ce qui est équivalent à $\mathrm{Tr}(gf) = 0$).

Lemme 3.2. *Les morphismes négligeables forment un idéal tensoriel $\mathscr{N}$ dans $\mathscr{C}$.*

Définition 3.3. On définit la *semi-simplification* d'une catégorie tensorielle symétrique $\mathscr{C}$ comme la catégorie quotient $\overline{\mathscr{C}} = \mathscr{C}/\mathscr{N}$, où $\mathscr{N}$ est l'idéal tensoriel des morphismes négligeables.

Proposition 3.4. *La catégorie $\overline{\mathscr{C}}$ est une catégorie tensorielle symétrique semi-simple, dont les objets simples sont les objets indécomposables de $\mathscr{C}$ dont la dimension est $\neq 0$ dans $\Bbbk$.*

Si $\mathscr{C}$ est déjà semi-simple, alors $\mathscr{N} = 0$ et $\overline{\mathscr{C}} = \mathscr{C}$.

Le lemme suivant permet de comprendre la notion de négligeabilité en termes des objets indécomposables. Il est démontré dans (Etingof et Ostrik, 2022, Lemma 2.2), où (Benson, 1984, §2.18, Exercise 3(ii)) est cité. Je pense qu'on a besoin de supposer que la catégorie $\mathscr{C}$ a des objets de longueur finie (et donc, par la rigidité, il s'ensuit que les Hom sont de dimension finie), pour que $\mathscr{C}$ ait la propriété de Krull–Schmidt. En particulier, pour X indécomposable, l'algèbre d'endomorphismes $\mathrm{End}_{\mathscr{C}}(X)$ est locale : on a $\mathrm{End}(X) = \Bbbk\,\mathrm{id} \oplus \mathfrak{m}$, où $\mathfrak{m}$, le radical, est un idéal nilpotent. Un endomorphisme f de X est donc soit inversible, soit nilpotent, auquel cas sa trace est nulle. En effet, la catégorie $\mathscr{C}$ étant supposée abélienne, on peut utiliser la filtration par les noyaux des puissances successives de f et le fait que la trace ne change pas quand on passe au gradué.

Lemme 3.5. (i) *Si X et Y sont indécomposables dans $\mathscr{C}$, alors un morphisme $f\colon X \to Y$ est non-négligeable si et seulement si $\dim(X) \neq 0$ et f est un isomorphisme.*

(ii) *Si $X = \bigoplus_{i=1}^{n} X_i$ et $Y = \bigoplus_{j=1}^{m}$ sont des décompositions en objets indécomposables, alors un morphisme $f = (f_{ij})\colon X \to Y$, avec $f_{ij}\colon X_i \to Y_j$, est négligeable si et seulement si tous les f_{ij} le sont.*

Démonstration. Soient X et Y indécomposables. Supposons d'abord que $f\colon X \to Y$ ne soit pas un isomorphisme. Alors pour tout $g\colon Y \to X$, le composé $gf \in \mathrm{End}_{\mathscr{C}}(X)$ n'est pas un isomorphisme non plus : si c'en était un, alors f serait un monomorphisme (et donc pas un épimorphisme) et on aurait une décomposition non triviale $Y \simeq \mathrm{im}\, f \oplus \ker g$, contredisant l'indécomposabilité de Y. Donc gf est nilpotent et par conséquent de trace nulle. Donc f est négligeable.

Supposons maintenant que f soit un isomorphisme. Si $\dim(X) \neq 0$, alors $\mathrm{Tr}(f^{-1}f) = \mathrm{Tr}(\mathrm{id}_X) = \dim(X) \neq 0$, donc f n'est pas négligeable. Enfin, si $\dim(X) = 0$, pour $g\colon Y \to X$, on écrit $fg = \lambda\, \mathrm{id}_Y + h$ avec $\lambda \in \Bbbk$ et $h \in \mathrm{End}_{\mathscr{C}}(Y)$ nilpotent. Alors $\mathrm{Tr}(fg) = \lambda \dim Y = \lambda \dim(X) = 0$. Donc f est négligeable. On a montré (i).

Pour (ii), supposons d'abord que les f_{ij} soient négligeables. Alors pour tout $g = (g_{ji})$, on a

$$\mathrm{Tr}(fg) = \sum_{i,j} \mathrm{Tr}(f_{ij}g_{ji}) = 0,$$

donc f est négligeable. Réciproquement, supposons f négligeable ; fixons un couple (k,l) et montrons que f_{kl} est négligeable. Étant donné un morphisme $h\colon Y_l \to X_k$, soit $g = (g_{ji})$ avec $g_{ji} = h$ si $(i,j) = (k,l)$, et 0 sinon. Alors $\mathrm{Tr}(fg) = \mathrm{Tr}(f_{ij}h) = 0$ et ceci pour tout h. Donc f_{ij} est négligeable. □

Démonstration de la proposition 3.4. Si X est indécomposable dans $\mathscr{C}$, alors $\mathrm{End}_{\mathscr{C}}(X) = \Bbbk\, \mathrm{id}_X \oplus \mathfrak{m}$ et le nilradical $\mathfrak{m}$ est un idéal maximal. Comme $\mathfrak{m} \subset \mathscr{N}(X,X) \subset \mathrm{End}_{\mathscr{C}}(X)$, l'idéal $\mathscr{N}(X,X)$ est forcément égal soit à $\mathfrak{m}$, soit à $\mathrm{End}_{\mathscr{C}}(X)$. D'après le lemme, la première possibilité a lieu quand $\dim(X) \neq 0$, la deuxième quand $\dim(X) = 0$. On a donc

$$\mathrm{End}_{\overline{\mathscr{C}}}(\overline{X}) = \begin{cases} \Bbbk & \text{si } \dim(X) \neq 0 \\ 0 & \text{si } \dim(X) = 0. \end{cases}$$

D'autre part, si X et Y sont indécomposables et non isomorphes, alors $\mathrm{Hom}_{\overline{\mathscr{C}}}(X,Y) = 0$ d'après le lemme.

On en conclut que $\overline{\mathscr{C}}$ est semi-simple et que ses objets simples sont les indécomposables de dimension non nulle de $\mathscr{C}$. □

Remarque 3.6. On n'a pas eu besoin du fait que $\mathscr{C}$ est abélienne, sauf dans l'utilisation de la filtration pour montrer qu'un endomorphisme nilpotent a une trace nulle. Par conséquent, Etingof et Ostrik (2022) remarquent que les résultats précédents sont encore valables si $\mathscr{C}$ est seulement $\Bbbk$-linéaire, karoubienne, monoïdale, rigide, dont tous les espaces de morphismes sont de dimension finie, et telle que tous les endomorphismes nilpotents ont une trace nulle.

3.2. Catégorie de Verlinde Ver_p

Dans ce qui suit, $\Bbbk$ est un corps algébriquement clos de caractéristique $p > 0$.

Définition 3.7. La catégorie de Verlinde Ver_p est la semi-simplification de $\mathrm{Rep}_\Bbbk C_p$, où $C_p = \langle g \rangle$ est un groupe cyclique d'ordre p, et $\Bbbk$ est de caractéristique p.

On peut aussi la voir comme la semi-simplification de la catégorie $\mathrm{Tilt}\,\mathbf{SL}_2$ des modules basculants de $\mathbf{SL}_2$, qui n'est pas abélienne mais vérifie les conditions de la remarque 3.6. La définition intrinsèque de module basculant pour un groupe réductif est d'admettre à la fois une filtration dont les sous-quotients sont standard (modules de Weyl) et une autre filtration dont tous les sous-quotients sont costandard (modules induits). Mais on peut aussi caractériser $\mathrm{Tilt}\,\mathbf{SL}_2$ simplement comme la sous-catégorie monoïdale symétrique de $\mathrm{Rep}_\Bbbk \mathbf{SL}_2$ engendrée par $V = \Bbbk^2$ (en prenant des produits tensoriels, des sommes directes, et des facteurs directs). Pour plus d'informations sur les modules basculants, on pourra consulter Jantzen (2003, II.E)

Commençons par remarquer que l'algèbre de groupe de C_p vérifie

$$\Bbbk C_p = \Bbbk[x]/(x^p - 1) = \Bbbk[x]/(x-1)^p \simeq \Bbbk[T]/(T^p).$$

Il est facile de voir que ses représentations indécomposables sont de la forme $J_i = \Bbbk^i$, où le générateur g agit par un bloc de Jordan de taille i, pour $1 \leqslant i \leqslant p$. D'après Green, 1962, on a les relations suivantes :

$$J_2 \otimes J_s = \begin{cases} J_2 & \text{si } s = 1, \\ J_{s-1} \oplus J_{s+1} & \text{si } s = 2, \dots, p-1, \\ J_p \oplus J_p & \text{si } s = p \end{cases}$$

et

$$J_{p-1} \otimes J_s = J_{p-s} \oplus (s-1) J_p.$$

Dans la semi-simplification $\mathrm{Ver}_p := \overline{\mathrm{Rep}_\Bbbk C_p}$, on a donc $p-1$ objets simples,

$$L_i := \overline{J_i}, \quad 1 \leqslant i \leqslant p-1.$$

Définition 3.8. Une catégorie de fusion symétrique est une catégorie prétannakienne avec un nombre fini de modules simples.

La catégorie Ver_p est donc un exemple de catégorie de fusion symétrique. La « règle de fusion » décrivant le produit tensoriel est donnée par une troncation de la règle de Clebsch–Gordan, qui donne les coefficients des produits tensoriels de représentations irréductibles de $\mathbf{SL}_2$ en caractéristique 0 : si on désigne temporairement par L_i, $i \in \mathbb{N}$, le module simple de dimension i pour $\mathbf{SL}_2(\mathbb{C})$, alors on a

$$L_m \otimes L_n = \bigoplus_{i=1}^{\min(m,n)} L_{|m-n|+2i-1}.$$

Dans Ver_p, on a « $L_p = 0$ », et la formule devient :

$$L_m \otimes L_n = \bigoplus_{i=1}^{\min(m,n,p-m,p-n)} L_{|m-n|+2i-1}. \tag{6}$$

Par exemple, pour $p = 5$, on a

$$L_3 \otimes L_3 \simeq L_1 \oplus L_3.$$

Cela montre déjà que Ver_5 ne peut pas avoir de foncteur fibre F à valeurs dans $\mathrm{Vec}_{\Bbbk}$, ou même dans $\mathrm{sVec}_{\Bbbk}$: en effet, dans ce cas la dimension $d := \dim_{\Bbbk} F(L_3)$ vérifierait

$$d^2 = d + 1,$$

mais cette équation n'a pas de solution entière !

On a des cas particuliers pour les petites valeurs de p.

Exemple 3.9. Pour $p = 2$, le seul objet simple est $\mathbf{1} = L_1$ et $\mathrm{Ver}_2 = \mathrm{Vec}_{\Bbbk}$.

Exemple 3.10. Pour $p = 3$, on a deux objets simples, $\mathbf{1} = L_1$ et $\overline{\mathbf{1}} =: L_2$; en fait, $\dim L_2 = 2 = -1 \in \Bbbk$ et $\mathrm{Ver}_3 = \mathrm{sVec}_{\Bbbk}$.

En général, $\overline{\mathbf{1}} := L_{p-1}$ vérifie $\overline{\mathbf{1}} \otimes \overline{\mathbf{1}} = \mathbf{1}$ et $\dim \overline{\mathbf{1}} = -1$; la sous-catégorie tensorielle $\langle \mathbf{1}, \overline{\mathbf{1}} \rangle$ est équivalente à $\mathrm{sVec}_{\Bbbk}$.

Pour la suite, nous aurons besoin du produit tensoriel $\boxtimes$ de catégories. Pour des catégories $\Bbbk$-linéaires $\mathscr{A}$ et $\mathscr{B}$, on note $\mathscr{A} \otimes \mathscr{B}$ la catégorie dont les objets sont les couples (X, Y) avec $X \in \mathscr{A}$ et $Y \in \mathscr{B}$, et dont les morphismes sont donnés par $\mathrm{Hom}((X,Y),(Z,W)) = \mathrm{Hom}_{\mathscr{A}}(X,Z) \otimes \mathrm{Hom}_{\mathscr{B}}(Y,W)$. Puis on note $\mathscr{A} \boxtimes \mathscr{B}$ l'enveloppe karoubienne de $\mathscr{A} \otimes \mathscr{B}$. Si $\mathscr{A}$ et $\mathscr{B}$ sont abéliennes et que $\mathscr{B}$ est semi-simple et schurienne (*i.e.* l'algèbre d'endomorphismes de chaque simple est $\Bbbk$), le produit $\mathscr{A} \boxtimes \mathscr{B}$ est encore une actégorie abélienne ; si $\mathscr{A}$ et $\mathscr{B}$ sont des catégories tensorielles symétriques, avec $\mathscr{B}$ semi-simple et schurienne, alors $\mathscr{A} \boxtimes \mathscr{B}$ est encore une catégorie tensorielle symétrique.

Proposition 3.11. *On suppose $p > 3$.*

(i) *On a une sous-catégorie de fusion* $\mathrm{Ver}_p^+ = \langle L_3 \rangle \subset \mathrm{Ver}_p$ *dont les objets simples sont* L_i *avec* i *impair.*

(ii) *La catégorie de fusion* Ver_p *admet exactement quatre sous-catégories de fusion :* $\mathrm{Vec}_{\Bbbk}$, $\mathrm{sVec}_{\Bbbk}$, Ver_p^+, Ver_p.

(iii) *On a une équivalence de catégories de fusion symétriques* $\mathrm{Ver}_p \simeq \mathrm{Ver}_p^+ \boxtimes \mathrm{sVec}_{\Bbbk}$.

Démonstration. D'après (6), il est clair que les sommes directes d'objets simples L_i avec i impair forment une sous-catégorie tensorielle symétrique et qu'elle est $\otimes$-engendrée par L_3.

Si une sous-catégorie de fusion contient un objet simple L_r avec $r \neq 1, p-1$, alors $\min(r, p-r) \geqslant 2$, donc d'après (6), l'objet simple L_3 apparaît dans la décomposition de L_r^2. Avec (i), cela implique (ii).

La factorisation (iii) résulte d'un lemme général sur les catégories de fusion symétriques, avec deux sous-catégories de fusion ayant seulement $\mathbf{1}$ pour objet simple en commun. □

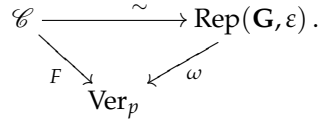
Du point de vue de $\mathbf{SL}_2$, la sous-catégorie Ver_p^+ correspond aux représentations qui se factorisent par $\mathbf{PGL}_2$.

La catégorie Ver_p^+, bien qu'elle soit à croissance modérée, n'est pas super-tannakienne dès que $p \geqslant 5$. De même qu'en caractéristique 0, l'existence de la catégorie non tannakienne $\mathrm{sVec}_{\Bbbk}$ conduit naturellement à la notion de catégorie super-tannakienne, il est naturel de considérer la notion suivante.

Définition 3.12. Une catégorie tensorielle symétrique $\mathscr{C}$ est Ver_p-tannakienne si elle admet un Ver_p-foncteur fibre, c'est-à-dire un foncteur tensoriel symétrique exact à valeurs dans Ver_p.

Par les remarques précédentes, cette classe contient les catégories tannakiennes et super-tannakiennes sur $\Bbbk$.

Si $\mathscr{C}$ admet un Ver_p-foncteur fibre $F\colon \mathscr{C} \to \mathrm{Ver}_p$, on définit $\mathbf{G} := \underline{\mathrm{Aut}}_{\otimes}(F)$, le schéma en groupes affines dans Ver_p des automorphismes de F. On a d'autre part le groupe fondamental $\pi(\mathrm{Ver}_p) := \underline{\mathrm{Aut}}_{\otimes}(\mathrm{id}_{\mathrm{Ver}_p})$. Son algèbre de Hopf (dans Ver_p) est $\mathscr{O}(\pi(\mathrm{Ver}_p)) = \bigoplus_i L_i \otimes L_i^*$. On a un morphisme canonique $\varepsilon\colon \pi(\mathrm{Ver}_p) \to \mathbf{G}$. On considère la catégorie $\mathrm{Rep}(\mathbf{G}, \varepsilon)$ des représentations de $\mathbf{G}$ dans Ver_p telles que $\pi(\mathrm{Ver}_p)$ agisse via ε. Le foncteur F se factorise par le foncteur d'oubli $\omega\colon \mathrm{Rep}(\mathbf{G}, \varepsilon) \to \mathrm{Ver}_p$, en une équivalence :



(Le groupe fondamental de $\mathrm{sVec}_{\Bbbk}$ est C_2, c'est pourquoi dans le cas super-tannakien on devait considérer un élément d'ordre divisant 2 dans $\mathbf{G}$.)

Théorème 3.13 (Ostrik, 2020). *Une catégorie de fusion symétrique sur* $\Bbbk$ (*catégorie tensorielle symétrique semi-simple avec un nombre fini d'objets simples*) *admet un* Ver_p*-foncteur fibre* (*unique à isomorphisme près*).

Un ingrédient important pour la preuve est l'introduction d'un foncteur de Frobenius $\mathrm{Fr}\colon \mathscr{C} \to \mathscr{C}^{(1)} \boxtimes \mathrm{Ver}_p$, où (1) indique un twist de Frobenius de la catégorie : pour avoir un foncteur $\Bbbk$-linéaire, il faut tordre l'action des scalaires par $\mathrm{Frob}\colon \lambda \mapsto \lambda^p$. Ostrik en donne une définition dans le cas des catégories de fusion symétriques ; il indique qu'une généralisation de cette notion devrait permettre de démontrer le théorème dans une plus grande généralité, ce qui sera fait dans (Coulembier, 2020 ; Coulembier, Etingof et Ostrik, 2023b ; Etingof et Ostrik, 2021).

Expliquons succinctement l'idée de ce foncteur de Frobenius. On commence par considérer le foncteur $P_0\colon \mathscr{C} \to \mathscr{C}$, $X \mapsto X^{\otimes p}$, qui n'est bien sûr absolument pas additif, mais qui a le mérite d'admettre une structure monoïdale symétrique évidente. De plus, il admet une action canonique par $\mathfrak{S}_p$ et, en particulier, par son p-sous-groupe de Sylow C_p, qui suffira pour les besoins (pour le moment). Cela permet de relever le foncteur P_0 en un foncteur P_1 à valeurs dans « l'équivariantisation » $\mathrm{Rep}(C_p, \mathscr{C})$, dont les objets sont les objets de $\mathscr{C}$ munis d'une action de C_p. Ensuite on peut passer à la semi-simplification de $\mathrm{Rep}(C_p, \mathscr{C})$, qui n'est autre que $\mathscr{C} \boxtimes \mathrm{Ver}_p$. L'idée est que dans le développement de $(X \oplus Y)^{\otimes p}$, les termes $X^{\otimes p}$ et $Y^{\otimes p}$ sont des points fixes sous C_p, alors que les autres termes sont permutés librement ; dans l'équivariantisation, ils ont un terme J_p en facteur, qui est tué dans la semi-simplification. C'est ce qui permet d'obtenir un foncteur additif et le twist de Frobenius permet d'avoir un foncteur $\Bbbk$-linéaire, noté Fr. (Pas besoin de twist pour la catégorie Ver_p, qui est définie sur le sous-corps premier.)

Une autre idée qui resservira : on dit que $\mathscr{C}$ est de type de Frobenius $\mathscr{A} \subset \mathrm{Ver}_p$ si $\mathrm{Fr}(\mathscr{C}) \subset \mathscr{C} \boxtimes \mathscr{A}$ et que $\mathscr{A}$ est minimale pour cette propriété. Pour $p > 3$, on a donc quatre possibilités pour $\mathscr{A}$ a priori.

Exemple 3.14. On a

$$\mathrm{Fr}(L_s) = \begin{cases} \mathbf{1} \boxtimes L_s & \text{si } s \text{ est impair,} \\ \bar{\mathbf{1}} \boxtimes L_{p-s} & \text{si } s \text{ est pair.} \end{cases} \tag{7}$$

Ainsi, Ver_p est de type de Frobenius Ver_p^+ et $\mathrm{sVec}_\Bbbk$ est de type de Frobenius $\mathrm{Vec}_\Bbbk$. D'autre part, le foncteur de Frobenius est compatible aux foncteurs tensoriels symétriques : si $F\colon \mathscr{C} \to \mathscr{D}$ est un tel foncteur à valeurs dans une autre catégorie de fusion symétrique, on a

$$\mathrm{Fr}\, F(X) = (F^{(1)} \boxtimes \mathrm{id})\, \mathrm{Fr}\, X \in \mathscr{D}^{(1)} \boxtimes \mathrm{Ver}_p\,.$$

Cela implique que si $\mathscr{C}$ est tannakienne ou super-tannakienne, alors $\mathscr{C}$ est de type de Frobenius $\mathrm{Vec}_\Bbbk$.

Un autre ingrédient important dans sa preuve est le fait qu'une catégorie de fusion symétrique non-dégénérée (de dimension globale non nulle) se relève en caractéristique 0, ce qui permet d'appliquer le théorème de Deligne (2002). La preuve de Coulembier, Etingof et Ostrik (2023b) n'utilise pas ce genre d'argument et est plus proche de l'esprit des travaux de Deligne.

4. Catégories Frobenius-exactes

4.1. Foncteurs de Frobenius

On a vu que dans le cas d'une catégorie de fusion symétrique, le twist de Frobenius d'Ostrik (2020) est un outil crucial pour construire un foncteur fibre vers Ver_p. Cette construction a été généralisée dans Etingof et Ostrik (2021) et Coulembier (2020) à n'importe quelle catégorie prétannakienne en caractéristique p. Coulembier, Etingof et Ostrik (2023b) construisent également une version enrichie, faisant intervenir tout le groupe symétrique $\mathfrak{S}_p$ et pas seulement le groupe cyclique C_p. La construction est en fait similaire dans l'esprit à celle d'Ostrik (2020).

Dans §2.5, ils rappellent qu'on peut encoder l'information d'une catégorie tensorielle symétrique $\mathscr{C}$ et d'un foncteur $\Bbbk$-linéaire fidèle et exact $F\colon \mathscr{C} \to \mathrm{Vec}_\Bbbk$ à l'aide d'une structure algébrique sur la coalgèbre correspondante, à savoir une structure de co-pseudo algèbre de Hopf cotriangulaire H (peu importe la définition précise pour ce qui nous concerne). Étant donné un groupe Γ, on peut interpréter H comme une co-pseudo algèbre de Hopf triangulaire dans la catégorie tensorielle symétrique $\mathrm{Ind}\,\mathrm{Rep}_\Bbbk \Gamma$, via l'inclusion $\mathrm{Vec}_\Bbbk \subset \mathrm{Rep}_\Bbbk \Gamma$. Alors on peut voir les représentations de Γ dans $\mathscr{C}$ (objets munis d'une action) comme des H-comodules dans $\mathrm{Rep}_\Bbbk \Gamma$.

On prend pour Γ un sous-groupe de $\mathfrak{S}_p$: ce sera soit C_p, soit $\mathfrak{S}_p$ tout entier. On construit un foncteur monoïdal symétrique $\mathrm{Fr}_\Gamma\colon \mathscr{C} \to \mathscr{C} \boxtimes \overline{\mathrm{Rep}\,\Gamma}$ via

$$\mathscr{C} \longrightarrow \mathrm{Rep}(\Gamma, \mathscr{C}) \simeq H\text{-}\mathrm{Comod}(\mathrm{Rep}_\Bbbk \Gamma) \longrightarrow H\text{-}\mathrm{Comod}(\overline{\mathrm{Rep}_\Bbbk \Gamma}) = \mathscr{C} \boxtimes \overline{\mathrm{Rep}_\Bbbk \Gamma},$$

où la première flèche est $X \mapsto X^{\otimes p}$ et la deuxième est induite par le foncteur de semi-simplification. On notera $\mathrm{Fr} := \mathrm{Fr}_{C_p}$ le foncteur de Frobenius (appelé twist de Frobenius externe dans Coulembier (2020)) et $\mathrm{Fr}^{\mathrm{en}} := \mathrm{Fr}_{\mathfrak{S}_p}$ le foncteur de Frobenius « enrichi ». On est amené à considérer la catégorie

$$\mathrm{Ver}_p^{\mathrm{en}} = \overline{\mathrm{Rep}\,\mathfrak{S}_p} = \mathrm{Ver}_p^+ \boxtimes \mathrm{Rep}(C_{2p-2}, z),$$

où z est l'élément d'ordre 2 dans C_{2p-2} (Etingof et Ostrik, 2022). Il résulte du fait que $|\mathfrak{S}_p : C_p|$ est inversible que le foncteur de restriction $\mathrm{Rep}_\Bbbk \mathfrak{S}_p \to \mathrm{Rep}_\Bbbk C_p$ envoie négligeable sur négligeable, donc induit un foncteur $R\colon \overline{\mathrm{Rep}_\Bbbk \mathfrak{S}_p} \to \overline{\mathrm{Rep}_\Bbbk C_p}$. Les deux versions du foncteur de Frobenius sont liées par :

$$\mathrm{Fr} = (\mathrm{id} \boxtimes R) \circ \mathrm{Fr}^{\mathrm{en}}. \tag{8}$$

On a aussi la définition plus naïve suivante d'un foncteur de Frobenius. On note DSym^n et Sym^n les foncteurs qui à X associent le plus grand sous-objet et le plus grand quotient $\mathfrak{S}_n$-invariant, respectivement (puissances divisées, puissances symétriques).

Définition 4.1. Pour X un objet d'une catégorie tensorielle symétrique $\mathscr{C}$ sur $\Bbbk$ de caractéristique $p > 0$, on définit le foncteur Frob^j-linéaire (*i.e.* l'action des scalaires est tordue par la j-ième puissance de l'endomorphisme de Frobenius) :

$$\begin{aligned}\mathrm{Fr}_+^{(j)} : \mathscr{C} &\longrightarrow \mathscr{C}\\ X &\longmapsto \mathrm{Triv}_{\mathfrak{S}_{p^j}}(X^{\otimes p^j}) := \mathrm{im}(\mathrm{DSym}^{p^j} X \hookrightarrow X^{\otimes p^j} \twoheadrightarrow \mathrm{Sym}^{p^j} X).\end{aligned}$$

On note $\mathrm{Fr}_+ := \mathrm{Fr}_+^{(1)}$.

Le foncteur Fr_+ est le facteur direct de $\mathrm{Fr}^{\mathrm{en}}$ correspondant à l'action triviale de $\mathfrak{S}_p$:

$$\mathrm{Fr}_+ = (\mathrm{id} \boxtimes \mathrm{Hom}(\mathbf{1}, -)) \circ \mathrm{Fr}^{\mathrm{en}}.$$

On en déduit que Fr_+ est canoniquement lax et oplax-monoïdal (même définition que foncteur monoïdal, sauf qu'on demande seulement d'avoir des morphismes, pas forcément des isomorphismes, entre $\mathrm{Fr}_+(X) \otimes \mathrm{Fr}_+(Y)$ et $\mathrm{Fr}_+(X \otimes Y)$, compatibles avec les contraintes).

Un résultat important de Coulembier, Etingof et Ostrik (2023b) est que les différentes notions d'exactitude de Frobenius qu'on pourrait envisager sont en fait les mêmes.

Proposition 4.2 (Coulembier, Etingof et Ostrik, 2023b, Prop. 3.4)**.** *Pour une catégorie prétannakienne $\mathscr{C}$, les conditions suivantes sont équivalentes :*

(i) *Le foncteur* Fr *est exact ;*

(ii) *Le foncteur* $\mathrm{Fr}^{\mathrm{en}}$ *est exact ;*

(iii) *Le foncteur* Fr_+ *est exact ;*

(iv) *Pour tout morphisme non nul $u : \mathbf{1} \to X$, on a $\mathrm{Fr}_+(u) \neq 0$.*

(v) *Il existe un foncteur monoïdal symétrique exact $\mathscr{C} \to \mathscr{D}$ vers une catégorie monoïdale symétrique abélienne $\Bbbk$-linéaire (avec $\otimes$ bilinéaire) qui envoie toute suite exacte courte dans $\mathscr{C}$ sur une suite exacte courte scindée dans $\mathscr{D}$.*

Définition 4.3. Une catégorie prétannakienne $\mathscr{C}$ satisfaisant les conditions équivalentes de la proposition est dite **Frobenius-exacte**.

Toute catégorie tensorielle symétrique admettant un foncteur tensoriel symétrique vers une catégorie tensorielle symétrique semi-simple (comme Ver_p et ses sous-catégories tensorielles symétriques) est nécessairement Frobenius-exacte.

4.2. Caractérisation interne des catégories Ver_p-tannakiennes

Théorème 4.4 (Coulembier, Etingof et Ostrik, 2023b, Theorem 1.2). *Pour une catégorie prétannakienne $\mathscr{C}$, les conditions suivantes sont équivalentes :*

(i) *$\mathscr{C}$ est tannakienne ;*

(ii) *$\mathscr{C}$ est Frobenius-exacte, à croissance modérée, et l'endofoncteur Fr_+, a priori seulement lax-monoïdal, est en fait monoïdal.*

C'est une étape importante dans la preuve du résultat principal, qui caractérise les catégories Ver_p-tannakiennes.

Théorème 4.5 (Coulembier, Etingof et Ostrik, 2023b, Theorem 1.1). *Pour une catégorie prétannakienne $\mathscr{C}$, les conditions suivantes sont équivalentes :*

(i) *$\mathscr{C}$ est Ver_p-tannakienne ;*

(ii) *$\mathscr{C}$ est Frobenius-exacte et à croissance modérée ;*

(iii) *$\mathscr{C}$ est Frobenius-exacte et, pour tout $X \in \mathscr{C}$, il existe un entier $n \in \mathbb{N}$ tel que $A^n X = 0$.*

De plus, un Ver_p-foncteur fibre, lorsqu'il existe, est unique à isomorphisme près.

4.3. Résumé des caractérisations internes

Pour $\mathscr{C}$ une catégorie tensorielle symétrique sur $\Bbbk = \overline{\Bbbk}$ de caractéristique 0, on a

$$\begin{aligned} \mathscr{C} \text{ tannakienne} &\iff (\forall X \in \mathscr{C},\ \dim(X) \in \mathbb{N}) \iff (\forall X \in \mathscr{C},\ \mathrm{ad}(X) < \infty), \\ \mathscr{C} \text{ super-tannakienne} &\iff (\forall X \in \mathscr{C},\ \mathrm{gd}(X) < \infty) \iff (\forall X \in \mathscr{C},\ \mathrm{sd}(X) < \infty). \end{aligned}$$

Pour $\mathscr{C}$ une catégorie tensorielle symétrique Frobenius-exacte sur $\Bbbk = \overline{\Bbbk}$ de caractéristique $p > 0$, on a

$$\mathscr{C} \text{ } \mathrm{Ver}_p\text{-tannakienne} \iff (\forall X \in \mathscr{C},\ \mathrm{gd}(X) < \infty) \iff (\forall X \in \mathscr{C},\ \mathrm{ad}(X) < \infty),$$

$$\begin{aligned} \mathscr{C} \text{ tannakienne} &\iff \mathrm{gd} = \mathrm{ad} < \infty \iff \mathrm{ad} < \infty \text{ multiplicative} \iff \mathrm{Fr}_+ \text{ monoïdal} \\ &\iff \mathrm{ad} < \infty \text{ et } (\forall X \in \mathscr{C}, \forall j, n \in \mathbb{N}, A^n \mathrm{Fr}_+^{(j)} X = 0 \Longrightarrow A^n X = 0). \end{aligned}$$

5. Une nouvelle famille de catégories tensorielles symétriques

Maintenant, on peut se poser la question : y a-t-il des catégories tensorielles symétriques à croissance modérée qui ne sont pas Ver_p-tannakiennes ? Au vu du théorème 4.5, si oui, elles ne peuvent être Frobenius-exactes.

La catégorie Ver_p peut être obtenue comme la semi-simplification $\overline{\mathrm{Tilt}\,\mathbf{SL}_2(\Bbbk)} = \mathrm{Tilt}\,\mathbf{SL}_2(\Bbbk)/\mathscr{N}$, où $\mathscr{N}$ est l'idéal des morphismes négligeables. S'il y avait d'autres idéaux dans cette catégorie, cela pourrait être une source de nouvelles catégories tensorielles symétriques.

Les modules basculants indécomposables sont classifiés par le plus haut poids : on les notera T_m, avec $m \in \mathbb{N}$. Parmi ceux-ci, les T_{p^n-1} sont très particuliers : ce sont les modules simples, de même plus haut poids, donc de dimension p^n. Il se trouve que $\mathscr{N}$ est l'idéal engendré par T_{p-1} (au sens où c'est le plus petit idéal tensoriel de $\mathscr{C}$ contenant $\mathrm{id}_{T_{p-1}}$).

Soit $\mathscr{I}_n$ l'idéal engendré par T_{p^n-1}. Il contient tous les T_j avec $j \geqslant p^n - 1$. On a $\mathscr{N} = \mathscr{I}_1 \supset \mathscr{I}_2 \supset \cdots$.

Théorème 5.1 (Coulembier, 2021). *Les idéaux tensoriels de* $\mathrm{Tilt}\,\mathbf{SL}_2(\Bbbk)$ *sont exactement les* $\mathscr{I}_n$.

La catégorie $\widetilde{\mathrm{Ver}}_{p^n} := \mathrm{Tilt}\,\mathbf{SL}_2(\Bbbk)/\mathscr{I}_n$ n'est pas abélienne pour $n \geqslant 2$. Mais on peut en construire une enveloppe abélienne Ver_{p^n} (Benson, Etingof et Ostrik, 2023 ; Coulembier, 2021), de telle sorte que les projectifs indécomposables soient les T_j pour $p^{n-1} - 1 \leqslant j \leqslant p^n - 2$.

Pour définir $M \otimes N$ pour deux objets dans Ver_{p^n}, on peut prendre des résolutions projectives $P_\bullet$ et $Q_\bullet$; on montre que $P_\bullet \otimes Q_\bullet$ est exact en dehors du degré 0, et on pose $M \otimes N := H^0(P_\bullet \otimes Q_\bullet)$.

Comme pour Ver_p, on a une factorisation :

$$\mathrm{Ver}_{p^n} = \mathrm{sVec}_\Bbbk \boxtimes \mathrm{Ver}_{p^n}^+ .$$

On a des inclusions $\mathrm{Ver}_{p^n} \subset \mathrm{Ver}_{p^{n+1}}$. On pose $\mathrm{Ver}_{p^\infty} := \bigcup_n \mathrm{Ver}_{p^n}$.

Une autre propriété remarquable est qu'on a un relèvement en caractéristique 0 de ces catégories tensorielles, toutefois il est tressé et pas symétrique : il s'agit de la catégorie définie de façon analogue mais pour le groupe quantique pour $\mathbf{SL}_2$ en une racine p^n-ième de l'unité. Lors de la réduction modulaire en caractéristique p, cette racine devient triviale, on obtient des représentations du groupe $\mathbf{SL}_2$, et le tressage devient symétrique.

Conjecture 5.2 (Benson, Etingof et Ostrik, 2023, Conjecture 1.4). *Toute catégorie tensorielle symétrique à croissance modérée sur* $\Bbbk$ *algébriquement clos de caractéristique* p *admet un foncteur fibre vers* Ver_{p^∞}.

Ce serait là un parfait analogue en caractéristique $p > 0$ du résultat de Deligne (2002) !

Définition 5.3. On dit qu'une catégorie tensorielle symétrique $\mathscr{C}$ est *incompressible* si tout foncteur tensoriel symétrique $F\colon \mathscr{C} \to \mathscr{D}$ est « injectif », *i.e.* pleinement fidèle.

En caractéristique 0, les seules catégories tensorielles symétriques incompressibles connues sont $\mathrm{Vec}_\Bbbk$ et $\mathrm{sVec}_\Bbbk$, on sait que ce sont les seules à croissance modérée et on conjecture qu'il n'y en a pas d'autre. (On sait que les catégories de Deligne (2007) n'admettent pas de foncteur tensoriel symétrique vers une catégorie incompressible.)

En caractéristique p, en plus de $\mathrm{Vec}_\Bbbk$ et $\mathrm{sVec}_\Bbbk$, toutes les catégories Ver_{p^n} et $\mathrm{Ver}^+_{p^n}$ sont incompressibles ($n \in \mathbb{N} \cup \{\infty\}$) ; Benson, Etingof et Ostrik (2023) conjecturent qu'il n'y en a pas d'autre. Ils conjecturent aussi que toute catégorie tensorielle symétrique à croissance modérée sur $\Bbbk$ admet un foncteur fibre vers Ver_{p^∞}.

Pour $p > 2$, voici un diagramme montrant les inclusions entre les catégories tensorielles symétriques incompressibles que nous avons rencontrées :

$$\begin{array}{ccccccccccc}
\mathrm{sVec}_\Bbbk & \subset & \mathrm{Ver}_p & \subset & \mathrm{Ver}_{p^2} & \subset & \mathrm{Ver}_{p^3} & \subset & \cdots & \subset & \mathrm{Ver}_{p^\infty} \\
\cup & & \cup & & \cup & & \cup & & & & \cup \\
\mathrm{Vec}_\Bbbk & \subset & \mathrm{Ver}^+_p & \subset & \mathrm{Ver}^+_{p^2} & \subset & \mathrm{Ver}^+_{p^3} & \subset & \cdots & \subset & \mathrm{Ver}^+_{p^\infty}
\end{array}$$

(Dans le cas particulier $p = 3$, on a $\mathrm{sVec}_\Bbbk = \mathrm{Ver}_3$.)

Pour $p = 2$, on a $\mathrm{Vec}_\Bbbk = \mathrm{sVec}_\Bbbk = \mathrm{Ver}_2$, de plus le diagramme serait plutôt une chaîne.

6. Applications

6.1. Extension de la notion de dimension de Frobenius–Perron

On rappelle que pour une catégorie de fusion $\mathscr{C}$, on a un unique caractère de l'anneau de Grothendieck de $\mathscr{C}$, appelé dimension de Frobenius–Perron et noté $\mathrm{FPdim}\colon [\mathscr{C}] \to \mathbb{R}$, prenant des valeurs positives sur les classes des objets simples (Etingof, Gelaki et al., 2015, §4.5). Il associe à chaque classe d'objet simple $[S]$ le rayon spectral de l'endomorphisme « multiplication par $[S]$ » sur $[\mathscr{C}]$. C'est un outil fondamental pour l'étude des catégories de fusion. Une première application des résultats de Coulembier, Etingof et Ostrik (2023b) est qu'il est possible d'étendre la notion de Frobenius–Perron de manière consistante à une classe bien plus large de

catégories tensorielles symétriques, à savoir les catégories prétannakiennes à croissance modérée, Frobenius-exactes si on est en caractéristique $p > 0$. Ce sera en fait la dimension de croissance gd. De plus, les foncteurs tensoriels symétriques entre telles catégories préservent cette notion de dimension. Le fait que ces catégories ont un foncteur fibre sur Ver_p (où par convention $\mathrm{Ver}_0 = \mathrm{sVec}_{\Bbbk}$ si $\Bbbk$ est de caractéristique 0) donne des informations très concrètes, car on peut utiliser tout ce qu'on sait sur $[\mathrm{Ver}_p]$. Notamment, cela limite l'ensemble des valeurs que peut prendre cet invariant. Soit $q = \exp(i\pi/p)$, posons

$$\mathscr{O}_p := \begin{cases} \mathbb{Z}[q+q^{-1}] & \text{si } p > 0, \\ \mathbb{Z} & \text{si } p = 0. \end{cases}$$

(On a donc $\mathscr{O}_p = \mathbb{Z}$ si et seulement si $p \leqslant 3$.) Pour $p > 2$, une $\mathbb{Z}$-base de $\mathscr{O}_p$ est donnée par :

$$[m]_q := \frac{q^m - q^{-m}}{q - q^{-1}}, \text{ avec } 1 \leqslant m \leqslant \frac{p-1}{2}.$$

Théorème 6.1 (Coulembier, Etingof et Ostrik, 2023b, Theorem 8.1). *Soit $\mathscr{C}$ une catégorie prétannakienne à croissance modérée, Frobenius-exacte si $\Bbbk$ est de caractéristique $p > 0$.*

(1) *On a un morphisme d'anneaux*

$$\begin{array}{ccc} [\mathscr{C}] & \longrightarrow & \mathscr{O}_p \\ [X] & \longmapsto & \mathrm{gd}(X) \end{array}$$

et $\mathrm{gd}(X) \geqslant [\dim(X)]_q$ pour $\dim(X) \in \{0, 1, \ldots, p-1\}$. De plus, si $p > 2$, alors $\mathrm{gd}(X)$ est une combinaison linéaire à coefficients dans $\mathbb{N}$ des $[m]_q$ avec $1 \leqslant m \leqslant \frac{p-1}{2}$.

(2) *Si $\mathscr{C}$ est une catégorie de fusion, alors la dimension de croissance n'est autre que la dimension de Frobenius–Perron, i.e. $\mathrm{gd}(X) = \mathrm{FPdim}\, X$.*

(3) *Pour le foncteur fibre $F\colon \mathscr{C} \to \mathrm{Ver}_p$, qui existe et est unique, on a $\mathrm{gd}(X) = \mathrm{FPdim}\, F(X)$.*

(4) *Si on a un foncteur tensoriel symétrique $H\colon \mathscr{C} \to \mathscr{C}'$, où $\mathscr{C}'$ vérifie les mêmes conditions que $\mathscr{C}$, alors $\mathrm{gd}(X) = \mathrm{gd}(H(X))$.*

(5) *$\mathscr{C}$ est tannakienne si et seulement si $\mathrm{gd}(X) = \mathrm{ad}(X)$ pour tout $X \in \mathscr{C}$, si et seulement si ad est multiplicative, i.e. $\mathrm{ad}(X \otimes Y) = \mathrm{ad}(X)\,\mathrm{ad}(Y)$ pour tous X et Y dans $\mathscr{C}$.*

(6) *$\mathscr{C}$ est super-tannakienne si et seulement si $\mathrm{gd}(X) \in \mathbb{N}$ pour tout $X \in \mathscr{C}$.*

Donnons juste la preuve du lemme suivant, dont l'hypothèse permet de l'appliquer non seulement immédiatement à Ver_p, mais aussi par exemple à Ver_{p^∞}.

Lemme 6.2 (Coulembier, Etingof et Ostrik, 2023b, Lemma 8.3). *Soit $\mathscr{D}$ une catégorie prétannakienne qui soit réunion de catégories prétannakiennes ayant un nombre fini d'objets simples. Alors pour tout objet X dans $\mathscr{D}$, on a $\mathrm{FPdim}\, X = \mathrm{gd}(X)$.*

Démonstration. On peut travailler dans la sous-catégorie tensorielle symétrique $\mathscr{C} := \langle X \rangle$ engendrée par X, qui par hypothèse a un nombre fini d'objets simples. Soit d la dimension de Frobenius–Perron maximale parmi les objets simples de $\mathscr{C}$. Comme FPdim est un morphismes d'anneaux $[\mathscr{C}] \to \mathbb{R}$ et qu'un objet simple a une dimension de Frobenius–Perron au moins égale à 1, on a

$$\ell(X^{\otimes n}) \leqslant \mathrm{FPdim}(X^{\otimes n}) = (\mathrm{FPdim}\, X)^n \leqslant d\ell(X^{\otimes n}).$$

La longueur de $X^{\otimes n}$ est donc comprise entre $(\mathrm{FPdim}\, X)^n/d$ et $(\mathrm{FPdim}\, X)^n$; on conclut en prenant la racine n-ième et en faisant tendre n vers $+\infty$. □

6.2. Taux de croissance pour les représentations modulaires des groupes finis

Coulembier, Etingof et Ostrik (2023b, §8.3) obtiennent des corollaires sur les représentations modulaires de groupes finis au sens classique.

L'étude des représentations modulaires indécomposables des groupes finis est très ardue : par exemple, un p-groupe ne peut avoir un nombre fini de $\Bbbk$-représentations indécomposables que s'il est cyclique, et il n'est de type de représentation modéré que si $|G : G'| \leqslant 4$; tous les autres cas sont sauvages (Bondarenko et Drozd, 1977). (On renvoie à Benson (1991, §4.4) pour une discussion du type de représentation d'une algèbre, qui peut être fini, modéré ou sauvage.)

Soit $\mathbf{G}$ un schéma en groupes affines sur $\Bbbk$ supposé de caractéristique $p > 0$ (par exemple un groupe fini) et soit V dans $\mathrm{Rep}\,\mathbf{G}$. On s'intéresse aux facteurs directs indécomposables apparaissant dans les puissances tensorielles $V^{\otimes n}$. On peut appliquer directement les résultats précédents pour obtenir des informations sur le nombre $\delta_n(V)$ de facteurs directs indécomposables de dimension première à p dans $V^{\otimes n}$, comptés avec multiplicités. En effet, on interprète ce nombre comme $d_n(X)$, où $X = \overline{V}$ est la semi-simplification de V dans $\overline{\mathrm{Rep}\,\mathbf{G}}$. La limite $\delta(V) := \lim_{n\to\infty} \delta_n(V)^{1/n}$ existe et est égale à $\mathrm{gd}(X)$. D'après le théorème 6.1(1), δ induit un morphisme d'anneaux $[\mathrm{Rep}\,\mathbf{G}] \to \mathbb{R}$ et

$$\delta(V) = \mathrm{gd}(X) \geqslant [\dim(\overline{V})]_q = [\dim(V)]_q = [\dim_{\Bbbk}(V)]_q.$$

D'après le théorème 4.5, on a un Ver_p-foncteur fibre

$$F \colon \mathscr{C} = \langle X \rangle \to \mathrm{Ver}_p\,.$$

On écrit

$$F(X) = \bigoplus_{k=1}^{p-1} m_k L_k,$$

alors

$$\delta(V) = \mathrm{gd}(X) = \mathrm{FPdim}(F(X)) = \sum_{k=1}^{p-1} m_k [k]_q.$$

Supposons $p > 2$. Avec cette seule équation, $\delta(V)$ ne détermine pas les m_k, mais seulement les $m_k + m_{p-k}$. En effet, rappelons que $\mathscr{O}_p$ est de rang $\frac{p-1}{2}$ et que $[k]_q = [p-k]_q$. Cependant en considérant la deuxième opération d'Adams ψ^2, en lien avec la notion de super-dimension de Frobenius–Perron (Etingof, Ostrik et Venkatesh, 2017), on peut aussi déterminer les $m_k - m_{p-k}$ et donc au final tous les m_k. Ainsi, des propriétés asymptotiques (taux de croissance) permettent de récupérer l'image dans Ver_p par le foncteur fibre et vice versa.

On renvoie à (Coulembier, Etingof et Ostrik, 2023b), ainsi qu'à (Benson, 2020a,b ; Benson et Symonds, 2020 ; Coulembier, Etingof et Ostrik, 2023a) pour plus de résultats et de conjectures.

Remerciements. — Je tiens à remercier les collaborateurs de Nicolas Bourbaki pour leur relecture et leur bienveillance, ainsi qu'Ofer Gabber pour ses corrections.

Références

André, Y. (2004). *Une introduction aux motifs (motifs purs, motifs mixtes, périodes)*. T. 17. Panoramas et Synthèses. Société Mathématique de France, Paris, p. xii+261.

Barrett, J. W. et Westbury, B. W. (1999). « Spherical categories », *Adv. Math.* **143** (2), p. 357-375.

Benson, D. (1984). *Modular representation theory : new trends and methods*. T. 1081. Lecture Notes in Mathematics. Springer-Verlag, Berlin, p. xi+231.

——— (1991). *Representations and cohomology. I*. T. 30. Cambridge Studies in Advanced Mathematics. Basic representation theory of finite groups and associative algebras. Cambridge University Press, Cambridge, p. xii+224.

——— (2020a). « Modular representation theory and commutative Banach algebras », *Mem. Amer. Math. Soc.* Sous presse.

——— (2020b). « Some conjectures and their consequences for tensor products of modules over a finite p-group », *J. Algebra* **558**, p. 24-42.

Benson, D. et Etingof, P. (2019). « Symmetric tensor categories in characteristic 2 », *Adv. Math.* **351**, p. 967-999.

Benson, D., Etingof, P. et Ostrik, V. (2023). « New incompressible symmetric tensor categories in positive characteristic », *Duke Math. J.* **172** (1), p. 105-200.

Benson, D. et Symonds, P. (2020). « The non-projective part of the tensor powers of a module », *J. Lond. Math. Soc.* (2) **101** (2), p. 828-856.

Bondarenko, V. M. et Drozd, Y. A. (1977). « The representation type of finite groups », *Zap. Naučn. Sem. Leningrad. Otdel. Mat. Inst. Steklov.* (*LOMI*) **71**. Modules and representations, p. 24-41, 282.

Coulembier, K. (2020). « Tannakian categories in positive characteristic », *Duke Math. J.* **169** (16), p. 3167-3219.

——— (2021). « Monoidal abelian envelopes », *Compos. Math.* **157** (7), p. 1584-1609.

Coulembier, K., Etingof, P. et Ostrik, V. (2023a). « Asymptotic properties of tensor powers in symmetric tensor categories ».

——— (2023b). « On Frobenius exact symmetric tensor categories », *Ann. of Math.* (2) **197** (3). Avec un appendice par Alexander Kleshchev, p. 1235-1279.

Deligne, P. (1990). « Catégories tannakiennes », in : *The Grothendieck Festschrift, Vol. II*. T. 87. Progr. Math. Birkhäuser Boston, Boston, MA, p. 111-195.

——— (2002). « Catégories tensorielles », *Mosc. Math. J.* **2** (2). Dedicated to Yuri I. Manin on the occasion of his 65th birthday, p. 227-248.

——— (2007). « La catégorie des représentations du groupe symétrique S_t, lorsque t n'est pas un entier naturel », in : *Algebraic groups and homogeneous spaces*. T. 19. Tata Inst. Fund. Res. Stud. Math. Tata Inst. Fund. Res., Mumbai, p. 209-273.

——— (2011). « Letter to A. Vasiu, Nov. 30, 2011 ».

Deligne, P. et Milne, J. S. (1982). « Tannakian categories », in : *Hodge cycles, motives, and Shimura varieties*. Sous la dir. de P. Deligne et al. T. 900. Lecture Notes in Mathematics. Springer-Verlag, Berlin-New York, p. 101-128.

Etingof, P., Gelaki, S. et al. (2015). *Tensor categories*. T. 205. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, p. xvi+343.

Etingof, P. et Ostrik, V. (2021). « On the Frobenius functor for symmetric tensor categories in positive characteristic », *J. Reine Angew. Math.* **773**, p. 165-198.

——— (2022). « On semisimplification of tensor categories », in : *Representation theory and algebraic geometry—a conference celebrating the birthdays of Sasha Beilinson and Victor Ginzburg*. Trends Math. Birkhäuser/Springer, Cham, p. 3-35.

Etingof, P., Ostrik, V. et Venkatesh, S. (2017). « Computations in symmetric fusion categories in characteristic p », *Int. Math. Res. Not. IMRN* (2), p. 468-489.

Gelfand, S. et Kazhdan, D. (1992). « Examples of tensor categories », *Invent. Math.* **109** (3), p. 595-617.

Georgiev, G. et Mathieu, O. (1994). « Fusion rings for modular representations of Chevalley groups », in : *Mathematical aspects of conformal and topological field theories and quantum groups* (*South Hadley, MA, 1992*). T. 175. Contemp. Math. Amer. Math. Soc., Providence, RI, p. 89-100.

Green, J. A. (1962). « The modular representation algebra of a finite group », *Illinois J. Math.* **6**, p. 607-619.

Grothendieck, A. (2021). *Récoltes et semailles I, II. Réflexions et témoignage sur un passé de mathématicien*. T. 437/438. Collect. Tel. Paris : Gallimard.

Harman, N. et Snowden, A. (2022). « Oligomorphic groups and tensor categories ».
James, G. et Kerber, A. (1981). *The representation theory of the symmetric group*. T. 16. Encyclopedia of Mathematics and its Applications. With a foreword by P. M. Cohn, With an introduction by Gilbert de B. Robinson. Addison-Wesley Publishing Co., Reading, Mass., p. xxviii+510.
Jannsen, U. (1992). « Motives, numerical equivalence, and semi-simplicity », *Invent. Math.* **107** (3), p. 447-452.
Jantzen, J. C. (2003). *Representations of algebraic groups*. Second. T. 107. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, p. xiv+576.
Knop, F. (2007). « Tensor envelopes of regular categories », *Adv. Math.* **214** (2), p. 571-617.
MacLane, S. (1971). *Categories for the working mathematician*. Graduate Texts in Mathematics, Vol. 5. Springer-Verlag, New York-Berlin, p. ix+262.
Ostrik, V. (2020). « On symmetric fusion categories in positive characteristic », *Selecta Math. (N.S.)* **26** (3), Paper No. 36, 19.
Saavedra Rivano, N. (1972). *Catégories Tannakiennes*. Lecture Notes in Mathematics, Vol. 265. Springer-Verlag, Berlin-New York, p. ii+418.

Daniel Juteau

LAMFA
Université de Picardie Jules Verne
CNRS
33 rue Saint-Leu, 80000 Amiens, France

E-mail : daniel.juteau@u-picardie.fr

Séminaire BOURBAKI Juin 2023
74e année, 2022–2023, no 1210, p. 481 à 499
doi: 10.24033/ast.1219

ROTATION INVARIANCE OF CRITICAL PLANAR PERCOLATION
[*after* Hugo Duminil-Copin, Karol Kajetan Kozlowski, Dmitry Krachun, Ioan Manolescu and Mendes Oulamara]

by **Vincent Tassion**

Introduction

Consider critical independent percolation on the square lattice $\mathbb{Z}^2$, viewed as a graph: For each edge, flip a coin, the edge is kept with probability $p = 1/2$, it is deleted otherwise. We thus obtain a random subgraph of $\mathbb{Z}^2$. The distribution of this random graph is invariant under rotation of angle $\pi/2$, as it inherits the symmetries of the lattice. But if we consider the large connected components, new symmetries emerge: Duminil-Copin et al. (2020) have shown that the distribution of these connected components is asymptotically invariant under all rotations. This result represents major progress towards understanding critical phenomena in planar statistical mechanics. The main conjecture in the field is that the distribution of large connected components is in fact invariant by conformal transformations, and it satisfies a principle of universality: this distribution does not depend on the underlying lattice. In this article, we give some general background on Bernoulli percolation, we state the new rotation invariance result and discuss some key aspects of it: what role does the parameter $1/2$ play? What heuristic reasons justify the emergence of these symmetries? What are the main ideas behind rotational invariance? We mainly focus on one important ingredient of the proof: the star-triangle transformation. Originated from the study of electrical networks, it allows the authors to relate percolation on the square lattice to other auxiliary graphs, and "import" extra symmetries satisfied by these graphs (namely symmetry under reflections).

Acknowledgment. I warmly thank Ioan Manolescu, who explained to me the main ideas behind the proof of rotation invariance of critical percolation, and was a precious help in the preparation of this article. As always, I could count on Sébastien Martineau who made a detailed list of valuable comments and suggestions on a previous version of this article, leading to substantial improvements. I am also grateful to Christoforos Panagiotis and the two anonymous readers for useful comments. This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 851565).

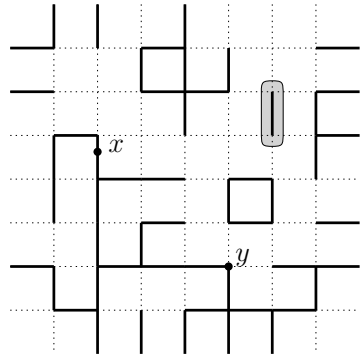
1. Phase transition of Bernoulli percolation

Bernoulli percolation was introduced in 1957 by Broadbent and Hammersley (1957) in order to understand the propagation of a fluid in a porous medium, modeled as follows. Consider the square lattice $\mathbb{Z}^2$, which we see as a planar graph embedded in the complex plane: its vertex set is $V = \{u + iv : u, v \in \mathbb{Z}\}$, and the edge set E is given by all linear segments $[u, v]$ with $|u - v| = 1$. Fix a parameter $p \in [0, 1]$, which represents the porosity of the material we want to model.

For each edge $e \in E$ toss a biased coin, and define

$$\omega_e = \begin{cases} 0 & \text{with probability } 1 - p, \\ 1 & \text{with probability } p, \end{cases}$$

independently of the other edges. We say that the edge e is open if $\omega_e = 1$ (solid edges in the figure below) and closed if $\omega_e = 0$ (dotted edges).



The terminology open/closed comes from the interpretation of ω as a porous material: the fluid can only travel through open edges, and percolation aims at describing the different paths that the fluid can follow. To this end, it is convenient to identify ω with the union of all the open edges. This way, we see ω as a closed subset of $\mathbb{C}$ and define its corresponding topological properties. We call open path a continuous path with support in ω. For example, in the picture above, there exists an open path from x to y. We emphasize that we do not impose that the path starts and ends at vertices of $\mathbb{Z}^2$. We call cluster a connected component of ω. For example, above, we surrounded a cluster made of a single edge. Despite this elementary mathematical description, Bernoulli percolation offers a natural probabilistic framework to develop and understand the theory of phase transitions, a key notion in statistical mechanics.

A natural question for Bernoulli percolation is whether there exists an infinite cluster in ω. The answer depends on the underlying parameter: if $p = 0$ we have

$\omega = \emptyset$ and there is no infinite cluster. For $p = 1$ all the edges are open, and there is a unique infinite cluster. When p varies continuously from 0 to 1, we observe a drastic change of behaviour at a certain critical value p_c. More precisely, elementary monotonicity and ergodic arguments show that there exists a critical parameter p_c such that

$$\begin{aligned} p < p_c \quad &\Longrightarrow \quad \text{all the clusters are finite almost surely,} \\ p > p_c \quad &\Longrightarrow \quad \text{there exists an infinite cluster almost surely.} \end{aligned}$$

In a groundbreaking work, Kesten (1980) proved that $p_c = 1/2$ for Bernoulli percolation on the square lattice and obtained a precise description of the subcritical phase ($p < p_c$) and the supercritical phase ($p > p_c$). The behaviour at $p = p_c = 1/2$ is still the object of famous conjectures in the field, and the present article reviews some recent progress in the study of this critical regime.

We refer to the manuscripts of Grimmett (1999), Bollobás and Riordan (2006) and Werner (2009) for general background on percolation theory.

Organization of this article. In Section 2, we state the new rotation invariance result of Duminil-Copin et al. (2020), and explain its relation to conformal invariance and universality of planar percolation in Section 3. The proof of rotation invariance relies on a discrete tool, the star-triangle transformation. In Section 4, we introduce this transformation, and in Section 5 we explain how it can be used to study the symmetries of certain percolation quantities. In Section 6, we discuss the role of the embedding of the graph and explain how the proof reduces to a key stability lemma.

2. Crossing probabilities and rotation invariance

In this section, we consider critical Bernoulli percolation at $p = p_c = 1/2$ and we discuss the rotation invariance result of Duminil-Copin et al. (2020). To keep this presentation light, we state a weaker version of the result: first we restrict to Bernoulli percolation, while the original result applies to more general models (FK percolation). Second, we state it in terms of rectangle crossings: the original result states that the whole collection of clusters is rotationally invariant, after a suitable truncation. Stating this strong result would require more background, in particular a careful definition of the state space for the collection of clusters.

For every a, b such that $0 \leqslant a \leqslant b$, we define the rectangle

$$R_{a,b} = [-a, a] \times [-b, b].$$

Through this article we identify $\mathbb{R}^2$ with the complex plane $\mathbb{C}$. In particular, we see $R_{a,b}$ as a subset of $\mathbb{C}$. Let ω be a critical Bernoulli percolation of the plane, seen as a

random closed subset of $\mathbb{C}$. We say that $R_{a,b}$ is crossed in ω if there exists an open path in $R_{a,b} \cap \omega$ from the left side $\{-a\} \times [-b, b]$ to the right side $\{a\} \times [-b, b]$. We write $R^{\theta}_{a,b}$ for the rotation of $R_{a,b}$ with angle θ around 0, and say that $R^{\theta}_{a,b}$ is crossed in ω if there exists an open path in $R^{\theta}_{a,b} \cap \omega$ connecting the images (under the θ-rotation) of the left and right sides of $R_{a,b}$. See Figure 1 for an illustration of this event. We emphasize that the connection probabilities are defined in terms of continuous subsets of the plane, hence the crossing events are well defined for arbitrary real numbers a, b, θ.

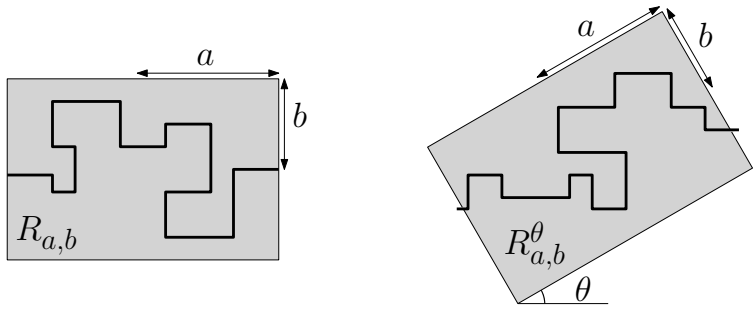


Figure 1: Diagrammatic representations of the events that $R_{a,b} = R^0_{a,b}$ is crossed (left) and $R^{\theta}_{a,b}$ is crossed with an arbitrary angle θ (right). In both cases, the solid path represents an open path connecting the left side to the right side of the rectangle.

Russo (1978), Seymour and Welsh (1978) proved that crossing probabilities with a fixed aspect ratio are non degenerated: For every fixed λ, θ, there exists $c > 0$ such that

$$\forall n \geqslant 1 \qquad c \leqslant \mathbb{P}[R^{\theta}_{\lambda n,n} \text{ is crossed in } \omega] \leqslant 1 - c.$$

The asymptotic behaviour of the critical crossing probabilities is not yet rigorously understood, and is the object of a major open problem (see e.g. Langlands, Pichet, et al., 1992), that we can state as follows.

Conjecture 2.1. *Consider a Bernoulli percolation ω on the square lattice with parameter $p = p_c = 1/2$.*

(*i*) *For every $\lambda \geqslant 1$, $\theta \in [0, \pi/2]$, the sequence $(\mathbb{P}[R^{\theta}_{\lambda n,n}$ is crossed in $\omega])_{n \geqslant 1}$ converges as n tends to infinity.*

(*ii*) *For every $\theta \in [0, \pi/2]$,*

$$\lim_{n \to \infty} P[R^{\theta}_{\lambda n,n} \text{ is crossed in } \omega] = \lim_{n \to \infty} P[R_{\lambda n,n} \text{ is crossed in } \omega].$$

The first part of the conjecture can be interpreted as a "dilatation invariance" of the model: the rectangle $R^{\theta}_{\lambda n,n}$ is a dilatation of the rectangle $R^{\theta}_{\lambda,1}$ by a factor n, and the crossing probabilities for large rectangles do not depend on the dilatation parameter n. The second part corresponds to a rotation invariance: the crossing probabilities for large rectangles do not depend on the angle θ of the rectangle.

Three years ago, Duminil-Copin et al. (2020) proved that crossing probabilities are invariant under rotation (which corresponds to the second item of the conjecture above). More precisely they establish the following theorem.

Theorem 2.2 (Duminil-Copin et al., 2020). *Consider a Bernoulli percolation ω on the square lattice with parameter $p = p_c = 1/2$. For every $\lambda \geqslant 1$ and every rotation angle $\theta \in [0, \pi/2]$, we have*

$$\mathbb{P}[R^{\theta}_{\lambda n,n} \text{ is crossed in } \omega] = \mathbb{P}[R_{\lambda n,n} \text{ is crossed in } \omega](1 + o(1))$$

as n tends to infinity.

Remarks:

- The case $\theta = \frac{\pi}{2}$ is easy because the lattice is already invariant under $\pi/2$-rotation. In contrast, the invariance for $\theta \in (0, \pi/2)$ is nontrivial and can not be deduced from the symmetries of the lattice.
- A self duality argument (see e.g. Grimmett, 1999) implies that the rectangles of the form $[0, n+1] \times [0, n]$ are crossed with probability $1/2$. Therefore, a direct corollary of Theorem 2.2 is that for every $\theta \in [0, \pi/2]$,

$$\lim_{n\to\infty} \mathbb{P}[R^{\theta}_{n,n} \text{ is crossed in } \omega] = \frac{1}{2}.$$

- The theorem does not state that the crossing probabilities converge and the first item in Conjecture 2.1 is still open.

3. Conformal invariance and universality

A much stronger symmetry of the crossing probabilities is conjectured, namely they are expected to be conformally invariant (see Langlands, Pouliot, and Saint-Aubin, 1994 and references therein). To state the conjecture, we use the notion of conformal rectangles, that we now define. Let $\lambda \geqslant 1$. We call conformal rectangle of modulus λ a pair (Ω, ϕ), where $\Omega \subset \mathbb{C}$ is a simply connected open set, and ϕ is a homeomorphism from the rectangle $R_{\lambda,1}$ to $\overline{\Omega}$ such that its restriction $\phi|_{(0,\lambda)\times(0,1)}$ is a conformal map from $(0, \lambda) \times (0, 1)$ to Ω.

For $n \geqslant 1$, notice that the blown up $(n \cdot \Omega, n \cdot \phi)$ is also a conformal rectangle of modulus λ, and in particular it has well-defined left and right sides. We say that $n \cdot \Omega$ is crossed if there exists an open path in $n \cdot \Omega$ from its left to its right side.

Conjecture 3.1 (Convergence to Cardy's formula). *Consider a Bernoulli percolation ω on the square lattice with parameter $p = p_c = 1/2$. For every $\lambda \geqslant 1$, there exists $f(\lambda)$ such that for every conformal rectangle (Ω, ϕ) of modulus λ,*

$$\mathbb{P}[n \cdot \Omega \text{ is crossed in } \omega] \text{ converges to } f(\lambda) \text{ as } n \text{ tends to infinity}.$$

Conjecture 3.1 directly implies Conjecture 2.1, since rotations are particular conformal maps. The explicit expression for $f(\lambda)$ was provided by Cardy (1992) in terms of hypergeometric functions. Later, Carleson noticed that Cardy's formula takes a simple form if one considers crossings in an equilateral triangle rather than in a rectangle. See Smirnov (2001, Corollary 1) or Beffara (2007).

One of the most famous results in the field is the proof of the convergence to Cardy's formula and conformal invariance for critical site percolation on the triangular lattice by Smirnov (2001). See also Khristoforov and Smirnov (2021) for a recent version of the proof. This percolation process is defined by first considering a regular tiling of the plane with hexagons, and then independently declaring each hexagon open or closed with probability $1/2$. Even though the model has a different local description, the asymptotic behaviour is expected to be the same as the one for Bernoulli bond percolation. This is related to the concept of universality, which we now discuss.

Universality. For every infinite planar graph, one can define a critical parameter p_c for the existence of an infinite cluster in Bernoulli percolation (site or bond). The value of p_c and the percolation properties at $p \neq p_c$ generally depend on the underlying graph. In contrast, it is expected that the behaviour at p_c is universal for a large class of planar graphs (see Beffara, 2008 and references therein for a modern discussion). In particular the convergence to Cardy's formula and the emergence of conformal invariance is also expected for critical Bernoulli percolation on $\mathbb{Z}^2$. Even though this statement is strongly supported by non-rigorous renormalization group methods, we are still lacking a rigorous derivation. The "magical" proof of Smirnov for the triangular lattice involves discrete holomorphicity: in some sense, conformal invariance is already present in the discrete model. For more general graphs, we expect this symmetry to emerge only in the scaling limit, and Smirnov's proof does not extend. A more robust approach (inspired by the original physics argument) would be to decompose the proof into two steps: first prove dilatation, translation and rotation invariance, and then extends it to conformal invariance using that a conformal map looks locally like a composition of a dilatation, a translation and a rotation. Some symmetries are already automatic for Bernoulli bond percolation on $\mathbb{Z}^2$, for example translation and $\pi/2$-rotation invariance are inherited from the symmetries of the underlying lattice. Dilatation invariance, which corresponds to the existence of the limit, is quite natural from the renormalization perspective, but today, there is no known rigorous argument. The result of Duminil-Copin et al. (2020) is particularly

impressive because the rotation symmetry was a priori the most mysterious symmetry, and also the most delicate to study (since it is very sensitive to the choice of the embedding of the underlying graph). It is definitely a major step towards a proof of the conformal invariance of critical planar percolation.

4. Star-triangle transformation

Let $p, q, r \in [0, 1]$. In this section, we consider inhomogeneous Bernoulli percolation on two simple weighted graphs, the triangle-graph $\triangle = (V, E)$ with weights (p, q, r) and the star-graph $Y = (W, F)$ with weights $(1-p, 1-q, 1-r)$ as defined on Figure 2. Inhomogeneous means here that the probability to be open can be different for different edges. If an edge has weight p_e, it is open with probability p_e, and closed with probability $1 - p_e$, independently of the other edges.

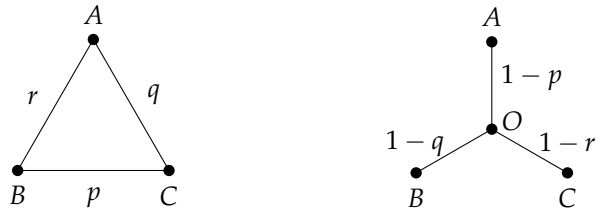


Figure 2: The triangle-graph Δ (left) and the star-graph Y (right).

We define the random partition $\xi^{\triangle}$ of $\{A, B, C\}$ associated to the Bernoulli percolation ω on the triangle graph $\triangle$ as follows: two vertices are in the same element of the partition if they are connected in ω. For example, if the edge AB is open and the two other edges are closed, we have $\xi^{\triangle} = \{AB, C\}$ (by abuse of notation, we write AB for $\{A, B\}$ and C for $\{C\}$). Similarly, we can define the random partition ξ^{Y} resulting from Bernoulli percolation on the star-graph. The star-triangle relation, stated below, asserts that these two partitions have the same distribution if the weights satisfy a certain relation. It was first discovered by Kennelly (1899) in the context of electrical networks. Also known as the Yang-Baxter equation in the physics literature, it was instrumental in work of Onsager (1944) who adapted it to the Ising model, and in the more general study of exact integrability (see e.g. Baxter, 1982). An important application for Bernoulli percolation was found by Sykes and Essam (1964), who used it to predict the critical values for bond percolation on triangular and hexagonal lattices.

Proposition 4.1. *If the edge weights satisfy*

$$p + q + r - pqr = 1, \tag{1}$$

then the two random partitions $\xi^{\triangle}$ and ξ^{Y} have the same distribution.

Proof. We prove that for each of the five partitions P of $\{A, B, C\}$ we have

$$\mathbb{P}[\xi^{\triangle} = P] = \mathbb{P}[\xi^{Y} = P].$$

Since all the probabilities sum to 1 and the three partitions with two elements play symmetric roles, it suffices to check the identity above for $P = \{AB, C\}$ and $P = \{ABC\}$.

Let us first consider the partition $\{AB, C\}$, where A, B are connected together, and C is isolated. On the triangle graph, we obtain this partition if and only if AB is open, and the two other edges are closed. Therefore, we have

$$\mathbb{P}[\xi^{\triangle} = \{AB, C\}] = r(1-p)(1-q).$$

On the star-graph, this happens if and only if OC is closed and the two other edges are open. Therefore, we also have

$$P[\xi^{Y} = \{AB, C\}] = r(1-p)(1-q).$$

Consider the partition $\{ABC\}$ where all the vertices are connected together. On the triangle graph, this happens if and only if at least two edges are open. Hence

$$\begin{aligned}\mathbb{P}[\xi^{\triangle} = \{ABC\}] &= pq(1-r) + p(1-q)r + (1-p)qr + pqr \\ &= pq + pr + qr - 2pqr.\end{aligned}$$

On the star-graph, we obtain this partition if and only if the three edges are open. Therefore

$$\mathbb{P}[\xi^{Y} = \{ABC\}] = (1-p)(1-q)(1-r) = 1 - p - q - r + pq + pr + qr - pqr,$$

and the relation (1) yields $\mathbb{P}[\xi^{\triangle} = \{ABC\}] = \mathbb{P}[\xi^{Y} = \{ABC\}]$. □

How star-triangle transformations can reveal hidden symmetries? On a graph without symmetry, it is hard to compare connection probabilities: for example, given three vertices O, A, B, can one compare the probability that O is connected to A with the probability that it it is connected to B? In the simple example below, we show how the star-triangle transformation can be used to reveal symmetries of the percolation probabilities.

Let $p \in (0,1)$ be the unique solution of the cubic equation $3p - p^3 = 1$ in $(0,1)$ and write $q = 1 - p$. Consider the weighted graph G represented on Figure 3, with vertices A, B, C, D, E. Consider a Bernoulli percolation ω, where the weight of an edge corresponds to the probability that the edge is open. We claim that for percolation on this graph, the probability of A being connected to B is equal to the probability of A being connected to E. To see this, consider the weighted graph G' with vertices A', B, C', E' drawn below.

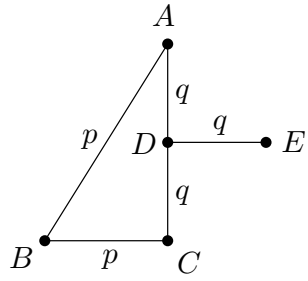


Figure 3: A graph with a hidden percolation symmetry

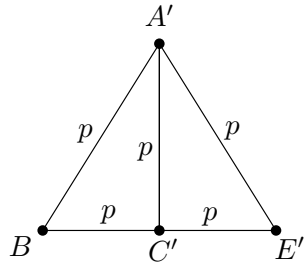


Figure 4: The symmetric graph G' obtained from G after a star-triangle transformation

This new graph can be obtained from G by applying a star-triangle transformation on the "star" in G bounded by ACE and replace it by a triangle $A'C'E'$ (the vertex D is simply deleted and the rest of the graph is left unchanged). By applying Proposition 4.1, we see that the connection probabilities between the vertices A, B, C, E are not affected by this transformation: If ω is a percolation on G and ω' a percolation on G', then we have

$$\mathbb{P}[A \overset{\omega}{\longleftrightarrow} B] = \mathbb{P}[A' \overset{\omega'}{\longleftrightarrow} B] \quad \text{and} \quad \mathbb{P}[A \overset{\omega}{\longleftrightarrow} E] = \mathbb{P}[A' \overset{\omega'}{\longleftrightarrow} E'],$$

where $X \overset{\eta}{\longleftrightarrow} Y$ means that X is connected to Y in η. Using the reflection symmetry of the new graph G', we obtain the desired identity.

Of course, on this simple example, one could have checked this identity by simply computing the probabilities, but for large graphs, exact computations are impossible in practice. In contrast, the star-triangle transformation can be applied repeatedly to compare connection probabilities on large graphs (and even infinite graphs), as discussed in the next section.

5. A hidden rotation symmetry

In this section, we give another concrete example of a graph G with no clear symmetry, and we show that symmetries emerge when we look at percolation properties. The graph G plays an important role in the approach of Duminil-Copin et al. (2020) to prove that critical percolation on $\mathbb{Z}^2$ is rotation invariant.

Let $n \in \mathbb{N}$, and write $\mathrm{k} = e^{\mathrm{i}\pi/4}$. The complex number k can be understood as a parameter that governs the embedding of the considered graphs. In this section, the precise embedding is not important and our choice of $\mathrm{k} = e^{\mathrm{i}\pi/4}$ is arbitrary. In Sections 6 and 7, we will discuss percolation properties that are affected by embedding and we will choose more general parameters of the form $\mathrm{k} = e^{\mathrm{i}\theta}$, $\theta \in (0, \pi/2)$.

Consider the graph H with vertex set $F_f \cup F_u \cup F_r$, where

$$\begin{aligned} F_f &= \{x + \mathrm{i}y \,:\, 0 \leqslant x, y \leqslant n\}, \\ F_u &= \{n\mathrm{i} + x + \mathrm{k}y \,:\, 0 \leqslant x, y \leqslant n\}, \\ F_r &= \{n + \mathrm{k}x + \mathrm{i}y \,:\, 0 \leqslant x, y \leqslant n\}. \end{aligned}$$

and edges between two vertices at Euclidean distance one of each other. As illustrated in Figure 5, the graph H can be seen as a planar representation of a three-dimensional $n \times n \times n$ "full" cube, filled with small cubes of size one. The vertices of F_f, F_u and F_r correspond to the front, upper, and right visual faces of the cube, respectively.

Notice that the graph H is planar and bipartite: its vertices can be partitioned into two sets V and V^c such that all the edges of H connect a vertex in V with a vertex in V^c. Fix V to be the unique such set which contains the origin 0. Construct the graph $G = (V, E)$ with vertex set V and edge set E given by the pairs of vertices bounding the same face (a small square or rhombus) in H. See Figure 5 for an illustration of the construction of G in the case $n = 2$. We define the boundary ∂G as the set of all the vertices of G with degree 1 or 2.

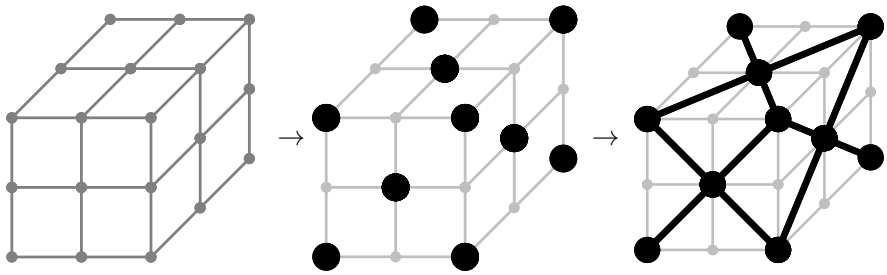


Figure 5: The left picture represents the graph H with $n = 2$. In the middle, the black dots represent the corresponding set V. On the right, the graph G is drawn in black.

We now associate some weights to the edges of G. Fix $p, q \in (0,1)$ such that

$$2p + 2q - pq = 1.$$

For each edge $e = \{v, w\}$, write $\vec{e} = v - w$ with the convention that the x-coordinate of $\vec{e}$ is positive, and define the weight (illustrated on Figure 6)

$$p_e = \begin{cases} 1/2 & \text{if } \vec{e} = 1 + \mathrm{i} \text{ or } 1 - \mathrm{i}, \\ p & \text{if } \vec{e} = 1 + \mathrm{k}, \\ 1 - p & \text{if } \vec{e} = 1 - \mathrm{k}, \\ q & \text{if } \vec{e} = \mathrm{k} + \mathrm{i}, \\ 1 - p & \text{if } \vec{e} = \mathrm{k} - \mathrm{i}. \end{cases} \tag{2}$$

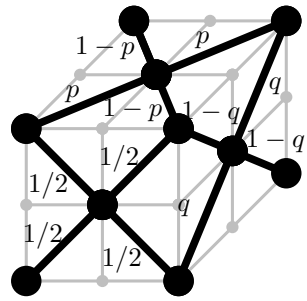


Figure 6: Representation of G and its associated weights for $n = 2$.

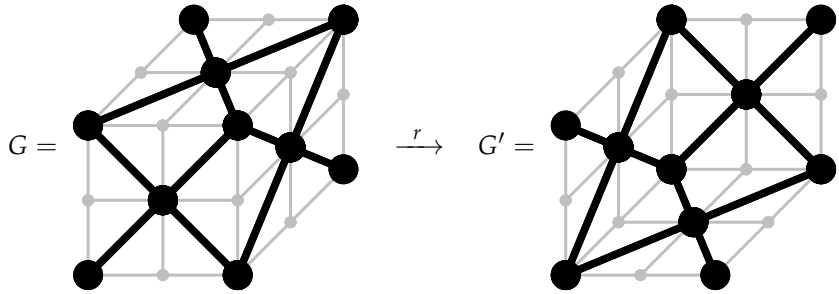
Notice that the edges between vertices of the front face F_f have the weight $1/2$, the edges on the upper face F_u receive the weights p or $1 - p$ and the edges on the right face F_r receive the weights q or $1 - q$.

Let r be the π-rotation of the plane with center $C = \frac{n}{2}(1 + \mathrm{i} + \mathrm{k})$. If n is an even integer, then the boundary ∂G is invariant under the rotation r. Therefore, for every partition P of G, we can define a rotated partition $r \cdot P$, the elements of which are the images of the elements of P under the rotation r. An inhomogeneous Bernoulli percolation ω on the weighted graph G gives rise to a random partition ξ of ∂G: two vertices of ∂G are in the same element of the partition ξ if they are connected together in ω. We can rotate ξ, and consider another random partition $r \cdot \xi$ of ∂G. At first, the distributions of ξ and $r \cdot \xi$ seem to be different, since the underlying graph is not invariant under r. Despite this lack of symmetry of G, we will be able to use the star-triangle transformation to show that the random partition ξ is rotation invariant, as formally stated in the following proposition.

Proposition 5.1. *Let $n \geqslant 2$ even and consider the graph G as above. Let ξ be the partition of ∂G generated by Bernoulli percolation on the weighted graph (G, p). For every partition P of ∂G, we have*

$$\mathbb{P}[\xi = P] = \mathbb{P}[r \cdot \xi = P].$$

Sketch of proof. Let $G' = r \cdot G$ be the image of G under the π-rotation around C.



Since n is even, G and G' have the same boundary. It is immediate that the partition ξ' of $\partial G'$ generated by percolation on G' has the same distribution as $r \cdot \xi$, since a Bernoulli percolation on G' can be obtained by rotating a percolation on G. As a consequence, for every partition P of ∂G, we have

$$\mathbb{P}[\xi' = P] = \mathbb{P}[r \cdot \xi = P].$$

We now show that G' can be alternatively obtained from G by using successive star-triangle transformations, without rotation. This will imply that ξ' and ξ have the same distribution, and therefore conclude the proof.

We first start by constructing a sequence of graphs interpolating from H to H', defined as the image of H under a π-rotation around C. As mentioned above, the graph H can be visualized as a three-dimensional $n \times n \times n$ cube, filled with n^3 small cubes of size one. Similarly, the graph H' can be seen as an emptied version of it. With this interpretation, a natural way to go from H to H' is to "remove small cubes" one by one as illustrated on Fig. 7 for $n = 2$. This defines a sequence of graphs

$$H = H_0, H_1, \ldots, H_{n^3-1}, H_{n^3} = H'.$$

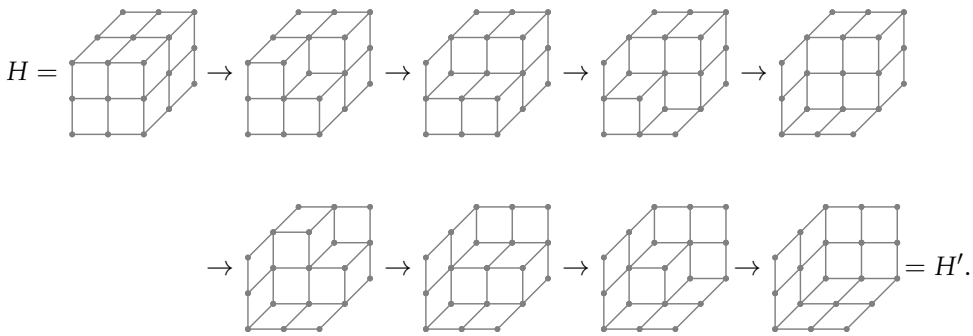


Figure 7: Transforming H to H' by "removing cubes" one by one

For every i, the graph H_i is bipartite and planar, and we can define a weighted graph G_i associated to H_i exactly as we defined G from H. In particular all the edges e of G_i are such that $\vec{e}$ belongs to $\{1+\mathrm{i}, 1-\mathrm{i}, \mathrm{k}+1, \mathrm{k}-1, \mathrm{k}+\mathrm{i}, \mathrm{k}-\mathrm{i}\}$, and we can define the weights of the edges of G_i following Equation (2).

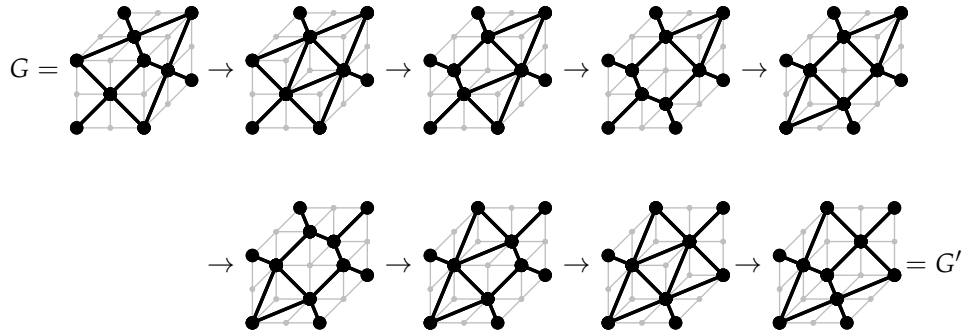


Figure 8: Local transformations from G to G'

The key observation is that G_{i+1} is obtained from G_i by applying one star-triangle transformation (at the place where the small cube is removed in H_i). If i is even, a star with weights $(1-p, 1-q, 1/2)$ is replaced by a triangle with weights $(p, q, 1/2)$ (as illustrated in Figure 9).

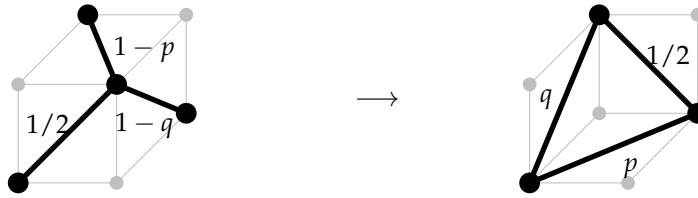


Figure 9: For i even, "removing one cube" corresponds to a star triangle transformation on G_i.

Analogously, if i is odd, a triangle with weights $(p, q, 1/2)$ is replaced by a star with weights $(1-p, 1-q, 1/2)$.

In particular, the random partitions ξ_i and ξ_{i+1} of ∂G resulting from percolation on the weighted graphs G_i and G_{i+1} have the same distribution (since none of the vertices of ∂G is removed by the transformation). By induction, we deduce that $\xi = \xi_0$ and $\xi' = \xi_{n^3}$ have the same distribution: for every partition P of ∂G, we have

$$\mathbb{P}[\xi = P] = \mathbb{P}[\xi' = P].$$

Since ξ' has the same distribution as $r \cdot \xi$, this concludes the proof. □

6. Isoradial embedding of the cube graph

In the previous sections, we saw that star-triangle transformations relate the percolation properties of two different graphs: the connection probabilities of the vertices left unchanged after several transformations are invariant. For example, when going from G to G', none of the vertices of ∂G is affected by the local transformations, and their connection probabilities are therefore left unchanged. However, analyzing the percolation properties of the *bulk vertices* (*i.e.*, the vertices of G' or G that are not at the boundary) is much more delicate. This complexity arises from the fact that when we go from G to G', all the bulk vertices of G are deleted. At some point, each bulk vertex becomes the center of a star-graph, which is transformed into a triangle, and the connection probabilities to such vertex are "lost" in the process.

In all the star-triangle statements we have seen so far, the identities were graph properties, and they were insensitive to the precise way the graphs were embedded in the plane. In Section 5, we chose the same embedding for all the graphs G, regardless of the values of p and q. However, in the study of bulk vertices, it becomes crucial to choose a suitable embedding for the graph G, which depends on the values of the weights p and q. We will now describe the isoradial embedding of the graph G, which represents the "correct" way to embed it in order to preserve both boundary and bulk connectivity properties.

Let $\theta \in [0, \pi/2]$. We consider the plane graph $G = G_n(\mathrm{k}, p, q)$ to be exactly the same graph as in Section 5, with the following choice of parameters:

$$\mathrm{k} = e^{i\theta}, \quad \frac{p}{1-p} = \frac{\sin(\frac{\theta}{3})}{\sin(\frac{\pi-\theta}{3})}, \quad \frac{q}{1-q} = \frac{\sin(\frac{\pi+2\theta}{6})}{\sin(\frac{\pi-2\theta}{6})}. \tag{3}$$

These weights originate from the interpretation of G as an isoradial graph. They were first introduced in the work of Kenyon (2004), and their significance was further emphasized in the work by Grimmett and Manolescu (2013a,b, 2014). For a precise definition of isoradial graphs and more details about these weights, we direct the interested reader to these references. An important feature of these weights is that they satisfy the star-triangle relation:

Lemma 6.1. *For every $\theta \in (0, \pi/2)$, the weights defined above satisfy*

$$2p + 2q - pq = 1.$$

Proof. First notice that the equation is equivalent to

$$x + y + 2xy = 1,$$

where $x = \frac{p}{1-p}$ and $y = \frac{q}{1-q}$. Using the explicit formula (3), this is again equivalent to

$$s_\theta s_{\pi/2-\theta} + s_{\pi-\theta} s_{\pi/2+\theta} + 2 s_\theta s_{\pi/2+\theta} = s_{\pi-\theta} s_{\pi/2-\theta},$$

where $s_\phi = \sin(\phi/3)$. Finally, this last equation can be deduced by elementary trigonometric computations. For example, one can replace each term using that for every a, b, c,

$$s_a s_b = \cos(\tfrac{a-b}{3}) - \cos(\tfrac{a+b}{3}). \qquad \square$$

We consider the following two regions in the plane

$$F = [0, n] + \mathrm{i} \cdot [0, n], \quad \text{and} \quad U = \mathrm{i}n + [0, n] + \mathrm{k} \cdot [0, n],$$

corresponding to the front and upper faces of G, respectively.

The part of G in the upper face U is invariant under the reflection with axis $\mathrm{i}n + \mathbb{R}e^{\mathrm{i}\theta/2}$ (thick line in Figure 10), which implies the following θ-rotation invariance property for rectangle crossings. Let $R = [a, b] \times [c, d] \subset U$ be a rectangle centered at $z = \mathrm{i}n + \frac{n}{2} + \frac{n}{2}e^{\mathrm{i}\theta}$ and R' be the image of R under the θ-rotation around z. Alternatively, the rectangle R' can also be seen as a reflected version of R through the axis $\mathrm{i}n + \mathbb{R}e^{\mathrm{i}\theta/2}$. The reflection symmetry mentioned above implies that R and R' are crossed with the same probability, as illustrated in Figure 10.

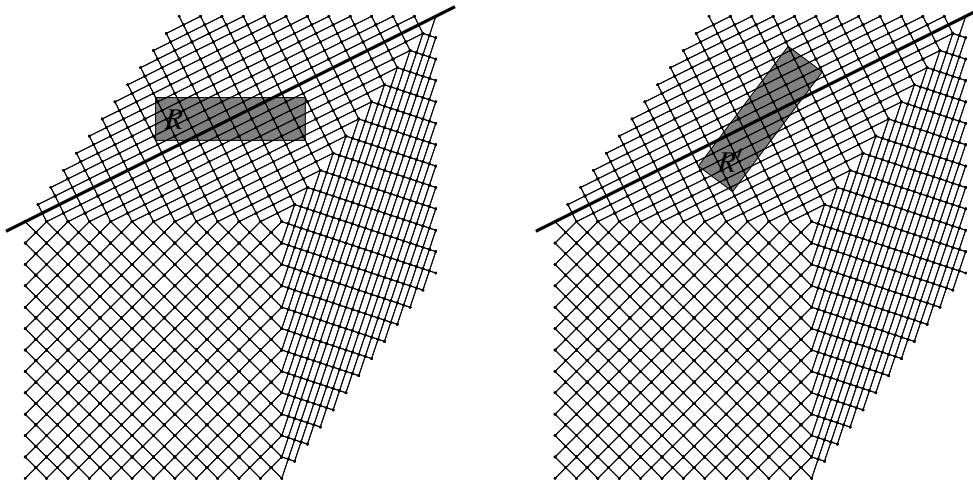


Figure 10: The rectangle R (filled in gray on the left) and its reflected version R' (right) have the same connection probabilities

Our aim is to "import" this symmetry from the upper face U to the front face F, in order to show that crossing probabilities in the front face F are also invariant under θ-rotation. To accomplish this, we rely on the star-triangle transformation to swap the two faces, together with a bulk stability result that ensures the preservation of percolation properties.

Let G' denote the π-rotated version of G, as discussed in Section 5. We define F' and U' as the images of F and U under the rotation that transforms G into G'.

Let $z' = \frac{n}{2} + \frac{n}{2}e^{i\theta}$ be the center of U' and consider the $\frac{n}{4} \times \frac{n}{8}$ rectangle centered at C', defined by

$$R_n = z' + [-\tfrac{n}{8}, \tfrac{n}{8}] \times [-\tfrac{n}{16}, \tfrac{n}{16}].$$

Notice that this rectangle has aspect ratio $\lambda = 2$ and it is a translated version of the rectangle $R_{n/8,n/16}$ introduced in Section 2. For every $\phi \in [0, \pi/2]$, we define R_n^ϕ as the image of R_n after a ϕ-rotation around C', the center of gravity of R_n. An important fact behind this choice is that all the rectangles R_n^ϕ belong to the F-face of G and the U'-face of G' if $\theta \in [\pi/4, \pi/2]$, as illustrated on Figure 11.

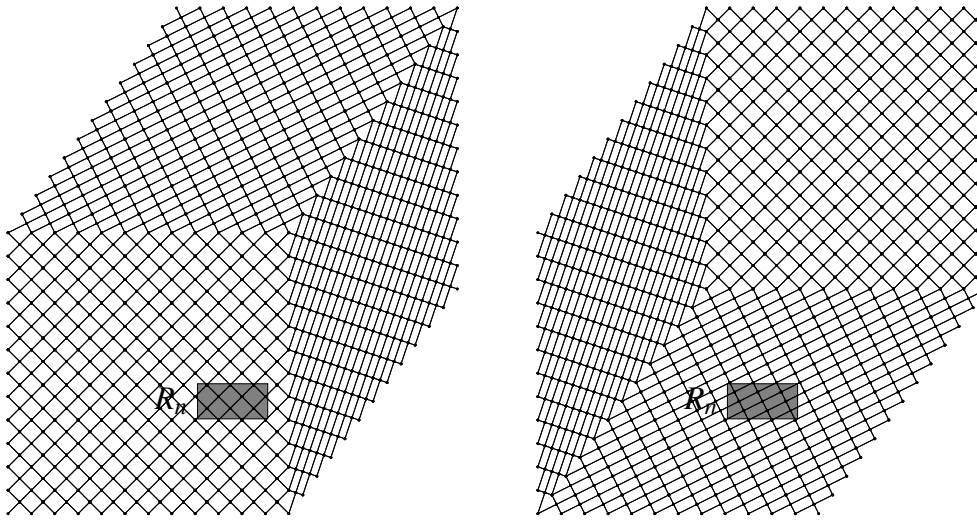


Figure 11: The rectangle R_n is a subset of the F-face of G (left) and is centered in the U'-face of G' (right).

The following statement constitutes the key lemma in the proof of Duminil-Copin et al. (2020):

Lemma 6.2 (Bulk stability of crossings). *Let $\theta \in [\frac{\pi}{4}, \frac{\pi}{2}]$. Let G be the weighted graph with parameters given by Equation (3), and G' its rotated version as above. Let ω and ω' be Bernoulli percolations on the weighted graphs G and G' respectively. For every $\phi \in [0, \frac{\pi}{2}]$, we have*

$$\mathbb{P}[R_n^\phi \text{ is crossed in } \omega] = \mathbb{P}[R_n^\phi \text{ is crossed in } \omega'](1 + o(1)) \tag{4}$$

as n tends to infinity.

We now explain how to deduce the proof of the main result (Theorem 2.2) from this lemma. We restrict ourselves to rectangles with aspect ratio $\lambda = 2$ for simplicity, the proof trivially extends to general λ after minor adjustments. Let us fix $\theta \in [\pi/4, \pi/2]$. As mentioned at the beginning of the section, the upper face U has a $\theta/2$-reflection symmetry. The U'-face of G' has the same property, which implies

$$\mathbb{P}[R_n \text{ is crossed in } \omega'] = \mathbb{P}[R_n^\theta \text{ is crossed in } \omega']. \tag{5}$$

We combine this observation with the stability result of Lemma 6.2. As n tends to infinity, we have

$$\begin{aligned}
\mathbb{P}[R_n^\theta \text{ is crossed in } \omega] &= \mathbb{P}[R_n^\theta \text{ is crossed in } \omega'](1+o(1)) && \text{by applying (4) to } \phi = \theta \\
&= \mathbb{P}[R_n \text{ is crossed in } \omega'](1+o(1)) && \text{by (5)} \\
&= \mathbb{P}[R_n \text{ is crossed in } \omega](1+o(1)) && \text{by applying (4) to } \phi = 0.
\end{aligned}$$

This proves that R_n and R_n^θ are asymptotically crossed with the same probability for every fixed $\theta \in [\pi/4, \pi/2]$. We finally extend it to all angles θ by using reflection invariance of the homogeneous square lattice.

7. A random walk argument

In this section, we present the strategy used by Duminil-Copin et al. (2020) to establish the bulk stability of crossing probabilities (key Lemma 6.2). The aim is to show that connection probabilities in ω (Bernoulli percolation on G) and ω' (Bernoulli percolation on G') are close to each other. To achieve this, we couple these two configurations, and construct a sequence of intermediate configurations $\omega = \omega_0, \omega_1, \dots, \omega_n = \omega'$ where ω_j is a Bernoulli percolation on $G_{n^2 j}$ (as in Section 5, G_i denotes the graph obtained from G after performing i successive star-triangle transformations, visually corresponding to removing i "small cubes" in the underlying graph H). In particular, the configuration ω_{j+1} is obtained from ω_j by performing $\simeq n^2$ star-triangle transformations.

In this coupling, we keep track of all the macroscopic clusters (say, the clusters of radius larger than $0.0001n$). Fix one such cluster C_0 in ω_0 and consider the sequence of corresponding clusters $C_1, C_2, \dots, C_n$ in the configurations $\omega_1, \dots, \omega_n$. To each cluster C_j of this sequence, associate its top-most left-most point X_j. The way X_j is affected by the star-triangle transformations is very local, and the authors show that $X_0, X_1, \dots, X_n$ behaves like an n-step random walk. Therefore, by the law of large numbers, there exists some $\delta \in \mathbb{C}$ such that almost surely, we have

$$X_n = \delta n + o(n)$$

as n tends to infinity. A key step is then to show that the drift δ of this random walk vanishes. In a first version of their paper, the authors managed to prove this fact

using a mapping to the six-vertex model, and the Bethe Ansatz. Recently the authors noticed that $\delta = 0$ follows from a simpler algebraic argument involving the symmetry of $\mathbb{Z}^2$. This argument will be presented in a forthcoming version of their paper. As a consequence, X_n is at distance $o(n)$ from X_0. This random walk argument can be extended to other points in order to show that open paths are not too much affected by the star-triangle transformations. Fix $\varepsilon > 0$. With probability $1 - o(1)$, the following holds: For every macroscopic open path γ in ω, there exists another open path γ' such that the Hausdorff distance between them satisfies

$$\mathrm{d}_{\mathrm{Hausdorff}}(\gamma, \gamma') \leqslant \varepsilon n.$$

The statement above asserts that ω and ω' can be coupled in such a way that all the macroscopic paths in ω are within εn distance from the macroscopic paths in ω'. From there, the desired result on crossing probabilities follows from a classical continuity argument in percolation theory.

References

Baxter, R. J. (1982). *Exactly solved models in statistical mechanics*. Academic Press, London.

Beffara, V. (2007). "Cardy's formula on the triangular lattice, the easy way", in: *Universality and renormalization*. Vol. 50. Fields Inst. Commun. Amer. Math. Soc., pp. 39–45.

——— (2008). "Is critical 2D percolation universal?", in: *In and out of equilibrium. 2*. Vol. 60. Progr. Probab. Birkhäuser, Basel, pp. 31–58.

Bollobás, B. and Riordan, O. (2006). *Percolation*. Cambridge University Press, New York.

Broadbent, S. R. and Hammersley, J. M. (1957). "Percolation processes. I. Crystals and mazes", *Proc. Cambridge Philos. Soc.* **53**, pp. 629–641.

Cardy, J. L. (1992). "Critical percolation in finite geometries", *J. Phys. A* **25** (4), pp. 201–206.

Duminil-Copin, H. et al. (2020). "Rotational invariance in critical planar lattice models", arXiv: 2012.11672.

Grimmett, G. R. (1999). *Percolation*. Second. Vol. 321. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin.

Grimmett, G. R. and Manolescu, I. (2013a). "Inhomogeneous bond percolation on square, triangular and hexagonal lattices", *Ann. Probab.* **41** (4), pp. 2990–3025.

——— (2013b). "Universality for bond percolation in two dimensions", *Ann. Probab.* **41** (5), pp. 3261–3283.

——— (2014). "Bond percolation on isoradial graphs: criticality and universality", *Probab. Theory Related Fields* **159** (1-2), pp. 273–327.

Kennelly, A. E. (1899). "The equivalence of triangles and three-pointed stars in conducting networks", *Electrical world and engineer* **34** (12), pp. 413–414.

Kenyon, R. (2004). "An introduction to the dimer model", in: *School and Conference on Probability Theory*. ICTP Lect. Notes, XVII. Abdus Salam Int. Cent. Theoret. Phys., Trieste, pp. 267–304.

Kesten, H. (1980). "The critical probability of bond percolation on the square lattice equals $\frac{1}{2}$", *Comm. Math. Phys.* **74** (1), pp. 41–59.

Khristoforov, M. and Smirnov, S. (2021). "Percolation and $O(1)$ loop model", arXiv: arXiv:2111.15612.

Langlands, R. P., Pichet, C., et al. (1992). "On the universality of crossing probabilities in two-dimensional percolation", *J. Statist. Phys.* **67** (3-4), pp. 553–574.

Langlands, R. P., Pouliot, P., and Saint-Aubin, Y. (1994). "Conformal invariance in two-dimensional percolation", *Bull. Amer. Math. Soc. (N.S.)* **30** (1), pp. 1–61.

Onsager, L. (1944). "Crystal statistics. I. A two-dimensional model with an order-disorder transition", *Phys. Rev. (2)* **65**, pp. 117–149.

Russo, L. (1978). "A note on percolation", *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **43** (1), pp. 39–48.

Seymour, P. D. and Welsh, D. J. A. (1978). "Percolation probabilities on the square lattice", *Ann. Discrete Math.* **3**, pp. 227–245.

Smirnov, S. (2001). "Critical percolation in the plane: conformal invariance, Cardy's formula, scaling limits", *C. R. Acad. Sci. Paris Sér. I Math.* **333** (3), pp. 239–244.

Sykes, M. F. and Essam, J. W. (1964). "Exact critical percolation probabilities for site and bond problems in two dimensions", *J. Mathematical Phys.* **5**, pp. 1117–1127.

Werner, W. (2009). *Percolation et modèle d'Ising*. Vol. 16. Cours Spécialisés. Société Mathématique de France, Paris.

Vincent Tassion

ETH Zürich,
Rämistrasse 101,
8092 Zurich, Switzerland

E-mail: vincent.tassion@math.ethz.ch

Séminaire BOURBAKI, 1948/49 à 2022/2023

Exposés 1 à 1210

Les volumes 1948/49 à 1967/68, Exposés 1 à 346, initialement publiés par *W.A. Benjamin, Inc.* New York, ont été réimprimés en 1996 par la Société mathématique de France sous forme d'un ensemble de 10 volumes hors série de la collection *Astérisque* :

vol. 1 : 1948/49, 1949/50, 1950/51 ;
vol. 2 : 1951/52-1952/53, 1953/54 ;
vol. 3 : 1954/55, 1955/56 ;
vol. 4 : 1956/57, 1957/58 ;
vol. 5 : 1958/59, 1959/60 ;
vol. 6 : 1960/61 ;
vol. 7 : 1961/62 ;
vol. 8 : 1962/63, 1963/64 ;
vol. 9 : 1964/65, 1965/66 ;
vol. 10 : 1966/67, 1967/68.

Les volumes 1968/69 à 1980/81, Exposés 347 à 578, ont été publiés par *Springer–Verlag*, collection Lecture Notes in Mathematics :

vol. 1968/69, n° 179, 1971 ;
vol. 1969/70, n° 180, 1971 ;
vol. 1970/71, n° 244, 1971 ;
vol. 1971/72, n° 317, 1973 ;
vol. 1972/73, n° 383, 1974 ;
vol. 1973/74, n° 431, 1975 ;
vol. 1974/75, n° 514, 1976 ;
vol. 1975/76, n° 567, 1977 ;
vol. 1976/77, n° 677, 1978 ;
vol. 1977/78, n° 710, 1979 ;
vol. 1978/79, n° 770, 1980 ;
vol. 1979/80, n° 842, 1981 ;
vol. 1980/81, n° 901, 1981.

Les volumes 1981/82 à 2022/23, Exposés 579 à 1210, ont été publiés par la Société mathématique de France dans la collection *Astérisque* :

vol. 1981/82, n° 92-93, 1982 ;
vol. 1982/83, n° 105-106, 1983 ;
vol. 1983/84, n° 121-122, 1985 ;
vol. 1984/85, n° 133-134, 1986 ;
vol. 1985/86, n° 145-146, 1987 ;
vol. 1986/87, n° 152-153, 1987 ;
vol. 1987/88, n° 161-162, 1988 ;
vol. 1988/89, n° 177-178, 1989 ;
vol. 1989/90, n° 189-190, 1990 ;
vol. 1990/91, n° 201-202-203, 1991 ;
vol. 1991/92, n° 206, 1992 ;
vol. 1992/93, n° 216, 1993 ;
vol. 1993/94, n° 227, 1995 ;
vol. 1994/95, n° 237, 1996 ;
vol. 1995/96, n° 241, 1997 ;
vol. 1996/97, n° 245, 1997 ;
vol. 1997/98, n° 252, 1998 ;
vol. 1998/99, n° 266, 2000 ;
vol. 1999/2000, n° 276, 2002 ;
vol. 2000/01, n° 282, 2002 ;
vol. 2001/02, n° 290, 2003 ;
vol. 2002/03, n° 294, 2004 ;
vol. 2003/04, n° 299, 2005 ;
vol. 2004/05, n° 307, 2006 ;
vol. 2005/06, n° 311, 2007 ;
vol. 2006/07, n° 317, 2008 ;
vol. 2007/08, n° 326, 2009 ;
vol. 2008/09, n° 332, 2010 ;
vol. 2009/10, n° 339, 2011 ;
vol. 2010/11, n° 348, 2012 ;
vol. 2011/12, n° 352, 2013 ;
vol. 2012/13, n° 361, 2014 ;
vol. 2013/14, n° 367–368, 2015 ;
vol. 2014/15, n° 380, 2016 ;
vol. 2015/16, n° 390, 2017 ;
vol. 2016/17, n° 407 2019 ;
vol. 2017/18, n° 414, 2019 ;
vol. 2018/19, n° 422, 2020 ;
vol. 2019/21, n° 430, 2022 ;
vol. 2021/22, n° 438, 2022 ;
vol. 2022/23, n° 446, 2023 ;

Une table des exposés du Séminaire Bourbaki, classée par noms d'auteurs, est disponible à l'adresse `https://bourbaki.fr/table.pdf`

ASTÉRISQUE

2023

445. F. Sukochev & D. Zanin – *The Connes character formula for locally compact spectral triples*
444. J. Szeftel – *Parametrix for wave equations on a rough background IV : Control of the error term*
443. J. Szeftel – *Parametrix for wave equations on a rough background I : Regularity of the phase at initial time. II : Construction and control at initial time*
442. G. David, J. Feneuil & S. Mayboroda – *Elliptic theory in domains with boundaries of mixed dimension*
441. J. Calvert, A. Hammond & M. Hegde – *Brownian structure in the KPZ fixed point*
440. S. Guillermou – *Sheaves and symplectic geometry of cotangent bundles*
439. F. Diamond, P. Kassaei & S. Sasakia – *A mod p Jacquet-Langlands relation and Serre filtration via the geometry of Hilbert modular varieties : Splicing and dicing*

2022

438. Séminaire Bourbaki, *volume 2021/2022, Exposés 1181–1196*
437. A. Borodin & M. Wheeler – *Coloured stochastic vertex models and their spectral theory*
436. S.-J. Oh & D. Tataru – *The Yang-Mills heat flow and the caloric gauge*
435. R. Donagi & T. Pantev – *Parabolic Hecke eigensheaves*
434. M. Bertolini, H. Darmon, V. Rotger, M. A. Seveso & R. Venerucci – *Heegner points, stark-Heegner points, and diagonal classes*
433. F. Binda, D. Park & P. A. Ostvaer – *Triangulated categories of logarithmic motives over a field*
432. Y. Wakabayashi – *A theory of dormant opers on pointed stable curves*
431. Q. Guignard – *Geometric local ε-factors*
430. Séminaire Bourbaki, *volume 2019/2021, Exposés 1166–1180*

2021

429. E. Gwynne & J. Miller – *Percolation on uniform quadrangulations and* SLE_6 *on* $\sqrt{8/3}$*-Liouville quantum gravity*
428. K. Prasanna & A. Venkatesh – *Automorphic cohomology, motivic cohomology, and the adjoint L-function*
427. B. Duplantier, J. Miller & S. Sheffield – *Liouville quantum gravity as a mating of trees*
426. P. Biran, O. Cornea & E. Shelukhin – *Lagrangian shadows and triangulated categories*
425. T. Bachmann & M. Hoyois – *Norms in motivic homotopy theory*
424. B. Bhatt, J. Lurie & A. Mathew – *Revisiting the de Rham-Witt complex*
423. K. Ardakov – *Equivariant $\mathscr{D}$-modules on rigid analytic spaces*

2020

422. Séminaire Bourbaki, *volume 2018/2019, Exposés 1151–1165*
421. J.H. Bruinier, B. Howard, S.S. Kudla, K. Madapusi Pera, M. Rapoport & T. Yang – *Arithmetic divisors on orthogonal and unitary Shimura varieties*

420. H. Ringström – *Linear systems of wave equations on cosmological backgrounds with convergent asymptotics*
419. V. Gorbounov, O. Gwilliam & B. Williams – *Chiral differential operators via quantization of the holomorphic σ-model*
418. R. Beuzart–Plessis – *A local trace formula for the Gan-Gross-Prasad conjecture for unitary groups : the Archimedean case*
417. J.D. Adams, M. Van Leeuwen, P.E. Trapa & D.A. Vogan, Jr. – *Unitary representations of real reductive groups*
416. S. Crovisier, R. Krikorian, C. Matheus & S. Senti (eds.) – *Some aspects of the theory of dynamical systems : A tribute to Jean-Christophe Yoccoz, II*
415. S. Crovisier, R. Krikorian, C. Matheus & S. Senti (eds.) – *Some aspects of the theory of dynamical systems : A tribute to Jean-Christophe Yoccoz, I*

2019

414. Séminaire Bourbaki, *Volume 2017/2018, Exposés 1136–1150*
413. M. Crainic, R. Loja Fernandes & D. Martínez Torres – *Regular Poisson manifolds of compact types*
412. E. Herscovich – *Renormalization in quantum field theory* (*after R. Borcherds*)
411. G. David – *Local regularity properties of almost- and quasiminimal sets with a sliding boundary condition*
410. P. Berger & J.-C. Yoccoz – *Strong regularity*
409. F. Calegari & A. Venkatesh – *A torsion Jacquet-Langlans correspondence*
408. D. Maulik & A. Okounkov – *Quantum groups and quantum cohomology*
407. Séminaire Bourbaki, *Volume 2016/2017, Exposés 1120–1135*

2018

406. L. Fargues & J.-M. Fontaine – *Courbes et fibrés vectoriels en théorie de Hodge p-adique* (*préface P. Colmez*)
405. J.-F. Bony, S. Fujiié, T. Ramond & M. Zerzeri – *Resonances for homoclinic trapped sets*
404. O. Matte & J. S. Møller – *Feynman-Kac formulas for the ultra-violet renormalized Nelson model*
403. M. Berti, T. Kappeler & R. Montalto – *Large Kam tori for perturbations of the defocusing NLS equation*
402. H. Bao & W. Wang – *A new approach to Kazhdan-Lustig theory of type B via quantum symmetric pairs*
401. J. Szeftel – *Parametrix for wave equations on a rough background III : space-time regularity of the phase*
400. A. Ducros – *Families of Berkovich spaces*
399. T. Lidman & C. Manolescu – *The equivalence of two Seiberg-Witten Floer homologies*
398. W. Teck Gan, F. Gao, W. H. Weissman – *L-groups and the Langlands program for covering groups*
397. S. Riche & G. Williamson – *Tilting modules and the p-canonical basis*