Bulletin

de la SOCIÉTÉ MATHÉMATIQUE DE FRANCE

Tome 150
Fascicule 1

2022

SOCIÉTÉ MATHÉMATIQUE DE FRANCE

Pages 1-249

Bull. Soc. Math. France
150 (1), 2022, p. 1-249

Sommaire

Bull. Soc. Math. France
150 (1), 2022, p. 1-249

Contents

Bull. Soc. Math. France
150 (1), 2022, p. 1-16

MAXIMAL ESTIMATES FOR THE KRAMERS–FOKKER–PLANCK OPERATOR WITH ELECTROMAGNETIC FIELD

by Bernard Helffer & Zeinab Karaki

Abstract. — In continuation of a former work by the first author with F. Nier (2009) and of a more recent work by the second author on the torus (2019), we consider the Kramers–Fokker–Planck operator (KFP) with an external electromagnetic field on $\mathbb{R}^d$. We show a maximal type estimate on this operator using a nilpotent approach for vector field polynomial operators and induced representations of a nilpotent graded Lie algebra. This estimate leads to an optimal characterization of the domain of the closure of the (KFP) operator and a criterion for the compactness of the resolvent.

Résumé (*Estimation maximale pour l'opérateur de Kramers-Fokker-Planck avec champ électromagnétique*). — Dans la continuité d'un travail antérieur du premier auteur avec F. Nier (2009) et d'un travail plus récent du deuxième auteur sur le tore (2019), nous considérons l'opérateur de Kramers-Fokker-Planck (KFP) avec un champ électromagnétique sur $\mathbb{R}^d$. Nous montrons une estimation de type maximal sur cet opérateur en utilisant une approche nilpotente pour les opérateurs polynômes de champs de vecteurs et des représentations induites d'une algèbre de Lie graduée nilpotente. Cette estimation conduit à une caractérisation optimale du domaine de la fermeture de l'opérateur (KFP) et à un critère de compacité de la résolvante.

Texte reçu le 25 janvier 2021, accepté le 20 juillet 2021.

Bernard Helffer, Université de Nantes, Laboratoire de Mathématiques Jean Leray, 2, rue de la Houssinière, BP 92208 F-44322 Nantes Cedex 3, France • *E-mail :* `bernard.helffer@univ-nantes.fr`

Zeinab Karaki, Université de Nantes, Laboratoire de Mathématiques Jean Leray, 2, rue de la Houssinière, BP 92208 F-44322 Nantes Cedex 3, France • *E-mail :* `zeinab.karaki@univ-nantes.fr`

Mathematical subject classification (2010). — 35H10, 47A67; 82C40, 82D40.

Key words and phrases. — Fokker–Planck equation, magnetic field, electric potential, Lie algebra, irreducible representation, maximal estimate.

doi:10.24033/bsmf.2842

1. Introduction and main results

1.1. Introduction. — The Fokker–Planck equation was introduced by Fokker and Planck at the beginning of the twentieth century to describe the evolution of the density of particles under Brownian motion. In recent years, global hypoelliptic estimates have led to new results motivated by applications to the kinetic theory of gases. In this direction, many authors have shown maximal estimates to deduce the compactness of the resolvent of the Fokker–Planck operator and to have resolvent estimates in order to address the issue of return to the equilibrium. F. Hérau and F. Nier in [5] highlighted the links between the Fokker–Planck operator with a confining potential and the associated Witten Laplacian. Later, this work was extended in the book of B. Helffer and F. Nier [2], and we refer more specifically to their Chapter 9 for a proof of the maximal estimate.

In this article, we continue the study of the model case of the operator of Fokker–Planck with an external magnetic field B_e, which was initiated in the case of the torus $\mathbb{T}^d$ ($d = 2, 3$) in [9, 10], by considering $\mathbb{R}^d$ and reintroducing an electric potential as in [2]. In this context, we establish a maximal-type estimate for this model, giving a characterization of the domain of its closed extension and giving sufficient conditions for the compactness of the resolvent.

1.2. Statement of the result. — For $d = 2$ or 3, we consider, for a given external electromagnetic field B_e defined on $\mathbb{R}^d$ with value in $\mathbb{R}^{d(d-1)/2}$ and a real valued electric potential V defined on $\mathbb{R}^d$, the associated Kramers–Fokker–Planck operator K (in short KFP) defined by:

$$K = v \cdot \nabla_x - \nabla_x V \cdot \nabla_v - (v \wedge B_e) \cdot \nabla_v - \Delta_v + v^2/4 - d/2 \,, \tag{1}$$

where $v \in \mathbb{R}^d$ represents the velocity, $x \in \mathbb{R}^d$ represents the space variable, and the notation $(v \wedge B_e) \cdot \nabla_v$ means:

$$(v \wedge B_e) \cdot \nabla_v = \begin{cases} b(x)\,(v_1\partial_{v_2} - v_2\partial_{v_1}) & \text{if } d = 2 \\ b_1(x)(v_2\partial_{v_3} - v_3\partial_{v_2}) + b_2(x)(v_3\partial_{v_1} - v_1\partial_{v_3}) & \\ \quad + b_3(x)(v_1\partial_{v_2} - v_2\partial_{v_1}) & \text{if } d = 3 \,. \end{cases}$$

The operator K is initially considered as an unbounded operator on the Hilbert space $L^2(\mathbb{R}^d \times \mathbb{R}^d)$, whose domain is $D(K) = C_0^\infty(\mathbb{R}^d \times \mathbb{R}^d)$.

We then denote by:

- $K_{\min}$ the minimal extension of K where $D(K_{min})$ is the closure of $D(K)$ with respect to the graph norm;
- $K_{\max}$ the maximal extension of K where $D(K_{\max})$ is given by:

$$D(K_{\max}) = \{u \in L^2(\mathbb{R}^d \times \mathbb{R}^d) \,|\, Ku \in L^2(\mathbb{T}^d \times \mathbb{R}^d)\} \,.$$

We will use the notation $\mathbf{K}$ for the operator K_{min} or $\mathbf{K}_{B_e,V}$ if we want to mention the reference to B_e and V.

The existence of a strongly continuous semigroup associated to operator $\mathbf{K}$ when the magnetic field is regular, and $V = 0$ is shown in [9]. We will improve this result by considering a much lower regularity. In order to obtain the maximal accretivity, we are led to substitute the hypoellipticity argument by a regularity argument for the operators with coefficients in L^∞_{loc}, which will be combined with the more classical results of Rothschild–Stein in [12] for Hörmander operators of type 2 (see [6] for more details of this subject). Our first result is:

Theorem 1.1. — *If $B_e \in L^\infty_{loc}(\mathbb{R}^d, \mathbb{R}^{d(d-1)/2})$ and $V \in W^{1,\infty}_{loc}(\mathbb{R}^d)$, then $\mathbf{K}_{B_e,V}$ is maximally accretive.*

The theorem implies that the domain of the operator $\mathbf{K} = K_{min}$ has the following property:

$$D(\mathbf{K}) = D(K_{\max})\,. \tag{2}$$

We are next interested in specifying the domain of the operator $\mathbf{K}$ introduced in (2). For this goal, we will establish a maximal estimate for $\mathbf{K}$, using techniques that were initially developed for the study of hypoellipticity of invariant operators on nilpotent groups and the proof of the Rockland conjecture. Before we state our main result, we introduce the following functional spaces:

- $B^2(\mathbb{R}^d)$ (or B^2_v to indicate the name of the variables) denotes the space:

$$B^2(\mathbb{R}^d) := \{u \in L^2(\mathbb{R}^d)\,|\,\forall(\alpha,\beta) \in \mathbb{N}^{2d},\ |\alpha| + |\beta| \leq 2\,,\ v^\alpha\,\partial_v^\beta\,u \in L^2(\mathbb{R}^d)\}\,,$$

  which is equipped with its natural Hilbertian norm.
- $\tilde{B}^2(\mathbb{R}^d\times\mathbb{R}^d)$ is the space $L^2_x\widehat{\otimes}B^2_v$ (in $L^2(\mathbb{R}^d\times\mathbb{R}^d)$ identified with $L^2_x\widehat{\otimes}L^2_v$) with its natural Hilbert norm.
- $\mathcal{H}^2(\mathbb{R}^{2d})$ is the Sobolev space of degree 2 associated with the vector fields $\frac{\partial}{\partial v_j}$ $(j = 1,\dots,d)$, $i\,v_\ell$ $(\ell = 1,\dots,d)$ with weight 1 and $v\cdot\nabla_x$ with weight 2 as introduced in [10, Section 2]. It also reads

$$\mathcal{H}^2(\mathbb{R}^{2d}) = \{u \in \tilde{B}^2(\mathbb{R}^{2d})\,,\ v\cdot\nabla_x u \in L^2(\mathbb{R}^{2d})\}\,.$$

- $\mathcal{H}^2_{loc}(\mathbb{R}^{2d})$ is the space of functions that are locally in $\mathcal{H}^2(\mathbb{R}^{2d})$.

We can now state the second theorem of this article:

Theorem 1.2. — *Let $d = 2$ or 3. We assume that $B_e \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R}^{d(d-1)/2}) \cap L^\infty(\mathbb{R}^d, \mathbb{R}^{d(d-1)/2})$, $V \in C^2(\mathbb{R}^d, \mathbb{R})$ and that there exist positive constants C, $\rho_0 > \frac{1}{3}$, and $\gamma_0 < \frac{1}{3}$, such that*

$$|\nabla_x B_e(x)| \leq C\langle\nabla V(x)\rangle^{\gamma_0}\,, \tag{3}$$

$$|D^\alpha_x V(x)| \leq C\langle\nabla V(x)\rangle^{1-\rho_0}\,,\ \forall\alpha\ s.t.\ |\alpha| = 2\,, \tag{4}$$

where

$$\langle\nabla V(x)\rangle = \sqrt{|\nabla V(x)|^2 + 1}\,.$$

Then there exists $C_1 > 0$ such that, for all $u \in C_0^\infty(\mathbb{R}^d \times \mathbb{R}^d)$, the operator K satisfies the following maximal estimate:

$$\begin{aligned}(5)\qquad & \left\| |\nabla V(x)|^{\frac{2}{3}}\, u \right\| + \|(v\cdot\nabla_x - \nabla_x V\cdot\nabla_v - (v\wedge B_e)\cdot\nabla_v)u\| + \|u\|_{\tilde{B}^2} \\ & \qquad \leq C_1(\|Ku\| + \|u\|)\ .\end{aligned}$$

The proofs will combine the previous works of [2] (in the case $B_e = 0$) and [10] (in the case $V = 0$) with, in addition, two differences:

- $\mathbb{T}^d$ is replaced by $\mathbb{R}^d$.
- The reference operator in the enveloping algebra of the nilpotent algebra is different.

Notice also that when $B_e = 0$, our assumptions are weaker than in [2] where the property that $|\nabla V(x)|$ tends to $+\infty$ as $|x| \to +\infty$ was used to construct the partition of unity.

Using the density of $C_0^\infty(\mathbb{R}^d \times \mathbb{R}^d)$ in the domain of $\mathbf{K}$, we obtain the following characterization of this domain:

Corollary 1.3. —

$$\begin{aligned}(6)\qquad D(\mathbf{K}) = \big\{u \in \tilde{B}^2(\mathbb{R}^d\times\mathbb{R}^d)\ \big|\ & (v\cdot\nabla_x - \nabla V\cdot\nabla_v - (v\wedge B_e)\cdot\nabla_v)u \\ & \textit{and } |\nabla V(x)|^{\frac{2}{3}}\, u \in L^2(\mathbb{R}^d\times\mathbb{R}^d)\big\}.\end{aligned}$$

In particular, this implies that under the assumptions of Theorem 1.2, the operator $\mathbf{K}_{B_e,V}$ has a compact resolvent if and only if $\mathbf{K}_{B_e=0,V}$ has the same property. This is, in particular, the case (see [2]) when

$$|\nabla V(x)| \to +\infty \text{ when } |x| \to +\infty\ ,$$

as can also be seen directly from (6).

Remark 1.4. — For more results in the case without a magnetic field, we refer the reader to [2] and recent results obtained in 2018 by Wei-Xi Li [11] and in 2019 by M. Ben Said [1] in connection with a conjecture of Helffer and Nier relating the compact resolvent property for the (KFP)-operator with the same property for the Witten Laplacian on (0)-forms: $-\Delta_{x,v} + \frac{1}{4}|\nabla\Phi|^2 - \frac{1}{2}\Delta\Phi$, with $\Phi(x,v) = V(x) + \frac{v^2}{2}$. Its proof also involves nilpotent techniques. One can then naturally ask about results when $B_e(x)$ is unbounded. In particular, the existence of (KFP)-magnetic bottles (i.e., the compact resolvent property for the (KFP)-operator) when $V = 0$ is natural. Here, we simply observe that Proposition 5.19 in [2] holds (with exactly the same proof) for $\mathbf{K}_{B_e,V}$ when B_e and V are C^∞. Hence, there are no (KFP)-magnetic bottles.

2. Maximal accretivity for the Kramers–Fokker–Planck operator with a weakly regular electromagnetic field

To prove Theorem 1.1, we will show the Sobolev regularity associated with the following problem

$$K^* f = g \text{ with } f, g \in L^2_{loc}(\mathbb{R}^{2d}) \,,$$

where K^* is the formal adjoint of K:

$$K^* = -v\cdot\nabla_x - \Delta_v + (v\wedge B_e + \nabla_x V)\cdot\nabla_v + v^2/4 - d/2 \,. \tag{7}$$

The result of Sobolev regularity is the following:

Theorem 2.1. — *Let $d = 2$ or 3. We suppose that $B_e \in L^\infty_{loc}(\mathbb{R}^d, \mathbb{R}^{d(d-1)/2})$, and $V \in W^{1,\infty}_{loc}(\mathbb{R}^d\times\mathbb{R}^d)$. Then, for all $f \in L^2_{loc}(\mathbb{R}^{2d})$, such that $K^*f = g$ with $g \in L^2_{loc}(\mathbb{R}^{2d})$, $f \in \mathcal{H}^2_{loc}(\mathbb{R}^{2d})$.*

Before proving Theorem 2.1, we recall the following result :

Proposition 2.2 (Proposition A.3 in [10]). — *Let $c_j \in L^{\infty,2}_{loc}(\mathbb{R}^d\times\mathbb{R}^d)$, $j = 1,\dots,d$, where $L^{\infty,2}_{loc}(\mathbb{R}^d\times\mathbb{R}^d) = \{u \in L^2_{loc}, \forall\varphi\in C_0^\infty(\mathbb{R}^d\times\mathbb{R}^d)$ such that $\varphi u \in L^\infty_x(L^2_v)\}$, such that*

$$\partial_{v_j}\left(c_j(x,v)\right) = 0 \ \text{ in } \mathcal{D}'(\mathbb{R}^{2d})\,, \quad \forall j = 1,\dots,d \,. \tag{8}$$

Let P_0 be the Kolmogorov operator

$$P_0 := -v\cdot\nabla_x - \Delta_v \,. \tag{9}$$

If $h \in L^2_{loc}(\mathbb{R}^{2d})$ satisfies

$$\begin{cases} P_0 h = \displaystyle\sum_{j=1}^d c_j(x,v)\,\partial_{v_j} h_j + \tilde g \\ h_j, \tilde g \in L^2_{loc}(\mathbb{R}^{2d}), \ \forall j = 1,\dots,d \,, \end{cases} \tag{10}$$

then $\nabla_v h \in L^2_{loc}(\mathbb{R}^{2d}, \mathbb{R}^d)$.

We can now give the proof of Theorem 2.1.

Proof of Theorem 2.1. — The proof is similar to that of Theorem A.2 in [10]. In the following, we will only focus on the differences appearing in our case. To show the Sobolev regularity of the problem $K^*\,f = g$ with f and $g \in L^2_{loc}(\mathbb{R}^{2d})$

we can reformulate the problem as follows:

$$\begin{cases} P_0\, f = \displaystyle\sum_{j=1}^{d} c_j(x,v)\,\partial_{v_j}\,h_j + \tilde g \\ \quad h_j = f \in L^2_{loc}(\mathbb{R}^{2d})\ , \\ \quad \tilde g = g - \dfrac{v^2}{4}\,f + \dfrac{d}{2}\,f \in L^2_{loc}(\mathbb{R}^{2d}),\, \forall j = 1,\dots,d\ . \end{cases}$$

Here, the coefficients c_j are defined by

$$c_j(x,v) = -(v \wedge B_e)_j - \partial_{x_j} V \in L^\infty(\mathbb{R}^d, L^2_{loc}(\mathbb{R}^d))\,, \forall j = 1,\dots,d\ .$$

We note that the coefficients c_j verify the condition (8) of the Proposition 2.2 because

$$\partial_{v_j}(v \wedge B_e)_j = 0 \ \text{ and } \ \partial_{v_j}\,\partial_{x_j}\,V = 0\,, \quad \forall j = 1,\dots,d\ . \qquad \square$$

2.1. Proof of Theorem 1.1. — The accretivity of the operator K is clear. To show that the operator $\mathbf{K}$ is maximally accretive, it suffices to show that there exists $\lambda_0 > 0$, such that the operator $T = K + \lambda_0\, Id$ has a dense range in $L^2(\mathbb{R}^d \times \mathbb{R}^d)$. As in [2], we take $\lambda_0 = \frac{d}{2} + 1$.

We have to prove that if $u \in L^2(\mathbb{R}^d \times \mathbb{R}^d)$ satisfies

$$\langle u, (K + \lambda_0 Id)\, w\rangle = 0, \quad \forall w \in D(K)\ , \tag{11}$$

then $u = 0$.

For this we observe that equation (11) implies that

$$K^*\, u = -\left(\frac{d}{2} + 1\right) u \quad \text{in } \mathcal{D}'(\mathbb{R}^d \times \mathbb{R}^d)\ ,$$

where K^* is the operator defined in (7).

Under the assumption that $B_e \in L^\infty_{loc}(\mathbb{R}^d, \mathbb{R}^{d(d-1)/2}), V \in W^{1,\infty}_{loc}(\mathbb{R}^d)$ and $u \in D(K^*) \subset L^2_{loc}(\mathbb{R}^d \times \mathbb{R}^d)$, Theorem 2.1 shows that $u \in \mathcal{H}^2_{loc}(\mathbb{R}^d \times \mathbb{R}^d)$. So, we have $\chi(x,v)u \in \mathcal{H}^2(\mathbb{R}^d \times \mathbb{R}^d)$, for any $\chi \in C_0^\infty(\mathbb{R}^d \times \mathbb{R}^d)$. The rest of the proof is standard. The regularity obtained for u allows us to justify the integrations by parts and the cut-off argument given in [2, Proposition 5.5].

Remark 2.3. — One can also prove that the operator $K = K_{V,B_e}$ is maximally accretive by using the Kato perturbation theory by an unbounded operator, but this can only be done under the stronger assumption that $B_e \in L^\infty(\mathbb{R}^d, \mathbb{R}^{d(d-1)/2})$ and $V \in W^{1,\infty}(\mathbb{R}^d)$. Under this assumption, one can prove that $\mathcal{B} = -\nabla_x V \cdot \nabla_v$ is a K_{B_e}-bounded operator with a K_{B_e}-bound that is strictly smaller than 1 (cf. [8, Section A.4]), where K_{B_e} is the same operator, with $V = 0$ already having been studied in [8].

3. Proof of Theorem 1.2

3.1. General strategy. — For technical reasons, it is easier to work with

$$\check{K} := K + \frac{d}{2}\,.$$

It is clear that it is equivalent to prove the maximal estimate for $\check{K}$.

The proof consists in constructing $\mathcal{G}$, a graded and stratified algebra of type 2, and, at any point $x \in \mathbb{R}^d$, a homogeneous element $\mathcal{F}_x$ in the enveloping algebra $\mathcal{U}_2(\mathcal{G})$, which satisfies the Rockland condition. We recall that Helffer and Nourrigat's proof is based on maximal estimates that not only hold for the operator but also (and uniformly) for $\pi(\mathcal{F}_x)$, where π is any induced representation of the Lie algebra.

It remains to find π_x, such that $\pi_x(\mathcal{F}_x) = \mathcal{K}_x + \frac{d}{2}$ is a good approximation of $\mathbf{K}$ in a suitable ball centered at x and to patch together the estimates through a partition of unity. This strategy was used in [2] in the case $B = 0$ and in [8, 9] in the case of the torus $\mathbb{T}^d$ with $V = 0$.

Actually, we first define $\mathcal{K}_x$ and then look for the Lie algebra, the operator, and the induced representation.

3.2. Maximal estimate. — For simplification purposes, let us present the approach when $d = 2$ (but this restriction is not important). The dependence on x will actually appear through the two parameters $(b, w) \in \mathbb{R} \times \mathbb{R}^2$, where

$$b = B_e(x) \text{ and } w = \nabla V(x)\,.$$

We then consider the following model:

$$\begin{aligned}(12)\quad \mathcal{K}_x + 1 &:= K_{w,b}\\ &= v \cdot \nabla_x - w.\nabla_v + b(v_1\partial_{v_2} - v_2\partial_{v_1}) - \Delta_v + v^2/4\,, \quad \text{in } \mathbb{R}^2 \times \mathbb{R}^2\,.\end{aligned}$$

We would like to have uniform estimates with respect to the parameters b and w. This is the object of the following proposition.

Proposition 3.1. — *For any compact interval I, there exists a constant C, such that for any $b \in I$, any $w \in \mathbb{R}^2$, and any $f \in \mathcal{S}(\mathbb{R}^4)$, we have the maximal estimate*

$$\begin{aligned}(13)\qquad |w|^{\frac{4}{3}}\|f\|^2 + |w|^{\frac{2}{3}}\|\nabla_v f\|^2 + |w|^{\frac{2}{3}}\|vf\|^2 + \sum_{|\alpha|+|\beta|\leq 2}\|v^\alpha \partial_v^\beta f\|^2 \\ \leq C\left(\|f\|^2 + \|K_{w,b}f\|^2\right)\,.\end{aligned}$$

3.3. Proof of Proposition 3.1. Step 1. — Following [2], we consider, after application of the $x \to \xi$ partial Fourier transform, the family indexed by w, b, ξ:

$$(14)\quad \hat{K}_{w,b,\xi} = iv \cdot \xi - w.\nabla_v + b(v_1\partial_{v_2} - v_2\partial_{v_1}) - \Delta_v + v^2/4, \quad \text{in } \mathbb{R}^2 \times \mathbb{R}^2\,.$$

We denote by σ the symbol of the operator $\hat{K}_{w,b,\xi}$ considered as acting in $L^2(\mathbb{R}^2_v)$:

$$\sigma(v,\eta) = i\xi \cdot v - iw \cdot \eta - ib(v_1\eta_2 - v_2\eta_1) + \eta^2 + v^2/4 \,.$$

We introduce the symplectic map on $\mathbb{R}^4$ associated with the matrix

$$A = \begin{pmatrix} \cos t_1 & -\sin t_1 & 0 & 0 \\ \sin t_1 & \cos t_1 & 0 & 0 \\ 0 & 0 & \cos t_2 & -\sin t_2 \\ 0 & 0 & \sin t_2 & \cos t_2 \end{pmatrix} .$$

Then there exists (see [2] for an explicit definition) an associate unitary metaplectic operator $\tilde{T}$ acting in $L^2(\mathbb{R}^4)$, such that

$$\begin{aligned} \tilde{T}^{-1}\,\hat{K}_{w,b,\xi}\,\tilde{T} = iv \cdot \xi'(t) - w'(t) \cdot \nabla_v + b'_1(t)(v_1\partial_{v_2} - v_2\partial_{v_1}) \\ + ib'_2(t)(v_1v_2 - \partial_{v_1}\partial_{v_2}) - \Delta_v + v^2/4 \,, \end{aligned} \tag{15}$$

with $t = (t_1, t_2) \in \mathbb{R}^2$

$$\begin{cases} w'_k(t) = w_k \cos t_k - \xi_k \sin t_k \\ \xi'_k(t) = w_k \sin t_k + \xi_k \cos t_k \\ b'_1(t) = b\cos(t_2 - t_1) \\ b'_2(t) = b\sin(t_2 - t_1) \end{cases} , \tag{16}$$

for $k = 1, 2$.

We now choose t_k so that

$$w'_k(t) = 0 \quad \text{and} \quad \xi'_k(t) = \sqrt{w_k^2 + \xi_k^2} := \rho_k \,.$$

With this choice of t_k, by (15) this leads to the analysis of a maximal estimate for

$$\check{K}_{\rho,b'} = i\,v \cdot \rho + b'_1(v_1\partial_{v_2} - v_2\partial_{v_1}) + ib'_2(v_1v_2 - \partial_{v_1}\partial_{v_2}) - \Delta_v + v^2/4 \,, \tag{17}$$

which is considered, for fixed (ρ, b'), as an operator on $\mathcal{S}(\mathbb{R}^2)$.

We notice that $|b'| = |b|$. Hence, if b belongs to a compact interval, then b' belongs to a compact set in $\mathbb{R}^2$.

It remains to show that for a suitable Lie algebra $\mathcal{G}$, this is the image by an induced representation π_ρ of an element $\mathcal{F}_{b'}$ satisfying the Rockland condition.

3.4. Nilpotent techniques for the analysis of $\check{K}_{\rho,b'}$. We construct a graded Lie algebra $\mathcal{G}$ of type 2, a subalgebra $\mathcal{H}$, for $b' \in \mathbb{R}^2$ an element $\mathcal{F}_{b'}$ in $\mathcal{U}_2(\mathcal{G})$, and for $\rho \in \mathbb{R}^2$ a linear form $\ell_\rho \in \mathcal{G}^*$, such that $\ell_\rho([\mathcal{H},\mathcal{H}|) = 0$, and

$$\pi_{\ell_\rho,\mathcal{H}}(\mathcal{F}_{b'}) = \check{K}_{\rho,b'} \;;$$

$\check{K}_{\rho,b'}$ can be written as a ρ-independent polynomial of five differential operators

$$\begin{aligned}\check{K}_{\rho,b'} = X_{1,2} - \sum_{k=1}^{2}\left((X'_{k,1})^2 + \frac{1}{4}(X''_{k,1})^2)\right) \\ - ib'_1\left(X'_{1,1}X''_{2,1} - X'_{2,1}X''_{1,1}\right) - ib'_2\left(X''_{1,1}X''_{2,1} + X'_{1,1}X'_{2,1}\right)\,,\end{aligned} \tag{18}$$

which are given by

$$X'_{1,1} = \partial_{v_1}\,,\quad X''_{1,1} = iv_1\,,\quad X'_{2,1} = \partial_{v_2}\,,\quad X''_{2,1} = iv_2\,,\, X_{1,2} = i\,v\cdot\rho\,. \tag{19}$$

We now look at the Lie algebra generated by these five operators and their brackets. This leads us to introduce three new elements that verify the following relations:

$$\begin{aligned}&X_{2,2} := [X'_{1,1}, X''_{1,1}] = [X'_{2,1}, X''_{2,1}] = i\,,\\ &X_{1,3} := [X_{1,2}, X'_{1,1}] = -i\rho_1\,,\, X_{2,3} := [X_{1,2}, X'_{2,1}] = -i\rho_2\,.\end{aligned}$$

We also observe that we have the following properties:

$$\begin{aligned}&[X'_{1,1}, X'_{2,1}] = [X''_{1,1}, X''_{2,1}] = 0\,,\\ &[X'_{j,1}, X_{k,3}] = [X''_{j,1}, X_{k,3}] = [X_{k,3}, X_{2,2}] = \cdots = 0\,,\quad \forall j,k = 1,2\,.\end{aligned}$$

We then construct a graded Lie algebra $\mathcal{G}$ verifying the same commutator relations. More precisely, $\mathcal{G}$ is stratified of type 2, nilpotent of rank 3, its underlying vector space is $\mathbb{R}^8$, and $\mathcal{G}_1$ is generated by four elements $Y'_{1,1}, Y'_{2,1}, Y''_{1,1}$, and $Y''_{2,1}$, $\mathcal{G}_2$ is generated by $Y_{1,2}$ and $Y_{2,2}$, and $\mathcal{G}_3$ is generated by $Y_{1,3}$ and $Y_{2,3}$.

$$\begin{aligned}&Y_{2,2} := [Y'_{1,1}, Y''_{1,1}] = [Y'_{2,1}, Y''_{2,1}],\\ &Y_{1,3} := [Y_{1,2}, Y'_{1,1}]\,,\, Y_{2,3} := [Y_{1,2}, Y'_{2,1}]\,,\end{aligned}$$

and

$$\begin{aligned}&[Y'_{1,1}, Y'_{2,1}] = [Y''_{1,1}, Y''_{2,1}] = 0\,,\\ &[Y'_{j,1}, Y_{k,3}] = [Y''_{j,1}, Y_{k,3}] = [Y_{k,3}, Y_{2,2}] = \cdots = 0\qquad,\quad \forall j,k = 1,2\,.\end{aligned}$$

We note that $\mathcal{G}$ is the same graded Lie algebra given in [10] and is, indeed, independent of the parameters (ρ, b').

We have to check that for a given $\rho = (\rho_1, \rho_2)$, the representation π (with the convention that if $\diamond = \emptyset$, there is no exponent) defined on its basis by

$$\pi(Y^{\diamond}_{i,j}) = X^{\diamond}_{i,j}\quad\text{with } i = 1,2,\ j = 1,2,3 \text{ and } \diamond \in \{\emptyset,\ \prime, \prime\prime\} \tag{20}$$

defines an induced representation of the Lie algebra $\mathcal{G}$.

By applying the same steps given in [10, page 11] (following the techniques of [4]), we actually obtain $\pi = \pi_{\ell,\mathcal{H}}$ with

$$\mathcal{H} = \mathrm{Vect}(Y''_{1,1}, Y''_{2,1}, Y_{1,2}, Y_{2,2}, Y_{1,3}, Y_{2,3})\,, \tag{21}$$

and $\ell_\rho \in \mathcal{G}^*$ defined by 0 for the elements of the basis of $\mathcal{G}$ except

$$\ell_\rho(Y_{1,3}) = -\rho_1\,,\, \ell_\rho(Y_{2,3}) = -\rho_2\quad\text{and}\quad \ell_\rho(Y_{2,2}) = 1\,. \tag{22}$$

With (18) in mind, we introduce

$$\begin{aligned}\mathcal{F}_{b'} = Y_{1,2} - \sum_{k=1}^{2}\left((Y'_{k,1})^2 + \frac{1}{4}(Y''_{k,1})^2\right)\\ - ib'_1\left(Y'_{1,1}Y''_{2,1} - Y'_{2,1}Y''_{1,1}\right) - ib'_2\left(Y''_{1,1}Y''_{2,1} + Y'_{1,1}Y'_{2,1}\right)\end{aligned}\tag{23}$$

and get

$$\pi_{\ell_\rho,\mathcal{H}}(\mathcal{F}_{b'}) = \check{K}_{\rho,b'}\ .\tag{24}$$

The verification of the Rockland condition is the same as in [10]. For any nontrivial irreducible representation π, we consider a C^∞-vector of the representation, such that $\pi(\mathcal{F}_{b'})u = 0$ and write $\Re\langle\pi(\mathcal{F}_{b'})u\,,\,u\rangle = 0$. Because the operator $\pi(Y)$ is a formally skew-adjoint operator for all $Y \in \mathcal{G}$ (see [10, Proposition 2.7]), we first get that $\pi(Y)u = 0$, for any $Y \in \mathcal{G}_1$ and by difference $\pi(Y_{1,2})u = 0$.

Observing that $\mathcal{G}$ is the algebra generated by $\mathcal{G}_1$ and Y_{12}, we get $\pi(Y)u = 0$, for any $Y \in \mathcal{G}$, which implies, π being nontrivial, that $u = 0$.

Therefore, according to the Helffer–Nourrigat theorem [3] the operator $\mathcal{F}_{b'}$ is maximal hypoelliptic, and this also implies that $\pi(\mathcal{F}_{b'})$ satisfies a maximal estimate for any induced representation π with a constant that is independent of π. By applying this argument to $\check{K}_{\rho,b'} = \pi_{\ell,\mathcal{H}}(\mathcal{F}_{b'})$, we obtain for any compact $K_0 \subset \mathbb{R}^2$ the existence of $C > 0$, such that, $\forall b' \in K_0$, $\forall \rho \in \mathbb{R}^2$ and $\forall u \in \mathcal{S}(\mathbb{R}^2)$,

$$\begin{aligned}\|X_{1,2}u\|^2 + \sum_{k=1}^{2}\left(\|(X'_{k,1})^2u\|^2 + \|(X''_{k,1})^2u\|^2\right) + \sum_{k,\ell=1}^{2}\|X'_{k,1}X''_{\ell,1}u\|^2\\ \le C\left(\|\check{K}_{\rho,b'}u\|^2 + \|u\|^2\right)\ .\end{aligned}\tag{25}$$

Notice here that we first prove the inequality for fixed b' and then show the local uniformity with respect to b'.

In particular, we have

$$\|(v\cdot\rho)\,u\|^2 + \sum_{|\alpha|+|\beta|\le 2}\|v^\alpha\partial_v^\beta u\|^2 \le C\left(\|\check{K}_{\rho,b'}u\|^2 + \|u\|^2\right)\ .\tag{26}$$

Using the treatment of the operator introduced in (V5.52) in [2], there exists $\check{C}$, such that, $\forall \rho \in \mathbb{R}^2$ and $\forall u \in \mathcal{S}(\mathbb{R}^2)$, we have

$$|\rho|^{\frac{4}{3}}\|u\|^2 \le \check{C}(\|(-\Delta + v\cdot\rho)u\|^2)\ .\tag{27}$$

Combining (26) and (27), we finally obtain, for any compact $K_0 \subset \mathbb{R}^2$, the existence of $\hat{C} > 0$, such that $\forall b' \in K_0$, $\forall \rho \in \mathbb{R}^2$, and $\forall u \in \mathcal{S}(\mathbb{R}^2)$,

$$\|(v\cdot\rho)\,u\|^2 + |\rho|^{\frac{4}{3}}\|u\|^2 + \sum_{|\alpha|+|\beta|\le 2}\|v^\alpha\partial_v^\beta u\|^2 \le \hat{C}\left(\|\check{K}_{\rho,b'}u\|^2 + \|u\|^2\right)\ .\tag{28}$$

3.5. End of the proof of Proposition 3.1. — Coming back to the initial coordinates, we get for a new constant $C > 0$,

$$\|f\|^2 + \|\hat{K}_{w,b,\xi} f\|^2 \geq \frac{1}{C}\Big(|w|^{\frac{4}{3}}\|f\|^2 + \sum_{|\alpha|+|\beta|\leq 2} \|v^\alpha \partial_v^\beta f\|^2\Big),\ \forall f \in \mathcal{S}(\mathbb{R}^2)\ . \tag{29}$$

Note that by complex interpolation, this implies also for $f \in \mathcal{S}(\mathbb{R}^2)$:

$$\|f\|^2 + \|\hat{K}_{w,b,\xi} f\|^2 \geq \frac{1}{\tilde{C}}\Big(|w|^{\frac{2}{3}}\|vf\|^2 + |w|^{\frac{2}{3}}\|\nabla_v f\|^2 + \sum_{|\alpha|+|\beta|\leq 2} \|v^\alpha \partial_v^\beta f\|^2\Big). \tag{30}$$

This achieves the proof of the proposition.

3.6. Proof of Theorem 1.2. —

3.6.1. *Step 1: construction of the partition of unity.* — We start with the following lemma.

Lemma 3.2. — *Let V satisfy (4), for some $\rho_0 \in (\frac{1}{3}, 1)$ and, for $s > 0$, let us define*

$$r(x) := \langle \nabla_x V(x)\rangle^{-s}\ . \tag{31}$$

Then there exists $\delta_0 > 0$, such that if $|x-y| \leq \delta_0 r(x)$, then

$$\delta_0 \leq \frac{r(x)}{r(y)} \leq \frac{1}{\delta_0}\ .$$

Proof. — For $c > 0$, let $\hat{v}(x,c) := \sup_{|x-y|\leq cr(x)} \langle \nabla V(y)\rangle$. Using our assumptions, there exists C_0, such that, for any $c > 0$, we have

$$\hat{v}(x,c) \leq \langle \nabla V(x)\rangle + C_0\, c\, r(x)\hat{v}(x,c)^{1-\rho_0}\ .$$

Hence,

$$\begin{aligned}\frac{\hat{v}(x,c)}{\langle \nabla V(x)\rangle} &\leq 1 + C_0\, c\, \langle \nabla V(x)\rangle^{-1-s}\hat{v}(x,c)^{1-\rho_0}\\ &\leq 1 + C_0\, c\, \big(\frac{\hat{v}(x,c)}{\langle \nabla V(x)\rangle}\big)^{1-\rho_0}\ .\end{aligned}$$

This inequality implies, for all $c > 0$, the existence of $C_1(c)$, such that, $\forall x \in \mathbb{R}^2$,

$$1 \leq \frac{\hat{v}(x,c)}{\langle \nabla V(x)\rangle} \leq C_1(c)\,,$$

and, in particular,

$$\frac{\langle \nabla V(y)\rangle}{\langle \nabla V(x)\rangle} \leq C_1(c)\,,\ \text{for } |x-y| \leq cr(x)\ .$$

To get the reverse inequality, we fix some c_0 and consider $c \in (0, c_0)$. We then get, for $|x-y| \leq cr(x)|$,

$$\begin{aligned}\langle \nabla V(x)\rangle &\leq \langle \nabla V(y)\rangle + C_0\, c\, |\nabla V(x)|^{-s}\, C_1(c_0)\, |\nabla V(x)|^{1-\rho_0}\\ &\leq \langle \nabla V(y)\rangle + C_0\, c\, C_1(c_0)\, |\nabla V(x)|\ .\end{aligned}$$

Choosing c small enough we obtain

$$\langle \nabla V(x)\rangle \leq (1 - C_0\, c\, C_1(c_0))^{-1} |\nabla V(y)|\ .$$

Hence, we have found $c_1 \in (0, c_0)$ and $C_2 > 1$, such that $|x-y| \leq c_1 r(x)|$,

$$\frac{1}{C_2} \leq \frac{\langle \nabla V(y)\rangle}{\langle \nabla V(x)\rangle} \leq C_2\ . \tag{32}$$

It is then easy to get the lemma for $\delta_0 \in (0, c_1)$ small enough. □

The parameter δ_0 is now fixed by the lemma, and we now consider $\delta \in (0, \delta_0)$ and $r(x, \delta) = \delta\, r(x)$. According to Lemma 18.4.4 (and around this lemma in Section 8.4) in [7], one can now introduce, for any $\delta \in (0, \delta_0]$, a δ-dependent partition of unity ϕ_j in the x variable corresponding to a covering by balls $B(x_j, r(x_j, \delta))$ (with the property of uniform finite intersection N_δ), where, for $r > 0$ and $\hat{x} \in \mathbb{R}^2$, the ball $B(\hat{x}, r)$ is defined by $B(\hat{x}, r) := \{x \in \mathbb{R}^2 \mid |x - \hat{x}| < r\}$.

So the support of each ϕ_j is contained in $B(x_j, \delta r(x_j))$, and we have

$$\sum_j \phi_j^2(x) = 1\ , \tag{33}$$

and

$$|\nabla \phi_j(x)| \leq C_\delta \langle \nabla V(x_j)\rangle^s \leq \hat{C}_\delta \langle \nabla V(x)\rangle^s\ . \tag{34}$$

This implies (using the finite intersection property)

$$\sum_j |\nabla \phi_j(x)|^2 \leq \check{C}_\delta \langle \nabla V(x)\rangle^{2s}\ . \tag{35}$$

3.6.2. *Step 2.* — The proof is inspired by Chapter 9 of [2] but is different due to the presence of the magnetic field and the absence of the assumption $|\nabla V(x)| \to +\infty$ at infinity.

We start, for $u \in C_0^\infty(\mathbb{R}^2 \times \mathbb{R}^2)$, from

$$\begin{aligned}
\|\check{K}u\|^2 &= \sum_j \|\phi_j \check{K}u\|^2 \\
&= \sum_j \|\check{K}\phi_j u\|^2 - \sum_j \|[\check{K}, \phi_j]u\|^2 \\
&= \sum_j \|\check{K}\phi_j u\|^2 - \sum_j \|(X_0\phi_j)u\|^2 \\
&= \sum_j \|\check{K}\phi_j u\|^2 - \sum_j \|(\nabla\phi_j)\cdot v\, u\|^2 \\
&\geq \sum_j \|\check{K}\phi_j u\|^2 - \check{C}_\delta \|\langle \nabla V\rangle^s v u\|^2 \,.
\end{aligned}$$

For the analysis of $\|\check{K}\phi_j u\|^2$, let us now write, with $w_j = \nabla V(x_j)$ and $b_j = B_e(x_j)$,

$$\check{K} = \check{K} - K_{w_j,b_j} + K_{w_j,b_j} \,.$$

We verify that, by using the construction of the partition of unity and the assumptions of Theorem 1.2,

$$\begin{aligned}
&\|(\check{K} - K_{w_j,b_j})\phi_j u\|^2 \\
&\quad \leq 2\|\phi_j(x)(\nabla V(x) - w_j)\cdot\nabla_v u\|^2 + 2\|\phi_j(x)(B_e(x) - b_j)(v_1\partial_{v_2} - v_2\partial_{v_1})u\|^2 \\
&\quad \leq C\delta^2\left(\left\||\nabla V|^{1-\rho_0-s}\phi_j\nabla_v u\right\|^2 + \left\||\nabla V|^{-s+\gamma_0}\phi_j(v_1\partial_{v_2} - v_2\partial_{v_1})u\right\|^2\right).
\end{aligned}$$

These errors have to be controlled by the main term.

We note that by using the inequality (30), we have

$$\begin{aligned}
\|\phi_j u\|^2 + \|K_{w_j,b_j}\phi_j u\|^2 \geq{}& \frac{1}{C}\left\||\nabla V|^{\frac{2}{3}}\phi_j u\right\|^2 + \frac{1}{C}\left\||\nabla V|^{\frac{1}{3}}\phi_j\nabla_v u\right\|^2 \\
&+ \frac{1}{C}\left\||\nabla V|^{\frac{1}{3}}\phi_j v u\right\|^2 + \frac{1}{C}\sum_{|\alpha|+|\beta|\leq 2}\|v^\alpha\partial_v^\beta\phi_j u\|^2 \,.
\end{aligned}$$

Finally, we observe that

$$\|\check{K}\phi_j u\|^2 \geq \frac{1}{2}\|K_{w_j,b_j}\phi_j u\|^2 - \|(\check{K} - K_{w_j,b_j})\phi_j u\|^2 \,.$$

We now choose

$$\max\left(\frac{2}{3} - \rho_0, \gamma_0\right) < s < \frac{1}{3} \,. \tag{36}$$

Summing up over j, we have obtained the existence of a constant C, such that for all $u \in C_0^\infty(\mathbb{R}^2 \times \mathbb{R}^2)$

$$
\begin{aligned}
(37)\quad & \|u\|^2 + \|\check{K}u\|^2 \\
& \geq \frac{1}{C}\left\||\nabla V|^{\frac{2}{3}}\, u\right\|^2 + \frac{1}{C}\left\||\nabla V|^{\frac{1}{3}}\, \nabla_v u\right\|^2 + \frac{1}{C}\left\||\nabla V|^{\frac{1}{3}}\, vu\right\|^2 \\
& \quad + \frac{1}{C}\sum_{|\alpha|+|\beta|\leq 2} \|v^\alpha \partial_v^\beta u\|^2 \\
& \quad - C\delta^2 \left\||\nabla V|^{1-s-\rho_0}\, \nabla_v u\right\|^2 - C\delta^2\left\||\nabla V|^{-s+\gamma_0}(v_1\partial_{v_2} - v_2\partial_{v_1})u\right\|^2 \\
& \quad - C\left(\int_{\mathbb{R}^d}\int_{\Omega(R)} |\nabla V|^{2s}\, |\nabla_v u|^2\, dxdv + \int_{\mathbb{R}^d}\int_{\Omega(R)^c} |\nabla V|^{2s}\, |\nabla_v u|^2\, dxdv\right) \\
& \quad - \hat{C}\left(\int_{\mathbb{R}^d}\int_{\Omega(R)} |\nabla V|^{2s}\, |vu|^2\, dxdv + \int_{\mathbb{R}^d}\int_{\Omega(R)^c} |\nabla V|^{2s}\, |vu|^2\, dxdv\right) \\
& \quad - C\delta^2(\|vu\|^2 + \|u\|^2 + \|\nabla_v u\|^2)\ ,
\end{aligned}
$$

where $\Omega(R)$ is defined by $\Omega(R) = \{x \in \mathbb{R}^d \,|\, |\nabla V(x)| < R\}$, and $\Omega(R)^c := \mathbb{R}^d \setminus \Omega(R)$ is the complement of $\Omega(R)$ in $\mathbb{R}^d$.

Using the definition of $\Omega(R)$ and the assumption that $s < \frac{1}{3}$ we obtain

$$
\begin{aligned}
\int_{\mathbb{R}^d}\int_{\Omega(R)} |\nabla V|^{2s} |\nabla_v u|^2 dxdv &\leq R^{2s} \int_{\mathbb{R}^d}\int_{\Omega(R)} |\nabla_v u|^2 dxdv, \\
\int_{\mathbb{R}^d}\int_{\Omega(R)^c} |\nabla V|^{2s} |\nabla_v u|^2 dxdv &\leq R^{2s-\frac{2}{3}} \int_{\mathbb{R}^d}\int_{\Omega(R)^c} |\nabla V|^{\frac{2}{3}} |\nabla_v u|^2 dxdv\ ,
\end{aligned}
$$

and

$$
\begin{aligned}
\int_{\mathbb{R}^d}\int_{\Omega(R)} |\nabla V|^{2s} |vu|^2 dxdv &\leq R^{2s} \int_{\mathbb{R}^d}\int_{\Omega(R)} |vu|^2 dxdv, \\
\int_{\mathbb{R}^d}\int_{\Omega(R)^c} |\nabla V|^{2s} |vu|^2 dxdv &\leq R^{2s-\frac{2}{3}} \int_{\mathbb{R}^d}\int_{\Omega(R)^c} |\nabla V|^{\frac{2}{3}} |vu|^2 dxdv\ .
\end{aligned}
$$

Similarly, for $\left\||\nabla V|^{1-s-\rho_0} \nabla_v u\right\|^2$, we have, for $1 > s + \rho_0 > \frac{2}{3}$,

$$
\begin{aligned}
\int_{\mathbb{R}^d}\int_{\Omega(R)} |\nabla V|^{2-2s-2\rho_0} |\nabla_v u|^2 dxdv &\leq R^{2-2s-2\rho_0} \int_{\mathbb{R}^d}\int_{\Omega(R)} |\nabla_v u|^2 dxdv, \\
\int_{\mathbb{R}^d}\int_{\Omega(R)^c} |\nabla V|^{2-2s-2\rho_0} |\nabla_v u|^2 dxdv &\leq R^{\frac{4}{3}-2s-2\rho_0} \int_{\mathbb{R}^d}\int_{\Omega(R)^c} |\nabla V|^{\frac{2}{3}} |\nabla_v u|^2 dxdv\ ,
\end{aligned}
$$

By combining the previous inequalities in (37), we obtain for a possibly new larger $C > 0$ and a (δ, R)-dependent constant

$$\begin{aligned}
&\|u\|^2 + \|\check{K}u\|^2 \\
&\quad \geq \frac{1}{C}\left\||\nabla V|^{\frac{2}{3}}\, u\right\|^2 \\
&\qquad + \left(\frac{1}{C} - \check{C}_\delta(R^{2s-2/3} + R^{\frac{4}{3}-2s-2\rho_0})\right)\left(\left\||\nabla V|^{\frac{1}{3}}\,\nabla_v u\right\|^2 + \left\||\nabla V|^{\frac{1}{3}}\, vu\right\|^2\right) \\
&\qquad + \left(\frac{1}{C} - C\delta^2\right)\|(v_1\partial_{v_2} - v_2\partial_{v_1})u\|^2 \\
&\qquad - C_{\delta,R}(\|\nabla_v u\|^2 + \|vu\|^2 + \|u\|^2)\,.
\end{aligned}$$

We can achieve the proof by observing our conditions on s, ρ_0, and γ_0 and by choosing fist δ small enough and then R large enough. With this choice of δ and R, we obtain the existence of a constant $C > 0$, such that

$$\begin{aligned}
&\|u\|^2 + \|\check{K}u\|^2 \\
&\quad \geq \frac{1}{C}\left(\left\||\nabla V|^{\frac{2}{3}}\, u\right\|^2 + \left\||\nabla V|^{\frac{1}{3}}\,\nabla_v u\right\|^2 + \left\||\nabla V|^{\frac{1}{3}}\, vu\right\|^2 + \|(v_1\partial_{v_2} - v_2\partial_{v_1})u\|^2\right) \\
&\qquad - C(\|\nabla_v u\|^2 + \|vu\|^2 + \|u\|^2)\,.
\end{aligned}$$

Combining this inequality with the standard inequality

$$\Re\langle \check{K}u, u\rangle \geq \|\nabla_v u\|^2 + \|vu\|^2\,, \quad \forall u \in C_0^\infty(\mathbb{R}^{2d})\,,$$

we can achieve the proof of the theorem, keeping in mind that the maximal estimate for $\check{K}$ is equivalent to the maximal estimate for K.

BIBLIOGRAPHY

[1] M. Ben Said – "Kramers–Fokker–Planck operators with homogeneous potentials", *Mathematical Methods in the Applied Sciences* **45** (2022), no. 2, p. 914–927.

[2] B. Helffer & F. Nier – *Hypoelliptic estimates and spectral theory for Fokker–Planck operators and Witten Laplacians*, Lecture Notes in Mathematics, no. 1862, Springer, 2005.

[3] B. Helffer & J. Nourrigat – "Hypoellipticité pour des groupes nilpotents de rang de nilpotence 3", *Communications in Partial Differential Equations* **3** (1978), no. 8, p. 643–743.

[4] B. Helffer & J. Nourrigat – *Hypoellipticité maximale pour des opérateurs polynômes de champs de vecteurs*, Progress in Mathematics, no. 58, Birkhäuser, 1985.

[5] F. Hérau & F. Nier – "Isotropic hypoellipticity and trend to equilibrium for the Fokker–Planck equation with a high-degree potential", *Archive for Rational Mechanics and Analysis* **171** (2004), no. 2, p. 151–218.

[6] L. Hörmander – "Hypoelliptic second order differential equations", *Acta Mathematica* **119** (1967), no. 1, p. 147–171.

[7] ———, *The analysis of linear partial differential equations. Vol. III*, Springer, 1985.

[8] Z. Karaki – "Équations cinétiques avec champ magnétique", Ph.D. thesis, Université de Nantes, 2019.

[9] ———, "Trend to the equilibrium for the Fokker–Planck system with an external magnetic field", *Kinetic & Related Models* **13** (2020), no. 2, p. 309.

[10] ———, "Maximal estimates for the Fokker–Planck operator with magnetic field", *J. Spectr. Theory* **11** (2021), no. 3, p. 1179–1213.

[11] W.-X. Li – "Compactness of the resolvent for the Witten Laplacian", *Annales Henri Poincaré* **19** (2018), p. 1259–1282.

[12] L. Rothschild – "A criterion for hypoellipticity of operators constructed from vector fields", *Communications in Partial Differential Equations* **4** (1979), no. 6, p. 645–699.

Bull. Soc. Math. France
150 (1), 2022, p. 17-52

SINGULAR VECTORS AND GEOMETRY AT INFINITY OF PRODUCTS OF HYPERBOLIC SPACES

by Toshiaki Hattori

Abstract. — Let $\mathbf{k}$ be a number field and $\mathbf{k}_M$ the Minkowski space associated to $\mathbf{k}$. Dirichlet's theorem in Diophantine approximation is generalized to the case of approximations of vectors in $\mathbf{k}_M$ by elements of $\mathbf{k}$. We study the set of singular elements of $\mathbf{k}_M$ in this setting and calculate its Hausdorff dimension, by relating the inequalities to Tits geometry of the geometric boundary of the symmetric space naturally associated to $\mathbf{k}$.

Résumé (*Vecteurs singuliers et géométrie à l'infini des produits d'espaces hyperboliques*). — Soient $\mathbf{k}$ un corps de nombres algébriques et $\mathbf{k}_M$ l'espace de Minkowski associé à $\mathbf{k}$. Le théorème de Dirichlet en approximation diophantienne se généralise au cas de l'approximation de vecteurs dans $\mathbf{k}_M$ par des éléments de $\mathbf{k}$. Nous étudions l'ensemble des éléments singuliers de $\mathbf{k}_M$ dans ce cadre et nous calculons sa dimension de Hausdorff, en reliant les inégalités définissant les vecteurs singuliers à la géométrie de Tits du bord géométrique de l'espace symétrique naturellement associé à $\mathbf{k}$.

1. Introduction

In the classical theory of Diophantine approximation, a real number x is said to be badly approximable if there exists a positive constant C such that

$$|q|\,|qx-p| \geq C$$

Texte reçu le 24 mai 2021, modifié le 15 septembre 2021, accepté le 22 septembre 2021.

Toshiaki Hattori, Department of Mathematics, Tokyo Institute of Technology, Oh-okayama, Meguro-ku, Tokyo 152-8551, Japan • *E-mail :* `hattori@math.titech.ac.jp`

Mathematical subject classification (2010). — 11J25, 53C35.

Key words and phrases. — Singular vector, geodesic ray, limit point, Tits building.

BULLETIN DE LA SOCIÉTÉ MATHÉMATIQUE DE FRANCE 0037-9484/2022/17/\$ 5.00
doi:10.24033/bsmf.2843

for all integers p, q with $q \neq 0$; a real number y is called singular if, for every $\varepsilon > 0$, there exists a positive constant $C_0(\varepsilon)$ such that the set of inequalities

$$0 < q \leq C, \quad |qy - p| < \varepsilon C^{-1}$$

has an integral solution (p, q) for all C greater than $C_0(\varepsilon)$. The set of badly approximable numbers has zero Lebesgue measure and is thick ([4], [26]). By contrast, the set of singular numbers is identical to the set of rational numbers, and hence its Hausdorff dimension is equal to zero (see [7, p. 94]).

The goal of this paper is to generalize the latter result into the setting of algebraic number fields. More precisely, let $\mathbf{k}$ be a number field of degree $d = l + 2m$ with l real places and m complex places. Let $\iota_1, \ldots, \iota_l : \mathbf{k} \longrightarrow \mathbf{R}$ be the real embeddings if l is positive, and let $\iota_{l+1}, \ldots, \iota_{l+m} : \mathbf{k} \longrightarrow \mathbf{C}$ be the complex embeddings that are not complex conjugate to each other if m is positive. We approximate elements of the Minkowski space $\mathbf{k}_M = \mathbf{R}^l \times \mathbf{C}^m$ associated to $\mathbf{k}$ by elements of $\mathbf{k}$ through the twisted diagonal embedding $\iota_{\mathbf{k}} : \mathbf{k} \longrightarrow \mathbf{k}_M$ given by

$$\iota_{\mathbf{k}}(a) = (\iota_1(a), \ldots, \iota_{l+m}(a)) \quad \text{for } a \in \mathbf{k} \, .$$

For $\xi = (\xi_1, \ldots, \xi_{l+m}), \mu = (\mu_1, \ldots, \mu_{l+m}) \in \mathbf{k}_M$, we define their sum and product by

$$\xi + \mu = (\xi_1 + \mu_1, \, \ldots, \, \xi_{l+m} + \mu_{l+m}) \, , \quad \xi \cdot \mu = (\xi_1 \mu_1, \, \ldots, \, \xi_{l+m} \mu_{l+m})$$

and we equip $\mathbf{k}_M$ with the sup norm

$$\|\xi\| = \max_{1 \leq i \leq l+m} |\xi_i| \, ,$$

where $|\cdot|$ is the usual Euclidean absolute value on $\mathbf{R}$ or $\mathbf{C}$. Let $\mathcal{O}_{\mathbf{k}}$ be the ring of integers of $\mathbf{k}$.

The following generalization of Dirichlet's theorem (cf. [32, Chapter I]) is obtained from a result of R. Quême ([30]).

Theorem 1.1 (cf. [30]). — *There exists a positive constant C depending only on $\mathbf{k}$ such that for every $\xi \in \mathbf{k}_M - \iota_{\mathbf{k}}(\mathbf{k})$, there are infinitely many $\beta = p/q$; $p \in \mathcal{O}_{\mathbf{k}}$, $q \in \mathcal{O}_{\mathbf{k}} - \{0\}$ satisfying*

$$\|\iota_{\mathbf{k}}(q) \cdot \xi - \iota_{\mathbf{k}}(p)\| < C \|\iota_{\mathbf{k}}(q)\|^{-1} \, . \tag{1}$$

A vector $\xi \in \mathbf{k}_M$ is called $\mathbf{k}$-badly approximable if there exists a positive constant C depending on ξ such that

$$\|\iota_{\mathbf{k}}(q)\| \, \|\iota_{\mathbf{k}}(q) \cdot \xi - \iota_{\mathbf{k}}(p)\| \geq C \tag{2}$$

for any $p \in \mathcal{O}_{\mathbf{k}}$ and $q \in \mathcal{O}_{\mathbf{k}} - \{0\}$ (see [13]). Let $\mathrm{Bad}(\mathbf{k})$ be the set of $\mathbf{k}$-badly approximable vectors in $\mathbf{k}_M$.

In this paper, we say that $\xi \in \mathbf{k}_M$ is a $\mathbf{k}$-singular vector if, for each $\varepsilon > 0$, the set of inequalities

$$\|\iota_{\mathbf{k}}(q)\| \leq C, \quad \|\iota_{\mathbf{k}}(q) \cdot \xi - \iota_{\mathbf{k}}(p)\| < \varepsilon C^{-1}$$

has a solution $(p, q) \in (\mathcal{O}_{\mathbf{k}})^2$ with $q \neq \mathbf{0}$ for all C greater than some positive constant $C_0(\varepsilon)$ depending on ξ and ε. Otherwise, we say that ξ is $\mathbf{k}$-regular. Note that $\mathbf{k}$-badly approximable vectors are $\mathbf{k}$-regular. Let $\mathrm{Sing}(\mathbf{k})$ be the set of $\mathbf{k}$-singular vectors.

The set $\mathrm{Bad}(\mathbf{k})$ has already been studied in [4], [10], [13], [15] and [27]. Generalizing the results in the case of the rational field and the results [10, Theorem 5.2], [15] in the case of imaginary quadratic fields with class number 1, M. Einsiedler, A. Ghosh, and B. Lytle showed the following.

Theorem 1.2 ([13]). — *The set* $\mathrm{Bad}(\mathbf{k})$ *has zero Lebesgue measure in* $\mathbf{k}_M$ *when we regard* $\mathbf{k}_M$ *as* $\mathbf{R}^d$. *It is also thick, and its Hausdorff dimension* $\dim_{\mathrm{H}}(\mathrm{Bad}(\mathbf{k}))$ *is equal to* d.

To state our results on $\mathrm{Sing}(\mathbf{k})$ we introduce an integral-valued function $f_{\mathbf{k}}$ on $\mathbf{k}_M$. For any finite set S, let $\#S$ denote the cardinality of S. For any nonnegative integer q, let $\mathbf{N}(q)$ be the set of all positive integers smaller than $q+1$. We also write $a^{(j)}$ instead of $\iota_j(a)$ for $a \in \mathbf{k}$. Let $\xi = (\xi_1, \ldots, \xi_{l+m}) \in \mathbf{k}_M$. We define a subset $A(\xi)$ of $\{0, \ldots, l\} \times \{0, \ldots, m\}$ as follows: $(\lambda, \mu) \in A(\xi)$ if and only if there exist a subset I_1 of $\mathbf{N}(l)$, a subset I_2 of $\mathbf{N}(l+m) \smallsetminus \mathbf{N}(l)$, and an element η of $\mathbf{k}$ such that $\#I_1 = \lambda$, $\#I_2 = \mu$, and

$$\xi_k \neq \eta^{(k)} \text{ for } k \in I_1 \cup I_2\ , \quad \xi_k = \eta^{(k)} \text{ for } k \notin I_1 \cup I_2\ .$$

We define a function $f_{\mathbf{k}} : \mathbf{k}_M \longrightarrow \mathbf{Z}$ by

$$f_{\mathbf{k}}(\xi) = \min\{\lambda + 2\mu \mid (\lambda,\, \mu) \in A(\xi)\} \quad \text{for } \xi \in \mathbf{k}_M\ . \tag{3}$$

We remark that $\xi \in \iota_{\mathbf{k}}(\mathbf{k})$ if and only if $f_{\mathbf{k}}(\xi) = 0$.

Generalizing the result in the case $\mathbf{k} = \mathbf{Q}$, we show the following.

Theorem 1.3. — *Let* $\xi \in \mathbf{k}_M = \mathbf{R}^l \times \mathbf{C}^m$. *Then* ξ *is* $\mathbf{k}$*-singular if and only if* $f_{\mathbf{k}}(\xi) < d/2$.

Theorem 1.4. — *Let*

$$d' = \begin{cases} (d-1)/2 & \textit{if } d \textit{ is odd} \\ d/2 - 2 & \textit{if } d \textit{ is even, } l = 0 \textit{ and } m \textit{ is even} \\ d/2 - 1 & \textit{otherwise.} \end{cases}$$

Then the Hausdorff dimension of the subset $\mathrm{Sing}(\mathbf{k})$ *of* $\mathbf{k}_M$ *is equal to* d' *when we regard* $\mathbf{k}_M$ *as* $\mathbf{R}^d$. *The set* $\mathrm{Sing}(\mathbf{k})$ *is identical to* $\iota_{\mathbf{k}}(\mathbf{k})$ *if and only if* $\mathbf{k}$ *is the rational field* $\mathbf{Q}$ *or a quadratic field or a totally complex quartic field.*

Theorem 1.3 also provides a criterion of being "not $\mathbf{k}$-badly approximable" in the case $\mathrm{Sing}(\mathbf{k}) \neq \iota_{\mathbf{k}}(\mathbf{k})$. Although measure theoretic properties of $\mathrm{Bad}(\mathbf{k})$ are well studied, it is another problem to know whether or not a given element of $\mathbf{k}_M$ is $\mathbf{k}$-badly approximable.

In the classical case, it is well known that a real number is badly approximable if and only if the partial quotients in its continued fraction expansion are bounded (cf. Theorem 5F of [32, Chapter I]). This was generalized to the case of $\mathbf{Q}(\sqrt{-3})$ in [9], the case of arbitrary imaginary quadratic field in [25]. E. Burger and R. Hines found methods to construct $\mathbf{k}$-badly approximable vectors for any $\mathbf{k}$ different from $\mathbf{Q}$ ([6], [24]). There is another method to construct $\mathbf{k}$-badly approximable vectors in the case $l+m \geq 2$: Proposition 3.1 of [13] and the argument in the proof of Proposition 8.5 of [20] show that $(b, \dots, b) \in \mathbf{k}_M$ is $\mathbf{k}$-badly approximable for any badly approximable real number b.

On the other hand, there were no known concrete criteria to ensure that a vector in $\mathbf{k}_M$ is not $\mathbf{k}$-badly approximable in the case $l+m \geq 2$, even though $\mathrm{Bad}(\mathbf{k})$ is a set of measure zero.

Example. — Let ζ be a primitive complex 7th root of unity and $\mathbf{k}$ the cyclotomic field $\mathbf{Q}(\zeta)$. Let $a \in \mathbf{k}$ and b be a badly approximable real number that is not contained in $\mathbf{k}$. Then $(a^{(1)}, a^{(2)}, b), (a^{(1)}, b, a^{(3)}), (b, a^{(2)}, a^{(3)}) \in \mathbf{k}_M = \mathbf{C}^3$ are not $\mathbf{k}$-badly approximable, while $(b, b, b) \in \mathbf{k}_M$ is $\mathbf{k}$-badly approximable.

We outline the proofs of the main results. Our approach is based on Riemannian geometry of nonpositively curved manifolds (see [1], [11] and [12]).

In [17], L.R. Ford found that Dirichlet's theorem is closely related to the geometry of the hyperbolic plane $\mathbf{H}$. He considered horoballs in the upper half-plane $\mathbf{H}$ together with the action of $SL(2, \mathbf{Z})$ on $\mathbf{H}$. After Ford, connections between Diophantine approximation problems and the geometry of hyperbolic spaces, or more generally Gromov hyperbolic spaces, were studied extensively by many authors (see, for example, [23], [16] and references therein). Some product spaces of hyperbolic spaces were also used in [20] and [28] for approximation by algebraic numbers.

From this point of view, it is natural in our case to consider horoballs in the product of l copies of the hyperbolic plane and m copies of the 3-dimensional hyperbolic space together with the action of $\Gamma = SL(2, \mathcal{O}_{\mathbf{k}})$ on this product.

Let $V = SL(2, \mathbf{R})/SO(2)$ and $\widehat{V} = SL(2, \mathbf{C})/SU(2)$. We identify V (respectively $\widehat{V}$) with the upper half-plane $\mathbf{H}$ (respectively the three-dimensional upper half-space $\mathcal{H}$) in the usual manner (see Section 2 for a more precise description), and equip V (respectively $\widehat{V}$) with the Poincaré metric (respectively the metric that is twice the Poincaré metric on $\mathcal{H}$). The group $G = SL(2, \mathbf{R})^l \times SL(2, \mathbf{C})^m$ acts on the Riemannian product $\widetilde{V} = V^l \times \widehat{V}^m$ by

$$g \cdot z = (g_1 \cdot x_1, \dots, g_l \cdot x_l, g_{l+1} \cdot \widehat{x}_{l+1}, \dots, g_{l+m} \cdot \widehat{x}_{l+m}) \tag{4}$$

for $z = (x_1, \ldots, x_l, \widehat{x}_{l+1}, \ldots, \widehat{x}_{l+m}) \in \widetilde{V}$ and $g = (g_1, \ldots, g_{l+m}) \in G$, where $x_1, \ldots, x_l \in V$ and $\widehat{x}_{l+1}, \ldots, \widehat{x}_{l+m} \in \widehat{V}$. Note that, in the case $l = 0$ or $m = 0$, an obvious modification should be made to this formula (4). However, in the rest of this paper, we omit to write such modified formulae in order to avoid complicating the description. We extend each embedding ι_j to an embedding of $SL(2, \mathbf{k})$ into $SL(2, \mathbf{R})$ or $SL(2, \mathbf{C})$ by

$$\iota_j(g) = \begin{pmatrix} \iota_j(p) & \iota_j(r) \\ \iota_j(q) & \iota_j(s) \end{pmatrix} \quad \text{for } g = \begin{pmatrix} p & r \\ q & s \end{pmatrix} \in SL(2, \mathbf{k}) \; . \tag{5}$$

Then the twisted diagonal embedding $\iota_{\mathbf{k}}$ can be extended to an embedding $SL(2, \mathbf{k}) \longrightarrow G$ by

$$\iota_{\mathbf{k}}(g) = (\iota_1(g), \ldots, \iota_{l+m}(g)) \quad \text{for } g \in SL(2, \mathbf{k}) \; . \tag{6}$$

The group $SL(2, \mathbf{k})$ acts isometrically on $\widetilde{V}$ through this embedding:

$$\iota_{\mathbf{k}}(g) \cdot z = (\iota_1(g) \cdot x_1, \; \ldots \; , \iota_l(g) \cdot x_l, \; \iota_{l+1}(g) \cdot \widehat{x}_{l+1}, \; \ldots, \; \iota_{l+m}(g) \cdot \widehat{x}_{l+m}) \tag{7}$$

for $z = (x_1, \ldots, x_l, \widehat{x}_{l+1}, \ldots, \widehat{x}_{l+m}) \in \widetilde{V}$ and $g \in SL(2, \mathbf{k})$.

Let $x_0 \in V$ be the coset of the identity element of $SL(2, \mathbf{R})$, $\widehat{x}_0 \in \widehat{V}$ the coset of the identity element of $SL(2, \mathbf{C})$, and let

$$z_0 = (x_0, \ldots, x_0, \widehat{x}_0, \ldots, \widehat{x}_0) \in \widetilde{V} \; . \tag{8}$$

Then the isotropy subgroup of G at z_0 is $K = SO(2)^l \times SU(2)^m$, and $\widetilde{V}$ is diffeomorphic to the quotient space G/K.

We associate each element of $\mathbf{k}_M$ with a certain geodesic ray of $\widetilde{V}$ as follows. Let $\xi = (\xi_1, \ldots, \xi_{l+m}) \in \mathbf{k}_M$. We define an element u_ξ of G by

$$u_\xi = \left(\begin{pmatrix} 1 & \xi_1 \\ 0 & 1 \end{pmatrix}, \ldots, \begin{pmatrix} 1 & \xi_{l+m} \\ 0 & 1 \end{pmatrix} \right) \tag{9}$$

and define a geodesic ray $\gamma_\xi : [0, \infty) \longrightarrow \widetilde{V}$ by

$$\gamma_\xi(t) = u_\xi g_t \cdot z_0 \quad \text{for } t \geq 0 \; , \tag{10}$$

where

$$g_t = \left(\begin{pmatrix} e^{-t/\sqrt{4d}} & \\ & e^{t/\sqrt{4d}} \end{pmatrix}, \ldots, \begin{pmatrix} e^{-t/\sqrt{4d}} & \\ & e^{t/\sqrt{4d}} \end{pmatrix} \right) \in G \; . \tag{11}$$

Let $\Pi : \widetilde{V} \longrightarrow \iota_{\mathbf{k}}(\mathit{\Gamma}) \backslash \widetilde{V}$ be the projection to the noncompact quotient space. For any geodesic ray γ of $\widetilde{V}$, we say that $\Pi \circ \gamma$ is divergent if, for any given compact subset W of $\iota_{\mathbf{k}}(\mathit{\Gamma}) \backslash \widetilde{V}$, there exists $t_0 \geq 0$ such that $\Pi \circ \gamma(t) \notin W$ for $t \geq t_0$. This condition is a paraphrase of another one written in terms of horoballs in $\widetilde{V}$ and the action of $\iota_{\mathbf{k}}(\mathit{\Gamma})$ on $\widetilde{V}$ (see Proposition 3.2).

Proposition 1.5. — *$\Pi \circ \gamma_\xi$ is divergent if and only if ξ is a $\mathbf{k}$-singular vector.*

This proposition is related to Dani's correspondence [8, Theorem 2.14] as follows. The trajectory $\{\iota_{\mathbf{k}}(\mathit{\Gamma})u_\xi g_t \,|\, t \geq 0\}$ in $\iota_{\mathbf{k}}(\mathit{\Gamma})\backslash G$ is said to be divergent if, for any given compact subset D of $\iota_{\mathbf{k}}(\mathit{\Gamma})\backslash G$, there exists $t_0 \geq 0$ such that $\iota_{\mathbf{k}}(\mathit{\Gamma})u_\xi g_t \notin D$ for $t \geq t_0$. Since the natural projection $G \longrightarrow \widetilde{V} = G/K$ is a fiber bundle with compact fiber K, Proposition 1.5 is equivalent to the following, which is a variant of [8, Theorem 2.14].

Proposition 1.6. — *The trajectory* $\{\iota_{\mathbf{k}}(\mathit{\Gamma})u_\xi g_t \,|\, t \geq 0\}$ *is divergent if and only if* ξ *is a* $\mathbf{k}$*-singular vector.*

In the case $d = 2$, $l = 0$, $m = 1$, $\mathbf{k}$ is an imaginary quadratic field, and $\widetilde{V} = \widehat{V}$ is the three-dimensional upper half space equipped with the rescaled Poincaré metric. Then, it is well known in three-dimensional hyperbolic geometry that $\Pi \circ \gamma_\xi$ is divergent if and only if $\xi \in \mathbf{k}$. From Proposition 1.5, ξ is $\mathbf{k}$-singular if and only if $\xi \in \mathbf{k}$, and $\dim_{\mathrm{H}}(\mathrm{Sing}(\mathbf{k})) = 0$. Hence it suffices to prove the case $l + m \geq 2$.

We now turn our attention to the geometric boundary $\widetilde{V}(\infty)$ of $\widetilde{V}$, which is defined as follows. Let $d_{\widetilde{V}}$ be the distance on $\widetilde{V}$ induced from the product metric of $\widetilde{V}$. Two geodesic rays γ_1 and γ_2 of $\widetilde{V}$ are called asymptotic if $d_{\widetilde{V}}(\gamma_1(t), \gamma_2(t))$ is uniformly bounded on $[0, \infty)$. Being asymptotic is an equivalence relation. The geometric boundary $\widetilde{V}(\infty)$ is the set of asymptotic classes of geodesic rays of $\widetilde{V}$. The equivalence class represented by a geodesic ray γ is denoted by $\gamma(\infty)$, which we call the point at infinity of γ. There is a natural topology, the cone topology, on $\widetilde{V} = \widetilde{V} \cup \widetilde{V}(\infty)$ (see [11], [12]).

It is well known that $\widetilde{V}(\infty)$ admits the structure of a spherical Tits building. We may suppose that G is the topological identity component of the group of real points of some semisimple linear algebraic group $\mathbf{G}$ defined over $\mathbf{Q}$ with $\mathbf{Q}$-rank 1 and that $\iota_{\mathbf{k}}(\mathit{\Gamma})$ is isomorphic to the group $\mathbf{G}_{\mathbf{Z}}$ of integral points of $\mathbf{G}$, by using the method of "restriction of scalars" of Weil ([35]). Hence, $\widetilde{V}(\infty)$ contains a smaller building $\mathcal{W}_{\mathbf{Q}}$, which is the spherical Tits building constructed from the set of all parabolic $\mathbf{Q}$-subgroups of $\mathbf{G}$. We call $\mathcal{W}_{\mathbf{Q}}$ the rational Tits building of $\mathbf{G}$. We also recall that there exists a natural distance Td on $\widetilde{V}(\infty)$ ([1], [11]).

We say that a point $v \in \widetilde{V}(\infty)$ is a conical limit point of $\iota_{\mathbf{k}}(\mathit{\Gamma})$ if for some (and hence every) geodesic ray γ with $\gamma(\infty) = v$, there exists a sequence $(h_i)_{i \in \mathbf{N}}$ of different elements of $\mathit{\Gamma}$ such that $d_{\widetilde{V}}(\iota_{\mathbf{k}}(h_i) \cdot z_0, \gamma([0, \infty)))$ are uniformly bounded (see [19, Definition B]). Since $\Pi \circ \gamma_\xi$ is divergent if and only if the point $\gamma_\xi(\infty)$ at infinity of the geodesic ray γ_ξ is not a conical limit point of $\iota_{\mathbf{k}}(\mathit{\Gamma})$ (see Proposition 6.4), we obtain the following from Theorems A, B of [19] and Proposition 1.5.

Theorem 1.7. — ξ *is* $\mathbf{k}$*-singular if and only if* $Td(\gamma_\xi(\infty), \mathcal{W}_{\mathbf{Q}}) < \pi/2$.

Theorem 1.3 now follows from the estimate of the distance $Td(\gamma_\xi(\infty), \mathcal{W}_{\mathbf{Q}})$. Theorems 1.3 and 1.7 hold also in the case $l+m=1$, so we prove them without the assumption on $l+m$.

Finally, we deduce from Theorem 1.3 that $\mathrm{Sing}(\mathbf{k})$ is a countable union of embedded copies of $\mathbf{R}^{d'}$, which shows that $\dim_H(\mathrm{Sing}(\mathbf{k})) = d'$.

We found Theorem 1.3 by visualizing the set $\mathrm{Sing}(\mathbf{k})$ as a subset of the boundary $\widetilde{V}(\infty)$ and trying to estimate the distance $Td(\gamma_\xi(\infty), \mathcal{W}_{\mathbf{Q}})$. We believe that this geometric approach would be useful for finding new results in Diophantine approximation.

This paper is organized as follows. In Section 2 we collect some facts on geometry of V, $\widehat{V}$ and the product $\widetilde{V} = V^l \times \widehat{V}^m$. In Section 3 we recall some facts concerning fundamental sets for $\iota_{\mathbf{k}}(\mathit{\Gamma})$. The proof of Proposition 1.5 is given in Section 4. In Section 5 we describe the geometric boundary of $\widetilde{V}$, which admits the structure of a spherical Tits building, and recall the definition of the distance Td on $\widetilde{V}(\infty)$. In Section 6 we describe conical limit points of $\iota_{\mathbf{k}}(\mathit{\Gamma})$, rational Tits building and prove Theorem 1.7. We prove Theorem 1.3 in Section 7 and derive Theorem 1.4 from Theorem 1.3 in the final section.

2. Geometry of $(SL(2,\mathrm{R})/SO(2))^l \times (SL(2,\mathrm{C})/SU(2))^m$

The group $SL(2,\mathbf{R})$ acts transitively on the upper half-plane

$$\mathbf{H} = \left\{x + \sqrt{-1}y \mid x, y \in \mathbf{R}; y > 0\right\}$$

as a group of linear fractional transformations:

$$\begin{aligned} g \cdot (x + \sqrt{-1}y) &= \frac{p(x+\sqrt{-1}y)+r}{q(x+\sqrt{-1}y)+s} \\ &= \frac{(px+r)(qx+s)+pqy^2}{(qx+s)^2+q^2y^2} + \sqrt{-1}\,\frac{y}{(qx+s)^2+q^2y^2} \\ &\qquad \text{for } x+\sqrt{-1}y \in \mathbf{H} \text{ and } g = \begin{pmatrix} p & r \\ q & s \end{pmatrix} \in SL(2,\mathbf{R})\,. \end{aligned}$$

Since the isotropy group at $\sqrt{-1}$ is $SO(2)$, we may identify $V = SL(2,\mathbf{R})/SO(2)$ with the upper half-plane. Under this identification, we equip V with the Poincaré metric $(dx^2+dy^2)/y^2$, so that V has constant sectional curvature -1. This metric is the left $SL(2,\mathbf{R})$-invariant Riemannian metric induced from the half of the Killing form of the Lie algebra of $SL(2,\mathbf{R})$. Similarly, $SL(2,\mathbf{C})$ acts transitively on the three-dimensional upper half-space

$$\mathcal{H} = \{(z,t) \in \mathbf{C} \times \mathbf{R} \mid t > 0\}$$

by $g\cdot(z,t)=(z',t')$,

$$z'=\frac{(pz+r)(\overline{qz}+\overline{s})+p\overline{q}t^2}{|qz+s|^2+|q|^2t^2}\ ,\quad t'=\frac{t}{|qz+s|^2+|q|^2t^2}$$
$$\text{for } (z,t)\in\mathcal{H} \text{ and } g=\begin{pmatrix} p & r\\ q & s\end{pmatrix}\in SL(2,\mathbf{C})$$

(see [14, 1.1 of Chapter 1]). The isotropy group at $(0,1)$ is $SU(2)$, and hence we may identify $\widehat{V}=SL(2,\mathbf{C})/SU(2)$ with $\mathcal{H}$. Under this identification, we equip $\widehat{V}$ with the left $SL(2,\mathbf{C})$-invariant metric that is twice the Poincaré metric $(dzd\overline{z}+dt^2)/t^2$. This metric is the left $SL(2,\mathbf{C})$-invariant Riemannian metric induced from the half of the Killing form of the Lie algebra of $SL(2,\mathbf{C})$, and $\widehat{V}$ has constant sectional curvature $-1/2$.

Let $G=SL(2,\mathbf{R})^l\times SL(2,\mathbf{C})^m$ and $K=SO(2)^l\times SU(2)^m$ as in Section 1. Then the Riemannian product $\widetilde{V}=V^l\times\widehat{V}^m$ is the symmetric space G/K of noncompact type equipped with the left G-invariant Riemannian metric induced from the half of the Killing form of the Lie algebra of G. This symmetric space is a Hadamard manifold, that is, a simply connected, complete Riemannian manifold of nonpositive sectional curvature (see [11], [22] for more details on symmetric spaces of noncompact type). The group $\Gamma=SL(2,\mathcal{O}_{\mathbf{k}})$ acts isometrically, properly discontinuously on $\widetilde{V}$ by (7), and the volume of the quotient space $\iota_{\mathbf{k}}(\Gamma)\backslash\widetilde{V}$ is finite.

Let M be a Hadamard manifold and d_M the distance on M induced from the Riemannian metric of M. A map $\gamma:[0,\infty)\longrightarrow M$ is a geodesic ray if it realizes the distance between any two points on it: $d_M(\gamma(t),\gamma(t'))=|t-t'|$ for any $t,t'\geq 0$. Any unit speed geodesic $[0,\infty)\longrightarrow M$ is a geodesic ray.

Definition 2.1 (cf. [1], [11], [12]). — (1) Let $\gamma:[0,\infty)\longrightarrow M$ be a geodesic ray. Then the function $b(\gamma)$ on M defined by

$$b(\gamma)(x)=\lim_{t\to\infty}\{d_M(x,\gamma(t))-t\}\quad\text{for } x\in M$$

is called the Busemann function associated to γ.

(2) For any geodesic ray γ and any real number τ, the set $B(\gamma,\tau)=b(\gamma)^{-1}(-\infty,-\tau)$ is called a horoball in M.

For an element g of the group $I(M)$ of isometries of M and $x\in M$, we denote by $g\cdot x$ the image of x under g. We use the similar notation for subsets of M and geodesics of M. Then we have the following for any $g\in I(M)$ and $\tau\in\mathbf{R}$.

$$b(g\cdot\gamma)(x)=b(\gamma)(g^{-1}\cdot x)\quad\text{for } x\in M\ ,\tag{12}$$

$$g\cdot B(\gamma,\tau)=B(g\cdot\gamma,\,\tau)\ .\tag{13}$$

Let $x_0 \in V$ be the coset of the identity element of $SL(2,\mathbf{R})$ and let $\widehat{x}_0 \in \widehat{V}$ be the coset of the identity element of $SL(2,\mathbf{C})$. In other words, $x_0 = \sqrt{-1} \in \mathbf{H}$ and $\widehat{x}_0 = (0,1) \in \mathcal{H}$. We define geodesic rays $\gamma_0 : [0,\infty) \longrightarrow V$ and $\widehat{\gamma}_0 : [0,\infty) \longrightarrow \widehat{V}$ by

$$\begin{aligned} \gamma_0(t) &= \begin{pmatrix} e^{t/2} & \\ & e^{-t/2} \end{pmatrix} \cdot x_0 = e^t\sqrt{-1}\ , \\ \widehat{\gamma}_0(t) &= \begin{pmatrix} e^{t/\sqrt{8}} & \\ & e^{-t/\sqrt{8}} \end{pmatrix} \cdot \widehat{x}_0 = (0, e^{t/\sqrt{2}})\ . \end{aligned} \tag{14}$$

Then we have

$$b(\gamma_0)(x+\sqrt{-1}y) = -\log y \quad \text{for } x+\sqrt{-1}y \in \mathbf{H}\ ,$$
$$B(\gamma_0,\, \tau) = b(\gamma_0)^{-1}(-\infty,\, -\tau) = \left\{ x+\sqrt{-1}y \in \mathbf{H} \,\middle|\, y > e^{\tau} \right\}$$

and

$$g \cdot B(\gamma_0,\, \tau) = \left\{ x+\sqrt{-1}y \in \mathbf{H} \,\middle|\, \left(x - \frac{p}{q}\right)^2 + \left(y - \frac{1}{2q^2e^{\tau}}\right)^2 < \frac{1}{4q^4e^{2\tau}} \right\}$$
$$\text{for } g = \begin{pmatrix} p & r \\ q & s \end{pmatrix} \in SL(2,\mathbf{R}) \text{ with } q \neq 0\ .$$

We also have

$$b(\widehat{\gamma}_0)((z,t)) = -\sqrt{2}\log t \quad \text{for } (z,t) \in \mathcal{H}\ ,$$
$$B(\widehat{\gamma}_0,\, \tau) = b(\widehat{\gamma}_0)^{-1}(-\infty,\, -\tau) = \left\{ (z,t) \in \mathcal{H} \,\middle|\, t > e^{\tau/\sqrt{2}} \right\}$$

and

$$g \cdot B(\widehat{\gamma}_0,\, \tau) = \left\{ (z,t) \in \mathcal{H} \,\middle|\, \left|z - \frac{p}{q}\right|^2 + \left(t - \frac{1}{2|q|^2e^{\tau/\sqrt{2}}}\right)^2 < \frac{1}{4|q|^4e^{\sqrt{2}\tau}} \right\}$$
$$\text{for } g = \begin{pmatrix} p & r \\ q & s \end{pmatrix} \in SL(2,\mathbf{C}) \text{ with } q \neq 0\ .$$

Let

$$\square_1(x+\sqrt{-1}y) = \frac{1}{y} \quad \text{for } x+\sqrt{-1}y \in \mathbf{H}$$

and

$$\square_1((w,t)) = \frac{1}{t} \quad \text{for } (w,t) \in \mathcal{H}\ .$$

Then we have

$$b(\gamma_0)(z) = \log \square_1(z) \quad \text{for } z \in V \tag{15}$$

and

$$b(\widehat{\gamma}_0)(z) = \sqrt{2}\log \square_1(z) \quad \text{for } z \in \widehat{V}\ . \tag{16}$$

If

$$z = g \cdot x_0 \ \text{ and } \ g = \begin{pmatrix} p & r \\ q & s \end{pmatrix} \in SL(2, \mathbf{R}) \ ,$$

then $\Box_1(z)$ is the $(2,2)$-entry of the matrix $g\,{}^t g$: $\Box_1(z) = q^2 + s^2$. If

$$z = g \cdot \widehat{x}_0 \ \text{ and } \ g = \begin{pmatrix} p & r \\ q & s \end{pmatrix} \in SL(2, \mathbf{C}) \ ,$$

then $\Box_1(z)$ is the $(2,2)$-entry of the matrix $g\,{}^t\overline{g}$: $\Box_1(z) = |q|^2 + |s|^2$ (cf. [18, Lemma 2.4], [21, Lemma 5.1]).

We extend the geodesic rays γ_0, $\widehat{\gamma}_0$ defined by (14) to geodesics $\mathbf{R} \longrightarrow V$, $\mathbf{R} \longrightarrow \widehat{V}$, respectively, and denote them by the same symbols:

$$\gamma_0(t) = \begin{pmatrix} e^{t/2} & \\ & e^{-t/2} \end{pmatrix} \cdot x_0 \ , \quad \widehat{\gamma}_0(t) = \begin{pmatrix} e^{t/\sqrt{8}} & \\ & e^{-t/\sqrt{8}} \end{pmatrix} \cdot \widehat{x}_0 \quad \text{for } t \in \mathbf{R} \ .$$

We define a geodesic $\gamma_{\mathbf{k}} : \mathbf{R} \longrightarrow \widetilde{V}$ by

$$\gamma_{\mathbf{k}}(t) = \left(\gamma_0\left(\frac{t}{\sqrt{d}}\right), \ldots, \gamma_0\left(\frac{t}{\sqrt{d}}\right), \widehat{\gamma}_0\left(\frac{\sqrt{2}t}{\sqrt{d}}\right), \ldots, \widehat{\gamma}_0\left(\frac{\sqrt{2}t}{\sqrt{d}}\right) \right) \ . \tag{17}$$

In the right-hand side of (17), the first l entries are equal to

$$\gamma_0\left(\frac{t}{\sqrt{d}}\right) = \operatorname{diag}\left(e^{t/\sqrt{4d}}, \ e^{-t/\sqrt{4d}}\right) \cdot x_0 = e^{t/\sqrt{d}}\sqrt{-1} \ ,$$

and the last m entries are equal to

$$\widehat{\gamma}_0\left(\frac{\sqrt{2}t}{\sqrt{d}}\right) = \operatorname{diag}\left(e^{t/\sqrt{4d}}, \ e^{-t/\sqrt{4d}}\right) \cdot \widehat{x}_0 = (0, \ e^{t/\sqrt{d}}) \ .$$

Since

$$l \cdot \left(\frac{1}{\sqrt{d}}\right)^2 + m \cdot \left(\frac{\sqrt{2}}{\sqrt{d}}\right)^2 = 1,$$

$\gamma_{\mathbf{k}}$ is a unit speed geodesic. We define a geodesic ray $\gamma^* : [0, \infty) \longrightarrow \widetilde{V}$ by

$$\gamma^*(t) = \gamma_{\mathbf{k}}(t) \quad \text{for } t \geq 0 \ . \tag{18}$$

It follows from [1, §3.8] and (15), (16) that

$$b(\gamma^*)(z_1,\ldots,z_{l+m}) = \frac{1}{\sqrt{d}}\left\{\sum_{j=1}^{l} b(\gamma_0)(z_j)\right\} + \frac{\sqrt{2}}{\sqrt{d}}\left\{\sum_{j=l+1}^{l+m} b(\widehat{\gamma}_0)(z_j)\right\} \tag{19}$$

$$= \frac{1}{\sqrt{d}}\left\{\sum_{j=1}^{l} \log \Box_1(z_j)\right\} + \frac{2}{\sqrt{d}}\left\{\sum_{j=l+1}^{l+m} \log \Box_1(z_j)\right\}$$

$$= \frac{1}{\sqrt{d}} \log\left\{\left(\prod_{j=1}^{l} \Box_1(z_j)\right)\left(\prod_{j=l+1}^{l+m} \Box_1(z_j)\right)^2\right\}$$

for $z_1,\ldots,z_l \in V$ and $z_{l+1},\ldots,z_{l+m} \in \widehat{V}$. In this formula, γ_0 is regarded as a geodesic ray $[0,\infty) \longrightarrow V$ and $\widehat{\gamma}_0$ is regarded as a geodesic ray $[0,\infty) \longrightarrow \widehat{V}$.

Let $\xi = (\xi_1,\ldots,\xi_{l+m}) \in \mathbf{k}_M$. We define an element u_ξ of G by

$$u_\xi = \left(\begin{pmatrix}1 & \xi_1\\ 0 & 1\end{pmatrix},\ldots,\begin{pmatrix}1 & \xi_l\\ 0 & 1\end{pmatrix},\begin{pmatrix}1 & \xi_{l+1}\\ 0 & 1\end{pmatrix},\ldots,\begin{pmatrix}1 & \xi_{l+m}\\ 0 & 1\end{pmatrix}\right)$$

and define a geodesic ray $\gamma_\xi : [0,\infty) \longrightarrow \widetilde{V}$ by

$$\gamma_\xi(t) = u_\xi \cdot \gamma_{\mathbf{k}}(-t) = \Big(\xi_1 + e^{-t/\sqrt{d}}\sqrt{-1},\ldots,\, \xi_l + e^{-t/\sqrt{d}}\sqrt{-1}, \tag{20}$$
$$(\xi_{l+1},\, e^{-t/\sqrt{d}}),\ldots,\,(\xi_{l+m},\, e^{-t/\sqrt{d}})\Big) \quad \text{for } t \geq 0\,.$$

3. Compactness criterion by horoballs

Let h be the class number of $\mathbf{k}$. For any subset $F = \{f_1,\ldots,f_n\}$ of $\mathcal{O}_{\mathbf{k}}$, we denote by $(f_1,\ldots,f_n)$ the ideal of $\mathcal{O}_{\mathbf{k}}$ generated by F. We choose, in the h ideal classes of $\mathbf{k}$, fixed integral ideals $\mathfrak{a}_1 = (a_1,b_1)$, $\ldots$, $\mathfrak{a}_h = (a_h,b_h)$ with $a_i, b_i \in \mathcal{O}_{\mathbf{k}}$, so that each $\mathfrak{a}_i$ is of minimum norm among all the integral ideals of its class. Since $\mathfrak{a}_i(\mathfrak{a}_i)^{-1} = \mathcal{O}_{\mathbf{k}}$, we choose $c_i, d_i \in (\mathfrak{a}_i)^{-1}$ such that $a_id_i - b_ic_i = 1$ and put

$$g_i = \begin{pmatrix}a_i & c_i\\ b_i & d_i\end{pmatrix} \quad \text{for each } i = 1,\ldots,h\,.$$

Let B be the subgroup of $SL(2,\mathbf{k})$ consisting of all the upper triangular matrices in $SL(2,\mathbf{k})$. The following should be well known to experts (cf. [33, §2 of Chapter III] for the case of totally real number fields). Since we could not find suitable references, we provide a proof for the sake of completeness.

Proposition 3.1. — *The set $\{g_1,\ldots,g_h\}$ is a complete representative system of the set of double coset classes $\Gamma\backslash SL(2,\mathbf{k})/B$.*

Proof. — We first show that $SL(2,\mathbf{k}) = \bigcup_{i=1}^{h} \Gamma g_i B$. Let

$$g = \begin{pmatrix} a & c \\ b & d \end{pmatrix} \in SL(2,\mathbf{k}) .$$

If $a \neq 0$, take $p, q \in \mathcal{O}_{\mathbf{k}}$ such that

$$\frac{b}{a} = \frac{p}{q} .$$

Suppose that the ideal (p, q) is equivalent to $\mathfrak{a}_j$. Then there exists $\theta \in \mathbf{k}$ such that

$$\mathfrak{a}_j = (\theta)(p,q) = (\theta p,\, \theta q) .$$

Take $r, s \in (\mathfrak{a}_j)^{-1}$ such that

$$g' = \begin{pmatrix} \theta q & r \\ \theta p & s \end{pmatrix} \in SL(2,\mathbf{k}) .$$

Let $g'' = g' g_j^{-1}$. Then

$$g'' = \begin{pmatrix} \theta q & r \\ \theta p & s \end{pmatrix} \begin{pmatrix} d_j & -c_j \\ -b_j & a_j \end{pmatrix} = \begin{pmatrix} (\theta q) d_j - b_j r & -(\theta q) c_j + a_j r \\ (\theta p) d_j - b_j s & -(\theta p) c_j + a_j s \end{pmatrix} \in \Gamma .$$

On the other hand, we have

$$g' \begin{pmatrix} 1/(\theta q) & -r \\ 0 & \theta q \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ p/q & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ b/a & 1 \end{pmatrix} = \begin{pmatrix} a & c \\ b & d \end{pmatrix} \begin{pmatrix} a & c \\ 0 & a^{-1} \end{pmatrix}^{-1} .$$

Therefore

$$g = g' \begin{pmatrix} 1/(\theta q) & -r \\ 0 & \theta q \end{pmatrix} \begin{pmatrix} a & c \\ 0 & a^{-1} \end{pmatrix} = g'' g_j \begin{pmatrix} 1/(\theta q) & -r \\ 0 & \theta q \end{pmatrix} \begin{pmatrix} a & c \\ 0 & a^{-1} \end{pmatrix} \in \Gamma g_j B .$$

If $a = 0$, then $b \neq 0$ and

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} g = \begin{pmatrix} b & d \\ 0 & -c \end{pmatrix} .$$

From the above argument, this shows that $g \in \Gamma g_j B$ for some j. Thus we conclude that $SL(2,\mathbf{k}) = \bigcup_{i=1}^{h} \Gamma g_i B$.

Suppose that $\Gamma g_i B = \Gamma g_j B$. Then there exist

$$g = \begin{pmatrix} p & r \\ q & s \end{pmatrix} \in \Gamma, \quad u = \begin{pmatrix} a & c \\ 0 & a^{-1} \end{pmatrix} \in B$$

such that $g_i = g g_j u$. Then we have

$$\begin{pmatrix} a_i & c_i \\ b_i & d_i \end{pmatrix} = \begin{pmatrix} p & r \\ q & s \end{pmatrix} \begin{pmatrix} a_j & c_j \\ b_j & d_j \end{pmatrix} \begin{pmatrix} a & c \\ 0 & a^{-1} \end{pmatrix} = \begin{pmatrix} pa_j + rb_j & pc_j + rd_j \\ qa_j + sb_j & qc_j + sd_j \end{pmatrix} \begin{pmatrix} a & c \\ 0 & a^{-1} \end{pmatrix}$$

and

$$a_i = a(pa_j + rb_j), \quad b_i = a(qa_j + sb_j) .$$

Therefore $\mathfrak{a}_j = (a_j, b_j) = (pa_j + rb_j,\ qa_j + sb_j)$ is equivalent to $\mathfrak{a}_i = (a_i, b_i)$ and $i = j$. □

Let $\Pi : \widetilde{V} \longrightarrow \iota_{\mathbf{k}}(\Gamma)\backslash\widetilde{V}$ be the natural projection. For any positive number τ, let

$$\widetilde{W}(\tau) = \widetilde{V} \smallsetminus \bigcup_{i=1}^{h} \bigcup_{g\in\Gamma} \iota_{\mathbf{k}}(g)\iota_{\mathbf{k}}(g_i)\cdot B(\gamma^*,\,\tau) \tag{21}$$

and $W(\tau) = \Pi(\widetilde{W}(\tau))$.

Using the method of "restriction of scalars" of Weil ([35]) (cf. Proposition 6.1.3 and Corollary 6.1.4 of [36]), one can find an algebraic group $\mathbf{G}$ defined over $\mathbf{Q}$ of $\mathbf{Q}$-rank 1 such that the group $\mathbf{G}_{\mathbf{R}}$ of real points of $\mathbf{G}$ is isomorphic to $G = SL(2,\mathbf{R})^l \times SL(2,\mathbf{C})^m$ and the group $\mathbf{G}_{\mathbf{Z}}$ of integral points of $\mathbf{G}$ is isomorphic to $\iota_{\mathbf{k}}(\Gamma)$. Since $\iota_{\mathbf{k}}(\Gamma)$ is a lattice of G, from Borel's result ([3, Chapter III], [2, Théorème 4.4]) and [29, Proposition 2.1] (see also [31, Chapter XIII]), there exists a positive number τ_0 such that the following hold: For any $\tau \geq \tau_0$, $W(\tau)$ is a nonempty compact subset of $\iota_{\mathbf{k}}(\Gamma)\backslash\widetilde{V}$, and the inverse image under Π of each connected component of the complement of $W(\tau)$ coincides with $\bigcup_{g\in\Gamma} \iota_{\mathbf{k}}(g)\iota_{\mathbf{k}}(g_i)\cdot B(\gamma^*,\,\tau)$ for some i. Moreover,

$$\widetilde{V} = \bigcup_{\tau_0\leq\tau} \widetilde{W}(\tau) \tag{22}$$

and $\iota_{\mathbf{k}}(\Gamma)\backslash\widetilde{V} = \bigcup_{\tau_0\leq\tau} W(\tau)$ is an exhaustion of $\iota_{\mathbf{k}}(\Gamma)\backslash\widetilde{V}$ by compact subsets. This shows the following.

Proposition 3.2. — *Let $\gamma : [0,\infty) \longrightarrow \widetilde{V}$ be a geodesic ray. Then the following two conditions are equivalent.*

(1) *For any $\tau > \tau_0$, there exists a positive number $T_0 = T_0(\tau)$ such that the following holds: For any $t \geq T_0$, there exists a horoball of the form $\iota_{\mathbf{k}}(g)\cdot B(\gamma^*,\,\tau)$, where g is an element of $\bigcup_{i=1}^{h} \Gamma g_i$, containing $\gamma(t)$.*
(2) *$\Pi\circ\gamma$ is divergent in $\iota_{\mathbf{k}}(\Gamma)\backslash\widetilde{V}$.*

4. Proof of Proposition 1.5

We first calculate the value of $b(\gamma^*)(\iota_{\mathbf{k}}(g)^{-1}\cdot\gamma_\xi(\lambda))$ for $g \in SL(2,\mathbf{k})$ and $\lambda \geq 0$.

Let

$$g = \begin{pmatrix} p & r \\ q & s \end{pmatrix} \in SL(2,\mathbf{k}), \quad \gamma_\xi(\lambda) = (\sigma_1(\lambda),\ldots,\sigma_{l+m}(\lambda)).$$

Then we have

$$\sigma_j(\lambda) = \begin{cases} \begin{pmatrix} e^{-\lambda/\sqrt{4d}} & \xi_j e^{\lambda/\sqrt{4d}} \\ 0 & e^{\lambda/\sqrt{4d}} \end{pmatrix} \cdot x_0 & \text{if } 1 \le j \le l \\ \begin{pmatrix} e^{-\lambda/\sqrt{4d}} & \xi_j e^{\lambda/\sqrt{4d}} \\ 0 & e^{\lambda/\sqrt{4d}} \end{pmatrix} \cdot \widehat{x}_0 & \text{if } l+1 \le j \le l+m\,, \end{cases}$$

and

$$\iota_j(g)^{-1} = \begin{pmatrix} s^{(j)} & -r^{(j)} \\ -q^{(j)} & p^{(j)} \end{pmatrix} .$$

Let $1 \le j \le l$. Then

$$\iota_j(g)^{-1} \cdot \sigma_j(\lambda) = \begin{pmatrix} s^{(j)} e^{-\lambda/\sqrt{4d}} & -(r^{(j)} - s^{(j)}\xi_j) e^{\lambda/\sqrt{4d}} \\ -q^{(j)} e^{-\lambda/\sqrt{4d}} & (p^{(j)} - q^{(j)}\xi_j) e^{\lambda/\sqrt{4d}} \end{pmatrix} \cdot x_0$$

and

$$\Box_1(\iota_j(g)^{-1} \cdot \sigma_j(\lambda)) = \left\{ q^{(j)} \right\}^2 e^{-\lambda/\sqrt{d}} + \left\{ q^{(j)}\xi_j - p^{(j)} \right\}^2 e^{\lambda/\sqrt{d}} . \tag{23}$$

Let $l+1 \le j \le l+m$. Then, similarly, we obtain

$$\Box_1(\iota_j(g)^{-1} \cdot \sigma_j(\lambda)) = |q^{(j)}|^2 e^{-\lambda/\sqrt{d}} + |q^{(j)}\xi_j - p^{(j)}|^2 e^{\lambda/\sqrt{d}} . \tag{24}$$

Therefore we have

$$\begin{aligned} & b(\gamma^*)(\iota_{\mathbf{k}}(g)^{-1} \cdot \gamma_\xi(\lambda)) \\ &= \frac{1}{\sqrt{d}} \log \left[\left\{ \prod_{j=1}^{l} \left(|q^{(j)}|^2 e^{-\lambda/\sqrt{d}} + |q^{(j)}\xi_j - p^{(j)}|^2 e^{\lambda/\sqrt{d}} \right) \right\} \right. \\ &\qquad \left. \cdot \left\{ \prod_{j=l+1}^{l+m} \left(|q^{(j)}|^2 e^{-\lambda/\sqrt{d}} + |q^{(j)}\xi_j - p^{(j)}|^2 e^{\lambda/\sqrt{d}} \right)^2 \right\} \right] . \end{aligned} \tag{25}$$

In the case $l+m \ge 2$, let $\varepsilon_1, \dots, \varepsilon_{l+m-1}$ be a system of fundamental units of $\mathcal{O}_{\mathbf{k}}$ and let

$$\begin{aligned} C_1 = \max \{ \, & |\iota_j(\varepsilon_i)|^{2(l+m-1)}, |\iota_j(\varepsilon_i)|^{-2(l+m-1)} \\ & | \, i = 1, \dots, l+m-1;\, j = 1, \dots, l+m \, \} . \end{aligned} \tag{26}$$

In the case $l+m = 1$, let $C_1 = 1$.

Proof of Proposition 1.5. — Suppose that $\Pi \circ \gamma_\xi$ is divergent in $\iota_{\mathbf{k}}(\Gamma)\backslash \widetilde{V}$. Then, from Proposition 3.2, for any $\tau > \tau_0$ there exists a positive number $T_0 = T_0(\tau)$ such that the following holds: For any $t \ge T_0$ there exists an element $g \in \bigcup_{j=1}^{h} \Gamma g_j$ satisfying the inequality

$$b(\gamma^*)(\iota_{\mathbf{k}}(g)^{-1} \cdot \gamma_\xi(t)) < -\tau . \tag{27}$$

We show that for any given $\varepsilon > 0$ there exists a positive number $C_0(\varepsilon)$ such that the system of inequalities

$$\|\iota_{\mathbf{k}}(a)\| \leq C, \quad \|\iota_{\mathbf{k}}(a)\xi - \iota_{\mathbf{k}}(b)\| < \varepsilon C^{-1}$$

has a solution $(a, b) \in (\mathcal{O}_{\mathbf{k}})^2$ with $a \neq 0$ for any C greater than $C_0(\varepsilon)$. We remark that it suffices to show this for sufficiently small ε.

Suppose that the given positive number ε is smaller than $C_1 e^{-\tau_0/\sqrt{d}}$. Let

$$\tau = \sqrt{d}\,\log\frac{C_1}{\varepsilon}\ . \tag{28}$$

Since $\tau > \tau_0$, we take the positive number $T_0 = T_0(\tau)$ and put

$$C_0(\varepsilon) = \sqrt{C_1}\,e^{-\tau/\sqrt{4d}}e^{T_0/\sqrt{4d}}\ . \tag{29}$$

We may assume that $C_0(\varepsilon) > 1$ by replacing T_0 with a larger number if necessary. From (29), we have

$$T_0 = 2\sqrt{d}\,\log\frac{C_0(\varepsilon)}{\sqrt{\varepsilon}}\ .$$

For any given C greater than $C_0(\varepsilon)$, we put

$$t = 2\sqrt{d}\,\log\frac{C}{\sqrt{\varepsilon}}\ .$$

Then we have $t > T_0$, and there exists

$$g = \begin{pmatrix} p & r \\ q & s \end{pmatrix} \in \bigcup_{j=1}^{h} \mathit{\Gamma} g_j \subset SL(2, \mathbf{k})$$

such that the inequality (27) holds. We remark that $p, q \in \mathcal{O}_{\mathbf{k}}$.

If $q = 0$, then $p \neq 0$, and it follows from (25) that

$$\begin{aligned} b(\gamma^*)(\iota_{\mathbf{k}}(g)^{-1}\cdot\gamma_\xi(t)) &= \frac{1}{\sqrt{d}}\,\log\left\{\left(\prod_{j=1}^{l}|p^{(j)}|^2\right)\left(\prod_{j=l+1}^{l+m}|p^{(j)}|^4\right)\cdot e^{\sqrt{d}\,t}\right\} \\ &= \frac{1}{\sqrt{d}}\,\log|N_{\mathbf{k}}(p)|^2 + t \geq t > 0 > -\tau\ , \end{aligned}$$

where $N_{\mathbf{k}}(p)$ is the norm of $p \in \mathcal{O}_{\mathbf{k}}$. This is a contradiction. Hence $q \neq 0$.

Let $G = SL(2,\mathbf{R})^l \times SL(2,\mathbf{C})^m$, $K = SO(2)^l \times SU(2)^m$ as in Section 1. Let

$$A = \left\{ (a_1, \ldots, a_{l+m}) \in G \,\middle|\, a_j = \begin{pmatrix} \alpha_j & \\ & \alpha_j^{-1} \end{pmatrix}, \alpha_j > 0 \text{ for each } j \right\}\ , \tag{30}$$

$$N = \left\{ \left(\begin{pmatrix} 1 & \eta_1 \\ 0 & 1 \end{pmatrix}, \ldots, \begin{pmatrix} 1 & \eta_{l+m} \\ 0 & 1 \end{pmatrix} \right) \in G \,\middle|\, \eta_1, \ldots, \eta_l \in \mathbf{R}; \eta_{l+1}, \ldots, \eta_{l+m} \in \mathbf{C} \right\}\ .$$

Then $G = NAK$ is an Iwasawa decomposition of G and $\widetilde{V} = NA \cdot z_0$, where

$$z_0 = (x_0, \dots, x_0, \widehat{x}_0, \dots, \widehat{x}_0) \in \widetilde{V}$$

as in (8). Let

$$A_{\mathbf{Q}} = \left\{ (a, \dots, a) \in A \;\middle|\; a = \begin{pmatrix} \alpha & \\ & \alpha^{-1} \end{pmatrix},\ \alpha > 0 \right\} \tag{31}$$

and

$$A' = \left\{ \left(\begin{pmatrix} \beta_1 & \\ & \beta_1^{-1} \end{pmatrix}, \dots, \begin{pmatrix} \beta_{l+m} & \\ & \beta_{l+m}^{-1} \end{pmatrix} \right) \in A \;\middle|\; (\beta_1 \cdots \beta_l)(\beta_{l+1}^2 \cdots \beta_{l+m}^2) = 1 \right\} .$$

Then $A_{\mathbf{Q}}$ is the topological identity component of the group of real points of a maximal $\mathbf{Q}$-split torus of $\mathbf{G}$, and $A = A' A_{\mathbf{Q}}$ is the topological identity component of the group of real points of a maximal $\mathbf{R}$-torus of G. We also have

$$\gamma_{\mathbf{k}}(\mathbf{R}) = A_{\mathbf{Q}} \cdot z_0 .$$

It follows from (19) that

$$b(\gamma^*)(ua' \cdot z) = b(\gamma^*)(z) \quad \text{for } z \in \widetilde{V},\ u \in N \text{ and } a' \in A' . \tag{32}$$

Let $\iota_{\mathbf{k}}(g)^{-1} \cdot \gamma_\xi(t) = ua'a \cdot z_0$, where

$$\begin{aligned} u &= \left(\begin{pmatrix} 1 & \mu_1 \\ 0 & 1 \end{pmatrix}, \dots, \begin{pmatrix} 1 & \mu_{l+m} \\ 0 & 1 \end{pmatrix} \right) \in N, \\ a' &= \left(\begin{pmatrix} \alpha_1 & \\ & \alpha_1^{-1} \end{pmatrix}, \dots, \begin{pmatrix} \alpha_{l+m} & \\ & \alpha_{l+m}^{-1} \end{pmatrix} \right) \in A', \\ a &= \left(\begin{pmatrix} \beta & \\ & \beta^{-1} \end{pmatrix}, \dots, \begin{pmatrix} \beta & \\ & \beta^{-1} \end{pmatrix} \right) \in A_{\mathbf{Q}} . \end{aligned}$$

For any unit $\nu \in \mathcal{O}_{\mathbf{k}}$, we put

$$k_\nu = \begin{pmatrix} \nu & \\ & \nu^{-1} \end{pmatrix} \in \Gamma$$

and

$$a_\nu = \iota_{\mathbf{k}}(k_\nu) = \left(\begin{pmatrix} \iota_1(\nu) & \\ & \iota_1(\nu)^{-1} \end{pmatrix}, \dots, \begin{pmatrix} \iota_{l+m}(\nu) & \\ & \iota_{l+m}(\nu)^{-1} \end{pmatrix} \right) .$$

Let

$$b'_\nu = \left(\begin{pmatrix} |\iota_1(\nu)| & \\ & |\iota_1(\nu)|^{-1} \end{pmatrix}, \dots, \begin{pmatrix} |\iota_{l+m}(\nu)| & \\ & |\iota_{l+m}(\nu)|^{-1} \end{pmatrix} \right) .$$

Then $b'_\nu \in A'$, since the norm $N_{\mathbf{k}}(a)$ of an element a of $\mathbf{k}$ is given by

$$N_{\mathbf{k}}(a) = \prod_{j=1}^{l} \iota_j(a) \prod_{j=l+1}^{l+m} |\iota_j(a)|^2 .$$

Let $\rho_{\nu,j} = |\iota_j(\nu)|\alpha_j\beta$ for each j and

$$b_\nu = \left(\begin{pmatrix} \rho_{\nu,1} & \\ & \rho_{\nu,1}^{-1} \end{pmatrix}, \dots, \begin{pmatrix} \rho_{\nu,l+m} & \\ & \rho_{\nu,l+m}^{-1} \end{pmatrix} \right) \in A .$$

Since

$$\begin{pmatrix} \iota_j(\nu) & \\ & \iota_j(\nu)^{-1} \end{pmatrix} = \begin{pmatrix} |\iota_j(\nu)| & \\ & |\iota_j(\nu)|^{-1} \end{pmatrix} \begin{pmatrix} \varepsilon_j & \\ & \varepsilon_j^{-1} \end{pmatrix}, \quad \varepsilon_j = \frac{\iota_j(\nu)}{|\iota_j(\nu)|}$$

if $1 \le j \le l$, and

$$\begin{pmatrix} \iota_j(\nu) & \\ & \iota_j(\nu)^{-1} \end{pmatrix} = \begin{pmatrix} |\iota_j(\nu)| & \\ & |\iota_j(\nu)|^{-1} \end{pmatrix} \begin{pmatrix} e^{\theta_j\sqrt{-1}} & \\ & e^{-\theta_j\sqrt{-1}} \end{pmatrix} ,$$

where $\iota_j(\nu) = |\iota_j(\nu)|e^{\theta_j\sqrt{-1}}$, if $l+1 \le j \le l+m$, we have

$$a_\nu a' a \cdot z_0 = b_\nu \cdot z_0 = b'_\nu a' a \cdot z_0 . \tag{33}$$

Let $u' = a_\nu u a_\nu{}^{-1} \in N$. Then it follows from (32) and (33) that

$$\begin{aligned} \iota_{\mathbf{k}}(k_\nu)\iota_{\mathbf{k}}(g)^{-1} \cdot \gamma_\xi(t) &= a_\nu u a' a \cdot z_0 = (a_\nu u a_\nu{}^{-1}) a_\nu a' a \cdot z_0 \\ &= u' b'_\nu a' a \cdot z_0 = u' b_\nu \cdot z_0 \end{aligned}$$

and

$$\begin{aligned} b(\gamma^*)(\iota_{\mathbf{k}}(k_\nu)\iota_{\mathbf{k}}(g)^{-1} \cdot \gamma_\xi(t)) &= b(\gamma^*)(u' b'_\nu a' a \cdot z_0) = b(\gamma^*)(a \cdot z_0) \\ &= b(\gamma^*)(u a' a \cdot z_0) = b(\gamma^*)(\iota_{\mathbf{k}}(g)^{-1} \cdot \gamma_\xi(t)) . \end{aligned} \tag{34}$$

Let $U_{\mathbf{k}}$ be the group of units of $\mathcal{O}_{\mathbf{k}}$ and let

$$\varphi : U_{\mathbf{k}} \longrightarrow W_{\mathbf{k}} = \left\{ (y_1, \dots, y_{l+m}) \in \mathbf{R}^{l+m} \;\middle|\; y_1 + \dots + y_{l+m} = 0 \right\}$$

be a homomorphism defined by

$$\varphi(\nu) = \left(\log|\iota_1(\nu)|, \dots, \log|\iota_l(\nu)|, \log|\iota_{l+1}(\nu)|^2, \dots, \log|\iota_{l+m}(\nu)|^2 \right)$$

for $\nu \in U_{\mathbf{k}}$. Let

$$\begin{aligned} \mathcal{D} = \{ \lambda_1\varphi(\varepsilon_1) + \dots + \lambda_{l+m-1}\varphi(\varepsilon_{l+m-1}) \mid \\ -1/2 \le \lambda_i < 1/2 \quad \text{for } i = 1, \dots, l+m-1 \} . \end{aligned}$$

Then the image $\varphi(U_{\mathbf{k}})$ is a cocompact lattice of $W_{\mathbf{k}}$ with a fundamental domain $\mathcal{D}$ due to Dirichlet's unit theorem. Therefore, there exists a unit ω such that

$$(\log\alpha_1, \dots, \log\alpha_l, 2\log\alpha_{l+1}, \dots, 2\log\alpha_{l+m}) + \varphi(\omega) \in \mathcal{D} .$$

Let

$$C_2 = \max\{\log|\iota_j(\varepsilon_i)|, \; -\log|\iota_j(\varepsilon_i)| \mid i = 1, \dots, l+m-1; j = 1, \dots, l+m\}$$

if $l+m \geq 2$, and let $C_2 = 0$ if $l+m=1$. Then we have $C_1 = e^{2(l+m-1)C_2}$. For each $j = 1, \ldots, l+m$, there exist $\lambda_1, \ldots, \lambda_{l+m-1} \in [-1/2, 1/2)$ such that

$$\log \alpha_j + \log |\iota_j(\omega)| = \sum_{i=1}^{l+m-1} \lambda_i \log |\iota_j(\varepsilon_i)| \,,$$

which implies that

$$|\log(|\iota_j(\omega)|\alpha_j)| \leq \sum_{i=1}^{l+m-1} |\lambda_i| \, |\log|\iota_j(\varepsilon_i)|| \leq \frac{(l+m-1)C_2}{2} \,. \tag{35}$$

Hence we have

$$-(l+m-1)C_2 \leq \log(|\iota_j(\omega)|\alpha_j) - \log(|\iota_{j'}(\omega)|\alpha_{j'}) \leq (l+m-1)C_2$$

and
(36)

$$e^{-(l+m-1)C_2} \leq \frac{\rho_{\omega,j}}{\rho_{\omega,j'}} = \frac{|\iota_j(\omega)|\alpha_j}{|\iota_{j'}(\omega)|\alpha_{j'}} \leq e^{(l+m-1)C_2} \quad \text{for } j, j' \in \{1, \ldots, l+m\} \,.$$

We remark that if $1 \leq j \leq l$, then

$$\iota_j(k_\omega)\iota_j(g)^{-1} \cdot \sigma_j(t) = \begin{pmatrix} \iota_j(\omega)\alpha_j\beta & \mu_j\iota_j(\omega)\alpha_j^{-1}\beta^{-1} \\ 0 & \iota_j(\omega)^{-1}\alpha_j^{-1}\beta^{-1} \end{pmatrix} \cdot x_0$$

and

$$\Box_1(\iota_j(k_\omega)\iota_j(g)^{-1} \cdot \sigma_j(t)) = |\iota_j(\omega)|^{-2}\alpha_j^{-2}\beta^{-2} \,.$$

Similarly, if $l+1 \leq j \leq l+m$, then

$$\Box_1(\iota_j(k_\omega)\iota_j(g)^{-1} \cdot \sigma_j(t)) = |\iota_j(\omega)|^{-2}\alpha_j^{-2}\beta^{-2}.$$

Therefore we have

$$\frac{1}{C_1} \leq \frac{\Box_1(\iota_j(k_\omega)\iota_j(g)^{-1} \cdot \sigma_j(t))}{\Box_1(\iota_{j'}(k_\omega)\iota_{j'}(g)^{-1} \cdot \sigma_{j'}(t))} = \left(\frac{|\iota_{j'}(\omega)|\alpha_{j'}}{|\iota_j(\omega)|\alpha_j}\right)^2 \leq C_1 \tag{37}$$

from (36). This means that

$$\Box_1(\iota_j(k_\omega)\iota_j(g)^{-1} \cdot \sigma_j(t)) \geq \frac{1}{C_1}\left\{\max_{1\leq j'\leq l+m} \Box_1(\iota_{j'}(k_\omega)\iota_{j'}(g)^{-1} \cdot \sigma_{j'}(t))\right\}$$

for each j. We have

$$\begin{aligned} d \cdot \frac{1}{\sqrt{d}} \log\left[\frac{1}{C_1} \max_{1\leq j\leq l+m} \left\{\Box_1(\iota_j(k_\omega)\iota_j(g)^{-1} \cdot \sigma_j(t))\right\}\right] \\ \leq b(\gamma^*)(\iota_{\mathbf{k}}(k_\omega)\iota_{\mathbf{k}}(g)^{-1} \cdot \gamma_\xi(t)) < -\tau \,, \end{aligned}$$

where the first inequality follows from (19) and the second one follows from (27) and (34). This shows that

$$\max_{1\leq j\leq l+m} \left\{\Box_1(\iota_j(k_\omega)\iota_j(g)^{-1} \cdot \sigma_j(t))\right\} < C_1 e^{-\tau/\sqrt{d}} \,.$$

From (23), (24) we have

$$\begin{aligned}(38)\qquad &\Box_1(\iota_j(k_\omega)\iota_j(g)^{-1}\cdot\sigma_j(t))\\ &= |\iota_j(q\omega^{-1})|^2 e^{-t/\sqrt{d}} + |\iota_j(q\omega^{-1})\xi_j - \iota_j(p\omega^{-1})|^2 e^{t/\sqrt{d}} < C_1 e^{-\tau/\sqrt{d}}\end{aligned}$$

for each j. This implies that

$$\|\iota_{\mathbf{k}}(q\omega^{-1})\|^2 e^{-t/\sqrt{d}} = \left(\max_{1\le j\le l+m} |\iota_j(q\omega^{-1})|^2\right) e^{-t/\sqrt{d}} < C_1 e^{-\tau/\sqrt{d}}$$

and

$$\begin{aligned}\|\iota_{\mathbf{k}}(q\omega^{-1})\cdot\xi - \iota_{\mathbf{k}}(p\omega^{-1})\|^2 e^{t/\sqrt{d}} &= \left(\max_{1\le j\le l+m} |\iota_j(q\omega^{-1})\xi_j - \iota_j(p\omega^{-1})|^2\right) e^{t/\sqrt{d}}\\ &< C_1 e^{-\tau/\sqrt{d}}\ .\end{aligned}$$

We obtain

$$\begin{gathered} q\omega^{-1}, p\omega^{-1} \in \mathcal{O}_{\mathbf{k}};\ q\neq 0\ ,\\ \|\iota_{\mathbf{k}}(q\omega^{-1})\| < \sqrt{C_1} e^{-\tau/\sqrt{4d}} e^{t/\sqrt{4d}} = C\end{gathered}$$

and

$$\|\iota_{\mathbf{k}}(q\omega^{-1})\cdot\xi - \iota_{\mathbf{k}}(p\omega^{-1})\| < \sqrt{C_1} e^{-\tau/\sqrt{4d}} e^{-t/\sqrt{4d}} = \frac{\varepsilon}{C}\ .$$

Therefore ξ is a $\mathbf{k}$-singular vector.

Suppose that ξ is $\mathbf{k}$-singular. It suffices to show that, for any $\tau > \tau_0$, there exists a positive number $T_0 = T_0(\tau)$ such that the following holds: For any $t \ge T_0$ there exists a horoball of the form $\iota_{\mathbf{k}}(g)\cdot B(\gamma^*, \tau)$ containing $\gamma(t)$, where g is an element of $\bigcup_{j=1}^h \mathit{\Gamma} g_j$.

We first remark that the following holds.

Lemma 4.1. — *Let $p, q \in \mathcal{O}_{\mathbf{k}}$ and $|p| + |q| \neq 0$. Then there exist $\theta \in \mathbf{k}$ and*

$$g' = \begin{pmatrix} p' & r' \\ q' & s'\end{pmatrix} \in \bigcup_{i=1}^h \mathit{\Gamma} g_i$$

such that $p' = p\theta$, $q' = q\theta$ and $\|\iota_{\mathbf{k}}(\theta)\| \le C_3$, where $C_3 = e^{(l+m-1)C_2/2} = (C_1)^{1/4}$.

Proof. — Since $(p, q) \neq (0)$, there exist a number $j \in \{1, \dots, h\}$ and $\theta' \in \mathbf{k}$ such that $(p, q)(\theta') = \mathfrak{a}_j$. Since the norm of $\mathfrak{a}_j$ is minimum among the integral ideals of its class, we have $|N_{\mathbf{k}}(\theta')| \le 1$. Let $p'' = p\theta'$ and $q'' = q\theta'$, so that $\mathfrak{a}_j = (p'', q'')$. Let $r'', s'' \in (\mathfrak{a}_j)^{-1}$ such that $p''s'' - q''r'' = 1$.

Let $b \in \mathbf{k} - \{0\}$. By the same argument as the one in the proof of Lemma 2.4 of [13], one can find a unit ω of $\mathbf{k}$ such that

$$\|\iota_{\mathbf{k}}(\omega b)\| \le C_3 |N_{\mathbf{k}}(b)|^{1/d}\ .$$

Indeed, for any $(a_1, \dots, a_{l+m}) \in \mathbf{k}_M$ with

$$|(a_1 \cdots a_l)(a_{l+1}^2 \cdots a_{l+m}^2)| = 1,$$

there exists a unit ω' of $\mathbf{k}$ such that

$$|a_j \iota_j(\omega')| \le C_3 \quad \text{for } j = 1, \dots, l+m$$

due to Dirichlet's theorem (cf. (35)). Then it suffices to apply this to

$$\left(\iota_1(b)/|N_{\mathbf{k}}(b)|^{1/d}, \dots, \iota_{l+m}(b)/|N_{\mathbf{k}}(b)|^{1/d}\right) .$$

We take a unit ν such that

$$\|\iota_{\mathbf{k}}(\nu\theta')\| \le C_3 |N_{\mathbf{k}}(\theta')|^{1/d} \le C_3 \tag{39}$$

and put $\theta = \theta'\nu$. Let

$$p' = p\theta = p''\nu, \quad q' = q\theta = q''\nu$$

and

$$r' = r''\nu^{-1}, \quad s' = s''\nu^{-1}, \quad g' = \begin{pmatrix} p' & r' \\ q' & s' \end{pmatrix} .$$

Then

$$g' g_j^{-1} = \begin{pmatrix} \nu p'' d_j - \nu^{-1} r'' b_j & -\nu p'' c_j + \nu^{-1} r'' a_j \\ \nu q'' d_j - \nu^{-1} s'' b_j & -\nu q'' c_j + \nu^{-1} s'' a_j \end{pmatrix} \in SL(2, \mathcal{O}_{\mathbf{k}})$$

and $g' \in \mathit{\Gamma} g_j$. □

If there exist $p, q \in \mathcal{O}_{\mathbf{k}}$ such that $q \ne 0$ and $\|\iota_{\mathbf{k}}(q) \cdot \xi - \iota_{\mathbf{k}}(p)\| = 0$, then take $\theta \in \mathbf{k}$ and $g' \in \bigcup_{i=1}^h \mathit{\Gamma} g_i$ as in Lemma 4.1. We have

$$\begin{aligned} \|\iota_{\mathbf{k}}(q') \cdot \xi - \iota_{\mathbf{k}}(p')\| &= \|\iota_{\mathbf{k}}(\theta)(\iota_{\mathbf{k}}(q) \cdot \xi - \iota_{\mathbf{k}}(p))\| \\ &\le \|\iota_{\mathbf{k}}(\theta)\| \, \|\iota_{\mathbf{k}}(q) \cdot \xi - \iota_{\mathbf{k}}(p)\| \le C_3 \|\iota_{\mathbf{k}}(q) \cdot \xi - \iota_{\mathbf{k}}(p)\| = 0 \end{aligned}$$

and

$$\begin{aligned} b(\gamma^*)(\iota_{\mathbf{k}}(g')^{-1} \cdot \gamma_\xi(\lambda)) &= \frac{1}{\sqrt{d}} \log \left\{ \left(\prod_{i=1}^{l} |(q')^{(i)}|^2 \right) \left(\prod_{i=l+1}^{l+m} |(q')^{(i)}|^4 \right) \cdot e^{-\sqrt{d}\,\lambda} \right\} \\ &= \frac{1}{\sqrt{d}} \log |N_{\mathbf{k}}(q')|^2 - \lambda \end{aligned}$$

by (25). For any $\varepsilon > 0$, we put

$$T_0 = T_0(\tau) = \frac{1}{\sqrt{d}} \log |N_{\mathbf{k}}(q')|^2 + \tau + 1 .$$

Then we have

$$b(\gamma^*)(\iota_{\mathbf{k}}(g')^{-1} \cdot \gamma_\xi(t)) \le \frac{1}{\sqrt{d}} \log |N_{\mathbf{k}}(q')|^2 - T_0 = -\tau - 1 < -\tau$$

and

$$\gamma_\xi(t) \in \iota_{\mathbf{k}}(g') \cdot B(\gamma^*, \tau)$$

if $t \geq T_0$.

Therefore we suppose that $\|\iota_{\mathbf{k}}(q) \cdot \xi - \iota_{\mathbf{k}}(p)\| \neq 0$ for any $p, q \in \mathcal{O}_{\mathbf{k}}$ with $q \neq 0$. Then, for each $\varepsilon > 0$, the set of inequalities

$$0 < \|\iota_{\mathbf{k}}(q)\| \leq C, \quad 0 < \|\iota_{\mathbf{k}}(q) \cdot \xi - \iota_{\mathbf{k}}(p)\| < \varepsilon C^{-1}$$

has a solution $(p, q) \in (\mathcal{O}_{\mathbf{k}})^2$ for all C greater than $C_0(\varepsilon)$.

For a given positive number $\tau > \tau_0$, let

$$\kappa = e^{-\tau/\sqrt{d}}, \quad \varepsilon = \frac{\kappa}{2(C_3)^2} .$$

For this ε, we take the positive number $C_0(\varepsilon)$. Let $C_0'(\varepsilon) = C_0(\varepsilon) + 1$ and

$$T_0 = -\tau + 2\sqrt{d} \log C_0'(\varepsilon) + \sqrt{d} \log \frac{1}{\varepsilon^2} - \sqrt{d} \log 2 - 2\sqrt{d} \log C_3 . \tag{40}$$

In other words,

$$C_0'(\varepsilon) = \sqrt{2} \varepsilon e^{\tau/\sqrt{4d}} e^{T_0/\sqrt{4d}} C_3 .$$

We may assume that $T_0 > 0$ by replacing $C_0(\varepsilon)$ with a larger number if necessary. For a given number $t \geq T_0$, we put

$$C = \sqrt{2} \varepsilon e^{\tau/\sqrt{4d}} e^{t/\sqrt{4d}} C_3 .$$

Then we have $C \geq C_0'(\varepsilon) > C_0(\varepsilon)$, and hence there exist $p, q \in \mathcal{O}_{\mathbf{k}}$ with $q \neq 0$ such that

$$0 < \|\iota_{\mathbf{k}}(q)\| \leq C, \quad 0 < \|\iota_{\mathbf{k}}(q) \cdot \xi - \iota_{\mathbf{k}}(p)\| < \frac{\varepsilon}{C} .$$

Take $\theta \in \mathbf{k}$ and $g' \in \bigcup_{i=1}^h \mathit{\Gamma} g_i$ as in Lemma 4.1. Then we have

$$\begin{aligned} 0 &< \|\iota_{\mathbf{k}}(q') \cdot \xi - \iota_{\mathbf{k}}(p')\| = \|\iota_{\mathbf{k}}(\theta)(\iota_{\mathbf{k}}(q) \cdot \xi - \iota_{\mathbf{k}}(p))\| \\ &\leq \|\iota_{\mathbf{k}}(\theta)\| \, \|\iota_{\mathbf{k}}(q) \cdot \xi - \iota_{\mathbf{k}}(p)\| \leq C_3 \|\iota_{\mathbf{k}}(q) \cdot \xi - \iota_{\mathbf{k}}(p)\| < C_3 \frac{\varepsilon}{C} \end{aligned}$$

and

$$0 < \|\iota_{\mathbf{k}}(q')\| = \|\iota_{\mathbf{k}}(\theta) \iota_{\mathbf{k}}(q)\| \leq \|\iota_{\mathbf{k}}(\theta)\| \, \|\iota_{\mathbf{k}}(q)\| \leq C_3 C .$$

From (25), we have

$$\begin{aligned} &b(\gamma^*)(\iota_{\mathbf{k}}(g')^{-1} \cdot \gamma_\xi(\lambda)) \\ &\qquad \leq \sqrt{d} \log \left\{ \|\iota_{\mathbf{k}}(q')\|^2 e^{-\lambda/\sqrt{d}} + \|\iota_{\mathbf{k}}(q') \cdot \xi - \iota_{\mathbf{k}}(p')\|^2 e^{\lambda/\sqrt{d}} \right\} \end{aligned} \tag{41}$$

for $\lambda > 0$. Let

$$X = e^{\lambda/\sqrt{d}}, \quad \eta = \|\iota_{\mathbf{k}}(q') \cdot \xi - \iota_{\mathbf{k}}(p')\|^2$$

and

$$f(X) = \eta X^2 - \kappa X + \|\iota_{\mathbf{k}}(q')\|^2 \ .$$

Then

$$\|\iota_{\mathbf{k}}(q')\|^2 e^{-t/\sqrt{d}} + \|\iota_{\mathbf{k}}(q') \cdot \xi - \iota_{\mathbf{k}}(p')\|^2 e^{t/\sqrt{d}} < e^{-\tau/\sqrt{d}}$$

if and only if $f(X) < 0$ for $\lambda = t$.

Let $\lambda = t$. Then we have

$$X = e^{t/\sqrt{d}} = \frac{C^2 \kappa}{2\varepsilon^2 (C_3)^2}$$

and

$$f(X) < \left(\frac{C_3 \varepsilon}{C}\right)^2 X^2 - \kappa X + (C_3 C)^2 = \frac{(C_3)^2 \varepsilon^2}{C^2} \cdot \frac{C^2 \kappa}{2\varepsilon^2 (C_3)^2} X - \kappa X + C^2 (C_3)^2$$
$$= -\frac{\kappa}{2} X + C^2 (C_3)^2 = -\frac{\kappa^2 C^2}{4\varepsilon^2 (C_3)^2} + C^2 (C_3)^2 = C^2 \left\{ (C_3)^2 - \frac{\kappa^2}{4\varepsilon^2 (C_3)^2} \right\} = 0 \ .$$

It follows from (41) that

$$b(\gamma^*)(\iota_{\mathbf{k}}(g')^{-1} \cdot \gamma_\xi(t)) \leq \sqrt{d} \log \left\{ \|\iota_{\mathbf{k}}(q')\|^2 e^{-t/\sqrt{d}} + \|\iota_{\mathbf{k}}(q') \cdot \xi - \iota_{\mathbf{k}}(p')\|^2 e^{t/\sqrt{d}} \right\}$$
$$< \sqrt{d} \log e^{-\tau/\sqrt{d}} = -\tau \ .$$

This shows that for any $t \geq T_0$, there exists an element g of $\bigcup_{j=1}^{h} \Gamma g_j$ such that

$$\gamma_\xi(t) \in \iota_{\mathbf{k}}(g) \cdot B(\gamma^*, \tau) \ .$$

Since $\tau > \tau_0$ is arbitrary, it follows from Proposition 3.2 that $\Pi \circ \gamma_\xi$ is divergent. This completes the proof of Proposition 1.5. □

5. Geometric boundary of $(SL(2,\mathrm{R})/SO(2))^l \times (SL(2,\mathrm{C})/SU(2))^m$

5.1. Geometric boundaries of V and $\widehat{V}$. — Let M be an n-dimensional Hadamard manifold and d_M the distance function on M induced from its Riemannian metric. We say that two geodesic rays γ and σ are asymptotic if the convex function $t \longmapsto d_M(\gamma(t), \sigma(t))$ is uniformly bounded on $[0, \infty)$. In this case we also have $d_M(\gamma(t), \sigma(t)) \leq d_M(\gamma(0), \sigma(0))$ for all $t \geq 0$. Being asymptotic is an equivalence relation. Let $M(\infty)$ be the set of equivalence classes of geodesic rays of M. The equivalence class of γ is denoted by $\gamma(\infty)$. If $w = \gamma(\infty)$, we say that γ tends to w, or w is the point at infinity of γ. The union $M \cup M(\infty)$ equipped with the "cone topology" (see [12], [11, §1.7]) is homeomorphic to the n-dimensional closed ball, and this provides a natural compactification of M. The boundary $M(\infty)$ equipped with the restriction of the cone topology, which is called the geometric boundary of M, is homeomorphic to the $(n-1)$-dimensional sphere. Any isometry $F : M \longrightarrow M$ induces a homeomorphism $M(\infty) \longrightarrow M(\infty)$:

$$F(\gamma(\infty)) = (F \circ \gamma)(\infty) \quad \text{for any geodesic ray } \gamma.$$

In the case of V, the geometric boundary $V(\infty)$ can be regarded as the real line $\mathbf{R}$ compactified by adding one point ∞. More precisely, for a real number a, let $\sigma_a : [0,\infty) \longrightarrow V$ be the geodesic ray defined by

(42)
$$\sigma_a(t) = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \begin{pmatrix} e^{-t/2} & \\ & e^{t/2} \end{pmatrix} \cdot x_0 = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \cdot \gamma_0(-t) = a + e^{-t}\sqrt{-1} \ \text{ for } t \geq 0 \ .$$

We also define a geodesic ray $\sigma_\infty : [0,\infty) \longrightarrow V$ by

$$\sigma_\infty(t) = \begin{pmatrix} e^{t/2} & \\ & e^{-t/2} \end{pmatrix} \cdot x_0 = \gamma_0(t) = e^t\sqrt{-1} \quad \text{for } t \geq 0 \ . \tag{43}$$

Then the map $\Phi : \mathbf{R} \cup \{\infty\} \longrightarrow V(\infty)$ defined by

$$\Phi(a) = \sigma_a(\infty) \quad \text{for } a \in \mathbf{R} \cup \{\infty\}$$

is a bijection. We identify $V(\infty)$ with $\mathbf{R} \cup \{\infty\}$ through this map.

Similarly, the geometric boundary of $\widehat{V}$ can be regarded as the complex plane $\mathbf{C}$ compactified by adding one point ∞. For a complex number a, let $\widehat{\sigma}_a$ be the geodesic ray of $\widehat{V}$ defined by

$$\begin{aligned} \widehat{\sigma}_a(t) &= \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \begin{pmatrix} e^{-t/\sqrt{8}} & \\ & e^{t/\sqrt{8}} \end{pmatrix} \cdot \widehat{x}_0 = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \cdot \widehat{\gamma}_0(-t) \\ &= (a,\, e^{-t/\sqrt{2}}) \quad \text{for } \ t \geq 0 \ . \end{aligned} \tag{44}$$

We also define a geodesic ray $\widehat{\sigma}_\infty$ of $\widehat{V}$ by

$$\widehat{\sigma}_\infty(t) = \begin{pmatrix} e^{t/\sqrt{8}} & \\ & e^{-t/\sqrt{8}} \end{pmatrix} \cdot \widehat{x}_0 = \widehat{\gamma}_0(t) = (0,\, e^{t/\sqrt{2}}) \quad \text{for } t \geq 0 \ . \tag{45}$$

Then the map $\widehat{\Phi} : \mathbf{C} \cup \{\infty\} \longrightarrow \widehat{V}(\infty)$ defined by

$$\widehat{\Phi}(a) = \widehat{\sigma}_a(\infty) \quad \text{for } a \in \mathbf{C} \cup \{\infty\}$$

is a bijection. We identify $\widehat{V}(\infty)$ with $\mathbf{C} \cup \{\infty\}$ through this map.

Under the identification $V(\infty) = \mathbf{R} \cup \{\infty\}$, the homeomorphic action of $SL(2,\mathbf{R})$ on $V(\infty)$ induced from the isometric action of $SL(2,\mathbf{R})$ on V is given by

$$g \cdot a = \frac{pa+r}{qa+s} \quad \text{for } a \in \mathbf{R} \cup \{\infty\}\,,\ g = \begin{pmatrix} p & r \\ q & s \end{pmatrix} \in SL(2,\mathbf{R}) \ .$$

Similarly, under the identification $\widehat{V}(\infty) = \mathbf{C} \cup \{\infty\}$, the homeomorphic action of $SL(2,\mathbf{C})$ on $\widehat{V}(\infty)$ induced from the isometric action of $SL(2,\mathbf{C})$ on $\widehat{V}$ is given by

$$g \cdot a = \frac{pa+r}{qa+s} \quad \text{for } a \in \mathbf{C} \cup \{\infty\}\,,\ g = \begin{pmatrix} p & r \\ q & s \end{pmatrix} \in SL(2,\mathbf{C})\,.$$

5.2. Geometric boundary of $\widetilde{V}$. — In order to describe the geometric boundary of $\widetilde{V} = V^l \times \widehat{V}^m$, we recall that any geodesic ray of $\widetilde{V}$ can be written as

$$\gamma(t) = (\gamma_1(a_1 t), \ldots, \gamma_{l+m}(a_{l+m} t))\,, \tag{46}$$

where $a_j \geq 0$ for each j, $\sum_{j=1}^{l+m}(a_j)^2 = 1$, γ_j is a geodesic ray of V if $j \leq l$ and a geodesic ray of $\widehat{V}$ if $j \geq l+1$. By the same argument as that in the proof of [19, Lemma 6.1] we obtain the following.

Lemma 5.1. — *The numbers $a_1, \ldots, a_{l+m}$ in* (46) *depend only on the equivalence class of γ. If $a_j \neq 0$, then the point $\gamma_j(\infty)$ is uniquely determined by the equivalence class of γ.*

Proof. — Let $\gamma' : [0,\infty) \longrightarrow \widetilde{V}$ be another geodesic ray such that $\gamma'(\infty) = \gamma(\infty)$. Let

$$\gamma'(t) = (\gamma'_1(b_1 t), \ldots, \gamma'_{l+m}(b_{l+m} t))\,,$$

where $b_j \geq 0$ for each j, $\sum_{j=1}^{l+m}(b_j)^2 = 1$, γ'_j is a geodesic ray of V if $j \leq l$ and a geodesic ray of $\widehat{V}$ if $j \geq l+1$.

Let $j \leq l$. Then the function

$$t \longmapsto d_V(\gamma_j(a_j t),\, \gamma'_j(b_j t)) \leq d_{\widetilde{V}}(\gamma(t),\, \gamma'(t)) \leq d_{\widetilde{V}}(\gamma(0),\, \gamma'(0))$$

is bounded on $[0,\infty)$. If $a_j > b_j$, then we have

$$d_V(\gamma_j(a_j t),\, \gamma'_j(b_j t)) \geq (a_j - b_j)t - d_V(\gamma_j(0),\, \gamma'_j(0))$$

from the triangle inequalities. This implies that

$$\lim_{t\to\infty} d_V(\gamma_j(a_j t),\, \gamma'_j(b_j t)) = \infty\,,$$

a contradiction. So $a_j \leq b_j$, and similarly $a_j \geq b_j$. Hence we have $a_j = b_j$. Since the function $t \longmapsto d_V(\gamma_j(a_j t),\, \gamma'_j(b_j t))$ is bounded, we have $\gamma_j(\infty) = \gamma'_j(\infty)$ if $a_j \neq 0$.

The case $j \geq l+1$ is similar. □

Let

$$\mathbf{S} = \left\{ \mathbf{a} = (a_1, \ldots, a_{l+m}) \in \mathbf{R}^{l+m} \;\middle|\; \sum_{j=1}^{l+m}(a_j)^2 = 1;\ a_j \geq 0 \text{ for each } j \right\}.$$

This set is a subset of the $(l+m-1)$-dimensional sphere S^{l+m-1} of radius 1, and topologically it can be regarded as an $(l+m-1)$-dimensional simplex. For $\alpha_1, \dots, \alpha_l \in \mathbf{R} \cup \{\infty\}$, $\alpha_{l+1}, \dots, \alpha_{l+m} \in \mathbf{C} \cup \{\infty\}$ and $\mathbf{a} = (a_1, \dots, a_{l+m}) \in \mathbf{S}$, we denote by

$$[\alpha_1, \dots, \alpha_l, \alpha_{l+1}, \dots \alpha_{l+m}; \mathbf{a}]$$

the point at infinity of the geodesic ray

$$[0, \infty) \ni t \longmapsto (\sigma_{\alpha_1}(a_1 t), \dots, \sigma_{\alpha_l}(a_l t), \widehat{\sigma}_{\alpha_{l+1}}(a_{l+1} t), \dots, \widehat{\sigma}_{\alpha_{l+m}}(a_{l+m} t)) \in \widetilde{V} ,$$

where σ_{α_i}, $\widehat{\sigma}_{\alpha_j}$ are as in (42), (43), (44) and (45). Then

$$\begin{aligned}\widetilde{V}(\infty) = \big\{ [\alpha_1, \dots, \alpha_l, \alpha_{l+1}, \dots, \alpha_{l+m}; \mathbf{a}] \;\big|\; & \alpha_1, \dots, \alpha_l \in \mathbf{R} \cup \{\infty\} ; \\ & \alpha_{l+1}, \dots, \alpha_{l+m} \in \mathbf{C} \cup \{\infty\} ; \mathbf{a} \in \mathbf{S} \big\} .\end{aligned}$$

For $\alpha_1, \dots, \alpha_l \in \mathbf{R} \cup \{\infty\}$ and $\alpha_{l+1}, \dots, \alpha_{l+m} \in \mathbf{C} \cup \{\infty\}$, let

$$\mathcal{C}_{\alpha_1, \dots, \alpha_{l+m}} = \{ [\alpha_1, \dots, \alpha_{l+m}; \mathbf{a}] \,|\, \mathbf{a} \in \mathbf{S} \} \subset \widetilde{V}(\infty) .$$

The map $\mathbf{a} \longmapsto [\alpha_1, \dots, \alpha_{l+m}; \mathbf{a}]$ is a bijection from $\mathbf{S}$ to $\mathcal{C}_{\alpha_1, \dots, \alpha_{l+m}}$. We regard the set $\mathcal{C}_{\alpha_1, \dots, \alpha_{l+m}}$ as a topological space equipped with the topology induced from the one of $\mathbf{S}$ through this bijection. Each $\mathcal{C}_{\alpha_1, \dots, \alpha_{l+m}}$ is called a closed Weyl chamber at infinity of $\widetilde{V}$. The geometric boundary $\widetilde{V}(\infty)$ is a union of closed Weyl chambers at infinity as a set:

$$\widetilde{V}(\infty) = \bigcup_{\substack{\alpha_1, \dots, \alpha_l \in \mathbf{R} \cup \{\infty\} \\ \alpha_{l+1}, \dots, \alpha_{l+m} \in \mathbf{C} \cup \{\infty\}}} \mathcal{C}_{\alpha_1, \dots, \alpha_{l+m}} .$$

We first consider the case $l+m = 1$. If $l = 1$, then $\widetilde{V} = V$ and $\mathcal{C}_\alpha$ consists of one point $[\alpha; 1]$ for each $\alpha \in \mathbf{R} \cup \{\infty\}$. Since $\mathcal{C}_\alpha$ meets $\mathcal{C}_\beta$ if and only if $\alpha = \beta$, $V(\infty)$ is a disjoint union of 0-dimensional simplices and can be regarded as a 0-dimensional simplicial complex. Similarly, if $m = 1$, then $\widetilde{V} = \widehat{V}$ and $\mathcal{C}_\alpha$ consists of one point $[\alpha; 1]$ for each $\alpha \in \mathbf{C} \cup \{\infty\}$. Since $\mathcal{C}_\alpha$ meets $\mathcal{C}_\beta$ if and only if $\alpha = \beta$, $\widehat{V}(\infty)$ is a disjoint union of 0-dimensional simplices and can be regarded as a 0-dimensional simplicial complex.

Next, we consider the case $l + m \geq 2$. Then $\widetilde{V}(\infty)$ is obtained by pasting $(l+m-1)$-dimensional simplices together as follows. Let

$$\alpha_1, \dots, \alpha_l, \beta_1, \dots, \beta_l \in \mathbf{R} \cup \{\infty\} ; \alpha_{l+1}, \dots, \alpha_{l+m}, \beta_{l+1}, \dots, \beta_{l+m} \in \mathbf{C} \cup \{\infty\}$$

and

$$\mathbf{a} = (a_1, \dots, a_{l+m}),\ \mathbf{b} = (b_1, \dots, b_{l+m}) \in \mathbf{S} .$$

We remark that $[\alpha_1, \dots, \alpha_{l+m}; \mathbf{a}] = [\beta_1, \dots, \beta_{l+m}; \mathbf{b}]$ if and only if $\mathbf{a} = \mathbf{b}$ and $\alpha_i = \beta_i$ for all i such that $a_i \neq 0$. Let J be the subset of $\{1, \dots, l+m\}$ such that $\alpha_j = \beta_j$ if and only if $j \notin J$. Then we have

$$\begin{aligned}&\mathcal{C}_{\alpha_1,\dots,\,\alpha_{l+m}} \cap \mathcal{C}_{\beta_1,\dots,\,\beta_{l+m}}\\ &= \left\{\, [\alpha_1, \dots, \alpha_{l+m}; (a_1, \dots, a_{l+m})] \in \mathcal{C}_{\alpha_1,\dots,\alpha_{l+m}} \,\middle|\, a_j = 0 \text{ for } j \in J \right\}\\ &= \left\{\, [\beta_1, \dots, \beta_{l+m}; (b_1, \dots, b_{l+m})] \in \mathcal{C}_{\beta_1,\dots,\beta_{l+m}} \,\middle|\, b_j = 0 \text{ for } j \in J \right\} .\end{aligned}$$

In other words, if $\mathcal{C}_{\alpha_1,\dots,\,\alpha_{l+m}}$ meets $\mathcal{C}_{\beta_1,\dots,\,\beta_{l+m}}$, then there exists an integer i such that $\alpha_i = \beta_i$, and the intersection is a boundary simplex of both $\mathcal{C}_{\alpha_1,\dots,\,\alpha_{l+m}}$ and $\mathcal{C}_{\beta_1,\dots,\,\beta_{l+m}}$. If $\mathcal{C}_{\alpha_1,\dots,\,\alpha_{l+m}} \neq \mathcal{C}_{\beta_1,\dots,\,\beta_{l+m}}$, then any point in the interior of $\mathcal{C}_{\alpha_1,\dots,\,\alpha_{l+m}}$ does not belong to $\mathcal{C}_{\beta_1,\dots,\,\beta_{l+m}}$. Thus $\widetilde{V}(\infty)$ becomes an $(l+m-1)$-dimensional simplicial complex.

With $\mathbf{G}$ as defined in Section 3, the simplicial complex $\widetilde{V}(\infty)$ is a geometric realization of the spherical Tits building constructed from the set of all parabolic $\mathbf{R}$-subgroups of $\mathbf{G}$. See [5], [34], [1] and [11] for more information on Tits buildings.

For any subset D of $\widetilde{V}$, let

$$D(\infty) = \left\{\gamma(\infty) \,\middle|\, \gamma \text{ is a geodesic ray of } \widetilde{V} \text{ contained in } D\right\} .$$

We recall that the group A defined by (30) is the topological identity component of the group of real points of a maximal $\mathbf{R}$-split torus of $\mathbf{G}$. For any geodesic ray γ of $\widetilde{V}$ contained in $A \cdot z_0$, there exists a point $(s_1, \dots, s_{l+m})$ of S^{l+m-1} such that the geodesic ray γ' defined by

$$\gamma'(t) = (\gamma_0(s_1 t), \dots, \gamma_0(s_l t), \widehat{\gamma}_0(s_{l+1} t), \dots, \widehat{\gamma}_0(s_{l+m} t)) \quad \text{for } t \geq 0$$

is asymptotic to γ. Let $\varepsilon_i = 0$ if $s_i \leq 0$, $\varepsilon_i = \infty$ if $s_i > 0$. Then $\gamma'(\infty)$ coincides with

$$[\,\varepsilon_1, \dots, \varepsilon_{l+m}; (s_1, \dots, s_{l+m})\,] \in \mathcal{C}_{\varepsilon_1,\dots,\,\varepsilon_{l+m}} .$$

Thus we have

$$(A \cdot z_0)(\infty) = \bigcup_{\varepsilon_i \in \{0,\,\infty\}} \mathcal{C}_{\varepsilon_1,\dots,\,\varepsilon_{l+m}}$$

and $(A \cdot z_0)(\infty)$ consists of 2^{l+m} closed Weyl chambers at infinity. The map

$$\begin{aligned}(A \cdot z_0)(\infty) \ni [\varepsilon_1,\, \dots\,,\, \varepsilon_{l+m}; (s_1, \dots, s_{l+m})\,]&\\ \longmapsto (\delta_1 s_1,\, \dots\,,\, \delta_{l+m} s_{l+m}) \in S^{l+m-1}&\,,\end{aligned} \tag{47}$$

where $\delta_i = 1$ if $\varepsilon_i = \infty$, and $\delta_i = -1$ if $\varepsilon_i = 0$, is a homeomorphism, when we equip $(A \cdot z_0)(\infty)$ with the restriction of the cone topology of $\widetilde{V}(\infty)$.

For $g = (g_1, \dots, g_{l+m}) \in G$, we have

$$g \cdot [\alpha_1,\, \dots\,,\, \alpha_{l+m};\, \mathbf{a}] = [g_1 \cdot \alpha_1,\, \dots\,,\, g_{l+m} \cdot \alpha_{l+m}; \mathbf{a}]$$

and

$$g \cdot \mathcal{C}_{\alpha_1,\dots,\,\alpha_{l+m}} = \mathcal{C}_{g_1\cdot\alpha_1,\dots,\,g_{l+m}\cdot\alpha_{l+m}}\ . \tag{48}$$

Any apartment of the Tits building $\widetilde{V}(\infty)$ is of the form

$$(g \cdot (A \cdot z_0))(\infty) = g \cdot ((A \cdot z_0)(\infty))$$

for some $g \in G$. The apartment $(g \cdot (A \cdot z_0))(\infty)$ consists of 2^{l+m} closed Weyl chambers at infinity:

$$(g \cdot (A \cdot z_0))(\infty) = \bigcup_{\varepsilon_i \in \{0,\,\infty\}} g \cdot \mathcal{C}_{\varepsilon_1,\dots,\,\varepsilon_{l+m}}\ .$$

For any two closed Weyl chambers $\mathcal{V}$, $\mathcal{V}'$ at infinity, there exists an apartment containing both $\mathcal{V}$ and $\mathcal{V}'$.

5.3. Tits metric. — For $\mathbf{a} = (a_1, \dots, a_{l+m}), \mathbf{b} = (b_1, \dots, b_{l+m}) \in \mathbf{S}$, we define $d_{\mathbf{S}}(\mathbf{a}, \mathbf{b})$ to be the minimum of the lengths of piecewise smooth curves in $\mathbf{S}$ joining $\mathbf{a}$ and $\mathbf{b}$, where the lengths of curves are measured in $\mathbf{R}^{l+m}$. Then $d_{\mathbf{S}}$ is a distance function on $\mathbf{S}$. In other words, $d_{\mathbf{S}}(\mathbf{a}, \mathbf{b})$ is the angle between $\mathbf{a}$ and $\mathbf{b}$, or the unique number $\theta \in [0, \pi]$ such that

$$\cos\theta = \sum_{j=1}^{l+m} a_j b_j\ .$$

On each $\mathcal{C}_{\alpha_1,\dots,\alpha_{l+m}}$ we have a distance induced from $d_{\mathbf{S}}$ under the identification of $\mathcal{C}_{\alpha_1,\dots,\alpha_{l+m}}$ with $\mathbf{S}$. This distance is extended to the distance on the whole $\widetilde{V}(\infty)$ in the usual manner by considering the lengths of curves in this connected simplicial complex if $l + m \geq 2$. We denote the resultant distance by Td. By the map (47), $(A \cdot z_0)(\infty)$ equipped with the restriction of Td is isometric to the sphere S^{l+m-1} equipped with the distance determined by the induced metric from the standard flat metric of $\mathbf{R}^{l+m}$.

If $l + m = 1$, then for $v, w \in \widetilde{V}(\infty)$, we have

$$Td(v, w) = \begin{cases} 0 & \text{if } v = w \\ \infty & \text{if } v \neq w\ . \end{cases} \tag{49}$$

In general, on the geometric boundary of a Hadamard manifold M, there exists a distance called the "Tits metric," which is preserved by the homeomorphism of $M(\infty)$ induced from any isometry of M (see [1, Section 4], [11, Chapter 3]). From Proposition 3.1.2 and Proposition 3.4.3 (2) of [11], the restriction of the Tits metric on $\widetilde{V}(\infty)$ to each apartment coincides with the restriction of the above distance Td to the apartment. Hence Td is the Tits metric on $\widetilde{V}(\infty)$.

6. Limit points of discrete subgroups of $SL(2, \mathrm{R})^l \times SL(2, \mathrm{C})^m$

We first recall the definitions of limit points and conical limit points.

Definition 6.1 (cf. [11, Definition 1.9.5]). — Let Λ be a discrete subgroup of G. We denote by $\overline{\Lambda \cdot z_0}$ the closure in $\widetilde{V} \cup \widetilde{V}(\infty)$ of the Λ-orbit of z_0 (with respect to the cone topology). We let $L(\Lambda) = \overline{\Lambda \cdot z_0} \cap \widetilde{V}(\infty)$ and call this set the limit set of Λ. We also call each point of $L(\Lambda)$ a limit point of Λ.

Remark 6.2. — Since $\iota_{\mathbf{k}}(\Gamma) = \iota_{\mathbf{k}}(SL(2, \mathcal{O}_{\mathbf{k}}))$ is a nonuniform lattice of G, we have $L(\Lambda) = \widetilde{V}(\infty)$ if $\Lambda = \iota_{\mathbf{k}}(\Gamma)$ (see [11, Proposition 1.9.32]).

Definition 6.3 (cf. [19, Definition B]). — Let Λ be a discrete subgroup of G. We say that $v \in \widetilde{V}(\infty)$ is a conical limit point of Λ if for some (and hence every) geodesic ray γ tending to v, there exists a sequence $(g_i)_{i \in \mathbf{N}}$ of different elements of Λ such that $d_{\widetilde{V}}(g_i \cdot z_0, \gamma([0, \infty)))$ are uniformly bounded.

The following is well known to experts. However, we provide a proof for the sake of completeness.

Proposition 6.4. — *Let $\gamma : [0, \infty) \longrightarrow \widetilde{V}$ be a geodesic ray. Then $\Pi \circ \gamma$ is divergent if and only if $\gamma(\infty)$ is not a conical limit point of $\iota_{\mathbf{k}}(\Gamma)$.*

Proof. — Let $\overline{z}$ be the image under the projection Π of $z \in \widetilde{V}$ and let

$$\overline{d} : (\iota_{\mathbf{k}}(\Gamma) \backslash \widetilde{V}) \times (\iota_{\mathbf{k}}(\Gamma) \backslash \widetilde{V}) \longrightarrow \mathbf{R}$$

be the distance on $\iota_{\mathbf{k}}(\Gamma) \backslash \widetilde{V}$ defined by

$$\overline{d}(\overline{z}, \overline{w}) = \inf_{g \in \Gamma} d_{\widetilde{V}}(z, \iota_{\mathbf{k}}(g) \cdot w) \quad \text{for } z, w \in \widetilde{V} .$$

Suppose that $\gamma(\infty)$ is a conical limit point. Then there exist a positive number C and a sequence $(g_j)_{j \in \mathbf{N}}$ of different elements of Γ such that

$$d_{\widetilde{V}}(\iota_{\mathbf{k}}(g_j) \cdot z_0, \gamma([0, \infty))) \leq C .$$

This means that there exists a sequence $(t_j)_{j \in \mathbf{N}}$ in $\mathbf{R}$ such that

$$d_{\widetilde{V}}(\iota_{\mathbf{k}}(g_j) \cdot z_0, \gamma(t_j)) = d_{\widetilde{V}}(\iota_{\mathbf{k}}(g_j) \cdot z_0, \gamma([0, \infty))) \leq C$$

for each j. Since Γ acts on $\widetilde{V}$ properly discontinuously, we have

$$\lim_{j \to \infty} d_{\widetilde{V}}(\iota_{\mathbf{k}}(g_j) \cdot z_0, z_0) = \infty$$

and $\lim_{j \to \infty} t_j = \infty$. Let

$$D = \left\{ \overline{x} \in \iota_{\mathbf{k}}(\Gamma) \backslash \widetilde{V} \,\middle|\, \overline{d}(\overline{x}, \overline{z}_0) \leq C \right\} .$$

Then D is compact and

$$\Pi \circ \gamma(t_j) = \Pi(\iota_{\mathbf{k}}(g_j)^{-1} \cdot \gamma(t_j)) \in D$$

for each j. Hence $\Pi \circ \gamma$ is not divergent.

Suppose that $\Pi \circ \gamma$ is not divergent. Then there exist a compact subset D of $\iota_{\mathbf{k}}(\Gamma)\backslash\widetilde{V}$ and a sequence $(t_j)_{j\in\mathbf{N}}$ in $\mathbf{R}$ such that $\lim_{j\to\infty} t_j = \infty$ and $\Pi\circ\gamma(t_j) \in D$ for each j. Since D is compact, there exists a positive number C such that

$$\overline{d}(\Pi\circ\gamma(t_j),\, \Pi(z_0)) \leq C \quad \text{for all } j\ .$$

This means that there exists a sequence $(g_j)_{j\in\mathbf{N}}$ in Γ such that

$$d_{\widetilde{V}}(z_0,\ \iota_{\mathbf{k}}(g_j)^{-1}\cdot\gamma(t_j)) = \overline{d}(\Pi(z_0),\, \Pi\circ\gamma(t_j)) \leq C\ ,$$

which shows that

$$d_{\widetilde{V}}(\iota_{\mathbf{k}}(g_j)\cdot z_0,\, \gamma([0,\infty))) \leq d_{\widetilde{V}}(\iota_{\mathbf{k}}(g_j)\cdot z_0,\, \gamma(t_j)) \leq C$$

for each j, and the sequence $(g_j)_{j\in\mathbf{N}}$ contains infinitely many different elements of Γ because

$$d_{\widetilde{V}}(\iota_{\mathbf{k}}(g_j)\cdot z_0,\, z_0) \geq d_{\widetilde{V}}(\gamma(t_j),\, z_0) - d_{\widetilde{V}}(\iota_{\mathbf{k}}(g_j)\cdot z_0,\, \gamma(t_j)) \geq t_j - C\ .$$

Hence $\gamma(\infty)$ is a conical limit point. □

The spherical Tits building constructed from the set of all parabolic $\mathbf{Q}$-subgroups of $\mathbf{G}$ is naturally realized as a subset of $\widetilde{V}(\infty)$. We recall that the subgroup $A_{\mathbf{Q}}$ of A defined by (31) is the topological identity component of the group of real points of a maximal $\mathbf{Q}$-split torus of $\mathbf{G}$. The group

$$G_{\mathbf{Q}} = \left\{(g_1,\ldots,g_{l+m}) \in G \,\middle|\, g_1 \in SL(2,\mathbf{k}); g_j = \iota_j(g_1) \text{ for each } j\right\}$$

is isomorphic to the group of rational points of $\mathbf{G}$. Let

$$(A_{\mathbf{Q}})^+ = \left\{(a,\ldots,a)\in G \,\middle|\, a = \begin{pmatrix}\alpha & \\ & \alpha^{-1}\end{pmatrix}, \alpha \geq 1 \text{ for each } j\right\} \subset A_{\mathbf{Q}}\ .$$

Definition 6.5 (cf. [19, Definition 2.2]). — We call the set $(g\cdot((A_{\mathbf{Q}})^+\cdot z_0))(\infty)$ a closed $\mathbf{Q}$-Weyl chamber at infinity if $g \in G_{\mathbf{Q}}$. We denote by $\mathcal{W}_{\mathbf{Q}}$ the union of all the closed $\mathbf{Q}$-Weyl chambers at infinity, which we call the rational Tits building of $\mathbf{G}$:

$$\mathcal{W}_{\mathbf{Q}} = \bigcup_{g\in G_{\mathbf{Q}}} (g\cdot((A_{\mathbf{Q}})^+\cdot z_0))(\infty)\ .$$

Let

$$\mathbf{a}_0 = \left(\frac{1}{\sqrt{d}},\ldots,\frac{1}{\sqrt{d}},\frac{\sqrt{2}}{\sqrt{d}},\ldots,\frac{\sqrt{2}}{\sqrt{d}}\right) \in \mathbf{S}\ ,$$

where the first l entries of $\mathbf{a}_0$ are $1/\sqrt{d}$ and the last m entries are $\sqrt{2}/\sqrt{d}$. From (17), (43) and (45) we have

$$((A_{\mathbf{Q}})^+\cdot z_0)(\infty) = \{\,[\infty,\ldots,\infty;\mathbf{a}_0]\,\}\ .$$

We also have

$$g \cdot [\infty, \dots, \infty; \mathbf{a}_0] = [\xi^{(1)}, \dots, \xi^{(l+m)}; \mathbf{a}_0]$$

if

$$g = (g_1, \dots, g_{l+m}) \in G_{\mathbf{Q}}, \quad g_1 = \begin{pmatrix} p & r \\ q & s \end{pmatrix} \in SL(2, \mathbf{k}), \quad q \neq 0 \quad \text{and} \ \xi = \frac{p}{q} \ .$$

Hence we obtain

PROPOSITION 6.6. —

$$\mathcal{W}_{\mathbf{Q}} = \left\{ [\xi^{(1)}, \dots, \xi^{(l+m)}; \mathbf{a}_0] \,\middle|\, \xi \in \mathbf{k} \right\} \cup \{[\infty, \dots, \infty; \mathbf{a}_0]\} \ .$$

PROPOSITION 6.7. — *The set of all conical limit points of $\iota_{\mathbf{k}}(\Gamma)$ is*

$$\left\{ v \in \widetilde{V}(\infty) \,\middle|\, Td(v, \mathcal{W}_{\mathbf{Q}}) \geq \pi/2 \right\} .$$

Proof. — In the case $l + m \geq 2$, this follows from Theorems A and B of [19], since $\mathbf{G}$ is of $\mathbf{Q}$-rank 1 and $\iota_{\mathbf{k}}(\Gamma)$ is an irreducible nonuniform lattice.

In the case $l+m = 1$, this follows from (49) and Proposition 6.6, since $\gamma_\xi(\infty)$ is conical if and only if $\xi \notin \mathbf{k}$. □

Theorem 1.7 follows from Propositions 1.5, 6.4 and 6.7.

7. Proof of Theorem 1.3

In this section we prove Theorem 1.3.

Let $\xi = (\xi_1, \dots, \xi_{l+m}) \in \mathbf{k}_M$. From Theorem 1.7 it suffices to show that $f_{\mathbf{k}}(\xi) < d/2$ if and only if $Td(\gamma_\xi(\infty), \mathcal{W}_{\mathbf{Q}}) < \pi/2$.

The case $l + m = 1$ is immediate from (49) and Theorem 1.7. For the case $l + m \geq 2$, it suffices to prove the following proposition.

PROPOSITION 7.1. — *Let $l + m \geq 2$ and $\xi \in \mathbf{k}_M$. Then we have*

$$\cos\left(Td(\gamma_\xi(\infty), \mathcal{W}_{\mathbf{Q}})\right) = \frac{d - 2f_{\mathbf{k}}(\xi)}{d} \ . \tag{50}$$

Proof. — Since the point $\gamma_\xi(\infty)$ and all the points of $\mathcal{W}_{\mathbf{Q}}$ are of the form

$$[\alpha_1, \dots, \alpha_{l+m}; \mathbf{a}_0]; \ \alpha_1, \dots, \alpha_l \in \mathbf{R} \cup \{\infty\}; \ \alpha_{l+1}, \dots, \alpha_{l+m} \in \mathbf{C} \cup \{\infty\} \ ,$$

the set

$$\{Td(\gamma_\xi(\infty), v) \,|\, v \in \mathcal{W}_{\mathbf{Q}}\}$$

is a discrete subset of $[0, \infty)$. Hence there exists a point v' of $\mathcal{W}_{\mathbf{Q}}$ such that

$$Td(\gamma_\xi(\infty), v') = Td(\gamma_\xi(\infty), \mathcal{W}_{\mathbf{Q}}) \ .$$

Then

$$v' = [\eta^{(1)}, \dots, \eta^{(l+m)}; \mathbf{a}_0] \quad \text{for some } \eta \in \mathbf{k}$$

or $v' = [\infty, \dots, \infty; \mathbf{a}_0]$ by Proposition 6.6.

In the case where $v' = [\infty, \dots, \infty; \mathbf{a}_0]$, take

$$g' = \begin{pmatrix} a & c \\ b & d \end{pmatrix} \in SL(2, \mathbf{Z})$$

such that $b \neq 0$ and

$$b\xi_i + d \neq 0 \quad \text{for } i = 1, \dots, l+m \ .$$

Let

$$\xi' = (\xi'_1, \dots, \xi'_{l+m}) \in \mathbf{k}_M; \ \xi'_i = \frac{a\xi_i + c}{b\xi_i + d} \quad \text{for } i = 1, \dots, l+m \ .$$

Then it is easily seen that $f_{\mathbf{k}}(\xi') = f_{\mathbf{k}}(\xi)$. Let $v'' = \iota_{\mathbf{k}}(g') \cdot v'$. Then we have

$$\gamma_{\xi'}(\infty) = \iota_{\mathbf{k}}(g') \cdot \gamma_\xi(\infty)$$

and

$$v'' = [a/b, \dots, a/b; \mathbf{a}_0] \ .$$

Since $\iota_{\mathbf{k}}(g') \cdot \mathcal{W}_{\mathbf{Q}} = \mathcal{W}_{\mathbf{Q}}$, we have

$$\begin{aligned} &Td(\gamma_{\xi'}(\infty), v'') = Td(\iota_{\mathbf{k}}(g') \cdot \gamma_\xi(\infty), \iota_{\mathbf{k}}(g') \cdot v') = Td(\gamma_\xi(\infty), v') \\ &= Td(\gamma_\xi(\infty), \mathcal{W}_{\mathbf{Q}}) = Td(\iota_{\mathbf{k}}(g') \cdot \gamma_\xi(\infty), \iota_{\mathbf{k}}(g') \cdot \mathcal{W}_{\mathbf{Q}}) = Td(\gamma_{\xi'}(\infty), \mathcal{W}_{\mathbf{Q}}) \ . \end{aligned}$$

Hence this case is reduced to the first case.

Suppose that

$$v' = [\eta^{(1)}, \dots, \eta^{(l+m)}; \mathbf{a}_0], \quad \eta \in \mathbf{k} \ .$$

Take the subsets $I_1 \subset \mathbf{N}(l)$, $I_2 \subset \mathbf{N}(l+m) \setminus \mathbf{N}(l)$ such that

$$\xi_k \neq \eta^{(k)} \quad \text{if and only if } k \in I_1 \cup I_2 \ .$$

Let $\lambda_0 = \# I_1$, $\mu_0 = \# I_2$,

$$\nu = \iota_{\mathbf{k}}(\eta) = (\eta^{(1)}, \dots, \eta^{(l+m)}) \in \iota_{\mathbf{k}}(\mathbf{k}) \ ,$$

and

$$\mathcal{V} = \mathcal{C}_{\xi_1, \dots, \xi_{l+m}}, \quad \mathcal{V}' = \mathcal{C}_{\eta^{(1)}, \dots, \eta^{(l+m)}} \ .$$

Then $\gamma_\xi(\infty) \in \mathcal{V}$ and $v' = \gamma_\nu(\infty) \in \mathcal{V}'$. We take an apartment $\mathcal{A}$ that contains both $\mathcal{V}$ and $\mathcal{V}'$. Suppose that

$$\mathcal{A} = g \cdot ((A \cdot z_0)(\infty)), \ \text{ where } \ g = (g_1, \dots, g_{l+m}) \in G \ .$$

For each j, let

$$g_j \cdot \infty = \omega_j^\infty, \quad g_j \cdot 0 = \omega_j^0 \ .$$

It follows from (48) that

$$\mathcal{A} = \bigcup_{\varepsilon_i \in \{0,\infty\}} g \cdot \mathcal{C}_{\varepsilon_1,\dots,\,\varepsilon_{l+m}} = \bigcup_{\varepsilon_i \in \{0,\infty\}} \mathcal{C}_{g_1 \cdot \varepsilon_1,\dots,\, g_{l+m}\cdot \varepsilon_{l+m}}$$
$$= \bigcup_{\varepsilon_i \in \{0,\infty\}} \mathcal{C}_{\omega_1^{\varepsilon_1},\dots,\,\omega_{l+m}^{\varepsilon_{l+m}}} .$$

Suppose that

$$\mathcal{V} = \mathcal{C}_{\omega_1^{\nu_1},\dots,\,\omega_{l+m}^{\nu_{l+m}}}$$

and let $g' = (g'_1, \dots, g'_{l+m})$ be the element of G such that

$$g'_i = g_i \text{ if } \nu_i = \infty \quad \text{and} \quad g'_i = g_i \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \text{ if } \nu_i = 0 .$$

Then $\mathcal{A} = g' \cdot ((A \cdot z_0)(\infty))$. Since

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \cdot 0 = \infty, \text{ hence } g_j \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \cdot 0 = \omega_j^{\infty} ,$$

we may suppose that

$$\mathcal{V} = \mathcal{C}_{\xi_1,\dots,\,\xi_{l+m}} = \mathcal{C}_{\omega_1^{\infty},\dots,\,\omega_{l+m}^{\infty}}$$

by replacing g with g' if necessary. Then, under the identification of $(A \cdot z_0)(\infty)$ with S^{l+m-1} by the map (47), we have

$$g^{-1} \cdot \mathcal{V} = \mathcal{C}_{\infty,\dots,\,\infty} = \mathbf{S}, \qquad g^{-1} \cdot \gamma_\xi(\infty) = [\infty, \dots,\, \infty; \mathbf{a}_0] = \mathbf{a}_0 .$$

The point $g^{-1} \cdot v'$ belongs to the subset

$$\left\{ \left(\frac{\varepsilon_1}{\sqrt{d}}, \dots, \frac{\varepsilon_l}{\sqrt{d}}, \frac{\varepsilon_{l+1}\sqrt{2}}{\sqrt{d}}, \dots, \frac{\varepsilon_{l+m}\sqrt{2}}{\sqrt{d}} \right) \,\middle|\, \varepsilon_j = \pm 1 \text{ for } j = 1, \dots, l+m \right\}$$

of S^{l+m-1}. Suppose that

$$g^{-1} \cdot v' = \mathbf{x} = \left(\frac{\delta_1}{\sqrt{d}}, \dots, \frac{\delta_l}{\sqrt{d}}, \frac{\delta_{l+1}\sqrt{2}}{\sqrt{d}}, \dots, \frac{\delta_{l+m}\sqrt{2}}{\sqrt{d}} \right) .$$

Then λ_0 coincides with the cardinality of the set $\{i \,|\, \delta_i = -1,\ 1 \le i \le l\}$, and μ_0 coincides with the cardinality of the set $\{i \,|\, \delta_i = -1,\ l+1 \le i \le l+m\}$. Let θ be the angle between $\mathbf{x}$ and $\mathbf{a}_0$. Then we have

$$\theta = Td(\gamma_\xi(\infty), v') = Td(\gamma_\xi(\infty), \mathcal{W}_{\mathbf{Q}})$$

and

$$\cos\theta = \frac{l - 2\lambda_0 + 2(m - 2\mu_0)}{d} = \frac{d - 2(\lambda_0 + 2\mu_0)}{d} . \tag{51}$$

We claim that $f_{\mathbf{k}}(\xi) = \lambda_0 + 2\mu_0$. If not, there exist integers λ'_0, μ'_0, subsets $I'_1 \subset \mathbf{N}(l)$, $I'_2 \subset \mathbf{N}(l+m) \setminus \mathbf{N}(l)$, and an element η' of $\mathbf{k}$ such that

$$\lambda'_0 = \# I'_1, \quad \mu'_0 = \# I'_2, \quad \lambda'_0 + 2\mu'_0 < \lambda_0 + 2\mu_0$$

and

$$\xi_k \neq (\eta')^{(k)} \quad \text{if and only if } k \in I_1' \cup I_2' \,.$$

Let

$$\xi'' = \iota_{\mathbf{k}}(\eta') = ((\eta')^{(1)}, \ldots, (\eta')^{(l+m)}) \in \iota_{\mathbf{k}}(\mathbf{k}) \,,$$
$$w = \gamma_{\xi''}(\infty) = [(\eta')^{(1)}, \ldots, (\eta')^{(l+m)}; \mathbf{a}_0] \in \mathcal{W}_{\mathbf{Q}}$$

and $\theta' = Td(\gamma_\xi(\infty), w)$. Repeating the same argument, we obtain

$$\cos\theta' = \frac{d - 2(\lambda_0' + 2\mu_0')}{d} \,.$$

This shows that $\cos\theta' > \cos\theta$ and

$$Td(\gamma_\xi(\infty), w) = \theta' < \theta = Td(\gamma_\xi(\infty), \mathcal{W}_{\mathbf{Q}}) \,,$$

which is a contradiction. Therefore we obtain $f_{\mathbf{k}}(\xi) = \lambda_0 + 2\mu_0$.

The equation (50) now follows from (51). □

8. Proof of Theorem 1.4

We can now prove Theorem 1.4.

First, we remark that d' is the maximum value of the elements of the set

$$\{\#I_1 + 2\cdot(\#I_2) \,|\, I_1 \subset \mathbf{N}(l),\, I_2 \subset \mathbf{N}(l+m) \smallsetminus \mathbf{N}(l)\} \cap [0,\, d/2) \,.$$

For $\eta \in \mathbf{k}$ and subsets $I_1 \subset \mathbf{N}(l)$, $I_2 \subset \mathbf{N}(l+m) \smallsetminus \mathbf{N}(l)$, let

$$\mathcal{S}(\eta; I_1, I_2) = \left\{\xi = (\xi_1, \ldots, \xi_{l+m}) \in \mathbf{k}_M \,|\, \xi_i = \eta^{(i)} \text{ for } i \notin I_1 \cup I_2\right\} \,.$$

For integers λ, μ such that $0 \le \lambda \le l$, $0 \le \mu \le m$, let

$$\mathcal{S}(\eta; \lambda, \mu) = \bigcup_{\#I_1=\lambda,\, \#I_2=\mu} \mathcal{S}(\eta; I_1, I_2) \,.$$

We define

$$\mathcal{S}(\eta) = \bigcup_{\lambda+2\mu=d'} \mathcal{S}(\eta; \lambda, \mu), \quad \mathcal{S}(\mathbf{k}) = \bigcup_{\eta\in\mathbf{k}} \mathcal{S}(\eta) \,.$$

Then we have

$$\mathrm{Sing}(\mathbf{k}) = \mathcal{S}(\mathbf{k}) \,.$$

Since

$$\dim_{\mathrm{H}}(\mathcal{S}(\eta; I_1, I_2)) = \#I_1 + 2\cdot(\#I_2) \,,$$

we obtain

$$\dim_{\mathrm{H}}(\mathcal{S}(\eta; \lambda, \mu)) = \lambda + 2\mu, \quad \dim_{\mathrm{H}}(\mathcal{S}(\eta)) = d' \,.$$

This shows that

$$\dim_{\mathrm{H}}(\mathrm{Sing}(\mathbf{k})) = \dim_{\mathrm{H}}(\mathcal{S}(\mathbf{k})) = d' ,$$

because $\mathbf{k}$ is a countable set.

We have $\mathrm{Sing}(\mathbf{k}) = \iota_{\mathbf{k}}(\mathbf{k})$ if and only if $d' = 0$, and the latter is equivalent to the condition that $\mathbf{k}$ is the rational field $\mathbf{Q}$ or a quadratic field or a totally complex quartic field. This completes the proof.

Acknowledgements. — The author would like to thank the anonymous referee for the careful reading of the manuscript and many valuable suggestions.

BIBLIOGRAPHY

[1] W. Ballmann, M. Gromov & V. Schroeder – *Manifolds of Nonpositive Curvature*, Progress in Math., no. 61, Birkhäuser, Boston, 1985.

[2] A. Borel – "Ensembles fondamentaux pour les groupes arithmétiques", in *Colloque sur la théorie des groupes algébriques*, Librairie Gauthier-Villars, Paris, 1962, p. 23–40.

[3] ———, *Introduction aux groupes arithmétiques*, Hermann, Paris, 1969.

[4] É. Borel – "Les probabilités dénombrables et leurs applications arithmétiques", *Rend. Circ. Math. Palermo* **27** (1909), p. 247–271.

[5] K. S. Brown – *Buildings*, Springer-Verlag, Berlin, 1991.

[6] E. Burger – "Homogeneous Diophantine approximation in S-integers", *Pacific J. Math.* **152** (1992), p. 211–253.

[7] J. W. S. Cassels – *An Introduction to Diophantine Approximation*, Cambridge Univ. Press, 1957.

[8] S. G. Dani – "Divergent trajectories of flows on homogeneous spaces and Diophantine approximation", *J. reine angew. Math.* **359** (1985), p. 55–89.

[9] ———, "Convergents as approximants in continued fraction expansions of complex numbers with Eisenstein integers", 2017, arXiv:1703.07672.

[10] M. M. Dodson & S. Kristensen – "Hausdorff dimension and Diophantine approximation", Proc. Sympos. Pure Math., no. 72, p. 305–347, Amer. Math. Soc., Providence, RI, 2004.

[11] P. Eberlein – *Geometry of Nonpositively Curved Manifolds*, The University of Chicago Press, Chicago, 1996.

[12] P. Eberlein & B. O'Neil – "Visibility manifolds", *Pacific J. Math.* **46** (1973), p. 45–109.

[13] M. Einsiedler, A. Ghosh & B. Lytle – "Badly approximable vectors, C^1 curves and number fields", *Ergodic Theory Dynam. Systems* **36** (2016), p. 1851–1864.

[14] J. Elstrodt, F. Grunewald & J. Mennicke – *Groups Acting on Hyperbolic Space: Harmonic Analysis and Number Theory*, Springer-Verlag, Berlin, 1998.

[15] R. Esdahl-Schou & S. Kristensen – "On badly approximable complex numbers", *Glasg. Math. J.* **52** (2010), p. 349–355.

[16] L. Fishman, D. Simmons & M. Urbański – "Diophantine approximation and the geometry of limit sets in Gromov hyperbolic metric spaces", *Mem. Amer. Math. Soc.* **254** (2018), no. 1215.

[17] L. R. Ford – "A geometric proof of a theorem of Hurwitz", *Proc. Edinburgh Math. Soc* **35** (1917), p. 59–65.

[18] T. Hattori – "Collapsing of quotient spaces of $SO(n) \setminus SL(n, \mathbf{R})$ at infinity", *J. Math. Soc. Japan* **47** (1995), p. 193–225.

[19] ———, "Geometric limit sets of higher rank lattices", *Proc. London Math. Soc.* **90** (2005), p. 689–710.

[20] ———, "Some Diophantine approximation inequalities and products of hyperbolic spaces", *J. Math. Soc. Japan* **59** (2007), p. 239–264.

[21] ———, "Shimizu's lemma for classical groups", *Proc. London Math. Soc.* **117** (2018), p. 1281–1320.

[22] S. Helgason – *Differential Geometry, Lie Groups, and Symmetric Spaces*, Academic Press, New York, 1978.

[23] S. Hersonsky & F. Paulin – "Diophantine approximation on negatively curved manifolds and in the Heisenberg group", in *Rigidity in Dynamics and Geometry*, Springer-Verlag, Berlin, 2002, p. 203–226.

[24] R. Hines – "Examples of badly approximable vectors over number fields", 2018, arXiv:1809.07404.

[25] ———, "Badly approximable numbers over imaginary quadratic fields", *Acta Arith.* **190** (2019), p. 102–125.

[26] V. Jarník – "Zur metrischen Theorie der diophantischen Approximationen", *Prace Mat.-Fiz.* **36** (1928), p. 91–106.

[27] D. Kleinbock & T. Ly – "Badly approximable S-numbers and absolute Schmidt games", *J. Number Theory* **164** (2016), p. 13–42.

[28] F. Maucourant – "Arithmetical and geometrical aspects of homogeneous Diophantine approximation by algebraic numbers in a given number field", Sémin. Congr., no. 19, p. 37–48, Soc. Math. France, 2009.

[29] G. Prasad – "Strong rigidity of $\mathbf{Q}$-rank 1 lattices", *Invent. math.* **21** (1973), p. 255–286.

[30] R. Quême – "On Diophantine approximatiom by algebraic numbers of a given number field: a new generalization of Dirichlet approximation theorem", Astérisque, no. 198–200, p. 273–283, 1991.

[31] M. S. Raghunathan – *Discrete Subgroups of Lie Groups*, Springer-Verlag, New York, 1972.

[32] W. M. Schmidt – *Diophantine Approximation*, Lecture Notes in Math., no. 785, Springer-Verlag, 1980.

[33] C. L. Siegel – *Advanced Analytic Number Theory*, Tata Institute of Fundamental Research, Bombay, 1980.

[34] J. Tits – *Buildings of Spherical Type and Finite BN-Pairs*, Lecture Notes in Mathematics, no. 386, Springer-Verlag, Berlin, 1974.
[35] A. Weil – *Adeles and Algebraic Groups*, Progress in Math., no. 23, Birkhäuser, Boston, 1982.
[36] R. J. Zimmer – *Ergodic Theory and Semisimple Groups*, Birkhäuser, Boston, 1984.

Bull. Soc. Math. France
150 (1), 2022, p. 53-75

PSEUDOVALUATIONS ON THE DE RHAM–WITT COMPLEX

by Rubén Muñoz--Bertrand

Abstract. — For a polynomial ring over a commutative ring of positive characteristic, we define on the associated de Rham–Witt complex a set of functions, and show that they are pseudovaluations in the sense of Davis, Langer and Zink. To achieve this, we explicitly compute products of basic elements on the complex. We also prove that the overconvergent de Rham–Witt complex can be recovered using these pseudovaluations.

Résumé (*Pseudovaluations sur le complexe de de Rham–Witt*). — Pour tout anneau polynomial sur un anneau commutatif de caractéristique strictement positive, on définit sur le complexe de de Rham–Witt associé un ensemble de fonctions, et l'on démontre que ce sont des pseudovaluations au sens de Davis, Langer et Zink. Pour y parvenir, on calcule explicitement des produits d'éléments basiques du complexe. On prouve également que le complexe de de Rham–Witt surconvergent peut être retrouvé en employant ces pseudovaluations.

Texte reçu le 13 décembre 2020, modifié le 26 août 2021, accepté le 14 octobre 2021.

Rubén Muñoz--Bertrand, Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405 Orsay, France • *E-mail :* `ruben.munoz-bertrand@universite-paris-saclay.fr` • *Url :* `https://www.imo.universite-paris-saclay.fr/~munoz/`

Mathematical subject classification (2010). — 14F30; 13F35.

Key words and phrases. — Overconvergent de Rham–Witt cohomology, p-adic cohomology.

This article was both written and revised while the author was affiliated with Normandie Univ, UNICAEN, CNRS, Laboratoire de Mathématiques Nicolas Oresme, 14000 Caen, France.

doi:10.24033/bsmf.2844

Introduction

Davis, Langer and Zink introduced the overconvergent de Rham–Witt complex in [2]. It is a complex of sheaves defined on any smooth variety X over a perfect field k of positive characteristic. It can be used to compute both the Monsky–Washnitzer and the rigid cohomology of the variety. This comparison was first established by [2] for quasi-projective smooth varieties; the assumption of quasi-projectiveness was then removed by Lawless [7].

This complex is defined as a differential graded algebra (dga) contained in the de Rham–Witt complex $W\Omega^{\bullet}_{X/k}$ of Deligne and Illusie. In order to achieve this they defined for any $\varepsilon > 0$, in the case where X is the spectrum of a polynomial ring $k[\underline{X}]$ over k, an order function $\gamma_{\varepsilon}\colon W\Omega^{\bullet}_{k[\underline{X}]/k} \to \mathbb{R} \cup \{+\infty, -\infty\}$. The overconvergent de Rham–Witt complex of X is the set of all $x \in W\Omega^{\bullet}_{k[\underline{X}]/k}$, such that $\gamma_{\varepsilon}(x) \neq -\infty$, for some $\varepsilon > 0$. In the general case, it is defined as the functional image of this set for a surjective morphism of smooth commutative algebras over k.

In degree zero (that is, for Witt vectors), these maps have nice properties. One of these is that they are pseudovaluations. We recall the definition. A **pseudovaluation** on a ring R is a function $v\colon R \to \mathbb{R} \cup \{+\infty, -\infty\}$, such that:

$$v(0) = +\infty,$$
$$v(1) = 0,$$
$$\forall r \in R,\ v(r) = v(-r)\,,$$
$$\forall r, s \in R,\ v(r+s) \geqslant \min\{v(r)\,, v(s)\}\,,$$
$$\forall r, s \in R,\ (v(r) \neq -\infty) \wedge (v(s) \neq -\infty) \implies (v(rs) \geqslant v(r) + v(s))\,.$$

The last formula will be referred to as the product formula in the remainder of this article.

Pseudovaluations and their behaviour have been studied in [3]. It appears that they form a convenient framework to study the overconvergence of recursive sequences. However, there are counterexamples showing that in positive degree, the maps γ_{ε} are not pseudovaluations. This becomes an obstacle when one wants to study the local structure of the overconvergent de Rham–Witt complex, or when one tries to find an interpretation of F-isocrystals for the overconvergent de Rham–Witt complex following the work of [4].

In this paper, we define new mappings

$$\zeta_{\varepsilon}\colon W\Omega^{\bullet}_{k[\underline{X}]/k} \to \mathbb{R} \cup \{+\infty, -\infty\}\ ,$$

for all $\varepsilon > 0$ and prove that these are pseudovaluations. Moreover, we show that the set of all $x \in W\Omega^{\bullet}_{k[\underline{X}]/k}$, such that $\zeta_{\varepsilon}(x) \neq -\infty$, for some $\varepsilon > 0$, also define the overconvergent de Rham–Witt complex.

In order to do so, in the first section we recall the main results concerning the de Rham–Witt complex, especially in the case of a polynomial algebra.

The second section, which is the most technical one, consists of computations of products of specific elements of the de Rham–Witt complex. The results are explicit and proven in the case where k is any commutative $\mathbb{Z}_{\langle p\rangle}$-algebra. This enables us in the last section to define the pseudovaluations and prove that in the case of a perfect field of positive characteristic, we retrieve with these functions the overconvergent de Rham–Witt complex.

The product formula comes in handy to control the overconvergence of sequences defined by recursion. This is the main motivation for this work, which will allow us in subsequent papers to study the structure of the overconvergent de Rham–Witt complex and, eventually, to give an interpretation of F-isocrystals as overconvergent de Rham–Witt connections.

1. The de Rham–Witt complex for a polynomial ring

Let p be a prime number. Let k be a commutative $\mathbb{Z}_{\langle p\rangle}$-algebra. Throughout this article, for any $i, j \in \mathbb{N}$, we shall write:

$$[\![i,j]\!] := \mathbb{N} \cap [i,j] \,.$$

Let $n \in \mathbb{N}$ and write $k[\underline{X}] := k[X_1, \dots, X_n]$. We will first recall basic properties of the de Rham–Witt complex of $k[\underline{X}]$, denoted $\left(W\Omega^{\bullet}_{k[\underline{X}]/k}, d\right)$ (for an introduction, see [5] or [6]). In degree zero, $W\Omega^0_{k[\underline{X}]/k}$ is isomorphic as a $W(k)$-algebra to $W(k[\underline{X}])$, the ring of Witt vectors over $k[\underline{X}]$.

There is a morphism of graded rings $F\colon W\Omega^{\bullet}_{k[\underline{X}]/k} \to W\Omega^{\bullet}_{k[\underline{X}]/k}$ called the **Frobenius endormorphism**, a morphism $V\colon W\Omega^{\bullet}_{k[\underline{X}]/k} \to W\Omega^{\bullet}_{k[\underline{X}]/k}$ of graded groups called the ***Verschiebung*** **endormorphism**, as well as a morphism of monoids $[\bullet] : (k[\underline{X}], \times) \to (W(k[\underline{X}]), \times)$ called the **Teichmüller lift** such that:

$$\forall r \in k[\underline{X}]\,,\ F([r]) = [r^p]\,, \tag{1}$$

$$\forall m \in \mathbb{N},\ \forall x, y \in W\Omega^{\bullet}_{k[\underline{X}]/k},\ V^m(xF^m(y)) = V^m(x)\, y, \tag{2}$$

$$\forall m \in \mathbb{N},\ \forall x \in W\Omega^{\bullet}_{k[\underline{X}]/k},\ d(F^m(x)) = p^m F^m(d(x))\,, \tag{3}$$

$$\forall m \in \mathbb{N},\ \forall P \in k[\underline{X}]\,,\ F^m(d([P])) = \left[P^{p^m-1}\right] d([P])\,, \tag{4}$$

$$\begin{gathered}\forall i, j \in \mathbb{N},\ \forall x \in W\Omega^i_{k[\underline{X}]/k},\ \forall y \in W\Omega^j_{k[\underline{X}]/k},\\ d(xy) = (-1)^i\, xd(y) + (-1)^{(i+1)j}\, yd(x)\,,\end{gathered} \tag{5}$$

$$\begin{gathered}\forall m \in \mathbb{N},\ \forall (x_i)_{i \in [\![1,m]\!]} \in (W(k[\underline{X}]))^m\,,\\ d\left(\prod_{i=1}^m x_i\right) = \sum_{i=1}^m \left(\prod_{j \in [\![1,m]\!] \smallsetminus \{i\}} x_j\right) d(x_i)\,.\end{gathered} \tag{6}$$

We shall introduce basic elements on the de Rham–Witt complex, called basic Witt differentials, and recall how any de Rham–Witt differential on $k[\underline{X}]$ can be expressed as a series using these elements. We mostly follow [6].

Definition 1.1. — A **weight function** is a mapping $a\colon [\![1,n]\!] \to \mathbb{N}\left[\frac{1}{p}\right]$; for all $i \in [\![1,n]\!]$, its values shall be written as a_i. We define:

$$|a| \coloneqq \sum_{i=1}^{n} a_i,$$

and:

$$\underline{X}^a \coloneqq \prod_{i=1}^{n} {X_i}^{a_i}.$$

For any weight function a and any $J \subset [\![1,n]\!]$, we denote by $a|_J$ the weight function that for all $i \in [\![1,n]\!]$ satisfies:

$$a|_J(i) = \begin{cases} a_i & \text{if } i \in J, \\ 0 & \text{otherwise.} \end{cases}$$

The **support** of a weight function a is the following set:

$$\operatorname{Supp}(a) \coloneqq \{i \in [\![1,n]\!] \mid a_i \neq 0\}.$$

A **partition** I of a weight function a is a subset $I \subset \operatorname{Supp}(a)$. Its **size** is its cardinal. We will denote by $\mathcal{P}$ the set of all pairs (a, I), where a is a weight function, and I is a partition of a.

In all this paper, the p-adic valuation shall be denoted v_p. For a weight function a, we fix the following total order $\preceq$ on $\operatorname{Supp}(a)$:

$$\begin{aligned} &\forall i, i' \in \operatorname{Supp}(a),\ i \preceq i' \\ &\qquad \iff ((\mathrm{v}_p(a_i) \leqslant \mathrm{v}_p(a_{i'})) \wedge ((\mathrm{v}_p(a_i) = \mathrm{v}_p(a_{i'})) \implies (i \leqslant i'))). \end{aligned}$$

We will denote by $\prec$ the associated strict total order and we also let $\min(a) \in \operatorname{Supp}(a)$ be the only element such that $\min(a) \preceq i$, for any $i \in \operatorname{Supp}(a)$.

Let $m \in [\![0,n]\!]$. Let $I \coloneqq \{i_j\}_{j \in [\![1,m]\!]}$ be a partition of a weight function a. We will always suppose that $i_j \prec i_{j'}$, for all $j, j' \in [\![1,m]\!]$, such that $j < j'$. By convention, we will say that $i_0 \preceq i$ and $i \prec i_{m+1}$ whenever $i \in \operatorname{Supp}(a)$. We define the following $m+1$ subsets of $\operatorname{Supp}(a)$ for any $l \in [\![0,m]\!]$:

$$I_l \coloneqq \{i \in \operatorname{Supp}(a) \mid i_l \preceq i \prec i_{l+1}\}.$$

Let a be a weight function. We set:

$$\begin{aligned} \mathrm{v}_p(a) &\coloneqq \min\{\mathrm{v}_p(a_i) \mid i \in [\![1,n]\!]\}, \\ u(a) &\coloneqq \max\{0, -\mathrm{v}_p(a)\}. \end{aligned}$$

If a is not the zero function, we put:

$$g(a) := F^{u(a)+\mathrm{v}_p(a)}\Big(d\Big(V^{u(a)}\Big(\Big[\underline{X}^{p^{-\mathrm{v}_p(a)}a}\Big]\Big)\Big)\Big).$$

Furthermore, if I is a partition of a, and η is any element in $W(k)$, we set:

(7)
$$e(\eta, a, I) := \begin{cases} d\Big(V^{u(a)}\Big(\eta\Big[\underline{X}^{p^{u(a)}a|_{I_0}}\Big]\Big)\Big)\prod_{l=2}^{\#I} g(a|_{I_l}) & \text{if } I_0 = \emptyset \text{ and } u(a) \neq 0, \\ V^{u(a)}\Big(\eta\Big[\underline{X}^{p^{u(a)}a|_{I_0}}\Big]\Big)\prod_{l=1}^{\#I} g(a|_{I_l}) & \text{otherwise.} \end{cases}$$

When $I_0 = \emptyset$ and $u(a) = 0$, then $V^{u(a)}\Big(\eta\Big[\underline{X}^{p^{u(a)}a|_{I_0}}\Big]\Big) = \eta$. So one can notice that, if one ignores η, the element defined above is a product of $\#I$ factors whenever $I_0 = \emptyset$, and of $\#I + 1$ factors otherwise, the factors being the images of an element through one of the functions d, g or V. We will use this fact later, when we define the pseudovaluations on the de Rham–Witt complex of a polynomial ring.

We recall the action of d, V and F on these elements.

Proposition 1.2. — *For any $(a, I) \in \mathcal{P}$ and any $\eta \in W(k)$, we have:*

$$d(e(\eta, a, I)) = \begin{cases} 0 & \text{if } I_0 = \emptyset, \\ e(\eta, a, I \cup \{\min(a)\}) & \text{if } I_0 \neq \emptyset \text{ and } \mathrm{v}_p(a) \leqslant 0, \\ p^{\mathrm{v}_p(a)} e(\eta, a, I \cup \{\min(a)\}) & \text{if } I_0 \neq \emptyset \text{ and } \mathrm{v}_p(a) > 0. \end{cases}$$

Proof. — See [6, proposition 2.6]. □

Proposition 1.3. — *For any $(a, I) \in \mathcal{P}$ and any $\eta \in W(k)$, we have:*

$$F(e(\eta, a, I)) = \begin{cases} e(\eta, pa, I) & \text{if } \mathrm{v}_p(a) < 0 \text{ and } I_0 = \emptyset, \\ e(p\eta, pa, I) & \text{if } \mathrm{v}_p(a) < 0 \text{ and } I_0 \neq \emptyset, \\ e(F(\eta), pa, I) & \text{if } \mathrm{v}_p(a) \geqslant 0. \end{cases}$$

Proof. — See [6, proposition 2.5]. □

Proposition 1.4. — *For any $(a, I) \in \mathcal{P}$ and any $\eta \in W(k)$, we have:*

$$V(e(\eta, a, I)) = \begin{cases} e\Big(V(\eta), \frac{a}{p}, I\Big) & \text{if } \mathrm{v}_p(a) > 0, \\ e\Big(p\eta, \frac{a}{p}, I\Big) & \text{if } \mathrm{v}_p(a) \leqslant 0 \text{ and } I_0 = \emptyset, \\ e\Big(\eta, \frac{a}{p}, I\Big) & \text{if } \mathrm{v}_p(a) \leqslant 0 \text{ and } I_0 \neq \emptyset. \end{cases}$$

Proof. — See [6, proposition 2.5]. □

The de Rham–Witt complex is endowed with a topology called the standard topology [5, I. 3.1.]. In this article, it will not be necessary to recall its definition, as we will only need the fact that a series of the form $\sum_{(a,I)\in\mathcal{P}} e(\eta_{a,I}, a, I)$, with $\eta_{a,I} \in W(k)$, for all $(a,I) \in \mathcal{P}$, converges in $W\Omega^{\bullet}_{k[\underline{X}]/k}$ if and only if for any $m \in \mathbb{N}$, we have $V^{u(a)}(\eta_{a,I}) \in V^m(W(k))$ except for a finite number of $(a,I) \in \mathcal{P}$.

The following theorem is essential to the definition of the overconvergent de Rham–Witt complex, and to its decomposition as a $W(k)$-module in the case of a polynomial algebra.

Theorem 1.5. — *For any differential $x \in W\Omega^{\bullet}_{k[\underline{X}]/k}$, there exists a unique function*

$$\eta\colon \begin{array}{c} \mathcal{P} \to W(k) \\ (a,I) \mapsto \eta_{a,I} \end{array}$$

such that:

$$x = \sum_{(a,I)\in\mathcal{P}} e(\eta_{a,I}, a, I)\,.$$

Proof. — See [6, theorem 2.8]. □

2. Computations

The goal of this section is to make explicit computations of the product of two basic Witt differentials, that is, elements of the form (7).

Let k be a commutative $\mathbb{Z}_{\langle p\rangle}$-algebra. Let $n \in \mathbb{N}$. In what follows, we shall denote $k[\underline{X}] \coloneqq k[X_1, \ldots, X_n]$. For any $(a,I) \in \mathcal{P}$ with a taking values in $\mathbb{N}$, we will write:

$$h(a,I) \coloneqq \prod_{i\in\operatorname{Supp}(a)\smallsetminus I} [X_i]^{k_i} \prod_{j\in I} g\big(a|_{\{j\}}\big)\,.$$

We will use the elements h defined above in order to achieve this, as they appear to be more convenient for calculations.

Lemma 2.1. — *Let R be a commutative k-algebra. Let $x \in R$ and let $m, m' \in \mathbb{N}$, such that $m + m' \neq 0$. Put $a \coloneqq \mathrm{v}_p(m+m')$ and $b \coloneqq p^{-a}(m+m')$. Then we have in the de Rham–Witt complex $W\Omega^{\bullet}_{R/k}$ of R:*

$$[x]^m\, d\Big([x]^{m'}\Big) = \frac{m'}{b} F^a\Big(d\Big([x]^b\Big)\Big)\,.$$

Proof. — Using (1) and (3), we get:

$$d\Big([x]^{m+m'}\Big) = d\Big(F^a\Big([x]^b\Big)\Big) = p^a F^a\Big(d\Big([x]^b\Big)\Big)\,.$$

Moreover, as:

$$(m+m')\,[x]^m\, d\Big([x]^{m'}\Big) = m' d\Big([x]^{m+m'}\Big)\,,$$

we obtain in the case where $k=\mathbb{Z}_{\langle p\rangle}$, $R=\mathbb{Z}_{\langle p\rangle}[X]$ and $x=X$ the formula:

$$p^a b[x]^m\, d\Big([x]^{m'}\Big)=p^a m' F^a\Big(d\Big([x]^b\Big)\Big)\,.$$

The ring $W\big(\mathbb{Z}_{\langle p\rangle}[X]\big)$ has no p-torsion, as $\mathbb{Z}_{\langle p\rangle}[X]$ itself has no p-torsion. So we deduce from theorem 1.5 that $W\Omega^{\bullet}_{k[\underline{X}]/k}$ also has no p-torsion, which allows us to conclude in this situation. For the general case, using the canonical commutative diagram:

$$\begin{tikzcd}
\mathbb{Z}_{\langle p\rangle}[X] \arrow[r] & R \\
\mathbb{Z}_{\langle p\rangle} \arrow[u] \arrow[r] & k, \arrow[u]
\end{tikzcd}$$

where the upper arrow sends X to x, we conclude using the morphism of $W\big(\mathbb{Z}_{\langle p\rangle}\big)$-dgas $W\Omega^{\bullet}_{\mathbb{Z}_{\langle p\rangle}[X]/\mathbb{Z}_{\langle p\rangle}}\to W\Omega^{\bullet}_{R/k}$ obtained by functoriality of $W\Omega^{\bullet}_{\bullet/\bullet}$. □

The next proposition gives a simple formula for products of values of h. The goal of the subsequent lemmas will be to use it in order to get a formula for products of elements of the form (7).

Proposition 2.2. — *Let $(a,I),(b,J)\in\mathcal{P}$, such that a and b take values in $\mathbb{N}$. There exists $m\in\mathbb{Z}_{\langle p\rangle}$, such that:*

$$h(a,I)\,h(b,J)=\begin{cases} m\,h(a+b,I\cup J) & \textit{if } I\cap J=\emptyset,\\ 0 & \textit{otherwise.}\end{cases}$$

Proof. — First, we have by definition:

$$h(a,I)\,h(b,J)\\ =\prod_{i\in\operatorname{Supp}(a)\smallsetminus I}[X_i]^{a_i}\prod_{i'\in I} g\big(a|_{\{i'\}}\big)\prod_{j\in\operatorname{Supp}(b)\smallsetminus J}[X_j]^{b_j}\prod_{j'\in J} g\big(b|_{\{j'\}}\big)\,.$$

Since $W\Omega^{\bullet}_{k[\underline{X}]/k}$ is alternating, this product is zero whenever $I\cap J\neq\emptyset$. Otherwise, we get:

$$h(a,I)\,h(b,J)\\ =\prod_{i\in\operatorname{Supp}(a+b)\smallsetminus(I\cup J)}[X_i]^{a_i+b_i}\prod_{i'\in I}[X_{i'}]^{b_{i'}}\, g\big(a|_{\{i'\}}\big)\prod_{j'\in J}[X_{j'}]^{a_{j'}}\, g\big(b|_{\{j'\}}\big)\,.$$

Moreover, for any $i' \in I$ we obtain:

$$[X_{i'}]^{b_{i'}}\, g\left(a|_{\{i'\}}\right) \stackrel{(4)}{=} [X_{i'}]^{\left(1-p^{-\mathrm{v}_p\left(a|_{\{i'\}}\right)}\right)a_{i'}+b_{i'}}\, d\left([X_i]^{p^{-\mathrm{v}_p\left(a|_{\{i'\}}\right)}a_{i'}}\right)$$

$$\stackrel{2.1}{=} \frac{p^{-\mathrm{v}_p\left(a|_{\{i'\}}\right)}a_{i'}}{p^{-\mathrm{v}_p(a_{i'}+b_{i'})}(a_{i'}+b_{i'})} g\left((a+b)\,|_{\{i'\}}\right).$$

Using the same argument, for any $j' \in J$, one successfully gets:

$$[X_{j'}]^{a_{j'}}\, g\left(b|_{\{j'\}}\right) = \frac{p^{-\mathrm{v}_p\left(b|_{\{j'\}}\right)}b_{j'}}{p^{-\mathrm{v}_p\left(a_{j'}+b_{j'}\right)}(a_{j'}+b_{j'})} g\left((a+b)\,|_{\{j'\}}\right).$$

This concludes the proof, because:

$$\prod_{i\in\mathrm{Supp}(a+b)\smallsetminus(I\cup J)} [X_i]^{a_i+b_i} \prod_{i'\in I} g\left((a+b)\,|_{\{i'\}}\right) \prod_{j'\in J} g\left((a+b)\,|_{\{j'\}}\right)$$
$$= h(a+b, I\cup J). \quad \square$$

Lemma 2.3. — *Let $(a, I) \in \mathcal{P}$, such that a takes values in $\mathbb{N}$. We have:*

$$g(a) = \sum_{j\in\mathrm{Supp}(a)} p^{\mathrm{v}_p(a_j)-\mathrm{v}_p(a)} h(a, \{j\}).$$

Proof. — Write $S \coloneqq \mathrm{Supp}(a)$ for simplicity. We compute:

$$\begin{aligned}
&F^{\mathrm{v}_p(a)}\left(d\left(\left[\underline{X}^{p^{-\mathrm{v}_p(a)}a}\right]\right)\right)\\
&\stackrel{(4)}{=} \left[\underline{X}^{\left(1-p^{-\mathrm{v}_p(a)}\right)a}\right] d\left(\left[\underline{X}^{p^{-\mathrm{v}_p(a)}a}\right]\right)\\
&\stackrel{(6)}{=} \left[\underline{X}^{\left(1-p^{-\mathrm{v}_p(a)}\right)a}\right] \sum_{j\in S}\left(\prod_{j'\in S\smallsetminus\{j\}}\left[X_{j'}{}^{p^{-\mathrm{v}_p(a)}a_{j'}}\right]\right) d\left(\left[X_j{}^{p^{-\mathrm{v}_p(a)}a_j}\right]\right)\\
&\stackrel{(4)}{=} \sum_{j\in S}\left(\prod_{j'\in S\smallsetminus\{j\}}\left[X_{j'}{}^{a_{j'}}\right]\right) F^{\mathrm{v}_p(a)}\left(d\left(\left[X_j{}^{p^{-\mathrm{v}_p(a)}a_j}\right]\right)\right)\\
&\stackrel{\substack{(1)\\(3)}}{=} \sum_{j\in S}\left(\prod_{j'\in S\smallsetminus\{j\}}\left[X_{j'}{}^{a_{j'}}\right]\right) p^{\mathrm{v}_p(a_j)-\mathrm{v}_p(a)} F^{\mathrm{v}_p(a_j)}\left(d\left(\left[X_j{}^{p^{-\mathrm{v}_p(a_j)}a_j}\right]\right)\right)\\
&= \sum_{j\in S} p^{\mathrm{v}_p(a_j)-\mathrm{v}_p(a)} h(a, \{j\}).
\end{aligned}$$

This ends the proof because by definition $g(a) = F^{\mathrm{v}_p(a)}\left(d\left(\left[\underline{X}^{p^{-\mathrm{v}_p(a)}a}\right]\right)\right)$. $\square$

The next lemma will be used to write any value of the function h defined above as a linear combination of elements of the form (7). The previous lemma can be seen as a kind of reciprocal.

Lemma 2.4. — *Let $(a, I) \in \mathcal{P}$, such that a takes values in $\mathbb{N}$. Denote by P the set of partitions of $\operatorname{Supp}(a)$ of size $\#I$. Then there exists a function $s\colon P \to \mathbb{N} \subset W(k)$, such that:*

$$h(a, I) = \sum_{J \in P} e(s(J), a, J) \,.$$

Proof. — Put $m := \#I$. If $m = 0$, then obviously $h(a, I) = e(1, a, I)$. Thus, suppose that $m \neq 0$. Write $\{i_l\}_{l \in [\![1,m]\!]} := I$, with $i_j \prec i_{j'}$, for any pair $j < j'$ in $[\![1, m]\!]$, and for all $j \in \operatorname{Supp}(a)$ put $v_j := \mathrm{v}_p(a_j)$ and $b_j = p^{-v_j} a_j$. By definition:

$$h(a, I) = \prod_{i \in \operatorname{Supp}(a) \smallsetminus I} [X_i]^{a_i} \prod_{j \in I} F^{v_j}\Big(d\Big([X_j]^{b_j}\Big)\Big) \,.$$

So we can write:

$$h(a, I) = h\big(a|_{\operatorname{Supp}(a) \smallsetminus I_m}, I \smallsetminus \{i_m\}\big)\, F^{v_{i_m}}\Big(d\Big([X_{i_m}]^{b_{i_m}}\Big)\Big) \prod_{i \in I_m \smallsetminus \{i_m\}} [X_i]^{a_i} \,.$$

Moreover, we can compute:

$$\begin{aligned}
&F^{v_{i_m}}\Big(d\Big([X_{i_m}]^{b_{i_m}}\Big)\Big) \prod_{i \in I_m \smallsetminus \{i_m\}} [X_i]^{a_i} \\
&\overset{(1)}{=} F^{v_{i_m}}\left(d\Big([X_{i_m}]^{b_{i_m}}\Big) \prod_{i \in I_m \smallsetminus \{i_m\}} [X_i]^{p^{-v_{i_m}} a_i}\right) \\
&\overset{(5)}{=} F^{v_{i_m}}\left(d\left(\prod_{i \in I_m} [X_i]^{p^{-v_{i_m}} a_i}\right) - [X_{i_m}]^{b_{i_m}}\, d\left(\prod_{i \in I_m \smallsetminus \{i_m\}} [X_i]^{p^{-v_{i_m}} a_i}\right)\right) \\
&\overset{\substack{(1)\\(3)}}{=} g(a|_{I_m}) - F^{v_{i_m}}\Big([X_{i_m}]^{b_{i_m}}\Big)\, p^{\mathrm{v}_p(a|_{I_m \smallsetminus \{i_m\}}) - v_{i_m}}\, g\big(a|_{I_m \smallsetminus \{i_m\}}\big) \\
&\overset{(1)}{=} g(a|_{I_m}) - p^{\mathrm{v}_p(a|_{I_m \smallsetminus \{i_m\}}) - v_{i_m}}\, [X_{i_m}]^{a_{i_m}}\, g\big(a|_{I_m \smallsetminus \{i_m\}}\big) \,.
\end{aligned}$$

So we get:

$$\begin{aligned}
h(a, I) = {} & h\big(a|_{\operatorname{Supp}(a) \smallsetminus I_m}, I \smallsetminus \{i_m\}\big)\, g(a|_{I_m}) \\
& - p^{\mathrm{v}_p(a|_{I_m \smallsetminus \{i_m\}}) - v_{i_m}}\, h\big(a|_{\{i_m\} \cup \operatorname{Supp}(a) \smallsetminus I_m}, I \smallsetminus \{i_m\}\big)\, g\big(a|_{I_m \smallsetminus \{i_m\}}\big) \,.
\end{aligned}$$

We can then deduce the lemma by induction on $m = \#I$. Indeed, if we suppose that $h\big(a|_{\operatorname{Supp}(a) \smallsetminus I_m}, I \smallsetminus \{i_m\}\big)$ can be written as a linear combination of elements of the form $e\big(1, a|_{\operatorname{Supp}(a) \smallsetminus I_m}, J'\big)$, where J' is a partition of

$\mathrm{Supp}(a) \smallsetminus I_m$ of size $m-1$, then since

$$e\big(1, a|_{\mathrm{Supp}(a)\smallsetminus I_m}, J'\big)\, g(a|_{I_m}) = e(1, a, J' \cup \{i_m\})\,,$$

the lemma is proven for the minuend of the above difference, and one can conclude for the subtrahend by using the same reasoning. □

Lemma 2.5. — *Let $(a, I), (b, J) \in \mathcal{P}$, such that a and b take values in $\mathbb{N}$. Let $\eta, \eta' \in W(k)$. Denoting by P the set of partitions of* $\mathrm{Supp}(a+b)$ *of size $\#I + \#J$, then there exists a function $s\colon P \to \mathbb{Z}_{\langle p\rangle}$ such that:*

$$e(\eta, a, I)\, e(\eta', b, J) = \sum_{L\in P} e(s(L)\,\eta\eta', a+b, L)\,.$$

Proof. — By definition we have $e(\eta, a, I) = \eta\Big[\underline{X}^{a|_{I_0}}\Big] \prod_{i=1}^{\#I} g(a|_{I_i})$. There is also a similar equation defining $e(\eta', b, J)$. Using lemma 2.3, for any $i \in [\![1, \#I]\!]$, we can write $g(a|_{I_i})$ as a linear combination of elements of the form $h(a|_{I_i}, \{j_i\})$ with $j_i \in I_i$. Also, by definition, $\Big[\underline{X}^{a|_{I_0}}\Big] = h(a|_{I_0}, \emptyset)$. Thus, we can write $e(\eta, a, I)$ as a linear combination of products of elements of the form $\eta h(a|_{I_0}, \emptyset) \prod_{i=1}^{\#I} h(a|_{I_i}, \{j_i\})$, where all $j_i \in I_i$, for any $i \in [\![1, \#I]\!]$. Again, we can do the same with $e(\eta', b, J)$. We can conclude by using proposition 2.2 and lemma 2.4. □

Lemma 2.6. — *Let $(a, I), (b, J) \in \mathcal{P}$, such that $u(a) \geqslant u(b)$ and $I_0 \neq \emptyset$. Denote by P the set of partitions of size $\#I + \#J$ of* $\mathrm{Supp}(a+b)$*, and put:*

$$v \coloneqq \begin{cases} u(b) & \text{if } J_0 \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

Then for any $\eta, \eta' \in W(k)$, there exists a function $s\colon P \to \mathbb{Z}_{\langle p\rangle}$ with:

$$\forall L \in P,\ \begin{cases} p^{v+u(a+b)} \mid s(L) & \text{if } L_0 = \emptyset, \\ p^{v} \mid s(L') & \text{otherwise,} \end{cases}$$

such that:

$$e(\eta, a, I)\, e(\eta', b, J) = \sum_{L\in P} e\Big(s(L)\, V^{u(a)-u(a+b)}\Big(\eta F^{u(a)-u(b)}(\eta')\Big), a+b, L\Big)\,.$$

Proof. — Put $\tilde{I} \coloneqq \bigcup_{i \in [\![1,\#I]\!]} I_i$. We can compute:

$$\begin{aligned}
&e(\eta, a, I)\, e(\eta', b, J) \\
&\overset{(2)}{=} V^{u(a)}\Big(\eta\Big[\underline{X}^{p^{u(a)}(a|_{I_0})}\Big]\, F^{u(a)}(e(1, a|_{\tilde{I}}, I)\, e(\eta', b, J))\Big) \\
&\overset{1.3}{=} V^{u(a)}\Big(\eta\Big[\underline{X}^{p^{u(a)}(a|_{I_0})}\Big]\, e\Big(1, p^{u(a)} a|_{\tilde{I}}, I\Big)\, e\Big(p^v F^{u(a)-u(b)}(\eta'), p^{u(a)} b, J\Big)\Big) \\
&= V^{u(a)}\Big(e\Big(\eta, p^{u(a)} a, I\Big)\, e\Big(p^v F^{u(a)-u(b)}(\eta'), p^{u(a)} b, J\Big)\Big).
\end{aligned}$$

These computations have been done so that the basic Witt differentials appearing in the last line are integral. In particular, we are now in position to apply lemma 2.5. That is, there is a function $s' \colon P \to \mathbb{Z}_{\langle p \rangle}$ such that:

$$e(\eta, a, I)\, e(\eta', b, J) = V^{u(a)}\left(\sum_{L \in P} e\Big(p^v s'(L)\, \eta F^{u(a)-u(b)}(\eta'), p^{u(a)}(a+b), L\Big)\right).$$

We can conclude by using proposition 1.4 and the fact that the *Verschiebung* endomorphism is additive. □

In the last two statements of this section, we are interested in the case where k has characteristic p. The results become clearer in this situation because we have $p = V(F(1))$.

Lemma 2.7. — *Suppose k has characteristic p. Let $(a, I), (b, J) \in \mathcal{P}$, such that $u(a) \geqslant u(b)$ and $I_0 \neq \emptyset$. Denote by P the set of partitions of size $\#I + \#J$ of* $\operatorname{Supp}(a+b)$*, and put:*

$$v \coloneqq \begin{cases} u(b) & \text{if } J_0 \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\alpha, \beta \in \mathbb{N}$. Then for any $\eta \in V^\alpha(W(k))$ and any $\eta' \in V^\beta(W(k))$, there exists a function $s \colon P \to W(k)$ with:

$$\forall L \in P, \; \begin{cases} s(L) \in V^{\alpha+\beta+v+u(a)}(W(k)) & \text{if } L_0 = \emptyset, \\ s(L) \in V^{\alpha+\beta+v+u(a)-u(a+b)}(W(k)) & \text{otherwise,} \end{cases}$$

such that:

$$e(\eta, a, I)\, e(\eta', b, J) = \sum_{L \in P} e(s(L), a+b, L).$$

Proof. — This is a special case of lemma 2.6 when k has characteristic p, because in that case, we have $px = F(V(x)) = V(F(x))$ for any $x \in W(k)$, but also $\eta\eta' \in V^{\alpha+\beta}(W(k))$ [1, proposition 5. p. IX.15]. □

Proposition 2.8. — *Suppose k has characteristic p. Let $(a,I),(b,J)\in\mathcal{P}$ with $I_0\neq\emptyset$. Denote by P the set of partitions of size $\#I+\#J$ of* $\operatorname{Supp}(a+b)$. *Let $\alpha,\beta\in\mathbb{N}$. Then, for any $\eta\in V^{\alpha}(W(k))$ and any $\eta'\in V^{\beta}(W(k))$, there exists a function $s\colon P\to W(k)$ with:*

$$\forall L\in P,\ \begin{cases} s(L)\in V^{\alpha+\beta+\min\{u(a),u(b)\}}(W(k)) & \text{if } L_0=\emptyset,\\ s(L)\in V^{\alpha+\beta+\max\{u(a),u(b)\}-u(a+b)}(W(k)) & \text{otherwise,}\end{cases}$$

such that:

$$e(\eta,a,I)\,e(\eta',b,J)=\sum_{L\in P}e(s(L)\,,a+b,L)\,.$$

Proof. — This statement is just a special case of lemma 2.7, except when $u(b)>u(a)$ and $J_0=\emptyset$. In that situation, if $J':=J\smallsetminus\{\min(b)\}$, we deduce from proposition 1.2 that:

$$\begin{aligned} &e(\eta,a,I)\,e(\eta',b,J)\\ &\qquad=e(\eta,a,I)\,d(e(\eta',b,J'))\\ &\qquad=(-1)^{\#I}\left(d(e(\eta,a,I)\,e(\eta',b,J'))-d(e(\eta,a,I))\,e(\eta',b,J')\right).\end{aligned}$$

This enables us to conclude using lemma 2.7 again. □

3. Pseudovaluations

We shall now consider the case where k is a commutative ring of characteristic p. Let $n\in\mathbb{N}$ and let $k[\underline{X}]:=k[X_1,\dots,X_n]$. Recall that theorem 1.5 says that any $x\in W\Omega^{\bullet}_{k[\underline{X}]/k}$ can be uniquely written as $\sum_{(a,I)\in\mathcal{P}}e(\eta_{a,I},a,I)$, where all $\eta_{a,I}\in W(k)$. This allows us to define specific $W(k)$-submodules of the de Rham–Witt complex.

Definition 3.1. — An element $x=\sum_{(a,I)\in\mathcal{P}}e(\eta_{a,I},a,I)\in W\Omega^{\bullet}_{k[\underline{X}]/k}$ is said to be **integral** if $\eta_{a,I}=0$, for all a with $u(a)\neq 0$. We denote by $W\Omega^{\mathrm{int},\bullet}_{k[\underline{X}]/k}$ the subset of all integral elements of the de Rham–Witt complex.

The element x is said to be **fractional** if $\eta_{a,I}=0$ for all a with $u(a)=0$. We denote by $W\Omega^{\mathrm{frac},\bullet}_{k[\underline{X}]/k}$ the subset of all fractional elements of the de Rham–Witt complex.

The element x is said to be **pure fractional** if $\eta_{a,I}=0$ for all (a,I), such that $u(a)=0$ or $I_0=\emptyset$. We denote by $W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}$ the subset of all pure fractional elements of the de Rham–Witt complex.

Notice that we have the following decomposition as $W(k)$-modules:

$$W\Omega^{\bullet}_{k[\underline{X}]/k}\cong W\Omega^{\mathrm{int},\bullet}_{k[\underline{X}]/k}\oplus W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}\oplus d\Big(W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}\Big)\,. \tag{8}$$

This is a refinement of Langer and Zink's decomposition into integral and fractional parts [6, (3.9)]. Indeed, we have:

$$W\Omega^{\mathrm{frac},\bullet}_{k[\underline{X}]/k} \cong W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k} \oplus d\Big(W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}\Big).$$

In the whole chapter, for any $x \in W\Omega^{\bullet}_{k[\underline{X}]/k}$, we will denote by $x|_{\mathrm{int}}$, $x|_{\mathrm{frac}}$, $x|_{\mathrm{frp}}$ and $x|_{\mathrm{d(frp)}}$ the obvious projections for these decompositions.

We will also denote by v_V the V-adic pseudovaluation on $W(k)$. Davis, Langer and Zink defined the following functions for any $\varepsilon > 0$ [2, (0.3)]:

$$\gamma_\varepsilon : \begin{array}{c} W\Omega^{\bullet}_{k[\underline{X}]/k} \to \mathbb{R} \cup \{+\infty, -\infty\} \\ \sum_{(a,I)\in\mathcal{P}} e(\eta_{a,I}, a, I) \mapsto \inf_{(a,I)\in\mathcal{P}}\{\mathrm{v}_V(\eta_{a,I}) + u(a) - \varepsilon|a|\}. \end{array}$$

To see that this definition coincides with the one given by Davis, Langer and Zink, it is necessary to see that in their definition of a basic Witt differential, they ask that $\mathrm{v}_V(\eta_{a,I}) \geqslant u(a)$, for all $(a, I) \in \mathcal{P}$ [6, p. 261]. The definition given in this article has been modified, which is why we need to add $u(a)$ in the definition of γ_ε.

The overconvergent de Rham–Witt complex of $k[\underline{X}]$ is the set of all $x \in W\Omega^{\bullet}_{k[\underline{X}]/k}$, such that there exists $\varepsilon > 0$ with $\gamma_\varepsilon(x) \neq -\infty$.

One of the main obstacles to studying the overconvergence of recursive sequences containing products of de Rham–Witt differentials is that these functions are not pseudovaluations. We will first study two counterexamples to the product rule in the case where $k[\underline{X}] \cong k[X, Y]$ as k-algebras. That is, we will find $x, y \in W\Omega^{\bullet}_{k[\underline{X}]/k}$, such that for all $\varepsilon > 0$, we have $\gamma_\varepsilon(x) \neq -\infty$, $\gamma_\varepsilon(y) \neq -\infty$ and $\gamma_\varepsilon(xy) < \gamma_\varepsilon(x) + \gamma_\varepsilon(y)$.

Example 3.2. — For any $m \in \mathbb{N}$, notice that:

$$\begin{aligned} V^m\Big(\Big[X^{p^m-1}\Big]\Big)\, d(V^m([X])) &= p^m d([X]), \\ \gamma_\varepsilon\Big(V^m\Big(\Big[X^{p^m-1}\Big]\Big)\Big) &= m - \frac{\varepsilon(p^m-1)}{p^m}, \\ \gamma_\varepsilon(d(V^m([X]))) &= m - \frac{\varepsilon}{p^m}, \\ \gamma_\varepsilon(p^m d([X])) &= m - \varepsilon < 2m - \varepsilon. \end{aligned}$$

This first counterexample illustrates what happens when one takes the product of two fractional elements. The phenomenon occurring here with $x = V^m\big(\big[X^{p^m-1}\big]\big)$ and $y = d(V^m([X]))$ is that the power of the denominator of the weight functions (which we denoted $a \mapsto u(a)$) can get smaller when taking products of differentials. Indeed, lemmas 2.6 and 2.7 and proposition 2.8 show that multiplying basic elements translates as an addition for weight functions. However, we notice in this example that the V-adic pseudovaluation we have to calculate gets bigger; it is just not big enough, so it compensates the decrease of u. In this example, to get a function satisfying the product formula of pseudovaluations, it seems to be enough to multiply the V-adic pseudovaluation by 2 in the definition of γ_ε. It is still not sufficient in general, as can be seen in the following counterexample.

Example 3.3. — Let $m \in \mathbb{N}$. Then:

$$\gamma_\varepsilon\Big(V^m\Big(\Big[X^{p^m-1}\Big]\Big)\Big) = m - \frac{\varepsilon(p^m-1)}{p^m},$$
$$\gamma_\varepsilon(d(V^m([Y]))) = m - \frac{\varepsilon}{p^m},$$
$$\gamma_\varepsilon\Big(V^m\Big(\Big[X^{p^m-1}\Big]\Big)\, d(V^m([Y]))\Big) = m - \varepsilon < 2m - \varepsilon.$$

Another type of counterexample thus appears taking $x = V^m\big(\big[X^{p^m-1}\big]\big)$ and $y = d(V^m([Y]))$. In this situation, x, y and xy are basic Witt differentials, and the image through u of their associated weight functions is always m. This happens to be the main reason why the product formula fails with γ_ε in this context, as we need to add $2m$ when computing $\gamma_\varepsilon(x) + \gamma_\varepsilon(x)$, but m only appears once in the computation of $\gamma_\varepsilon(xy)$. So, in order for the product formula to work in general, we need to multiply the value of u in the definition of γ_ε by the number of factors in the definition of (7). As this number is smaller than n, as remarked after the first counterexample, we also have to multiply the V-adic pseudovaluation by $2n$.

This leads us to the definition below, which is a modification of Davis, Langer and Zink's definition. From now, $n \in \mathbb{N}$ is an arbitrary integer.

Definition 3.4. — For any $\varepsilon > 0$ put:

$$\zeta_\varepsilon \colon \begin{array}{l} W\Omega^\bullet_{k[\underline{X}]/k} \to \mathbb{R} \cup \{+\infty, -\infty\} \\ x \mapsto \begin{cases} \inf_{(a,I)\in\mathcal{P}}\{2n\, \mathrm{v}_V(\eta_{a,I}) + \#I u(a) - \varepsilon|a|\} & \text{if } I_0 = \emptyset, \\ \inf_{(a,I)\in\mathcal{P}}\{2n\, \mathrm{v}_V(\eta_{a,I}) + (\#I+1)\, u(a) - \varepsilon|a|\} & \text{if } I_0 \neq \emptyset, \end{cases} \end{array}$$

for $x = \sum_{(a,I)\in\mathcal{P}} e(\eta_{a,I}, a, I)$.

We will prove that these functions are pseudovaluations. Before we demonstrate the product formula, we first give a few basic properties. It is, for instance, immediate that:

$$\forall x, y \in W\Omega^\bullet_{k[\underline{X}]/k},\ \zeta_\varepsilon(x+y) \geqslant \min\{\zeta_\varepsilon(x), \zeta_\varepsilon(y)\}\,. \tag{9}$$

Also, a consequence of proposition 1.2 is that:

$$\forall x \in W\Omega^\bullet_{k[\underline{X}]/k},\ \zeta_\varepsilon(d(x)) \geqslant \zeta_\varepsilon(x)\,. \tag{10}$$

The following proposition tells us that we recover the definition of the overconvergent de Rham–Witt complex with these functions.

PROPOSITION 3.5. — *Let* $x \in W\Omega^{\bullet}_{k[\underline{X}]/k}$. *There exists* $\varepsilon > 0$, *such that* $\gamma_\varepsilon(x) \neq -\infty$ *if and only if* $\zeta_{\varepsilon'}(x) \neq -\infty$, *for some* $\varepsilon' > 0$.

Proof. — Notice that whenever $n \neq 0$, we have:

$$\forall x \in W\Omega^{\bullet}_{k[\underline{X}]/k},\ 2n\gamma_{\frac{\varepsilon}{2n}}(x) \geqslant \zeta_\varepsilon(x) \geqslant \gamma_\varepsilon(x)\,.$$

This ends the proof except when $n = 0$. However, when $n = 0$, then $W\Omega^{\bullet}_{k[\underline{X}]/k} \cong W(k)$ as $W(k)$-dgas, so there is nothing to do. □

We will now prove the product formula. We are doing this by exhaustion using the decomposition (8). Even though most of the proofs below follow the same, simple strategy, it is still interesting to carry them out in detail as one gets stronger formulas in some cases.

PROPOSITION 3.6. — *For any* $\varepsilon > 0$ *and any* $x, y \in W\Omega^{\mathrm{int},\bullet}_{k[\underline{X}]/k}$, *we have:*

$$(\zeta_\varepsilon(x) \neq -\infty \wedge \zeta_\varepsilon(y) \neq -\infty) \implies \zeta_\varepsilon(xy) \geqslant \zeta_\varepsilon(x) + \zeta_\varepsilon(y)\,.$$

Proof. — By definition of $W\Omega^{\mathrm{int},\bullet}_{k[\underline{X}]/k}$, we know that for all $(a, I), (b, J) \in \mathcal{P}$, there exists $\eta_{a,I}, \eta'_{b,J} \in W(k)$, such that:

$$x = \sum_{\substack{(a,I)\in\mathcal{P}\\ u(a)=0}} e(\eta_{a,I}, a, I)\,,$$

$$y = \sum_{\substack{(b,J)\in\mathcal{P}\\ u(b)=0}} e\big(\eta'_{b,J}, b, J\big)\,.$$

For any $(a, I), (b, J) \in \mathcal{P}$, such that $u(a) = u(b) = 0$, using lemma 2.5 we get:

$$\zeta_\varepsilon\big(e(\eta_{a,I}, a, I)\, e\big(\eta'_{b,J}, b, J\big)\big) \geqslant 2n\,\mathrm{v}_V\big(\eta_{a,I}\eta'_{b,J}\big) + (\#I + \#J)\,u(a+b) - \varepsilon|a+b|.$$

Since $u(a+b) = 0$ and $\mathrm{v}_V\Big(\eta_{a,I}\eta'_{b,J}\Big) \geqslant \mathrm{v}_V(\eta_{a,I}) + \mathrm{v}_V\Big(\eta'_{b,J}\Big)$ because k has characteristic p [1, proposition 5. p. IX.15], we can conclude. □

PROPOSITION 3.7. — *Let* $\varepsilon > 0$. *For any* $x \in W\Omega^{\mathrm{int},\bullet}_{k[\underline{X}]/k}$ *and any* $y \in W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}$, *we have:*

$$(\zeta_\varepsilon(x) \neq -\infty \wedge \zeta_\varepsilon(y) \neq -\infty) \implies \zeta_\varepsilon\big((xy)\,|_{\mathrm{d(frp)}}\big) \geqslant \zeta_\varepsilon(x) + \zeta_\varepsilon(y) + 1.$$

Proof. — By definition of integral and pure fractional elements, we know that, for all $(a, I), (b, J) \in \mathcal{P}$, there exists $\eta'_{a,I}, \eta_{b,J} \in W(k)$ such that:

$$x = \sum_{\substack{(b,J)\in\mathcal{P}\\ u(b)=0}} e(\eta_{b,J}, b, J),$$

$$y = \sum_{\substack{(a,I)\in\mathcal{P}\\ u(a)>0\\ I_0\neq\emptyset}} e(\eta'_{a,I}, a, I).$$

Then, for any $(a, I), (b, J) \in \mathcal{P}$, such that $u(a) > 0$, $I_0 \neq \emptyset$ and $u(b) = 0$, lemma 2.7 gives us:

$$\begin{aligned}\zeta_\varepsilon\big(\big(e(\eta_{b,J}, b, J)\, e\big(\eta'_{a,I}, a, I\big)\big)\,|_{\mathrm{d(frp)}}\big)&\\ \geqslant 2n\big(\mathrm{v}_V(\eta_{b,J}) + \mathrm{v}_V\big(\eta'_{a,I}\big) + u(a)\big) &+ (\#I + \#J)\, u(a+b) - \varepsilon|a+b|.\end{aligned}$$

However, $u(a+b) = u(a)$, so

$$\zeta_\varepsilon\big(\big(e(\eta_{b,J}, b, J)\, e\big(\eta'_{a,I}, a, I\big)\big)\,|_{\mathrm{d(frp)}}\big) \geqslant \zeta_\varepsilon(x) + \zeta_\varepsilon(y) + 1,$$

as needed. □

Proposition 3.8. — *For any $\varepsilon > 0$, any $j \in \mathbb{N}$, any $x \in W\Omega^{\mathrm{int},j}_{k[\underline{X}]/k}$ and any $y \in W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}$, we get:*

$$(\zeta_\varepsilon(x) \neq -\infty \wedge \zeta_\varepsilon(y) \neq -\infty) \implies \zeta_\varepsilon((xy)\,|_{\mathrm{frp}}) \geqslant \zeta_\varepsilon(x) + \zeta_\varepsilon(y) + j.$$

Proof. — By definition of integral and pure fractional elements, we know that for all $(a, I), (b, J) \in \mathcal{P}$, there exists $\eta'_{a,I}, \eta_{b,J} \in W(k)$, such that:

$$x = \sum_{\substack{(b,J)\in\mathcal{P}\\ u(b)=0\\ \#J=j}} e(\eta_{b,J}, b, J),$$

$$y = \sum_{\substack{(a,I)\in\mathcal{P}\\ u(a)>0\\ I_0\neq\emptyset}} e(\eta'_{a,I}, a, I).$$

Using lemma 2.7, we know that for any $(a, I), (b, J) \in \mathcal{P}$, such that $u(a) > 0$, $I_0 \neq \emptyset$ and $u(b) = 0$, we have:

$$\begin{aligned}\zeta_\varepsilon\big(\big(e(\eta_{b,J}, b, J)\, e\big(\eta'_{a,I}, a, I\big)\big)\,|_{\mathrm{frp}}\big)&\\ \geqslant 2n\big(\mathrm{v}_V(\eta_{b,J}) + \mathrm{v}_v\big(\eta'_{a,I}\big)\big) &+ (\#I + \#J + 1)\, u(a+b) - \varepsilon|a+b|.\end{aligned}$$

Furthermore, notice that $u(a+b) = u(a) > 0$. Therefore, we obtain that $\zeta_\varepsilon\big(\big(e(\eta_{b,J}, b, J)\, e\big(\eta'_{a,I}, a, I\big)\big)\,|_{\mathrm{frp}}\big) \geqslant \zeta_\varepsilon(x) + \zeta_\varepsilon(y) + \#J$, which ends this proof. □

Proposition 3.9. — *Let $\varepsilon > 0$. For any $x \in W\Omega^{\mathrm{int},\bullet}_{k[\underline{X}]/k}$ and any $y \in W\Omega^{\bullet}_{k[\underline{X}]/k}$, we have:*

$$(\zeta_\varepsilon(x) \neq -\infty \wedge \zeta_\varepsilon(y) \neq -\infty) \implies \zeta_\varepsilon(xy) \geqslant \zeta_\varepsilon(x) + \zeta_\varepsilon(y)\,.$$

Proof. — Recall that $y = y|_{\mathrm{int}} + y|_{\mathrm{frp}} + y|_{\mathrm{d(frp)}}$ and notice that $\zeta_\varepsilon(y|_{\mathrm{int}}) \geqslant \zeta_\varepsilon(y)$, $\zeta_\varepsilon(y|_{\mathrm{frp}}) \geqslant \zeta_\varepsilon(y)$ and $\zeta_\varepsilon\big(y|_{\mathrm{d(frp)}}\big) \geqslant \zeta_\varepsilon(y)$. Therefore, using proposition 3.7 we get:

$$(\zeta_\varepsilon(x) \neq -\infty \wedge \zeta_\varepsilon(y) \neq -\infty) \implies \zeta_\varepsilon\big((x(y|_{\mathrm{frp}}))\,|_{\mathrm{d(frp)}}\big) \geqslant \zeta_\varepsilon(x) + \zeta_\varepsilon(y) + 1.$$

Applying proposition 3.8 and (9) yields:

$$(\zeta_\varepsilon(x) \neq -\infty \wedge \zeta_\varepsilon(y) \neq -\infty) \implies \zeta_\varepsilon((x(y|_{\mathrm{frp}}))\,|_{\mathrm{frp}}) \geqslant \zeta_\varepsilon(x) + \zeta_\varepsilon(y)\,.$$

Using lemma 2.7, we obtain $x(y|_{\mathrm{frp}}) \in W\Omega^{\mathrm{frac},\bullet}_{k[\underline{X}]/k}$. Thus, formula (9) implies that:

$$(\zeta_\varepsilon(x) \neq -\infty \wedge \zeta_\varepsilon(y) \neq -\infty) \implies \zeta_\varepsilon(x(y|_{\mathrm{frp}})) \geqslant \zeta_\varepsilon(x) + \zeta_\varepsilon(y)\,.$$

Moreover, using proposition 3.6 we get:

$$(\zeta_\varepsilon(x) \neq -\infty \wedge \zeta_\varepsilon(y) \neq -\infty) \implies \zeta_\varepsilon(x(y|_{\mathrm{int}})) \geqslant \zeta_\varepsilon(x) + \zeta_\varepsilon(y)\,.$$

By applying (9) once more, we see that it only remains to show that:

$$(\zeta_\varepsilon(x) \neq -\infty \wedge \zeta_\varepsilon(y) \neq -\infty) \implies \zeta_\varepsilon\big(x\big(y|_{\mathrm{d(frp)}}\big)\big) \geqslant \zeta_\varepsilon(x) + \zeta_\varepsilon(y)\,.$$

Again by (9), it is sufficient to prove this in the case where $x \in W\Omega^{i}_{k[\underline{X}]/k}$, for some $i \in \mathbb{N}$, and $y \in W\Omega^{j}_{k[\underline{X}]/k}$, for some $j \in \mathbb{N}$. Let $y' \in W\Omega^{\mathrm{frp},j-1}_{k[\underline{X}]/k}$ be the element such that $d(y') = y|_{\mathrm{d(frp)}}$. Using proposition 1.2 we get $\zeta_\varepsilon(y') = \zeta_\varepsilon\big(y|_{\mathrm{d(frp)}}\big)$. However, by (5) we find:

$$x\big(y|_{\mathrm{d(frp)}}\big) = xd(y') = (-1)^i\left(d(xy') - (-1)^{(i+1)(j-1)}\,y'd(x)\right).$$

So one can conclude using (9), (10) as well as propositions 3.7 and 3.8. ☐

Proposition 3.10. — *For any $\varepsilon > 0$ and any $x, y \in W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}$, we get:*

$$(\zeta_\varepsilon(x) \neq -\infty \wedge \zeta_\varepsilon(y) \neq -\infty) \implies \zeta_\varepsilon((xy)\,|_{\mathrm{frp}}) \geqslant \zeta_\varepsilon(x) + \zeta_\varepsilon(y) + 1.$$

Proof. — Reasoning as in the proof of proposition 3.8, it is enough to prove that for any $(a, I), (b, J) \in \mathcal{P}$ with $u(a) \neq 0$, $I_0 \neq \emptyset$, $u(b) \neq 0$ and $J_0 \neq \emptyset$, and any Witt vectors $\eta_{a,I}, \eta_{b,J} \in W(k)$, we have:

$$\zeta_\varepsilon((e(\eta_{a,I}, a, I)\, e(\eta_{b,J}, b, J))\,|_{\mathrm{frp}}) \geqslant \zeta_\varepsilon(e(\eta_{a,I}, a, I)) + \zeta_\varepsilon(e(\eta_{b,J}, b, J)) + 1.$$

A consequence of lemma 2.7 is that:

$$\begin{aligned}\zeta_\varepsilon((e(\eta_{a,I}, a, I)\, e(\eta_{b,J}, b, J))\,|_{\mathrm{frp}}) &\\ \geqslant 2n(\mathrm{v}_V(\eta_{a,I}) &+ \mathrm{v}_V(\eta_{b,J}) + u(a) + u(b) - u(a+b))\\ &+ (\#I + \#J + 1)\, u(a+b) - \varepsilon|a+b|.\end{aligned}$$

Therefore, one can conclude if one has:

$$2n(u(a)+u(b)-u(a+b))+(\#I+\#J+1)\,u(a+b)\\ \geqslant(\#I+1)\,u(a)+(\#J+1)\,u(b)+1.$$

Notice that $\#I+1\leqslant n$ and $\#J+1\leqslant n$ because we assumed that $I_0\neq\emptyset$ and $J_0\neq\emptyset$. Since $u(a+b)\leqslant\max\{u(a)\,,u(b)\}$, we get:

$$2n(u(a)+u(b))+(\#I+\#J+1-2n)\,u(a+b)\\ \geqslant 2n\min\{u(a)\,,u(b)\}+(\#I+\#J+1)\max\{u(a)\,,u(b)\}\,.$$

This ends the proof whenever $n\neq 0$. If $n=0$, there is nothing to show. □

Proposition 3.11. — *Let $\varepsilon>0$, $x\in W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}$ and $y\in d\Big(W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}\Big)$. We have:*

$$(\zeta_\varepsilon(x)\neq-\infty\wedge\zeta_\varepsilon(y)\neq-\infty)\implies\zeta_\varepsilon((xy)\,|_{\mathrm{frp}})\geqslant\zeta_\varepsilon(x)+\zeta_\varepsilon(y)\,.$$

Proof. — We only have to demonstrate that for any $(a,I)\,,(b,J)\in\mathcal{P}$ with $u(a)\neq 0$, $I_0\neq\emptyset$, $u(b)\neq 0$ and $J_0=\emptyset$, and any Witt vectors $\eta_{a,I},\eta_{b,J}\in W(k)$, we have $\zeta_\varepsilon((e(\eta_{a,I},a,I)\,e(\eta_{b,J},b,J))\,|_{\mathrm{frp}})\geqslant\zeta_\varepsilon(e(\eta_{a,I},a,I))+\zeta_\varepsilon(e(\eta_{b,J},b,J))$. Using proposition 2.8, one obtains:

$$\zeta_\varepsilon((e(\eta_{a,I},a,I)\,e(\eta_{b,J},b,J))\,|_{\mathrm{frp}})\\ \geqslant 2n(\mathrm{v}_V(\eta_{a,I})+\mathrm{v}_V(\eta_{b,J})+\max\{u(a)\,,u(b)\}-u(a+b))\\ +(\#I+\#J+1)\,u(a+b)-\varepsilon|a+b|.$$

So we can conclude if we show:

$$2n(\max\{u(a)\,,u(b)\}-u(a+b))+(\#I+\#J+1)\,u(a+b)\\ \geqslant(\#I+1)\,u(a)+\#Ju(b)\,.$$

Notice that $\#I+1\leqslant n$ because we assumed that $I_0\neq\emptyset$ and $\#J\leqslant n$. As $u(a+b)\leqslant\max\{u(a)\,,u(b)\}$, we can see that:

$$2n\max\{u(a)\,,u(b)\}+(\#I+\#J+1-2n)\,u(a+b)\\ \geqslant(\#I+\#J+1)\max\{u(a)\,,u(b)\}\,.\quad\square$$

Proposition 3.12. — *For any $\varepsilon>0$ and any $x,y\in W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}$, we have:*

$$(\zeta_\varepsilon(x)\neq-\infty\wedge\zeta_\varepsilon(y)\neq-\infty)\implies\zeta_\varepsilon\big((xy)\,|_{\mathrm{d(frp)}}\big)\geqslant\zeta_\varepsilon(x)+\zeta_\varepsilon(y)+3.$$

Proof. — We only have to verify that for any $(a,I)\,,(b,J)\in\mathcal{P}$, such that $u(a)\neq 0$, $I_0\neq\emptyset$, $u(b)\neq 0$ and $J_0\neq\emptyset$, and any $\eta_{a,I},\eta_{b,J}\in W(k)$, we have

$\zeta_\varepsilon\big((e(\eta_{a,I},a,I)\,e(\eta_{b,J},b,J))\,|_{\mathrm{d(frp)}}\big) \geqslant \zeta_\varepsilon(e(\eta_{a,I},a,I))+\zeta_\varepsilon(e(\eta_{b,J},b,J))+3$. Due to lemma 2.7 we can see that:

$$\begin{aligned}&\zeta_\varepsilon\big((e(\eta_{a,I},a,I)\,e(\eta_{b,J},b,J))\,|_{\mathrm{d(frp)}}\big)\\ &\quad\geqslant 2n(\mathrm{v}_V(\eta_{a,I})+\mathrm{v}_V(\eta_{b,J})+u(a)+u(b))+(\#I+\#J)\,u(a+b)-\varepsilon|a+b|.\end{aligned}$$

Therefore, the proof is complete if:

$$2n(u(a)+u(b))+(\#I+\#J)\,u(a+b)\geqslant(\#I+1)\,u(a)+(\#J+1)\,u(b)+3.$$

In the fractional part that we are studying, we necessarily have $u(a)>0$, $u(b)>0$ and $u(a+b)>0$. Moreover, $\#I+1\leqslant n$ and $\#J+1\leqslant n$ as we assumed that $I_0\neq\emptyset$ and $J_0\neq\emptyset$. So we get:

$$\begin{aligned}&2n(u(a)+u(b))+(\#I+\#J)\,u(a+b)\\ &\qquad\qquad\geqslant(\#I+1)\,u(a)+(\#J+1)\,u(b)+2+\#I+\#J.\end{aligned}$$

If $\#I+\#J\neq 0$, the proof is complete. Otherwise, it means that we are multiplying two Witt vectors. In particular, the projection on $d\Big(W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}\Big)$ is 0, but $\zeta_\varepsilon(0)=+\infty$, so the proof becomes obvious. □

Proposition 3.13. — *For any $\varepsilon>0$ and any $x,y\in W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}$, we have:*

$$(\zeta_\varepsilon(x)\neq-\infty\wedge\zeta_\varepsilon(y)\neq-\infty)\implies\zeta_\varepsilon((xy)\,|_{\mathrm{int}})\geqslant\zeta_\varepsilon(x)+\zeta_\varepsilon(y)+2.$$

Proof. — It is enough to prove that for any $(a,I),(b,J)\in\mathcal{P}$, such that $u(a)\neq 0$, $I_0\neq\emptyset$, $u(b)\neq 0$ and $J_0\neq\emptyset$, and for any $\eta_{a,I},\eta_{b,J}\in W(k)$, we have $\zeta_\varepsilon((e(\eta_{a,I},a,I)\,e(\eta_{b,J},b,J))\,|_{\mathrm{int}})\geqslant\zeta_\varepsilon(e(\eta_{a,I},a,I))+\zeta_\varepsilon(e(\eta_{b,J},b,J))+2$. However, using lemma 2.7 one gets:

$$\begin{aligned}&\zeta_\varepsilon((e(\eta_{a,I},a,I)\,e(\eta_{b,J},b,J))\,|_{\mathrm{int}})\\ &\qquad\qquad\geqslant 2n(\mathrm{v}_V(\eta_{a,I})+\mathrm{v}_V(\eta_{b,J})+u(a)+u(b))-\varepsilon|a+b|\end{aligned}$$

because in the integral part, we always have $u(a+b)=0$, which ends the proof as $n\geqslant\#I+1$ and $n\geqslant\#J+1$ since we assumed that $I_0\neq\emptyset$ and $J_0\neq\emptyset$. □

Proposition 3.14. — *For any $\varepsilon>0$ and any $x,y\in W\Omega^{\mathrm{frac},\bullet}_{k[\underline{X}]/k}$, we have:*

$$(\zeta_\varepsilon(x)\neq-\infty\wedge\zeta_\varepsilon(y)\neq-\infty)\implies\zeta_\varepsilon((xy)\,|_{\mathrm{int}})\geqslant\zeta_\varepsilon(x)+\zeta_\varepsilon(y)\,.$$

Proof. — We will first show that for any $(a,I),(b,J)\in\mathcal{P}$, such that $u(a)\neq 0$, $I_0\neq\emptyset$ and $u(b)\neq 0$, and any $\eta_{a,I},\eta_{b,J}\in W(k)$, we always have:

$$\zeta_\varepsilon((e(\eta_{a,I},a,I)\,e(\eta_{b,J},b,J))\,|_{\mathrm{int}})\geqslant\zeta_\varepsilon(e(\eta_{a,I},a,I))+\zeta_\varepsilon(e(\eta_{b,J},b,J))\,.$$

Due to proposition 2.8 we can see that:

$$\begin{aligned}&\zeta_\varepsilon((e(\eta_{a,I},a,I)\,e(\eta_{b,J},b,J))\,|_{\mathrm{int}})\\ &\qquad\qquad\geqslant 2n(\mathrm{v}_V(\eta_{a,I})+\mathrm{v}_V(\eta_{b,J})+\min\{u(a)\,,u(b)\})-\varepsilon|a+b|\end{aligned}$$

because in the integral part we have $u(a+b)=0$, which is only possible if $u(a)=u(b)$. This proves this specific case because $2n \geqslant (\#I+1+\#J)$ if $J_0=\emptyset$, and $2n \geqslant (\#I+1+\#J+1)$ otherwise.

For the general case, notice that if $I_0=\emptyset$, then proposition 1.2 gives us the equality:

$$\begin{aligned} e(\eta_{a,I},a,I)\,e(\eta_{b,J},b,J) &= d(e(\eta_{a,I},a,I\smallsetminus\{\min(a)\})\,e(\eta_{b,J},b,J)) \\ &\quad - (-1)^{\#I-1}\,e(\eta_{a,I},a,I\smallsetminus\{\min(a)\})\,d(e(\eta_{b,J},b,J)). \end{aligned}$$

Therefore, we can conclude using the first paragraph as well as (10). □

Proposition 3.15. — *Let $\varepsilon>0$, $x\in W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}$ and $y\in d\Big(W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}\Big)$. We have:*

$$(\zeta_\varepsilon(x)\neq-\infty\wedge\zeta_\varepsilon(y)\neq-\infty)\implies \zeta_\varepsilon\big((xy)\,|_{\mathrm{d(frp)}}\big)\geqslant\zeta_\varepsilon(x)+\zeta_\varepsilon(y)+1.$$

Proof. — We only have to show that for any $(a,I),(b,J)\in\mathcal{P}$, such that $u(a)\neq0$, $I_0\neq\emptyset$, $u(b)\neq0$ and $J_0=\emptyset$, and any $\eta_{a,I},\eta_{b,J}\in W(k)$, we have $\zeta_\varepsilon\big((e(\eta_{a,I},a,I)\,e(\eta_{b,J},b,J))\,|_{\mathrm{d(frp)}}\big)\geqslant\zeta_\varepsilon(e(\eta_{a,I},a,I))+\zeta_\varepsilon(e(\eta_{b,J},b,J))+1$. Using proposition 2.8 one finds that:

$$\begin{aligned} &\zeta_\varepsilon\big((e(\eta_{a,I},a,I)\,e(\eta_{b,J},b,J))\,|_{\mathrm{d(frp)}}\big) \\ &\qquad\geqslant 2n(\mathrm{v}_V(\eta_{a,I})+\mathrm{v}_V(\eta_{b,J})+\min\{u(a)\,,u(b)\}) \\ &\qquad\quad+(\#I+\#J)\,u(a+b)-\varepsilon|a+b|. \end{aligned}$$

So the proof is over if:

$$2n\min\{u(a)\,,u(b)\}+(\#I+\#J)\,u(a+b)\geqslant(\#I+1)\,u(a)+\#Ju(b)+1.$$

Since $\#I+1\leqslant n$ and $1\leqslant\#J\leqslant n$ because we assumed that $I_0\neq\emptyset$ and $J_0=\emptyset$, and since $u(a+b)\neq0$ because we study the fractional part, this inequality becomes obvious whenever $u(a)=u(b)$; if not, then $u(a+b)=\max\{u(a)\,,u(b)\}$, and we are done. □

Proposition 3.16. — *For any $\varepsilon>0$ and any $x,y\in d\Big(W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}\Big)$, we get:*

$$(\zeta_\varepsilon(x)\neq-\infty\wedge\zeta_\varepsilon(y)\neq-\infty)\implies \zeta_\varepsilon\big((xy)\,|_{\mathrm{d(frp)}}\big)\geqslant\zeta_\varepsilon(x)+\zeta_\varepsilon(y)\,.$$

Proof. — One can suppose without any loss of generality that $x\in W\Omega^i_{k[\underline{X}]/k}$, for some $i\in\mathbb{N}$. Put $y'\in W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}$, such that $d(y')=y$. Using proposition 1.2 we get $\zeta_\varepsilon(y')=\zeta_\varepsilon(y)$. However, $xy=(-1)^i\,d(xy')$, so we can conclude due to (9), (10) as well as proposition 3.11. □

We will now study the cases not treated in the previous statements of this section. Notice that if one takes $x, y \in W\Omega^{\mathrm{int},\bullet}_{k[\underline{X}]/k}$, then $xy \in W\Omega^{\mathrm{int},\bullet}_{k[\underline{X}]/k}$. In particular, $(xy)\,|_{\mathrm{frp}} = (xy)\,|_{\mathrm{d(frp)}} = 0$, which implies:

$$\forall \varepsilon > 0,\ \zeta_\varepsilon((xy)\,|_{\mathrm{frp}}) = \zeta_\varepsilon\big((xy)\,|_{\mathrm{d(frp)}}\big) = +\infty.$$

In a similar fashion, if $x \in W\Omega^{\mathrm{int},\bullet}_{k[\underline{X}]/k}$ and $y \in W\Omega^{\mathrm{frac},\bullet}_{k[\underline{X}]/k}$, lemma 2.7 implies that:

$$\forall \varepsilon > 0,\ \zeta_\varepsilon((xy)\,|_{\mathrm{int}}) = +\infty.$$

Also, if $x, y \in d\Big(W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}\Big)$, then as xy lies in the image of d, we get $(xy)\,|_{\mathrm{frp}} = 0$, which in turn implies that:

$$\forall \varepsilon > 0,\ \zeta_\varepsilon((xy)\,|_{\mathrm{frp}}) = +\infty.$$

The following table compiles all of the propositions that we have shown concerning the function ζ_ε, for any $\varepsilon > 0$ (we will always suppose that $\zeta_\varepsilon(x) \neq -\infty$ and $\zeta_\varepsilon(y) \neq -\infty$).

	$\zeta_\varepsilon((xy)\,\|_{\mathrm{int}}) \geqslant$	$\zeta_\varepsilon((xy)\,\|_{\mathrm{frp}}) \geqslant$	$\zeta_\varepsilon\big((xy)\,\|_{\mathrm{d(frp)}}\big) \geqslant$
$x \in W\Omega^{\mathrm{int},\bullet}_{k[\underline{X}]/k}$ $y \in W\Omega^{\mathrm{int},\bullet}_{k[\underline{X}]/k}$	$\zeta_\varepsilon(x) + \zeta_\varepsilon(y)$	$+\infty$	$+\infty$
$x \in W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}$ $y \in W\Omega^{\mathrm{int},\bullet}_{k[\underline{X}]/k}$	$+\infty$	$\zeta_\varepsilon(x) + \zeta_\varepsilon(y)$	$\zeta_\varepsilon(x) + \zeta_\varepsilon(y) + 1$
$x \in d\Big(W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}\Big)$ $y \in W\Omega^{\mathrm{int},\bullet}_{k[\underline{X}]/k}$	$+\infty$	$\zeta_\varepsilon(x) + \zeta_\varepsilon(y)$	$\zeta_\varepsilon(x) + \zeta_\varepsilon(y)$
$x \in W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}$ $y \in W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}$	$\zeta_\varepsilon(x) + \zeta_\varepsilon(y) + 2$	$\zeta_\varepsilon(x) + \zeta_\varepsilon(y) + 1$	$\zeta_\varepsilon(x) + \zeta_\varepsilon(y) + 3$
$x \in d\Big(W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}\Big)$ $y \in W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}$	$\zeta_\varepsilon(x) + \zeta_\varepsilon(y)$	$\zeta_\varepsilon(x) + \zeta_\varepsilon(y)$	$\zeta_\varepsilon(x) + \zeta_\varepsilon(y) + 1$
$x \in d\Big(W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}\Big)$ $y \in d\Big(W\Omega^{\mathrm{frp},\bullet}_{k[\underline{X}]/k}\Big)$	$\zeta_\varepsilon(x) + \zeta_\varepsilon(y)$	$+\infty$	$\zeta_\varepsilon(x) + \zeta_\varepsilon(y)$

In particular, this proves the main theorem of this paper:

THEOREM 3.17. — *For any $\varepsilon > 0$, the function ζ_ε is a pseudovaluation.*

Proof. — This is straightforward using (9) and the previous table. □

COROLLARY 3.18. — *Let $\varepsilon > 0$. Let $\varphi\colon k[\underline{X}] \to R$ be a surjective morphism of commutative k-algebras. Then:*

$$\zeta_{\varepsilon,\varphi}\colon \begin{array}{rl} W\Omega^\bullet_{R/k} & \to \mathbb{R}\cup\{+\infty,-\infty\} \\ x & \mapsto \sup\{\zeta_\varepsilon(y) \mid y\in\varphi^{-1}(\{x\})\} \end{array}$$

is a pseudovaluation.

Proof. — According to [3, p. 4], this map is a pseudovaluation if and only if $\zeta_{\varepsilon,\varphi}(1) \neq +\infty$. We will show that $\zeta_{\varepsilon,\varphi}(1) \leqslant 0$. Let $y \in W\Omega^\bullet_{k[\underline{X}]/k}$ such that $\zeta_\varepsilon(y) > 0$. Write $y = \sum_{(a,I)\in\mathcal{P}} e(\eta_{a,I}, a, I)$ with $\eta_{a,I} \in W(k)$, for all $(a, I) \in \mathcal{P}$ using theorem 1.5. Then, by definition of ζ_ε, for all $(a, I) \in \mathcal{P}$, such that $\eta_{a,I} \neq 0$, we must have $2n\,\mathrm{v}_V(\eta_{a,I}) + u(a) > 0$. If $n = 0$, this cannot happen, otherwise then either $\mathrm{v}_V(\eta_{a,I}) > 0$ or $u(a) > 0$. In all cases, this implies that $e(\eta_{a,I}, a, I)$ is in the image of V. In turn, y is also in the image of V, and by functoriality of $W\Omega^\bullet_{\bullet/k}$, so is $\varphi(y)$. In particular, $\varphi(y) \neq 1$, so $\zeta_{\varepsilon,\varphi}(1) \leqslant 0$, and we are finished. □

In subsequent papers, we will use these results and this table in order to study the local structure of the overconvergent de Rham–Witt complex, and give an interpretation of F-isocrystals in this context.

Acknowledgements. — This work is a slight generalization of a part of my PhD thesis. As such, I am very indebted to my advisor Daniel Caro. I would also like to thank all the members of the jury Andreas Langer, Tobias Schmidt, Andrea Pulita, Christine Huyghe and Jérôme Poineau for all their comments. I also had the luck to talk about these topics with Bernard Le Stum.

I am also grateful to the two anonymous reviewers, whose many comments greatly contributed to improve the overall quality of this article.

BIBLIOGRAPHY

[1] N. BOURBAKI – *Algèbre commutative : Chapitres 8 et 9*, Springer-Verlag, 2006.

[2] C. DAVIS, A. LANGER & T. ZINK – “Overconvergent de Rham–Witt cohomology”, *Annales Scientifiques de l'École Normale Supérieure quatrième série* **44** (2011), no. 2, p. 197–262.

[3] ______, “Overconvergent Witt vectors”, *Journal für die reine und angewandte Mathematik* **668** (2012), p. 1–34.

[4] V. Ertl – "Comparison between Rigid and Overconvergent Cohomology with Coefficients", arXiv:1310.3237v3, 2016.
[5] L. Illusie – "Complexe de de Rham–Witt et cohomologie cristalline", *Annales scientifiques de l'É.N.S. 4 e série* **12** (1979), no. 4, p. 501–661.
[6] A. Langer & T. Zink – "De Rham–Witt cohomology for a proper and smooth morphism", *Journal of the Institute of Mathematics of Jussieu* **3** (2004), no. 2, p. 231–314.
[7] N. Lawless – "A Comparison of Overconvergent Witt de–Rham Cohomology and Rigid Cohomology on Smooth Schemes", arXiv:1810.10059v1, 2018.

Bull. Soc. Math. France
150 (1), 2022, p. 77-109

FINITENESS AND PERIODICITY OF CONTINUED FRACTIONS OVER QUADRATIC NUMBER FIELDS

by Zuzana Masáková, Tomáš Vávra & Francesco Veneziano

Abstract. — In this paper, we prove a periodicity theorem for certain continued fractions with partial quotients in the ring of integers of a fixed quadratic field. This theorem generalizes the classical theorem of Lagrange to a large set of continued fraction expansions.

As an application we consider the β-continued fractions and show that for any quadratic Perron number β, the β-continued fraction expansion of elements in $\mathbb{Q}(\beta)$ is either finite or eventually periodic.

More generally, we examine the finiteness and periodicity of the β-continued fractions for all quadratic integers β, thus studying problems raised by Rosen and Bernat.

Résumé (*Finitude et périodicité des fractions continues sur des corps de nombres quadratiques*). — Dans cet article, nous prouvons un théorème de périodicité pour certaines fractions continues avec les quotients incomplets dans l'anneau des entiers d'un corps quadratique fixé, qui généralise le théorème classique de Lagrange.

Texte reçu le 29 mars 2021, modifié le 2 novembre 2021, accepté le 29 novembre 2021.

Zuzana Masáková, Department of Mathematics, FNSPE, Czech Technical University in Prague, Trojanova 13, 120 00 Praha 2, Czech Republic • *E-mail :* `zuzana.masakova@fjfi.cvut.cz`

Tomáš Vávra, Department of Pure Mathematics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1 • *E-mail :* `tvavra@uwaterloo.ca`

Francesco Veneziano, Department of mathematics, University of Genova, Via Dodecaneso 35, 16146 Genova, Italy • *E-mail :* `veneziano@dima.unige.it`

Mathematical subject classification (2010). — 11A55, 11J70, 11A63.

Key words and phrases. — Continued fraction, Perron number, quadratic Pisot numbers, finiteness.

This work was supported by the project CZ.02.1.01/0.0/0.0/16_019/0000778 of the Czech Technical University in Prague and projects PRIMUS/20/SCI/002 and UNCE/SCI/022 from Charles University. We thank the Centro di Ricerca Matematica Ennio De Giorgi for support. The third author is a member of the GNSAGA research group of INdAM..

doi:10.24033/bsmf.2845

Comme application, nous considérons les β-fractions continues et montrons que pour tout nombre de Perron quadratique β, le développement en β-fractions continues des éléments dans $\mathbb{Q}(\beta)$ est soit fini, soit éventuellement périodique.

Plus généralement, nous examinons la finitude et la périodicité des β-fractions continues pour tous les entiers quadratiques β, étudiant ainsi des problèmes soulevés par Rosen et Bernat.

1. Introduction

1.1. A periodicity result for continued fractions with partial quotients in a real quadratic field. — As is well known, a theorem due to Lagrange states that the simple continued fraction expansion of quadratic irrationals is periodic. In this paper, we present an extension of Lagrange's theorem, which applies to many different continued fraction expansions. Our main theorem is the following:

Theorem 1.1. — *Let $\xi \in \mathbb{R}$ be a quadratic irrational number, let $\mathbb{Q}(\xi) = \mathbb{Q}(\sqrt{D})$ (where D a positive squarefree integer) be the field generated by ξ over $\mathbb{Q}$ and denote by $\mathcal{O}$ its ring of integers.*

Let $\xi = [a_0, a_1, \dots]$ be any infinite continued fraction converging to ξ and such that the following conditions are satisfied for every $n \geq 0$:

$$a_n \in \mathcal{O}, \qquad a_n \geq 1, \qquad |a'_n| \leq a_n,$$

where a'_n denotes the image of a_n under the nontrivial Galois automorphism of the field $\mathbb{Q}(\xi)$.

Then the sequence $(a_n)_{n\in\mathbb{N}}$ is eventually periodic, and either all partial quotients in the period belong to $\mathbb{Z}$, or they all belong to $\sqrt{D}\mathbb{Z}$.

Moreover, there is an effective constant C_ξ, which bounds from above the lengths of the period and of the pre-period.

If every partial quotient a_i belongs to $\mathbb{Z}$, we recover the statement of the classical theorem of Lagrange. Theorem 1.1, however, also applies to many other representations of ξ as a continued fraction with partial quotients in a quadratic field. We will present a class of such expansions and we will focus in particular on the β-continued fractions; in this setting, we will apply our theorem to study some open problems on the subject.

We remark that there are, indeed, nontrivial examples in which all partial quotients belong to $\sqrt{D}\mathbb{Z}$, as we will see in Section 6.2.

The proof of Theorem 1.1 is given in Section 3 and relies on diophantine approximation and algebraic number theory. An explicit expression of the bound C_ξ is discussed in Remark 3.4. In the Appendix, we present a more elementary argument, which is nevertheless capable of proving a weaker version of the main theorem.

1.2. A question of Rosen and β-continued fractions. — In 1977, Rosen [18] stated the following research problem: "Is it possible to devise a continued fraction that represents uniquely all real numbers, so that the finite continued fractions represent the elements of an algebraic number field, and conversely, every element of the number field is represented by a finite continued fraction"? The classical regular continued fraction has this property with respect to the field of rational numbers.

For $\lambda = 2\cos\frac{\pi}{q}$ with $q \geq 3$ odd, Rosen gave a definition of λ-continued fractions, whose partial quotients are integral multiples of λ. Rosen showed, as a consequence of his own work [17], that if $q = 5$ (i.e. $\lambda = \varphi = \frac{1}{2}(1+\sqrt{5})$ the golden ratio), the λ-continued fraction satisfies his desired property.

A rather different construction was presented by Bernat [2]. He defined φ-continued fractions whose partial quotients belong to the set of the so-called φ-integers, i.e. numbers whose greedy expansion in base φ uses only non-negative powers of the base. Bernat showed that his φ-continued fractions also represent every element of $\mathbb{Q}(\sqrt{5})$ finitely. His proof is established using a very detailed and tedious analysis of the behaviour of φ-integers under arithmetic operations. This approach crucially depends on the arithmetic properties of φ-integers, descending from the fact that φ is a quadratic Pisot number.

It is natural to ask whether the analogously-defined continued fraction expansion based on the β-integers would provide finite representation of $\mathbb{Q}(\beta)$ for any other choice of a quadratic Pisot number β. When trying to adapt Bernat's proof to other values of β, already in the case of the next smallest quadratic Pisot number $\beta = 1+\sqrt{2}$, the necessary analysis becomes even more technical, preventing one from proving the finiteness of the expansions. Bernat remarked that we do not even know whether for some other value of β other than φ, this construction provides at least an eventually periodic representation of all elements of $\mathbb{Q}(\beta)$.

In this paper, we have taken a different approach, considering more general continued fractions whose partial quotients belong to some discrete subset M of the ring of integers in a real quadratic field K. The β-continued fraction of Bernat is a special case of such M-continued fractions when M is chosen to be the set $\mathbb{Z}_\beta$ of β-integers (see Section 5). With the aim of answering Bernat's problem and classifying all quadratic numbers β according to the qualitative behaviour of the β-continued fraction expansion of elements of $\mathbb{Q}(\beta)$, we introduce the following definitions.

Let $\beta > 1$ be a real algebraic integer.

- (CFF) We say that β has the continued fraction finiteness property (CFF), if every element of $\mathbb{Q}(\beta)$ has a finite β-continued fraction expansion.
- (CFP) We say that β has the continued fraction periodicity property (CFP), if every element of $\mathbb{Q}(\beta)$ has a finite or eventually periodic β-continued fraction expansion.

We show that all quadratic Perron numbers and the square roots of positive integers satisfy (CFP) (Theorems 6.4 and 6.9). We prove (CFF) for four quadratic Perron numbers, including the golden ratio φ (Theorem 6.5). Moreover, assuming a conjecture stated by Mercat [14], we show that these four Perron numbers are the only ones with (CFF).

We also consider the case of non-Perron quadratic β. In this case, if the algebraic conjugate of β is positive, we are able to construct a class of elements in $\mathbb{Q}(\beta)$ having aperiodic β-continued fraction expansion, thus showing that neither (CFF) nor (CFP) hold (Theorem 6.10). To achieve a full characterisation of the numbers for which neither (CFF) nor (CFP) hold, it remains to give a complete answer when β is a non-Perron number with a negative algebraic conjugate. Computer experiments suggest that (CFF) is not satisfied by any such β, with some of them having (CFP) and some not. In Theorem 6.14, we are able to construct counterexamples to (CFF) for a large class of such β, but even assuming Mercat's conjecture the matter is not fully settled.

The full picture of what we are able to prove is summarized in Table 1.1, where $\beta > 1$ is a quadratic integer, and β' is its algebraic conjugate.

TABLE 1.1. Entries in black are unconditional, entries in blue assume the validity of Mercat's conjecture.

	(CFP)	(CFF)
$\beta' > \beta$	NO (Theorem 6.10)	NO (Theorem 6.10)
$\lvert\beta'\rvert < \beta$	YES (Theorem 6.4)	YES if $\beta = \frac{1+\sqrt{5}}{2}, 1+\sqrt{2}, \frac{1+\sqrt{13}}{2}, \frac{1+\sqrt{17}}{2}$ NO otherwise (Theorem 6.5 and Corollary 6.7)
$\beta' = -\beta$	YES (Theorem 6.9)	NO (Theorem 6.9)
$-\beta - 3 \leq \beta' < -\beta$	Open	Probably NO but still open for 20 values of β (Table 6.2)
$\beta' \leq -\beta - 4$	Open	NO (Theorem 6.14)

These theorems on β-continued fractions are obtained with a wide range of different techniques, ranging from diophantine approximation to algebraic number theory and dynamics.

1.3. Further lines of inquiry. — One could also study the properties (CFF), (CFP) for algebraic integers β of degree bigger than 2. The issue is considerably more intricate, and already in the cubic Pisot case, we can find instances of different β's that seem to have (CFF), (CFP), and aperiodic β-continued fraction expansions. A modification of some of our arguments may be possible in the case of higher degree fields, but its application is likely to be non-trivial.

Another setting in which the same problem might be stated is that of function fields; a function field analogue of β-continued fractions was studied in [11].

We also mention, as a different path of inquiry that could lead to the study of M-continued fractions for other choices of M, the recent work [4]. There, the authors studied periodic continued fractions with fixed lengths of pre-period and period and with partial quotients in the ring of S-integers of a number field, describing such continued fractions as S-integral points on some suitable affine varieties.

2. Preliminaries

We will denote by $\mathbb{N}$ the set of non-negative integers.

2.1. Continued fractions. — Let $(a_i)_{i\in\mathbb{N}}$ be a sequence of real numbers, such that $a_i > 0$ for $i \geq 1$, and define two sequences $(p_n)_{n\geq -2}$, $(q_n)_{n\geq -2}$ by the linear second-order recurrences

$$\begin{aligned} p_n &= a_n p_{n-1} + p_{n-2}, \quad p_{-1} = 1,\ p_{-2} = 0\,, \\ q_n &= a_n q_{n-1} + q_{n-2}, \quad q_{-1} = 0,\ q_{-2} = 1\,. \end{aligned} \tag{1}$$

This recurrence can be written in matrix form as

$$\begin{pmatrix} p_{n+1} & p_n \\ q_{n+1} & q_n \end{pmatrix} = \begin{pmatrix} p_n & p_{n-1} \\ q_n & q_{n-1} \end{pmatrix} \begin{pmatrix} a_{n+1} & 1 \\ 1 & 0 \end{pmatrix}, \quad n \geq -1\,.$$

Taking determinants, it can be easily shown that

$$p_{n-1}q_n - p_n q_{n-1} = (-1)^n, \quad n \geq -1\,. \tag{2}$$

By induction, one can also show that

$$[a_0, a_1, \ldots, a_n] := a_0 + \cfrac{1}{a_1 + \cfrac{1}{\ddots + \cfrac{1}{a_{n-1} + \cfrac{1}{a_n}}}} = \frac{p_n}{q_n}\,.$$

Notice that the assumption of positivity of a_i ensures that the expression on the right-hand side is well defined for every n.

2.1.1. *Continuants.* — The numerator and denominator p_n, q_n can be expressed in terms of the a_i using the so-called continuants: multivariate polynomials defined by the recurrence

$$\begin{aligned}&K_{-1} = 0, \quad K_0 = 1,\\ &K_n(t_1,\dots,t_n) = t_n K_{n-1}(t_1,\dots,t_{n-1}) + K_{n-2}(t_1,\dots,t_{n-2})\,.\end{aligned}$$

Then $p_n = K_{n+1}(a_0,\dots,a_n)$ and $q_n = K_n(a_1,\dots,a_n)$. It is clear from the definition that K_n is a polynomial with positive integer coefficients in $t_1,\dots,t_n$, and that each of the t_i appears in at least one monomial with a non-zero coefficient.

The continuants satisfy a number of useful properties. We will, in particular, need that $K_n(t_1,\dots,t_n) = K_n(t_n,\dots,t_1)$ and that, for all $k,l \geq 1$, we have

$$\begin{aligned}K_{k+l}(t_1,\dots,t_{k+l}) = {}&K_k(t_1,\dots,t_k)K_l(t_{k+1},\dots,t_{k+l})\\ &+ K_{k-1}(t_1,\dots,t_{k-1})K_{l-1}(t_{k+2},\dots,t_{k+l})\,.\end{aligned} \tag{3}$$

We refer the reader to [10, Chapter X] for the classical theory of continued fractions and the proofs of the properties that we list in this section. Properties of continuants are given in [8, Section 6.7].

2.1.2. *Convergence.* — For an infinite sequence of a_i, $i \geq 1$, the continued fraction $[a_0, a_1, a_2, \dots]$ is defined as the limit

$$[a_0, a_1, a_2, \dots] := \lim_{n\to\infty} \frac{p_n}{q_n}\,, \tag{4}$$

if the limit exists. The numbers a_i are called partial quotients, and the fractions $\frac{p_n}{q_n}$ the convergents of the continued fraction $[a_0, a_1, a_2, \dots]$.

Under our assumptions (positivity of the a_i), the limit (4) exists if and only if $\lim_{n\to\infty} q_n = +\infty$; this is implied by the first equality in (7) below (see [10, Sections 10.7 and 10.8]). A sufficient condition to guarantee the convergence is that $\inf_i a_i > 0$; indeed, if $c = \inf_i a_i$, then $q_n \gg \left(1 + \frac{c}{2}\right)^n$.

2.1.3. *Approximation properties.* — Suppose that the continued fraction converges. For each $i \geq -1$, we have that

$$\xi = a_0 + \cfrac{1}{a_1 + \cfrac{1}{\dots + \cfrac{1}{a_i + \cfrac{1}{\xi_{i+1}}}}} = \frac{\xi_{i+1}p_i + p_{i-1}}{\xi_{i+1}q_i + q_{i-1}}\,, \tag{5}$$

where the $\xi_i = [a_i, a_{i+1}, \dots]$ are the so-called complete quotients.

From (5), we derive

$$\xi - \frac{p_i}{q_i} = \frac{p_{i-1}q_i - q_{i-1}p_i}{q_i(\xi_{i+1}q_i + q_{i-1})} = \frac{(-1)^i}{q_i(\xi_{i+1}q_i + q_{i-1})}\,, \tag{6}$$

which is different from zero; the last equality follows from (2).

As $(q_j)_{j\geq 1}$ tends to infinity, and the convergents p_i/q_i tend to ξ, equation (6) allows us to quantify the rate of convergence with the following estimates:

$$\left|\xi - \frac{p_i}{q_i}\right| = \frac{1}{q_i(\xi_{i+1}q_i + q_{i-1})} \leq \frac{1}{q_i(a_{i+1}q_i + q_{i-1})} = \frac{1}{q_iq_{i+1}} \leq \frac{1}{a_{i+1}q_i^2}\,. \tag{7}$$

If $\inf_i a_i \geq 1$, then $\xi_{i+1} \leq a_{i+1} + 1$, and we also have the lower estimate

$$\left|\xi - \frac{p_i}{q_i}\right| \geq \frac{1}{q_iq_{i+2}}\,.$$

2.1.4. *Uniqueness of the expansion.* — When the partial quotients of a continued fraction $[a_0, a_1, \dots]$ take values in the integers and $a_i \geq 1$ for $i \geq 1$, then the so-called regular continued fractions represent uniquely any real number $\xi = \lim_{n\to\infty} \frac{p_n}{q_n}$, except for the ambiguity $[a_0, \dots, a_n] = [a_0, \dots, a_n - 1, 1]$ in the finite continued fractions representing rational numbers.

Something similar can be shown in a slightly more general setup.

Lemma 2.1. — *Let $m > 0$ and $A \subseteq \mathbb{R}$, such that*

- $a \geq m \quad \forall a \in A$;
- $|a - b| \geq 1/m \quad \forall a, b \in A$ *distinct.*

Then every real number has at most one expression as an infinite continued fraction with partial quotients in A.

Proof. — Let $\xi = [a_0, a_1, a_2, \dots]$ be a convergent continued fraction with partial quotients in A. We have that

$$a_0 < \xi = a_0 + \frac{1}{\xi_1} < a_0 + \frac{1}{a_1} \leq a_0 + \frac{1}{m}\,,$$

which implies

$$\xi - \frac{1}{m} < a_0 < \xi\,.$$

Due to the hypothesis that no two elements of A are less than $1/m$ apart, this identifies uniquely the value of a_0 and thus also that of ξ_1. The statement now follows by induction. □

2.1.5. *Finiteness and periodicity.* — It is a well-known fact that finite regular continued fractions represent precisely the rational numbers. Lagrange's theorem states that irrational quadratic numbers have an eventually periodic regular continued fraction expansion. Galois proved that a quadratic number ξ has a purely periodic regular continued fraction, if and only if $\xi > 1$, and its algebraic conjugate ξ' belongs to the interval $(-1, 0)$.

2.2. The Weil height of algebraic numbers. — The Weil height is an important tool that measures, broadly speaking, the arithmetic complexity of algebraic numbers. It can be described in different ways, but the usual definition involves an infinite product decomposition in local factors.

Let K be a number field and $\mathcal{M}_K$ the set of representatives for the places of K (i.e. equivalence classes of non-trivial absolute values over K), suitably normalized in such a way that a product formula $\prod_{v\in\mathcal{M}_K} |x|_v = 1$ holds for all $x \in K^*$. Then we can define the (multiplicative) Weil height of any $x \in K$ as

$$H(x) = \prod_{v\in\mathcal{M}_K} \max\{1, |x|_v\},$$

where it is easily seen that all but finitely many factors of the infinite product are equal to 1. Due to the choice of the normalization, this definition extends to a function $H : \overline{\mathbb{Q}} \to [1, +\infty)$, where $\overline{\mathbb{Q}}$ denotes the algebraic closure of rational numbers. The function H, so normalized, has the following properties:

Proposition 2.2. — *For all $x, y \in \overline{\mathbb{Q}}$, we have*

(i) $H(x + y) \leq 2H(x)H(y)$;
(ii) $H(xy) \leq H(x)H(y)$;
(iii) $H(x^n) = H(x)^{|n|}$ *for all* $n \in \mathbb{Z}$;
(iv) $H(\sigma(x)) = H(x)$ *for all* $\sigma \in \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$;
(v) $H(x) = 1$ *if and only if x is zero or root of unity (Kronecker's theorem);*
(vi) *for all $B, D > 0$, the set $\{\xi \in \overline{\mathbb{Q}} \mid H(\xi) \leq B$ and $[\mathbb{Q}(\xi) : \mathbb{Q}] \leq D\}$ is finite (Northcott's theorem).*

We refer the reader to the first two chapters of [3] for a thorough introduction to the theory of heights and the proof of the properties listed above.

Remark 2.3. — Let $\xi \in \overline{\mathbb{Q}}$, and let d be the (positive) leading coefficient of the minimal polynomial of ξ over $\mathbb{Z}$. Then

$$H(\xi)^{[\mathbb{Q}(\xi):\mathbb{Q}]} = d \prod_{\sigma:\mathbb{Q}(\xi)\to\overline{\mathbb{Q}}} \max\{1, |\sigma(\xi)|\},$$

where σ in the product ranges over all field homomorphisms from $\mathbb{Q}(\xi)$ into $\overline{\mathbb{Q}}$.

We will only use heights of numbers in real quadratic fields, so we give the exact normalization of the absolute values in this case. Let K be a real quadratic number field and let $\mathfrak{p}$ be a prime ideal of $\mathcal{O}_K$ lying over a prime number p. Then we normalize $|\cdot|_\mathfrak{p}$ in such a way that $|p|_\mathfrak{p} = p^{-1}$, if the prime p is inert or ramifies in $\mathcal{O}_K$, and $|p|_\mathfrak{p} = p^{-1/2}$, if the prime p splits. The two non-Archimedean absolute values are normalized as follows: $|x|_+ = |x|^{1/2}$ and $|x|_- = |x'|^{1/2}$, where by x' we denote the algebraic conjugate of x in K.

3. Continued Fractions over Quadratic Fields

In this section, we focus on continued fractions with positive partial quotients in the ring of integers of a real quadratic field. Let K be a real quadratic field and let $\mathcal{O}_K$ be its ring of integers. For an element $x \in K$ we denote by x' its image under the unique non-trivial automorphism of K. We will prove the main Theorem 1.1.

We start with a lemma controlling the growth of the complete quotients.

Lemma 3.1. — *Let K be a real quadratic field. Let $\xi = [a_0, a_1, \dots]$ be an infinite continued fraction with $a_n \in \mathcal{O}_K$, such that $a_n > 0$ for $n \geq 0$. Assume that $\xi \in K$. Then the height of the complete quotients ξ_n can be estimated as*

$$H(\xi_{n+1}) \leq \frac{H(\xi)}{\min\{q_n, q_{n+1}\}^{1/2}} \max\left\{|\xi' q_n' - p_n'|, |\xi' q_{n-1}' - p_{n-1}'|\right\}^{1/2}.$$

If furthermore $a_n \geq 1$ and $|a_n'| \leq a_n$, for all $n \geq 0$, then

$$\begin{aligned} H(\xi_n) &\leq \sqrt{3} H(\xi), \\ H(a_n) &\leq a_n \leq 3H(\xi)^2, \end{aligned}$$

and, therefore, only finitely many distinct complete quotients and partial quotients may occur.

Proof. — Denote for simplicity $A_n := \xi q_n - p_n \in \mathcal{O}_K + \xi \mathcal{O}_K$. By (7) we have $|A_n| \leq q_{n+1}^{-1}$. From the ultra-metric inequality and the integrality of p_n, q_n, it follows that, for all non-Archimedean absolute values of K,

$$|A_n|_v \leq \max\{|\xi q_n|_v, |p_n|_v\} \leq \max\{|\xi|_v, 1\}. \tag{8}$$

Due to (5), the complete quotients ξ_n can be expressed as

$$\xi_{n+1} = -\frac{A_{n-1}}{A_n} \quad \text{for any } n \geq -1.$$

Consider now, for all $n \geq 0$, the following chain of inequalities

$$\begin{aligned} H(\xi_{n+1}) &= \prod_{v \in \mathcal{M}_K} \max\{|A_{n-1}/A_n|_v, 1\} = \prod_{v \in \mathcal{M}_K} \max\{|A_{n-1}|_v, |A_n|_v\} \\ &= \max\{|A_{n-1}|, |A_n|\}^{1/2} \max\left\{|A_{n-1}'|, |A_n'|\right\}^{1/2} \prod_{v \text{ finite}} \max\{|A_{n-1}|_v, |A_n|_v\} \\ &\leq (\min\{q_n, q_{n+1}\})^{-1/2} \max\left\{|A_n'|, |A_{n-1}'|\right\}^{1/2} H(\xi), \end{aligned}$$

where the second equality follows from the product formula and the last inequality from (8). This proves the first part of the statement.

Under the condition that $a_n \geq 1$ we know that the sequences of the p_n and q_n are nondecreasing, and under the assumption that $|a_n'| \leq a_n$, we have that

$|p_n'| \leq p_n$ and $|q_n'| \leq q_n$. Therefore,

$$\max\{|A_{n-1}|, |A_n|\} \leq \max\left\{q_n^{-1}, q_{n+1}^{-1}\right\} = q_n^{-1},$$

and

$$\max\left\{\left|A_{n-1}'\right|, |A_n'|\right\} \leq \max\{|\xi'|\, q_{n-1} + p_{n-1}, |\xi'|\, q_n + p_n\} = |\xi'|\, q_n + p_n\,.$$

Hence,

$$\begin{aligned} &H(\xi_{n+1}) \\ &= \max\{|A_{n-1}|, |A_n|\}^{1/2} \max\left\{\left|A_{n-1}'\right|, |A_n'|\right\}^{1/2} \prod_{v \text{ finite}} \max\{|A_{n-1}|_v, |A_n|_v\} \\ &\leq \left(|\xi'| + \frac{p_n}{q_n}\right)^{\frac{1}{2}} \prod_{v \text{ finite}} \max\{|\xi|_v, 1\} \leq (|\xi'| + |\xi| + 1)^{1/2} \prod_{v \text{ finite}} \max\{|\xi|_v, 1\} \\ &\leq \sqrt{3} \max\{|\xi'|, 1\}^{1/2} \max\{|\xi|, 1\}^{1/2} \prod_{v \text{ finite}} \max\{|\xi|_v, 1\} = \sqrt{3} H(\xi)\,, \end{aligned}$$

where we used the elementary inequality $1 + a + b \leq 3\max\{1, a\}\max\{1, b\}$, for all $a, b \geq 0$.

Finally,

$$H(a_n) = \max\{1, |a_n|\}^{1/2} \max\{1, |a_n'|\}^{1/2} \leq a_n < \xi_n \leq H(\xi_n)^2\,.$$

Since the heights of the partial and complete quotients are bounded, Northcott's theorem (item (vi) of Proposition 2.2) implies that they can take only finitely many values. □

The following proposition concerns the elements of the field K that can be expressed as a purely periodic continued fraction. By an algebraic argument we show that, in order for the value of the continued fraction to lie in K, the numbers p_n, q_n need to satisfy a certain arithmetic condition.

PROPOSITION 3.2. — *Let $K = \mathbb{Q}(\sqrt{D})$ be a real quadratic field (with D a positive squarefree integer). Let $\xi = [\overline{a_0, a_1, \ldots, a_n}]$ be a purely periodic continued fraction with $a_i \in \mathcal{O}_K$. Assume that $\xi \in K$. Then $p_n + q_{n-1} \in \mathbb{Z} \cup \sqrt{D}\mathbb{Z}$.*

Proof. — Assume that $\xi = [\overline{a_0, \ldots, a_n}]$ is a purely periodic continued fraction with partial quotients in $\mathcal{O}_K$ representing an element in K. Then by (5), ξ satisfies a quadratic equation over K, namely,

$$q_n\xi^2 + (q_{n-1} - p_n)\xi - p_{n-1} = 0\,.$$

The discriminant of this equation,

$$\Delta = (q_{n-1} - p_n)^2 + 4q_n p_{n-1} = (q_{n-1} + p_n)^2 + 4(-1)^n = y^2\,,$$

must be a square of an element y of K, because $\xi \in K$.

Let us write $x = p_n + q_{n-1}$, so we have

$$4(-1)^n = (y+x)(y-x)\,, \tag{9}$$

where $x, y \in \mathcal{O}_K$.

Let us show that this implies that $x' = \pm x$.

If the prime 2 remains inert in K, then 2 must divide $y+x$ or $y-x$ by definition of a prime ideal. Then it must divide the other, because it divides their sum. So we have

$$(-1)^n = \frac{y+x}{2}\frac{y-x}{2}\,,$$

and both $\frac{y+x}{2}, \frac{y-x}{2}$ are units in $\mathcal{O}_K$. Let us write $u = \frac{y+x}{2}$. Then $\frac{y-x}{2} = (-1)^n u^{-1} = (-1)^n \epsilon u'$, where $\epsilon = N_{K/\mathbb{Q}}(u) = \pm 1$. So $x = u - (-1)^n \epsilon u'$, which means that $x' = \pm x$.

If, instead, the prime 2 ramifies as $2\mathcal{O}_K = \mathfrak{p}^2$, then as ideals $\mathfrak{p}^4 = (y+x)(y-x)$, so at least one of $y+x$ or $y-x$ must be divisible by $\mathfrak{p}^2 = (2)$, and the argument proceeds as in the previous case.

If the prime 2 splits as $2\mathcal{O}_K = \mathfrak{p}\mathfrak{p}'$, then as ideals $\mathfrak{p}^2{\mathfrak{p}'}^2 = (y+x)(y-x)$, so either one of $y+x$ or $y-x$ is divisible by (2), and we argue again as above, or (without loss of generality) $\mathfrak{p}^2 = (y+x)$ and ${\mathfrak{p}'}^2 = (y-x)$ as ideals. This means that there is a unit u such that $y - x = u(y'+x')$. Substituting back in (9) we obtain $\pm 4 = uN_{K/\mathbb{Q}}(x+y)$, which implies that $u \in \mathbb{Q}$, and so $u = \pm 1$. Now $y - uy' = x + ux'$ with $u = \pm 1$ implies that $x' = \pm x$. However, clearly, the elements of $\mathbb{Z} \cup \sqrt{D}\mathbb{Z}$ are the only numbers in $\mathcal{O}_K$ satisfying $x' = \pm x$. $\square$

Proposition 3.3. — *Let $K = \mathbb{Q}(\sqrt{D})$ be a real quadratic field. Let $\xi = [\overline{a_0, \ldots, a_n}]$ be a purely periodic continued fraction with $a_k \in \mathcal{O}_K$ and $a_k \geq 1$, for all $k \geq 0$. Assume that $\xi \in K$ and that $|a_k'| \leq a_k$, for $k = 0, \ldots, n$.*

Then either $a_k \in \mathbb{Z}$, for all $k = 0, \ldots, n$, or $a_k \in \sqrt{D}\mathbb{Z}$ for all $k = 0, \ldots, n$.

Proof. — By Proposition 3.2 we have that $p_n + q_{n-1} \in \mathbb{Z} \cup \sqrt{D}\mathbb{Z}$, which implies that $p_n + q_{n-1} = \left|p_n' + q_{n-1}'\right|$. As explained in Section 2.1, $p_n + q_{n-1} = K_{n+1}(a_0, \ldots, a_n) + K_{n-1}(a_1, \ldots, a_{n-1})$ is a polynomial with positive integer coefficients in the variables $a_0, \ldots, a_n$; therefore

$$\begin{aligned} p_n + q_{n-1} = \left|p_n' + q_{n-1}'\right| &\leq K_{n+1}(|a_0'|, \ldots, |a_n'|) + K_{n-1}(|a_1'|, \ldots, \left|a_{n-1}'\right|) \\ &\leq K_{n+1}(a_0, \ldots, a_n) + K_{n-1}(a_1, \ldots, a_{n-1}) = p_n + q_{n-1}\,. \end{aligned}$$

This implies that equality must hold at every point, so, in particular, $|a_k'| = a_k$, for all k, which shows that each a_k is in $\mathbb{Z} \cup \sqrt{D}\mathbb{Z}$.

Assume now that the period $\overline{a_0, \ldots, a_n}$ contains both an $a_i \in \mathbb{Z}$ and an $a_j \in \sqrt{D}\mathbb{Z}$. Take the minimal $k \in \{0, \ldots, n\}$, such that $a_k a_{k+1} \in \sqrt{D}\mathbb{Z}$. Then by (3)

$$\begin{aligned}p_n &= K(a_0,\dots,a_k)K(a_{k+1},\dots,a_n) + K(a_0,\dots,a_{k-1})K(a_{k+2},\dots,a_n)\\ &= \big(a_kK(a_0,\dots,a_{k-1}) + K(a_0,\dots,a_{k-2})\big)\big(a_{k+1}K(a_{k+2},\dots,a_n)\\ &\quad + K(a_{k+3},\dots,a_n)\big) + K(a_0,\dots,a_{k-1})K(a_{k+2},\dots,a_n)\\ &= (a_ka_{k+1}+1)K(a_0,\dots,a_{k-1})K(a_{k+2},\dots,a_n)\\ &\quad + K(a_0,\dots,a_{k-2})K(a_{k+3},\dots,a_n) + a_kK(a_0,\dots,a_{k-1})K(a_{k+3},\dots,a_n)\\ &\quad + a_{k+1}K(a_0,\dots,a_{k-2})K(a_{k+2},\dots,a_n)\,.\end{aligned}$$

Note that we have omitted the indices of the continuants since they are clear from the context. By assumption, we have $a_ka_{k+1}+1 \in \mathbb{Z}^+ + \sqrt{D}\mathbb{Z}^+$. Since all the partial quotients are positive, we see that $x = p_n + q_{n-1} \in \mathbb{Z}^+ + \sqrt{D}\mathbb{Z}^+$. Then obviously, $|x'| < x$, which is a contradiction. $\square$

Proof of Theorem 1.1. — We are in the hypotheses of the second part of Lemma 3.1, so only finitely many distinct complete quotients occur in the continued fraction $[a_0, a_1, a_2, \dots]$. Note that their number can be effectively bounded in a way that depends only on the height of ξ and not on the number field K, because the degree of K over $\mathbb{Q}$ is fixed, see item (vi) of Proposition 2.2. Necessarily at least one complete quotient occurs infinitely many times. Consider any two occurrences of the same complete quotient, say $\xi_r = \xi_{r+s}$. Then we have

$$\xi_r = [a_r, a_{r+1}, \dots] = [a_r, a_{r+1}, \dots, a_{r+s-1}, \xi_{r+s}] = [\overline{a_r, a_{r+1}, \dots, a_{r+s-1}}]\,.$$

By Proposition 3.3, we have $a_{r+i} \in \mathbb{Z}$ for all $i = 0, \dots, s-1$, or $a_{r+i} \in \sqrt{D}\mathbb{Z}$, for all $i = 0, \dots, s-1$.

Since the same consideration can be done for any two occurrences of the complete quotient ξ_r, we derive that there exist k_0 such that $a_k \in \mathbb{Z}$, for all $k \geq k_0$, or $a_k \in \sqrt{D}\mathbb{Z}$, for all $k \geq k_0$. In both cases, we have $\xi_r = [a_{k_0}, a_{k_0+1}, \dots]$, with partial quotients in a discrete set with distances between consecutive elements ≥ 1. Such a sequence represents the number ξ uniquely, see Lemma 2.1. Therefore the continued fraction of ξ_r is purely periodic, and the continued fraction of ξ is eventually periodic. $\square$

Remark 3.4. — A possible value for the constant C_ξ of Theorem 1.1 is given by $\#\{x \in K \mid H(x) \leq \sqrt{3}H(\xi)\}$. We remark that for a real quadratic field K, the cardinality of the finite set $\{x \in K \mid H(x) \leq B\}$ is asymptotic (for large B) to c_KB^4, where c_K has an explicit expression in terms of the discriminant, regulator, number of ideal classes, number of roots of unity and the Dedekind zeta function of the field K (see [19, Corollary]).

A bound that does not depend on the field K is given by

$$\#\{x \in \overline{\mathbb{Q}} \mid H(x) \leq B \text{ and } [\mathbb{Q}(x):\mathbb{Q}] = 2\} \leq \frac{8}{\zeta(3)} B^6 + 16690 B^4 \log B\,,$$

for $B \geq \sqrt{2}$ (see [9, Theorem 11.1]).

In the Appendix, we present a different proof of Theorem 1.1 in the case that $|a_n'| < a_n$, which gives a better estimate for the number of irrational partial quotients.

4. M-Continued Fractions

In the classical theory of continued fractions, the partial quotients are taken to be positive integers (except possibly the first one), and there is a canonical algorithm defined through the Gauss map that attaches to every $\xi \in \mathbb{R}$ a continued fraction converging to ξ. We now define a different expansion, using, instead of the classical integral part, the M-integral part for certain subsets M of the real line.

Let $M = (m_n)_{n \in \mathbb{Z}}$ be an infinite subset of the reals without any accumulation points, enumerated in increasing order. Assume that $m_0 = 0$, $m_1 \leq 1$ and $m_{n+1} - m_n \leq 1/m_1$, for all $n \in \mathbb{Z}$.

For $\xi \in \mathbb{R}$, we define its M-integral part and its M-fractional part as

$$\lfloor \xi \rfloor_M := \max\{y \in M : y \leq \xi\}, \qquad \{\xi\}_M := \xi - \lfloor \xi \rfloor_M\,,$$

respectively.

The M-continued fraction expansion of a number $\xi \in \mathbb{R}$ is now defined as a direct generalization of the regular continued fraction expansion obtained if $M = \mathbb{Z}$, as explained below.

Let $\xi \in \mathbb{R}$. Set $\xi_0 := \xi$. For $n \geq 0$ define inductively:

$$\begin{aligned} a_n &= \lfloor \xi_n \rfloor_M, \\ \xi_{n+1} &= \frac{1}{\xi_n - a_n} && \text{if } \xi_n \notin M\,, \\ p_n &= a_n p_{n-1} + p_{n-2}, && p_{-1} = 1,\ p_{-2} = 0\,, \\ q_n &= a_n q_{n-1} + q_{n-2}, && q_{-1} = 0,\ q_{-2} = 1\,. \end{aligned} \tag{10}$$

If some $\xi_n \in M$, then the algorithm stops, and we say that the M-continued fraction $[a_0, \ldots, a_n]$ is finite. Note that conditions on the set M ensure that $a_n > 0$ for $n \geq 1$.

Remark 4.1. — In the classical case ($M = \mathbb{Z}$), it is possible to prove that any infinite sequence of positive integers is the continued fraction expansion of some real number. For a general set M, it might happen that certain sequences of positive partial quotients in M are not allowed, i.e. they never occur in the output of the iteration (10), for any $\xi \in \mathbb{R}$. We will see an example of this later in Lemma 5.6.

Remark 4.2. — It is clear that if the M-continued fraction expansion of ξ is finite, then $\xi \in \mathbb{Q}(M)$. If the continued fraction expansion of ξ is eventually periodic, then ξ is quadratic over $\mathbb{Q}(M)$.

5. β-Integers and β-Continued Fractions

5.1. β-integers. — In view of generalizing the result of [2], one of the interests of the present paper is to describe properties of M-continued fraction expansions for a special class of sets M, the so-called β-integers, as defined in [5]. Consider a real base $\beta > 1$. Any real x can be expanded in the form $x = \pm\sum_{i=-\infty}^{k} x_i\beta^i$, where the digits $x_i \in \mathbb{Z}$ satisfy $0 \le x_i < \beta$. Under the condition that the inequality

$$\sum_{i=-\infty}^{j} x_i\beta^i < \beta^{j+1} \tag{11}$$

holds for each $j \le i$, we have that such a representation is unique up to the leading zeroes; this is called the greedy β-expansion of x, see Rényi [16]. For the greedy expansion of x, we write

$$(x)_\beta = x_k x_{k-1}\cdots x_0 \bullet x_{-1}x_{-2}\cdots$$

If $\beta \notin \mathbb{N}$, then not every sequence of digits in $\{k \in \mathbb{N} : 0 \le x < \beta\}$ corresponds to the greedy β-expansion of a real number. Admissibility of digit sequences as β-expansions is described by the so-called Parry condition [15].

The β-expansions respect the natural order of real numbers in the radix ordering. In particular, if $x = \sum_{i=-\infty}^{k} x_i\beta^i$ and $y = \sum_{i=-\infty}^{l} y_i\beta^i$ with $x_k, y_l \neq 0$ are the β-expansions of x, y, respectively, then $x < y$ if and only if $k < l$, or $k = l$ and $x_k x_{k-1}\cdots$ is lexicographically smaller than $y_k y_{k-1}\cdots$.

A real number x is called a β-integer if the greedy β-expansion of its absolute value $|x|$ uses only non-negative powers of the base β; we denote by $\mathbb{Z}_\beta$ the set of β-integers and by $\mathbb{Z}_\beta^+$ the set of non-negative β-integers, i.e.

$$\mathbb{Z}_\beta = \{x \in \mathbb{R} : (|x|)_\beta = x_k x_{k-1}\cdots x_0 \bullet 00\cdots\},$$
$$\mathbb{Z}_\beta^+ = \{x \in \mathbb{R} : (x)_\beta = x_k x_{k-1}\cdots x_0 \bullet 000\cdots\}.$$

If the base β is in $\mathbb{Z}$, the β-integers are rational integers, i.e. $\mathbb{Z}_\beta = \mathbb{Z}$. Otherwise, it is an aperiodic set of points that can be ordered into a sequence $(t_j)_{j=-\infty}^{\infty}$, such that $t_i < t_{i+1}$ for $i \in \mathbb{Z}$ and $t_0 = 0$. The smallest positive β-integers are

$$1,\ 2, \dots, \lfloor\beta\rfloor,\ \beta,\ \dots$$

We can derive the following property of β-integers.

LEMMA 5.1. — *Let $\beta > 1$ be an algebraic number and let σ be a Galois embedding of $\mathbb{Q}(\beta)$ in $\mathbb{C}$.*

(i) *Assume that $|\sigma(\beta)| < \beta$. Then for any $x \in \mathbb{Z}_\beta^+ \setminus \{0, 1, \dots, \lfloor\beta\rfloor\}$, we have*
$$\frac{|\sigma(x)|}{x} \le \frac{\lfloor\beta\rfloor + |\sigma(\beta)|}{\lfloor\beta\rfloor + \beta} < 1\,.$$

(ii) *Assume that $\sigma(\beta) \in \mathbb{R}$ and $\sigma(\beta) > \beta$. Then for any $x \in \mathbb{Z}_\beta^+ \setminus \{0, 1, \dots, \lfloor\beta\rfloor\}$, we have*
$$\frac{\sigma(x)}{x} \ge \frac{\lfloor\beta\rfloor + \sigma(\beta)}{\lfloor\beta\rfloor + \beta} > 1\,.$$
Moreover, if $\lfloor\sigma(\beta)\rfloor > \lfloor\beta\rfloor$, then for any pair $x, y \in \mathbb{Z}_\beta$, $x \neq y$, we have $|\sigma(x) - \sigma(y)| \ge 1$.

(iii) *Assume that $c = \frac{|\sigma(\beta)|}{\beta} > 1 + 2\frac{\lfloor\beta\rfloor}{\beta - 1}$. Then for any $x \in \mathbb{Z}_\beta^+ \setminus \{0, 1, \dots, \lfloor\beta\rfloor\}$, we have*
$$\frac{|\sigma(x)|}{x} > \frac{c(\beta - 1) - \lfloor\beta\rfloor}{\beta - 1 + \lfloor\beta\rfloor} > 1\,.$$

(iv) *Assume that $|\sigma(\beta)| = \beta$. Then for any $x \in \mathbb{Z}_\beta^+$, we have $|\sigma(x)| \le x$.*

In cases (i),(ii) *and* (iii), *we have that $\mathbb{Z}_\beta^+ \cap \mathbb{Q} = \{0, 1, \dots, \lfloor\beta\rfloor\}$.*

Proof. — Let $x = \sum_{i=0}^k x_i\beta^i \in \mathbb{Z}_\beta^+ \setminus \{0, 1, \dots, \lfloor\beta\rfloor\}$, so that $x_0 \le \lfloor\beta\rfloor$, and $\sum_{i=1}^k x_i\beta^i \ge \beta$. Let us prove the first item (i). Denote $c := |\sigma(\beta)|/\beta < 1$. Then for the algebraic conjugate $\sigma(x)$ of x, we have
$$\begin{aligned}\frac{|\sigma(x)|}{x} &= \frac{\left|x_0 + \sum_{i=1}^k x_i\sigma(\beta)^i\right|}{x_0 + \sum_{i=1}^k x_i\beta^i} \le \frac{x_0 + \sum_{i=1}^k x_i c^i\beta^i}{x_0 + \sum_{i=1}^k x_i\beta^i}\\ &\le \frac{x_0 + c\sum_{i=1}^k x_i\beta^i}{x_0 + \sum_{i=1}^k x_i\beta^i} \le \frac{\lfloor\beta\rfloor + c\beta}{\lfloor\beta\rfloor + \beta} = \frac{\lfloor\beta\rfloor + |\sigma(\beta)|}{\lfloor\beta\rfloor + \beta} < 1\,.\end{aligned}$$

Very similarly, to prove part (ii) denote $c := \sigma(\beta)/\beta > 1$. Then for the algebraic conjugate $\sigma(x)$ of x, we have
$$\begin{aligned}\frac{\sigma(x)}{x} &= \frac{x_0 + \sum_{i=1}^k x_i\sigma(\beta)^i}{x_0 + \sum_{i=1}^k x_i\beta^i} = \frac{x_0 + \sum_{i=1}^k x_i c^i\beta^i}{x_0 + \sum_{i=1}^k x_i\beta^i}\\ &\ge \frac{x_0 + c\sum_{i=1}^k x_i\beta^i}{x_0 + \sum_{i=1}^k x_i\beta^i} \ge \frac{\lfloor\beta\rfloor + c\beta}{\lfloor\beta\rfloor + \beta} = \frac{\lfloor\beta\rfloor + \sigma(\beta)}{\lfloor\beta\rfloor + \beta} > 1\,.\end{aligned}$$

Suppose now that $\lfloor\sigma(\beta)\rfloor = M > m = \lfloor\beta\rfloor$. We shall estimate the distance $|\sigma(x) - \sigma(y)|$ of two distinct β-integers $x = \sum_{i=0}^k x_i\beta^i, y = \sum_{i=0}^l y_i\beta^i$. If

$k = l = 0$, obviously $|\sigma(x) - \sigma(y)| \geq 1$. Otherwise, we can without loss of generality assume that $k > l$, $x_k > 0$. We have

$$|\sigma(x) - \sigma(y)| = \Big|\sum_{i=0}^{k} x_i\sigma(\beta)^i - \sum_{i=0}^{l} y_i\sigma(\beta)^i\Big| \geq \sigma(\beta)^k - \sum_{i=0}^{k-1} m\sigma(\beta)^i\,. \tag{12}$$

Here, we have used that $m = \lfloor\beta\rfloor$ is the maximal digit allowed in the β-expansion. Now consider the expansion in base $\sigma(\beta)$. The maximal allowed digit is $M = \lfloor\sigma(\beta)\rfloor > m$. It can be easily derived from the Parry condition [15] that $\sum_{i=1}^{k-1} m\sigma(\beta)^i + (m+1)$ is an admissible greedy $\sigma(\beta)$-expansion. By definition (11), we have

$$\sum_{i=1}^{k-1} m\sigma(\beta)^i + (m+1) < \sigma(\beta)^k\,,$$

which is equivalent to

$$\sigma(\beta)^k - \sum_{i=0}^{k-1} m\sigma(\beta)^i > 1\,.$$

Comparing with (12) we have the statement of item (ii).

In the same way as the proof of part (i), we can show part (iii) as follows:

$$\begin{aligned}\frac{|\sigma(x)|}{x} &\geq \frac{x_k(c\beta)^k - \left|\sum_{i=0}^{k-1} x_i\sigma(\beta)^i\right|}{x_k\beta^k + \sum_{i=0}^{k-1} x_i\beta^i} \geq \frac{(c\beta)^k - \left|\sum_{i=0}^{k-1} x_i\sigma(\beta)^i\right|}{\beta^k + \sum_{i=0}^{k-1} x_i\beta^i}\\ &\geq \frac{(c\beta)^k - \sum_{i=0}^{k-1} x_i c^i\beta^i}{\beta^k + \sum_{i=0}^{k-1} x_i\beta^i} \geq c^{k-1}\frac{c\beta^k - \sum_{i=0}^{k-1} x_i\beta^i}{\beta^k + \sum_{i=0}^{k-1} x_i\beta^i}\\ &\geq \frac{c\beta^k - \lfloor\beta\rfloor\frac{\beta^k-1}{\beta-1}}{\beta^k + \lfloor\beta\rfloor\frac{\beta^k-1}{\beta-1}} > \frac{c - \frac{\lfloor\beta\rfloor}{\beta-1}}{1 + \frac{\lfloor\beta\rfloor}{\beta-1}} = \frac{c(\beta-1) - \lfloor\beta\rfloor}{\beta - 1 + \lfloor\beta\rfloor} > 1\,.\end{aligned}$$

The part (iv) is easily proved, by checking that

$$|\sigma(x)| = \Big|\sum_{i=0}^{k} x_i\sigma(\beta)^i\Big| \leq \sum_{i=0}^{k} x_i\,|\sigma(\beta)|^i = x\,.$$

In order to show that in cases (i) and (ii), it holds that $\mathbb{Z}_\beta^+ \cap \mathbb{Q} = \{0, 1, \ldots, \lfloor\beta\rfloor\}$, we realize that in these cases, any $x \in \mathbb{Z}_\beta^+ \setminus \{0, 1, \ldots, \lfloor\beta\rfloor\}$ is not fixed by the Galois isomorphism, and therefore it cannot be rational. $\square$

Remark 5.2. — Notice that the final assertion of the lemma, namely that $\mathbb{Z}_\beta^+ \cap \mathbb{Q} = \{0, 1, \ldots, \lfloor\beta\rfloor\}$, does not hold in general. Obvious counterexamples are given by β's, which are square roots of rational integers, in which case all even powers of β are in $\mathbb{Z}_\beta^+ \cap \mathbb{Q}$. Less trivial counterexamples can occur for some quadratic β's with $\beta' < -\beta$. For instance, if β is the positive root of $X^2 + 2X - 9 = 0$, then $\beta^2 + 2\beta = 9 \in \mathbb{Z}_\beta^+ \cap \mathbb{Q}$.

More generally, if β is the positive root of $X^2 + mX - (2m^2 + 1)$, for $m \geq 2$, then it can be shown that $\lfloor \beta \rfloor = m$ and that $\beta^2 + m\beta = 2m^2 + 1 \in \mathbb{Z}_\beta^+ \cap \mathbb{Q}$.

A class of algebraic numbers satisfying the assumptions of Lemma 5.1 (i), is formed by Perron numbers, i.e. algebraic integers $\beta > 1$, such that every conjugate $\sigma(\beta)$ of β satisfies $|\sigma(\beta)| < \beta$. Perron numbers appear as dominant eigenvalues of primitive integer matrices. A special subclass of Perron numbers are Pisot numbers, i.e. algebraic integers $\beta > 1$ whose conjugates lie in the interior of the unit disc. Note that when β is chosen to be a Pisot number, the Galois conjugates of the β-integers are uniformly bounded, and, consequently, the β-integers enjoy many interesting properties, especially from the arithmetical point of view; see, e.g., [7] or [1]. We also mention that Perron numbers in any given number field form a discrete subset.

5.2. β-continued fractions. — When setting M to be the set $\mathbb{Z}_\beta$ of β-integers, it is clear that any finite β-continued fraction belongs to the field $\mathbb{Q}(\beta)$. The opposite, however, is not obvious, as we will see in the sequel. For that, we define properties (CFF) and (CFP).

DEFINITION 5.3. — For a real number $\beta > 1$ set $M = \mathbb{Z}_\beta$. The M-continued fraction in this case is said to be the β-continued fraction. We say that a real number $\beta > 1$ has the continued fraction finiteness property (CFF), if every element of $\mathbb{Q}(\beta)$ has a finite β-continued fraction expansion. We say that $\beta > 1$ has the continued fraction periodicity property (CFP), if every element of $\mathbb{Q}(\beta)$ has either finite or eventually periodic β-continued fraction expansion.

Note that properties (CFF) and (CFP) could also be studied for other sets M.

REMARK 5.4. — If $\beta \in \mathbb{N}$, then $\mathbb{Z}_\beta = \mathbb{Z}$, and the β-continued fraction expansion is the classical regular continued fraction expansion; this implies that integer β's satisfy (CFF).

REMARK 5.5. — Let $\beta > 1$ and $\xi > 0$. If the regular continued fraction expansion of ξ only involves partial quotients strictly smaller than $\lfloor \beta \rfloor$, then it coincides with the β-continued fraction expansion of ξ.

Based on the above remark, any sequence of integers in $\{1, 2, \dots, \lfloor \beta \rfloor - 1\}$ is a β-continued fraction expansion of a real number x. In general, however, not any sequence of β-integers will occur as some β-continued fraction expansion. The following statement describes a sequence of partial quotients in $\mathbb{Z}_\beta$ that is not admissible in a β-continued fraction expansion.

LEMMA 5.6. — *Let $\beta > 1$ satisfy $\beta - \lfloor\beta\rfloor \leq \beta^{-1}$. Let a_i, a_{i+1} be two consecutive partial quotients in the β-continued fraction expansion of a real number ξ. If $a_i = \lfloor\beta\rfloor$, then $a_{i+1} \geq \beta$.*

Proof. — If $a_i = \lfloor\xi_i\rfloor_{\mathbb{Z}_\beta} = \lfloor\beta\rfloor$, then $\lfloor\beta\rfloor < \xi_i < \beta$. Thus $\xi_{i+1} = (\xi_i - \lfloor\beta\rfloor)^{-1} > (\beta - \lfloor\beta\rfloor)^{-1} \geq \beta$. □

In what follows, we will study property (CFF) in quadratic fields. For quadratic Pisot units β, the set of all rules for admissibility of strings of partial quotients is given in [6]. Let us mention that a similar study can already be found in [12]. In Lemma 5.6, we have cited only the rule that will be needed later.

6. Properties of β-Continued Fractions for Quadratic Numbers

We aim to characterize which quadratic integers have properties (CFF) and (CFP). We will apply the general results of Section 3 to the set $M = \mathbb{Z}_\beta$ defined above. In view of Remark 5.5, it is clear that results on the classical regular continued fractions can have implications on β-continued fractions for β big enough. In this spirit, the following proposition shows that the (CFF) property among quadratic numbers is rather rare.

PROPOSITION 6.1. — *For every real quadratic field K, there is a positive bound $m_K > 1$, such that no irrational number $\beta > m_K$ in K has property* (CFF)*. In particular, in a given real quadratic field K, only finitely many Perron numbers can have property* (CFF)*.*

Proof. — Let $\xi \in K \setminus \mathbb{Q}$. The regular continued fraction expansion of ξ is eventually periodic. Denote by c the maximum of the partial quotients appearing in the periodic part. Let $\beta > c + 1$ be an element of $\mathcal{O}_K \setminus \mathbb{Z}$. Then β does not have (CFF), because by Remark 5.5 the complete quotient of ξ defined by the purely periodic tail does not have a finite β-continued fraction expansion. We can thus take $m_K = c + 1$. □

The matter of studying continued fractions with small partial quotients is already very complicated for the classical regular continued fractions. In [13], where the issue under the point of view of dynamics and geodesics on arithmetic manifolds is studied, McMullen poses the following problem:

PROBLEM ([13, p.22]). — *Does every real quadratic field contain infinitely many periodic continued fractions with partial quotients equal to 1 or 2?*

Mercat conjectures an affirmative answer to a weaker version of this problem:

CONJECTURE 6.2 ([14, Conjecture 1.6]). — *Every real quadratic field contains a periodic continued fraction with partial quotients equal to 1 or 2.*

Remark 6.3. — If $K = \mathbb{Q}(\sqrt{d})$, then standard estimates on the continued fraction expansion of $\lfloor\sqrt{d}\rfloor + \sqrt{d}$ imply that for the constant m_K of Proposition 6.1, we can take $m_K = 2\lfloor\sqrt{d}\rfloor + 1$; see [14, Proposition 7.11].

If Mercat's conjecture is true, we can take $m_K = 3$, for all quadratic fields K.

6.1. Pisot numbers and Perron numbers. — Bernat [2] showed that the golden ratio $\varphi = \frac{1}{2}(1+\sqrt{5})$ has property (CFF); we will now study whether other quadratic Pisot numbers with (CFF) can be found. The quadratic Pisot numbers can be characterized as the larger roots of one of the polynomials with integer coefficients

$$X^2 - aX - b,\ a \geq b \geq 1, \qquad X^2 - aX + b,\ a \geq b + 2 \geq 3\,.$$

The golden ratio φ with minimal polynomial $x^2 - x - 1$ is the smallest among quadratic Pisot numbers.

We consider the more general quadratic Perron numbers. It is not difficult to see that these are the larger roots of the polynomials

$$X^2 - aX - b,\ a \geq 1, \quad \text{such that } a^2 + 4b > 0,\ \sqrt{a^2 + 4b} \notin \mathbb{Q}\,.$$

In view of Remark 6.3, we will focus on quadratic Perron numbers smaller than 3.

An intermediate step towards establishing property (CFF) is formulated in the following statement, which is a special case of Theorem 1.1.

Theorem 6.4. — *Let β be a quadratic Perron number. Then the β-continued fraction of any $\xi \in \mathbb{Q}(\beta)$ contains at most finitely many partial quotients in $\mathbb{Z}_\beta \backslash \mathbb{Z}$ and, thus, it is either finite or eventually periodic with all partial quotients in the period being rational integers.*

In particular, any quadratic Perron number satisfy (CFP).

Proof. — It suffices to check that the assumptions of Theorem 1.1 are satisfied, if the β-continued fraction expansion of ξ is not finite. Possibly after replacing ξ by the first complete quotient ξ_1, this is shown by item (i) of Lemma 5.1. By Theorem 1.1, the partial quotients in the period belong to $\mathbb{Z}$ or $\sqrt{D}\mathbb{Z}$. The second possibility is, however, not possible, again by item (i) of Lemma 5.1. □

We apply Theorem 6.4 to establish (CFF) for the four smallest quadratic Perron numbers. Note that Bernat [2] showed the statement for φ by completely different methods.

Theorem 6.5. — *The four Perron numbers*

$$\varphi = \frac{1+\sqrt{5}}{2}, \qquad 1+\sqrt{2}, \qquad \frac{1+\sqrt{13}}{2}, \qquad \frac{1+\sqrt{17}}{2}$$

have property (CFF).

Proof. — Let β be one of the four Perron numbers above. From Theorem 6.4, we derive that the β-continued fraction expansion of any $\xi \in \mathbb{Q}(\beta)$ is either finite, or eventually periodic, with the partial quotients appearing in the period being integers smaller than β.

If $\beta = \varphi$, then the only possible periodic tail is $[\overline{1}]$, but by Lemma 5.6 it is not an admissible φ-continued fraction expansion.

If β is one of the other three values appearing in the statement, then $2 < \beta < \frac{18}{7}$, and the partial quotients in a periodic tail necessarily belong to $\{1, 2\}$.

Assume that the period contains at least one partial quotient equal to 1 and one equal to 2. Up to replacing ξ by one of its complete quotients, we can assume that the β-continued fraction expansion of ξ begins with $[2, 1, a_2, \dots]$, so that writing ξ_3 for the third complete quotient we have

$$\xi = [2, 1, a_2, \xi_3] = 2 + \cfrac{1}{1 + \cfrac{1}{a_2 + \cfrac{1}{\xi_3}}} = \frac{3a_2\xi_3 + 2\xi_3 + 3}{a_2\xi_3 + \xi_3 + 1},$$

where $a_2 \in \{1, 2\}$, and $1 < \xi_3 < 3$. This rational expression is easily seen to be strictly increasing with a_2 and strictly decreasing with ξ_3, so we have that $\xi > \frac{18}{7} > \beta$; this is a contradiction, because then the expansion would not start with a 2.

The only possible periodic tails are, therefore, $[\overline{1}] = \frac{1+\sqrt{5}}{2}$ and $[\overline{2}] = 1 + \sqrt{2}$. The first possibility is excluded because $\varphi \notin \mathbb{Q}(\beta)$; for the same reason, the second possibility could only occur if $\beta = 1 + \sqrt{2}$, and in this case, Lemma 5.6 again shows that $[\overline{2}]$ cannot be a β-continued fraction expansion.

We conclude that every $\xi \in \mathbb{Q}(\beta)$ has a finite β-continued fraction expansion. □

Remark 6.6. — Notice that the first three numbers in the statement of Theorem 6.5 satisfy the hypothesis of Lemma 5.6. This allows us, in the proof of the theorem, to argue that the periodic tails cannot contain any partial quotient equal to 2, which leads to a quicker conclusion. However, this argument does not work for $\beta = \frac{1+\sqrt{17}}{2}$. For example,

$$\frac{164 + 65\sqrt{17}}{251} = [\overline{1, 1, 2, 1, 1, 2, 2, 2, 2}]$$

is a number whose classical continued fraction expansion is periodic and uses only partial quotients equal to 1 and 2; but this is not its β-expansion, which is

$$\frac{164 + 65\sqrt{17}}{251} = [1, 1, \beta, 2\beta^3 + \beta^2 + 1, \beta^3 + \beta + 1, 2, \beta + 1].$$

Corollary 6.7. — *The Perron numbers*

$$\varphi = \frac{1+\sqrt{5}}{2}, \qquad 1+\sqrt{2}, \qquad \frac{1+\sqrt{13}}{2}, \qquad \frac{1+\sqrt{17}}{2}$$

are the only quadratic Perron numbers smaller than 3 having property (CFF)*.*

By assuming Mercat's conjecture 6.2, they are the only quadratic Perron numbers with property (CFF)*.*

Proof. — Table 6.1 gives a full list of quadratic Perron numbers smaller than 3.

Table 6.1. Quadratic Perron numbers smaller than 3.

β	Approximate value	Minimal polynomial	Pisot unit	(CFF)
$\frac{1}{2}(1+\sqrt{5})$	1.618033988...	x^2-x-1	yes	yes
$\frac{1}{2}(1+\sqrt{13})$	2.302775637...	x^2-x-3	no	yes
$1+\sqrt{2}$	2.414213562...	x^2-2x-1	yes	yes
$\frac{1}{2}(1+\sqrt{17})$	2.561552812...	x^2-x-4	no	yes
$\frac{1}{2}(3+\sqrt{5})$	2.618033988...	x^2-3x+1	yes	no
$1+\sqrt{3}$	2.732050807...	x^2-2x-2	no	no
$\frac{1}{2}(1+\sqrt{21})$	2.791287847...	x^2-x-5	no	no

Theorem 6.5 shows that the first four of them have property (CFF). The following counterexamples, which can respectively be shown to be β-continued fraction expansions for the last three values of β in the list, show that these values do not have property (CFF):

$$\begin{aligned} [\overline{1}] &= \frac{1+\sqrt{5}}{2}, \\ [\overline{1,1,1,1,1,1,1,2,1,1,2,1,2,1,1,1,2,2}] &= \frac{11055+10864\sqrt{3}}{18471}, \\ [\overline{1,1,1,2,1,2,1,2,2,2,1,1,2,2}] &= \frac{117+44\sqrt{21}}{202}. \end{aligned}$$

According to Remark 6.3, under Mercat's conjecture 6.2, no other quadratic Perron number can have property (CFF). □

Remark 6.8. — Refuting (CFF) for quadratic integers bigger than 3 depends on the validity of Mercat's conjecture. However, for the subclass of quadratic Pisot units, we can provide explicit examples of bases β for which we can disprove property (CFF). In this way, we are able to determine unconditionally that the only quadratic Pisot units with property (CFF) are $\frac{1}{2}(1+\sqrt{5})$ and

$1+\sqrt{2}$. The examples of infinite β-continued fractions in the field $\mathbb{Q}(\beta)$ for other quadratic Pisot units β are the following:

- $\beta > 1$, root of $X^2 - mX - 1$, $m \geq 3$:
  m even:
  $$[\overline{(m-2)/2, 1, 1}] = \frac{m-2+\sqrt{m^2+4}}{4} = \frac{\beta-1}{2} \in \mathbb{Q}(\beta)\,.$$
  m odd, $m \geq 5$:
  $$[\overline{(m-3)/2, (m+1)/2, 3, 1}] = \frac{m^2-3m-3+m\sqrt{m^2+4}}{4m+6} \in \mathbb{Q}(\beta)\,.$$
  $m = 3$:
  $$[\overline{1, 1, 2, 2, 2}] = \frac{11+5\sqrt{13}}{17} \in \mathbb{Q}(\sqrt{13})\,.$$
- $\beta > 1$, root of $X^2 - mX + 1$, $m \geq 3$:
  $$[\overline{1, m-2}] = \frac{m-2+\sqrt{m^2-4}}{2(m-2)} \in \mathbb{Q}(\beta)\,.$$

This last example was already given in [12].

6.2. Square roots. — The simplest example of a non-Perron quadratic integer is an irrational square root of a rational integer D, i.e. $\beta = \sqrt{D} \notin \mathbb{Q}$. Such β satisfies $\beta' = -\beta$. For the study of finiteness and periodicity of β-continued fractions, in this case, we apply results of Section 3.

Theorem 6.9. — *Let D be a positive integer with an irrational square root. Denote $\beta = \sqrt{D}$. Then β satisfies* (CFP) *and, under Mercat's conjecture, does not satisfy* (CFF).

Proof. — The fact that $\beta = \sqrt{D}$ satisfies (CFP) follows from Theorem 1.1 with the use of Lemma 5.1 item (iv). Let us now focus on property (CFF). Assuming Mercat's conjecture, we only need to check $\beta \in \{\sqrt{2}, \sqrt{3}, \sqrt{5}, \sqrt{6}, \sqrt{7}, \sqrt{8}\}$. To disprove (CFF), in these cases, it suffices to consider periodic β-continued fractions

$$\begin{aligned} [\overline{4\sqrt{2}}] &= 3+2\sqrt{2}, & [\overline{8\sqrt{6}, 2\sqrt{6}}] &= 2(5+2\sqrt{6})\,,\\ [\overline{3,4}] &= \frac{3+2\sqrt{3}}{2}, & [\overline{\sqrt{7}, 2\sqrt{7}}] &= \frac{3+\sqrt{7}}{2}\,,\\ [\overline{32\sqrt{5}, 2\sqrt{5}}] &= 16(9+\sqrt{5}), & [\overline{2\sqrt{8}}] &= 3+\sqrt{8}\,. \end{aligned}$$ □

6.3. Non-Perron quadratic integers with positive conjugate. — Consider now as base β a quadratic integer that is not Perron, so that Theorem 1.1 may no longer be applied. We are, nevertheless, able to obtain information on the β-continued fractions by other means. We will show that neither (CFF) nor (CFP) hold by showing the existence of elements in the quadratic field $\mathbb{Q}(\beta)$ with aperiodic infinite β-continued fraction expansion.

THEOREM 6.10. — *Let $\beta > 1$ be a quadratic integer with conjugate β' satisfying $\beta' > \beta$. Then each $\xi \in \mathbb{Q}(\beta)$, such that $\xi > \beta$ and $\xi' \in (-1, 0)$ has an aperiodic β-continued fraction expansion.*

In particular, β satisfies neither (CFF) *nor* (CFP).

The first step towards the proof is an application of the algebraic argument in Proposition 3.2.

PROPOSITION 6.11. — *Let $\beta > 1$ be a quadratic integer, such that $\beta' > \beta$. Let ξ be an element in $\mathbb{Q}(\beta)$ whose β-continued fraction expansion is eventually periodic. Then the period consists only of partial quotients in $\{1, \ldots, \lfloor\beta\rfloor\}$.*

Proof. — Assume (replacing it by a complete quotient if needed) that $\xi = [\overline{a_0, \ldots, a_n}]$ is the purely periodic β-continued fraction expansion of an element in $\mathbb{Q}(\beta)$. Then by Proposition 3.2, for $x = p_n + q_{n-1}$, we have that $x' = \pm x$.

Recall now that p_n, q_{n-1} arise from the continuants, and thus also x is a polynomial with positive coefficients in the partial quotients $a_0, \ldots, a_n$, all of them appearing with positive degree. By item (ii) of Lemma 5.1, for each partial quotient a_i in $\mathbb{Z}_\beta^+ \setminus \{1, \ldots, \lfloor\beta\rfloor\}$, its conjugate a_i' satisfies $a_i' > a_i$. Thus, we obtain that $x' \neq \pm x$, unless all partial quotients appearing in the period belong to $\{1, \ldots, \lfloor\beta\rfloor\}$, thus proving the statement. $\square$

REMARK 6.12. — Let β be a quadratic integer. Notice that $|\beta - \beta'|$ is at least 1, because it is equal to the absolute value of the square root of the discriminant of the minimal polynomial of β.

Proof of Theorem 6.10. — First notice that if $\xi_i' \in (-1, 0)$, then $\xi_{i+1}' = \frac{1}{\xi_i' - a_i'} \in (-1, 0)$ because $a_i' \geq 1$. Thus, the β-continued fraction expansion of any $\xi \in \mathbb{Q}(\beta)$ with $\xi' \in (-1, 0)$ is infinite.

Next we prove that if ξ has eventually periodic β-continued fraction, then it is, in fact, purely periodic. Then, by Proposition 6.11, all its partial quotients belong to $\{1, \ldots, \lfloor\beta\rfloor\}$. However, $a_0 = \lfloor\xi\rfloor_\beta \geq \beta$, which gives a contradiction.

To show pure periodicity of the β-continued fraction of ξ we argue as in the characterization of purely periodic classical continued fractions. Assume that ξ has an eventually periodic β-continued fraction expansion. Take $k < l$ to be the minimal indices, such that $\xi_k = \xi_l$. Then we either have $k = 0$, and the proof is finished, or $k > 0$, and we have $\xi_{k-1}' = a_{k-1}' + \frac{1}{\xi_k'} \in (-1, 0)$, which implies

$$-1 - \frac{1}{\xi'_k} < a'_{k-1} < -\frac{1}{\xi'_k}\,. \tag{13}$$

By Remark 6.12, we have $\lfloor \beta' \rfloor > \lfloor \beta \rfloor$, and, therefore, by item (ii) of Lemma 5.1, the distances between the conjugates x', y' of β-integers x, y are at least 1. Consequently, (13) defines $a'_{k-1} = a'_{l-1}$ uniquely. Thus,

$$\xi_{k-1} = a_{k-1} + \frac{1}{\xi_k} = a_{l-1} + \frac{1}{\xi_l} = \xi_{l-1}\,,$$

which contradicts the minimality of indices k, l. $\square$

REMARK 6.13. — Let us mention that in refuting (CFF) for β with $\beta' > \beta$ we did not use Mercat's conjecture. Another interesting fact to mention is that the infinitely many elements of $\mathbb{Q}(\beta)$ with aperiodic β-continued fraction expansion found in Theorem 6.10 belong to the family of numbers with purely periodic regular continued fraction.

6.4. Non-Perron quadratic integers with negative conjugate. — It remains to treat the case when $\beta > 1$ is a quadratic integer such that $\beta' < -\beta$. These β's are the positive irrational roots of the polynomials of the shape

$$X^2 + bX - c, \quad \text{where } b \geq 1 \text{ and } c \geq b + 2 \text{ are integers.} \tag{14}$$

Notice that in this case, $\beta' = -\beta - b$. Here the situation appears to be more complicated than in the other cases already treated. We shall prove the following statement.

THEOREM 6.14. — *Let $\beta > 1$ be a root of* (14) *with $b \geq 4$, i.e. β is a quadratic integer with conjugate β' satisfying $\beta' \leq -\beta - 4$. Let $\xi \in \mathbb{Q}(\beta)$ be such that $\xi' \in (-1, 0)$. Then ξ does not have a finite β-expansion.*

In particular, β does not satisfy property (CFF)*.*

LEMMA 6.15. — *Let $b \geq 1$ be integers and assume that $\beta' = -\beta - b$. Then the function*

$$F(k) = \beta'^{2k+1} + \lfloor \beta \rfloor \sum_{i=0}^{k} \beta'^{2i}\,, \quad k \geq 0\,,$$

is decreasing with k, and the function

$$G(k) = \beta'^{2k} + \lfloor \beta \rfloor \sum_{i=0}^{k-1} \beta'^{2i+1}\,, \quad k \geq 1\,,$$

is increasing with k.

Proof. — We see that

$$F(k+1)-F(k)=\beta'^{2k+1}\left(\beta'^2-1+\lfloor\beta\rfloor\beta'\right)$$
$$G(k+1)-G(k)=\beta'^{2k}\left(\beta'^2-1+\lfloor\beta\rfloor\beta'\right).$$

However, $\lfloor\beta\rfloor\leq\beta$, which implies that

$$\beta'^2+\lfloor\beta\rfloor\beta'-1\geq\beta'^2+\beta\beta'-1=(\beta+b)b-1>1\,,$$

which completes the proof. □

Proposition 6.16. — *Assume that $\beta'\leq-\beta-4$. Then for every non-zero $x\in\mathbb{Z}_\beta^+$, we have that either $x=1$, or $x'\geq 2$, or $x'<-4$.*

Proof. — Write $\beta'=-\beta-b$, for $b\geq 4$ an integer. Let $x=\sum_{i=0}^n x_i\beta^i\in\mathbb{Z}_\beta^+$. Assume first that the highest power of β appearing in x is even. If this power is 0, then $x\in\{1,\ldots,\lfloor\beta\rfloor\}$, so $x'=1$ or $x'\geq 2$. Otherwise, write $n=2k$ with $k\geq 1$. Then

$$\begin{aligned}x'=\sum_{i=0}^n x_i\beta'^i&\geq\beta'^{2k}+\lfloor\beta\rfloor\sum_{i=0}^{k-1}\beta'^{2i+1}\\&\geq\beta'^2+\lfloor\beta\rfloor\beta'\geq\beta'(\beta'+\beta)=(\beta+b)b\geq 2\,,\end{aligned}$$

where the second inequality follows from Lemma 6.15. Similarly, if the highest power appearing in x is $n=2k+1$ odd, we have

$$\begin{aligned}x'=\sum_{i=0}^n x_i\beta'^i&\leq\beta'^{2k+1}+\lfloor\beta\rfloor\sum_{i=0}^{k}\beta'^{2i}\\&\leq\beta'+\lfloor\beta\rfloor=\lfloor\beta\rfloor-\beta-b<-b\leq-4\,.\end{aligned}$$

□

Proof of Theorem 6.14. — By the previous proposition, we have that the conjugate of any partial quotient is either equal to 1, or is at least 2, or at most −4. This allows us to track the position of the conjugates of the complete quotients and to check that they never become zero. From the relation $\xi_{k+1}=1/(\xi_k-a_k)$, we can derive the following list of implications, which is graphically represented in Figure 6.1.

(1) if $\xi'_k\in(-1,0)$, then
  (a) $a'_k\geq 1\implies\xi'_{k+1}\in(-1,0)$,
  (b) $a'_k\leq -3\implies\xi'_{k+1}\in(0,\frac{1}{2})$;
(2) if $\xi'_k\in(0,\frac{1}{2})$, then
  (a) $a'_k\geq\frac{3}{2}\implies\xi'_{k+1}\in(-1,0)$,
  (b) $a'_k\leq -2\implies\xi'_{k+1}\in(0,\frac{1}{2})$,
  (c) $a_k=1\implies\xi'_{k+1}\in(-2,-1)$;
(3) if $\xi'_k\in(-2,-1)$, then
  (a) $a'_k>0\implies\xi'_{k+1}\in(-1,0)$,
  (b) $a'_k\leq -4\implies\xi'_{k+1}\in(0,\frac{1}{2})$. □

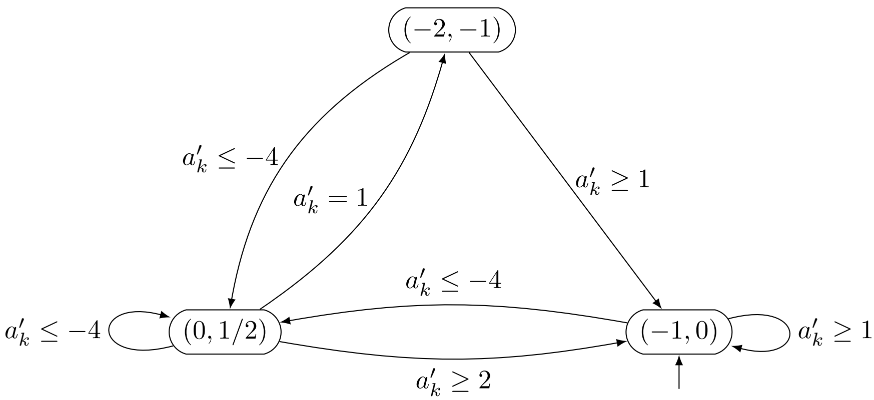

FIGURE 6.1. Labeled graph representing the dynamics of the conjugates of the complete quotients in the β-continued fraction expansion of ξ.

Theorem 6.14 covers the cases when β is a root of (14) with $b \geq 4$, but for $b = 1, 2, 3$, we have to disprove property (CFF) by other means. Mercat's conjecture implies that if $\beta > 3$, (CFF) cannot hold. This leaves only 30 values of β: the positive roots of polynomials

$$X^2 + bX - c, \quad b = 1, 2, 3, \quad b + 2 \leq c < 3(b + 3),\ c \neq 2(b + 2)\,.$$

TABLE 6.2. Expected behaviour of the β-continued fractions for non-Perron quadratic numbers with negative conjugate in terms of the coefficients of its minimal polynomial $x^2 + bx - c$.

		(CFF)	Reason
$b = 1$	$c \geq 12$	No?	Implied by Mercat's conjecture
$b = 1$	$5 \leq c \leq 11$, $c \neq 6$	No	We found a periodic expansion
$b = 1$	$c = 3, 4$	Open	
$b = 2$	$c \geq 16$	No?	Implied by Mercat's conjecture
$b = 2$	$12 \leq c \leq 14$	No	We found a periodic expansion
$b = 2$	$4 \leq c \leq 11$, $c \neq 8$	Open	
$b = 3$	$c \geq 18$	No?	Implied by Mercat's conjecture
$b = 3$	$c = 17$	No	We found a periodic expansion
$b = 3$	$5 \leq c \leq 16$, $c \neq 10$	Open	
$b \geq 4$	$c \geq b + 2$	No	Theorem 6.14

For 10 of these values, we have found elements of the corresponding quadratic field $\mathbb{Q}(\beta)$ with periodic β-continued fraction expansion, thus refuting (CFF) directly. However, there remain 20 cases for which the question of property (CFF) remains open. The results supported by of our computational experiments are summarized in Table 6.2.

Let us comment on what our computational experiments suggest about eventually periodic β-continued fraction expansions of elements of $\mathbb{Q}(\beta)$. Given $b \geq 1$, for sufficiently large c, we expect that property (CFP) holds. For intermediate values of c, we can find in $\mathbb{Q}(\beta)$ both elements with eventually periodic and (probably) aperiodic β-continued fraction expansion. For the smallest admissible values of $c \geq b+2$, it is likely that all elements of $\mathbb{Q}(\beta)$ have either finite or aperiodic β-continued fraction expansion. The exact behaviour is yet to be investigated.

Appendix A.

Let us present here an alternative proof of the finiteness result in Theorem 1.1, which is more elementary in nature and provides a better bound for the number of irrational partial quotients.

In the entire Appendix, let $\xi = [a_0, a_1, \ldots]$ be an infinite continued fraction with the value in a quadratic field K, $a_n \in \mathcal{O}_K$ and $a_n \geq 1$, for all n. Let d be the leading coefficient of the minimal polynomial of ξ over $\mathbb{Z}$ (which we take to be positive). We will always denote by x' the Galois conjugate of $x \in K$.

Theorem A.1. — *Assume that $|a_n'| < a_n$, for all n with $a_n \notin \mathbb{Z}$. Then there are at most*

$$36H(\xi)^2 \log H(\xi) + 9\log(3)H(\xi)^2$$

irrational partial quotients, and, therefore, the given continued fraction is eventually periodic.

Before plunging into the proof, let us briefly recall why all rational numbers have a finite regular continued fraction expansion.

On the one hand, continued fractions give very good rational approximations, and if p/q is a convergent of a rational number a/b, then the majoration $\left|\frac{a}{b} - \frac{p}{q}\right| \leq \frac{1}{q^2}$ holds. On the other hand, rational numbers are badly approximated by other rational numbers, and if $\frac{a}{b} \neq \frac{p}{q}$, then a minoration $\left|\frac{a}{b} - \frac{p}{q}\right| \geq \frac{|aq-bp|}{bq} \geq \frac{1}{bq}$ holds. The two inequalities taken together imply that for a fixed a/b, only finitely many distinct convergents p/q may exist.

The same majoration also holds for any convergent continued fraction, as seen with (7), so in order to repeat the classical proof, we need an argument of diophantine approximation to supply a minoration; this is given by the following proposition.

Proposition A.2. — *For all $n \geq 0$ the following inequality holds:*

$$\left|\frac{q'_n}{q_n}\right| \left|\xi' - \frac{p'_n}{q'_n}\right| > \frac{a_{n+1}}{d^2}\,.$$

Proof. — The quantity $q_n\xi - p_n$ is not equal to zero because otherwise the expansion of ξ would be finite. As d is the leading coefficient of the minimal polynomial of ξ over $\mathbb{Z}$, we see that $d\xi \in \mathcal{O}_K$. Then the norm of $d(q_n\xi - p_n)$ must be at least 1 in absolute value, because p_n and q_n are algebraic integers. So we have that

$$1 \leq |N(d(q_n\xi - p_n))| = d^2\,|q_n\xi - p_n| \cdot |q'_n\xi' - p'_n| < \frac{d^2}{a_{n+1}} \left|\frac{q'_n}{q_n}\right| \cdot \left|\xi' - \frac{p'_n}{q'_n}\right|,$$

where the last inequality follows from (7). □

As the Euclidean absolute value over $\mathbb{Q}$ splits into two Archimedean absolute values over the quadratic number field K, our minoration involves both the convergents p/q and their conjugates p'/q'. The conjugates p' and q' obey the same recurrence relations as p and q, with the partial quotients replaced by their conjugates. These conjugates, however, need not be bounded away from 0, or even be positive, and we cannot guarantee in general that the ratio p'/q' will converge to some value. In order to proceed with the proof we need a way of controlling their growth, which we will do through an accurate analysis of the recurrences (1).

We further state here two easily checked remarks, which we will use in the proof below:

Remark A.3. — Let $0 < A < B$ be real numbers. The function $f(x) = \frac{A+x}{B+x}$ is strictly increasing on $(0, +\infty)$.

Remark A.4. — Let $(\alpha_n)_{n\in\mathbb{N}}$ satisfy a linear recurrence of order 2. Then for every $n \geq 0$, $\alpha_n = f_n\alpha_1 + g_n\alpha_0$, where $(f_n)_{n\in\mathbb{N}}$, $(g_n)_{n\in\mathbb{N}}$ satisfy the same linear recurrence with initial conditions $f_0 = g_1 = 0$, $f_1 = g_0 = 1$.

Lemma A.5. — *Let $(\alpha_n)_{n\in\mathbb{N}}$, $(\beta_n)_{n\in\mathbb{N}}$ be two real sequences satisfying the same linear recurrence relation of order 2*

$$x_n = r_n x_{n-1} + x_{n-2}, \qquad n \geq 2\,, \tag{15}$$

with $r_i \geq 1$, for $i \geq 1$, with initial conditions $\alpha_1 > \beta_1 > 0$, $\alpha_0 = \beta_0 > 0$. Then for all $n \geq 3$, we have

$$\frac{\beta_n}{\alpha_n} < \frac{\beta_1 + \beta_0}{\alpha_1 + \alpha_0}\,.$$

Proof. — Let $(f_n)_{n\in\mathbb{N}}$, $(g_n)_{n\in\mathbb{N}}$ be the sequences from Remark A.4, such that $\alpha_n = f_n\alpha_1 + g_n\alpha_0$, and $\beta_n = f_n\beta_1 + g_n\beta_0 = f_n\beta_1 + g_n\alpha_0$. Since $(f_n)_{n\in\mathbb{N}}$, $(g_n)_{n\in\mathbb{N}}$ satisfy (15), they are both positive sequences for $n \geq 2$. By induction, with the use of $r_i \geq 1$, one can show that $g_n < f_n$ for $n \geq 3$. Using Remark A.3, we then have

$$\frac{\beta_n}{\alpha_n} = \frac{f_n\beta_1 + g_n\beta_0}{f_n\alpha_1 + g_n\alpha_0} < \frac{f_n\beta_1 + f_n\beta_0}{f_n\alpha_1 + f_n\alpha_0} = \frac{\beta_1 + \beta_0}{\alpha_1 + \alpha_0}\,,$$

where we have used that $\alpha_0 = \beta_0$. □

We now prove a proposition that allows us to compare recurrent sequences of the shape (1). We will apply it to the sequences $(p_n)_{n\in\mathbb{N}}, (q_n)_{n\in\mathbb{N}}$ and their conjugates.

Proposition A.6. — *Let $(a_n)_{n\in\mathbb{N}}$ be a sequence of positive reals and $(b_n)_{n\in\mathbb{N}}$ be a sequence of complex numbers, such that $a_n \geq 1$ and $|b_n| \leq a_n$, for every $n \geq 0$.*

Let $(s_n)_n$, $(t_n)_n$ satisfy for every $n \geq 0$

$$\begin{aligned} s_n &= a_n s_{n-1} + s_{n-2}, & s_{-1} &= t_{-1} > 0\,,\\ t_n &= b_n t_{n-1} + t_{n-2}, & s_{-2} &= t_{-2} = 0\,. \end{aligned}$$

Let $0 < C < 1$ and $r_n = \#\{0 \leq k \leq n \mid |b_k| < Ca_k\}$. Then for all $n \geq 4$, we have

$$\frac{|t_n|}{s_n} \leq \left(\frac{C+2}{3}\right)^{r_{n-3}}.$$

In particular, if $\frac{|b_k|}{a_k} < C$ for infinitely many $k \in \mathbb{N}$, then

$$\lim_{n\to+\infty} \frac{t_n}{s_n} = 0\,.$$

Proof. — Let us define for each $k \in \mathbb{N}$ an auxiliary sequence $(\gamma_n^{(k)})_{n\in\mathbb{N}}$ as follows,

$$\begin{aligned} \gamma_{-2}^{(k)} &= 0, & \gamma_{-1}^{(k)} &= s_{-1}\,,\\ \gamma_n^{(k)} &= |b_n|\gamma_{n-1}^{(k)} + \gamma_{n-2}^{(k)}\,, & &\text{for } 0 \leq n < k\,,\\ \gamma_n^{(k)} &= a_n\gamma_{n-1}^{(k)} + \gamma_{n-2}^{(k)}\,, & &\text{for } n \geq k\,. \end{aligned}$$

Clearly, each sequence $(\gamma^{(k)})_{n\in\mathbb{N}}$ is strictly increasing for $n \geq 0$, and we also have

$$|t_n| = \gamma_n^{(n+1)} \leq \cdots \leq \gamma_n^{(k+1)} \leq \gamma_n^{(k)} \leq \cdots \leq \gamma_n^{(0)} = s_n \quad \text{for every } n \in \mathbb{N}\,. \tag{16}$$

Moreover, we have $\gamma_n^{(k+1)} = \gamma_n^{(k)}$ for $n = 0, \ldots, k-1$, and if $|b_k| < a_k$, then $\gamma_k^{(k+1)} < \gamma_k^{(k)}$.

Now consider a fixed index $k \geq 2$, such that $\frac{|b_k|}{a_k} < C$. We will show that there exists a constant $T < 1$, independent of k, such that

$$\frac{\gamma_n^{(k+1)}}{\gamma_n^{(k)}} < T \qquad \text{for every } n \geq k+2\,. \tag{17}$$

We will apply Lemma A.5 with $\alpha_n = \gamma_{k+n-1}^{(k)}$, and $\beta_n = \gamma_{k+n-1}^{(k+1)}$, for $n \geq 0$. These sequences satisfy the same recurrence with $r_n = a_{k+n-1}$, and the assumptions $\alpha_1 > \beta_1 > 0$, $\alpha_0 = \beta_0 > 0$ on initial conditions are satisfied.

For all $n \geq k+2$, we obtain

$$\begin{aligned}\frac{\gamma_n^{(k+1)}}{\gamma_n^{(k)}} &= \frac{\beta_{n-k+1}}{\alpha_{n-k+1}} < \frac{\beta_1+\beta_0}{\alpha_1+\alpha_0}\frac{\gamma_k^{(k+1)}+\gamma_{k-1}^{(k+1)}}{\gamma_k^{(k)}+\gamma_{k-1}^{(k)}} = \frac{|b_k|\gamma_{k-1}^{(k+1)}+\gamma_{k-2}^{(k+1)}+\gamma_{k-1}^{(k+1)}}{a_k\gamma_{k-1}^{(k)}+\gamma_{k-2}^{(k)}+\gamma_{k-1}^{(k)}}\\ &= \frac{(|b_k|+1)\gamma_{k-1}^{(k+1)}+\gamma_{k-2}^{(k+1)}}{(a_k+1)\gamma_{k-1}^{(k)}+\gamma_{k-2}^{(k)}} = \frac{|b_k|+1+\frac{\gamma_{k-2}^{(k+1)}}{\gamma_{k-1}^{(k+1)}}}{a_k+1+\frac{\gamma_{k-2}^{(k)}}{\gamma_{k-1}^{(k)}}}\,,\end{aligned}$$

where we have used that $\gamma_{k-1}^{(k+1)} = \gamma_{k-1}^{(k)}$. Since the sequences $(\gamma_n^{(j)})_{n\in\mathbb{N}}$ are strictly increasing, we derive by Remark A.3 that

$$\frac{\gamma_n^{(k+1)}}{\gamma_n^{(k)}} < \frac{|b_k|+2}{a_k+2} = \frac{\frac{|b_k|}{a_k}+\frac{2}{a_k}}{1+\frac{2}{a_k}} \leq \frac{\frac{|b_k|}{a_k}+2}{3} < \frac{C+2}{3} =: T\,,$$

where we use again Remark A.3 with the fact that $a_k \geq 1$ and k is such that $\frac{|b_k|}{a_k} < C$. Since $C < 1$, also $T < 1$, which shows (17) is true.

By (16) and (17), for any $k \geq 2$ and $n \geq k+2$, we have

$$\frac{|t_n|}{s_n} \leq \frac{\gamma_n^{(k)}}{\gamma_n^{(0)}} = \prod_{j=0}^{k-1}\frac{\gamma_n^{(j+1)}}{\gamma_n^{(j)}} \leq \prod_{j=0}^{k-1} T_j\,,$$

where $T_j = T$ for j, such that $\frac{|b_j|}{a_j} < C$, and $T_j = 1$ otherwise. Thus, for every $n \geq 4$, by substituting $k = n-2$ above, we obtain

$$0 \leq \frac{|t_n|}{s_n} \leq \prod_{j=0}^{n-3} T_j = T^{r_{n-3}}\,.$$

If r_n tends to infinity with n, we obtain $\lim_{n\to\infty}\frac{|t_n|}{s_n} = 0$. □

We are now ready for the final part of the argument.

Proof of Theorem A.1. — We first show that if $|a_n'| < a_n$ for every irrational partial quotient a_n, then there exists a constant $0 < C < 1$, such that for all $n \geq 0$ with $a_n \notin \mathbb{Z}$, we have $\frac{|a_n'|}{a_n} \leq C$.

Since $|a'_n| \leq a_n$, for all $n \geq 0$, by Lemma 3.1 the partial quotients satisfy $a_n \leq 3H(\xi)^2$. Since a_n, a'_n are algebraic integers, the quantities $a_n + a'_n$ and $a_n - a'_n$ are in $\mathbb{Z} \cup \sqrt{D}\mathbb{Z}$, so that $a_n - |a'_n| \geq 1$, whenever it is not equal to 0. Dividing by a_n we see that

$$\frac{|a'_n|}{a_n} \leq 1 - \frac{1}{a_n} \leq 1 - \frac{1}{3H(\xi)^2} =: C\,.$$

We can now apply Proposition A.6 with $(a'_n)_n$ as the sequence $(b_n)_n$, and $(p_n)_n, (p'_n)_n$ as the sequences $(s_n)_n, (t_n)_n$. Analogously but shifting all indices by 1, so that the initial conditions are verified, we can do it for $(q_{n+1})_n, (q'_{n+1})_n$. Therefore, for all $n \geq 5$, we have that

$$\frac{|p'_n|}{p_n}, \frac{|q'_n|}{q_n} \leq \left(\frac{C+2}{3}\right)^{r_{n-4}}\,,$$

where $r_n = \#\{0 \leq k \leq n \mid a_k \notin \mathbb{Z}\}$.

This implies that

$$\left|\frac{q'_n}{q_n}\right| \cdot \left|\xi' - \frac{p'_n}{q'_n}\right| \leq \left|\frac{q'_n}{q_n}\right| |\xi'| + \left|\frac{p'_n}{p_n}\right| \left|\frac{p_n}{q_n}\right| \leq \left(\frac{C+2}{3}\right)^{r_{n-4}} (|\xi'| + |\xi| + 1)\,.$$

On the other hand, by Proposition A.2

$$\left|\frac{q'_n}{q_n}\right| \cdot \left|\xi' - \frac{p'_n}{q'_n}\right| > \frac{1}{d^2}\,,$$

so combining them we obtain

$$\left(\frac{C+2}{3}\right)^{-r_{n-4}} < d^2(|\xi'| + |\xi| + 1) \leq 3H(\xi)^4\,,$$

where the second inequality comes from Remark 2.3, and the elementary inequality $1 + a + b \leq 3\max\{1, a\}\max\{1, b\}$, for all $a, b \geq 0$.

Taking logarithms we obtain

$$-r_{n-4} \log\left(1 - \frac{1}{9H(\xi)^2}\right) < 4\log H(\xi) + \log 3\,,$$

and remembering that $\log(1 + x) \leq x$, for all $x > -1$, we obtain

$$r_{n-4} < 36H(\xi)^2 \log H(\xi) + 9\log(3)H(\xi)^2\,.$$

Letting n go to infinity this shows that the number of irrational partial quotients is finite and does not exceed the stated bound.

As soon as the tail of the expansion contains only positive integers, the classical theory of simple continued fractions and the fact that ξ lies in a quadratic field imply that the expansion is eventually periodic. □

Notice that, while this theorem gives a bound for the *number* of irrational partial quotients, it does not tell us *where* they are. In comparison, Theorem 1.1

shows that the irrational partial quotients can be effectively found. Nevertheless, we wanted to include this Appendix because the proof given here is self contained, rather elementary, and gives a bound that is substantially lower than those alluded to in Remark 3.4, which rely on much more sophisticated techniques.

BIBLIOGRAPHY

[1] S. Akiyama – "Pisot numbers and greedy algorithm", in *Number theory (Eger, 1996)*, de Gruyter, Berlin, 1998, p. 9–21.

[2] J. Bernat – "Continued fractions and numeration in the Fibonacci base", *Discrete Math.* **306** (2006), no. 22, p. 2828–2850.

[3] E. Bombieri & W. Gubler – *Heights in Diophantine geometry*, New Mathematical Monographs, vol. 4, Cambridge University Press, Cambridge, 2006.

[4] B. W. Brock, N. D. Elkies & B. W. Jordan – "Periodic continued fractions over S-integers in number fields and Skolem's p-adic method", *Acta Arith.* **197** (2021), no. 4, p. 379–420.

[5] v. Burdík, C. Frougny, J.-P. Gazeau & R. Krejcar – "Beta-integers as natural counting systems for quasicrystals", *J. Phys. A* **31** (1998), no. 30, p. 6449–6472.

[6] J. Feng, C. Ma & S. Wang – "Metric theorems for continued β-fractions", *Monatsh. Math.* **190** (2019), p. 281–299.

[7] C. Frougny & B. Solomyak – "Finite beta-expansions", *Ergodic Theory Dynam. Systems* **12** (1992), no. 4, p. 713–723.

[8] R. L. Graham, D. E. Knuth & O. Patashnik – *Concrete mathematics: A foundation for computer science*, 2nd ed., Addison-Wesley Longman Publishing Co., Inc., USA, 1994.

[9] R. Grizzard & J. Gunther – "Slicing the stars: counting algebraic numbers, integers, and units by degree and height", *Algebra Number Theory* **11** (2017), no. 6, p. 1385–1436.

[10] G. H. Hardy & E. M. Wright – *An introduction to the theory of numbers*, sixth ed., Oxford University Press, Oxford, 2008, Revised by D. R. Heath-Brown and J. H. Silverman, With a foreword by Andrew Wiles.

[11] M. Hbaib & R. Kammoun – "Continued β-fractions with formal power series over finite fields", *Ramanujan J.* **39** (2016), no. 1, p. 95–105.

[12] J. Kolář – *Continued fractions based on beta-integers (in czech)*, Bachelor's project – Czech Technical University in Prague, 2011.

[13] C. T. McMullen – "Uniformly Diophantine numbers in a fixed real quadratic field", *Compos. Math.* **145** (2009), no. 4, p. 827–844.

[14] P. Mercat – "Construction de fractions continues périodiques uniformément bornées", *J. Théor. Nombres Bordeaux* **25** (2013), no. 1, p. 111–146.

[15] W. Parry – "On the β-expansions of real numbers", *Acta Math. Acad. Sci. Hungar.* **11** (1960), p. 401–416.

[16] A. Rényi – "Representations for real numbers and their ergodic properties", *Acta Math. Acad. Sci. Hungar* **8** (1957), p. 477–493.

[17] D. Rosen – "A class of continued fractions associated with certain properly discontinuous groups", *Duke Math. J.* **21** (1954), p. 549–563.

[18] ______, "Research Problems: Continued Fractions in Algebraic Number Fields", *Amer. Math. Monthly* **84** (1977), no. 1, p. 37–39.

[19] S. H. Schanuel – "Heights in number fields", *Bull. Soc. Math. France* **107** (1979), no. 4, p. 433–449.

Bull. Soc. Math. France
150 (1), 2022, p. 111-207

ABOUT PLANE PERIODIC WAVES OF THE NONLINEAR SCHRÖDINGER EQUATIONS

by Corentin Audiard & L. Miguel Rodrigues

Abstract. — The present contribution contains a quite extensive theory for the stability analysis of plane periodic waves of general Schrödinger equations. On the one hand, we put the one-dimensional theory, or in other words the stability theory for longitudinal perturbations, on par with the one available for systems of Korteweg type, including results on coperiodic spectral instability, nonlinear coperiodic orbital stability, sideband spectral instability and linearized large-time dynamics in relation with modulation theory, and resolutions of all the involved assumptions in both the small-amplitude and large-period regimes. On the other hand, we provide extensions of the spectral part of the latter to the multidimensional context. Notably, we provide suitable multidimensional modulation formal asymptotics, validate those at the spectral level, and use them to prove that waves are always spectrally unstable in both the small-amplitude and the large-period regimes.

Texte reçu le 17 août 2021, accepté le 16 décembre 2021.

Corentin Audiard, Sorbonne Université, CNRS, Université de Paris, Laboratoire Jacques-Louis Lions (LJLL), F-75005 Paris, France • *E-mail :* corentin.audiard@upmc.fr
L. Miguel Rodrigues, Univ Rennes & IUF, CNRS, IRMAR – UMR 6625, F-35000 Rennes, France • *E-mail :* luis-miguel.rodrigues@univ-rennes1.fr

Mathematical subject classification (2010). — 35B10, 35B35, 35P05, 35Q55, 37K45.

Key words and phrases. — Schrödinger equations, periodic traveling waves, spectral stability, orbital stability, abbreviated action integral, harmonic limit, soliton asymptotics, modulation systems, Hamiltonian dynamics.

Research of C.A. was partially supported by the French ANR Project NABUCO ANR-17-CE40-0025.

 0037-9484/2022/111/$ 5.00
 doi:10.24033/bsmf.2846

Résumé (*À propos des ondes planes périodiques des équations de Schrödinger non linéaires*). — Le travail présenté ici comprend une théorie relativement complète permettant l'analyse de la stabilité des ondes planes périodiques des équations de Schrödinger générales. D'une part, nous mettons la théorie unidimensionnelle, ou autrement dit la théorie de stabilité sous perturbations longitudinales, au niveau de celle disponible pour les systèmes de type Korteweg, en y incluant des résultats sur l'instabilité spectrale co-périodique, la stabilité orbitale non linéaire co-périodique, l'instabilité spectrale latérale et la dynamique linéarisée en temps long et ses relations avec la théorie de la modulation, et en résolvant toutes les hypothèses associées dans les régimes de petite amplitude et de grande période. D'autre part, nous étendons la partie spectrale de cette analyse au contexte multidimensionnel. En particulier, nous développons une asymptotique formelle de modulation multidimensionnelle, validons celle-ci au niveau spectral et l'utilisons pour démontrer que les ondes sont toujours spectralement instables à la fois dans les régimes de petite amplitude et de grande période.

1. Introduction

We consider Schrödinger equations in the form

$$\mathrm{i}\,\partial_t f = -\,\mathrm{div}_{\mathbf{x}}\left(\kappa(|f|^2)\,\nabla_{\mathbf{x}} f\right) + \kappa'(|f|^2)\,\|\nabla_{\mathbf{x}} f\|^2\, f + 2\,W'(|f|^2)\, f \tag{1}$$

(or some anisotropic generalizations) with W a smooth real-valued function and κ a smooth positive-valued function, bounded away from zero, where the unknown f is complex valued, $f(t,\mathbf{x}) \in \mathbf{C}$, and $(t,\mathbf{x}) \in \mathbf{R}\times\mathbf{R}^d$. Note that the sign assumption on κ may be replaced with the assumption that κ is real valued and far from zero since one may change the sign of κ by replacing (f,κ,W) with $(\overline{f},-\kappa,-W)$.

Since the nonlinearity is not holomorphic in f, it is convenient to adopt a real point of view and introduce real and imaginary parts $f = a + \mathrm{i}\,b$, $\mathbf{U} = \begin{pmatrix} a \\ b \end{pmatrix}$. Multiplication by $-\,\mathrm{i}$ is, thus, encoded in

$$\mathbf{J} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \tag{2}$$

and Equation (1) takes the form

$$\partial_t \mathbf{U} = \mathbf{J}\left(-\,\mathrm{div}_{\mathbf{x}}\left(\kappa(\|\mathbf{U}\|^2)\,\nabla_{\mathbf{x}}\mathbf{U}\right) + \kappa'(\|\mathbf{U}\|^2)\,\|\nabla_{\mathbf{x}}\mathbf{U}\|^2\,\mathbf{U} + 2\,W'(\|\mathbf{U}\|^2)\,\mathbf{U}\right). \tag{3}$$

The problem has a Hamiltonian structure

$$\partial_t \mathbf{U} = \mathbf{J}\,\delta\mathscr{H}_0[\mathbf{U}] \qquad \text{with} \qquad \mathscr{H}_0\,[\mathbf{U}] = \tfrac{1}{2}\kappa(\|\mathbf{U}\|^2)\|\nabla_{\mathbf{x}}\mathbf{U}\|^2\,, +W(\|\mathbf{U}\|^2),$$

with δ denoting the variational gradient[1]. Indeed, our interest in (1) originates in the fact that we regard the class of equations (1) as the most natural class of

1. See the notational section at the end of the present Introduction for a definition.

isotropic quasi/linear dispersive Hamiltonian equations, including most classical semilinear Schrödinger equations. See [54] for a comprehensive introduction to the latter. In Appendix C, we also show how to treat some anisotropic versions of the equations.

Note that in the above form, invariances are embedded with respect to rotations, time translations and space translations; if f is a solution, so is $\widetilde{f}$ when

$$\begin{aligned}
\widetilde{f}(t,\mathbf{x}) &= \mathrm{e}^{-\,\mathrm{i}\,\phi_0}\, f(t,\mathbf{x})\,, && \phi_0 \in \mathbf{R}\,, && \text{rotational invariance}\,,\\
\widetilde{f}(t,\mathbf{x}) &= f(t-t_0,\mathbf{x})\,, && t_0 \in \mathbf{R}\,, && \text{time translation invariance}\,,\\
\widetilde{f}(t,\mathbf{x}) &= f(t,\mathbf{x}-\mathbf{x}_0)\,, && \mathbf{x}_0 \in \mathbf{R}^d\,, && \text{space translation invariance}\,.
\end{aligned}$$

Actually, rotations and time and space translations leave the Hamiltonian $\mathscr{H}_0$ essentially unchanged, in a sense made explicit in Appendix A. Thus, through a suitable version of Noether's principle, they are associated with conservation laws, respectively on mass $\mathscr{M}[\mathbf{U}] = \frac{1}{2}\|\mathbf{U}\|^2$, Hamiltonian $\mathscr{H}_0[\mathbf{U}]$ and momentum $\mathbb{Q}[\mathbf{U}] = (\mathbb{Q}_j[\mathbf{U}])_j$, with $\mathbb{Q}_j[\mathbf{U}] = \frac{1}{2}\mathbf{J}\mathbf{U}\cdot\partial_j\mathbf{U}$, $j=1,\dots,d$. Namely, invariance by rotation implies that any solution $\mathbf{U}$ to (3) satisfies the mass conservation law

$$\partial_t \mathscr{M}(\mathbf{U}) = \sum_j \partial_j\left(\mathbf{J}\delta\mathscr{M}[\mathbf{U}]\cdot\nabla_{\mathbf{U}_{x_j}}\mathscr{H}_0[\mathbf{U}]\right). \tag{4}$$

Likewise, invariance by time translation implies that (3) contains the conservation law

$$\partial_t \mathscr{H}_0[\mathbf{U}] = \sum_j \partial_j\left(\nabla_{\mathbf{U}_{x_j}}\mathscr{H}_0[\mathbf{U}]\cdot\mathbf{J}\delta\mathscr{H}_0[\mathbf{U}]\right). \tag{5}$$

Finally, invariance by spatial translation implies that from (3) stems

(6)
$$\partial_t\left(\mathbb{Q}[\mathbf{U}]\right) = \nabla_{\mathbf{x}}\left(\frac{1}{2}\mathbf{J}\mathbf{U}\cdot\mathbf{J}\delta\mathscr{H}_0[\mathbf{U}] - \mathscr{H}_0[\mathbf{U}]\right) + \sum_\ell \partial_\ell(\mathbf{J}\delta\,\mathbb{Q}[\mathbf{U}]\cdot\nabla_{\mathbf{U}_{x_\ell}}\mathscr{H}_0[\mathbf{U}])\,.$$

The reader is referred to Appendix A for a derivation of the latter.

We are interested in the analysis of the dynamics of near-plane periodic uniformly traveling waves of (1). Let us first recall that a (uniformly traveling) wave is a solution whose time evolution occurs through the action of symmetries. We say that the wave is a plane wave when in a suitable frame it is constant in all but one direction, and that it is periodic if it is periodic up to symmetries. Given the foregoing set of symmetries, after choosing for the sake of concreteness, the direction of propagation as[2] $\mathbf{e}_1$ and the normalizing period

2. Throughout the text, we denote as $\mathbf{e}_j$ the jth vector of the canonical basis of $\mathbf{R}^d$. In particular, $\mathbf{e}_1 = (1,0,\dots,0)$.

to be 1 through the introduction of wavenumbers, we are interested in solutions to (1) of the form

$$\begin{aligned} f(t,\mathbf{x}) &= \mathrm{e}^{-\,\mathrm{i}\left(\underline{k}_\phi\,(x-\underline{c}_x\,t)+\underline{\omega}_\phi\,t\right)}\,\underline{f}(\underline{k}_x\,(x-\underline{c}_x\,t))\\ &= \mathrm{e}^{-\,\mathrm{i}(\underline{k}_\phi\,x+(\underline{\omega}_\phi-\underline{k}_\phi\,\underline{c}_x)\,t)}\,\underline{f}(\underline{k}_x\,x+\underline{\omega}_x\,t)\,, \end{aligned}$$

with profile $\underline{f}$ 1-periodic, wavenumbers $(\underline{k}_\phi,\underline{k}_x)\in\mathbf{R}^2$, $\underline{k}_x>0$, time frequencies $(\underline{\omega}_\phi,\underline{\omega}_x)\in\mathbf{R}^2$, spatial speed $\underline{c}_x\in\mathbf{R}$, where

$$\mathbf{x}=(x,\mathbf{y})\qquad \underline{\omega}_x=-\underline{k}_x\,\underline{c}_x\,.$$

In other terms, we consider solutions to (3) in the form

$$\mathbf{U}(t,\mathbf{x})=\mathrm{e}^{\left(\underline{k}_\phi\,(x-\underline{c}_x\,t)+\underline{\omega}_\phi\,t\right)\mathbf{J}}\,\underline{\mathcal{U}}(\underline{k}_x\,(x-\underline{c}_x\,t))\,,\tag{7}$$

with $\underline{\mathcal{U}}$ 1-periodic (and nonconstant). More general periodic plane waves are also considered in Appendix D. Beyond references to results involved in our analysis given along the text and comparison to the literature provided near each main statement, in order to place our contribution in a bigger picture, we refer the reader to [37] for a general background on nonlinear wave dynamics and to [46, 30, 14] for material more specific to Hamiltonian systems.

To set the frame for linearization, we observe that going to a frame adapted to the background wave in (7) by

$$\mathbf{U}(t,\mathbf{x})=\mathrm{e}^{\left(\underline{k}_\phi\,(x-\underline{c}_x\,t)+\underline{\omega}_\phi\,t\right)\mathbf{J}}\,\mathbf{V}(t,\underline{k}_x\,(x-\underline{c}_x\,t),\mathbf{y})\,,$$

changes (3) into

$$\begin{aligned} &\partial_t\mathbf{V}=\mathbf{J}\delta\mathscr{H}[\mathbf{V}]\,,\\ &\mathscr{H}[\mathbf{V}]:=\mathscr{H}_0(\mathbf{V},(\underline{k}_x\partial_x+\underline{k}_\phi\mathbf{J})\mathbf{V},\nabla_{\mathbf{y}}\mathbf{V})-\underline{\omega}_\phi\mathscr{M}[\mathbf{V}]+\underline{c}_x\mathbb{Q}_1(\mathbf{V},(\underline{k}_x\partial_x+\underline{k}_\phi\mathbf{J})\mathbf{V})\\ &\qquad=\mathscr{H}_0(\mathbf{V},(\underline{k}_x\partial_x+\underline{k}_\phi\mathbf{J})\mathbf{V},\nabla_{\mathbf{y}}\mathbf{V})-\left(\underline{\omega}_\phi-\underline{k}_\phi\,\underline{c}_x\right)\mathscr{M}[\mathbf{V}]-\underline{\omega}_x\mathbb{Q}_1[\mathbf{V}]\,, \end{aligned}\tag{8}$$

and that $(t,x,\mathbf{y})\mapsto\underline{\mathcal{U}}(x)$ is a stationary solution to (8). Direct linearization of (8) near this solution provides the linear equation $\partial_t\mathbf{V}=\mathcal{L}\,\mathbf{V}$ with $\mathcal{L}$ defined by

$$\mathcal{L}\mathbf{V}=\mathbf{J}\,\mathrm{Hess}\,\mathscr{H}[\underline{\mathcal{U}}](\mathbf{V})\,,\tag{9}$$

where Hess denotes the variational Hessian, that is, Hess $=L\delta$ with L denoting linearization. Incidentally, we point out that the natural splitting

$$\mathscr{H}_0=\mathscr{H}_0^x+\mathscr{H}^{\mathbf{y}}\,,\qquad \mathscr{H}^{\mathbf{y}}\,[\mathbf{U}]=\tfrac12\kappa(\|\mathbf{U}\|^2)\|\nabla_{\mathbf{y}}\mathbf{U}\|^2\,,$$

may be followed all the way through frame change and linearization

$$\begin{aligned} \mathscr{H}&=\mathscr{H}^x+\mathscr{H}^{\mathbf{y}}\,,\\ \mathcal{L}&=\mathbf{J}\,\mathrm{Hess}\,\mathscr{H}^x[\underline{\mathcal{U}}]+\mathbf{J}\,\mathrm{Hess}\,\mathscr{H}^{\mathbf{y}}[\underline{\mathcal{U}}]=:\mathcal{L}^x+\mathcal{L}^{\mathbf{y}}\,, \end{aligned}$$

with $\mathcal{L}^{\mathbf{y}}=-\kappa(\|\underline{\mathcal{U}}\|^2)\mathbf{J}\,\Delta_{\mathbf{y}}$.

As made explicit in Section 3.1 at the spectral and linear level, to make the most of the spatial structure of periodic plane waves, it is convenient to introduce a suitable Bloch–Fourier integral transform. As a result, one may analyze the action of $\mathcal{L}$ defined on $L^2(\mathbf{R})$ through[3] the actions of $\mathcal{L}_{\xi,\boldsymbol{\eta}}$ defined on $L^2((0,1))$ with periodic boundary conditions, where $(\xi,\boldsymbol{\eta})\in[-\pi,\pi]\times\mathbf{R}^{d-1}$, ξ being a longitudinal Floquet exponent and $\boldsymbol{\eta}$ a transverse Fourier frequency. The operator $\mathcal{L}_{\xi,\boldsymbol{\eta}}$ encodes the action of $\mathbf{J}\,\mathrm{Hess}\,\mathscr{H}[\underline{\mathcal{U}}]$ on functions of the form

$$\mathbf{x}=(x,\mathbf{y})\mapsto \mathrm{e}^{\mathrm{i}\,\xi x+\mathrm{i}\,\boldsymbol{\eta}\cdot\mathbf{y}}\,\mathbf{W}(x)\,,\qquad \mathbf{W}(\cdot+1)=\mathbf{W}\,,$$

through

$$\mathbf{J}\,\mathrm{Hess}\,\mathscr{H}[\underline{\mathcal{U}}]\left((x,\mathbf{y})\mapsto \mathrm{e}^{\mathrm{i}\,\xi x+\mathrm{i}\,\boldsymbol{\eta}\cdot\mathbf{y}}\,\mathbf{W}(x)\right)(\mathbf{x})=\mathrm{e}^{\mathrm{i}\,\xi x+\mathrm{i}\,\boldsymbol{\eta}\cdot\mathbf{y}}\,(\mathcal{L}_{\xi,\boldsymbol{\eta}}\mathbf{W})(x)\,.$$

In particular, the spectrum of $\mathcal{L}$ coincides with the union over $(\xi,\boldsymbol{\eta})$ of the spectra of $\mathcal{L}_{\xi,\boldsymbol{\eta}}$. In turn, as recalled in Section 3.3, generalizing the analysis of Gardner [25], the spectrum of each $\mathcal{L}_{\xi,\boldsymbol{\eta}}$ may be conveniently analyzed with the help of an Evans function $D_\xi(\,\cdot\,,\boldsymbol{\eta})$, an analytic function whose zeros agree in location and algebraic multiplicity[4] with the spectrum of $\mathcal{L}_{\xi,\boldsymbol{\eta}}$. A large part of our spectral analysis hinges on the derivation of an expansion of $D_\xi(\lambda,\boldsymbol{\eta})$ when $(\lambda,\xi,\boldsymbol{\eta})$ is small (Theorem 3.2).

As derived in Section 2, families of plane periodic profiles in a fixed direction – here taken to be $\mathbf{e}_1$ – form four-dimensional manifolds when identified up to rotational and spatial translations, parametrized by $(\mu_x,c_x,\omega_\phi,\mu_\phi)$, where (μ_x,μ_ϕ) are constants of integration of profile equations associated with conservation laws (4) and (6) (or, more precisely, its first component since we consider waves propagating along $\mathbf{e}_1$). The averages along wave profiles of quantities of interest are expressed in terms of an action integral $\Theta(\mu_x,c_x,\omega_\phi,\mu_\phi)$ and its derivatives. This action integral plays a prominent role in our analysis. A significant part of our analysis, indeed, aims at reducing properties of operators acting on infinite-dimensional spaces to properties of this finite-dimensional function.

After these preliminary observations, we here give a brief account of each of our main results and provide only later in the text more specialized comments around precise statements. Our main achievements are essentially twofold. On the one hand, we provide counterparts to the main upshots of [12, 13, 9, 10, 11, 50] – derived for one-dimensional Hamiltonian equations of Korteweg type – for one-dimensional Hamiltonian equations of Schrödinger type. On the other hand, we extend parts of this analysis to the present multidimensional framework.

3. As, by using Fourier transforms on constant-coefficient operators one reduces their action on functions over the whole space to finite-dimensional operators parametrized by Fourier frequencies.

4. Defined, for the spectrum, as the rank of the residue of the resolvent map.

1.1. Longitudinal perturbations. — To describe the former, we temporarily restrict ourselves to longitudinal perturbations or somewhat equivalently to the case $d = 1$. At the linear level, this amounts to setting $\boldsymbol{\eta} = 0$.

The first set of results that we prove concerns perturbations that in the above adapted moving frame are spatially periodic with the same period as the background waves, so-called coperiodic perturbations. At the linear level, this amounts to restricting ourselves to $(\xi, \boldsymbol{\eta}) = (0, 0)$. In Theorem 4.1, as in [9], we prove that a wave of parameters $(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi)$, such that $\operatorname{Hess}(\Theta)(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi)$ is invertible, is

1. H^1 (conditionally) nonlinearly (orbitally) stable under coperiodic longitudinal perturbations if $\operatorname{Hess}(\Theta)(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi)$ has negative signature 2 and $\partial^2_{\mu_x}\Theta(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi) \neq 0$;
2. spectrally (exponentially) unstable under coperiodic longitudinal perturbations if this negative signature is either 1 or 3, or equivalently if $\operatorname{Hess}(\Theta)(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi)$ has negative determinant.

The main upshot here is that instead of the rather long list of assumptions that would be required by directly applying the abstract general theory [27, 14], assumptions are both simple and expressed in terms of the finite-dimensional Θ.

Then, as in [10], we elucidate these criteria in two limits of interest, the solitary-wave limit when the spatial period tends to infinity and the harmonic limit when the amplitude of the wave tends to zero. To describe the solitary-wave regime, let us point out that the solitary-wave profiles under consideration are naturally parametrized by (c_x, ρ, k_ϕ), where $\rho > 0$ is the limiting value at spatial infinities of its mass, and that families of solitary waves also come with an action integral $\Theta_{(s)}(c_x, \rho, k_\phi)$, known as the Boussinesq momentum of stability [16, 5, 6] and for Schrödinger-like equations associated with the famous Vakhitov–Kolokolov slope condition [55]. The reader may consult [58, 41, 14] as entering gates in the quite extensive mathematical literature on the latter. In Theorem 4.3, we prove that

1. in nondegenerate small-amplitude regimes, waves are nonlinearly stable to coperiodic perturbations;
2. in the large-period regime near a solitary wave of parameters $(\underline{c}_x^{(0)}, \underline{\rho}^{(0)}, \underline{k}_\phi^{(0)})$, coperiodic spectral instability occurs when $\partial^2_{c_x}\Theta_{(s)}(\underline{c}_x^{(0)}, \underline{\rho}^{(0)}, \underline{k}_\phi^{(0)}) < 0$, whereas coperiodic nonlinear orbital stability holds when $\partial^2_{c_x}\Theta_{(s)}(\underline{c}_x^{(0)}, \underline{\rho}^{(0)}, \underline{k}_\phi^{(0)}) > 0$.

Both results appear to be new in this context. Note in particular that our small-amplitude regime is disjoint from the cubic semilinear one considered in [24] since there the constant asymptotic mass is taken to be zero. Yet, in the large-period regime, the spectral instability result could also be partly recovered by combining a spectral instability result for solitary waves available in the above-mentioned literature for some semilinear equations, with a nontrivial spectral perturbation argument from [26, 52, 57].

The rest of our results on longitudinal perturbations concerns sideband longitudinal perturbations, that is, perturbations corresponding to $(\xi, \boldsymbol{\eta}) = (\xi, 0)$ with ξ small (but nonzero) and geometrical optics *à la* Whitham [56].

The latter is derived by inserting in (3) the two-phase slow/fast-oscillatory *ansatz*

$$\mathbf{U}^{(\varepsilon)}(t,x) = \mathrm{e}^{\frac{1}{\varepsilon}\varphi_\phi^{(\varepsilon)}(\varepsilon\, t, \varepsilon\, x)\,\mathbf{J}}\,\mathcal{U}^{(\varepsilon)}\left(\varepsilon\, t, \varepsilon\, x; \frac{\varphi_x^{(\varepsilon)}(\varepsilon\, t, \varepsilon\, x)}{\varepsilon}\right);, \tag{10}$$

with, for any (T, X), $\zeta \mapsto \mathcal{U}^{(\varepsilon)}(T, X; \zeta)$ periodic of period 1 and, as $\varepsilon \to 0$,

$$\begin{aligned}
\mathcal{U}^{(\varepsilon)}(T,X;\zeta) &= \mathcal{U}_0(T,X;\zeta) + \varepsilon\,\mathcal{U}_1(T,X;\zeta) + o(\varepsilon)\,,\\
\varphi_\phi^{(\varepsilon)}(T,X) &= (\varphi_\phi)_0(T,X) + \varepsilon\,(\varphi_\phi)_1(T,X) + o(\varepsilon)\,,\\
\varphi_x^{(\varepsilon)}(T,X) &= (\varphi_x)_0(T,X) + \varepsilon\,(\varphi_x)_1(T,X) + o(\varepsilon)\,.
\end{aligned}$$

Arguing heuristically and identifying orders of ε as detailed in Section 4.2, one obtains that the foregoing *ansatz* may describe the behavior of solutions to (3) provided that the leading profile $\mathcal{U}_0$ stems from a slow modulation of wave parameters

$$\mathcal{U}_0(T,X;\zeta) = \mathcal{U}^{(\mu_x, c_x, \omega_\phi, \mu_\phi)(T,X)}(\zeta)\,, \tag{11}$$

where $\mathcal{U}^{(\mu_x, c_x, \omega_\phi, \mu_\phi)}$ denotes a wave profile of parameters $(\mu_x, c_x, \omega_\phi, \mu_\phi)$, with local wavenumbers $(k_\phi, k_x) = (\partial_X(\varphi_\phi)_0, \partial_X(\varphi_x)_0)$, and the slow evolution of local parameters obeys

$$k_x \mathbf{A}_0\, \mathrm{Hess}\,\Theta\;(\partial_T + c_x \partial_X)\begin{pmatrix}\mu_x\\ c_x\\ \omega_\phi\\ \mu_\phi\end{pmatrix} = \mathbf{B}_0\,\partial_X\begin{pmatrix}\mu_x\\ c_x\\ \omega_\phi\\ \mu_\phi\end{pmatrix}, \tag{12}$$

where

$$\mathbf{A}_0 := \begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&-1&0\\0&0&0&-1\end{pmatrix}, \qquad \mathbf{B}_0 := \begin{pmatrix}0&1&0&0\\1&0&0&0\\0&0&0&1\\0&0&1&0\end{pmatrix}.$$

Let us point out that actually, in the derivation sketched above, System (12) is first obtained in the equivalent form

$$(13)\qquad \begin{cases} \partial_T k_x = \partial_X \omega_x \\ \partial_T \mathscr{q} = \partial_X \left(\mu_x - c_x \mathscr{q}\right) \\ \partial_T \mathscr{m} = \partial_X \left(\mu_\phi - c_x \mathscr{m}\right) \\ \partial_T k_\phi = \partial_X \left(\omega_\phi - c_x\, k_\phi\right) \end{cases},$$

with $\mathscr{m}$ and $\mathscr{q}$ denoting averages over one period of, respectively, $\mathscr{M}(\mathcal{U})$ and $\mathbb{Q}_1(\mathcal{U}, (k_\phi\, \mathbf{J} + k_x \partial_\zeta)\mathcal{U})$, with $\mathcal{U} = \mathcal{U}^{(\mu_x, c_x, \omega_\phi, \mu_\phi)}$. Note that two of the equations of (13) are so-called conservations of waves, whereas the two others arise as averaged equations. For a more thorough introduction to modulation systems such as (13) in the context of Hamiltonian systems, we refer the reader to the introduction of [10] and references therein.

Our second set of results concerns spectral validation of the foregoing formal arguments in the slow/sideband regime. More explicitly, as in [13], we obtain in the specialization of Theorem 3.2 to $\boldsymbol{\eta} = 0$ that

$$(14)\qquad D_\xi(\lambda, 0) \overset{(\lambda,\xi)\to(0,0)}{=} \det\left(\lambda\, \mathbf{A}_0 \text{ Hess } \Theta(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi) - \mathrm{i}\, \xi\, \mathbf{B}_0\right) + \mathcal{O}\left(\left(|\lambda|^4 + |\xi|^4\right) |\lambda|\right),$$

for a wave of parameter $(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi)$. This connects the slow/sideband Bloch spectral dispersion relation for the wave profile $\underline{\mathcal{U}}$ as a stationary solution to (8) with the slow/slow Fourier dispersion relation for $(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi)$ as a solution to (12). Among the direct consequences of (14) derived in Corollary 4.5, we point out that this implies that if $\text{Hess}(\Theta)(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi)$ is invertible, and (12) fails to be weakly hyperbolic at $(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi)$, then the wave is spectrally exponentially unstable to sideband perturbations. Afterwards, as in [11], in Theorem 4.6, we combine asymptotics for (12) with the foregoing instability criterion to derive that waves are spectrally exponentially unstable to longitudinal sideband perturbations

1. in nondegenerate small-amplitude regimes near a harmonic wave train, such that $\delta_{hyp} < 0$ or $\delta_{BF} < 0$, with indices $(\delta_{hyp}, \delta_{BF})$ defined explicitly in (71) and (72);
2. in the large-period regime near a solitary wave of parameters $(\underline{c}_x^{(0)}, \underline{\rho}^{(0)}, \underline{k}_\phi^{(0)})$, such that $\partial_{c_x}^2 \Theta_{(s)}(\underline{c}_x^{(0)}, \underline{\rho}^{(0)}, \underline{k}_\phi^{(0)}) < 0$.

Again, these results are new in this context, except for the corollary about weak hyperbolicity that overlaps with the recent preprint[5] [20] – based on the recent

5. During the editorial process of the present paper it became a published paper.

[40] –, which appeared during the preparation of the present contribution. Note, however, that our proof of the corollary is different, and our assumptions are considerably weaker.

Our third set of results concerning longitudinal perturbations shows that for spectrally stable waves in a suitable dispersive sense, by including higher-order corrections in (12), one obtains a version of (11)–(12) that captures at any arbitrary order the large-time asymptotics for the slow/sideband part of the linearized dynamics. Besides the oscillatory-integral analysis directly borrowed from [50], this hinges on a spectral validation of the formal asymptotics – obtained in Theorem 4.7 – as predictors for expansions of spectral projectors (and not only of spectral curves) in the slow/sideband regime. The identified decay is inherently of dispersive type, and we refer the curious reader to [42, 23] for comparisons with the well-known theory for constant-coefficient operators. Let us stress that deriving global-in-time dispersive estimates for nonconstant, nonnormal operators is a considerably harder task and that the analysis in [50] has provided the first-ever dispersive estimates for the linearized dynamics about a periodic wave. We also point out that a large-time dynamical validation of modulation systems for general data – as opposed to a spectral validation or a validation for well-prepared data – requires the identification of effective initial data for modulation systems, a highly nontrivial task that cannot be guessed from the formal arguments sketched above.

At this stage, the reader could wonder how, in a not-so-large number of pages, Schrödinger-like counterparts to Korteweg-like results, originally requiring a quite massive body of literature [12, 13, 9, 10, 11, 50], can be obtained. There are at least two phenomena at work. On the one hand, we have actually left a significant part of [11, 50] without counterparts. Results in [11] were mainly motivated by the study of dispersive shocks, and the few stability results adapted here from [11] were obtained there almost in passing. The analysis in [50] studies the full linearized dynamics for the Korteweg–de Vries equation. Yet, the underlying arguments being technically demanding, we have chosen to adapt here only the part of the analysis directly related to modulation behavior, for the sake of both consistency and brevity. On the other hand, some of the results proved here are actually deduced from the results derived for some Korteweg-like systems rather than proved from scratch.

The key to these deductions is a suitable study of Madelung's transformation [43]. As we develop in Section 2.3, even at the level of generality considered here, Madelung's transformation provides a convenient hydrodynamic formulation of (1) of Korteweg type. A solution $\mathbf{U}$ to (3) is related to a solution $(\rho, \mathbf{v})$, with curl-free velocity $\mathbf{v}$, of a Euler–Korteweg system through

$$\mathbf{U} = \sqrt{2\,\rho}\;\mathrm{e}^{\theta\,\mathbf{J}}\begin{pmatrix}1\\0\end{pmatrix}, \qquad \mathbf{v} = \nabla_{\mathbf{x}}\theta\,.$$

We refer the reader to [18] for some background on the transformation and its mathematical use. Let us stress that the transformation changes the geometric structure of the equations dramatically, in both its group of symmetries and its conservation laws. A basic observation that makes the Madelung's transformation particularly efficient here is that nonconstant periodic wave profiles stay away from zero. Consistently, the asymptotic regimes that we consider also lie in the far-from-zero zone. Our coperiodic nonlinear orbital stability result is in particular proved here by studying in Lemma 4.2 correspondences through the Madelung's transformation. Even more efficiently, the identification of respective action integrals also reduces the asymptotic expansions of Hess Θ required here to those already obtained in [10, 11]. For the sake of completeness, in Section 3.2, we also carry out a detailed study of spectral correspondences. Yet those fail to fully elucidate spectral behavior near $(\lambda, \xi, \boldsymbol{\eta}) = (0,0,0)$ and, thus, they play no role in our spectral and linear analyses.

1.2. General perturbations. — In the second part of our analysis, we extend the spectral results of the longitudinal part to genuinely multidimensional perturbations.

To begin with we provide an instability criterion for perturbations that are longitudinally coperiodic, that is, that corresponds to $\xi = 0$. The corresponding result, Corollary 5.1-(1), is made somewhat more explicit in Lemma 5.2. Yet we do not investigate the corresponding asymptotics because in the multidimensional context, we are more interested in determining whether waves may be stable against any perturbation, and the present coperiodic instability criterion turns out to be weaker than the slow/sideband one contained in Corollary 5.1-(2), which we describe now.

The second, and main, set of results of this second part focuses on slow/sideband perturbations corresponding to the regime $(\lambda, \xi, \boldsymbol{\eta})$ small. In the latter regime, generalizing the longitudinal analysis, we derive an instability criterion, interpret it in terms of formal geometrical optics, and elucidate it in both the small-amplitude and large-period asymptotics.

Concerning geometrical optics, a key observation is that even if one is merely interested in the stability of waves in the specific form (7), the relevant modulation theory involves more general waves in the form

$$\mathbf{U}(t,\mathbf{x}) = \mathrm{e}^{\left(\underline{\mathbf{k}}_\phi\cdot(\mathbf{x}-\underline{c}_x\,\underline{\mathbf{e}}_x\,t)+\underline{\omega}_\phi\,t\right)\mathbf{J}}\ \underline{\mathcal{U}}(\underline{\mathbf{k}}_x\cdot(\mathbf{x}-\underline{c}_x\,\underline{\mathbf{e}}_x\,t))\,, \tag{15}$$

with $\underline{\mathcal{U}}$ 1-periodic and $\underline{\mathbf{k}}_x$ nonzero of unitary direction $\underline{\mathbf{e}}_x$. The main departure in (15) from (7) is that $\underline{\mathbf{k}}_x$ and $\underline{\mathbf{k}}_\phi$ are no longer assumed to be colinear. To stress comparisons with (7), let us decompose $(\underline{\mathbf{k}}_x, \underline{\mathbf{k}}_\phi)$ as

$$\underline{\mathbf{k}}_x = \underline{k}_x\,\underline{\mathbf{e}}_x\,,\qquad \underline{\mathbf{k}}_\phi = \underline{k}_\phi\,\underline{\mathbf{e}}_x + \widetilde{\underline{\mathbf{k}}}_\phi\,,$$

with $\underline{\widetilde{\mathbf{k}}}_\phi$ orthogonal to $\underline{\mathbf{e}}_x$. In Section 2.6, we show that this more general set of plane waves may be conveniently parametrized by $(\mu_x, c_x, \omega_\phi, \mu_\phi, \mathbf{e}_x, \widetilde{\mathbf{k}}_\phi)$, with $(\mathbf{e}_x, \widetilde{\mathbf{k}}_\phi)$ varying in the $2(d-1)$-dimensional manifold of vectors, such that $\mathbf{e}_x$ is unitary, and $\widetilde{\mathbf{k}}_\phi$ is orthogonal to $\mathbf{e}_x$.

With this in hand, adding possible slow dependence on $\mathbf{y}$ in (10) through

$$\mathbf{U}^{(\varepsilon)}(t,\mathbf{x}) = \mathrm{e}^{\frac{1}{\varepsilon}\varphi_\phi^{(\varepsilon)}(\varepsilon\, t, \varepsilon\, \mathbf{x})\,\mathbf{J}}\,\mathcal{U}^{(\varepsilon)}\left(\varepsilon\, t, \varepsilon\, \mathbf{x}; \frac{\varphi_x^{(\varepsilon)}(\varepsilon\, t, \varepsilon\, \mathbf{x})}{\varepsilon}\right) \tag{16}$$

and arguing as before leads to the modulation behavior

$$\mathcal{U}_0(T, \mathbf{X}; \zeta) = \mathcal{U}^{(\mu_x, c_x, \omega_\phi, \mu_\phi, \mathbf{e}_x, \widetilde{\mathbf{k}}_\phi)(T, \mathbf{X})}(\zeta)\,, \tag{17}$$

with local wavevectors $(\mathbf{k}_\phi, \mathbf{k}_x) = (\nabla_{\mathbf{X}}(\varphi_\phi)_0, \nabla_{\mathbf{X}}(\varphi_x)_0)$, and the slow evolution of local parameters obeys

$$\begin{cases} \partial_T \mathbf{k}_x = \nabla_{\mathbf{X}} \omega_x \\ \partial_T \boldsymbol{q} = \nabla_{\mathbf{X}}\left(\mu_x - c_x \boldsymbol{q} + \dfrac{1}{2}\,\tau_0\,\|\widetilde{\mathbf{k}}_\phi\|^2\right) \\ \qquad\qquad + \operatorname{div}_{\mathbf{X}}\left(\tau_1\,\widetilde{\mathbf{k}}_\phi \otimes \widetilde{\mathbf{k}}_\phi + \tau_2\,(\widetilde{\mathbf{k}}_\phi \otimes \mathbf{e}_x + \mathbf{e}_x \otimes \widetilde{\mathbf{k}}_\phi) + \tau_3\,(\mathbf{e}_x \otimes \mathbf{e}_x - \mathrm{I}_d)\right) \\ \partial_T m = \operatorname{div}_{\mathbf{X}}\left((\mu_\phi - c_x m)\,\mathbf{e}_x + \tau_1\,\widetilde{\mathbf{k}}_\phi\right) \\ \partial_T \mathbf{k}_\phi = \nabla_{\mathbf{X}}\,(\omega_\phi - c_x\,k_\phi) \end{cases}, \tag{18}$$

with extra constraints (propagated by the time evolution) that $\mathbf{k}_x$ and $\mathbf{k}_\phi$ are curl free. In System (18), $\mathbf{a} \otimes \mathbf{b}$ denotes the matrix of (j,ℓ)-coordinate $\mathbf{a}_\ell\,\mathbf{b}_j$, $\operatorname{div}_{\mathbf{X}}$ acts on matrix-valued maps row-wise, and $\boldsymbol{q}$, τ_0, τ_1, τ_2, and τ_3 denote the averages over one period of, respectively,

$$\begin{aligned} &\mathbb{Q}(\mathcal{U}, (k_\phi\,\mathbf{J} + k_x\partial_\zeta)\mathcal{U}), \quad \kappa'(\|\mathcal{U}\|^2)\,\|\mathcal{U}\|^2, \quad \kappa(\|\mathcal{U}\|^2)\,\|\mathcal{U}\|^2, \\ &\kappa(\|\mathcal{U}\|^2)\,\mathbf{J}\mathcal{U}\cdot(k_\phi\,\mathbf{J} + k_x\partial_\zeta)\mathcal{U} \quad \text{and} \quad \kappa(\|\mathcal{U}\|^2)\,\|(k_\phi\,\mathbf{J} + k_x\partial_\zeta)\mathcal{U}\|^2\,, \end{aligned}$$

with $\mathcal{U} = \mathcal{U}^{(\mu_x, c_x, \omega_\phi, \mu_\phi, \mathbf{e}_x, \widetilde{\mathbf{k}}_\phi)}$. Linearizing System (18) about the constant $(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi, \mathbf{e}_1, 0)$ yields, after a few manipulations,

$$\begin{cases} \underline{k}_x \mathbf{A}_0\,\mathrm{Hess}\,\Theta\,\left(\partial_T + \underline{c}_x\partial_X\right)\begin{pmatrix}\mu_x\\ c_x\\ \omega_\phi\\ \mu_\phi\end{pmatrix} = \mathbf{B}_0\,\partial_X\begin{pmatrix}\mu_x\\ c_x\\ \omega_\phi\\ \mu_\phi\end{pmatrix} + \begin{pmatrix}0 & 0\\ \underline{\tau}_3 & \underline{\tau}_2\\ \underline{\tau}_2 & \underline{\tau}_1\\ 0 & 0\end{pmatrix}\begin{pmatrix}\operatorname{div}_{\mathbf{X}}(\mathbf{e}_x)\\ \operatorname{div}_{\mathbf{X}}(\widetilde{\mathbf{k}}_\phi)\end{pmatrix} \\ \qquad (\partial_T + \underline{c}_x\partial_X)\,\mathbf{e}_x = -\,(\nabla_{\mathbf{X}} - \mathbf{e}_1\,\partial_X)\,c_x \\ \qquad (\partial_T + \underline{c}_x\partial_X)\,\widetilde{\mathbf{k}}_\phi = (\nabla_{\mathbf{X}} - \mathbf{e}_1\,\partial_X)\,\omega_\phi \end{cases} \tag{19}$$

with extra constraints that $\widetilde{\mathbf{k}}_\phi$ and $\mathbf{e}_x$ are orthogonal to $\underline{\mathbf{e}}_x = \mathbf{e}_1$ and that $k_x\underline{\mathbf{e}}_x + \underline{k}_x\,\mathbf{e}_x$ and $k_\phi\underline{\mathbf{e}}_x + \underline{k}_\phi\,\mathbf{e}_x + \widetilde{\mathbf{k}}_\phi$ are curl free, where (k_x, k_ϕ) are deviations given explicitly as

$$k_x = -\underline{k}_x^2\,\mathrm{d}\,(\partial_{\mu_x}\Theta)(\mu_x, c_x, \omega_\phi, \mu_\phi)\,, \quad k_\phi = \frac{k_x}{\underline{k}_x}\,\underline{k}_\phi - \underline{k}_x\,\mathrm{d}\,(\partial_{\mu_\phi}\Theta)(\mu_x, c_x, \omega_\phi, \mu_\phi)\,,$$

where total derivatives are taken with respect to $(\mu_x, c_x, \omega_\phi, \mu_\phi)$, and evaluation is at $(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi, \mathbf{e}_1, 0)$. In System 19, $\mathrm{Hess}\,\Theta = \mathrm{Hess}_{(\mu_x,c_x,\omega_\phi,\mu_\phi)}\,\Theta$ is likewise evaluated at $(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi, \mathbf{e}_1, 0)$, and $\mathbf{A}_0$ and $\mathbf{B}_0$ are as in System (12).

As made explicit in Section 5.1, our Theorem 3.2 provides a spectral validation of (18) in the form

$$\lambda^{2(d-1)} \times D_\xi(\lambda, \boldsymbol{\eta}) \overset{(\lambda,\xi,\boldsymbol{\eta})\to(0,0,0)}{=}$$
$$\det\left(\lambda\begin{pmatrix}\mathrm{I}_{2(d-1)} & 0\\ 0 & \mathbf{A}_0\ \mathrm{Hess}\,\Theta\end{pmatrix} - \mathrm{i}\,\xi\begin{pmatrix}0 & 0\\ 0 & \mathbf{B}_0\end{pmatrix} + \left(\begin{array}{cc|cccc} 0 & 0 & 0 & -\mathrm{i}\,\boldsymbol{\eta} & 0 & 0\\ 0 & 0 & 0 & 0 & \mathrm{i}\,\boldsymbol{\eta} & 0\\ \hline 0 & 0 & 0 & 0 & 0 & 0\\ \frac{\tau_3}{\underline{k}_x}\ \mathrm{i}\,\boldsymbol{\eta}^\mathsf{T} & \frac{\tau_2}{\underline{k}_x}\ \mathrm{i}\,\boldsymbol{\eta}^\mathsf{T} & 0 & 0 & 0 & 0\\ \frac{\tau_2}{\underline{k}_x}\ \mathrm{i}\,\boldsymbol{\eta}^\mathsf{T} & \frac{\tau_1}{\underline{k}_x}\ \mathrm{i}\,\boldsymbol{\eta}^\mathsf{T} & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right)\right)$$
$$+\,\mathcal{O}\left(|\lambda|^{2(d-1)}\,(|\lambda| + |\xi| + \|\boldsymbol{\eta}\|)^5\right),$$

or equivalently in the form

$$D_\xi(\lambda, \boldsymbol{\eta}) \overset{(\lambda,\xi,\boldsymbol{\eta})\to(0,0,0)}{=} \det\left(\lambda\,\mathbf{A}_0\ \mathrm{Hess}\,\Theta - \mathrm{i}\,\xi\,\mathbf{B}_0 + \frac{\|\boldsymbol{\eta}\|^2}{\lambda}\,\mathbf{C}_0\right) + \mathcal{O}\left((|\lambda| + |\xi| + \|\boldsymbol{\eta}\|)^5\right), \tag{20}$$

with

$$\mathbf{C}_0 := \begin{pmatrix} 0 & 0 & 0 & 0\\ 0 & -\sigma_3 & \sigma_2 & 0\\ 0 & -\sigma_2 & \sigma_1 & 0\\ 0 & 0 & 0 & 0\end{pmatrix}, \qquad \sigma_j = \frac{\tau_j}{\underline{k}_x}, \quad j \in \{1,2,3\}\,.$$

In the foregoing, again $\mathrm{Hess}\,\Theta = \mathrm{Hess}_{(\mu_x,c_x,\omega_\phi,\mu_\phi)}\,\Theta(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi, \mathbf{e}_1, 0)$. Note that, consistently with the equality, the structure of $\mathbf{B}_0$ and $\mathbf{C}_0$ implies that the apparent singularity in λ of the left-hand side of (20) is, indeed, spurious, each factor $\|\boldsymbol{\eta}\|^2/\lambda$ being necessarily paired with a factor λ in the expansion of the determinant. The only other rigorous spectral validation of a multidimensional modulation system that we are aware of is [45], which deals with systems of parabolic conservation laws.

It follows directly from (20) that if (19) fails to be weakly hyperbolic at $(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi, \mathbf{e}_1, 0)$, then the corresponding wave is spectrally exponentially unstable. In Section 5.2, besides this most general instability criterion, we provide two instability criteria, more specific but easier to check, corresponding to the breaking of multiple roots near $\boldsymbol{\eta} = 0$ (Proposition 5.4) and near $\xi = 0$ (Proposition 5.5)m respectively.

Afterwards, we turn to the elucidation of the full instability criterion in the asymptotic regimes already studied in the longitudinal part. Our striking conclusion is that, when $d \geq 2$, in nondegenerate cases, plane waves of the form (7) are spectrally exponentially unstable in both the small-amplitude (Theorem 5.8) and the large-period (Theorem 5.6) regimes. More explicitly, we prove that such waves are spectrally exponentially unstable to slow/sideband perturbations

1. in nondegenerate, small-amplitude regimes near an harmonic wave train, such that $\delta_{hyp} \neq 0$ and $\delta_{BF} \neq 0$, with indices defined explicitly in (71) and (72);
2. in the large-period regime near a solitary wave of parameters $(\underline{c}_x^{(0)}, \underline{\rho}^{(0)}, \underline{k}_\phi^{(0)})$, such that $\partial_{c_x}^2 \Theta_{(s)}(\underline{c}_x^{(0)}, \underline{\rho}^{(0)}, \underline{k}_\phi^{(0)}) \neq 0$.

Let us stress that to obtain the latter we derive various instability scenarios – all hinging on expansion (20), thus occurring in the region $(\lambda, \xi, \boldsymbol{\eta})$ small – corresponding to different instability criteria. The point is that the union of these criteria covers all possibilities. In particular, in the harmonic limit, the argument requires the full strength of the joint expansion in $(\lambda, \xi, \boldsymbol{\eta})$, and it is relatively elementary – see Appendix B – to check that the instability is non-trivial in the sense that it occurs even in cases when the limiting constant states is spectrally stable. We also stress that both asymptotic results are derived by extending to the multidimensional context some of the finest properties of longitudinal modulated systems proved in [11] from the asymptotic expansions of $\operatorname{Hess}\Theta$ obtained in [10].

All the results about general perturbations are new, including this form of the formal derivation of a modulation system. The only small overlap of which we are aware is with [40] appearing during the preparation of the present contribution and studying to leading order the spectrum of $\mathcal{L}_{(0,\boldsymbol{\eta})}$ near $\lambda = 0$, when $\boldsymbol{\eta}$ is small. Even for this partial result, our proof is different, and our assumptions are considerably weaker. Let us also stress that [40] discusses neither modulation systems nor asymptotic regimes. Finally, we point out that the operator $\mathcal{L}_{(0,\boldsymbol{\eta})}$ depends on $\boldsymbol{\eta}$ only through the scalar parameter $\|\boldsymbol{\eta}\|^2$, so that the problem studied in [40] fits the frame of spectral analysis of analytic one-parameter perturbations, a subpart of general spectral perturbation theory that is considerably more regular and simpler, even compared to two-parameter perturbations like we consider here. Concerning the latter, we refer the reader to [38, 21] for a general background on spectral theory. In addition to [40], in the large-period regime, we again expect that the spectral instability result could be partly recovered by combining a spectral instability result for solitary waves available in the literature for some specific semilinear equations [51], with a nontrivial spectral perturbation argument as mentioned above [26, 52, 57].

Extensions and open problems. — Since such plane waves play a role in the nearby modulation theory, the reader may wonder whether our main results extend to more general plane waves in the form (15). As was pointed out in Section 2.6, it is straightforward to check that it is so for all results concerning longitudinal perturbations. Concerning instability under general perturbations, a first obvious answer is that instabilities persist under perturbations and, thus, extend to waves associated with small $\widetilde{\mathbf{k}}_\phi$. In Appendix D, we show how to extend the results to all waves in the semilinear case, that is, when κ is constant, and in the high-dimensional case, that is, when $d \geq 3$.

Finally, in Appendix C, we show how to extend our results to anisotropic equations, even with dispersion of mixed signature, for waves propagating in a principal direction.

Although our results strongly hint at the multidimensional spectral instability of any periodic plane wave, they do leave this question unanswered, even for semilinear versions of (1). In the reverse direction of leaving some hope for stability, we stress that there are known natural examples of classes of one-dimensional equations for which both small-amplitude and large-period waves are unstable, but there are bands of stable periodic waves. The reader is referred to [2, 35, 1] for examples on the Korteweg–de Vries/Kuramoto–Sivashinsky equation and to [4, 3] for examples on shallow-water Saint-Venant equations. We regard the elucidation of this possibility, even numerically, as an important open question. We point out, as an intermediate issue whose resolution would already be interesting, and probably more tractable, the determination of whether there exist periodic waves of (1) associated with wave parameters at which the modulation system (18) is weakly hyperbolic.

Let us conclude the global presentation of our main results by recalling that more specialized discussions, including more technical comparisons to the literature, are provided throughout the text.

Outline. — The next two sections contain general preliminary material: the first one on the structure of wave profile manifolds and the following one on adapted spectral theory. The latter, however, contains two highly nontrivial results: spectral conjugations through a linearized Madelung's transform (Section 3.2) and the slow/sideband expansion of the Evans function (Theorem 3.2) – a key block of our spectral analysis. After these two sections follow two sections devoted, respectively, to longitudinal perturbations and to general perturbations. Appendices contain key algebraic relations stemming from invariances and symmetries used throughout the text (Appendix A), the examination of constant-state spectral stability (Appendix B), extensions to more general equations (Appendix C), and more general profiles (Appendix D) and a table of symbols (Appendix E).

Subsections of the two main sections are in clear correspondence with various sets of results described in the Introduction, so that the reader interested in a specific class of results may jump to the relevant part of the analysis and refer to the table of symbols to search for the definitions involved.

Notation. — Before engaging in more concrete analysis, we here make our conventions for vectorial, differential and variational notation explicit.

Throughout we identify vectors with columns. The partial derivative with respect to a variable a is denoted ∂_a, or ∂_j when variables are numbered and a is the jth one. The piece of notation d stands for differentiation, so that $\mathrm{d}\,g(x)(h)$ denotes the derivative of g at x in the direction h. The Jacobian matrix $\mathrm{Jac}\,g(x)$ is the matrix associated with the linear map $\mathrm{d}\,g(x)$ in the canonical basis. The gradient $\nabla g(x)$ is the adjoint matrix of $\mathrm{Jac}\,g(x)$, and we sometimes use suffix $_a$ to denote the gradient with respect to a. The Hessian operator Hess is given as the Jacobian of the gradient, $\mathrm{Hess}\,g = \mathrm{Jac}(\nabla g)$. The divergence operator div is the opposite of the dual of the ∇ operator. We say that a vector field is curl free if its Jacobian is valued in symmetric matrices.

For any two vectors $\mathbf{V}$ and $\mathbf{W}$ in $\mathbf{R}^{d_0}$, thought of as column vectors, $\mathbf{V}\otimes\mathbf{W}$ stands for the rank-1, square matrix of size d_0

$$\mathbf{V}\otimes\mathbf{W} = \mathbf{V}\,\mathbf{W}^{\mathsf{T}},$$

whatever d_0, where $^{\mathsf{T}}$ stands for matrix transposition. Acting on square-valued maps, div acts row-wise. A dot $\cdot$ denotes the standard scalar product. Since, as a consequence of invariance by rotational changes, our differential operators act mostly component-wise, we believe that no confusion is possible and do not mark differences of meaning of $\cdot$ even when two vectorial structures coexist. The convention is that summation in scalar products is taken over compatible dimensions. For instance,

$$\mathbf{V}\cdot\nabla_{\mathbf{U}}\mathscr{H}_0(\mathbf{U},\nabla_{\mathbf{x}}\mathbf{U}) = \sum_{j=1}^{2}\mathbf{V}_j\,\partial_{U_j}\mathscr{H}_0(\mathbf{U},\nabla_{\mathbf{x}}\mathbf{U}),$$

$$\mathbf{e}_x\cdot\nabla_{\nabla_{\mathbf{x}}\mathbf{U}}\mathscr{H}_0(\mathbf{U},\nabla_{\mathbf{x}}\mathbf{U}) = \sum_{j=1}^{d}(\mathbf{e}_x)_j\,\nabla_{\partial_j\mathbf{U}}\mathscr{H}_0(\mathbf{U},\nabla_{\mathbf{x}}\mathbf{U}),$$

$$\nabla_{\mathbf{x}}\mathbf{U}\cdot\nabla_{\nabla_{\mathbf{x}}\mathbf{U}}\mathscr{H}_0(\mathbf{U},\nabla_{\mathbf{x}}\mathbf{U}) = \sum_{j=1}^{d}\sum_{\ell=1}^{2}\partial_j U_\ell\,\nabla_{\partial_j U_\ell}\mathscr{H}_0(\mathbf{U},\nabla_{\mathbf{x}}\mathbf{U}).$$

We also use notation for differential calculus on functional spaces (thus, in infinite dimensions), mostly in variational form. We use L to denote linearization, analogously to d, so that $L(\mathcal{F})[\mathbf{U}]\mathbf{V}$ denotes the linearization of $\mathcal{F}$ at $\mathbf{U}$ in the direction $\mathbf{V}$. Notation δ stands for variational derivative and plays a role analogous to gradient, except that we use it on functional densities instead of functionals. With suitable boundary conditions, this would be the gradient for the L^2 structure of the functional associated with the given functional density at hand. We only consider functional densities depending on derivatives up to order 1, so that this is explicitly given as

$$\delta\mathcal{A}[\mathbf{U}] = \nabla_{\mathbf{U}}\mathcal{A}(\mathbf{U},\nabla_{\mathbf{x}}\mathbf{U}) - \operatorname{div}_{\mathbf{x}}\left(\nabla_{\nabla_{\mathbf{x}}\mathbf{U}}\mathcal{A}(\mathbf{U},\nabla\mathbf{U})\right).$$

In this context, Hess denotes the linearization of the variational derivative, $\operatorname{Hess} = L\delta$, here explicitly

$$\begin{aligned}\operatorname{Hess}\mathcal{A}[\mathbf{U}]\mathbf{V} &= \mathrm{d}_{(\mathbf{U},\nabla_{\mathbf{x}}\mathbf{U})}(\nabla_{\mathbf{U}}\mathcal{A})(\mathbf{U},\nabla_{\mathbf{x}}\mathbf{U})(\mathbf{V},\nabla_{\mathbf{x}}\mathbf{V})\\ &\quad - \operatorname{div}_{\mathbf{x}}\left(\mathrm{d}_{(\mathbf{U},\nabla_{\mathbf{x}}\mathbf{U})}(\nabla_{\nabla_{\mathbf{x}}\mathbf{U}}\mathcal{A})(\mathbf{U},\nabla_{\mathbf{x}}\mathbf{U})(\mathbf{V},\nabla_{\mathbf{x}}\mathbf{V})\right).\end{aligned}$$

Even when one is interested in a single wave, nearby waves enter in stability considerations. We almost systematically use underlining to denote quantities associated with the particular given background wave under study. In particular, when a wave parametrization is available, underlining denotes evaluation at the parameters of the particular wave under study.

2. Structure of periodic wave profiles

To begin with, we gather some facts about plane traveling wave manifolds. Up to Section 2.6, we restrict to waves in the form (7). Consistently, here, for concision, we may set $\mathbb{Q} = \mathbb{Q}_1$.

2.1. Radius equation. — To analyze the structure of the wave profiles, we step back from (7) and look for profiles in the form

$$\mathbf{U}(t,\mathbf{x}) = \mathrm{e}^{\omega_\phi\, t\mathbf{J}}\ \mathcal{V}(x - c_x\, t)\,, \tag{21}$$

without normalizing to enforce 1-periodicity. The profile equation becomes

$$0 = \delta\mathscr{H}_{\mathrm{u}}[\mathcal{V}]\,, \qquad \text{with} \qquad \mathscr{H}_{\mathrm{u}}[\mathcal{V}] = \mathscr{H}_0[\mathcal{V}] - \omega_\phi\mathscr{M}[\mathcal{V}] + c_x\mathbb{Q}[\mathcal{V}]\,. \tag{22}$$

Moreover, we note that, as a consequence of the rotational and spatial translation invariances of $\mathscr{H}_{\mathrm{u}}$, (22) also contains the following form of mass and momentum conservations

$$0 = -\frac{\mathrm{d}}{\mathrm{d}\,x}\left(\mathcal{V}\cdot\mathbf{J}\nabla_{\mathbf{U}_x}\mathscr{H}_{\mathrm{u}}[\mathcal{V}]\right), \tag{23}$$

$$0 = \frac{\mathrm{d}}{\mathrm{d}\,x}\left(-\mathscr{H}_{\mathrm{u}}[\mathcal{V}] + \partial_x\mathcal{V}\cdot\nabla_{\mathbf{U}_x}\mathscr{H}_{\mathrm{u}}[\mathcal{V}]\right) \tag{24}$$

and introduce μ_ϕ and μ_x corresponding constants of integration, so that

$$\mu_\phi = \mathbf{J}\mathcal{V} \cdot \nabla_{\mathbf{U}_x} \mathscr{H}_{\mathrm{u}}[\mathcal{V}]\,, \tag{25}$$

$$\mu_x = \frac{\mathrm{d}\,\mathcal{V}}{\mathrm{d}\,x} \cdot \nabla_{\mathbf{U}_x} \mathscr{H}_{\mathrm{u}}[\mathcal{V}] - \mathscr{H}_{\mathrm{u}}[\mathcal{V}]\,. \tag{26}$$

Observe that reciprocally by differentiating (25)–(26) one obtains

$$\mathbf{J}\mathcal{V} \cdot \delta\mathscr{H}_{\mathrm{u}}[\mathcal{V}] = 0 \qquad \text{and} \qquad \left(\mathcal{V} \cdot \frac{\mathrm{d}\,\mathcal{V}}{\mathrm{d}\,x}\right) \mathcal{V} \cdot \delta\mathscr{H}_{\mathrm{u}}[\mathcal{V}] = 0\,,$$

which yields (22) provided that the set where $\mathcal{V} \cdot \frac{\mathrm{d}\,\mathcal{V}}{\mathrm{d}\,x}$ vanishes has an empty interior.

We now check that the above-mentioned condition on $\mathcal{V} \cdot \frac{\mathrm{d}\,\mathcal{V}}{\mathrm{d}\,x}$ excludes only solutions of the form (21) that have constant modulus and travel uniformly in phase. Since (22) is a differential equation, it is already clear that if $\mathcal{V}$ vanishes on some nontrivial interval, then $\mathcal{V} \equiv 0$, and from now on we exclude this case from our analysis. Then, if $\mathcal{V} \cdot \frac{\mathrm{d}\,\mathcal{V}}{\mathrm{d}\,x}$ vanishes on some nontrivial interval, it follows that on this interval $\|\mathcal{V}\|$ is a constant equal to some $r_0 > 0$ and from (25) that

$$\mathcal{V}(x) = \mathrm{e}^{\frac{2\mu_\phi - c_x\, r_0^2}{2\,\kappa(r_0^2)\, r_0^2}\, x\,\mathbf{J}} \left(r_0\; \mathrm{e}^{\varphi_\phi \mathbf{J}}\,\mathbf{e}_1\right),$$

for some $\varphi_\phi \in \mathbf{R}$. Since the formula provides a solution to (22) everywhere, this holds everywhere, and henceforth we also exclude this case. However, these constant solutions are discussed further in Appendix B.

Now to analyze (22) further we first recast (25)–(26) in a more explicit form,

$$\begin{aligned}
\mu_\phi &= \kappa(\|\mathcal{V}\|^2)\,\mathbf{J}\mathcal{V} \cdot \frac{\mathrm{d}\,\mathcal{V}}{\mathrm{d}\,x} + \frac{c_x}{2}\|\mathcal{V}\|^2\,,\\
\mu_x &= \frac{1}{2}\kappa(\|\mathcal{V}\|^2)\left\|\frac{\mathrm{d}\,\mathcal{V}}{\mathrm{d}\,x}\right\|^2 - W(\|\mathcal{V}\|^2) + \frac{\omega_\phi}{2}\|\mathcal{V}\|^2\,.
\end{aligned}$$

Then we set $\alpha = \|\mathcal{V}\|^2$ and observe that

$$\begin{aligned}
\alpha\,\frac{\mathrm{d}\,\mathcal{V}}{\mathrm{d}\,x} &= \frac{1}{2}\frac{\mathrm{d}\,\alpha}{\mathrm{d}\,x}\,\mathcal{V} + \mathbf{J}\mathcal{V} \cdot \frac{\mathrm{d}\,\mathcal{V}}{\mathrm{d}\,x}\,\mathbf{J}\mathcal{V}\,,\\
\alpha\left\|\frac{\mathrm{d}\,\mathcal{V}}{\mathrm{d}\,x}\right\|^2 &= \frac{1}{4}\left(\frac{\mathrm{d}\,\alpha}{\mathrm{d}\,x}\right)^2 + \left(\mathbf{J}\mathcal{V} \cdot \frac{\mathrm{d}\,\mathcal{V}}{\mathrm{d}\,x}\right)^2\,.
\end{aligned}$$

In particular, from (25)–(26) stems

$$\frac{1}{8}\kappa(\alpha)\left(\frac{\mathrm{d}\,\alpha}{\mathrm{d}\,x}\right)^2 + \mathcal{W}_\alpha(\alpha; c_x, \omega_\phi, \mu_\phi) = \mu_x\,\alpha\,, \tag{27}$$

with

$$\mathcal{W}_\alpha(\alpha; c_x, \omega_\phi, \mu_\phi) := -W(\alpha)\,\alpha + \frac{\omega_\phi}{2}\,\alpha^2 + \frac{1}{8}\frac{(2\mu_\phi - c_x\,\alpha)^2}{\kappa(\alpha)}\,. \tag{28}$$

Consistently going back to (22), one derives

$$\frac{1}{4}\kappa(\alpha)\frac{\mathrm{d}^2\alpha}{\mathrm{d}x^2}+\frac{\kappa'(\alpha)}{\kappa(\alpha)}(\mu_x\,\alpha-\mathcal{W}_\alpha(\alpha))+\partial_\alpha\mathcal{W}_\alpha(\alpha)=\mu_x\,. \tag{29}$$

As a consequence, since $\alpha\geq 0$, if α vanishes at some point, then its derivative also vanishes there, and $\mu_\phi=0$. From this we deduce near the same point

$$\frac{\mathrm{d}\alpha}{\mathrm{d}x}=\mathcal{O}(\alpha)\qquad\text{and}\qquad \mathbf{J}\mathcal{V}\cdot\frac{\mathrm{d}\mathcal{V}}{\mathrm{d}x}=\mathcal{O}(\alpha)\qquad\qquad\text{hence}\qquad \frac{\mathrm{d}\mathcal{V}}{\mathrm{d}x}=\mathcal{O}(\sqrt{\alpha})\,,$$

This implies $\mu_x=-W(0)$ and corresponds to the trivial solution to (22) given by $\mathcal{V}\equiv 0$ that we have already ruled out. Note that this exclusion may be enforced by requiring $(\mu_x,\mu_\phi)\neq(-W(0),0)$.

The foregoing discussion ensures that $\mathcal{V}$ actually does not vanish, so that in particular $r=\sqrt{\alpha}=\|\mathcal{V}\|$ is a smooth function solving

$$\frac{1}{2}\kappa(r^2)\left(\frac{\mathrm{d}r}{\mathrm{d}x}\right)^2+\mathcal{W}_r(r;c_x,\omega_\phi,\mu_\phi)=\mu_x, \tag{30}$$

where $\mathcal{W}_r$ is defined by

$$\begin{aligned}\mathcal{W}_r(r;c_x,\omega_\phi,\mu_\phi)&:=\frac{1}{r^2}\mathcal{W}_\alpha(r^2;c_x,\omega_\phi,\mu_\phi)\\&=-W(r^2)+\frac{\omega_\phi}{2}\,r^2+\frac{1}{8}\frac{(2\mu_\phi-c_x\,r^2)^2}{\kappa(r^2)\,r^2}\,,\end{aligned} \tag{31}$$

and

$$\kappa(r^2)\frac{\mathrm{d}^2r}{\mathrm{d}x^2}+2r\,\frac{\kappa'(r^2)}{\kappa(r^2)}(\mu_x-\mathcal{W}_r(r))+\partial_r\mathcal{W}_r(r)=0\,. \tag{32}$$

Note that the excluded case where r is constant equal to some r_0 happens only when

$$\mu_x=-\mathcal{W}_r(r_0;c_x,\omega_\phi,\mu_\phi)\qquad\text{and}\qquad 0=\partial_r\mathcal{W}_r(r_0;c_x,\omega_\phi,\mu_\phi)\,.$$

When coming back from (30) to (22) some care is needed when μ_ϕ is zero since then $\mathcal{W}_r$ may be extended to $\mathbf{R}$, but solutions to (30) taking negative values must still be discarded. Except for that point, one readily obtains from (25) that the family of solutions to (22)–(25)–(26) is associated with any solution r to (30)

$$\mathcal{V}(x)=\mathrm{e}^{\left(\int_0^{x+\varphi_x}\frac{2\mu_\phi-c_x\,r(y)}{2\,\kappa(r(y)^2)\,r(y)^2}\,\mathrm{d}y\right)\mathbf{J}}\left(r(x+\varphi_x)\;\mathrm{e}^{\varphi_\phi\mathbf{J}}\,\mathbf{e}_1\right),$$

parametrized by rotational and spatial shifts $(\varphi_\phi,\varphi_x)\in\mathbf{R}^2$.

Classical arguments show that if parameters are such that (30) defines[6] a nontrivial closed curve in phase-space that is included in the half-plane $r>0$,

6. The relation could define many connected components but implicitly we discuss them one by one. See Figures 2.1 and 2.2.

then the above construction yields a wave of the sought form, unique up to translations in rotational and spatial positions.

2.2. Jump map. — Rather than focusing on the existence of periodic waves, we now turn our focus to their parametrization, assuming the existence of a given reference wave $\underline{\mathcal{V}}$. As announced, parameters associated with $\underline{\mathcal{V}}$ are underlined, and, more generally, any functional $\mathcal{F}$ evaluated at the reference wave is denoted $\underline{\mathcal{F}}$.

In the unscaled framework, instead of wavenumbers (k_x, k_ϕ), we rather manipulate the spatial period $X_x := 1/k_x$ and the rotational shift[7] $\xi_\phi := k_\phi/k_x$ that satisfy

$$\mathcal{V}(\cdot + X_x) = \mathrm{e}^{\xi_\phi\,\mathbf{J}}\,\mathcal{V}(\cdot).$$

Our goal is to show the existence of nearby waves and to determine which parameters are suitable for wave parametrization among

ω_ϕ ,	rotational pulsation
c_x ,	spatial speed
$\mathcal{V}(0),\ \dfrac{\mathrm{d}\,\mathcal{V}}{\mathrm{d}\,x}(0),$	initial data for the wave profile ODE
$\mu_\phi,\ \mu_x,$	constants of integration associated with conservation laws
$X_x,$	spatial period
$\xi_\phi,$	rotational shift after a period
$\varphi_\phi,\ \varphi_x,$	rotational and spatial translations.

It follows from the Cauchy–Lipschitz theory that functions $\mathcal{V}$ satisfying equation (22) are uniquely and smoothly determined by initial data $(\mathcal{V}(0), \frac{\mathrm{d}\,\mathcal{V}}{\mathrm{d}\,x}(0)) = (\mathcal{V}_0, \mathcal{V}_1)$, and parameters of the equation (ω_ϕ, c_x), on some common neighborhood of $[0, \underline{X}_x]$ provided that $(\mathcal{V}_0, \mathcal{V}_1, \omega_\phi, c_x)$ is sufficiently close to $(\underline{\mathcal{V}}(0), \frac{\mathrm{d}\,\underline{\mathcal{V}}}{\mathrm{d}\,x}(0), \underline{\omega}_\phi, \underline{c}_x)$. Note that the point 0 plays no particular role, and we may use a spatial translation to replace it with another nearby point so as to ensure suitable conditions on $(\underline{\mathcal{V}}(0), \frac{\mathrm{d}\,\underline{\mathcal{V}}}{\mathrm{d}\,x}(0))$. In particular, there is no loss in generality in assuming that $\underline{\mathcal{V}}(0) \cdot \frac{\mathrm{d}\,\underline{\mathcal{V}}}{\mathrm{d}\,x}(0) \neq 0$.

At this stage, to carry out algebraic manipulations it is convenient to introduce notation

$$\begin{aligned}
\mathcal{S}_\phi[\mathbf{U}] &:= \mathbf{J}\mathbf{U} \cdot \nabla_{\mathbf{U}_x} \mathscr{H}_0[\mathbf{U}]\,,\\
\mathcal{S}_x[\mathbf{U}] &:= -\mathscr{H}_0[\mathbf{U}] + \mathbf{U}_x \cdot \nabla_{\mathbf{U}_x} \mathscr{H}_0[\mathbf{U}]\,.
\end{aligned}$$

7. We refrain from using the word Floquet exponent for ξ_ϕ to avoid confusion with Floquet exponents involved in integral transforms.

so that (25)–(26) is written as

$$\mu_\phi = \mathcal{S}_\phi[\mathcal{V}] + c_x\mathcal{M}[\mathcal{V}]\,, \tag{33}$$

$$\mu_x = \mathcal{S}_x[\mathcal{V}] + \omega_\phi\mathcal{M}[\mathcal{V}]\,. \tag{34}$$

Now we observe that $\mathrm{d}_{\mathbf{U}_x}(\mathcal{S}_\phi,\mathcal{S}_x)(\underline{\mathcal{V}}(0),\frac{\mathrm{d}\,\underline{\mathcal{V}}}{\mathrm{d}\,x}(0))$ has the determinant $(\kappa(\|\underline{\mathcal{V}}(0)\|^2))^2\,\underline{\mathcal{V}}(0)\cdot\frac{\mathrm{d}\,\underline{\mathcal{V}}}{\mathrm{d}\,x}(0)\neq 0$. In particular, as a consequence of the implicit function theorem, for $(\mathcal{V}_0,\mathcal{V}_1,c_x,\omega_\phi,\mu_\phi,\mu_x)$ near $(\underline{\mathcal{V}}(0),\frac{\mathrm{d}\,\underline{\mathcal{V}}}{\mathrm{d}\,x}(0),\underline{c}_x,\underline{\omega}_\phi,\underline{\mu}_\phi,\underline{\mu}_x)$,

$$\begin{aligned}\mu_\phi &= \mathcal{S}_\phi(\mathcal{V}_0,\mathcal{V}_1) + c_x\mathcal{M}(\mathcal{V}_0)\,,\\ \mu_x &= \mathcal{S}_x(\mathcal{V}_0,\mathcal{V}_1) + \omega_\phi\mathcal{M}(\mathcal{V}_0)\,,\end{aligned}$$

is smoothly (and equivalently) solved as

$$\mathcal{V}_1 = \mathcal{V}^1(\mathcal{V}_0;c_x,\omega_\phi,\mu_\phi,\mu_x)\,.$$

The same is true near $(\underline{\mathcal{V}}(X_x),\frac{\mathrm{d}\,\underline{\mathcal{V}}}{\mathrm{d}\,x}(X_x),\underline{c}_x,\underline{\omega}_\phi,\underline{\mu}_\phi,\underline{\mu}_x)$. This implies that, on the one hand, one may replace $(\mathcal{V}_0,\mathcal{V}_1,\omega_\phi,c_x)$ with $(\mathcal{V}_0,\omega_\phi,c_x,\mu_\phi,\mu_x)$ in the parametrization of solutions to (22) and, on the other hand, since values of $(\mathcal{S}_\phi[\mathcal{V}]+c_x\mathcal{M}[\mathcal{V}],\mathcal{S}_x[\mathcal{V}]+\omega_\phi\mathcal{M}[\mathcal{V}])$ are invariant under the flow of (22), that, as a consequence of the Cauchy–Lipschitz theory, solutions to (22) defined on a neighborhood of $[0,\underline{X}_x]$ extend as solutions on $\mathbf{R}$ such that $\mathcal{V}(\cdot+X_x)=e^{\xi_\phi\,\mathbf{J}}\mathcal{V}(\cdot)$ if and only if $\mathcal{V}(X_x)=e^{\xi_\phi\,\mathbf{J}}\mathcal{V}(0)$.

We now show that we may replace $(\mathcal{V}_0,\omega_\phi,c_x,\mu_\phi,\mu_x)$ with $(\varphi_\phi,\varphi_x,\omega_\phi,c_x,\mu_\phi,\mu_x)$ by taking the solution corresponding to $\mathcal{V}_0=\underline{\mathcal{V}}(0)$ and acting with rotational and spatial translations. The action of rotational and spatial translations is $\mathcal{V}(\cdot)\mapsto\mathcal{V}_{\varphi_\phi,\varphi_x}:=\mathrm{e}^{\varphi_\phi\mathbf{J}}\,\mathcal{V}(\cdot+\varphi_x)$. Obviously, it leaves the set of periodic-wave profiles invariant and, among parameters, interacts only with initial data, thus, after the elimination of $\mathcal{V}_1$, only with $\mathcal{V}_0$. Let us denote by $\mathcal{V}^{(\mu_x,c_x,\omega_\phi,\mu_\phi)}$ the solution to (22), such that

$$\mathcal{V}^{(\mu_x,c_x,\omega_\phi,\mu_\phi)}(0)=\underline{\mathcal{V}}(0)\,,\quad \frac{\mathrm{d}}{\mathrm{d}\,x}\mathcal{V}^{(\mu_x,c_x,\omega_\phi,\mu_\phi)}(0)=\mathcal{V}^1(\underline{\mathcal{V}}(0);c_x,\omega_\phi,\mu_\phi,\mu_x)\,.$$

At background parameters the map

$$(\varphi_\phi,\varphi_x,\mu_x,c_x,\omega_\phi,\mu_\phi)\mapsto(\mathcal{V}^{(\mu_x,c_x,\omega_\phi,\mu_\phi)})_{\varphi_\phi,\varphi_x}(0)$$

has Jacobian determinant with respect to variations in (φ_ϕ,φ_x) equal to $\underline{\mathcal{V}}(0)\cdot\frac{\mathrm{d}\,\underline{\mathcal{V}}}{\mathrm{d}\,x}(0)\neq 0$. Thus, as claimed, as a consequence of the implicit function theorem, one may smoothly and invertibly replace $(\mathcal{V}_0,\omega_\phi,c_x,\mu_\phi,\mu_x)$ with $(\varphi_\phi,\varphi_x,\omega_\phi,c_x,\mu_\phi,\mu_x)$ to parametrize solutions to (22) near the background profile.

As a conclusion, when identified up to rotational and spatial translations, periodic-wave profiles are smoothly identified as the zero level set of the map

$$(\mu_x,c_x,\omega_\phi,\mu_\phi,X_x,\xi_\phi)\mapsto\mathcal{V}^{(\mu_x,c_x,\omega_\phi,\mu_\phi)}(X_x)-\mathrm{e}^{\xi_\phi\mathbf{J}}\,\mathcal{V}^{(\mu_x,c_x,\omega_\phi,\mu_\phi)}(0)\,.$$

Now, at background parameters, the foregoing map has Jacobian determinant with respect to variations in (X_x, ξ_ϕ) equal to $\underline{\mathcal{V}}(0) \cdot \frac{\mathrm{d}\,\underline{\mathcal{V}}}{\mathrm{d}\,x}(0) \neq 0$. Therefore, a third application of the implicit function theorem achieves the proof of the following proposition.

Proposition 2.1. — *Near a periodic-wave profile with nonconstant mass, periodic wave profiles form a six-dimensional manifold smoothly parametrized as*

$$(\varphi_\phi, \varphi_x, \omega_\phi, c_x, \mu_\phi, \mu_x) \mapsto (\mathcal{V}^{(\omega_\phi, c_x, \mu_\phi, \mu_x)}_{\varphi_\phi, \varphi_x}, X_x(\omega_\phi, c_x, \mu_\phi, \mu_x), \xi_\phi(\omega_\phi, c_x, \mu_\phi, \mu_x))\,,$$

with, for any $(\varphi_\phi, \varphi_x)$,

$$\mathcal{V}^{(\omega_\phi, c_x, \mu_\phi, \mu_x)}_{\varphi_\phi, \varphi_x}(\cdot) = \mathrm{e}^{\varphi_\phi \mathbf{J}}\, \mathcal{V}^{(\omega_\phi, c_x, \mu_\phi, \mu_x)}_{0,0}(\cdot + \varphi_x)\,.$$

2.3. Madelung's transformation. — To ease comparisons with the analyses in [12, 13, 9, 10, 11] for dispersive systems of Korteweg type, including Euler–Korteweg systems and quasi-linear Korteweg–de Vries equations, we now provide hydrodynamic formulations of (1)/(3) and correspondences between the respective periodic waves. The reader is referred to [8] for similar discussions concerning other kinds of traveling waves.

In the present section, we temporarily go back to the general multidimensional framework. On the one hand, we consider for $f = a + \mathrm{i}\, b$, $\mathbf{U} = \begin{pmatrix} a \\ b \end{pmatrix}$, a system in the form

$$\partial_t \mathbf{U} = \mathbf{J}\, \delta \mathscr{H}_\#[\mathbf{U}] \qquad \text{with} \qquad \mathscr{H}_\#\,[\mathbf{U}] = \mathscr{H}_{\mathrm{eff}}\left(\mathscr{M}[\mathbf{U}], \mathfrak{Q}[\mathbf{U}], \frac{1}{2}\,\|\nabla_{\mathbf{x}} \mathbf{U}\|^2\right). \tag{35}$$

Then we introduce

$$\mathcal{U}(\rho, \theta) := \sqrt{2\,\rho}\; \mathrm{e}^{\theta\, \mathbf{J}}(\mathbf{e}_1)\;, \qquad (\rho, \theta) \in \mathbf{R}_+ \times \mathbf{R}\,,$$

and

$$H_\#[(\rho, \mathbf{v})] := \mathscr{H}_{\mathrm{eff}}\left(\rho, \rho\,\mathbf{v}, \frac{1}{4\rho}\|\nabla_{\mathbf{x}}\rho\|^2 + \rho\,\|\mathbf{v}\|^2\right)$$

and observe that

$$\rho = \mathscr{M}[\mathcal{U}(\rho(\cdot), \theta(\cdot))]\,, \quad \nabla_{\mathbf{x}}\theta = \frac{\mathfrak{Q}[\mathcal{U}(\rho(\cdot), \theta(\cdot))]}{\mathscr{M}[\mathcal{U}(\rho(\cdot), \theta(\cdot))]}\,,$$
$$H_\#[(\rho, \nabla_{\mathbf{x}}\theta)] = \mathscr{H}_\#\,[\mathcal{U}(\rho(\cdot), \theta(\cdot))]\;.$$

We also point out that

$$\mathcal{U}(\rho, \theta) \cdot \mathbf{J}\, \delta \mathscr{H}_\#\,[\mathcal{U}(\rho(\cdot), \theta(\cdot))] = \mathrm{div}_{\mathbf{x}}\,(\delta_{\mathbf{v}} H_\#[(\rho, \nabla_{\mathbf{x}}\theta)]) \tag{36}$$

$$\frac{1}{2\rho}\mathbf{J}\,\mathcal{U}(\rho, \theta) \cdot \mathbf{J}\, \delta \mathscr{H}_\#\,[\mathcal{U}(\rho(\cdot), \theta(\cdot))] = \delta_\rho H_\#[(\rho, \nabla_{\mathbf{x}}\theta)]\,, \tag{37}$$

so that if $\mathbf{U}$ solves (35) and is bounded away from zero, then

$$(\rho, \mathbf{v}) := \left(\mathscr{M}[\mathbf{U}], \frac{\mathbb{Q}[\mathbf{U}]}{\mathscr{M}[\mathbf{U}]}\right) \tag{38}$$

solves

$$\partial_t \begin{pmatrix} \rho \\ \mathbf{v} \end{pmatrix} = \mathcal{J}\, \delta H_{\#}[(\rho, \mathbf{v})]\,, \tag{39}$$

with the constraint that $\mathbf{v}$ is curl free, where $\mathcal{J}$ denotes the skew-symmetric operator

$$\mathcal{J} := \begin{pmatrix} 0 & \mathrm{div}_{\mathbf{x}} \\ \nabla_{\mathbf{x}} & 0 \end{pmatrix}.$$

Note that the curl-free constraint is preserved by the time evolution, so that it is sufficient to prescribe it on the initial data.

Reciprocally, if $(\rho, \mathbf{v})$ solves (39), and ρ is bounded below away from zero, then for any θ such that

$$\partial_t \theta = \delta_\rho H_{\#}[(\rho, \mathbf{v})]\,, \tag{40}$$

we have $\nabla_{\mathbf{v}} \theta = \mathbf{v}$, and $\mathbf{U} := \mathcal{U}(\rho(\cdot), \theta(\cdot))$ solves (35). Note, moreover, that under such conditions, for any $(t_0, \mathbf{x}_0, \theta_0)$, (40) possesses a unique solution, such that $\theta(t_0, \mathbf{x}_0) = \theta_0$ and that, for any $\mathbf{x}_0$, (40) could alternatively be replaced by: for any t, $\partial_t \theta(t, \mathbf{x}_0) = \delta_\rho H_{\#}[(\rho, \mathbf{v})](t, \mathbf{x}_0)$ and $\nabla_{\mathbf{v}} \theta(t, \cdot\,) = \mathbf{v}(t, \cdot\,)$.

We point out that whereas the Madelung transformation $\mathbf{U} \mapsto (\rho, \mathbf{v})$ quotients the rotational invariance, it preserves the time and space translation invariances. With respect to the latter, we consider

$$Q_j[\rho, \mathbf{v}] := \rho\, \mathbf{v} \cdot \mathbf{e}_j\,, \qquad j = 1, \dots, d,$$

and observe that, on the one hand, Q_j generates spatial translations along the direction $\mathbf{e}_j$ in the sense that if $\mathbf{v}$ is curl-free, then

$$\mathbf{e}_j \cdot \nabla \begin{pmatrix} \rho \\ \mathbf{v} \end{pmatrix} = \mathcal{J}\, \delta Q_j[(\rho, \mathbf{v})]\,,$$

and that, on the other hand,

$$Q_j[(\rho, \nabla\theta)] = \mathbb{Q}_j\left[\mathcal{U}(\rho(\cdot), \theta(\cdot))\right]\,.$$

We also note that (39) implies

$$\begin{aligned} \partial_t\left(Q_j(\rho, \mathbf{v})\right) = \partial_j\left(\rho\, \partial_\rho H_{\#}[(\rho, \mathbf{v})] - H_{\#}[(\rho, \mathbf{v})]\right) \\ + \mathrm{div}_{\mathbf{x}}\left(v_j\, \nabla_{\mathbf{v}} H_{\#}[(\rho, \mathbf{v})] + \rho_{x_j}\, \nabla_{\nabla_{\mathbf{x}}\rho} H_{\#}[(\rho, \mathbf{v})]\right), \end{aligned}$$

and, for comparison with (6), that when $\mathbf{U} = \mathcal{U}(\rho(\cdot), \theta(\cdot))$, $\mathbf{v} = \nabla_{\mathbf{x}} \theta$,

$$\begin{aligned} \nabla_{\mathbf{U}_{x_j}} \mathbb{Q}_j[\mathbf{U}] \cdot \mathbf{J}\delta\mathscr{H}_{\#}[\mathbf{U}] &= \rho\, \partial_\rho H_{\#}[(\rho, \mathbf{v})]\,, \\ \mathbf{J}\delta\mathbb{Q}_j[\mathbf{U}] \cdot \nabla_{\mathbf{U}_{x_\ell}} \mathscr{H}_{\#}[\mathbf{U}] &= v_j\, \partial_{v_\ell} H_{\#}[(\rho, \mathbf{v})] + \rho_{x_j}\, \partial_{\rho_{x_\ell}} H_{\#}[(\rho, \mathbf{v})]\,. \end{aligned}$$

Concerning the time translation invariance, we note that (39) implies

$$\partial_t\left(H_\#(\rho,\mathbf{v})\right)=\mathrm{div}_{\mathbf{x}}\Big(\delta_\rho H_\#[(\rho,\mathbf{v})]\,\nabla_{\mathbf{v}}H_\#[(\rho,\mathbf{v})]\\ +\mathrm{div}_{\mathbf{x}}(\nabla_{\mathbf{v}}H_\#[(\rho,\mathbf{v})])\,\nabla_{\nabla_{\mathbf{x}}\rho}H_\#[(\rho,\mathbf{v})]\Big)\,;$$

and, for comparison with (5), that when $\mathbf{U}=\mathcal{U}(\rho(\cdot),\theta(\cdot))$, $\mathbf{v}=\nabla_{\mathbf{x}}\theta$,

$$\nabla_{\mathbf{U}_{x_j}}\mathscr{H}_\#[\mathbf{U}]\cdot\mathbf{J}\delta\mathscr{H}_\#[\mathbf{U}]=\delta_\rho H_\#[(\rho,\mathbf{v})]\,\partial_{v_j}H_\#[(\rho,\mathbf{v})]\\ +\mathrm{div}_{\mathbf{x}}(\nabla_{\mathbf{v}}H_\#[(\rho,\mathbf{v})])\,\partial_{\rho_{x_j}}H_\#[(\rho,\mathbf{v})]\,.$$

In the hydrodynamic formulation, what replaces to some extent the rotational invariance and its accompanying conservation law for $\mathscr{M}[\mathbf{U}]$ is the fact that the time evolution in (39) obeys a system of $d+1$ conservation laws and that one may add to $H_\#$ any affine function of $(\rho,\mathbf{v})$ without changing (39). With this respect, to compare (4) with the equation on $\partial_t\rho$, we note that when $\mathbf{U}=\mathcal{U}(\rho(\cdot),\theta(\cdot))$, $\mathbf{v}=\nabla_{\mathbf{x}}\theta$,

$$\mathbf{J}\delta\mathscr{M}[\mathbf{U}]\cdot\nabla_{\mathbf{U}_{x_j}}\mathscr{H}_\#[\mathbf{U}]=\partial_{v_j}H_\#[(\rho,\mathbf{v})]\,.$$

To make the discussion slightly more concrete, we compute that when $\mathscr{H}_\#=\mathscr{H}_0$, one receives

$$H_0[(\rho,\mathbf{v})]:=H_\#[(\rho,\mathbf{v})]=\kappa(2\,\rho)\,\rho\,\|\mathbf{v}\|^2+\frac{\kappa(2\,\rho)}{4\,\rho}\|\nabla\rho\|^2+W(2\,\rho)\,,\tag{41}$$

and that when $\mathscr{H}_\#=\mathscr{H}_{\mathrm{u}}$, one receives

$$H_{\mathrm{u}}[(\rho,\mathbf{v})]:=H_\#[(\rho,\mathbf{v})]=H_0[(\rho,\mathbf{v})]-\omega_\phi\rho+c_xQ_1(\rho,\mathbf{v})\,.$$

Turning to the identification of periodic traveling waves moving in the direction $\mathbf{e}_1$, we now restrict the spatial variable to dimension 1 and consider functions independent of time. We point out that $\mathcal{V}$ is a solution to

$$0=\delta\mathscr{H}_{\mathrm{u}}[\mathcal{V}]\,,\qquad\qquad \mu_\phi=\mathbf{J}\mathcal{V}\cdot\nabla_{\mathbf{U}_x}\mathscr{H}_{\mathrm{u}}[\mathcal{V}]\,,$$

bounded away from zero if and only if $\mathcal{V}=\mathcal{U}(\rho(\cdot),\theta(\cdot))$, with ρ bounded below away from zero, $v=\frac{\mathrm{d}\,\theta}{\mathrm{d}\,x}$, and $0=\delta H_{\mathrm{EK}}[(\rho,v)]$, where

$$H_{\mathrm{EK}}[(\rho,v)]:=H_{\mathrm{u}}[(\rho,v)]-\mu_\phi\,v=H_0[(\rho,v)]-\omega_\phi\rho-\mu_\phi\,v+c_xQ(\rho,v)\,.\tag{42}$$

Moreover, then with μ_x as in (26),

$$\mu_x=-H_{\mathrm{u}}[(\rho,v)]+v\,\mu_\phi+\rho_x\,\partial_{\rho_x}H_{\mathrm{u}}[(\rho,v)]\\ =\rho_x\,\partial_{\rho_x}H_{\mathrm{EK}}[(\rho,v)]-H_{\mathrm{EK}}[(\rho,v)]\,,$$

and

$$v=\nu(\rho;c_x,\mu_\phi):=\frac{\mu_\phi-c_x\,\rho}{2\,\rho\,\kappa(2\,\rho)}\,.\tag{43}$$

Furthermore, we stress that under these circumstances, there exists k_ϕ such that $x \mapsto \mathrm{e}^{-k_\phi\, x\, \mathbf{J}}\, \mathcal{V}(x)$ is periodic of period X_x if and only (ρ, v) is periodic of period X_x, and when this happens, k_ϕ is the average of v over one period.

With notation[8] from [11], we have untangled the correspondences in parameters

$$\begin{aligned} c_x &= c\,, & \mu_x &= \mu\,, & X_x &= \Xi\,,\\ \omega_\phi &= -\lambda_\rho\,, & \mu_\phi &= -\lambda_v\,, & k_\phi &= \langle v\rangle\,, \end{aligned}$$

in addition to the pointwise correspondences of mass, momentum, and Hamiltonian.

We also point out that we have recovered the reduction of profile equations to a two-dimensional Hamiltonian system associated with

$$\frac{\kappa(2\,\rho)}{4\,\rho}\left(\frac{\mathrm{d}\,\rho}{\mathrm{d}\,x}\right)^2 + \mathcal{W}_\rho(\rho; c_x, \omega_\phi, \mu_\phi) = \mu_x\,, \tag{44}$$

where

$$\begin{aligned} \mathcal{W}_\rho(\rho; c_x, \omega_\phi, \mu_\phi) := &-W(2\,\rho) - \kappa(2\,\rho)\,\rho\,(\nu(\rho))^2\\ &+ \omega_\phi\,\rho + \mu_\phi\,\nu(\rho) - c_x Q(\rho, \nu(\rho))\,, \end{aligned} \tag{45}$$

with $\nu(\rho) = \nu(\rho; c_x, \mu_\phi)$.

2.4. Action integral. — Motivated by the foregoing sections we introduce

$$\Theta(\mu_x, c_x, \omega_\phi, \mu_\phi) := \int_0^{X_x}\left(\mathscr{H}_0[\mathcal{V}] + c_x \mathbb{Q}[\mathcal{V}] - \omega_\phi\,\mathscr{M}[\mathcal{V}] - \mu_\phi\,\frac{\mathbb{Q}[\mathcal{V}]}{\mathscr{M}[\mathcal{V}]} + \mu_x\right)\mathrm{d}\,x\,, \tag{46}$$

with $(X_x, \mathcal{V})$ associated with $(\mu_x, c_x, \omega_\phi, \mu_\phi)$ as in Section 2.2. Note that Θ is, indeed, independent of $(\varphi_\phi, \varphi_x)$, and since

$$k_\phi\, X_x = \xi_\phi = \int_0^{X_x} \frac{\mathbb{Q}[\mathcal{V}]}{\mathscr{M}[\mathcal{V}]}\,\mathrm{d}\,x\,,$$

we also have

$$\Theta(\mu_x, c_x, \omega_\phi, \mu_\phi) = \int_0^{X_x}\left(\mathscr{H}_0[\mathcal{V}] + c_x \mathbb{Q}[\mathcal{V}] - \omega_\phi\,\mathscr{M}[\mathcal{V}] - \mu_\phi\, k_\phi + \mu_x\right)\mathrm{d}\,x\,,$$

with $(X_x, k_\phi, \mathcal{V})$ associated with $(\mu_x, c_x, \omega_\phi, \mu_\phi)$ as in Section 2.2.

Based on (44) we stress the following basic alternative formula

$$\Theta(\mu_x, c_x, \omega_\phi, \mu_\phi) = 2\int_{\rho_{\min}}^{\rho_{\max}} \sqrt{\mu_x - \mathcal{W}_\rho(\rho; c_x, \omega_\phi, \mu_\phi))}\sqrt{\frac{\kappa(2\,\rho)}{\rho}}\,\mathrm{d}\,\rho\,,$$

8. Except that (ρ, v) plays the role of (v, u) in [11].

where $\rho_{\min} = \rho_{\min}(\mu_x, c_x, \mu_\phi, \omega_\phi)$ and $\rho_{\max} = \rho_{\max}(\mu_x, c_x, \mu_\phi, \omega_\phi)$ are, respectively, the minimum and maximum values of $\mathscr{M}[\mathcal{V}]$. Note that $\rho_{\min}$ and $\rho_{\max}$ are (locally) characterized by

$$\mu_x = \mathcal{W}_\rho(\rho_{\min}; c_x, \omega_\phi, \mu_\phi)\,, \qquad \mu_x = \mathcal{W}_\rho(\rho_{\max}; c_x, \omega_\phi, \mu_\phi)\,.$$

A fundamental observation, intensively used in [12, 13, 9, 10, 11], is that

$$\begin{cases} \partial_{\mu_x}\Theta(\mu_x, c_x, \omega_\phi, \mu_\phi) = X_x\,, \\ \partial_{c_x}\Theta(\mu_x, c_x, \omega_\phi, \mu_\phi) = \displaystyle\int_0^{X_x} \mathbb{Q}[\mathcal{V}]\,\mathrm{d}\,x\,, \\ \partial_{\omega_\phi}\Theta(\mu_x, c_x, \omega_\phi, \mu_\phi) = -\displaystyle\int_0^{X_x} \mathscr{M}[\mathcal{V}]\,\mathrm{d}\,x\,, \\ \partial_{\mu_\phi}\Theta(\mu_x, c_x, \omega_\phi, \mu_\phi) = -\displaystyle\int_0^{X_x} \frac{\mathbb{Q}[\mathcal{V}]}{\mathscr{M}[\mathcal{V}]}\,\mathrm{d}\,x\,. \end{cases} \tag{47}$$

See, for instance, [12, Proposition 1] for a proof[9] of this elementary fact. For each of those, we also have

$$\partial_\#\Theta(\mu_x, c_x, \omega_\phi, \mu_\phi) = \int_{\rho_{\min}}^{\rho_{\max}} \frac{\partial_\#(\mu_x - \mathcal{W}_\rho)(\rho; c_x, \omega_\phi, \mu_\phi)}{\sqrt{\mu_x - \mathcal{W}_\rho(\rho; c_x, \omega_\phi, \mu_\phi))}} \sqrt{\frac{2\,\kappa(2\,\rho)}{2\,\rho}}\,\mathrm{d}\,\rho\,.$$

2.5. Asymptotic regimes. — As in [10, 11], we shall specialize most of the general results to two asymptotic regimes: small-amplitude and large-period asymptotics.

We make explicit here the descriptions of both regimes in terms of parameters. Let $(\underline{\rho}^{(0)}, \underline{k}_\phi^{(0)}) \in (0,\infty) \times \mathbf{R}$. Then, for any $\phi^{(0)} \in \mathbf{R}$,

$$\mathcal{V}^{(0)}(x) = \sqrt{2\underline{\rho}^{(0)}}\,\mathrm{e}^{(\phi^{(0)} + \underline{k}_\phi^{(0)}\,x)\mathbf{J}}(\mathbf{e}_1)$$

defines an unscaled profile with parameters $(\underline{\mu}_x^{(0)}, \underline{c}_x^{(0)}, \underline{\omega}_\phi^{(0)}, \underline{\mu}_\phi^{(0)})$ determined by (see (25),(26))

$$\begin{aligned} \underline{\mu}_\phi^{(0)} &= \underline{c}_x^{(0)}\underline{\rho}^{(0)} + \kappa(2\underline{\rho}^{(0)})\,2\,\underline{\rho}^{(0)}\,\underline{k}_\phi^{(0)}\,, \\ \underline{\omega}_\phi^{(0)} &= \underline{c}_x^{(0)}\underline{k}_\phi^{(0)} + \left(\kappa'(2\underline{\rho}^{(0)})\,2\,\underline{\rho}^{(0)} + \kappa(2\underline{\rho}^{(0)})\right)(\underline{k}_\phi^{(0)})^2 + 2W'(2\underline{\rho}^{(0)}) \\ \underline{\mu}_x^{(0)} &= -\frac{1}{2}\,\kappa(2\underline{\rho}^{(0)})\,2\underline{\rho}^{(0)}\,(\underline{k}_\phi^{(0)})^2 - W(2\underline{\rho}^{(0)}) - \underline{c}_x^{(0)}\underline{\rho}^{(0)}\underline{k}_\phi^{(0)} + \underline{\omega}_\phi^{(0)}\,\underline{\rho}^{(0)} + \underline{\mu}_\phi^{(0)}\underline{k}_\phi^{(0)}\,, \end{aligned}$$

except for $\underline{c}_x^{(0)} \in \mathbf{R}$, which may be chosen arbitrarily. Using ν, and $\mathcal{W}_\rho$ introduced in (43)–(45), the determination of parameters is equivalently written

9. Let us recall that in this reference, the role of (ρ, v) is played by (v, u). Note that the proof given there uses $\underline{\mathcal{V}} \cdot \underline{\mathcal{V}}_x(0) = 0$, but this may be assumed up to an harmless spatial translation.

as

$$\begin{aligned}\underline{k}_\phi^{(0)} &= \nu(\underline{\rho}^{(0)};\underline{c}_x^{(0)},\underline{\mu}_\phi^{(0)})\,,\\ 0 &= \partial_\rho \mathcal{W}_\rho(\underline{\rho}^{(0)};\underline{c}_x^{(0)},\underline{\omega}_\phi^{(0)},\underline{\mu}_\phi^{(0)})\,,\\ \underline{\mu}_x^{(0)} &= \mathcal{W}_\rho(\underline{\rho}^{(0)};\underline{c}_x^{(0)},\underline{\omega}_\phi^{(0)},\underline{\mu}_\phi^{(0)})\,.\end{aligned}$$

In this alternate formulation, it is clear that we could instead fix $(\underline{\rho}^{(0)},\underline{\mu}_\phi^{(0)},\underline{c}_x^{(0)}) \in (0,\infty)\times\mathbf{R}^2$ and determine $(\underline{k}_\phi^{(0)},\underline{\omega}_\phi^{(0)},\underline{\mu}_x^{(0)})$ correspondingly.

We are only interested in nondegenerate constant solutions and, thus, assume

$$\partial_\rho^2 \mathcal{W}_\rho(\underline{\rho}^{(0)};\underline{c}_x^{(0)},\underline{\omega}_\phi^{(0)},\underline{\mu}_\phi^{(0)}) \neq 0.$$

Under this condition, for any $(c_x^{(0)},\,\omega_\phi^{(0)},\,\mu_\phi^{(0)})$ in some neighborhood of $(\underline{c}_x^{(0)},\,\underline{\omega}_\phi^{(0)},\,\underline{\mu}_\phi^{(0)})$, there is a unique corresponding

$$(\rho^{(0)},k_\phi^{(0)},\mu_x^{(0)}) := (\rho^{(0)},k_\phi^{(0)},\mu_x^{(0)})(c_x^{(0)},\omega_\phi^{(0)},\mu_\phi^{(0)})\,,$$

in some neighborhood of $(\underline{\rho}^{(0)},\underline{k}_\phi^{(0)},\underline{\mu}_x^{(0)})$.

When

$$\partial_\rho^2 \mathcal{W}_\rho(\underline{\rho}^{(0)};\underline{c}_x^{(0)},\underline{\omega}_\phi^{(0)},\underline{\mu}_\phi^{(0)}) > 0\,,$$

to any $(c_x,\omega_\phi,\mu_\phi,\mu_x)$ sufficiently close to $(\underline{c}_x^{(0)},\underline{\omega}_\phi^{(0)},\underline{\mu}_\phi^{(0)},\underline{\mu}_x^{(0)})$ and satisfying

$$\mu_x > \mu_x^{(0)}(c_x,\omega_\phi,\mu_\phi)$$

corresponds a unique – up to rotational and spatial translations invariances – periodic traveling wave with mass close[10] to $\rho^{(0)}(c_x,\omega_\phi,\mu_\phi)$. The small-amplitude limit denotes the asymptotics $\mu_x-\mu_x^{(0)}(c_x,\omega_\phi,\mu_\phi)\to 0$, and the *small-amplitude regime* is the zone where $\mu_x-\mu_x^{(0)}(c_x,\omega_\phi,\mu_\phi)$ is small but positive. Incidentally, we point out that the limiting small amplitude period is given by

$$X_x^{(0)}(c_x,\omega_\phi,\mu_\phi) := 2\pi\sqrt{\frac{\kappa(2\,\rho^{(0)})}{2\rho^{(0)}\partial_\rho^2\mathcal{W}_\rho(\rho^{(0)};c_x,\omega_\phi,\mu_\phi)}}\,, \tag{48}$$

with $\rho^{(0)}=\rho^{(0)}(c_x,\omega_\phi,\mu_\phi)$.

When

$$\partial_\rho^2 \mathcal{W}_\rho(\underline{\rho}^{(0)};\underline{c}_x^{(0)},\underline{\omega}_\phi^{(0)},\underline{\mu}_\phi^{(0)}) < 0\,,$$

there are at most two solitary wave profiles with parameters $(\underline{c}_x^{(0)},\underline{\omega}_\phi^{(0)},\underline{\mu}_\phi^{(0)})$, namely at most one with $\underline{\rho}^{(0)}$ as both an infimum and an end state for its mass

10. Recall that there could be various branches corresponding to the same parameters. We give this precision to exclude other branches; see Figure 2.1.

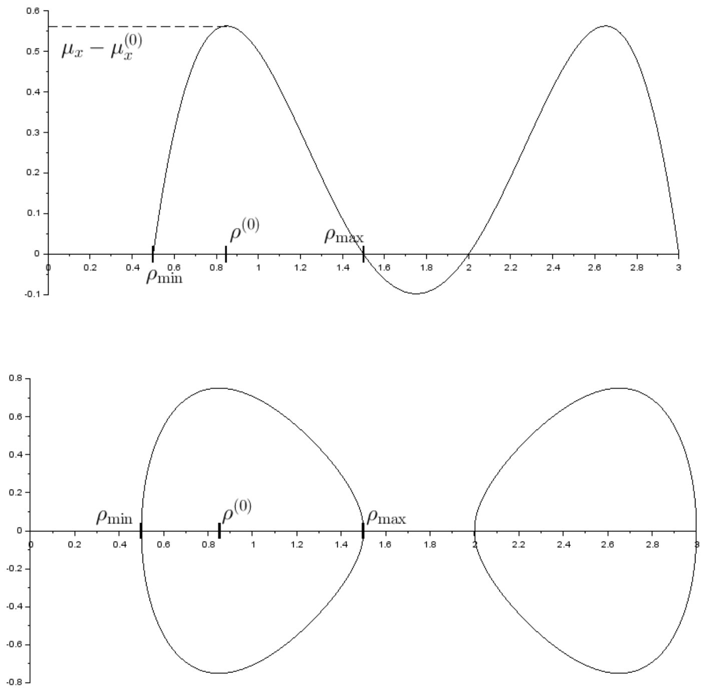

FIGURE 2.1. **Small-amplitude limit** with two branches with the same parameters. The upper graph is the graph of $\mu_x - \mathcal{W}_\rho$ as a function of ρ. The lower graph is, in the $(\rho, \frac{\mathrm{d}\,\rho}{\mathrm{d}\,x})$ phase plane, the level set defined by (44). The two closed curves correspond to two periodic waves, the curve of interest being the one circling $\rho^{(0)}$.

and at most one with $\underline{\rho}^{(0)}$ as both a supremum and an end state for its mass; see Figure 2.2. Concerning the large-period regime, we restrict ourselves to the case when the periodic-wave profile asymptotes a single solitary-wave profile and leave aside the case[11] when the periodic wave profile is asymptotically obtained by gluing two pieces of distinct solitary wave profiles sharing the same end state. From now on, we focus on the case where $\underline{\rho}^{(0)}$ is an infimum. Note that when there are two solitary waves with the same endstate/parameters,

11. There are yet more possibilities (involving fronts/kinks in addition to solitary waves), but they may be thought of as degenerate in the sense that they form a manifold of a smaller dimension. The two-bump case is nondegenerate but was left aside in [10] as *a priori* significantly different from the single-bump case dealt with here and there.

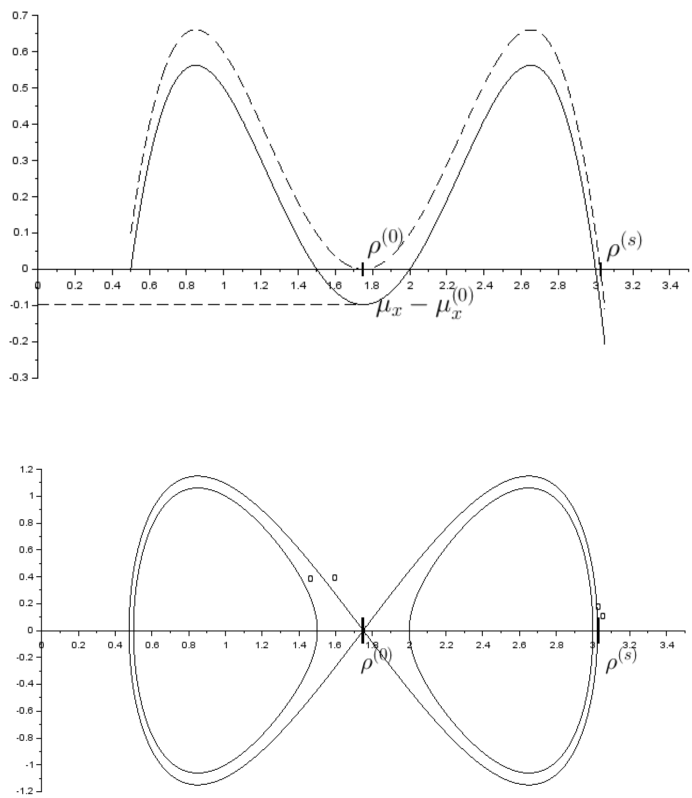

FIGURE 2.2. **Solitary-wave limit** with two branches with the same parameters. The upper graph is the graph of $\mu_x - \mathcal{W}_\rho$ as a function of ρ. The lower graph is, in the $(\rho, \frac{\mathrm{d}\,\rho}{\mathrm{d}\,x})$ phase plane, the level set defined by (44). In both graphs, we superimpose images corresponding to parameters of the solitary-wave limit and nearby parameters corresponding to periodic waves of a large period. The curves of interest are the right-hand ones.

they generate distinct branches of (single-bump) periodic waves, thus, may be analyzed independently. Moreover, we point out that the related analysis of the supremum case is completely analogous. The existence of a solitary wave of such a type is equivalent to the existence of $\underline{\rho}^{(s)} > \underline{\rho}^{(0)}$, such that

$$\mathcal{W}_\rho(\underline{\rho}^{(s)}; \underline{c}_x^{(0)}, \underline{\omega}_\phi^{(0)}, \underline{\mu}_\phi^{(0)}) = \underline{\mu}_x^{(0)}\,, \qquad \partial_\rho \mathcal{W}_\rho(\underline{\rho}^{(s)}; \underline{c}_x^{(0)}, \underline{\omega}_\phi^{(0)}, \underline{\mu}_\phi^{(0)}) > 0\,,$$

and

$$\forall \rho \in (\underline{\rho}^{(0)}, \underline{\rho}^{(s)}), \qquad \mathcal{W}_\rho(\rho; \underline{c}_x^{(0)}, \underline{\omega}_\phi^{(0)}, \underline{\mu}_\phi^{(0)}) \neq \underline{\mu}_x^{(0)},$$

where $(\underline{\rho}^{(0)}, \underline{\mu}_x^{(0)}) = (\rho^{(0)}, \mu_x^{(0)})(\underline{c}_x^{(0)}, \underline{\omega}_\phi^{(0)}, \underline{\mu}_\phi^{(0)})$. The situation is stable by perturbation of parameters $(\underline{c}_x^{(0)}, \underline{\omega}_\phi^{(0)}, \underline{\mu}_\phi^{(0)})$. Assuming the latter, one deduces that to any $(\mu_x, c_x, \omega_\phi, \mu_\phi)$ sufficiently close to $(\underline{\mu}_x^{(0)}, \underline{c}_x^{(0)}, \underline{\omega}_\phi^{(0)}, \underline{\mu}_\phi^{(0)})$ and satisfying

$$\mu_x < \mu_x^{(0)}(c_x, \omega_\phi, \mu_\phi),$$

corresponds a unique – up to rotational and spatial translations invariances – periodic traveling wave with mass average and mass minimum close to $\rho^{(0)}(c_x, \omega_\phi, \mu_\phi)$. The large period limit denotes the asymptotics $\mu_x - \mu_x^{(0)}(c_x, \omega_\phi, \mu_\phi) \to 0$ and the *large period regime* is the zone where $\mu_x - \mu_x^{(0)}(c_x, \omega_\phi, \mu_\phi)$ is sufficiently small but negative.

From the point of view of the solitary waves themselves, it is actually both more natural and more convenient to keep a parametrization by $(c_x, \rho_{(0)}, k_\phi)$ rather than by $(c_x, \omega_\phi, \mu_\phi)$, with $\rho_{(0)}$ the end state. This is consistent with the fact that variations in the end state (thus in (ρ, k_ϕ)) play no role in the classical stability analysis of solitary waves (under localized perturbations). Assuming as above that there is a $\underline{\rho}^{(s)}$ associated with $(\underline{c}_x^{(0)}, \underline{\rho}_{(0)}, \underline{k}_\phi^{(0)})$, one deduces that for any $(c_x, \rho_{(0)}, k_\phi)$ sufficiently close to $(\underline{c}_x^{(0)}, \underline{\rho}^{(0)}, \underline{k}_\phi^{(0)})$, there exists $\rho^{(s)} = \rho^{(s)}(c_x, \rho_{(0)}, k_\phi)$ close to $\underline{\rho}^{(s)}$ such that

$$\mathcal{W}_\rho(\rho^{(s)}; c_x, \omega_\phi^{(0)}, \mu_\phi^{(0)}) = (\mu_x)_{(0)}, \qquad \partial_\rho \mathcal{W}_\rho(\rho^{(s)}; c_x, \omega_\phi^{(0)}, \mu_\phi^{(0)}) > 0,$$

and

$$\forall \rho \in (\rho_{(0)}, \rho^{(s)}), \qquad \mathcal{W}_\rho(\rho; c_x, \omega_\phi^{(0)}, \mu_\phi^{(0)}) \neq (\mu_x)_{(0)},$$

where $(\omega_\phi^{(0)}, \mu_\phi^{(0)}) = (\omega_\phi^{(0)}, \mu_\phi^{(0)})(c_x, \rho^{(0)}, k_\phi)$ is defined implicitly by

$$(\rho_{(0)}, k_\phi) = (\rho^{(0)}, k_\phi^{(0)})(c_x, \omega_\phi^{(0)}, \mu_\phi^{(0)}),$$

and $(\mu_x)_{(0)} = (\mu_x)_{(0)}(c_x, \rho_{(0)}, k_\phi) := \mu_x^{(0)}(c_x, \omega_\phi^{(0)}(c_x, \rho_{(0)}, k_\phi), \mu_\phi^{(0)}(c_x, \rho_{(0)}, k_\phi))$. The mass of the corresponding solitary-wave profile $\rho_{(s)} = \rho_{(s)}(\,\cdot\,; c_x, \rho_{(0)}, k_\phi)$ is then obtained by solving

$$\begin{aligned}\frac{\kappa(2\,\rho_{(s)})}{2\,\rho_{(s)}} \frac{\mathrm{d}^2\,\rho_{(s)}}{\mathrm{d}\,x^2} = -\left(\frac{\kappa'(2\,\rho_{(s)})}{2\,\rho_{(s)}} - \frac{\kappa(2\,\rho_{(s)})}{4\,\rho_{(s)}^2}\right)\left(\frac{\mathrm{d}\,\rho_{(s)}}{\mathrm{d}\,x}\right)^2 \\ - \partial_\rho \mathcal{W}_\rho(\rho_{(s)}; c_x, \omega_\phi^{(0)}, \mu_\phi^{(0)}),\end{aligned}$$

with $\rho_{(s)}(\,0\,;c_x,\rho_{(0)},k_\phi)=\rho^{(s)}(c_x,\rho_{(0)},k_\phi)$. Then the unscaled profile $\mathcal{V}_{(s)}=\mathcal{V}_{(s)}(\,\cdot\,;c_x,\rho_{(0)},k_\phi)$ is obtained through[12]

$$\begin{aligned}\mathcal{V}_{(s)}&=\sqrt{2\,\rho_{(s)}}\;\mathrm{e}^{\theta_{(s)}\,\mathbf{J}}(\mathbf{e}_1)\,,\\ \theta_{(s)}(x)&=\int_0^x\nu(\rho_{(s)}(\,\cdot\,;c_x,\rho_{(0)},k_\phi);c_x,\mu_\phi^{(0)}(c_x,\rho,k_\phi))\,.\end{aligned}$$

Stability conditions are expressed in terms of

$$\begin{aligned}\Theta_{(s)}(c_x,\rho,k_\phi):=\int_{-\infty}^{\infty}\Bigg(&\mathscr{H}_0[\mathcal{V}_{(s)}]+c_x\mathbb{Q}[\mathcal{V}_{(s)}]-\omega_\phi^{(0)}\,\mathscr{M}[\mathcal{V}_{(s)}]\\ &-\mu_\phi^{(0)}\,\frac{\mathbb{Q}[\mathcal{V}_{(s)}]}{\mathscr{M}[\mathcal{V}_{(s)}]}+(\mu_x)_{(0)}\Bigg)\,.\end{aligned}\tag{49}$$

Concerning the small-amplitude limit, although this is less crucial, at some point it will also be convenient to adopt a parametrization of limiting harmonic wave trains by $(k_x,\rho_{(0)},k_\phi)$ (rather than by $(c_x,\omega_\phi,\mu_\phi)$). Our starting point was a parametrization by $(c_x,\rho_{(0)},k_\phi)$, so that we only need to examine the invertibility of the relation $c_x\mapsto 1/X_x^{(0)}$ at fixed (ρ,k_ϕ). The equation to invert is

$$\partial_\rho^2\mathcal{W}_\rho(\rho;c_x,\omega_\phi,\mu_\phi)=\frac{1}{2}\frac{\kappa(2\,\rho)}{2\rho}\,(2\pi\,k_x)^2\,,$$

with (ω_ϕ,μ_ϕ) associated with (c_x,ρ,k_ϕ) through

$$k_\phi=\nu(\rho;c_x,\underline{\mu}_\phi)\,,\qquad\qquad 0=\partial_\rho\mathcal{W}_\rho(\rho;c_x,\omega_\phi,\mu_\phi)\,.\tag{50}$$

Straightforward computations detailed in [11, Appendix A] show that

$$\begin{aligned}\partial_\rho^2\mathcal{W}_\rho(\rho;c_x,\omega_\phi,\mu_\phi)&=\frac{1}{\kappa(2\,\rho)\,2\rho}\,\det(\mathbf{B}\,\mathrm{Hess}\,H^{(0)}(\rho,k_\phi)+c_x\,\mathrm{I}_2)\\ -2\,\kappa(2\,\rho)\,2\rho\,\partial_\rho\nu(\rho;c_x,\mu_\phi)&=2c_x+\mathrm{Tr}(\mathbf{B}\,\mathrm{Hess}\,H^{(0)}(\rho,k_\phi))\,,\end{aligned}$$

where

$$\mathbf{B}:=\begin{pmatrix}0&1\\1&0\end{pmatrix},\qquad H^{(0)}(\rho,v):=\kappa(2\,\rho)\rho\,v^2+W(2\,\rho)\,.\tag{51}$$

Let us stress incidentally that $H^{(0)}$ is the zero dispersion limit of the Hamiltonian H_0 of the hydrodynamic formulation of the Schrödinger equation, and $\mathbf{B}$ is the self-adjoint matrix involved in this formulation. As a result, if $\partial_\rho\nu(\underline{\rho}^{(0)};\underline{c}_x^{(0)},\underline{\mu}_\phi^{(0)})\neq 0$, then locally one may, indeed, parametrize waves by (k_x,ρ,k_ϕ), and we shall denote

$$c_x=c_x^{(0)}(k_x,\rho,k_\phi)\,,\qquad\qquad \omega_x^{(0)}(k_x,\rho,k_\phi):=-k_x\,c_x^{(0)}(k_x,\rho,k_\phi)\,,$$

the corresponding harmonic phase speed and associated spatial time frequency.

12. The choice of the point where the value $\rho^{(s)}(c_x,\rho,k_\phi)$ is achieved (respectively of $\theta_{(s)}(0;c_x,\rho,k_\phi)$), quotients the invariance by spatial (respectively rotational) translation.

2.6. General plane waves. — We now explain how to extend the foregoing analysis to more general plane waves in the form (15). So far, we have discussed explicitly the case when $\mathbf{k}_\phi$ and $\mathbf{k}_x$ point in the direction of $\mathbf{e}_1$. The main task is to show how to reduce to the case when $\mathbf{k}_\phi$ and $\mathbf{k}_x$ are colinear, that is, when $\widetilde{\mathbf{k}}_\phi = 0$.

Let us first observe that for any vector $\widetilde{\mathbf{k}}_\phi$, the frame change

$$\mathbf{U}(t,\mathbf{x}) = \mathrm{e}^{\widetilde{\mathbf{k}}_\phi\cdot\mathbf{x}\,\mathbf{J}}\;\widetilde{\mathbf{U}}(t,\mathbf{x})$$

changes (3) into

$$\partial_t\widetilde{\mathbf{U}} = \mathbf{J}\delta\mathscr{H}_{\widetilde{\mathbf{k}}_\phi}[\widetilde{\mathbf{U}}]\,,$$

$$\begin{aligned}(52)\quad \mathscr{H}_{\widetilde{\mathbf{k}}_\phi}[\widetilde{\mathbf{U}}] &:= \mathscr{H}_0(\widetilde{\mathbf{U}},(\nabla_{\mathbf{x}}+\widetilde{\mathbf{k}}_\phi\mathbf{J})\widetilde{\mathbf{U}})\\ &= \mathscr{H}_0(\widetilde{\mathbf{U}},\nabla_{\mathbf{x}}\widetilde{\mathbf{U}}) + \frac{1}{2}\,\|\widetilde{\mathbf{U}}\|^2\,\kappa(\|\widetilde{\mathbf{U}}\|^2)\,\|\widetilde{\mathbf{k}}_\phi\|^2 + \kappa(\|\widetilde{\mathbf{U}}\|^2)\,\widetilde{\mathbf{k}}_\phi\cdot\,\mathbb{Q}[\widetilde{\mathbf{U}}]\,.\end{aligned}$$

As a consequence, if one is simply interested in analyzing the structure of waves or the behavior of solutions arising from longitudinal perturbations or more generally from perturbations depending only on directions orthogonal to $\widetilde{\mathbf{k}}_\phi$, it is sufficient to fix $\mathbf{e}_x$ and $\widetilde{\mathbf{k}}_\phi$ (orthogonal to each other) and replace $\mathscr{H}_0$ with $\mathscr{H}_{0,\widetilde{\mathbf{k}}_\phi}$ defined by

$$\mathscr{H}_{0,\widetilde{\mathbf{k}}_\phi}[\widetilde{\mathbf{U}}] := \mathscr{H}_0(\widetilde{\mathbf{U}},\nabla_{\mathbf{x}}\widetilde{\mathbf{U}}) + \frac{1}{2}\,\|\widetilde{\mathbf{U}}\|^2\,\kappa(\|\widetilde{\mathbf{U}}\|^2)\,\|\widetilde{\mathbf{k}}_\phi\|^2\,,$$

or equivalently to replace W with $W_{\widetilde{\mathbf{k}}_\phi}$ defined through

$$W_{\widetilde{\mathbf{k}}_\phi}(\alpha) := W(\alpha) + \frac{1}{2}\,\alpha\,\kappa(\alpha)\,\|\widetilde{\mathbf{k}}_\phi\|^2\,.$$

With this point of view, all quantities manipulated in previous subsections of the present section should be thought of as implicitly depending on $\mathbf{e}_x$ and $\widetilde{\mathbf{k}}_\phi$. Note, however, that actually they do not depend on $\mathbf{e}_x$ and depend on $\widetilde{\mathbf{k}}_\phi$ only through $\|\widetilde{\mathbf{k}}_\phi\|^2$. In particular, their first-order derivatives with respect to $\widetilde{\mathbf{k}}_\phi$ vanish at $\widetilde{\mathbf{k}}_\phi = 0$.

To prepare the analysis of stability under general perturbations, let us make explicit the relations defining constants of integration and averaged quantities for general plane waves taken in the form

$$\mathbf{U}(t,\mathbf{x}) = \mathrm{e}^{\left(\widetilde{\mathbf{k}}_\phi\cdot(\mathbf{x}-\underline{c}_x\,\underline{\mathbf{e}}_x\,t)+\underline{\omega}_\phi\,t\right)\mathbf{J}}\;\mathcal{V}(\mathbf{e}_x\,\cdot(\mathbf{x}-\underline{c}_x\;\mathbf{e}_x\;t))\,,$$

generalizing (21). We still have

$$\mu_\phi = \mathcal{S}_\phi[\mathcal{V}] + c_x\mathscr{M}[\mathcal{V}]\,,$$

but $\mathcal{S}_\phi$ should be taken as

$$\mathcal{S}_\phi[\mathcal{V}] = \mathbf{J}\mathcal{V}\cdot(\mathbf{e}_x\cdot\nabla_{\mathbf{U}_\mathbf{x}})\,\mathscr{H}_{0,\widetilde{\mathbf{k}}_\phi}(\mathcal{V},\mathbf{e}_x\;\partial_\zeta\mathcal{V})\,.$$

Likewise

$$\mu_x = \mathcal{S}_x[\mathcal{V}] + \omega_\phi \mathscr{M}[\mathcal{V}]\,,$$

with

$$\mathcal{S}_x[\mathbf{V}] = -\mathscr{H}_{0,\widetilde{\mathbf{k}}_\phi}(\mathcal{V}, \mathbf{e}_x\ \partial_\zeta \mathcal{V}) + \partial_\zeta \mathcal{V}\ \cdot (\mathbf{e}_x \cdot \nabla_{\mathbf{U_x}})\, \mathscr{H}_{0,\widetilde{\mathbf{k}}_\phi}(\mathcal{V}, \mathbf{e}_x\ \partial_\zeta \mathcal{V})\,.$$

This implies that

$$\begin{aligned}
\mathbb{Q}[\mathbf{U}] &= \mathbf{e}_x \cdot \mathbb{Q}(\mathcal{V}, \mathbf{e}_x\ \partial_\zeta \mathcal{V})\ \mathbf{e}_x + \mathscr{M}[\mathcal{V}]\, \widetilde{\mathbf{k}}_\phi\,,\\
\mathbf{J}\mathbf{U} \cdot \nabla_{\mathbf{U_x}} \mathscr{H}_0[\mathbf{U}] &= \mathbf{e}_x\ (\mu_\phi - c_x \mathscr{M}[\mathcal{V}]) + \widetilde{\mathbf{k}}_\phi\, \kappa(\|\mathcal{V}\|^2)\, 2\, \mathscr{M}[\mathcal{V}]\,,\\
\frac{1}{2}\mathbf{J}\mathbf{U} \cdot \mathbf{J}\delta\mathscr{H}_0[\mathbf{U}] - \mathscr{H}_0[\mathbf{U}] &= \mu_x - c_x\, \mathbf{e}_x \cdot \mathbb{Q}(\mathcal{V}, \mathbf{e}_x\ \partial_\zeta \mathcal{V})\\
&\quad - \kappa(\|\mathcal{V}\|^2)\, \|\partial_\zeta \mathcal{V}\|^2 + \kappa'(\|\mathcal{V}\|^2)\, \mathscr{M}[\mathcal{V}]\, \|\widetilde{\mathbf{k}}_\phi\|^2\,,\\
\mathbf{J}\delta\, \mathbb{Q}_j[\mathbf{U}] \cdot \nabla_{\mathbf{U}_{x_\ell}} \mathscr{H}_0[\mathbf{U}] &= \kappa(\|\mathcal{V}\|^2)\ \Big((\mathbf{e}_x)_j\, \partial_\zeta \mathcal{V} + (\widetilde{\mathbf{k}}_\phi)_j\, \mathbf{J}\mathcal{V}\Big)\\
&\quad \cdot \Big((\mathbf{e}_x)_\ell\, \partial_\zeta \mathcal{V} + (\widetilde{\mathbf{k}}_\phi)_\ell\, \mathbf{J}\mathcal{V}\Big)\,,
\end{aligned}$$

with right-hand terms evaluated at $(\mathbf{e}_x \cdot (\mathbf{x} - \underline{c}_x\ \mathbf{e}_x\ t))$.

3. Structure of the spectrum

Now we turn to gathering key facts about the spectrum of operators arising from linearization, in suitable frames, about periodic plane waves.

3.1. The Bloch transform. — Our first observation is that, due to a suitable integral transform, the spectrum of the linearized operator $\mathcal{L}$ defined in (9) may be studied through normal-mode analysis.

To begin with, we introduce a suitable Fourier–Bloch transform, as a mix of a Floquet/Bloch transform in the x variable and the Fourier transform in the $\mathbf{y}$ variable:

$$\mathcal{B}(g)(\xi, x, \boldsymbol{\eta}) = \check{g}(\xi, x, \boldsymbol{\eta}) := \sum_{j \in \mathbf{Z}} \mathrm{e}^{\mathrm{i}\ 2j\pi x}\ \widehat{g}(\xi + 2j\pi, \boldsymbol{\eta})\,, \tag{53}$$

where $\widehat{g}$ is the usual Fourier transform normalized so that for $\mathbf{x} = (x, \mathbf{y})$

$$\begin{aligned}
\mathcal{F}(g)(\xi, \boldsymbol{\eta}) = \hat{g}(\xi, \boldsymbol{\eta}) &:= \frac{1}{(2\pi)^d} \int_{\mathbf{R}} \mathrm{e}^{-\mathrm{i}\,\xi x - \mathrm{i}\,\boldsymbol{\eta}\cdot\mathbf{y}}\, g(\mathbf{x})\ \mathrm{d}\,\mathbf{x},\\
g(\mathbf{x}) &= \int_{\mathbf{R}} \mathrm{e}^{\mathrm{i}\,\xi x + \mathrm{i}\,\boldsymbol{\eta}\cdot\mathbf{y}}\ \hat{g}(\xi, \boldsymbol{\eta})\ \ \mathrm{d}\,\xi\ \mathrm{d}\,\boldsymbol{\eta}\,.
\end{aligned}$$

Obviously, $\check{g}(\xi, \cdot, \boldsymbol{\eta})$ is periodic of period 1 for any $(\xi, \boldsymbol{\eta})$, that is,

$$\forall x \in \mathbf{R}, \qquad \check{g}(\xi, x+1, \boldsymbol{\eta}) = \check{g}(\xi, x, \boldsymbol{\eta})\,.$$

As follows readily from (53) and basic Fourier theory, $(2\pi)^{d/2}\,\mathcal{B}$ is a total isometry from $L^2(\mathbf{R}^d)$ to $L^2((-\pi,\pi)\times(0,1)\times\mathbf{R}^{d-1})$, and it satisfies the inversion formula

$$g(\mathbf{x}) = \int_{-\pi}^{\pi}\int_{\mathbf{R}^{d-1}} \mathrm{e}^{\mathrm{i}\,\xi x+\mathrm{i}\,\boldsymbol{\eta}\cdot\mathbf{y}}\;\check{g}(\xi,x,\boldsymbol{\eta})\,\mathrm{d}\,\xi\,\mathrm{d}\,\boldsymbol{\eta}\,. \tag{54}$$

The Poisson summation formula provides an alternative equivalent formula for (53)

$$\check{g}(\xi,x,\boldsymbol{\eta}) = \frac{1}{2\pi}\sum_{\ell\in\mathbf{Z}} \mathrm{e}^{-\,\mathrm{i}\,\xi(x+\ell)}\,\mathcal{F}_{\mathbf{y}}(g)(x+\ell,\boldsymbol{\eta})\,,$$

where $\mathcal{F}_{\mathbf{y}}$ denotes the Fourier transform in the $\mathbf{y}$-variable only.

The key feature of the transform $\mathcal{B}$ is that in some sense it diagonalizes differential operators whose coefficients do not depend on $\mathbf{y}$ and are 1-periodic in x. For large classes of such operators $\mathcal{P} = P(x,\partial_x,\nabla_{\mathbf{y}})$, stands

$$\mathcal{B}(\mathcal{P}u)(\xi,x,\boldsymbol{\eta}) = P(x,\partial_x+\mathrm{i}\,\xi,\mathrm{i}\,\boldsymbol{\eta})\,\mathcal{B}(u)(\xi,x,\boldsymbol{\eta})\,,$$

so that the action of such operators on functions defined on $\mathbf{R}^d$ is reduced to the action of $\mathcal{P}_{\xi,\boldsymbol{\eta}} = P(x,\partial_x+\mathrm{i}\,\xi,\mathrm{i}\,\boldsymbol{\eta})$ on 1-periodic functions, parametrized by $(\xi,\boldsymbol{\eta})$.

In particular, for $\mathcal{L}$ as in (9) we have

$$(\mathcal{L}g)(\mathbf{x}) = \int_{-\pi}^{\pi}\int_{\mathbf{R}^{d-1}} \mathrm{e}^{\mathrm{i}\,\xi\mathbf{x}+\mathrm{i}\,\boldsymbol{\eta}\cdot\mathbf{y}}\,(\mathcal{L}_{\xi,\boldsymbol{\eta}}\check{g}(\xi,\,\cdot\,,\boldsymbol{\eta}))(x)\,,\,\mathrm{d}\,\xi\,\mathrm{d}\,\boldsymbol{\eta},$$

where $\mathcal{L}_{\xi,\boldsymbol{\eta}}$ acts on 1-periodic functions and from $\mathscr{H} = \mathscr{H}^x + \mathscr{H}^{\mathbf{y}}$ inherits the splitting

$$\mathcal{L}_{\xi,\boldsymbol{\eta}} := \mathcal{L}^x_{\xi} + \mathcal{L}^{\mathbf{y}}_{\boldsymbol{\eta}}\,,\qquad\text{with}\quad \mathcal{L}^{\mathbf{y}}_{\boldsymbol{\eta}} := \|\boldsymbol{\eta}\|^2\,\kappa(\|\mathbf{U}\|^2)\,\mathbf{J}\,,$$

and $\mathcal{L}^x_{\xi}$ given by

$$\begin{aligned}\mathcal{L}^x_{\xi}\mathbf{V} = \mathbf{J}\,\Big(&\,\mathrm{d}_{(\mathbf{U},\mathbf{U}_x)}(\nabla_{\mathbf{U}}\mathscr{H})(\underline{\mathbf{U}},\nabla_{\mathbf{x}}\underline{\mathbf{U}})(\mathbf{V},(\partial_x+\mathrm{i}\,\xi)\mathbf{V})\\ &-(\partial_x+\mathrm{i}\,\xi)\,\big(\mathrm{d}_{(\mathbf{U},\mathbf{U}_x)}(\nabla_{\mathbf{U}_x}\mathscr{H})(\underline{\mathbf{U}},\nabla_{\mathbf{x}}\underline{\mathbf{U}})(\mathbf{V},(\partial_x+\mathrm{i}\,\xi)\mathbf{V})\big)\Big)\,.\end{aligned}$$

On[13] $L^2_{\mathrm{per}}((0,1))$ each $\mathcal{L}_{\xi,\boldsymbol{\eta}}$ has a compact resolvent and depends analytically on $(\xi,\boldsymbol{\eta})$ in the strong resolvent sense.

It is both classical and relatively straightforward to derive from the latter and the isometry of $(2\pi)^{d/2}\,\mathcal{B}$ that the spectrum of $\mathcal{L}$ on $L^2(\mathbb{R})$ coincides with the union over $(\xi,\boldsymbol{\eta})\in[-\pi,\pi]\times\mathbf{R}^{d-1}$ of the spectrum of each $\mathcal{L}_{\xi,\boldsymbol{\eta}}$ on $L^2_{\mathrm{per}}((0,1))$. The reader is referred to [48, p.30-31] for more details on the argument, to [39]

13. We insist on the subscript $_{\mathrm{per}}$ to emphasize that corresponding domains involve $H^s_{\mathrm{per}}((0,1))$ spaces, thus effectively encoding periodic boundary conditions when $s>1/2$. Notation "$\mathcal{L}^{\mathrm{per}}_{\xi,\boldsymbol{\eta}}$ acts on $L^2((0,1))$" would be mathematically more accurate but more cumbersome.

for a general background on the Bloch transform, and to [48, 49] for comments on its use in periodic-wave stability problems.

3.2. Linearizing Madelung's transformation. — We would like to point out here how the analysis of Section 2.3 may be extended to the spectral level. We stress that working with Bloch–Fourier symbols $\mathcal{L}_{\xi,\boldsymbol{\eta}}$ provides crucial simplifications in the arguments.

Firstly, we observe that linearizing (36)–(37) provides all the necessary algebraic identities. Secondly, we note that applying a Bloch–Fourier transform to both sides of the foregoing identities yields the required algebraic conjugations between the respective Bloch–Fourier symbols.

To go beyond algebraic relations, we start with a few notational or elementary considerations.

1. From elementary elliptic regularity arguments, it follows that the $L^2_{\rm per}$-spectrum of each $\mathcal{L}_{\xi,\boldsymbol{\eta}}$ coincides with its $H^1_{\rm per}$-spectrum.
2. With $L^2_{{\rm curl}_{\xi,\boldsymbol{\eta}}}((0,1))$ denoting the space of $L^2((0,1);\mathbf{C})^d$-functions $\mathbf{v}$, such that[14]
$$\begin{pmatrix}(\partial_x+\mathrm{i}\,\xi)\\ \mathrm{i}\,\boldsymbol{\eta}\end{pmatrix}\wedge\mathbf{v}=0\,,$$
we observe that when $(\xi,\boldsymbol{\eta})\in[-\pi,\pi]\times\mathbf{R}^{d-1}\setminus\{(0,0)\}$,
$$\mathcal{I}_{\xi,\boldsymbol{\eta}}\ :\quad H^1_{\rm per}((0,1))\longrightarrow L^2_{{\rm curl}_{\xi,\boldsymbol{\eta}}}((0,1))\,,\qquad \theta\mapsto\begin{pmatrix}(\partial_x+\mathrm{i}\,\xi)\\ \mathrm{i}\,\boldsymbol{\eta}\end{pmatrix}\theta$$
is a bounded invertible operator.
3. The linearization of the relation
$$\mathbf{U}=\mathrm{e}^{-\frac{k_\phi}{k_x}(\cdot)\mathbf{J}}\,\mathcal{U}(\rho(\cdot),\theta(\cdot))\,,$$
at $\underline{\mathcal{U}}$, $(\underline{\rho},\underline{\theta})$, is given by
$$\mathfrak{m}\ :\quad H^1_{\rm per}((0,1);\mathbf{C}^2)\longrightarrow H^1_{\rm per}((0,1);\mathbf{C})^2\,,\quad \mathbf{V}\mapsto\left(\underline{\mathcal{U}}\cdot\mathbf{V},\,\frac{\mathbf{J}\underline{\mathcal{U}}}{2\underline{\rho}}\cdot\mathbf{V}\right)$$
and is bounded and invertible with inverse
$$\mathfrak{m}^{-1}\ :\quad H^1_{\rm per}((0,1);\mathbf{C})^2\longrightarrow H^1_{\rm per}((0,1);\mathbf{C}^2)\,,\quad (\rho,\theta)\mapsto\rho\,\frac{\underline{\mathcal{U}}}{2\,\underline{\rho}}+\theta\,\mathbf{J}\underline{\mathcal{U}}\,.$$

Considering $\mathcal{L}_{\xi,\boldsymbol{\eta}}$ as an operator on $H^1_{\rm per}$ and denoting $L_{\xi,\boldsymbol{\eta}}$ the corresponding Bloch–Fourier symbol for the associated Euler–Korteweg system (39), we

14. The condition means: $(\partial_x+\mathrm{i}\,\xi)v_j=\mathrm{i}\,\eta_{j-1}v_1$ for $2\leq j\leq d$, and $\eta_j v_\ell=\eta_\ell v_j$ for $1\leq j,\,\ell\leq d-1$. When $d=1$, $L^2_{{\rm curl}_{\xi,\boldsymbol{\eta}}}((0,1))=L^2((0,1);\mathbf{C})$.

deduce when $(\xi,\boldsymbol{\eta})\in[-\pi,\pi]\times\mathbf{R}^{d-1}\setminus\{(0,0)\}$, the conjugation

$$\mathcal{L}_{\xi,\boldsymbol{\eta}}=\mathfrak{m}^{-1}\begin{pmatrix}\mathrm{I}&0\\0&\mathcal{I}_{\xi,\boldsymbol{\eta}}^{-1}\end{pmatrix}L_{\xi,\boldsymbol{\eta}}\begin{pmatrix}\mathrm{I}&0\\0&\mathcal{I}_{\xi,\boldsymbol{\eta}}\end{pmatrix}\mathfrak{m}\,.$$

Of course, the conjugation yields the identity of spectra, including algebraic multiplicities but also the identity of the detailed algebraic structure of each eigenvalue. When $d=1$, by continuity of the eigenvalues with respect to ξ, one also concludes that $\mathcal{L}_0$ and L_0 share the same spectrum, including algebraic multiplicities, but algebraic structures may differ (and as stressed below, in general, they do). Note that to go from spectral to linear stability it is actually crucial to examine semisimplicity of eigenvalues.

For this reason, we focus now a bit more on the case $(\xi,\boldsymbol{\eta})=(0,0)$. To begin with, denoting $L^2_0((0,1))$ the space of $L^2((0,1);\mathbf{C})^d$-functions $\mathbf{v}$ of the form

$$\begin{pmatrix}v\\0\end{pmatrix},\quad\text{with }\int_0^1 v=0\,,$$

we observe that

$$\mathcal{I}^{(0)}\ :\quad H^1_{\mathrm{per}}((0,1))\longrightarrow L^2_0((0,1))\,,\qquad \theta\mapsto\begin{pmatrix}\partial_x\\0\end{pmatrix}\theta$$

is a bounded invertible operator. Moreover, we point out that $L_{0,0}$ leaves $H^1_{\mathrm{per}}((0,1))\times L^2_0((0,1))$ invariant,[15] and its restriction is conjugated to $\mathcal{L}_{0,0}$ through

$$\mathcal{L}_{0,0}=\mathfrak{m}^{-1}\begin{pmatrix}\mathrm{I}&0\\0&(\mathcal{I}^{(0)})^{-1}\end{pmatrix}(L_{0,0})_{|\,H^1_{\mathrm{per}}((0,1))\times L^2_0((0,1))}\begin{pmatrix}\mathrm{I}&0\\0&\mathcal{I}^{(0)}\end{pmatrix}\mathfrak{m}\,.$$

Denoting π_0 the orthogonal projector of $H^1_{\mathrm{per}}((0,1))\times L^2_{\mathrm{curl}_{0,0}}((0,1))$ on $H^1_{\mathrm{per}}((0,1))\times L^2_0((0,1))$ we also note that $(\mathrm{I}-\pi_0)\,L_{0,0}\,(\mathrm{I}-\pi_0)$ is identically zero, and $\pi_0\,L_{0,0}\,(\mathrm{I}-\pi_0)$ is bounded. As a conclusion, one derives when λ is nonzero and does not belong to the spectrum of $\mathcal{L}_{(0,0)}$ or equivalently, when λ is nonzero and does not belong to the spectrum of $L_{(0,0)}$

$$\begin{aligned}(\lambda\mathrm{I}-L_{(0,0)})^{-1}&=\frac{1}{\lambda}\,(\mathrm{I}-\pi_0)\\&+\begin{pmatrix}\mathrm{I}&0\\0&\mathcal{I}^{(0)}\end{pmatrix}\mathfrak{m}^{-1}(\lambda\mathrm{I}-\mathcal{L}_{(0,0)})^{-1}\,\mathfrak{m}\begin{pmatrix}\mathrm{I}&0\\0&(\mathcal{I}^{(0)})^{-1}\end{pmatrix}\left(\pi_0+\frac{1}{\lambda}\pi_0\,L_{0,0}\,(\mathrm{I}-\pi_0)\right)\,,\end{aligned}$$

so that for nonzero eigenvalues, the algebraic structures[16] of $\mathcal{L}_{(0,0)}$ and $L_{0,0}$ are the same.

---

15. For an unbounded operator A defined on X with domain D, we say that Y, a subspace of X, is left invariant by A if $A(D\cap Y)\subset Y$, and, in this case, $A_{|Y}$ is defined on Y with domain $D\cap Y$.

16. Recall that the algebraic structure of an eigenvalue λ_0 of an operator A is read on the singular part of $\lambda\mapsto(\lambda\mathrm{I}-A)^{-1}$ at $\lambda=\lambda_0$.

As we comment further below, in general, 0 is an eigenvalue of $\mathcal{L}_{(0,0)}$ of algebraic multiplicity 4 with two Jordan blocks of height 2, whereas, when $d = 1$, 0 is an eigenvalue of L_0 of algebraic multiplicity 4 with geometric multiplicity 3 and one Jordan block of height 2.

3.3. The Evans function. — Since each $\mathcal{L}_{\xi,\boldsymbol{\eta}}$ acts on functions of a scalar variable, it is convenient to analyze their spectra by focusing on spatial dynamics, rewriting spectral problems in terms of ODEs of the spatial variable. Adapting the construction of Gardner [25] to the situation at hand, this leads to the introduction of a suitable Evans function.

To keep spectral ODEs as simple as possible, it is expedient to work with unscaled equations as in Section 2. Explicitly, with the notation from Section 2, for $\lambda \in \mathbf{C}$ and $\boldsymbol{\eta} \in \mathbf{R}^{d-1}$, we consider $R(\,\cdot\,, x_0; \lambda, \boldsymbol{\eta})$ the solution operator of the first-order four-dimensional differential operator canonically associated with the second-order two-dimensional operator $\mathbf{J}\,\mathrm{Hess}\,\mathscr{H}_{\mathrm{u}}[\underline{\mathcal{V}}] + \|\boldsymbol{\eta}\|^2\,\kappa(\|\underline{\mathcal{V}}\|^2)\mathbf{J} - \lambda$. Note that $R(x_0, x_0; \lambda, \boldsymbol{\eta}) = \mathrm{I}_4$. Accordingly, we introduce the Evans function

$$D_\xi^{x_0}(\lambda, \boldsymbol{\eta}) = \det\left(R(x_0 + \underline{X}_x, x_0; \lambda, \boldsymbol{\eta}) - \mathrm{e}^{\mathrm{i}\,\xi}\,\mathrm{diag}(\mathrm{e}^{\frac{\xi}{\underline{\phi}}\mathbf{J}}, \mathrm{e}^{\frac{\xi}{\underline{\phi}}\mathbf{J}})\right). \tag{55}$$

The choice of x_0 is immaterial; we shall set $x_0 = 0$ and drop the corresponding superscript in the following.

The backbone of the Evans function theory is that λ_0 belongs to the spectrum of $\mathcal{L}_{\xi,\boldsymbol{\eta}}$ if and only if λ_0 is a root of $D_\xi(\,\cdot\,, \boldsymbol{\eta})$, and that its (algebraic) multiplicity as an eigenvalue of $\mathcal{L}_{\xi,\boldsymbol{\eta}}$ agrees with its multiplicity as a root of $D_\xi(\,\cdot\,, \boldsymbol{\eta})$. The first part of the claim is a simple reformulation of the fact that the spectrum of $\mathcal{L}_{\xi,\boldsymbol{\eta}}$ contains only eigenvalues, whereas the second part may be derived from the expression of resolvents of $\mathcal{L}_{\xi,\boldsymbol{\eta}}$ at λ in terms of solution operators $R(\,\cdot\,, \cdot; \lambda, \boldsymbol{\eta})$ and the characterization/definition of algebraic multiplicity at λ_0 as the rank of the residue at λ_0 of the resolvent map.

To a large extent, the benefits from using an Evans function instead of directly studying spectra are the same as those arising from the consideration of characteristic polynomials to study finite-dimensional spectra.

3.4. High-frequency analysis. — It is quite straightforward to check that when $|\Re(\lambda)|$ is sufficiently large, λ does not belong to the spectrum of any $\mathcal{L}_{\xi,\boldsymbol{\eta}}$. When λ is real, and $\xi \in \{0, \pi\}$, $D_\xi(\lambda, \boldsymbol{\eta})$ is real valued, and we would like to go further and determine its sign when $(\lambda, \boldsymbol{\eta})$ is sufficiently large with λ real. This is useful in order to derive instability criteria based on the intermediate value theorem.

Since the principal part of $\mathcal{L}_{\xi,\boldsymbol{\eta}}$ has nonconstant coefficients, this is not completely trivial, but one may reduce the computation to the constant-coefficient case by a homotopy argument similar to the one in [9].

Proposition 3.1. — *Let $\underline{\mathcal{V}}$ be an unscaled wave profile (in the sense of* (22)*). There exists $R_0 > 0$ such that for any $(\lambda, \boldsymbol{\eta}) \in \mathbf{R} \times \mathbf{R}^{d-1}$ satisfying $|\lambda| + \|\boldsymbol{\eta}\|^2 \geq R_0$, we have $D_0(\lambda, \boldsymbol{\eta}) > 0$ and $D_\pi(\lambda, \boldsymbol{\eta}) > 0$.*

Proof. — An elementary Lax–Milgram type argument shows that when $(\lambda, \boldsymbol{\eta})$ is sufficiently large (with λ real) independently of $\theta \in [0,1]$, λ does not belong to the spectrum of

$$\mathcal{L}_{0,\boldsymbol{\eta}}^{(\theta)} := (1-\theta)\mathcal{L}_{0,\boldsymbol{\eta}} + \theta\,\mathbf{J}(-(\underline{k}_x\,\partial_x + \underline{k}_\phi\mathbf{J})^2 + \|\boldsymbol{\eta}\|^2)$$

on $L^2_{\mathrm{per}}((0,1))$, for any $\theta \in [0,1]$. The needed estimates stem from the form

$$\begin{aligned}\mathcal{L}_{0,\boldsymbol{\eta}}^{(\theta)} &= \mathbf{J}((\theta + (1-\theta)\kappa(\|\underline{\mathcal{U}}\|^2))\,(-\underline{k}_x^2\partial_x^2 + \|\boldsymbol{\eta}\|^2))\\ &\qquad + \text{lower order terms independent of } \lambda \text{ and } \boldsymbol{\eta}\end{aligned}$$

and the fact that $\min(\kappa(\|\underline{\mathcal{U}}\|^2)) > 0$. Indeed, for some positive constants c, C independent of $(\lambda, \boldsymbol{\eta}, \theta) \in \mathbf{R} \times \mathbf{R}^{d-1} \times [0,1]$

$$\begin{aligned}\langle(\mathbf{J}\mathbf{V} - \mathrm{sgn}(\lambda)\mathbf{V}, (\mathcal{L}_{0,\boldsymbol{\eta}}^{(\theta)} - \lambda)\mathbf{V})\rangle_{L^2} &\geq c\,(\|\mathbf{V}\|_{H^1}^2 + (\|\boldsymbol{\eta}\|^2 + |\lambda|)\|\mathbf{V}\|^2)\\ &\quad - C\,\|\mathbf{V}\|_{H^1}\,\|\mathbf{V}\|_{L^2}\\ &\geq \frac{c}{2}\,(\|\mathbf{V}\|_{H^1}^2 + (\|\boldsymbol{\eta}\|^2 + |\lambda|)\|\mathbf{V}\|^2)\,,\end{aligned}$$

provided that $(\lambda, \boldsymbol{\eta})$ is sufficiently large and $\mathbf{V} \in H^2_{\mathrm{per}}((0,1))$. A similar bound holds for the adjoint problem.

For the corresponding Evans functions, this implies that $(\lambda, \boldsymbol{\eta}, \theta) \mapsto D_0^{(\theta)}(\lambda, \boldsymbol{\eta})$ has a constant sign on

$$\{\,(\lambda, \boldsymbol{\eta}, \theta) \in \mathbf{R} \times \mathbf{R}^{d-1} \times [0,1]\,;\, |\lambda| + \|\boldsymbol{\eta}\| \geq R_0\,,\,\}$$

for some $R_0 > 0$. This sign is easily evaluated by considering either $D_0^{(1)}(\lambda, 0)$ when λ is large or[17] $D_0^{(1)}(0, \boldsymbol{\eta})$ when $\|\boldsymbol{\eta}\|$ is large. The foregoing computations can be made even more explicit running first another homotopy argument moving $\mathbf{J}(-(\underline{k}_x\,\partial_x + \underline{k}_\phi\mathbf{J})^2 + \|\boldsymbol{\eta}\|^2)$ to $\mathbf{J}(-\underline{k}_x^2\,\partial_x^2 + \|\boldsymbol{\eta}\|^2)$, thus reducing to $\underline{\xi}_\phi = 0$; in this case, we have $D_0^{(1)}(0, \boldsymbol{\eta}) = (\mathrm{e}^{\|\boldsymbol{\eta}\|\underline{X}_x} - 1)^2\,(\mathrm{e}^{-\|\boldsymbol{\eta}\|\underline{X}_x} - 1)^2$.

The study of $D_\pi(\lambda, \boldsymbol{\eta})$ is nearly identical and thus omitted. □

3.5. Low-frequency analysis. — Now we turn to the derivation of an expansion of $D_\xi(\lambda, \boldsymbol{\eta})$ when $(\lambda, \xi, \boldsymbol{\eta})$ is small. We begin with a few preliminary remarks to prepare such an expansion.

For the sake of brevity in algebraic manipulations, we introduce the notation

$$[A]_0 := A(\underline{X}_x) - \mathrm{e}^{\underline{\xi}_\phi\mathbf{J}}\,A(0)\,.$$

17. When $d = 1$, this requires us to first embed artificially the spectral problem at hand in a corresponding higher-dimensional problem.

Let us observe that if $\mathcal{A}[\mathbf{U}] = \mathcal{A}(\mathbf{U},\, \mathbf{U}_x)$ is rotationally invariant,

$$L\mathcal{A}[\mathbf{U}]\mathbf{V} = L\mathcal{A}[\mathrm{e}^{\varphi_\phi \mathbf{J}}\,\mathbf{U}]\,\mathrm{e}^{\varphi_\phi \mathbf{J}}\,\mathbf{V}\,,$$

for any $\varphi_\phi \in \mathbf{R}$; hence if $\mathcal{V}(\cdot + \underline{X}_x) = \mathrm{e}^{\underline{\xi}_\phi \mathbf{J}}\,\mathcal{V}(\cdot)$,

$$\begin{aligned}(L\mathcal{A}[\mathcal{V}]\boldsymbol{\psi})(0) &= \mathrm{d}_{(\mathbf{U},\mathbf{U}_x)}\,\mathcal{A}(\mathcal{V}(\underline{X}_x), \mathcal{V}_x(\underline{X}_x))(\mathrm{e}^{\underline{\xi}_\phi \mathbf{J}}\;\boldsymbol{\psi}(0), \mathrm{e}^{\underline{\xi}_\phi \mathbf{J}}\;\boldsymbol{\psi}_x(0))\,,\\ (L\mathcal{A}[\mathcal{V}]\boldsymbol{\psi})(\underline{X}_x) - (L\mathcal{A}[\mathcal{V}]\boldsymbol{\psi})(0) &= \mathrm{d}_{(\mathbf{U},\mathbf{U}_x)}\,\mathcal{A}(\mathcal{V}(\underline{X}_x), \mathcal{V}_x(\underline{X}_x))([\boldsymbol{\psi}]_0, [\boldsymbol{\psi}_x]_0)\,.\end{aligned}$$

All relevant quantities depend on $\boldsymbol{\eta}$ only through $\|\boldsymbol{\eta}\|^2$ and a wealth of information on the regime $(\lambda, \xi, \boldsymbol{\eta})$ small – used repeatedly below without mention – is obtained by differentiating (22), (33) and (34); for $a = \omega_\phi, c_x, \mu_\phi, \mu_x$,

$$[\partial_a \underline{\mathcal{V}}]_0 = \partial_a \underline{\xi}_\phi\;\mathrm{e}^{\underline{\xi}_\phi \mathbf{J}}\,\mathbf{J}\underline{\mathcal{V}}(0) - \partial_a \underline{X}_x\;\mathrm{e}^{\underline{\xi}_\phi \mathbf{J}}\;\underline{\mathcal{V}}_x(0)\,, \tag{56}$$

$$(L(\mathcal{S}_\phi + \underline{c}_x \mathcal{M})[\underline{\mathcal{V}}]\partial_a \underline{\mathcal{V}})(0) = \partial_a \underline{\mu}_\phi - \partial_a \underline{c}_x\, \mathcal{M}[\underline{\mathcal{V}}](0)\,, \tag{57}$$

$$(L(\mathcal{S}_x + \underline{\omega}_\phi \mathcal{M})[\underline{\mathcal{V}}]\partial_a \underline{\mathcal{V}})(0) = \partial_a \underline{\mu}_x - \partial_a \underline{\omega}_\phi\, \mathcal{M}[\underline{\mathcal{V}}](0)\,. \tag{58}$$

Finally, from Appendix A we derive that if

$$\lambda\boldsymbol{\psi} = \mathbf{J}\,\mathrm{Hess}\,\mathcal{H}_{\mathrm{u}}[\underline{\mathcal{V}}]\boldsymbol{\psi} + \|\boldsymbol{\eta}\|^2\,\kappa(\|\underline{\mathcal{V}}\|^2)\mathbf{J}\boldsymbol{\psi}\,,$$

then

$$\lambda L\mathcal{M}[\underline{\mathcal{V}}]\boldsymbol{\psi} = \partial_x\left(L(\mathcal{S}_\phi + \underline{c}_x\mathcal{M})[\underline{\mathcal{V}}]\boldsymbol{\psi}\right) - \|\boldsymbol{\eta}\|^2\,\kappa(\|\underline{\mathcal{V}}\|^2)\,\mathbf{J}\underline{\mathcal{V}}\cdot\boldsymbol{\psi}\,, \tag{59}$$

$$\begin{aligned}\lambda L\mathcal{Q}_1[\underline{\mathcal{V}}]\boldsymbol{\psi} = \partial_x&\left(L(\mathcal{S}_x + \underline{\omega}_\phi\mathcal{M})[\underline{\mathcal{V}}]\boldsymbol{\psi}\right) - \|\boldsymbol{\eta}\|^2\,\kappa(\|\underline{\mathcal{V}}\|^2)\,\underline{\mathcal{V}}_x\cdot\boldsymbol{\psi}\\ &+ \partial_x\left(\frac{1}{2}\mathbf{J}\underline{\mathcal{V}}\cdot\left(\lambda\boldsymbol{\psi} - \|\boldsymbol{\eta}\|^2\,\kappa(\|\underline{\mathcal{V}}\|^2)\mathbf{J}\boldsymbol{\psi}\right)\right).\end{aligned} \tag{60}$$

Moreover, as has already been mentioned in Section 2.2

$$\det(\mathrm{d}_{\mathbf{U}_x}(\mathcal{S}_\phi, \mathcal{S}_x)(\underline{\mathcal{V}}(\underline{X}_x), \underline{\mathcal{V}}_x(\underline{X}_x)) = (\kappa(\|\underline{\mathcal{V}}(0)\|^2))^2\,\underline{\mathcal{V}}(0)\cdot\underline{\mathcal{V}}_x(0)\,.$$

Theorem 3.2. — *With the notation from Section 2, consider an unscaled wave profile $\underline{\mathcal{V}}$ such that $\underline{\mathcal{V}}\cdot\underline{\mathcal{V}}_x \not\equiv 0$. Then the corresponding Evans function expands uniformly in $\xi \in [-\pi, \pi]$ as*

$$\begin{aligned}D_\xi(\lambda, \boldsymbol{\eta}) = \det&\left(\lambda\,\Sigma_t - (\mathrm{e}^{\mathrm{i}\,\xi} - 1)\mathrm{I}_4 + \tfrac{\|\boldsymbol{\eta}\|^2}{\lambda}\Sigma_{\mathbf{y}}\right)\\ &+ \mathcal{O}\left((|\lambda| + |\xi| + \|\boldsymbol{\eta}\|^2)\,(|\lambda|^2 + |\xi|^2 + \|\boldsymbol{\eta}\|^2)\,(|\lambda|(|\lambda| + |\xi|) + \|\boldsymbol{\eta}\|^2)\right),\end{aligned} \tag{61}$$

when $(\lambda,\boldsymbol{\eta})\to(0,0)$, *with*

$$\Sigma_t=\begin{pmatrix}\partial_a\underline{\xi}_\phi\\ \partial_a\underline{X}_x\\ \int_0^{\underline{X}_x}L\mathscr{M}[\underline{\mathcal{V}}]\partial_a\underline{\mathcal{V}}+\mathscr{M}[\underline{\mathcal{V}}](0)\,\partial_a\underline{X}_x\\ \int_0^{\underline{X}_x}L\mathbb{Q}_1[\underline{\mathcal{V}}]\partial_a\underline{\mathcal{V}}+\mathbb{Q}_1[\underline{\mathcal{V}}](0)\,\partial_a\underline{X}_x\end{pmatrix}_{a=\omega_\phi,c_x,\mu_\phi,\mu_x},$$

$$\Sigma_{\mathbf{y}}=\begin{pmatrix}0&0&0&0\\0&0&0&0\\ \int_0^{\underline{X}_x}\kappa(\|\underline{\mathcal{V}}\|^2)\,\|\underline{\mathcal{V}}\|^2 & -\int_0^{\underline{X}_x}\kappa(\|\underline{\mathcal{V}}\|^2)\,\mathbf{J}\underline{\mathcal{V}}\cdot\underline{\mathcal{V}}_x&0&0\\ \int_0^{\underline{X}_x}\kappa(\|\underline{\mathcal{V}}\|^2)\,\underline{\mathcal{V}}_x\cdot\mathbf{J}\underline{\mathcal{V}} & -\int_0^{\underline{X}_x}\kappa(\|\underline{\mathcal{V}}\|^2)\,\|\underline{\mathcal{V}}_x\|^2&0&0\end{pmatrix}.$$

Note that the structure of $\Sigma_{\mathbf{y}}$ is consistent with the fact that there is actually no singularity in the low-frequency expansion of the Evans function; every power of $\|\boldsymbol{\eta}\|^2/\lambda$ is balanced by a corresponding power of λ.

Proof. — By a density argument on the point where the Evans function is considered, we may reduce the analysis to the case when $\underline{\mathcal{V}}(0)\cdot\underline{\mathcal{V}}_x(0)\neq0$.

Guided by rotation and translation invariance, we introduce

$$\Psi_1(\,\cdot\,;\lambda,\boldsymbol{\eta})=R(\,\cdot\,,0;\lambda,\boldsymbol{\eta})\begin{pmatrix}\mathbf{J}\underline{\mathcal{V}}(0)\\ \mathbf{J}\underline{\mathcal{V}}_x(0)\end{pmatrix}\quad \Psi_2(\,\cdot\,;\lambda,\boldsymbol{\eta})=R(\,\cdot\,,0;\lambda,\boldsymbol{\eta})\begin{pmatrix}\underline{\mathcal{V}}_x(0)\\ \underline{\mathcal{V}}_{xx}(0)\end{pmatrix},$$

$$\Psi_3(\,\cdot\,;\lambda,\boldsymbol{\eta})=R(\,\cdot\,,0;\lambda,\boldsymbol{\eta})\begin{pmatrix}\partial_{\mu_\phi}\underline{\mathcal{V}}(0)\\ \partial_{\mu_\phi}\underline{\mathcal{V}}_x(0)\end{pmatrix}\ \Psi_4(\,\cdot\,;\lambda,\boldsymbol{\eta})=R(\,\cdot\,,0;\lambda,\boldsymbol{\eta})\begin{pmatrix}\partial_{\mu_x}\underline{\mathcal{V}}(0)\\ \partial_{\mu_x}\underline{\mathcal{V}}_x(0)\end{pmatrix},$$

so that in particular

$$\Psi_1(\,\cdot\,;0,0)=\begin{pmatrix}\mathbf{J}\underline{\mathcal{V}}\\ \mathbf{J}\underline{\mathcal{V}}_x\end{pmatrix},\qquad \Psi_2(\,\cdot\,;0,0)=\begin{pmatrix}\underline{\mathcal{V}}_x\\ \underline{\mathcal{V}}_{xx}\end{pmatrix}$$
$$\Psi_3(\,\cdot\,;0,0)=\begin{pmatrix}\partial_{\mu_\phi}\underline{\mathcal{V}}\\ \partial_{\mu_\phi}\underline{\mathcal{V}}_x\end{pmatrix},\quad \Psi_4(\,\cdot\,;0,0)=\begin{pmatrix}\partial_{\mu_x}\underline{\mathcal{V}}\\ \partial_{\mu_x}\underline{\mathcal{V}}_x\end{pmatrix}.$$

Then we set $\Psi=\begin{pmatrix}\Psi_1&\Psi_2&\Psi_3&\Psi_4\end{pmatrix}$ and observe that from the computations in Section 2.2 stems

$$D_\xi(\lambda,\boldsymbol{\eta})=(\kappa(\|\underline{\mathcal{V}}(0)\|^2))^2$$
$$\times\det\Big([\Psi(\,\cdot\,;\lambda,\boldsymbol{\eta})]_0-(\mathrm{e}^{\mathrm{i}\,\xi}-1)\,\mathrm{diag}(\mathrm{e}^{\underline{\xi}_\phi\mathbf{J}},\mathrm{e}^{\underline{\xi}_\phi\mathbf{J}})\Psi(0;\lambda,\boldsymbol{\eta})\Big)\,.$$

Note that each $\Psi_\ell(\,\cdot\,;\lambda,\boldsymbol{\eta})$ splits as $(\boldsymbol{\psi}_\ell(\,\cdot\,;\lambda,\boldsymbol{\eta}),(\boldsymbol{\psi}_\ell)_x(\,\cdot\,;\lambda,\boldsymbol{\eta}))$ for some $\boldsymbol{\psi}_\ell(\,\cdot\,;\lambda,\boldsymbol{\eta})$ and that

$$\lambda\boldsymbol{\psi}_\ell=\mathbf{J}\,\mathrm{Hess}\,\mathscr{H}_{\mathrm{u}}[\underline{\mathcal{V}}]\boldsymbol{\psi}_\ell+\|\boldsymbol{\eta}\|^2\,\kappa(\|\underline{\mathcal{V}}\|^2)\mathbf{J}\boldsymbol{\psi}_\ell\,.$$

We may now use the identities (59) (60)) and perform line combinations so as to obtain that $(\underline{\mathcal{V}}(0)\cdot\underline{\mathcal{V}}_x(0))\times D_\xi(\lambda,\boldsymbol{\eta})$ coincides with the determinant of a matrix of the form

$$\begin{pmatrix}[\boldsymbol{\psi}_\ell]_0-(\mathrm{e}^{\mathrm{i}\,\xi}-1)\,\mathrm{e}^{\frac{\xi}{\underline{\ }}{}_\phi\mathbf{J}}\,\boldsymbol{\psi}_\ell(0)\\ \lambda\int_0^{\underline{X}_x}L\mathscr{M}[\underline{\mathcal{V}}](\boldsymbol{\psi}_\ell)+\|\boldsymbol{\eta}\|^2\int_0^{\underline{X}_x}\kappa(\|\underline{\mathcal{V}}\|^2)\,\mathbf{J}\underline{\mathcal{V}}\cdot\boldsymbol{\psi}_\ell-(\mathrm{e}^{\mathrm{i}\,\xi}-1)\,(L(\mathcal{S}_\phi+\underline{c}_x\mathscr{M})[\underline{\mathcal{V}}]\boldsymbol{\psi}_\ell)(0)\\ \lambda\int_0^{\underline{X}_x}L\mathbb{Q}_1[\underline{\mathcal{V}}](\boldsymbol{\psi}_\ell)+\|\boldsymbol{\eta}\|^2\int_0^{\underline{X}_x}\kappa(\|\underline{\mathcal{V}}\|^2)\,\underline{\mathcal{V}}_x\cdot\boldsymbol{\psi}_\ell\\ -(\mathrm{e}^{\mathrm{i}\,\xi}-1)\,\left(\left(L(\mathcal{S}_x+\underline{\omega}_\phi\mathscr{M})[\underline{\mathcal{V}}]+\left(\frac{\lambda}{2}\mathbf{J}\underline{\mathcal{V}}-\|\boldsymbol{\eta}\|^2\,\kappa(\|\underline{\mathcal{V}}\|^2)\underline{\mathcal{V}}\right)\cdot\right)\boldsymbol{\psi}_\ell\right)(0)\end{pmatrix}_\ell,$$

in the limit $(\lambda,\boldsymbol{\eta})\to(0,0)$ (where we have left implicit the dependence of $\boldsymbol{\psi}_\ell$ on $(\lambda,\boldsymbol{\eta})$ for the sake of concision). Then we observe that it follows from invariances by rotational and spatial translations that the first two columns of the foregoing matrix are of the form

$$\begin{pmatrix}\mathcal{O}(|\lambda|+|\xi|+\|\boldsymbol{\eta}\|^2)\\ \mathcal{O}(|\lambda|+|\xi|+\|\boldsymbol{\eta}\|^2)\\ \mathcal{O}(|\lambda|\,(|\lambda|+|\xi|)+\|\boldsymbol{\eta}\|^2)\\ \mathcal{O}(|\lambda|\,(|\lambda|+|\xi|)+\|\boldsymbol{\eta}\|^2)\end{pmatrix},$$

when $(\lambda,\boldsymbol{\eta})\to(0,0)$ and that, as follows by comparing respective equations, both $\partial_\lambda\boldsymbol{\psi}_1(\,\cdot\,;0,0)$ and $\partial_{\omega_\phi}\underline{\mathcal{V}}$ on the one hand and $\partial_\lambda\boldsymbol{\psi}_2(\,\cdot\,;0,0)$, and $-\partial_{c_x}\underline{\mathcal{V}}$ on the other hand, differ only by a linear combination of $\boldsymbol{\psi}_1(\,\cdot\,;0,0)$, $\boldsymbol{\psi}_2(\,\cdot\,;0,0)$, $\boldsymbol{\psi}_3(\,\cdot\,;0,0)$ and $\boldsymbol{\psi}_4(\,\cdot\,;0,0)$.

Therefore, from a direct expansion and a column manipulation, one derives that

$$(-\underline{\mathcal{V}}(0)\cdot\underline{\mathcal{V}}_x(0))\times D_\xi(\lambda,\boldsymbol{\eta})=\det\left(C_1\ C_2\ C_3\ C_4\right),$$

with

$$\begin{aligned}C_1=&\begin{pmatrix}\lambda\,[\partial_{\omega_\phi}\underline{\mathcal{V}}]_0-(\mathrm{e}^{\mathrm{i}\,\xi}-1)\,\mathrm{e}^{\frac{\xi}{\underline{\ }}{}_\phi\mathbf{J}}\,\mathbf{J}\underline{\mathcal{V}}(0)\\ \lambda^2\int_0^{\underline{X}_x}L\mathscr{M}[\underline{\mathcal{V}}]\partial_{\omega_\phi}\underline{\mathcal{V}}+\|\boldsymbol{\eta}\|^2\int_0^{\underline{X}_x}\kappa(\|\underline{\mathcal{V}}\|^2)\,\|\underline{\mathcal{V}}\|^2\\ \lambda^2\int_0^{\underline{X}_x}L\mathbb{Q}_1[\underline{\mathcal{V}}]\partial_{\omega_\phi}\underline{\mathcal{V}}+\|\boldsymbol{\eta}\|^2\int_0^{\underline{X}_x}\kappa(\|\underline{\mathcal{V}}\|^2)\,\underline{\mathcal{V}}_x\cdot\mathbf{J}\underline{\mathcal{V}}\end{pmatrix}\\ &+\begin{pmatrix}\mathcal{O}(|\lambda|(|\lambda|+|\xi|)+\|\boldsymbol{\eta}\|^2)\\ \mathcal{O}((|\lambda|^2+\|\boldsymbol{\eta}\|^2)(|\lambda|+|\xi|)+\|\boldsymbol{\eta}\|^4)\\ \mathcal{O}((|\lambda|^2+\|\boldsymbol{\eta}\|^2)(|\lambda|+|\xi|)+\|\boldsymbol{\eta}\|^4)\end{pmatrix},\end{aligned}$$

$$C_2 = \begin{pmatrix} \lambda\,[\partial_{c_x}\underline{\mathcal{V}}]_0 + (\mathrm{e}^{\mathrm{i}\,\xi}-1)\ \mathrm{e}^{\frac{\xi}{\phi}\mathbf{J}}\ \underline{\mathcal{V}}_x(0) \\ \lambda^2\int_0^{X_x} L\mathscr{M}[\underline{\mathcal{V}}]\partial_{c_x}\underline{\mathcal{V}} + \lambda\,(\mathrm{e}^{\mathrm{i}\,\xi}-1)\,\mathscr{M}[\underline{\mathcal{V}}](0) - \|\boldsymbol{\eta}\|^2\int_0^{X_x}\kappa(\|\underline{\mathcal{V}}\|^2)\,\mathbf{J}\underline{\mathcal{V}}\cdot\underline{\mathcal{V}}_x \\ \lambda^2\int_0^{X_x} L\mathbb{Q}_1[\underline{\mathcal{V}}]\partial_{c_x}\underline{\mathcal{V}} + \lambda\,(\mathrm{e}^{\mathrm{i}\,\xi}-1)\,\mathbb{Q}_1[\underline{\mathcal{V}}](0) - \|\boldsymbol{\eta}\|^2\int_0^{X_x}\kappa(\|\underline{\mathcal{V}}\|^2)\,\|\underline{\mathcal{V}}_x\|^2 \end{pmatrix}$$
$$+ \begin{pmatrix} \mathcal{O}(|\lambda|(|\lambda|+|\xi|)+\|\boldsymbol{\eta}\|^2) \\ \mathcal{O}((|\lambda|^2+\|\boldsymbol{\eta}\|^2)(|\lambda|+|\xi|)+\|\boldsymbol{\eta}\|^4) \\ \mathcal{O}((|\lambda|^2+\|\boldsymbol{\eta}\|^2)(|\lambda|+|\xi|)+\|\boldsymbol{\eta}\|^4) \end{pmatrix},$$

$$C_3 = \begin{pmatrix} [\partial_{\mu_\phi}\underline{\mathcal{V}}]_0 - (\mathrm{e}^{\mathrm{i}\,\xi}-1)\ \mathrm{e}^{\frac{\xi}{\phi}\mathbf{J}}\ \partial_{\mu_\phi}\underline{\mathcal{V}}(0) \\ \lambda\int_0^{X_x} L\mathscr{M}[\underline{\mathcal{V}}]\partial_{\mu_\phi}\underline{\mathcal{V}} + \|\boldsymbol{\eta}\|^2\ \int_0^{X_x}\kappa(\|\underline{\mathcal{V}}\|^2)\,\mathbf{J}\underline{\mathcal{V}}\cdot\partial_{\mu_\phi}\underline{\mathcal{V}} - (\mathrm{e}^{\mathrm{i}\,\xi}-1) \\ \lambda\int_0^{X_x} L\mathbb{Q}_1[\underline{\mathcal{V}}]\partial_{\mu_\phi}\underline{\mathcal{V}} + \|\boldsymbol{\eta}\|^2\ \int_0^{X_x}\kappa(\|\underline{\mathcal{V}}\|^2)\,\underline{\mathcal{V}}_x\cdot\partial_{\mu_\phi}\underline{\mathcal{V}} \end{pmatrix}$$
$$+ \begin{pmatrix} \mathcal{O}(|\lambda|+\|\boldsymbol{\eta}\|^2) \\ \mathcal{O}((|\lambda|+\|\boldsymbol{\eta}\|^2)(|\lambda|+|\xi|+\|\boldsymbol{\eta}\|^2)) \\ \mathcal{O}((|\lambda|+\|\boldsymbol{\eta}\|^2)(|\lambda|+|\xi|+\|\boldsymbol{\eta}\|^2)) \end{pmatrix},$$

and

$$C_4 = \begin{pmatrix} [\partial_{\mu_x}\underline{\mathcal{V}}]_0 - (\mathrm{e}^{\mathrm{i}\,\xi}-1)\ \mathrm{e}^{\frac{\xi}{\phi}\mathbf{J}}\ \partial_{\mu_x}\underline{\mathcal{V}}(0) \\ \lambda\int_0^{X_x} L\mathscr{M}[\underline{\mathcal{V}}]\partial_{\mu_x}\underline{\mathcal{V}} + \|\boldsymbol{\eta}\|^2\ \int_0^{X_x}\kappa(\|\underline{\mathcal{V}}\|^2)\,\mathbf{J}\underline{\mathcal{V}}\cdot\partial_{\mu_x}\underline{\mathcal{V}} \\ \lambda\int_0^{X_x} L\mathbb{Q}_1[\underline{\mathcal{V}}]\partial_{\mu_x}\underline{\mathcal{V}} + \|\boldsymbol{\eta}\|^2\ \int_0^{X_x}\kappa(\|\underline{\mathcal{V}}\|^2)\,\underline{\mathcal{V}}_x\cdot\partial_{\mu_x}\underline{\mathcal{V}} - (\mathrm{e}^{\mathrm{i}\,\xi}-1) \end{pmatrix}$$
$$+ \begin{pmatrix} \mathcal{O}(|\lambda|+\|\boldsymbol{\eta}\|^2) \\ \mathcal{O}((|\lambda|+\|\boldsymbol{\eta}\|^2)(|\lambda|+|\xi|+\|\boldsymbol{\eta}\|^2)) \\ \mathcal{O}((|\lambda|+\|\boldsymbol{\eta}\|^2)(|\lambda|+|\xi|+\|\boldsymbol{\eta}\|^2)) \end{pmatrix}.$$

Then the result follows steadily from an expansion of the determinant and a few manipulations on the first two lines based on Formula (56) for $[\partial_a\underline{\mathcal{V}}]_0$. □

4. Longitudinal perturbations

We begin by completing and discussing consequences of the latter sections on the stability analysis for longitudinal perturbations. For results derived – *via* Madelung's transformation – from corresponding known results for larger classes of Euler–Korteweg systems, we also provide some hints about direct proofs.

4.1. Coperiodic perturbations. — As in [12, 9], we connect stability with respect to coperiodic longitudinal perturbations with properties of the Hessian of the action integral Θ. We remember that Θ is considered as a function of $(\mu_x, c_x, \omega_\phi, \mu_\phi)$, in that order.

At the spectral level, restricting ourselves to coperiodic longitudinal perturbations corresponds to focusing on $\mathcal{L}_{0,0}$, the Bloch–Fourier symbol at $(\xi,\boldsymbol{\eta}) = (0,0)$. It is thus worth pointing out that it follows from identities in (47) that the matrix Σ_t in Theorem 3.2 is such that

$$\Sigma_t = \begin{pmatrix} 0\,0\,0\,1 \\ 1\,0\,0\,0 \\ 0\,0\,1\,0 \\ 0\,1\,0\,0 \end{pmatrix} \begin{pmatrix} \mathrm{I}_2 & 0 \\ 0 & -\mathrm{I}_2 \end{pmatrix} \text{Hess}\,\Theta \begin{pmatrix} 0\,0\,0\,1 \\ 0\,1\,0\,0 \\ 1\,0\,0\,0 \\ 0\,0\,1\,0 \end{pmatrix} \tag{62}$$

so that

$$D_0(\lambda, 0) = \lambda^4 \det(\text{Hess}\,\Theta) + \mathcal{O}\left(|\lambda|^5\right) \tag{63}$$

as $|\lambda| \to 0$. Combining it with Proposition 3.1 provides the first half of the following theorem. Incidentally, we point out that we expect that this half could alternatively be derived by counting Krein signatures and refer the reader to [9, Remark 3] and [37, Chapter 13] for more information on the latter.

Theorem 4.1. — *Let $\underline{\mathcal{U}}$ be a wave profile of parameter $(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi)$, such that* Hess $\Theta(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi)$ *is nonsingular.*

1. *The number of eigenvalues of $\mathcal{L}_{0,0}$ in $(0, +\infty)$, counted with algebraic multiplicity, is*
   - *even if* $\det(\text{Hess}\,\Theta)(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi) > 0$ *;*
   - *odd if* $\det(\text{Hess}\,\Theta)(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi) < 0$.

   In particular, in the latter case, the wave is spectrally exponentially unstable to coperiodic longitudinal perturbations.
2. *Assume that $\partial^2_{\mu_x}\Theta(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi) \neq 0$ and that the negative signature of* Hess $\Theta(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi)$ *equals 2. Then the wave is conditionally orbitally stable in $H^1_{\text{per}}((0, \underline{X}_x))$.*

By conditional orbital stability in $H^1_{\text{per}}((0, \underline{X}_x))$, we mean that for any $\delta_0 > 0$, there exists $\varepsilon_0 > 0$ such that for any $\mathbf{U}_0$ satisfying

$$\inf_{(\varphi_\phi, \varphi_x) \in \mathbb{R}^2} \left\| \mathbf{U}_0 - \mathrm{e}^{\varphi_\phi \mathbf{J}}\, \underline{\mathcal{U}}(\,\cdot + \varphi_x\,) \right\|_{H^1_{\text{per}}((0,1))} \leq \varepsilon_0$$

and any solution[18] $\mathbf{U}$ to (8) defined on an interval I containing 0, starting from $\mathbf{U}(0, \cdot\,) = \mathbf{U}_0$ and sufficiently smooth to guarantee that

- $\mathbf{U} \in \mathcal{C}^0(I; H^1_{\text{per}}((0,1)))$;
- $t \mapsto \int_0^{\underline{X}_x} \mathscr{M}[\mathbf{U}(t, \cdot)]$, $t \mapsto \int_0^{\underline{X}_x} \mathbb{Q}_1[\mathbf{U}(t, \cdot)]$ and $t \mapsto \int_0^{\underline{X}_x} \mathscr{H}[\mathbf{U}(t, \cdot)]$ are constant on I;

then for any $t \in I$,

$$\inf_{(\varphi_\phi, \varphi_x) \in \mathbb{R}^2} \left\| \mathbf{U}(t, \cdot\,) - \mathrm{e}^{\varphi_\phi \mathbf{J}}\, \underline{\mathcal{U}}(\,\cdot + \varphi_x\,) \right\|_{H^1_{\text{per}}((0,1))} \leq \delta_0\,.$$

18. Knowing in which precise sense does not matter since only conservations are used in the stability argument.

To go from conditional orbital stability to orbital stability one needs to know that for the notion of solution at hand controlling the H^1-norm is sufficient to prevent finite-time blowup. This is in particular the case when κ is constant; see e.g., [19, Section 3.5]).

Proof. — The first point is a direct consequence of identity (63) and $D_0(\lambda, 0) > 0$ for λ real and large (Proposition 3.1).

The second part is deduced from a corresponding result for the Euler–Korteweg system (39): $(\underline{\rho}, \underline{u})$ is conditionally orbitally stable in $H^1_{per} \times L^2_{per}((0,1))$ if the negative signature of Hess $\Theta(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi)$ equals 2. See Theorem 3 and its accompanying remarks in [9, Section 4.2] (conveniently summarized as [10, Theorem 1]). The conversion to our setting stems from the following lemma and the fact that System (39) preserves the integral of v_1. □

Lemma 4.2. — 1. *For any $c_0 > 0$, there exist $\varepsilon > 0$ and C, such that if $\underline{\mathcal{U}} \in H^1_{\rm per}((0,1))$ satisfies $\|\underline{\mathcal{U}}\| \geq c_0$, then for any $(\varphi_\phi, \varphi_x) \in \mathbf{R}^2$ and any $\mathbf{U} \in H^1_{\rm per}((0,1))$ satisfying*

$$\left\|\mathbf{U} - \mathrm{e}^{\varphi_\phi \mathbf{J}}\, \underline{\mathcal{U}}(\,\cdot + \varphi_x\,)\right\|_{H^1_{\rm per}((0,1))} \leq \varepsilon\,,$$

with

$$(\rho, \widetilde{v}) = \left(\mathscr{M}[\mathbf{U}], \frac{\mathbb{Q}_1[\mathbf{U}]}{\mathscr{M}[\mathbf{U}]}\right), \qquad (\underline{\rho}, \underline{\widetilde{v}}) = \left(\mathscr{M}[\underline{\mathcal{U}}], \frac{\mathbb{Q}_1[\underline{\mathcal{U}}]}{\mathscr{M}[\mathbf{U}]}\right),$$

there holds $(\rho, \widetilde{v})$, $(\underline{\rho}, \underline{\widetilde{v}}) \in H^1_{\rm per}((0,1)) \times L^2((0,1))$, $\int_0^1 \widetilde{v} = \int_0^1 \underline{\widetilde{v}} = 0$, and

$$\begin{aligned}&\left\|(\rho, \widetilde{v}) - (\underline{\rho}, \underline{\widetilde{v}})(\,\cdot + \varphi_x\,)\right\|_{H^1_{\rm per} \times L^2_{\rm per}} \\ &\qquad \leq C\left(1 + \|\underline{\mathcal{U}}\|^3_{H^1_{\rm per}}\right)\left\|\mathbf{U} - \mathrm{e}^{\varphi_\phi \mathbf{J}}\, \underline{\mathcal{U}}(\,\cdot + \varphi_x\,)\right\|_{H^1_{\rm per}}.\end{aligned}$$

2. *There exists C such that if*

$$\underline{\mathcal{U}} = \sqrt{2\underline{\rho}}\; \mathrm{e}^{\underline{\theta} \mathbf{J}}(\mathbf{e}_1)\,, \qquad\qquad \mathbf{U} = \sqrt{2\rho}\; \mathrm{e}^{\theta \mathbf{J}}(\mathbf{e}_1)\,,$$

with $(\rho, \partial_x\theta)$, $(\underline{\rho}, \partial_x\underline{\theta}) \in H^1_{\rm per}((0,1)) \times L^2((0,1))$, $\int_0^1 \partial_x\theta = \int_0^1 \partial_x\underline{\theta} = 0$, then $\mathbf{U} \in H^1_{\rm per}((0,1))$, $\underline{\mathcal{U}} \in H^1_{\rm per}((0,1))$ and, for any $\varphi_x \in \mathbf{R}$,

$$\begin{aligned}&\left\|\mathbf{U} - \mathrm{e}^{\varphi_\phi \mathbf{J}}\, \underline{\mathcal{U}}(\,\cdot + \varphi_x\,)\right\|_{H^1_{\rm per}} \\ &\qquad \leq C\left(1 + \|(\rho, \partial_x\theta)\|^2_{H^1_{\rm per} \times L^2_{\rm per}} + \left\|(\underline{\rho}, \partial_x\underline{\theta})\right\|^2_{H^1_{\rm per} \times L^2_{\rm per}}\right) \\ &\qquad\quad \times \left\|(\rho, \partial_x\theta) - (\underline{\rho}, \partial_x\underline{\theta})(\,\cdot + \varphi_x\,)\right\|_{H^1_{\rm per} \times L^2_{\rm per}}\end{aligned}$$

where

$$\varphi_\phi = \int_0^1 (\theta(\zeta) - \underline{\theta}(\,\zeta + \varphi_x\,))d\zeta\,.$$

Proof. — The proof of the lemma is quite straightforward, using the continuous embedding

$$H^1_{\mathrm{per}}((0,1)) \hookrightarrow L^\infty((0,1))$$

and the Poincaré inequality. We use the latter in the following form. There exists C such that for any θ such that $\partial_x\theta \in L^2((0,1))$, $\int_0^1 \partial_x\theta = 0$, we have $\theta \in H^1_{\mathrm{per}}((0,1))$, and if $\int_0^1 \theta = 0$,

$$\|\theta\|_{L^2((0,1))} \leq C\,\|\partial_x\theta\|_{L^2((0,1))}\,. \qquad \square$$

The first part of the foregoing theorem could also be deduced from [12, 9] through Section 3.2. In the reverse direction, we expect that the second part could be deduced from abstract results directly concerning equations of the same type as the nonlinear Schrödinger equation – see [27, 14] – essentially as the conclusions in [9] used here were deduced there by combining an abstract result – [9, Theorem 3] – with a result proving connections with the action integral – [9, Theorem 7].

As in [10] for systems of Korteweg type, we now specialize Theorem 4.1 to two asymptotic regimes, small-amplitude and large-period. To state our result, in the small-amplitude regime, we need one more nondegeneracy index

(64)

$$\begin{aligned}\mathfrak{a}_0(c_x,\rho,k_\phi) := \frac{1}{8(\partial_\rho^2\mathcal{W}_\rho)^3}\Bigg[&\frac{5}{3}(\partial_\rho^3\mathcal{W}_\rho)^2 - \partial_\rho^2\mathcal{W}_\rho\,\partial_\rho^4\mathcal{W}_\rho\\ &- 4\,\partial_\rho^2\mathcal{W}_\rho\,\partial_\rho^3\mathcal{W}_\rho\left(\frac{\kappa'(2\rho)}{\kappa(\rho)} - \frac{1}{2\,\rho}\right)\\ &+ 16\,(\partial_\rho^2\mathcal{W}_\rho)^2\left(\frac{\kappa''(2\rho)}{\kappa(\rho)} - \frac{1}{2\,\rho}\frac{\kappa'(2\rho)}{\kappa(\rho)} + \frac{1}{2\,(\rho)^2}\right)\Bigg],\end{aligned}$$

with derivatives of $\mathcal{W}_\rho$ evaluated at $(\rho;c_x,\omega_\phi,\mu_\phi)$, (ω_ϕ,μ_ϕ) being associated with (c_x,ρ,k_ϕ) through (50). The following theorem is then merely a translation of Corollaries 1 and 2 in [10].

THEOREM 4.3. — 1. *In the small amplitude regime near a* $(\underline{c}_x^{(0)},\underline{\rho}^{(0)},\underline{k}_\phi^{(0)})$, *such that*[19]

$$\partial_\rho\nu(\underline{\rho}^{(0)};\underline{c}_x^{(0)},\underline{\mu}_\phi^{(0)}) \neq 0\,, \qquad \mathfrak{a}_0(\underline{c}_x^{(0)},\underline{\rho}^{(0)},\underline{k}_\phi^{(0)}) \neq 0\,,$$

we have that $\partial^2_{\mu_x}\Theta \neq 0$ *and that the negative signature of* $\mathrm{Hess}\,\Theta$ *equals 2, so that waves are conditionally orbitally stable in* $H^1_{\mathrm{per}}((0,\underline{X}_x))$.

19. With $\underline{\mu}_\phi^{(0)}$ associated with $(\underline{c}_x^{(0)},\underline{\rho}^{(0)},\underline{k}_\phi^{(0)})$ through (50).

2. *In the large period regime,* $\partial_{\mu_x}^2\Theta \neq 0$ *and*
   - *if* $\partial_{c_x}^2\Theta_{(s)}(\underline{c}_x^{(0)},\underline{\rho}^{(0)},\underline{k}_\phi^{(0)}) > 0$ *then in the large period regime near* $(\underline{c}_x^{(0)},\underline{\rho}^{(0)},\underline{k}_\phi^{(0)})$*, the negative signature of* Hess Θ *equals 2, so that waves are conditionally orbitally stable in* $H^1_{\rm per}((0,\underline{X}_x))$*;*
   - *if* $\partial_{c_x}^2\Theta_{(s)}(\underline{c}_x^{(0)},\underline{\rho}^{(0)},\underline{k}_\phi^{(0)}) < 0$*, then in the large-period regime near* $(\underline{c}_x^{(0)},\underline{\rho}^{(0)},\underline{k}_\phi^{(0)})$*, the negative signature of* Hess Θ *equals 3, so that waves are spectrally exponentially unstable to coperiodic longitudinal perturbations.*

A few comments are in order.

1. The condition $\mathfrak{a}_0 \neq 0$ is directly connected to the condition $\partial_{\mu_x}^2\Theta \neq 0$ since $\mathfrak{a}_0\,X_x^{(0)}$ is the limiting value of $\partial_{\mu_x}^2\Theta \neq 0$ in the small-amplitude regime; see [10, Theorem 4].
2. The small-amplitude regime considered here is disjoint from the one analyzed for the semilinear cubic Schrödinger equations in [24] since here the constant asymptotic mass is nonzero, namely $\underline{\rho}^{(0)} > 0$.
3. The condition on $\partial_{c_x}^2\Theta_{(s)}$ agrees with the usual criterion for stability of solitary waves, known as the Vakhitov–Kolokolov slope condition; see e.g., [27].

4.2. Sideband perturbations. — Sideband perturbations are perturbations corresponding to Floquet exponents ξ arbitrarily small but nonzero. As in [12, 13, 11], we analyze the spectrum of $\mathcal{L}_{\xi,0}$ near 0 when ξ is small. Some instability criteria associated with this part of the spectrum could be deduced readily from Theorem 3.2. Yet we postpone these conclusions slightly since we are more interested in proving that such rigorous conclusions agree with those guessed from formal geometrical optics considerations.

Thus, let us consider the two-phase slow/fast-oscillatory *ansatz*

$$\mathbf{U}^{(\varepsilon)}(t,x) = \mathrm{e}^{\frac{1}{\varepsilon}\varphi_\phi^{(\varepsilon)}(\varepsilon\,t,\varepsilon\,x)\,\mathbf{J}}\,\mathcal{U}^{(\varepsilon)}\left(\varepsilon\,t,\varepsilon\,x;\frac{\varphi_x^{(\varepsilon)}(\varepsilon\,t,\varepsilon\,x)}{\varepsilon}\right), \tag{65}$$

with, for any (T,X), $\zeta\mapsto\mathcal{U}^{(\varepsilon)}(T,X;\zeta)$ periodic of period 1 and, as $\varepsilon\to0$,

$$\begin{aligned}
\mathcal{U}^{(\varepsilon)}(T,X;\zeta) &= \mathcal{U}_0(T,X;\zeta) + \varepsilon\,\mathcal{U}_1(T,X;\zeta) + o(\varepsilon)\,,\\
\varphi_\phi^{(\varepsilon)}(T,X) &= (\varphi_\phi)_0(T,X) + \varepsilon\,(\varphi_\phi)_1(T,X) + o(\varepsilon)\,,\\
\varphi_x^{(\varepsilon)}(T,X) &= (\varphi_x)_0(T,X) + \varepsilon\,(\varphi_x)_1(T,X) + o(\varepsilon)\,.
\end{aligned}$$

Requiring (65) to solve (3) up to a remainder of size $o(1)$ is equivalent to $\zeta\mapsto\mathcal{U}_0(T,X;\zeta)$ being a scaled profile of a periodic traveling wave of (22).

Explicitly,

$$\mathbf{J}\delta\mathcal{H}_0(\mathcal{U}_0, \mathrm{e}_1\,(k_\phi\mathbf{J} + k_x\partial_\zeta)\mathcal{U}_0) = \omega_\phi\mathbf{J}\mathcal{U}_0 - c_x\,(k_\phi\mathbf{J} + k_x\partial_\zeta)\mathcal{U}_0\,, \tag{66}$$

with local parameters (depending on slow variables (T, X)) related to phases by

$$\partial_T(\varphi_\phi)_0 = \omega_\phi - k_\phi\,c_x\,, \quad \partial_X(\varphi_\phi)_0 = k_\phi\,, \quad \partial_T(\varphi_x)_0 = \omega_x\,, \quad \partial_X(\varphi_x)_0 = k_x\,.$$

Symmetry of derivatives already constrains the slow evolution of wave parameters with

$$\partial_T k_\phi = \partial_X\left(\omega_\phi - k_\phi\,c_x\right), \qquad \partial_T k_x = \partial_X\omega_x\,.$$

Since periodic profiles form a four-dimensional manifold (after discarding translation and rotation parameters), in order to determine the leading-order dynamics of (65), we need two more equations. The fastest way to obtain such equations is to also require (65) to solve (4) and (6) up to remainders of size $o(\varepsilon)$. Observing that all quantities in (4) and (6) are independent of phases,

$$\begin{aligned}
&\partial_T(\mathscr{M}(\mathcal{U}_0)) = \partial_X(\mathcal{S}_\phi(\mathcal{U}_0, \mathrm{e}_1\,(k_x\partial_\zeta + k_\phi\mathbf{J})\mathcal{U}_0)) + \partial_\zeta\,(*)\,,\\
&\partial_T(\mathbb{Q}_1(\mathcal{U}_0, \mathrm{e}_1\,(k_x\partial_\zeta + k_\phi\mathbf{J})\mathcal{U}_0))\\
&\qquad = \partial_X\big(\big(\nabla_{\mathbf{U}_x}\mathbb{Q}_1\cdot\mathbf{J}\delta\mathcal{H}_0 + \mathcal{S}_x\big)(\mathcal{U}_0, \mathrm{e}_1\,(k_x\partial_\zeta + k_\phi\mathbf{J})\mathcal{U}_0)\big) + \partial_\zeta\,(**)\,,
\end{aligned}$$

with omitted terms $*$ and $**$ 1-periodic in ζ. Averaging in ζ (using (66)) provides two more equations, completing the modulation system

$$\begin{cases}
\partial_T k_x = \partial_X\omega_x\\
\partial_T(\langle\mathbb{Q}(_1\mathcal{U}_0, \mathrm{e}_1\,(k_x\partial_\zeta + k_\phi\mathbf{J})\mathcal{U}_0)\rangle)\\
\qquad\qquad = \partial_X(\langle(\omega_\phi\mathcal{M} - c_x\mathbb{Q}_1 + \mathcal{S}_x)(\mathcal{U}_0, \mathrm{e}_1\,(k_x\partial_\zeta + k_\phi\mathbf{J})\mathcal{U}_0)\rangle)\\
\partial_T(\langle\mathscr{M}(\mathcal{U}_0)\rangle) = \partial_X(\langle\mathcal{S}_\phi(\mathcal{U}_0, \mathrm{e}_1\,(k_x\partial_\zeta + k_\phi\mathbf{J})\mathcal{U}_0)\rangle)\\
\partial_T k_\phi = \partial_X\left(\omega_\phi - k_\phi\,c_x\right)
\end{cases}, \tag{67}$$

where $\langle\,\cdot\,\rangle = \displaystyle\int_0^1 \cdot\, d\zeta$ is the average over a periodic cell.

The reader may wonder why in the foregoing formal derivation we asked for (3) to be satisfied at order 1 and for (4) and (6) to be satisfied at order ε. Alternatively, one may ask for (3) to be satisfied at order ε and check that requirements on (4) and (6) come as necessary conditions. One may also check that when (3) is satisfied at order 1, so are (4) and (6).

System (67) should be thought of as a system for functions defined on the manifold of periodic traveling waves (identified when coinciding up to rotational and spatial translations). To make this more concrete, we now rewrite it in terms of parameters $(\mu_x, c_x, \omega_\phi, \mu_\phi)$. To do so, with notation from Section 2, we introduce

$$
(68)\qquad \begin{cases} \mathscr{m}(\mu_x, c_x, \omega_\phi, \mu_\phi) := \langle \mathscr{M}[\mathcal{V}] \rangle = \dfrac{1}{X_x} \displaystyle\int_0^{X_x} \mathscr{M}[\mathcal{V}] dx\,, \\ \mathscr{q}(\mu_x, c_x, \omega_\phi, \mu_\phi) := \langle \mathbb{Q}[\mathcal{V}] \rangle = \dfrac{1}{X_x} \displaystyle\int_0^{X_x} \mathbb{Q}_1[\mathcal{V}] dx\,, \end{cases}
$$

where $\mathcal{V}$ is the unscaled profile associated with $(\mu_x, c_x, \omega_\phi, \mu_\phi)$, and X_x is the corresponding period. By making use of (33) and (34), one obtains

$$
(69)\qquad \begin{cases} \partial_T k_x = \partial_X \omega_x \\ \partial_T \mathscr{q} = \partial_X \left(\mu_x - c_x \mathscr{q}\right) \\ \partial_T \mathscr{m} = \partial_X \left(\mu_\phi - c_x \mathscr{m}\right) \\ \partial_T k_\phi = \partial_X \left(\omega_\phi - c_x\, k_\phi\right) \end{cases}
$$

as an alternative form of (67). To connect with the analysis of other sections in terms of the action integral Θ, we recall (47)

$$
k_x = \frac{1}{\partial_{\mu_x}\Theta}\,, \qquad \begin{pmatrix} 1 \\ \mathscr{q} \\ \mathscr{m} \\ k_\phi \end{pmatrix} = \frac{\mathbf{A}_0\, \nabla\Theta}{\partial_{\mu_x}\Theta}\,, \qquad \text{with} \quad \mathbf{A}_0 := \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}.
$$

Thus (for smooth solutions), System (69) takes the alternative form

$$
(70)\qquad k_x \mathbf{A}_0 \operatorname{Hess}\Theta\ \left(\partial_T + c_x \partial_X\right) \begin{pmatrix} \mu_x \\ c_x \\ \omega_\phi \\ \mu_\phi \end{pmatrix} = \mathbf{B}_0\, \partial_X \begin{pmatrix} \mu_x \\ c_x \\ \omega_\phi \\ \mu_\phi \end{pmatrix},
$$

with

$$
\mathbf{B}_0 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.
$$

Remark 4.4. — One may check that the modulated system (69), also often called Whitham's system, agrees with the one derived for the associated Euler–Korteweg system (39) by injecting a one-phase slow/fast-oscillatory *ansatz*. See [48, 13, 11] for a discussion of the latter. This may be achieved by direct comparisons of either formal *ansatz*, averaged forms, or more concrete parameterized forms.

We now specialize the use of System (69) to the discussion of the dynamics near a particular periodic traveling wave. Note that traveling-wave solutions fit the *ansatz* (65) and correspond to the case when phases φ_ϕ and φ_x are affine functions of the slow variables, and wave parameters are constant. Thus, when

$\underline{\mathbf{U}}$ is a wave profile of parameters $(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi)$, one may expect that the stability[20] of $(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi)$ as a solution to (69) is necessary for the stability of $\underline{\mathbf{U}}$ as a solution to (8). The literature proving such a claim at the spectral level is now quite extensive, and we refer the reader to [53, 44], [13, 17], [36], and [32] for respectively, results on parabolic systems, Hamiltonian systems of Korteweg type, lattice dynamical systems, and some hyperbolic systems with discontinuous waves. Yet this is the first time[21] that a result for a class of systems with symmetry group of dimension higher than 1 is established.

In the present case, the spectral validation of (69) is a simple corollary of Theorem 3.2 based on a counting root argument for analytic functions, since

$$\lambda\,\Sigma_t - (\mathrm{e}^{\mathrm{i}\,\xi} - 1)\mathrm{I}_4 = \begin{pmatrix} 0&0&0&1\\ 1&0&0&0\\ 0&0&1&0\\ 0&1&0&0 \end{pmatrix} \left(\lambda\,\mathbf{A}_0\ \mathrm{Hess}\,\Theta - (\mathrm{e}^{\mathrm{i}\,\xi} - 1)\,\mathbf{B}_0\right) \begin{pmatrix} 0&0&0&1\\ 0&1&0&0\\ 1&0&0&0\\ 0&0&1&0 \end{pmatrix}.$$

Corollary 4.5. — *Consider an unscaled wave profile $\underline{\mathcal{V}}$ such that $\underline{\mathcal{V}} \cdot \underline{\mathcal{V}}_x \not\equiv 0$, with associated parameters $(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi)$.*

1. *The following three statements are equivalent.*
   - 0 *is an eigenvalue of algebraic multiplicity* 4 *of* $\mathcal{L}_{0,0}$.
   - *The map* $(\mu_x, c_x, \omega_\phi, \mu_\phi) \mapsto (k_x, q, m, k_\phi)$ *is a local diffeomorphism near* $(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi)$.
   - $\mathrm{Hess}\,\Theta(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi)$ *is nonsingular.*
2. *Assume that* $\mathrm{Hess}\,\Theta(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi)$ *is nonsingular. Then there exist* $\lambda_0 > 0$, $\xi_0 > 0$ *and* C_0 *such that*
   - *for any* $\xi \in [-\xi_0, \xi_0]$, $\mathcal{L}_{\xi,0}$ *possesses four eigenvalues (counted with algebraic multiplicity) in the disk* $B(0, \lambda_0)$*;*
   - *if* $a - \underline{c}_x$ *is a characteristic speed of* (69) *at* $(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi)$ *of algebraic multiplicity* r*, that is, if* a *is an eigenvalue of* $(k_x \mathbf{A}_0\,\mathrm{Hess}\,\Theta(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi))^{-1}\mathbf{B}_0$ *of algebraic multiplicity* r*, then for any* $\xi \in [-\xi_0, \xi_0]$, $\mathcal{L}_{\xi,0}$ *possesses* r *eigenvalues (counted with algebraic multiplicity) in the disk* $B(\mathrm{i}\,\underline{k}_x \xi\, a\,,\, C_0|\xi|^{1+\frac{1}{r}})$.

   In particular, if System (69) *is not weakly hyperbolic at* $(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi)$*, that is, if* $(k_x \mathbf{A}_0\,\mathrm{Hess}\,\Theta(\underline{\mu}_x, \underline{c}_x, \underline{\omega}_\phi, \underline{\mu}_\phi))^{-1}\mathbf{B}_0$ *possesses a nonreal eigenvalue, then the wave is spectrally unstable to longitudinal sideband perturbations.*

A few comments are in order.

20. Incidentally, we point out that from the homogeneity of first-order systems it follows that ill-posedness and stability are essentially the same for systems such as (69).

21. Except for the almost simultaneous [20]. See the detailed comparison in Section 4.3.

1. Note that the subtraction of $\underline{c}_x$ in the second part of the corollary accounts for the fact that System 69 is not expressed in a comoving frame.
2. The second part of the foregoing corollary could also be deduced from results in [13] through Madelung's transformation.

We now turn to the small-amplitude and large-period regimes. To describe the small-amplitude regime, we need to introduce two instability indices

$$\delta_{hyp}(c_x,\omega_\phi,\mu_\phi) := W''(2\,\rho^{(0)}) + \Big(\kappa''(2\,\rho^{(0)})\,\rho^{(0)} + \kappa'(2\,\rho^{(0)})\Big)\,(k_\phi^{(0)})^2 \tag{71}$$

and

(72)
$$\begin{aligned}
\delta_{BF}(c_x,\omega_\phi,\mu_\phi) :=& \left(\frac{1}{2}\frac{\kappa(2\,\rho^{(0)})}{2\,\rho^{(0)}}\left(\frac{2\pi}{X_x^{(0)}}\right)^2\right)^3\\
&\times\left(-3\left(\frac{\kappa'(2\,\rho^{(0)})}{\kappa(2\,\rho^{(0)})}\right)^2 - 2\,\frac{\kappa'(2\,\rho^{(0)})}{\kappa(2\,\rho^{(0)})}\,\frac{1}{2\,\rho^{(0)}} + \frac{\kappa''(2\,\rho^{(0)})}{\kappa(2\,\rho^{(0)})}\right)\\
&+\left(\frac{1}{2}\frac{\kappa(2\,\rho^{(0)})}{2\,\rho^{(0)}}\left(\frac{2\pi}{X_x^{(0)}}\right)^2\right)^2\\
&\times\Bigg(W''(2\,\rho^{(0)})\Bigg(-12\left(\frac{\kappa'(2\,\rho^{(0)})}{\kappa(2\,\rho^{(0)})}\right)^2 - 6\,\frac{\kappa'(2\,\rho^{(0)})}{\kappa(2\,\rho^{(0)})}\,\frac{1}{2\,\rho^{(0)}}\\
&\qquad + 4\left(\frac{1}{2\,\rho^{(0)}}\right)^2 + 3\,\frac{\kappa''(2\,\rho^{(0)})}{\kappa(2\,\rho^{(0)})}\Bigg)\\
&\qquad + 4\,W'''(2\,\rho^{(0)})\left(\frac{\kappa'(2\,\rho^{(0)})}{\kappa(2\,\rho^{(0)})} + 2\,\frac{1}{2\,\rho^{(0)}}\right) + 2\,W''''(2\,\rho^{(0)})\Bigg)\\
&+\left(\frac{1}{2}\frac{\kappa(2\,\rho^{(0)})}{2\,\rho^{(0)}}\left(\frac{2\pi}{X_x^{(0)}}\right)^2\right)\\
&\times\Bigg(12\,(W''(2\,\rho^{(0)}))^2\left(\left(\frac{\kappa'(2\,\rho^{(0)})}{\kappa(2\,\rho^{(0)})}\right)^2 + 4\,\frac{\kappa'(2\,\rho^{(0)})}{\kappa(2\,\rho^{(0)})}\,\frac{1}{2\,\rho^{(0)}} + 3\left(\frac{1}{2\,\rho^{(0)}}\right)^2\right)\\
&\qquad + 8\,W''(2\,\rho^{(0)})\,W'''(2\,\rho^{(0)})\left(4\,\frac{\kappa'(2\,\rho^{(0)})}{\kappa(2\,\rho^{(0)})} + 5\,\frac{1}{2\,\rho^{(0)}}\right)\\
&\qquad + \frac{4}{3}\Big(W'''(2\,\rho^{(0)})\Big)^2 + 6\,W''(2\,\rho^{(0)})\,W''''(2\,\rho^{(0)})\Bigg)\\
&+ 8\,W''(2\,\rho^{(0)})\left(W'''(2\,\rho^{(0)}) + 3\,W''(2\,\rho^{(0)})\left(\frac{\kappa'(2\,\rho^{(0)})}{\kappa(2\,\rho^{(0)})} + \frac{1}{2\,\rho^{(0)}}\right)\right)^2,
\end{aligned}$$

where $(\rho^{(0)}, k_\phi^{(0)})$ are the associated limiting mass and rotational shift, and $X_x^{(0)}$ is the associated period.

The following theorem is a consequence of Corollary 4.5 and results in [11] for the Euler–Korteweg systems, namely Theorems 7 and 8, respectively, for the first and second points[22].

Theorem 4.6. — 1. *In the small-amplitude regime near a* $(\underline{c}_x^{(0)}, \underline{\rho}^{(0)}, \underline{k}_\phi^{(0)})$ *such that*

$$\partial_\rho \nu(\underline{\rho}^{(0)}; \underline{c}_x^{(0)}, \underline{\mu}_\phi^{(0)}) \neq 0\,,$$

$\operatorname{Hess}\Theta$ *is nonsingular, and if*

$$\delta_{hyp}(\underline{c}_x^{(0)}, \underline{\omega}_\phi^{(0)}, \underline{\mu}_\phi^{(0)}) < 0 \qquad \textit{or} \qquad \delta_{BF}(\underline{c}_x^{(0)}, \underline{\omega}_\phi^{(0)}, \underline{\mu}_\phi^{(0)}) < 0\,,$$

then waves are spectrally exponentially unstable to longitudinal sideband perturbations.

2. *If* $\partial_{c_x}^2 \Theta_{(s)}(\underline{c}_x^{(0)}, \underline{\rho}^{(0)}, \underline{k}_\phi^{(0)}) \neq 0$*, then, in the large-period regime near* $(\underline{c}_x^{(0)}, \underline{\rho}^{(0)}, \underline{k}_\phi^{(0)})$*,* $\operatorname{Hess}\Theta$ *is nonsingular, and if*

$$\partial_{c_x}^2 \Theta_{(s)}(\underline{c}_x^{(0)}, \underline{\rho}^{(0)}, \underline{k}_\phi^{(0)}) < 0\,,$$

then in the large-period regime near $(\underline{c}_x^{(0)}, \underline{\rho}^{(0)}, \underline{k}_\phi^{(0)})$*, waves are spectrally exponentially unstable to longitudinal sideband perturbations.*

A few comments are worth making. In particular, we borrow here some of the upshots of the much more comprehensive analysis in [11].

1. Again, we point out that the small-amplitude regime considered here is disjoint from the one analyzed for the semilinear cubic Schrödinger equations in [24]. Let us, however, stress that for the semilinear cubic Schrödinger equations our instability criterion provides instability if and only if the potential is focusing, independently of the particular limit value under consideration. This is consistent with the conclusions for the case $\rho^{(0)} = 0$ derived in [24].
2. In the small-amplitude limit, the characteristic velocities split into two groups of two. One of these groups converges to the linear group velocity at the limiting constant value and the sign of δ_{BF} precisely determines how this double root splits. The corresponding instability is often referred to as the Benjamin–Feir instability. The other group converges to the characteristic velocities of a dispersionless hydrodynamic system at the limiting constant value; see [11, Theorem 7]. The sign of δ_{hyp} decides the weak hyperbolicity of the latter system. When κ is constant, it is directly related to the focusing/defocusing nature of the potential W (namely W'' negative/positive).

22. In notation of [11], δ_{BF} is Δ_{MI}.

3. A similar scenario takes place in the large-period limit, with the phase velocity of the solitary wave replacing the linear group velocity. The sign of $\partial_{c_x}^2 \Theta_{(s)}$ determines how the double root splits. However, due to the nature of end states of solitary waves, the dispersionless system is always hyperbolic, hence the reduction to a single instability index. See Appendix B for some related details.

4.3. Large-time dynamics. — Our interest in modulated systems also hinges on the belief that they play a deep role in the description of the large-time dynamics. In other words, one expects that near stable waves the large-time dynamics is well approximated by simply varying wave parameters in a space–time dependent way and that the dynamics of these parameters is itself well captured by some (higher-order version of a) modulated system.

The latter scenario has been proved to occur at the nonlinear level for a large class of parabolic systems [33, 34] and at the linearized level for the Korteweg–de Vries equation [50]. The reader is also referred to [48, 49] for some more intuitive arguments supporting the general claim.

We would like to extend here a small part of the analysis in [50] to the class of equations under consideration. We begin by revisiting the second part of Corollary 4.5 from the point of view of Floquet symbols rather than Evans' functions. The goal is to provide a description of how eigenfunctions and spectral projectors behave near the quadruple eigenvalue at the origin. Once this is done, the arguments of [50] may be directly imported and provide different results (adapted to the presence of a two-dimensional group of symmetries) but with nearly identical – thus omitted – proofs.

In a certain way, we leave the point of view convenient for spatial dynamics to focus on time dynamics. To do so, it is expedient to use scaled variables so as to normalize period and to parameter waves not by phase-portrait parameters $(\mu_x, c_x, \omega_\phi, \mu_\phi)$ but by modulation parameters $(k_x, k_\phi, \mathpzc{q}, \mathpzc{m})$. The first part of Corollary 4.5 proves that the latter is possible when the eigenvalue at the origin is indeed of multiplicity 4. Therefore, in the present section, we consider scaled profiles $\mathcal{U}$ as in (8) and parameters $(\mu_x, c_x, \omega_\phi, \mu_\phi)$ as functions of $(k_x, k_\phi, \mathpzc{q}, \mathpzc{m})$. In scaled variables, the averaged mass and impulse from (68) take the form

$$\mathpzc{m} = \langle \mathscr{M}(\mathcal{U}) \rangle = \int_0^1 \frac{1}{2} \|\mathcal{U}\|^2 ,$$

$$\mathpzc{q} = \langle \mathbb{Q}_1(\mathcal{U}, \mathrm{e}_1(k_x\partial_x + k_\phi \mathbf{J})\mathcal{U}) \rangle = \int_0^1 \frac{1}{2} \mathbf{J}\mathcal{U} \cdot (k_x\partial_x + k_\phi \mathbf{J})\mathcal{U} .$$

Our focus is on the operator $\mathcal{L}_{\xi,0} = \mathcal{L}^x_\xi$. Correspondingly, we consider the Whitham matrix-valued map

$$\mathbf{W}(\mu_x, c_x, \omega_\phi, \mu_\phi) := \mathrm{Jac}\begin{pmatrix} \omega_x \\ \omega_\phi - c_x\, k_\phi \\ \mu_x - c_x \mathscr{q} \\ \mu_\phi - c_x \mathscr{m} \end{pmatrix}. \tag{73}$$

To connect both objects, we shall use various algebraic relations obtained from profile equations and conservation laws that we first derive.

Differentiating profile equation $\delta\mathcal{H}_u[\mathcal{U}] = 0$ with respect to rotational and spatial translation parameters (left implicit here) and to $(\mathscr{q}, \mathscr{m})$ yields

$$\mathcal{L}_{0,0}\,\underline{\mathcal{U}}_x = 0\,, \qquad \mathcal{L}_{0,0}\,\partial_{\mathscr{q}}\underline{\mathcal{U}} = \left(\partial_{\mathscr{q}}\underline{\omega}_\phi - \underline{k}_\phi\,\partial_{\mathscr{q}}\underline{c}_x\right)\mathbf{J}\,\underline{\mathcal{U}} - \underline{k}_x\partial_{\mathscr{q}}\underline{c}_x\,\underline{\mathcal{U}}_x\,, \tag{74}$$

$$\mathcal{L}_{0,0}\,\mathbf{J}\,\underline{\mathcal{U}} = 0\,, \qquad \mathcal{L}_{0,0}\,\partial_{\mathscr{m}}\underline{\mathcal{U}} = \left(\partial_{\mathscr{m}}\underline{\omega}_\phi - \underline{k}_\phi\,\partial_{\mathscr{m}}\underline{c}_x\right)\mathbf{J}\,\underline{\mathcal{U}} - \underline{k}_x\,\partial_{\mathscr{m}}\underline{c}_x\,\underline{\mathcal{U}}_x\,. \tag{75}$$

To highlight the role of $\partial_{k_x}\underline{\mathcal{U}}$ and $\partial_{k_\phi}\underline{\mathcal{U}}$ we expand

$$\mathcal{L}_{\xi,0} = \mathcal{L}_{0,0} + \mathrm{i}\,\underline{k}_x\xi\,\mathcal{L}_{(1)} + (\mathrm{i}\,\underline{k}_x\xi)^2\,\mathcal{L}_{(2)}\,.$$

Differentiating profile equations with respect to (k_x, k_ϕ) leaves

$$\begin{cases} \mathcal{L}_{0,0}\,\partial_{k_x}\underline{\mathcal{U}} = \left(\partial_{k_x}\underline{\omega}_\phi - \underline{k}_\phi\,\partial_{k_x}\underline{c}_x\right)\mathbf{J}\,\underline{\mathcal{U}} - \underline{k}_x\,\partial_{k_x}\underline{c}_x\,\underline{\mathcal{U}}_x - \mathcal{L}_{(1)}\,\underline{\mathcal{U}}_x\,, \\ \mathcal{L}_{0,0}\,\partial_{k_\phi}\underline{\mathcal{U}} = \left(\partial_{k_\phi}\underline{\omega}_\phi - \underline{k}_\phi\,\partial_{k_\phi}\underline{c}_x\right)\mathbf{J}\,\underline{\mathcal{U}} - \underline{k}_x\,\partial_{k_\phi}\underline{c}_x\,\underline{\mathcal{U}}_x - \mathcal{L}_{(1)}\,\mathbf{J}\,\underline{\mathcal{U}}\,. \end{cases} \tag{76}$$

By differentiating the definitions of mass and impulse averages, we also obtain that

$$\int_0^1 \delta\mathbb{Q}_1(\underline{\mathcal{U}}, \mathrm{e}_1(\underline{k}_x\partial_x + \underline{k}_\phi\mathbf{J})\underline{\mathcal{U}})\,\underline{\mathcal{U}}_x = 0\,, \tag{77}$$

$$\int_0^1 \delta\mathbb{Q}_1(\underline{\mathcal{U}}, \mathrm{e}_1(\underline{k}_x\partial_x + \underline{k}_\phi\mathbf{J})\underline{\mathcal{U}})\,\mathbf{J}\,\underline{\mathcal{U}} = 0\,,$$

$$\int_0^1 \delta\mathbb{Q}_1(\underline{\mathcal{U}}, \mathrm{e}_1(\underline{k}_x\partial_x + \underline{k}_\phi\mathbf{J})\underline{\mathcal{U}})\,\partial_{\mathscr{q}}\underline{\mathcal{U}} = 1\,, \tag{78}$$

$$\int_0^1 \delta\mathbb{Q}_1(\underline{\mathcal{U}}, \mathrm{e}_1(\underline{k}_x\partial_x + \underline{k}_\phi\mathbf{J})\underline{\mathcal{U}})\,\partial_{\mathscr{m}}\underline{\mathcal{U}} = 0\,,$$

$$\begin{aligned} \int_0^1 \delta\mathbb{Q}_1(\underline{\mathcal{U}}, \mathrm{e}_1(\underline{k}_x\partial_x + \underline{k}_\phi\mathbf{J})\underline{\mathcal{U}})\,\partial_{k_x}\underline{\mathcal{U}} &= -\int_0^1 \mathrm{d}_{\mathbf{U}_x}\,\mathbb{Q}_1(\underline{\mathcal{U}}, \mathrm{e}_1(\underline{k}_x\partial_x + \underline{k}_\phi\mathbf{J})\underline{\mathcal{U}})\,\underline{\mathcal{U}}_x \\ &= -\frac{1}{\underline{k}_x}(\underline{\mathscr{q}} - \underline{k}_\phi\underline{\mathscr{m}})\,, \end{aligned} \tag{79}$$

$$\begin{aligned} \int_0^1 \delta\mathbb{Q}_1(\underline{\mathcal{U}}, \mathrm{e}_1(\underline{k}_x\partial_x + \underline{k}_\phi\mathbf{J})\underline{\mathcal{U}})\,\partial_{k_\phi}\underline{\mathcal{U}} &= -\int_0^1 \mathrm{d}_{\mathbf{U}_x}\,\mathbb{Q}(\underline{\mathcal{U}}, \mathrm{e}_1(\underline{k}_x\partial_x + \underline{k}_\phi\mathbf{J})\underline{\mathcal{U}})\,\mathbf{J}\,\underline{\mathcal{U}} \\ &= -\underline{\mathscr{m}}\,, \end{aligned} \tag{80}$$

and

$$\int_0^1 \delta\mathscr{M}[\underline{\mathcal{U}}]\,\underline{\mathcal{U}}_x = 0\,, \qquad \int_0^1 \delta\mathscr{M}[\underline{\mathcal{U}}]\,\partial_q\underline{\mathcal{U}} = 0\,, \qquad \int_0^1 \delta\mathscr{M}[\underline{\mathcal{U}}]\,\partial_{k_x}\underline{\mathcal{U}} = 0\,, \tag{81}$$

$$\int_0^1 \delta\mathscr{M}[\underline{\mathcal{U}}]\,\mathbf{J}\,\underline{\mathcal{U}} = 0\,, \qquad \int_0^1 \delta\mathscr{M}[\underline{\mathcal{U}}]\,\partial_m\underline{\mathcal{U}} = 1\,, \qquad \int_0^1 \delta\mathscr{M}[\underline{\mathcal{U}}]\,\partial_{k_\phi}\underline{\mathcal{U}} = 0\,,\,. \tag{82}$$

Finally, linearizing conservation laws for mass and impulse provides for any smooth $\mathbf{V}$

$$\begin{aligned}
&\delta\mathscr{M}[\underline{\mathcal{U}}]\cdot\mathcal{L}_{\xi,0}\mathbf{V}\\
&\quad= \underline{k}_x(\partial_x+\mathrm{i}\,\xi)\Big(\nabla_{\mathbf{U}}(\mathcal{S}_\phi+\underline{c}_x\mathscr{M})(\underline{\mathcal{U}},\mathrm{e}_1(\underline{k}_x\partial_x+\underline{k}_\phi\mathbf{J})\underline{\mathcal{U}})\cdot\mathbf{V}\\
&\qquad\qquad+\nabla_{\mathbf{U}_x}(\mathcal{S}_\phi+\underline{c}_x\mathscr{M})(\underline{\mathcal{U}},\mathrm{e}_1(\underline{k}_x\partial_x+\underline{k}_\phi\mathbf{J})\underline{\mathcal{U}})\cdot(\underline{k}_x(\partial_x+\mathrm{i}\,\xi)+\underline{k}_\phi\mathbf{J})\mathbf{V}\Big)
\end{aligned}$$

and

$$\begin{aligned}
&\nabla_{\mathbf{U}}\mathbb{Q}_1(\underline{\mathcal{U}},\mathrm{e}_1(\underline{k}_x\partial_x+\underline{k}_\phi\mathbf{J})\underline{\mathcal{U}})\cdot\mathcal{L}_{\xi,0}\mathbf{V}\\
&\;+\nabla_{\mathbf{U}_x}\mathbb{Q}_1(\underline{\mathcal{U}},\mathrm{e}_1(\underline{k}_x\partial_x+\underline{k}_\phi\mathbf{J})\underline{\mathcal{U}})\cdot(\underline{k}_x(\partial_x+\mathrm{i}\,\xi)+\underline{k}_\phi\mathbf{J})\,\mathcal{L}_{\xi,0}\mathbf{V}\\
&\quad= \underline{k}_x(\partial_x+\mathrm{i}\,\xi)\Big(\nabla_{\mathbf{U}_x}\mathbb{Q}_1(\underline{\mathcal{U}},\mathrm{e}_1(\underline{k}_x\partial_x+\underline{k}_\phi\mathbf{J})\underline{\mathcal{U}})\cdot\mathcal{L}_{\xi,0}\mathbf{V}\\
&\qquad\qquad+\nabla_{\mathbf{U}}(\mathcal{S}_x+\underline{\omega}_\phi\mathscr{M})(\underline{\mathcal{U}},\mathrm{e}_1(\underline{k}_x\partial_x+\underline{k}_\phi\mathbf{J})\underline{\mathcal{U}})\cdot\mathbf{V}\\
&\qquad\qquad+\nabla_{\mathbf{U}_x}(\mathcal{S}_x+\underline{\omega}_\phi\mathscr{M})(\underline{\mathcal{U}},\mathrm{e}_1(\underline{k}_x\partial_x+\underline{k}_\phi\mathbf{J})\underline{\mathcal{U}})\cdot(\underline{k}_x(\partial_x+\mathrm{i}\,\xi)+\underline{k}_\phi\mathbf{J})\mathbf{V}\Big)\,.
\end{aligned}$$

Evaluating at $\xi=0$ and integrating show

$$\mathcal{L}_{0,0}^*\,\delta\mathbb{Q}_1(\underline{\mathcal{U}},(\underline{k}_x\partial_x+\underline{k}_\phi\mathbf{J})\underline{\mathcal{U}})=0, \qquad \mathcal{L}_{0,0}^*\,\delta\mathscr{M}[\underline{\mathcal{U}}]=0\,, \tag{83}$$

where $\mathcal{L}_{0,0}^*$ denotes the adjoint of $\mathcal{L}_{0,0}$. Alternatively, the latter may be checked by using explicit expressions of $\delta\mathbb{Q}_1(\underline{\mathcal{U}},\mathrm{e}_1(\underline{k}_x\partial_x+\underline{k}_\phi\mathbf{J})\underline{\mathcal{U}})$ and $\delta\mathscr{M}[\underline{\mathcal{U}}]$ in terms of $\mathbf{J}\underline{\mathcal{U}}$ and $\underline{\mathcal{U}}_x$ and Hamiltonian duality $\mathcal{L}_{0,0}^*=-\mathbf{J}^{-1}\,\mathcal{L}_{0,0}\,\mathbf{J}$. At next orders, for $\mathbf{V}$ smooth and periodic, we also deduce

$$\langle\delta\mathscr{M}[\underline{\mathcal{U}}];\mathcal{L}_{(1)}\mathbf{V}\rangle_{L^2}=\langle\delta(\mathcal{S}_\phi+\underline{c}_x\mathscr{M})(\underline{\mathcal{U}},\mathrm{e}_1(\underline{k}_x\partial_x+\underline{k}_\phi\mathbf{J})\underline{\mathcal{U}});\mathbf{V}\rangle_{L^2}\,, \tag{84}$$

$$\langle\delta\mathscr{M}[\underline{\mathcal{U}}];\mathcal{L}_{(2)}\mathbf{V}\rangle_{L^2}=\langle\nabla_{\mathbf{U}_x}(\mathcal{S}_\phi+\underline{c}_x\mathscr{M})(\underline{\mathcal{U}},\mathrm{e}_1(\underline{k}_x\partial_x+\underline{k}_\phi\mathbf{J})\underline{\mathcal{U}});\mathbf{V}\rangle_{L^2}\,, \tag{85}$$

and

$$\begin{aligned}
&\langle\delta\mathbb{Q}_1(\underline{\mathcal{U}},\mathrm{e}_1(\underline{k}_x\partial_x+\underline{k}_\phi\mathbf{J})\underline{\mathcal{U}});\mathcal{L}_{(1)}\mathbf{V}\rangle_{L^2}\\
&\qquad=\langle\delta(\mathcal{S}_x+\underline{\omega}_\phi\mathscr{M})(\underline{\mathcal{U}},\mathrm{e}_1(\underline{k}_x\partial_x+\underline{k}_\phi\mathbf{J})\underline{\mathcal{U}});\mathbf{V}\rangle_{L^2}\,,
\end{aligned}\tag{86}$$

$$\begin{aligned}
&\langle\delta\mathbb{Q}_1(\underline{\mathcal{U}},\mathrm{e}_1(\underline{k}_x\partial_x+\underline{k}_\phi\mathbf{J})\underline{\mathcal{U}});\mathcal{L}_{(2)}\mathbf{V}\rangle_{L^2}\\
&\qquad=\langle\nabla_{\mathbf{U}_x}(\mathcal{S}_x+\underline{\omega}_\phi\mathscr{M})(\underline{\mathcal{U}},\mathrm{e}_1(\underline{k}_x\partial_x+\underline{k}_\phi\mathbf{J})\underline{\mathcal{U}});\mathbf{V}\rangle_{L^2}\,,
\end{aligned}\tag{87}$$

In the foregoing relations, $\langle\cdot;\cdot\rangle_{L^2}$ denotes the canonical Hermitian scalar product on $L^2((0,1);\mathbf{C}^2)$, $\mathbf{C}$-linear on the right[23].

These are the key algebraic relations to prove the following proposition.

Theorem 4.7. — *Let $\underline{\mathcal{U}}$ be a wave profile such that $\underline{\mathcal{U}}\cdot\underline{\mathcal{U}}_x\not\equiv 0$ and that 0 has algebraic multiplicity exactly 4 as an eigenvalue of $\mathcal{L}_{0,0}$. Assume that eigenvalues of $\underline{\mathbf{W}}$ are distinct.*
There exist $\lambda_0>0$, $\xi_0\in(0,\pi)$, analytic curves $\lambda_j:[-\xi_0,\xi_0]\to B(0,\lambda_0)$, $j=1,2,3,4$, such that for $\xi\in[-\xi_0,\xi_0]$

$$\sigma(\mathcal{L}_{\xi,0})\cap B(0,\lambda_0)=\{\lambda_j(\xi)|j\in\{1,2,3,4\}\}$$

and associated left and right eigenfunctions $\widetilde{\psi}_j(\xi,\cdot\,)$ and $\psi_j(\xi,\cdot\,)$, $j=1,2,3,4$, satisfying pairing relations[24]

$$\langle\widetilde{\psi}_j(\xi,\cdot\,),\psi_\ell(\xi,\cdot\,)\rangle_{L^2}=\mathrm{i}\,\underline{k}_x\xi\,\delta_\ell^j,\qquad 1\le j,\ell\le 4\,,$$

obtained as

$$\begin{aligned}\psi_j(\xi,\cdot\,)&=\sum_{\ell=1}^{2}\beta_\ell^{(j)}(\xi)\,q_\ell(\xi,\cdot\,)+(\mathrm{i}\,\underline{k}_x\xi)\sum_{\ell=3}^{4}\beta_\ell^{(j)}(\xi)\,q_\ell(\xi,\cdot\,)\\ \widetilde{\psi}_j(\xi,\cdot\,)&=-(\mathrm{i}\,\underline{k}_x\xi)\sum_{\ell=1}^{2}\widetilde{\beta}_\ell^{(j)}(\xi)\,\widetilde{q}_\ell(\xi,\cdot\,)+\sum_{\ell=3}^{4}\widetilde{\beta}_\ell^{(j)}(\xi,\cdot\,)\,\widetilde{q}_\ell(\xi,\cdot\,)\,,\end{aligned}$$

where

- *$(q_j(\xi,\cdot\,))_{1\le j\le 4}$ and $(\widetilde{q}_j(\xi,\cdot\,))_{1\le j\le 4}$ are dual bases of spaces associated with the spectrum in $B(0,\lambda_0)$ of respectively $\mathcal{L}_{\xi,0}$ and its adjoint $\mathcal{L}^*_{\xi,0}$, that are analytic in ξ and such that*

$$\begin{aligned}(q_1(0,\cdot\,),q_2(0,\cdot\,),q_3(0,\cdot\,),q_4(0,\cdot\,))&=(\underline{\mathcal{U}}_x,\mathbf{J}\,\underline{\mathcal{U}},\partial_q\underline{\mathcal{U}},\partial_m\underline{\mathcal{U}})\,,\\ (\widetilde{q}_3(0,\cdot\,),\widetilde{q}_4(0,\cdot\,))&=(\delta\mathbb{Q}_1(\underline{\mathcal{U}},\mathrm{e}_1(\underline{k}_x\partial_x+\underline{k}_\phi\mathbf{J})\underline{\mathcal{U}}),\,\delta\mathscr{M}[\underline{\mathcal{U}}])\\ &=(-\underline{k}_x\,\mathbf{J}\,\underline{\mathcal{U}}_x+\underline{k}_\phi\,\underline{\mathcal{U}},\,\underline{\mathcal{U}})\,,\\ (\partial_\xi q_1(0,\cdot\,),\partial_\xi q_2(0,\cdot\,))&=\mathrm{i}\,\underline{k}_x\,(\partial_{k_x}\underline{\mathcal{U}},\,\partial_{k_\phi}\underline{\mathcal{U}})\,;\end{aligned}$$

- *$(\beta^{(j)}(\xi))_{1\le j\le 4}$ and $(\widetilde{\beta}^{(j)}(\xi))_{1\le j\le 4}$ are dual bases of $\mathbf{C}^4$ that are analytic in ξ and such that $(\beta^{(j)}(0))_{1\le j\le 4}$ and $(\widetilde{\beta}^{(j)}(0))_{1\le j\le 4}$ are dual right and left eigenbases of $\underline{c}_x\,\mathrm{I}_4+\underline{\mathbf{W}}$ associated with eigenvalues $(a_0^{(j)})_{1\le j\le 4}$ labeled so that*

$$\lambda_j(\xi)\stackrel{\xi\to 0}{=}\mathrm{i}\,\underline{k}_x\xi a_0^{(j)}+\mathcal{O}(|\xi|^3)\,,\quad 1\le j\le 4\,.$$

23. That is, $\langle f;g\rangle_{L^2}=\int_0^1\bar{f}\cdot g$.
24. With $\delta_\ell^j=1$ if $j=\ell$, and $\delta_\ell^j=0$ otherwise.

The way in which the eigenvalue 0 of multiplicity 4 breaks is highly non-generic from the point of view of abstract spectral theory. Indeed, we already know from Corollary 4.5 that the four arising eigenvalues are differentiable at $\xi=0$, and we obtain that when eigenvalues of $\underline{\mathbf{W}}$ are distinct, the four eigenvalues of $\mathcal{L}_{\xi,0}$ are analytic in ξ. This should be contrasted with the fact that eigenvalues arising from generic Jordan blocks of height 2 are no better than $\frac{1}{2}$-Hölder (and, in particular, are not Lipschitz).

Proof. — We make extensive use of standard spectral perturbation theory as expounded at length in [38]. To begin with, we introduce $\lambda_0>0$ and $\xi_0>0$, such that for $|\xi|\leq\xi_0$, the spectrum of $\mathcal{L}_{\xi,0}$ in $B(0,\lambda_0)$ has multiplicity 4 and denote by Π_ξ the corresponding Riesz spectral projector. From (74), the range of Π_0 is spanned by $(\underline{\mathcal{U}}_x,\mathbf{J}\,\underline{\mathcal{U}},\partial_q\underline{\mathcal{U}},\partial_m\underline{\mathcal{U}})$ and from (77),(81), we may choose a dual basis of the range of Π_0^* in the form $(*,**,\delta\mathbb{Q}_1(\underline{\mathcal{U}},\mathrm{e}_1(\underline{k}_x\partial_x+\underline{k}_\phi\mathbf{J})\underline{\mathcal{U}}),\,\delta\mathscr{M}[\underline{\mathcal{U}}])$. By Kato's perturbation method, we may extend these dual bases as dual bases $(q_j(\xi,\cdot\,))_{1\leq j\leq 4}$ and $(\widetilde{q}_j(\xi,\cdot))_{1\leq j\leq 4}$ of, respectively, the ranges of Π_ξ and Π_ξ^*.

One may use the corresponding coordinates to reduce the study of the spectrum of $\mathcal{L}_{\xi,0}$ to the consideration of the matrix

$$\Lambda_\xi:=\left(\langle\widetilde{q}_j(\xi,\cdot\,);\mathcal{L}_{\xi,0}\,q_\ell(\xi,\cdot\,)\rangle_{L^2}\right)_{(j,\ell)\in\{1,2,3,4\}^2}\,.$$

From relations expounded above stems

$$\Lambda_0=\begin{pmatrix}0&0&-\underline{k}_x\,\partial_q\underline{c}_x&-\underline{k}_x\,\partial_m\underline{c}_x\\ 0&0&\partial_q\underline{\omega}_\phi-\underline{k}_\phi\,\partial_q\underline{c}_x&\partial_m\underline{\omega}_\phi-\underline{k}_\phi\,\partial_m\underline{c}_x\\ 0&0&0&0\\ 0&0&0&0\end{pmatrix}.$$

Note, in particular, that Λ_0^2 is zero. From (76), we also derive

$$\langle\widetilde{q}_j(0,\cdot\,);\mathcal{L}_{(1)}\,q_\ell(0,\cdot\,)\rangle_{L^2}=0\,,\qquad 3\leq j\leq 4\,,\quad 1\leq\ell\leq 2\,.$$

Thus,

$$\widetilde{\Lambda}_\xi:=\frac{1}{\mathrm{i}\,\underline{k}_x\,\xi}P_\xi^{-1}\,\Lambda_\xi\,P_\xi\,,\qquad P_\xi:=\begin{pmatrix}\mathrm{I}_2&0\\ 0&\mathrm{i}\,\underline{k}_x\,\xi\,\mathrm{I}_2\end{pmatrix},\quad \xi\neq0 \tag{88}$$

defines a matrix $\widetilde{\Lambda}_\xi$ extending analytically to $\xi=0$.

Our main intermediate goal is to compute $\widetilde{\Lambda}_0$. We first show that we may enforce

$$\partial_\xi q_1(0,\cdot\,)=\mathrm{i}\,\underline{k}_x\,\partial_{k_x}\underline{\mathcal{U}},\qquad \partial_\xi q_2(0,\cdot\,)=\partial_{k_\phi}\underline{\mathcal{U}}\,. \tag{89}$$

To do so, for $\ell=1,2$, by expanding $\Pi_\xi(\mathcal{L}_{\xi,0}q_\ell(\xi,\cdot\,))=\mathcal{L}_{\xi,0}q_\ell(\xi,\cdot\,)$, we derive that $\mathcal{L}_{0,0}\partial_\xi q_\ell(0,\cdot\,)+\mathrm{i}\,\underline{k}_x\mathcal{L}_{(1)}q_\ell(0,\cdot\,)$ belongs to the range of Π_0. Comparing this with equations for $\partial_{k_x}\underline{\mathcal{U}}$ and $\partial_{k_\phi}\underline{\mathcal{U}}$ we deduce that $\mathcal{L}_{0,0}(\partial_\xi q_1(0,\cdot\,)-\mathrm{i}\,\underline{k}_x\partial_{k_x}\underline{\mathcal{U}})$

and $\mathcal{L}_{0,0}(\partial_\xi q_2(0,\cdot\,) - \mathrm{i}\,\underline{k}_x \partial_{k_\phi}\underline{\mathcal{U}})$, thus also $\partial_\xi q_1(0,\cdot\,) - \mathrm{i}\,\underline{k}_x \partial_{k_x}\underline{\mathcal{U}}$ and $\partial_\xi q_2(0,\cdot\,) - \mathrm{i}\,\underline{k}_x \partial_{k_\phi}\underline{\mathcal{U}}$ belong to the range of Π_0. Let $(\alpha_j^{(\ell)})_{1\leq j\leq 4,\,1\leq \ell\leq 2}$ be such that

$$\partial_\xi q_1(0,\cdot\,) - \mathrm{i}\,\underline{k}_x \partial_{k_x}\underline{\mathcal{U}} = \sum_{j=1}^{4} \alpha_j^{(1)} q_j(0,\cdot\,)\,,$$
$$\partial_\xi q_2(0,\cdot\,) - \mathrm{i}\,\underline{k}_x \partial_{k_\phi}\underline{\mathcal{U}} = \sum_{j=1}^{4} \alpha_j^{(2)} q_j(0,\cdot\,)\,.$$

Lessening ξ_0 if necessary, one may then replace $(q_j(\xi,\cdot\,))_{1\leq j\leq 4}$ with

$$q_1(\xi,\cdot\,) - \xi \sum_{j=1}^{4} \alpha_j^{(1)} q_j(\xi,\cdot\,)\,,\quad q_2(\xi,\cdot\,) - \xi \sum_{j=1}^{4} \alpha_j^{(2)} q_j(\xi,\cdot\,)\,,\quad q_3(\xi,\cdot\,)\,,\quad q_4(\xi,\cdot\,)\,,$$

and $(\widetilde{q}_j(\xi,\cdot\,))_{1\leq j\leq 4}$ with

$$\widetilde{q}_j(\xi,\cdot\,) + \xi \sum_{\ell=1}^{2} \widetilde{\alpha}_j^{(\ell)}(\xi)\widetilde{q}_\ell(\xi,\cdot\,)\,,\qquad j=1,2,3,4\,,$$

with $(\widetilde{\alpha}_j^{(\ell)}(\xi))_{1\leq j\leq 4,\,1\leq \ell\leq 2}$ tuned to preserve duality relations and have (89), which we assume from now on. To make the most of associated relations, we observe that from duality stems

$$\langle \partial_\xi \widetilde{q}_j(0,\cdot\,);\, q_\ell(0,\cdot\,)\rangle_{L^2} = -\langle \widetilde{q}_j(0,\cdot\,);\, \partial_\xi q_\ell(0,\cdot\,)\rangle_{L^2}\,,\qquad 1\leq j\,,\ell\leq 4\,.$$

Since

$$(\widetilde{\Lambda}_0)_{j,\ell} = \left\langle \widetilde{q}_j(0,\cdot\,); \frac{1}{\mathrm{i}\,\underline{k}_x}\mathcal{L}_{0,0}\,\partial_\xi\, q_\ell(0,\cdot\,) + \mathcal{L}_{(1)}\, q_\ell(0,\cdot\,)\right\rangle_{L^2},\ 1\leq j\leq 2\,,\ 1\leq \ell\leq 2\,,$$
$$(\widetilde{\Lambda}_0)_{j,\ell} = (\Lambda_0)_{j,\ell}\,,\qquad\qquad 1\leq j\leq 2\,,\ 3\leq \ell\leq 4\,,$$

from (76) this readily gives

$$\left((\widetilde{\Lambda}_0)_{j,\ell}\right)_{1\leq j\leq 2,\,1\leq \ell\leq 4} = \begin{pmatrix} -\underline{k}_x\,\partial_{k_x}\underline{c}_x & -\underline{k}_x\,\partial_{k_\phi}\underline{c}_x & -\underline{k}_x\,\partial_q\underline{c}_x & -\underline{k}_x\,\partial_m\underline{c}_x \\ \partial_{k_x}\underline{\omega}_\phi - \underline{k}_\phi\,\partial_{k_x}\underline{c}_x & \partial_{k_\phi}\underline{\omega}_\phi - \underline{k}_\phi\,\partial_{k_\phi}\underline{c}_x & \partial_q\underline{\omega}_\phi - \underline{k}_\phi\,\partial_q\underline{c}_x & \partial_m\underline{\omega}_\phi - \underline{k}_\phi\,\partial_m\underline{c}_x \end{pmatrix}.$$

The extra relations also carry

$$(\widetilde{\Lambda}_0)_{j,\ell} = \left\langle \widetilde{q}_j(0,\cdot\,); \frac{1}{\mathrm{i}\,\underline{k}_x}\mathcal{L}_{(1)}\,\partial_\xi\, q_\ell(0,\cdot\,) + \mathcal{L}_{(2)}\, q_\ell(0,\cdot\,)\right\rangle_{L^2}$$
$$+ (\underline{m}\,\partial_{k^{(\ell)}}\underline{\omega}_\phi - \underline{q}\partial_{k^{(\ell)}}\underline{c}_x)\,\delta_j^3\,,\qquad 3\leq j\leq 4\,,\ 1\leq \ell\leq 2\,,$$
$$(\widetilde{\Lambda}_0)_{j,\ell} = \left\langle \widetilde{q}_j(0,\cdot\,); \mathcal{L}_{(1)}\, q_\ell(0,\cdot\,)\right\rangle_{L^2}$$
$$+ (\underline{m}\,\partial_{m^{(\ell)}}\underline{\omega}_\phi - \underline{q}\partial_{m^{(\ell)}}\underline{c}_x)\,\delta_j^3\,,\qquad 3\leq j\leq 4\,,\ 3\leq \ell\leq 4\,,$$

with $k^{(1)} = k_x$, $k^{(2)} = k_\phi$, $m^{(3)} = q$, and $m^{(4)} = m$. Using (84) and (86) to evaluate the foregoing expressions leads to the final identification

$$\widetilde{\Lambda}_0 = \underline{c}_x \, \mathrm{I}_4 + \underline{\mathbf{W}} \, .$$

The proof is then completed by diagonalizing matrices $\widetilde{\Lambda}_\xi$, which have simple eigenvalues provided that ξ_0 is taken sufficiently small and undoing the various transformations. $\square$

We would like to make a few comments on the foregoing proof.

1. Although this is useless for our purposes, one may also compute explicitly $\widetilde{q}_1(0,\cdot\,)$ and $\widetilde{q}_2(0,\cdot\,)$ as combinations of $\mathbf{J}\underline{\mathcal{U}}_x$, $\underline{\mathcal{U}}$, $\mathbf{J}\partial_q\underline{\mathcal{U}}$ and $\mathbf{J}\partial_m\underline{\mathcal{U}}$. Indeed, it follows from Hamiltonian duality that the four vectors form a basis of the range of Π_0^*, and their scalar products with $\underline{\mathcal{U}}_x$, $\mathbf{J}\underline{\mathcal{U}}$, $\partial_q\underline{\mathcal{U}}$ and $\partial_m\underline{\mathcal{U}}$ are explicitly known.
2. The assumption that the eigenvalues of $\underline{\mathbf{W}}$ are distinct is only used at the very end of the proof. Removing it, the arguments still give an alternative proof of the second part of Corollary 4.5. For semilinear equations, this has already been carried out, to some extent, in the recent [40] with a few variations that we point out now.
   (a) The authors of [40] further assume that $\mathcal{L}_{0,0}$ exhibits two Jordan blocks of height 2 at 0; in other words, they assume that the above matrix Λ_0 has rank 2.
   (b) In [40], no formal interpretation is provided for the underlying instability criterion. In particular, no connection with geometrical optics and modulated systems is offered for the matrix $\widetilde{\Lambda}_0$. This connection was established in a recent preprint [20], building upon [40]. Hence, the next remarks also apply to [20].
   (c) The structure of eigenfunctions is left out of the discussion in [40], whereas this is our main motivation for reproving in a different way the second part of Corollary 4.5. In turn, the main focus of [40] is on spectral stability, and the authors supplement their analysis with numerical experiments for cubic and quintic semilinear equations.
   (d) We have taken advantage of the fact that we have already proven the first part of Corollary 4.5 to use modulation coordinates (k_x, k_ϕ, q, m), whereas the analysis in [40] is carried out with phase-portrait parameters $(\mu_x, c_x, \omega_\phi, \mu_\phi)$.

In the remaining part of this section, since we are only discussing longitudinal perturbations, we assume $d = 1$ for the sake of readability. Then, denoting by $(S(t))_{t\in\mathbf{R}}$ the group associated with the operator $\mathcal{L}$ on $L^2(\mathbf{R})$ and, for $\xi \in [-\pi,\pi]$, by $(S_\xi(t))_{t\in\mathbf{R}}$ the group associated with the operator $\mathcal{L}_\xi$ on $L^2_{\mathrm{per}}((0,1))$, we note that from Bloch inversion (54) stems

$$(S(t)g)(\mathbf{x}) = \int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}\,\xi x}\ (S_\xi(t)\check{g}(\xi,\cdot\,))(x)\ \mathrm{d}\,\xi\,.$$

Our main concern here is to analyze the large-time dynamics for the slow sideband part of the evolution

$$(S_{\mathrm{p}}(t)g)(\mathbf{x}) := \int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}\,\xi x}\ \chi(\xi)\,(S_\xi(t)\,\Pi_\xi\,\check{g}(\xi,\cdot\,))(x)\ \mathrm{d}\,\xi\,,$$

where χ is a smooth cut-off function equal to 1 on $[-\xi_0/2, \xi_0/2]$ and to 0 outside of $[-\xi_0, \xi_0]$, with $\xi_0 > 0$ as in the statement of Theorem 4.7 and Π_ξ the associated spectral projector, as in the proof of Theorem 4.7.

Let us explain in which sense this is expected to be the principal part of the linearized evolution for suitably spectrally stable waves. As a first remark we point out that when considering general perturbations on $\mathbf{R}$ (as opposed to coperiodic perturbations) we have to abandon not only stability in its strongest sense, which would require a control of $\|\mathbf{U} - \underline{\mathcal{U}}\|_X$ (in some functional space X of functions defined on $\mathbf{R}$), but also orbital stability, which here requires a control of

$$\inf_{(\varphi_\phi,\varphi_x)\in\mathbb{R}^2} \left\|\mathrm{e}^{-\varphi_\phi \mathbf{J}}\,\mathbf{U}(\,\cdot - \varphi_x\,) - \underline{\mathcal{U}}\right\|_X\,,$$

and, instead, to adopt the notion of *space-modulated stability* that is encoded by bounds on

$$\inf_{(\varphi_\phi,\varphi_x)\ \text{functions on }\mathbf{R}} \left(\left\|\mathrm{e}^{-\varphi_\phi(\cdot)\mathbf{J}}\,\mathbf{U}(\,\cdot - \varphi_x(\cdot)\,) - \underline{\mathcal{U}}\right\|_X + \|\partial_x\varphi_\phi\|_X + \|\partial_x\varphi_x\|_X\right).$$

Rather than bounding $\|\mathbf{V}\|_X$ or

$$\inf_{\substack{(\varphi_\phi,\varphi_x)\in\mathbb{R}^2\\ \mathbf{V}=\varphi_\phi\,\mathbf{J}\underline{\mathcal{U}}+\varphi_x\,\underline{\mathcal{U}}_x+\widetilde{\mathbf{V}}}} \|\widetilde{\mathbf{V}}\|_X\,,$$

at the linearized level, this consists in trying to bound

$$N_X(\mathbf{V}) := \inf_{\substack{(\varphi_\phi,\varphi_x)\ \text{functions on }\mathbf{R}\\ \mathbf{V}=\varphi_\phi\,\mathbf{J}\underline{\mathcal{U}}+\varphi_x\,\underline{\mathcal{U}}_x+\widetilde{\mathbf{V}}}} \left(\|\widetilde{\mathbf{V}}\|_X + \|\partial_x\varphi_\phi\|_X + \|\partial_x\varphi_x\|_X\right).$$

Note that N_X precisely quotients "locally" the unstable directions highlighted in (the proof of) Theorem 4.7, so that φ_ϕ, φ_x should be thought of as *local parameters*. Here we adapt to the case with a two-dimensional symmetry group the nonlinear notion formalized in [34] and its linearized counterpart introduced in [50]. Both notions have been proved to be sharp for a large class of parabolic systems in [34] and for the linearized Korteweg–de Vries equation in [50], respectively. The reader is also referred to [48, 49] for some more intuitive descriptions of the notions at hand.

Remark 4.8. — An incorrect choice of stability type would lead to a claim of instability in situations where the global shape is preserved but positions need to be resynchronized either uniformly in space in the orbitally stable case or in a slowly varying way in the space-modulated stable case. In the latter case, the underlying spurious growth is due to the presence of Jordan blocks in the spectrum, and this results in departures from the background profile that are algebraic in time (when no space-dependent synchronization is allowed). Thus, concluding to a *genuine* instability either at the linear or nonlinear requires extra care in the analysis. See, for instance, [22] for an example of the latter. Unfortunately, although it seems clear that some extra analysis could be carried out to fill this gap, the only general nonlinear instability result available so far [31] is expressed as an instability for the strongest sense of stability.

Following the lines of [50], one expects that for suitably spectrally stable waves the following bounds hold

$$\begin{aligned}
\|(S(t)-S_{\mathrm{p}}(t))\mathbf{V}_0\|_{H^s(\mathbf{R})} &\leq C_s\,N_{H^s(\mathbf{R})}(\mathbf{V}_0)\,, && t\in\mathbf{R}\,, s\in\mathbf{N}\,,\\
\|(S(t)-S_{\mathrm{p}}(t))\mathbf{V}_0\|_{L^\infty(\mathbf{R})} &\leq \frac{C}{|t|^{\frac12}}\,N_{L^1(\mathbf{R})}(\mathbf{V}_0)\,, && t\in\mathbf{R}
\end{aligned}$$

(with constants independent of $(t,\mathbf{V}_0)$). We shall not try to prove or even formulate more precisely the latter, but the reader should keep in mind the claimed $|t|^{-1/2}$ decay so as to compare it with bounds below. In particular, the conclusions of the next theorem contain that

$$N_{L^\infty(\mathbf{R})}(S_{\mathrm{p}}(t)\mathbf{V}_0)\leq \frac{C}{(1+|t|)^{\frac13}}\,N_{L^1(\mathbf{R})}(\mathbf{V}_0)\,,\qquad t\in\mathbf{R}\,.$$

Theorem 4.9 (Slow modulation behavior). — *Under the assumptions of Theorem 4.7 and with its set of notation, assume, moreover, that*

1. *for any* $\xi\in[-\xi_0,\xi_0]$, *for* $j\in\{1,2,3,4\}$, $\lambda_j(\xi)\in\mathrm{i}\,\mathbf{R}$;
2. *for* $j\in\{1,2,3,4\}$, $\partial_\xi^3\lambda_j(0)\neq 0$.

There exists $C>0$ *such that for any* $\mathbf{V}_0$ *such that* $N_{L^1(\mathbf{R})}(\mathbf{V}_0)<\infty$, *there exists local parameter functions* φ_x, φ_ϕ, $\mathtt{q}$ *and* $\mathtt{m}$ *such that for any time* $t\in\mathbf{R}$

$$\begin{aligned}
&\Big\|S_{\mathrm{p}}(t)\,(\mathbf{V}_0)-\varphi_x(t,\cdot\,)\,\underline{\mathcal{U}}_x-\varphi_\phi(t,\cdot\,)\,\mathbf{J}\underline{\mathcal{U}}\\
&\quad-\mathrm{d}_{k_x,k_\phi,\mathtt{q},\mathtt{m}}\,\underline{\mathcal{U}}\cdot(\underline{k}_x\partial_x\varphi_x(t,\cdot\,),\underline{k}_x\partial_x\varphi_\phi(t,\cdot\,),\mathtt{q}(t,\cdot\,),\mathtt{m}(t,\cdot\,))\Big\|_{L^\infty(\mathbf{R})}\\
&\qquad\leq \frac{C}{(1+|t|)^{\frac12}}\,N_{L^1(\mathbf{R})}(\mathbf{V}_0)\,,
\end{aligned}$$

where $\varphi_x(t,\cdot)$ *and* $\varphi_\phi(t,\cdot)$ *are centered,* $\varphi_x(t,\cdot)$, $\varphi_\phi(t,\cdot)$, $q(t,\cdot)$, *and* $m(t,\cdot)$ *are low frequency*[25], *and*

$$\|(\underline{k}_x\partial_x\varphi_x(t,\cdot),\underline{k}_x\partial_x\varphi_\phi(t,\cdot),q(t,\cdot),m(t,\cdot))\|_{L^\infty(\mathbf{R})} \leq \frac{C}{(1+|t|)^{\frac{1}{3}}}\,N_{L^1(\mathbf{R})}(\mathbf{V}_0)\,.$$

We omit the proof of Theorem 4.9 since with Theorem 4.7 in hand, the proof is identical to that of the corresponding result in [50]. Theorem 4.7 is the counterpart of [50, Proposition 2.1], while Theorem 4.9 is a low-frequency version of [50, Theorem 1.3] (which is why the decay factor is bounded at $t=0$); see, in particular, [50, Propositions 3.2 & 4.2]. Yet we would like to add some comments.

1. A choice of local parameters can be given explicitly:

$$\begin{aligned}\begin{pmatrix}\underline{k}_x\partial_x\varphi_x(t,\cdot)\\ \underline{k}_x\partial_x\varphi_\phi(t,\cdot)\\ q(t,\cdot)\\ m(t,\cdot)\end{pmatrix}(x) &= s^{\mathrm{p}}(t)(\mathbf{V}_0)(x)\\ &:= \sum_{j=1}^{4}\int_{-\pi}^{\pi}\mathrm{e}^{\mathrm{i}\,\xi x+\lambda_j(\xi)\,t}\,\chi(\xi)\,\beta^{(j)}(\xi)\,\langle\widetilde{\psi}_j(\xi,\cdot),\widecheck{\mathbf{V}}_0\,(\xi,\cdot)\rangle_{L^2}\,\mathrm{d}\,\xi\,.\end{aligned}$$

   This is motivated by the explicit diagonalization of $\mathcal{L}_\xi\Pi_\xi$ from Theorem 4.7, which implies

$$\begin{aligned}S_\xi(t)\,\Pi_\xi\,g =& \frac{1}{\mathrm{i}\,\underline{k}_x\xi}\left(\sum_{j=1}^{4}\mathrm{e}^{\lambda_j(\xi)\,t}\,\beta_1^{(j)}(\xi)\,\langle\widetilde{\psi}_j(\xi,\cdot),g\rangle_{L^2}\right)q_1(\xi,\cdot)\\ &+\frac{1}{\mathrm{i}\,\underline{k}_x\xi}\left(\sum_{j=1}^{4}\mathrm{e}^{\lambda_j(\xi)\,t}\,\beta_2^{(j)}(\xi)\,\langle\widetilde{\psi}_j(\xi,\cdot),g\rangle_{L^2}\right)q_2(\xi,\cdot)\\ &+\left(\sum_{j=1}^{4}\mathrm{e}^{\lambda_j(\xi)\,t}\,\beta_3^{(j)}(\xi)\,\langle\widetilde{\psi}_j(\xi,\cdot),g\rangle_{L^2}\right)q_3(\xi,\cdot)\\ &+\left(\sum_{j=1}^{4}\mathrm{e}^{\lambda_j(\xi)\,t}\,\beta_4^{(j)}(\xi)\,\langle\widetilde{\psi}_j(\xi,\cdot),g\rangle_{L^2}\right)q_4(\xi,\cdot)\,;\end{aligned}$$

   the choice of local parameters is then dictated by analyzing the various expressions (including remainders) arising from expansions with respect to ξ of $q_1(\xi,\cdot)$, $q_2(\xi,\cdot)$ at order 2 and $q_3(\xi,\cdot)$, $q_4(\xi,\cdot)$, at order 1. One may also replace χ with a cut-off function with support closer to the origin if required.

25. In the sense that their (distributional) Fourier transform has compact support that could be taken arbitrarily close to the origin.

2. Note that, since $(\partial_\xi\lambda_j(0))_{j\in\{1,2,3,4\}}$ are two-by-two distinct, assuming that for any $\xi\in[-\xi_0,\xi_0]$ and any j, $\lambda_j(\xi)\in \mathrm{i}\,\mathbf{R}$, from the Hamiltonian symmetry of the spectrum one derives that for $|\xi|\leq\xi_0$ (ξ_0 sufficiently small) and any $j\in\{1,2,3,4\}$ $\lambda_j(\xi)=\overline{\lambda_j(-\xi)}=-\lambda_j(-\xi)$. In particular, for any $j\in\{1,2,3,4\}$, $\lambda_j(\cdot)$ is an odd function, and thus $\partial_\xi^2\lambda_j(0)=0$. Therefore, the assumption that for $j\in\{1,2,3,4\}$, $\partial_\xi^3\lambda_j(0)\neq 0$ expresses that the dispersive effects on local parameters are as strong as possible. In contrast, the $|t|^{-1/2}$ decay claimed for the leftover part $S(t)-S_{\mathrm{p}}(t)$ is expected to be derivable from the assumption that outside the origin $(\lambda,\xi)=(0,0)$ second-order derivatives with respect to ξ of spectral curves do not vanish.
3. For the semilinear defocusing cubic Schrödinger equation, full spectral stability under longitudinal perturbations is known for all the waves, and we expect that the remaining assumptions may be checked by reliable elementary numerics by using explicit formula for the spectra obtained in [15].

Theorem 4.9 essentially proves that $S_{\mathrm{p}}(t)(\mathbf{V}_0)$ fits well with a large-time linearized version of the *ansatz* (65) with $\mathcal{U}_0(T,X;\cdot)$ being a periodic wave profile of parameters, such that $k_x=\partial_X\varphi_x$ and $k_\phi=\partial_X\varphi_\phi$. We now prove that some version of (67) drives the evolution of local parameters $(\underline{k}_x\partial_x\varphi_x,\underline{k}_x\partial_x\varphi_\phi,\mathpzc{q},\mathpzc{m})$ of Theorem 4.9. We need to modify (67) so as to account for dispersive effects.

Let P_0 diagonalize $\underline{\mathbf{W}}$ so that $P_0=\big(\beta^{(1)}(0)\ \beta^{(2)}(0)\ \beta^{(3)}(0)\ \beta^{(4)}(0)\big)$

$$P_0^{-1}=\begin{pmatrix}\widetilde{\beta}^{(1)}(0)\\ \widetilde{\beta}^{(2)}(0)\\ \widetilde{\beta}^{(3)}(0)\\ \widetilde{\beta}^{(4)}(0)\end{pmatrix},\quad P_0^{-1}\,\underline{\mathbf{W}}\,P_0=\mathrm{diag}(a_0^{(1)}-\underline{c}_x,a_0^{(2)}-\underline{c}_x,a_0^{(3)}-\underline{c}_x,a_0^{(4)}-\underline{c}_x)\,,$$

and define for q an integer

$$\begin{aligned}\underline{D}_q(\xi):=P_0\,\mathrm{diag}\,\big(&\lambda_1^{[q]}(\xi)-a_0^{(1)}\,\mathrm{i}\,\underline{k}_x\xi,\lambda_2^{[q]}(\xi)-a_0^{(2)}\,\mathrm{i}\,\underline{k}_x\xi,\\ &\lambda_3^{[q]}(\xi)-a_0^{(3)}\,\mathrm{i}\,\underline{k}_x\xi,\lambda_4^{[q]}(\xi)-a_0^{(4)}\,\mathrm{i}\,\underline{k}_x\xi\big)P_0^{-1}\,,\end{aligned}$$

where $\lambda_j^{[q]}(\xi)$ is the qth order Taylor expansion of $\lambda_j(\xi)$ near 0. By convention we also include the pseudo-differential case where $q=\infty$ by choosing $\lambda_j^{(\infty)}$ as a smooth purely imaginary-valued function that coincides with λ_j in a neighborhood of zero. Then consider the higher-order linearized modulation system

$$\partial_t\begin{pmatrix}k_x\\ k_\phi\\ \mathpzc{q}\\ \mathpzc{m}\end{pmatrix}=\underline{k}_x\ (\underline{\mathbf{W}}+\underline{c}_x\mathrm{I}_4)\,\partial_x\begin{pmatrix}k_x\\ k_\phi\\ \mathpzc{q}\\ \mathpzc{m}\end{pmatrix}+\underline{D}_q(\mathrm{i}^{-1}\,\partial_x)\begin{pmatrix}k_x\\ k_\phi\\ \mathpzc{q}\\ \mathpzc{m}\end{pmatrix}.\tag{90}$$

Note that when $q=3$, $\underline{D}_q(\mathrm{i}^{-1}\,\partial_x)$ takes the form $\underline{\mathbf{D}}_3(\underline{k}_x\partial_x)^3$, where $\underline{\mathbf{D}}_3$ is a real-valued matrix.

Theorem 4.10 (Averaged dynamics). — *Let q be an odd integer larger than 1, or $q=\infty$. Under the assumptions of Theorem 4.9, there exist $C>0$ and a cut-off function $\widetilde{\chi}$ such that for any $\mathbf{V}_0$ such that $N_{L^1(\mathbf{R})}(\mathbf{V}_0)<\infty$, there exist $(\varphi_x^{(0)},\varphi_\phi^{(0)})$ centered and low frequency such that with $\widetilde{\mathbf{V}}_0:=\mathbf{V}_0-\varphi_x^{(0)}\,\underline{\mathcal{U}}_x-\varphi_\phi^{(0)}\,\mathbf{J}\underline{\mathcal{U}}$*

$$\|\widetilde{\mathbf{V}}_0\|_{L^1(\mathbf{R})}+\|\partial_x\varphi_x^{(0)}\|_{L^1(\mathbf{R})}+\|\partial_x\varphi_\phi^{(0)}\|_{L^1(\mathbf{R})}\;\leq\;2\,N_{L^1(\mathbf{R})}(\mathbf{V}_0)\,,$$

and for any such $(\varphi_x^{(0)},\varphi_\phi^{(0)})$, the local parameters $(\underline{k}_x\partial_x\varphi_x,\underline{k}_x\partial_x\varphi_\phi,\mathscr{q},\mathscr{m})$ of Theorem 4.9 may be chosen in such a way that with

$$\begin{pmatrix}k_x^{(0)}\\ k_\phi^{(0)}\\ \mathscr{q}^{(0)}\\ \mathscr{m}^{(0)}\end{pmatrix}:=\widetilde{\chi}(\mathrm{i}^{-1}\,\partial_x)\begin{pmatrix}\underline{k}_x\partial_x\varphi_x^{(0)}\\ \underline{k}_x\partial_x\varphi_\phi^{(0)}\\ \delta\mathbb{Q}(\underline{\mathcal{U}},\underline{k}_x\partial_x\underline{\mathcal{U}}+\underline{k}_\phi\mathbf{J}\underline{\mathcal{U}})\,\widetilde{\mathbf{V}}_0\\ \delta\mathscr{M}[\underline{\mathcal{U}}]\,\widetilde{\mathbf{V}}_0\end{pmatrix},$$

for any time $t\in\mathbf{R}$

$$\begin{aligned}&\Big\|(\underline{k}_x\partial_x\varphi_x(t,\cdot),\underline{k}_x\partial_x\varphi_\phi(t,\cdot),\mathscr{q}(t,\cdot),\mathscr{m}(t,\cdot))\\&\quad-\Sigma_q^W(t)(k_x^{(0)},k_\phi^{(0)},\mathscr{q}^{(0)},\mathscr{m}^{(0)})\Big\|_{L^\infty(\mathbf{R})}\\&\qquad\leq\frac{C}{(1+|t|)^{\frac{q+1}{2(q+2)}}}\,N_{L^1(\mathbf{R})}(\mathbf{V}_0)\,,\end{aligned}$$

and

$$\begin{aligned}&\left\|(\varphi_x(t,\cdot),\varphi_\phi(t,\cdot))-\begin{pmatrix}\mathbf{e}_1\\ \mathbf{e}_2\end{pmatrix}\cdot\,(\underline{k}_x\partial_x)^{-1}\Sigma_q^W(t)(k_x^{(0)},k_\phi^{(0)},\mathscr{q}^{(0)},\mathscr{m}^{(0)})\right\|_{L^\infty(\mathbf{R})}\\&\qquad\leq C\,N_{L^1(\mathbf{R})}(\mathbf{V}_0)\begin{cases}\dfrac{1}{(1+|t|)^{\frac13}} & \textit{if } q\geq 5\\[2mm] \dfrac{1}{(1+|t|)^{\frac15}} & \textit{if } q=3\end{cases},\end{aligned}$$

where Σ_q^W is the solution operator to System (90).

Again, we omit the proof as nearly identical to the one of the corresponding result in [50], namely Theorems 1.4 ($q=3$) and 1.5 ($q>3$), but provide a few comments.

1. In the case $q=3$, one may drop the low-frequency cut-off $\widetilde{\chi}(\mathrm{i}^{-1}\,\partial_x)$ (provided that one restricts to times $|t|\geq 1$) since it is here only to compensate for the fact that when $q\geq 5$, one cannot infer good dispersive

properties of the Taylor expansions globally in frequency. This is somehow analogous to the fact that slow expansions of well-behaved parabolic systems may produce ill-posed systems. Similar estimates hold for $q=1$ but are somewhat pointless since the decay rate is then the same as in Theorem 4.9.

2. If one is willing to use a less explicit and pleasant formula for $(k_x^{(0)}, k_\phi^{(0)}, \mathscr{q}^{(0)}, \mathscr{m}^{(0)})$, then the description of $(\varphi_x(t,\cdot), \varphi_\phi(t,\cdot))$ may actually be achieved up to an error of size $(1+|t|)^{-\frac{1}{3}}$ if $q=3$ and $(1+|t|)^{-\frac{q-1}{2(q+2)}}$, if $q\geq 5$; see [50, Theorems 1.5 & 1.6].
3. If one removes the assumption that $(\varphi_x^{(0)}, \varphi_\phi^{(0)})$ is low frequency, then the formula for $(k_x^{(0)}, k_\phi^{(0)}, \mathscr{q}^{(0)}, \mathscr{m}^{(0)})$ should be modified as

$$\begin{pmatrix} k_x^{(0)} \\ k_\phi^{(0)} \\ \mathscr{q}^{(0)} \\ \mathscr{m}^{(0)} \end{pmatrix} := \widetilde{\chi}(\mathrm{i}^{-1}\,\partial_x) \\ \cdot \begin{pmatrix} \underline{k}_x \partial_x \varphi_x^{(0)} \\ \underline{k}_x \partial_x \varphi_\phi^{(0)} \\ \delta\mathbb{Q}(\underline{\mathcal{U}}, \underline{k}_x \partial_x \underline{\mathcal{U}} + \underline{k}_\phi \mathbf{J}\underline{\mathcal{U}})\, \widetilde{\mathbf{V}}_0 - (\mathscr{M}[\underline{\mathcal{U}}] - \underline{\mathscr{m}})\big(\underline{k}_\phi \partial_x \varphi_x^{(0)} - \underline{k}_x \partial_x \varphi_\phi^{(0)}\big) \\ \delta\mathscr{M}[\underline{\mathcal{U}}]\, \widetilde{\mathbf{V}}_0 - (\mathscr{M}[\underline{\mathcal{U}}] - \underline{\mathscr{m}})\partial_x \varphi_x^{(0)} \end{pmatrix}.$$

To illustrate how the high-frequency corrections arise let us point out that

$$\langle \delta\mathscr{M}[\underline{\mathcal{U}}]; \widetilde{\varphi_x^{(0)} \underline{\mathcal{U}}_x}(\xi,\cdot)\rangle_{L^2} = -[\widehat{(\mathscr{M}[\underline{\mathcal{U}}] - \underline{\mathscr{m}})\, \partial_x \varphi_x^{(0)}}](\xi) \\ + \mathrm{i}\,\xi\, \langle \mathscr{M}[\underline{\mathcal{U}}]; \widetilde{\varphi_x^{(0)}}(\xi,\cdot) - \widehat{\varphi_x^{(0)}}(\xi)\rangle_{L^2}\,,$$

and that extra ξ-factors bring extra decay.

4. We expect that in the case $q=3$, System (90) could be derived from higher-order versions of geometrical optics as in [44]. In contrast, the formal derivation of either System (90) in the general case or of effective data $(k_x^{(0)}, k_\phi^{(0)}, \mathscr{q}^{(0)}, \mathscr{m}^{(0)})$ (in particular when $(\varphi_x^{(0)}, \varphi_\phi^{(0)})$ is not low frequency) seems out of reach.
5. The foregoing construction of D_q closely follows the classical construction of artificial viscosity systems as large-time asymptotic equivalents to systems that are only parabolic in the hypocoercive sense of Kawashima. We refer the reader to, for instance, [29, Section 6], [47], [34, Appendix B], or [48, Appendix A] for a description of the latter. A notable

difference, however, is that in the diffusive context, higher-order expansions of dispersion relations beyond the second-order necessary to capture some dissipation do not provide any sharper description since the second-order expansion already provides the maximal rate compatible with a first-order expansion of eigenvectors. Here, one needs to use the full pseudo-differential dispersion relations so as to reach a description up to $\mathcal{O}(|t|^{-1/2})$ error terms.

6. We infer from Theorems 4.9 and 4.10 that at leading order, the behavior of $S_{\mathrm{p}}(t)(\mathbf{V}_0)$ is captured by a linear modulation of phases $\varphi_x(t,\cdot\,)\,\underline{\mathcal{U}}_x + \varphi_\phi(t,\cdot\,)\,\mathbf{J}\,\underline{\mathcal{U}}$, and the phase shifts (φ_x,φ_ϕ) are the antiderivative of the two first components of a four-dimensional vector $(\underline{k}_x\partial_x\varphi_x,\underline{k}_x\partial_x\varphi_\phi,\mathscr{q},\mathscr{m})$ that itself is at leading-order a sum of four linear dispersive waves of Airy type, each one traveling with its own velocity. In particular, three scales coexist: the oscillation of the background wave at scale 1 in $\underline{\mathcal{U}}_x$ and $\mathbf{J}\,\underline{\mathcal{U}}$, spatial separation of the four dispersive waves at linear hyperbolic scale t, the width of Airy waves of size $t^{1/3}$. We refer the reader to [2, 50] for an enlightening illustration by direct simulations of similar multiscale, large-time dynamics.

5. General perturbations

We now come back to the general spectral stability problem. We begin with a corollary to Theorem 3.2, from which we recall the following key formula. The Evans function $D_\xi(\lambda,\boldsymbol{\eta})$ expands as

$$\begin{aligned}D_\xi(\lambda,\boldsymbol{\eta}) &\overset{(\lambda,\boldsymbol{\eta})\to(0,0)}{=} \det\left(\lambda\,\Sigma_t-(\mathrm{e}^{\mathrm{i}\,\xi}-1)\mathrm{I}_4+\tfrac{\|\boldsymbol{\eta}\|^2}{\lambda}\Sigma_{\mathbf{y}}\right)\\ &\quad+\mathcal{O}\left((|\lambda|+|\xi|+\|\boldsymbol{\eta}\|^2)\,(|\lambda|^2+|\xi|^2+\|\boldsymbol{\eta}\|^2)\,(|\lambda|(|\lambda|+|\xi|)+\|\boldsymbol{\eta}\|^2)\right).\end{aligned}$$

For later use, we introduce the homogeneous fourth-order polynomial with real coefficients[26]

$$\Delta_0(\lambda,z,\zeta):=\det\left(\lambda\,\Sigma_t-z\,\mathrm{I}_4+\tfrac{\zeta^2}{\lambda}\Sigma_{\mathbf{y}}\right). \tag{91}$$

That the coefficients of Δ_0 are real may be seen directly or related to the fact that $D_0(\lambda,\boldsymbol{\eta})$ is real when λ and $\boldsymbol{\eta}$ are real. Likewise, note that for any $(\varepsilon,\lambda,z,\zeta)\in\mathbf{C}^4$

$$\begin{aligned}\Delta_0(\lambda,z,-\zeta)&=\Delta_0(\lambda,z,\zeta)\,, &\qquad \Delta_0(-\lambda,-z,\zeta)&=\Delta_0(\lambda,z,\zeta)\,,\\ \Delta_0(\varepsilon\lambda,\varepsilon z,\varepsilon\zeta)&=\varepsilon^4\Delta_0(\lambda,z,\zeta)\,, &\qquad \Delta_0(\overline{\lambda},\overline{z},\overline{\zeta})&=\overline{\Delta_0(\lambda,z,\zeta)}\,.\end{aligned}$$

The second and the fourth relations are inherited from original real and Hamiltonian symmetries.

26. The formula being extended by continuity to incorporate the cases when $\lambda=0$.

Since the longitudinal perturbations have already been analyzed at length, the following corollary focuses on perturbations that do have a transverse component.

Corollary 5.1. — *Consider an unscaled wave profile* $\underline{\mathcal{V}}$ *such that* $\underline{\mathcal{V}} \cdot \underline{\mathcal{V}}_x \not\equiv 0$.

1. *Assume that there exists* $(\lambda_0, \zeta_0) \in \mathbf{R}^2$, *such that* $\zeta_0 \neq 0$ *and* $\Delta_0(\lambda_0, 0, \zeta_0) < 0$. *Then the wave is spectrally exponentially unstable to perturbations that are longitudinally coperiodic and transversally arbitrarily slow, that is,* $\mathcal{L}_{0,\boldsymbol{\eta}}$ *has eigenvalues of positive real part for* $\boldsymbol{\eta}$ *arbitrarily small but nonzero.*
2. *Assume that there exists* $(\lambda_0, \xi_0, \zeta_0) \in \mathbf{C} \times \mathbf{R}^2$, *such that* $\zeta_0 \neq 0$ *and* λ_0 *is a root of* $\Delta_0(\,\cdot\,, \mathrm{i}\,\xi_0, \zeta_0)$ *of algebraic multiplicity* r. *Then there exist* C_0 *and* $\eta_0 > 0$, *such that for any* $\boldsymbol{\eta}$ *such that* $0 < \|\boldsymbol{\eta}\| \leq \eta_0$,
$$\mathcal{L}_{\frac{\|\boldsymbol{\eta}\|}{|\zeta_0|}\xi_0, \boldsymbol{\eta}}$$
possesses r *eigenvalues (counted with algebraic multiplicity) in the disk*
$$B\left(\frac{\|\boldsymbol{\eta}\|}{|\zeta_0|}\lambda_0\,,\, C_0\|\boldsymbol{\eta}\|^{1+\frac{1}{r}}\right).$$
In particular if $\Re(\lambda_0) \neq 0$, *then the wave is spectrally unstable.*

Proof. — By using symmetries of Δ_0 we may assume that $\lambda_0 \geq 0$ and $r_0 = 1$. Then since Theorem 3.2 ensures
$$D_0(\|\boldsymbol{\eta}\|\,\lambda_0, \boldsymbol{\eta}) \stackrel{\|\boldsymbol{\eta}\| \to 0}{=} \|\boldsymbol{\eta}\|^4\,\Delta_0(\lambda_0, 1) + \mathcal{O}(\|\boldsymbol{\eta}\|^5)\,,$$
we deduce that there exists $\boldsymbol{\eta}_0 > 0$, such that for any $0 < \|\boldsymbol{\eta}\| \leq \boldsymbol{\eta}_0$, $D_0(\|\boldsymbol{\eta}\|\,\lambda_0, \boldsymbol{\eta}) < 0$. Comparing this with Proposition 3.1 yields that when $0 < \|\boldsymbol{\eta}\| \leq \boldsymbol{\eta}_0$, the spectrum of $\mathcal{L}_{0,\boldsymbol{\eta}}$ intersects $(0, +\infty)$.

The second part stems from a counting root argument based directly on Theorem 3.2 and the symmetries of Δ_0 . □

At this stage, two more comments are worth stating.

1. Since both Theorem 3.2 and Proposition 3.1 include the case $\xi = \pi$ besides the case $\xi = 0$, one may obtain a $\xi = \pi$ counterpart to the first parts of Theorem 4.1 and Corollary 5.1. Yet the corresponding instability criteria are never met since $\Delta_0(0, -2, 0) > 0$. For a similar reason, although Theorem 3.2 deals with arbitrary Floquet exponents ξ, the second parts of Corollaries 4.5 and 5.1 involve Floquet exponents converging to $\xi = 0$. This is due to the fact that $\Delta_0(0, (\mathrm{e}^{\mathrm{i}\,\xi} - 1), 0) = 0$ if and only if $\xi \in 2\pi\mathbf{Z}$.

2. We point out that to some extent the restriction to $\xi_0 = 0$ of the second part of the corollary has already been derived in the recent [40] with a few variations that we point out now.
   (a) The authors restrict themselves to semilinear equations, a fact that comes with quite a few algebraic simplifications in computations.
   (b) They further assume that $\mathcal{L}_{0,0}$ exhibits two Jordan blocks of height 2 at 0.
   (c) Their proof goes by direct spectral perturbation of $\mathcal{L}_{0,0}$ rather than Evans function computations.

5.1. Geometrical optics. — Prior to studying at length properties of Δ_0, we show that the latter may be derived from a suitable version of geometrical optics *à la* Whitham. To start bridging the gap, we first recall (62) and observe that

(92)
$$\lambda\,\Sigma_t - z\mathrm{I}_4 + \frac{\zeta^2}{\lambda}\Sigma_{\mathbf{y}} = \begin{pmatrix} 0&0&0&1\\ 1&0&0&0\\ 0&0&1&0\\ 0&1&0&0 \end{pmatrix} \left(\lambda\,\mathbf{A}_0 \text{ Hess }\Theta - z\,\mathbf{B}_0 + \frac{\zeta^2}{\lambda}\,\mathbf{C}_0\right) \begin{pmatrix} 0&0&0&1\\ 0&1&0&0\\ 1&0&0&0\\ 0&0&1&0 \end{pmatrix},$$

with

$$\mathbf{A}_0 = \begin{pmatrix} \mathrm{I}_2 & 0\\ 0 & -\mathrm{I}_2 \end{pmatrix}, \qquad \mathbf{B}_0 = \begin{pmatrix} 0&1&0&0\\ 1&0&0&0\\ 0&0&0&1\\ 0&0&1&0 \end{pmatrix}, \qquad \mathbf{C}_0 := \begin{pmatrix} 0&0&0&0\\ 0&-\sigma_3&\sigma_2&0\\ 0&-\sigma_2&\sigma_1&0\\ 0&0&0&0 \end{pmatrix},$$

where

$$\sigma_1 := \int_0^{X_x} \kappa(\|\underline{\mathcal{V}}\|^2)\,\|\underline{\mathcal{V}}\|^2\,, \qquad \sigma_2 := \int_0^{X_x} \kappa(\|\underline{\mathcal{V}}\|^2)\,\mathbf{J}\underline{\mathcal{V}}\cdot\underline{\mathcal{V}}_x\,,$$
$$\sigma_3 := \int_0^{X_x} \kappa(\|\underline{\mathcal{V}}\|^2)\,\|\underline{\mathcal{V}}_x\|^2\,. \tag{93}$$

In particular

$$\Delta_0(\lambda,z,\zeta) = \det\left(\lambda\,\mathbf{A}_0 \text{ Hess }\Theta - z\,\mathbf{B}_0 + \frac{\zeta^2}{\lambda}\,\mathbf{C}_0\right). \tag{94}$$

Now let us start formal asymptotics with a multidimensional *ansatz* similar to (65)

$$\mathbf{U}^{(\varepsilon)}(t,\mathbf{x}) = \mathrm{e}^{\frac{1}{\varepsilon}\varphi_\phi^{(\varepsilon)}(\varepsilon\,t,\varepsilon\,\mathbf{x})\,\mathbf{J}}\,\mathcal{U}^{(\varepsilon)}\left(\varepsilon\,t,\varepsilon\,\mathbf{x};\,\frac{\varphi_x^{(\varepsilon)}(\varepsilon\,t,\varepsilon\,\mathbf{x})}{\varepsilon}\right), \tag{95}$$

with

$$\begin{aligned}\mathcal{U}^{(\varepsilon)}(T,\mathbf{X};\zeta) &= \mathcal{U}_0(T,\mathbf{X};\zeta) + \varepsilon\,\mathcal{U}_1(T,\mathbf{X};\zeta) + o(\varepsilon)\,,\\ \varphi_\phi^{(\varepsilon)}(T,\mathbf{X}) &= (\varphi_\phi)_0(T,\mathbf{X}) + \varepsilon\,(\varphi_\phi)_1(T,\mathbf{X}) + o(\varepsilon)\,,\\ \varphi_x^{(\varepsilon)}(T,\mathbf{X}) &= (\varphi_x)_0(T,\mathbf{X}) + \varepsilon\,(\varphi_x)_1(T,\mathbf{X}) + o(\varepsilon)\,,\end{aligned}$$

with $\mathcal{U}_0(T,\mathbf{X};\cdot)$ and $\mathcal{U}_1(T,\mathbf{X};\cdot)$ 1-periodic. Inserting (95) in (3) yields at leading order

$$(\partial_T(\varphi_\phi)_0\mathbf{J} + \partial_T(\varphi_x)_0\partial_\zeta)\mathcal{U}_0 = \mathbf{J}\delta\mathcal{H}_0\left(\mathcal{U}_0, (\nabla_{\mathbf{X}}(\varphi_\phi)_0\mathbf{J} + \nabla_{\mathbf{X}}(\varphi_x)_0\partial_\zeta)\,\mathcal{U}_0\right),$$

so that for each $(T,\mathbf{X})$, $\mathcal{U}_0(T,\mathbf{X};\cdot)$ must be a wave profile as in (15), such that

$$\partial_T(\varphi_\phi)_0 = \omega_\phi - c_x k_\phi\,,\quad \nabla_{\mathbf{X}}(\varphi_\phi)_0 = \mathbf{k}_\phi\,,\quad \partial_T(\varphi_x)_0 = \omega_x\,,\quad \nabla_{\mathbf{X}}(\varphi_x)_0 = \mathbf{k}_x\,.$$

As a consequence, $\mathbf{k}_x$ and $\mathbf{k}_\phi$ are curl free and

$$\partial_T\mathbf{k}_\phi = \nabla_{\mathbf{X}}\left(\omega_\phi - k_\phi\,c_x\right),\qquad \partial_T\mathbf{k}_x = \nabla_{\mathbf{X}}\omega_x\,. \tag{96}$$

Moreover, inserting (95) in (4) and (6) yields at leading order, respectively,

$$\partial_T(\mathscr{M}(\mathcal{U}_0)) = \mathrm{div}_{\mathbf{X}}\left(\mathbf{J}\mathcal{U}_0\cdot\nabla_{\mathbf{U_x}}\mathscr{H}_0\left(\mathcal{U}_0, (\mathbf{k}_\phi\mathbf{J} + \mathbf{k}_x\partial_\zeta)\,\mathcal{U}_0\right)\right) + \partial_\zeta(*)$$

and

$$\begin{aligned}&\partial_T\Big(\mathbb{Q}(\mathcal{U}_0, (\mathbf{k}_\phi\mathbf{J} + \mathbf{k}_x\partial_\zeta)\,\mathcal{U}_0)\Big)\\ &\quad= \nabla_{\mathbf{X}}\left(\frac{1}{2}\mathbf{J}\mathcal{U}_0\cdot\mathbf{J}\delta\mathscr{H}_0(\mathcal{U}_0, (\mathbf{k}_\phi\mathbf{J} + \mathbf{k}_x\partial_\zeta)\,\mathcal{U}_0) - \mathscr{H}_0(\mathcal{U}_0, (\mathbf{k}_\phi\mathbf{J} + \mathbf{k}_x\partial_\zeta)\,\mathcal{U}_0)\right)\\ &\qquad+ \sum_\ell \partial_{X_\ell}\Big(\mathbf{J}\delta\,\mathbb{Q}(\mathcal{U}_0, (\mathbf{k}_\phi\mathbf{J} + \mathbf{k}_x\partial_\zeta)\,\mathcal{U}_0)\cdot\nabla_{\mathbf{U}_{X_\ell}}\mathscr{H}_0(\mathcal{U}_0, (\mathbf{k}_\phi\mathbf{J} + \mathbf{k}_x\partial_\zeta)\,\mathcal{U}_0)\Big)\\ &\qquad+ \partial_\zeta(**)\,,\end{aligned}$$

with $*$ and $**$ 1-periodic in ζ. Averaging the foregoing equations using formulas in Section 2.6 provides equations that combined with (96) yield

$$\begin{cases}\partial_T\mathbf{k}_x = \nabla_{\mathbf{X}}\omega_x\\ \partial_T\boldsymbol{q} = \nabla_{\mathbf{X}}\left(\mu_x - c_x\boldsymbol{q} + \dfrac{1}{2}\,\tau_0\,\|\widetilde{\mathbf{k}}_\phi\|^2\right)\\ \qquad\qquad + \mathrm{div}_{\mathbf{X}}\left(\tau_1\,\widetilde{\mathbf{k}}_\phi\otimes\widetilde{\mathbf{k}}_\phi + \tau_2\,(\widetilde{\mathbf{k}}_\phi\otimes\mathbf{e}_x + \mathbf{e}_x\otimes\widetilde{\mathbf{k}}_\phi) + \tau_3\,\left(\mathbf{e}_x\otimes\mathbf{e}_x - \mathrm{I}_d\right)\right)\\ \partial_T m = \mathrm{div}_{\mathbf{X}}\left((\mu_\phi - c_x m)\,\mathbf{e}_x + \tau_1\,\widetilde{\mathbf{k}}_\phi\right)\\ \partial_T\mathbf{k}_\phi = \nabla_{\mathbf{X}}\left(\omega_\phi - c_x\,k_\phi\right)\end{cases}, \tag{97}$$

with curl-free $\mathbf{k}_x$ and $\mathbf{k}_\phi$, where

$$\tau_0 := \langle \kappa'(\|\mathcal{U}_0\|^2)\,\|\mathcal{U}_0\|^2\rangle\,, \qquad \tau_1 := \langle \kappa(\|\mathcal{U}_0\|^2)\,\|\mathcal{U}_0\|^2\rangle\,,$$
$$\tau_2 := \langle \kappa(\|\mathcal{U}_0\|^2)\,\mathbf{J}\mathcal{U}_0\cdot(k_\phi\,\mathbf{J}+k_x\partial_\zeta)\mathcal{U}_0\rangle\,, \quad \tau_3 := \langle \kappa(\|\mathcal{U}_0\|^2)\,\|(k_\phi\,\mathbf{J}+k_x\partial_\zeta)\mathcal{U}_0\|^2\rangle\,.$$

Before linearizing, in order to prepare System 97 for comparison, we recall that $\mathbf{k}_x = k_x\ \mathbf{e}_x$, $\mathbf{k}_\phi = k_\phi\ \mathbf{e}_x + \widetilde{\mathbf{k}}_\phi$, and $\boldsymbol{q} = q\ \mathbf{e}_x + m\,\widetilde{\mathbf{k}}_\phi$, with $\mathbf{e}_x$ unitary and $\widetilde{\mathbf{k}}_\phi$ orthogonal to $\mathbf{e}_x$ and write (97) in terms of $(k_x, k_\phi, m, q, \mathbf{e}_x, \widetilde{\mathbf{k}}_\phi)$. By using that, for any derivative $\partial_\sharp$,

$$\mathbf{e}_x\cdot\partial_\sharp\,\mathbf{e}_x = 0\,,\quad \mathbf{e}_x\cdot\partial_\sharp\widetilde{\mathbf{k}}_\phi = -\widetilde{\mathbf{k}}_\phi\cdot\partial_\sharp\,\mathbf{e}_x\,, \tag{98}$$

this yields

(99)
$$\begin{cases}
\partial_T\,\mathbf{e}_x = \dfrac{1}{k_x}(\nabla_{\mathbf{X}} - \mathbf{e}_x\ \mathbf{e}_x\cdot\nabla_{\mathbf{X}})\omega_x \\
\partial_T\widetilde{\mathbf{k}}_\phi = (\nabla_{\mathbf{X}} - \mathbf{e}_x\ \mathbf{e}_x\cdot\nabla_{\mathbf{X}})\,(\omega_\phi - c_x\,k_\phi) - \dfrac{k_\phi}{k_x}(\nabla_{\mathbf{X}} - \mathbf{e}_x\ \mathbf{e}_x\cdot\nabla_{\mathbf{X}})\omega_x \\
\qquad\qquad - \mathbf{e}_x\ \dfrac{\widetilde{\mathbf{k}}_\phi}{k_x}\cdot\nabla_{\mathbf{X}}\omega_x \\
\partial_T k_x = \mathbf{e}_x\cdot\nabla_{\mathbf{X}}\omega_x \\
\partial_T q = \mathbf{e}_x\cdot\nabla_{\mathbf{X}}\left(\mu_x - c_x q + \dfrac{1}{2}\,\tau_0\,\|\widetilde{\mathbf{k}}_\phi\|^2\right) + m\,\dfrac{\widetilde{\mathbf{k}}_\phi}{k_x}\cdot\nabla_{\mathbf{X}}\omega_x \\
\qquad\qquad + \mathbf{e}_x\cdot\operatorname{div}_{\mathbf{X}}\left(\tau_1\,\widetilde{\mathbf{k}}_\phi\otimes\widetilde{\mathbf{k}}_\phi + \tau_2\,(\widetilde{\mathbf{k}}_\phi\otimes\mathbf{e}_x + \mathbf{e}_x\otimes\widetilde{\mathbf{k}}_\phi) + \tau_3\ (\mathbf{e}_x\otimes\mathbf{e}_x - \mathrm{I}_d)\right) \\
\partial_T m = \operatorname{div}_{\mathbf{X}}\left((\mu_\phi - c_x m)\ \mathbf{e}_x + \tau_1\,\widetilde{\mathbf{k}}_\phi\right) \\
\partial_T k_\phi = \mathbf{e}_x\cdot\nabla_{\mathbf{X}}\,(\omega_\phi - c_x\,k_\phi) + \dfrac{\widetilde{\mathbf{k}}_\phi}{k_x}\cdot\nabla_{\mathbf{X}}\omega_x
\end{cases}$$

with $\mathbf{e}_x$ unitary, $\widetilde{\mathbf{k}}_\phi$ orthogonal to $\mathbf{e}_x$, and $k_x\ \mathbf{e}_x$ and $k_\phi\ \mathbf{e}_x + \widetilde{\mathbf{k}}_\phi$ curl free. System (99) may be simplified further by noticing that from curl-free conditions (and (98)) stem

$$(\nabla_{\mathbf{X}} - \mathbf{e}_x\ \mathbf{e}_x\cdot\nabla_{\mathbf{X}})k_x = k_x\,(\mathbf{e}_x\cdot\nabla_{\mathbf{X}})\,\mathbf{e}_x\,,$$
$$(\nabla_{\mathbf{X}} - \mathbf{e}_x\ \mathbf{e}_x\cdot\nabla_{\mathbf{X}})k_\phi = (\mathbf{e}_x\cdot\nabla_{\mathbf{X}})\widetilde{\mathbf{k}}_\phi + \nabla_{\mathbf{X}}\left({\mathbf{e}_x}^{\mathsf{T}}\right)\widetilde{\mathbf{k}}_\phi + k_\phi\,(\mathbf{e}_x\cdot\nabla_{\mathbf{X}})\,\mathbf{e}_x\ .$$

This leaves as an equivalent system

(100)
$$\begin{cases}(\partial_T+c_x(\mathbf{e}_x\cdot\nabla_{\mathbf{X}}))\,\mathbf{e}_x=-(\nabla_{\mathbf{X}}-\mathbf{e}_x\ \mathbf{e}_x\cdot\nabla_{\mathbf{X}})c_x\\ (\partial_T+c_x(\mathbf{e}_x\cdot\nabla_{\mathbf{X}}))\widetilde{\mathbf{k}}_\phi=(\nabla_{\mathbf{X}}-\mathbf{e}_x\ \mathbf{e}_x\cdot\nabla_{\mathbf{X}})\omega_\phi-\mathbf{e}_x\ \dfrac{\widetilde{\mathbf{k}}_\phi}{k_x}\cdot\nabla_{\mathbf{X}}\omega_x\\ \qquad\qquad -c_x\nabla_{\mathbf{X}}\left(\mathbf{e}_x{}^{\mathsf{T}}\right)\widetilde{\mathbf{k}}_\phi\\ \partial_T k_x=\mathbf{e}_x\cdot\nabla_{\mathbf{X}}\omega_x\\ \partial_T \mathscr{q}=\mathbf{e}_x\cdot\nabla_{\mathbf{X}}\left(\mu_x-c_x\mathscr{q}+\dfrac{1}{2}\,\tau_0\,\|\widetilde{\mathbf{k}}_\phi\|^2\right)+\mathscr{m}\ \dfrac{\widetilde{\mathbf{k}}_\phi}{k_x}\cdot\nabla_{\mathbf{X}}\omega_x\\ \qquad +\mathbf{e}_x\cdot\mathrm{div}_{\mathbf{X}}\Big(\tau_1\,\widetilde{\mathbf{k}}_\phi\otimes\widetilde{\mathbf{k}}_\phi+\tau_2\,(\widetilde{\mathbf{k}}_\phi\otimes\mathbf{e}_x+\mathbf{e}_x\otimes\widetilde{\mathbf{k}}_\phi)\\ \qquad\qquad +\tau_3\ (\mathbf{e}_x\otimes\mathbf{e}_x-\mathrm{I}_d)\Big)\\ \partial_T\mathscr{m}=\mathrm{div}_{\mathbf{X}}\left((\mu_\phi-c_x\mathscr{m})\ \mathbf{e}_x+\tau_1\,\widetilde{\mathbf{k}}_\phi\right)\\ \partial_T k_\phi=\mathbf{e}_x\cdot\nabla_{\mathbf{X}}\,(\omega_\phi-c_x\,k_\phi)+\dfrac{\widetilde{\mathbf{k}}_\phi}{k_x}\cdot\nabla_{\mathbf{X}}\omega_x\end{cases}$$

with $\mathbf{e}_x$ unitary, $\widetilde{\mathbf{k}}_\phi$ orthogonal to $\mathbf{e}_x$, and $k_x\ \mathbf{e}_x$ and $k_\phi\ \mathbf{e}_x+\widetilde{\mathbf{k}}_\phi$ curl free.

Linearizing System (100) about the constant $(\underline{\mu}_x,\underline{c}_x,\underline{\omega}_\phi,\underline{\mu}_\phi,\mathbf{e}_1,0)$ yields

(101)
$$\begin{cases}(\partial_T+\underline{c}_x\partial_X)\,\mathbf{e}_x=-\,(\nabla_{\mathbf{X}}-\mathbf{e}_1\ \partial_X)\,c_x\\ (\partial_T+\underline{c}_x\partial_X)\,\widetilde{\mathbf{k}}_\phi=(\nabla_{\mathbf{X}}-\mathbf{e}_1\ \partial_X)\,\omega_\phi\\ \underline{k}_x\mathbf{A}_0\,\mathrm{Hess}\,\Theta\ (\partial_T+\underline{c}_x\partial_X)\begin{pmatrix}\mu_x\\ c_x\\ \omega_\phi\\ \mu_\phi\end{pmatrix}=\mathbf{B}_0\,\partial_X\begin{pmatrix}\mu_x\\ c_x\\ \omega_\phi\\ \mu_\phi\end{pmatrix}+\begin{pmatrix}0&0\\ \underline{\tau}_3&\underline{\tau}_2\\ \underline{\tau}_2&\underline{\tau}_1\\ 0&0\end{pmatrix}\begin{pmatrix}\mathrm{div}_{\mathbf{X}}(\mathbf{e}_x)\\ \mathrm{div}_{\mathbf{X}}(\widetilde{\mathbf{k}}_\phi)\end{pmatrix}\end{cases}$$

with extra constraints that $\widetilde{\mathbf{k}}_\phi$ and $\mathbf{e}_x$ are orthogonal to $\underline{\mathbf{e}}_x=\mathbf{e}_1$ and that $k_x\underline{\mathbf{e}}_x+\underline{k}_x\,\mathbf{e}_x$ and $k_\phi\underline{\mathbf{e}}_x+\underline{k}_\phi\,\mathbf{e}_x+\widetilde{\mathbf{k}}_\phi$ are curl free, where (k_x,k_ϕ) are given explicitly as

(102)
$$\begin{aligned}k_x&=-\underline{k}_x^2\,\mathrm{d}\,(\partial_{\mu_x}\Theta)(\mu_x,c_x,\omega_\phi,\mu_\phi)\,,\\ k_\phi&=\frac{k_x}{\underline{k}_x}\,\underline{k}_\phi-\underline{k}_x\,\mathrm{d}\,(\partial_{\mu_\phi}\Theta)(\mu_x,c_x,\omega_\phi,\mu_\phi)\,,\end{aligned}$$

where total derivatives are taken with respect to $(\mu_x,c_x,\omega_\phi,\mu_\phi)$, and evaluation is at $(\underline{\mu}_x,\underline{c}_x,\underline{\omega}_\phi,\underline{\mu}_\phi,\mathbf{e}_1,0)$. In System 101, $\mathrm{Hess}\,\Theta=\mathrm{Hess}_{(\mu_x,c_x,\omega_\phi,\mu_\phi)}\,\Theta$ is likewise evaluated at $(\underline{\mu}_x,\underline{c}_x,\underline{\omega}_\phi,\underline{\mu}_\phi,\mathbf{e}_1,0)$. For the convenience of the reader, we detail some of the observations and manipulations used to go from (100) to (101).

1. As in going from (69) to (70), derivatives of Θ arise in (101) and (102) from (47).
2. As pointed out in Section 2.6, $\mathrm{d}_{(\mathbf{e}_x,\widetilde{\mathbf{k}}_\phi)}(\nabla_{(\mu_x,c_x,\omega_\phi,\mu_\phi)}\Theta)$ vanish at $(\underline{\mu}_x,\underline{c}_x,\underline{\omega}_\phi,\underline{\mu}_\phi,\mathbf{e}_1,0)$.
3. For any scalar-valued function a,
$$\underline{\mathbf{e}}_x\cdot\mathrm{div}_{\mathbf{X}}\left(a\;(\underline{\mathbf{e}}_x\otimes\underline{\mathbf{e}}_x-\mathrm{I}_d)\right)=0\,.$$
4. From orthogonality constraints stem that for any derivative $\partial_\sharp$,
$$\underline{\mathbf{e}}_x\cdot\partial_\sharp\,\mathbf{e}_x=0\,,\qquad\underline{\mathbf{e}}_x\cdot\partial_\sharp\widetilde{\mathbf{k}}_\phi=0\,,$$
so that
$$\begin{aligned}\underline{\mathbf{e}}_x\cdot\mathrm{div}_{\mathbf{X}}\left(\underline{\mathbf{e}}_x\otimes\mathbf{e}_x+\mathbf{e}_x\otimes\underline{\mathbf{e}}_x\right)&=\mathrm{div}_{\mathbf{X}}(\mathbf{e}_x)\,,\\ \underline{\mathbf{e}}_x\cdot\mathrm{div}_{\mathbf{X}}\left(\underline{\mathbf{e}}_x\otimes\widetilde{\mathbf{k}}_\phi+\widetilde{\mathbf{k}}_\phi\otimes\underline{\mathbf{e}}_x\right)&=\mathrm{div}_{\mathbf{X}}(\widetilde{\mathbf{k}}_\phi)\,.\end{aligned}$$

At this stage, noting that for $j\in\{1,2,3\}$, $\sigma_j=\underline{\tau}_j/\underline{k}_x$, a few line manipulations achieve proving the claimed relation between Δ_0 and modulated systems in the form

$$\begin{aligned}&\lambda^{2(d-1)}\times\Delta_0(\lambda,\mathrm{i}\,\xi,\|\boldsymbol{\eta}\|)\\ &=\det\left(\lambda\begin{pmatrix}\mathrm{I}_{2(d-1)}&0\\0&\mathbf{A}_0\,\mathrm{Hess}\,\Theta\end{pmatrix}-\mathrm{i}\,\xi\begin{pmatrix}0&0\\0&\mathbf{B}_0\end{pmatrix}+\left(\begin{array}{cc|cccc}0&0&0&-\mathrm{i}\,\boldsymbol{\eta}&0&0\\0&0&0&0&\mathrm{i}\,\boldsymbol{\eta}&0\\\hline 0&0&0&0&0&0\\\frac{\underline{\tau}_3}{\underline{k}_x}\,\mathrm{i}\,\boldsymbol{\eta}^\mathsf{T}&\frac{\underline{\tau}_2}{\underline{k}_x}\,\mathrm{i}\,\boldsymbol{\eta}^\mathsf{T}&0&0&0&0\\\frac{\underline{\tau}_2}{\underline{k}_x}\,\mathrm{i}\,\boldsymbol{\eta}^\mathsf{T}&\frac{\underline{\tau}_1}{\underline{k}_x}\,\mathrm{i}\,\boldsymbol{\eta}^\mathsf{T}&0&0&0&0\\0&0&0&0&0&0\end{array}\right)\right).\end{aligned}$$

5.2. Instability criteria. — The rest of the section is devoted to the study of instability criteria contained in Corollary 5.1 and its longitudinal counterparts.

We begin by rephrasing the main consequence of Corollary 5.1. A stable wave must satisfy

1. for any $(\lambda,\zeta)\in\mathbf{R}^2$, $\Delta_0(\lambda,0,\zeta)\geq0$;
2. for any $(\xi,\zeta)\in\mathbf{R}^2$, the roots of the (real) polynomial $\omega\mapsto\Delta_0(\mathrm{i}\,\omega,\mathrm{i}\,\xi,\zeta)$ are real.

Note that the latter condition may be expressed explicitly as inequality constraints on some polynomial expressions in $(\xi,\zeta)\in\mathbf{R}^2$, but the expressions involved are rather cumbersome.

The restriction to $\xi=0$ is much simpler to analyze. To do so, let us introduce notation for coefficients of Δ_0

$$\Delta_0(\lambda,z,\zeta)=\sum_{\substack{0\leq m,n,p\leq 4\\ m+n+p=4\\ p\leq m}}\delta_{(m,n,p)}\lambda^{m-p}\,z^n\,\zeta^{2p}\,, \tag{103}$$

and note that

$$\Delta_0(\lambda,0,\zeta)=\lambda^4\,\delta_{(4,0,0)}+\zeta^2\lambda^2\,\delta_{(3,0,1)}+\zeta^4\,\delta_{(2,0,2)}\,.$$

A straightforward consequence is the following stability condition.

Lemma 5.2. — *If the wave is stable, then for any $\zeta\in\mathbf{R}^2$, the roots of the polynomial $\omega\mapsto\Delta_0(\mathrm{i}\,\omega,0,\zeta)$ are real and*

$$\delta_{4,0,0}\geq 0\,,\qquad \delta_{3,0,1}\geq 2\sqrt{|\delta_{4,0,0}\,\delta_{2,0,2}|}\,,\qquad \delta_{2,0,2}\geq 0\,. \tag{104}$$

Moreover, if all the signs of the three inequalities in (104) *are strict, then $\Delta_0(\lambda,0,\zeta)>0$ when $(\lambda,\zeta)\in\mathbf{R}^2\setminus\{0\}$, and the roots of $\omega\mapsto\Delta_0(\mathrm{i}\,\omega,0,\zeta)$ are distinct when $\zeta\neq 0$.*

Note that the mere combination of the cases $\boldsymbol{\eta}=0$ – corresponding to longitudinal perturbations – (studied in Corollary 4.5 and in [40]) and $\xi=0$ – corresponding to longitudinally coperiodic perturbations – (studied in Lemma 5.2 and in [40]) is *a priori* insufficient to capture the full strength of Corollary 5.1. Indeed, note that the associated instability criteria do not involve coefficients $\delta_{(1,2,1)}$ and $\delta_{(2,1,1)}$. To illustrate why we expect that these coefficients matter, let us consider an abstract real polynomial π_0 of the form (103). Then

- fixing all coefficients except $\delta_{(2,1,1)}$ with $\delta_{4,0,0}\neq 0$ and choosing some $(\xi_0,\zeta_0)\in(\mathbf{R}^*)^2$, it follows that if $|\delta_{(2,1,1)}|$ is sufficiently large, then either $\omega\mapsto\pi_0(\mathrm{i}\,\omega,\mathrm{i}\,\xi_0,\zeta_0)$ or $\omega\mapsto\pi_0(\mathrm{i}\,\omega,-\mathrm{i}\,\xi_0,\zeta_0)$ possesses a nonreal root;
- fixing all coefficients except $\delta_{(1,2,1)}$ with $\delta_{4,0,0}\neq 0$ and choosing some $(\xi_0,\zeta_0)\in(\mathbf{R}^*)^2$, it follows that if $|\delta_{(1,2,1)}|$ is sufficiently large with $\delta_{(1,2,1)}<0$, then $\omega\mapsto\pi_0(\mathrm{i}\,\omega,\mathrm{i}\,\xi_0,\zeta_0)$ possesses a nonreal root.

A less pessimistic guess could be that the inspection of the regimes $|\xi|\ll|\zeta|$ and $|\zeta|\ll|\xi|$ (that are perturbations of $\xi=0$ and $\zeta=0$) could involve the missing coefficients and be sufficient to decide the instability criteria encoded by Corollary 5.1. For an example of a closely related situation where such a scenario occurs, the reader is referred to [44, 35]. Yet the following lemma suggests that even this weaker claim can be expected only in degenerate situations. Let us also anticipate a bit and stress that the coefficient $\delta_{(1,2,1)}$ plays a deep role in our analysis of the small-amplitude regime.

Lemma 5.3. — 1. *Assume that Σ_t is nonsingular and that the eigenvalues of Σ_t are real and distinct, or equivalently, that* Hess Θ *is nonsingular and that the characteristic values of the modulation system are real and* *distinct. Then there exists* $\varepsilon_0 > 0$*, such that when* $(\xi,\zeta) \in \mathbf{R}^2 \setminus \{(0,0)\}$ *is such that* $|\zeta| \leq \varepsilon_0\,|\xi|$*, the fourth-order real polynomial* $\omega \mapsto \Delta_0(\mathrm{i}\,\omega, \mathrm{i}\,\xi, \zeta)$ *possesses four distinct real roots.*

2. *Assume that* Σ_t *is nonsingular and that the eigenvalues of* $(\Sigma_t)^{-1}\Sigma_{\mathbf{y}}$ *are positive and distinct. Then there exists* $\varepsilon_0 > 0$*, such that when* $(\xi,\zeta) \in \mathbf{R}^2 \setminus \{(0,0)\}$ *is such that* $|\xi| \leq \varepsilon_0\,|\zeta|$*, the fourth-order real polynomial* $\omega \mapsto \Delta_0(\mathrm{i}\,\omega, \mathrm{i}\,\xi, \zeta)$ *possesses four distinct real roots.*

Recall that if Σ_t is nonsingular and either Σ_t possesses a nonreal eigenvalue, or $(\Sigma_t)^{-1}\Sigma_{\mathbf{y}}$ possesses an eigenvalue in $\mathbf{C} \setminus [0,+\infty)$, then the associated wave is unstable.

Proof. — The distinct character follows from continuity of polynomial roots (applied, respectively, at $\zeta = 0$ and $\xi = 0$). Then their reality is deduced from the stability by complex conjugation of the root set of real polynomials. □

More generally, the transition to nonreal roots of $\omega \mapsto \Delta_0(\mathrm{i}\,\omega, \mathrm{i}\,\xi, \zeta)$ may only occur near a (ξ_0, ζ_0) where the polynomial possesses a multiple root. With this in mind, we now elucidate how the breaking of a multiple root occurs near $\xi = 0$ and near $\zeta = 0$.

Proposition 5.4 (Breaking of a multiple root near $\boldsymbol{\eta} = 0$). — *Assume that* Σ_t *is nonsingular and that* ω_0 *is a real eigenvalue of* $(\Sigma_t)^{-1}$ *of algebraic multiplicity* r_0*. If*

$$\begin{array}{lll} \textit{either} & \Big(r_0 \geq 3 \quad \textit{and} \quad \delta_{(1,2,1)} + \delta_{(2,1,1)}\omega_0 + \delta_{(3,0,1)}\omega_0^2 \neq 0\Big), \\ \textit{or} & \left(r_0 = 2 \quad \textit{and} \quad \dfrac{\delta_{(1,2,1)} + \delta_{(2,1,1)}\omega_0 + \delta_{(3,0,1)}\omega_0^2}{\frac{1}{r_0!}\partial_\lambda^{r_0}\Delta_0(\omega_0,1,0)} < 0\right), \end{array}$$

then the corresponding wave is spectrally unstable.

Proof. — This follows from the Taylor expansion of Δ_0

$$\begin{aligned} \Delta_0(\lambda, \mathrm{i}\,\xi, \|\boldsymbol{\eta}\|) = (\lambda - \mathrm{i}\,\xi\omega_0)^{r_0}\,&\frac{\xi^{4-r_0}\partial_\lambda^{r_0}\Delta_0(\mathrm{i}\,\omega_0, \mathrm{i}, 0)}{r_0!} \\ &+ \|\boldsymbol{\eta}\|^2\left(\delta_{1,2,1}(\mathrm{i}\,\xi)^2 + \delta_{2,1,1}\lambda\,\mathrm{i}\,\xi + \delta_{3,0,1}\lambda^2\right) \\ &+ \mathcal{O}\left(|\lambda - \mathrm{i}\,\xi\omega_0|^{r_0+1} + \|\boldsymbol{\eta}\|^4\right). \end{aligned}$$

Hence from Theorem 3.2 and a continuity argument there are r_0 roots of $D_\xi(\,\cdot\,,\|\boldsymbol{\eta}\|)$ near $\mathrm{i}\,\xi\omega_0$ that expand as

$$\mathrm{i}\,\xi\,\omega_0+\mathrm{i}\,\xi\,\mathcal{Z}\,\left(\frac{\|\boldsymbol{\eta}\|}{|\xi|}\right)^{\frac{2}{r_0}}\left(\frac{\delta_{(1,2,1)}+\delta_{(2,1,1)}\omega_0+\delta_{(3,0,1)}\omega_0^2}{\frac{1}{r_0!}\partial_\lambda^{r_0}\Delta_0(\omega_0,1,0)}\right)^{\frac{1}{r_0}}$$
$$+\,\mathcal{O}\left(|\xi|\left(\frac{\|\boldsymbol{\eta}\|}{|\xi|}\right)^{\frac{2}{r_0}}\left(\left(\frac{\|\boldsymbol{\eta}\|}{|\xi|}\right)^{\frac{2}{r_0}}+\frac{|\xi|^3}{\|\boldsymbol{\eta}\|^2}\right)\right),$$

in the limit

$$\left(|\xi|,\frac{\|\boldsymbol{\eta}\|}{|\xi|},\frac{|\xi|^3}{\|\boldsymbol{\eta}\|^2}\right)\to(0,0,0)\,,$$

where $\mathcal{Z}$ runs over the r_0th roots of unity. □

Proposition 5.5 (Breaking of a multiple root near $\xi=0$). — *Assume that* $\delta_{(4,0,0)}\neq 0$ *and* $\delta_{(3,0,1)}^2=4\,\delta_{(4,0,0)}\,\delta_{(2,0,2)}$. *If*

$$\delta_{(2,1,1)}\,\delta_{(4,0,0)}-\frac{1}{2}\delta_{(3,1,0)}\,\delta_{(3,0,1)}\neq 0\,,$$

then the corresponding wave is spectrally unstable.

Proof. — From Lemma 5.2, stability requires $\delta_{(4,0,0)}>0$, $\delta_{(2,0,2)}\geq 0$, and $\delta_{(3,0,1)}=2\sqrt{\delta_{(4,0,0)}\,\delta_{(2,0,2)}}$, and we assume this from now on. The polynomial Δ_0 is then

$$\Delta_0(\lambda,\mathrm{i}\,\xi,\zeta)=\delta_{(4,0,0)}\left(\lambda^2+\sqrt{\frac{\delta_{(2,0,2)}}{\delta_{(4,0,0)}}}\zeta^2\right)^2+\mathrm{i}\,\xi(\delta_{(2,1,1)}\lambda\zeta^2+\delta_{(3,1,0)}\lambda^3)$$
$$+\,\mathcal{O}\left(\xi^2(\lambda^2+\zeta^2)).\right)$$

We begin with the case $\delta_{(2,0,2)}\neq 0$. To analyze it we introduce

$$\omega_0:=\left(\frac{\delta_{(2,0,2)}}{\delta_{(4,0,0)}}\right)^{1/4}=\sqrt{\frac{\delta_{(3,0,1)}}{2\delta_{(4,0,0)}}}\,.$$

Then for each $\sigma\in\{-1,1\}$, when $(\|\boldsymbol{\eta}\|,|\xi|/\|\boldsymbol{\eta}\|,\|\boldsymbol{\eta}\|^3/|\xi|^2)$ is sufficiently small, there are 2 roots of $D_\xi(\,\cdot\,,\|\boldsymbol{\eta}\|)$ near $\mathrm{i}\,\sigma\|\boldsymbol{\eta}\|\omega_0$ that expand as

$$\mathrm{i}\,\sigma\|\boldsymbol{\eta}\|\omega_0\left(1\pm\frac{1}{2\delta_{(4,0,0)}}\sqrt{\frac{\sigma\xi}{\|\boldsymbol{\eta}\|}\left(\delta_{(2,1,1)}\delta_{(4,0,0)}-\frac{1}{2}\delta_{(3,1,0)}\delta_{(3,0,1)}\right)}\right.$$
$$\left.+\,\mathcal{O}\left(\|\boldsymbol{\eta}\|\sqrt{\frac{|\xi|}{\|\boldsymbol{\eta}\|}}\left(\frac{|\xi|}{\|\boldsymbol{\eta}\|}+\frac{\|\boldsymbol{\eta}\|^2}{|\xi|}\right)\right)\right),$$

in the limit $\left(\|\boldsymbol{\eta}\|,\dfrac{|\xi|}{\|\boldsymbol{\eta}\|},\dfrac{\|\boldsymbol{\eta}\|^2}{|\xi|}\right)\to 0$.

When $\delta_{(2,0,2)} = 0$ – thus also $\delta_{(3,0,1)} = 0$ – and $(\|\boldsymbol{\eta}\|, |\xi|/\|\boldsymbol{\eta}\|, \|\boldsymbol{\eta}\|^2/|\xi|)$ is sufficiently small, three of the four roots of $D_\xi(\,\cdot\,, \|\boldsymbol{\eta}\|)$ near 0 expand as

$$\mathrm{i}\,\|\boldsymbol{\eta}\|^{\frac{2}{3}}\xi^{\frac{1}{3}}\,\mathcal{Z}\left(\frac{\delta_{(2,1,1)}}{\delta_{(4,0,0)}}\right)^{\frac{1}{3}} + \mathcal{O}\left(\|\boldsymbol{\eta}\|^{\frac{2}{3}}|\xi|^{\frac{1}{3}}\left(\left(\frac{|\xi|}{\|\boldsymbol{\eta}\|}\right)^{\frac{2}{3}} + \frac{\|\boldsymbol{\eta}\|^2}{|\xi|}\right)\right),$$

where $\mathcal{Z}$ runs over the 3rd roots of unity, in the limit

$$\left(\|\boldsymbol{\eta}\|, \frac{|\xi|}{\|\boldsymbol{\eta}\|}, \frac{\|\boldsymbol{\eta}\|^2}{|\xi|}\right) \to (0,0,0)\,. \qquad \square$$

5.3. Large-period regime. — We now examine consequences of Corollary 5.1 in asymptotic regimes described in from Section 2.5. Since the large-period regime turns out to be significantly simpler to analyze, we begin the solitary-wave limit.

We prove the following theorem.

Theorem 5.6. — *When $d \geq 2$, if $\partial_{c_x}^2 \Theta_{(s)}(\underline{c}_x^{(0)}, \underline{\rho}^{(0)}, \underline{k}_\phi^{(0)}) \neq 0$ then, in the large period regime near $(\underline{c}_x^{(0)}, \underline{\rho}^{(0)}, \underline{k}_\phi^{(0)})$, System 97 fails to be weakly hyperbolic, and waves are spectrally exponentially unstable to transversally slow, longitudinally coperiodic perturbations.*

Before turning to the proof of Theorem 5.6, we would like to add one comment. It is natural to wonder whether the proved instability corresponds to an instability of the limiting solitary-wave and even to expect that one could argue the other way around by proving the instability of solitary waves and deduce periodic-wave instability in the large-period regime by a spectral perturbation argument. When $\partial_{c_x}^2 \Theta_{(s)}(\underline{c}_x^{(0)}, \underline{\rho}^{(0)}, \underline{k}_\phi^{(0)}) < 0$, this is a well-known fact. We expect this to be true under the assumptions of Theorem 5.6. Yet, so far, general results for solitary-wave instabilities [7, 51] have been proven only for semilinear cases. More precisely, they have been proven for very specific forms of Schrödinger equations and for larger classes of Euler–Korteweg systems, sufficiently general to include all our semilinear cases. In the semilinear case, a different proof of Theorem 5.6 could thus be obtained by applying results from [7, 51] to solitary waves of the associated Euler–Korteweg systems, transferring those to large-period periodic waves of the same Euler–Korteweg systems through a suitable spectral perturbation theorem in the spirit of [26, 52, 57] and passing the latter to the Schrödinger systems by the results of Section 3.2.

Let us stress that instead our proof goes by examining the large-period limit of a periodic-wave criterion. Incidentally, we point out that our periodic-wave criterion is orthogonal to the arguments in [51] but shares some similarities with those in [7]. For the adaptation as a periodic-wave criterion of the arguments of the former, the reader is referred to [28].

One of the advantages of the way we have chosen is that it offers a symmetric treatment of both limits of interest, whereas the spectral perturbation argument fails in the harmonic limit. Another one is that we prove that the instability is of modulation type, being associated with the failure of weak hyperbolicity of System 97.

The rest of the present section is devoted to the proof of Theorem 5.6. This section and the next about the harmonic regime use formulation (94) and build upon intermediate[27] results from [10] and [11] on systems of Korteweg type. Indeed, in the solitary-wave limit, once relevant asymptotic expansions have been recalled, the proof shall be quite straightforward.

To ease notational translations, it is useful to recall that in Section 2.3 we have derived for (3) the hydrodynamic formulation (39)

$$\partial_t \begin{pmatrix} \rho \\ \mathbf{v} \end{pmatrix} = \mathcal{J}\,\delta H_0[(\rho,\mathbf{v})]\,, \tag{105}$$

with $\mathbf{v}$ curl free, where

$$\mathcal{J} = \begin{pmatrix} 0 & \mathrm{div} \\ \nabla & 0 \end{pmatrix}$$

and

$$H_0[(\rho,\mathbf{v})] = \kappa(2\,\rho)\,\rho\,\|\mathbf{v}\|^2 + \frac{\kappa(2\,\rho)}{4\,\rho}\|\nabla_{\mathbf{x}}\rho\|^2\,.$$

In turn, the Hamiltonian problems studied in [10, 11] include systems[28] in the form (105) but with a larger class of Hamiltonian densities, given in original notation from [10, 11] as

$$\mathcal{H}[(\rho,\mathbf{u}))] = \frac{1}{2}\tau(v)\|\mathbf{u}\|^2 + \frac{1}{2}\kappa(v)\|\nabla_{\mathbf{x}} v\|^2 + f(v)\,.$$

Thus, when importing results from [10, 11], we shall keep in mind the notational correspondence

$$(v,\,u) \to (\rho,\,v)\,,\quad \kappa(v) \to \frac{1}{2\rho}\kappa(2\rho)\,,\quad \tau(v) \to 2\rho\,\kappa(2\rho)\,,\quad f(v) \to W(2\rho)\,.$$

27. As opposed to main results.

28. Restricted to the one-dimensional case.

To derive expansions for σ_1, σ_2 and σ_3, it is convenient to use the profile equation (44) so as to write them as

$$\begin{aligned}
\sigma_1 &= \int_{\rho_{\min}(\mu_x)}^{\rho_{\max}(\mu_x)} \frac{f_1(\rho)}{\sqrt{\mu_x - \mathcal{W}_\rho(\rho)}} \sqrt{\frac{2\,\kappa(2\,\rho)}{2\,\rho}}\, \mathrm{d}\,\rho\,,\\
\sigma_2 &= \int_{\rho_{\min}(\mu_x)}^{\rho_{\max}(\mu_x)} \frac{f_2(\rho)}{\sqrt{\mu_x - \mathcal{W}_\rho(\rho)}} \sqrt{\frac{2\,\kappa(2\,\rho)}{2\,\rho}}\, \mathrm{d}\,\rho\,,\\
\sigma_3 &= \int_{\rho_{\min}(\mu_x)}^{\rho_{\max}(\mu_x)} \frac{f_3(\rho)}{\sqrt{\mu_x - \mathcal{W}_\rho(\rho)}} \sqrt{\frac{2\,\kappa(2\,\rho)}{2\,\rho}}\, \mathrm{d}\,\rho\\
&\quad + \int_{\rho_{\min}(\mu_x)}^{\rho_{\max}(\mu_x)} \sqrt{\mu_x - \mathcal{W}_\rho(\rho)} \sqrt{\frac{2\,\kappa(2\,\rho)}{2\,\rho}}\, \mathrm{d}\,\rho\,,
\end{aligned}$$

with

$$\begin{aligned}
f_1(\rho) &:= \kappa(2\,\rho)\,2\,\rho\,,\\
f_2(\rho; c_x, \mu_\phi) &:= \kappa(2\,\rho)\,2\,\rho\,\nu(\rho; c_x, \mu_\phi)\,,\\
f_3(\rho; c_x, \mu_\phi) &:= \kappa(2\,\rho)\,2\,\rho\,(\nu(\rho; c_x\,, \mu_\phi))^2\,.
\end{aligned}$$

In this form, [10, Proposition C.3] is directly applicable and yields the required expansions.

In the above and from now on, we mostly keep the dependence on $(c_x, \omega_\phi, \mu_\phi)$ implicit for the sake of readability. This is consistent with the fact that the limit is reached by holding $(c_x, \omega_\phi, \mu_\phi)$ fixed and taking μ_x sufficiently close to $\mu_x^{(0)}(c_x, \omega_\phi, \mu_\phi)$ (uniformly for $(c_x, \omega_\phi, \mu_\phi)$ in a compact neighborhood of $(\underline{c}_x^{(0)}, \underline{\omega}_\phi^{(0)}, \underline{\mu}_\phi^{(0)})$).

The solitary-wave expansions naturally involve a mass conjugated to the minimal[29] mass of periodic waves. Explicitly, in the large-period regime there exists $\rho_{\mathrm{dual}} = \rho_{\mathrm{dual}}(\mu_x; c_x, \omega_\phi, \mu_\phi)$, such that

$$\mu_x = \mathcal{W}_\rho(\rho_{\mathrm{dual}})\,,$$

with $\rho_{\mathrm{dual}} < \rho^{(0)}$, and $\mu_x - \mathcal{W}_\rho(\cdot)$ does not vanish on $(\rho_{\mathrm{dual}}, \rho_{\min})$. In other words, ρ_{dual} is the first cancellation point of $\mu_x - \mathcal{W}_\rho(\cdot)$ on the left of $\rho^{(0)}$; see Figure 2.2.

The following theorem gathers the relevant pieces of asymptotic expansions. Up to a slight extension to incorporate expansions of σ_1, σ_2, σ_3, it is the translation in our setting of results from [10, Theorem 3.16 & Lemma 4.1]. We

29. Recall that we have decided to focus only on the case when the end state of the mass of the limiting solitary wave is also its infimum.

borrow the statement and notation[30] from [11, Proposition 4 & Theorem 2], where relevant results from [10] are compactly summarized.

THEOREM 5.7 ([10]). — *In the large-period regime, there exist real numbers* $\mathfrak{a}_s$, $\mathfrak{b}_s$*, a positive number* $\mathfrak{h}_s$*, a vector* $\mathbf{X}_s$*, and a symmetric matrix* $\mathbb{O}_s$ *– depending smoothly on the parameters* $(c_x, \omega_\phi, \mu_\phi)$ *– such that, with*[31]

$$\varepsilon(\mu_x) := \frac{\rho_{\min}(\mu_x) - \rho_{\text{dual}}(\mu_x)}{\rho_{\max}(\mu_x) - \rho_{\min}(\mu_x)}\,, \qquad \mathfrak{c}_s := -\frac{1}{2\partial_\rho^2 \mathcal{W}_\rho(\rho^{(0)})}\,, \tag{106}$$

$$\begin{aligned}\frac{\pi}{X_x^{(s)}}\,\nabla\Theta = -\mathbf{V}_0\,\ln\varepsilon - \mathbf{X}_s + \frac{\varepsilon}{2}\,\mathbf{V}_0 - \frac{1}{2\mathfrak{h}_s}\,(\mathfrak{a}_s\,\mathbf{V}_0 + \mathfrak{b}_s\,\mathbf{W}_0 + \mathfrak{c}_s\,\mathbf{Z}_0)\,\varepsilon^2\ln\varepsilon \\ + \mathcal{O}\big(\varepsilon^2\big)\,,\end{aligned} \tag{107}$$

$$\begin{aligned}\frac{\pi}{X_x^{(s)}}\,\mathrm{Hess}\,\Theta = \mathfrak{h}_s\,\frac{1+\varepsilon}{\varepsilon^2}\,\mathbf{V}_0\otimes\mathbf{V}_0 \\ + (\mathfrak{a}_s\,\mathbf{V}_0\otimes\mathbf{V}_0 + \mathfrak{b}_s\,(\mathbf{V}_0\otimes\mathbf{W}_0 + \mathbf{W}_0\otimes\mathbf{V}_0))\ln\varepsilon \\ + (\mathbf{T}_0\otimes\mathbf{T}_0 + 2\mathfrak{c}_s\,\mathbf{W}_0\otimes\mathbf{W}_0 + \mathfrak{c}_s\,(\mathbf{Z}_0\otimes\mathbf{V}_0 + \mathbf{V}_0\otimes\mathbf{Z}_0))\ln\varepsilon \\ + \mathbb{O}_s + \mathcal{O}\big(\varepsilon\ln\varepsilon\big)\,,\end{aligned} \tag{108}$$

and, for $j = 1, 2, 3$,

$$\frac{\pi}{X_x^{(s)}}\,\sigma_j = -f_j(\rho^{(0)})\,\ln\varepsilon + \mathcal{O}(1)\,,$$

where

$$X_x^{(s)} := \sqrt{\frac{-\kappa(\rho^{(0)})}{\partial_\rho^2\mathcal{W}_\rho(\rho^{(0)})}}\,,$$

$$\mathbf{V}_0 := \begin{pmatrix}1\\ \mathfrak{q}(\rho^{(0)})\\ \rho^{(0)}\\ \nu(\rho^{(0)})\end{pmatrix}, \qquad \mathbf{W}_0 := \begin{pmatrix}0\\ \partial_\rho\mathfrak{q}(\rho^{(0)})\\ 1\\ \partial_\rho\nu(\rho^{(0)})\end{pmatrix},$$

$$\mathbf{Z}_0 := \begin{pmatrix}0\\ \partial_\rho^2\mathfrak{q}(\rho^{(0)})\\ 0\\ \partial_\rho^2\nu(\rho^{(0)})\end{pmatrix}, \qquad \mathbf{T}_0 := \frac{1}{\sqrt{\kappa(2\,\rho^{(0)})\,2\,\rho^{(0)}}}\begin{pmatrix}0\\ \rho^{(0)}\\ 0\\ 1\end{pmatrix},$$

30. Except for a few variations. We use $(X_x^{(s)}, \varepsilon)$ instead of (Ξ_s, ρ) and $(\rho_{\text{dual}}, \rho_{\min}, \rho_{\max})$ instead of (v_1, v_2, v_3). The subscript 0 was originally s.

31. The parameter ε goes to zero as $\sqrt{\mu_x^{(0)}(c_x, \omega_\phi, \mu_\phi) - \mu_x}$.

with $\mathfrak{q}$ *defined by*

$$\mathfrak{q}(\rho) := \rho\,\nu(\rho)\,.$$

Moreover, the vectors are such that

$$\begin{aligned}
&\mathbf{V}_0\cdot\mathbf{B}_0^{-1}\mathbf{A}_0\,\mathbf{V}_0 = 0\,,\quad \mathbf{V}_0\,\cdot\mathbf{B}_0^{-1}\mathbf{A}_0\,\mathbf{W}_0 = 0\,,\quad \mathbf{V}_0\,\cdot\mathbf{B}_0^{-1}\mathbf{A}_0\,\mathbf{T}_0 = 0\,,\\
&\mathbf{V}_0\,\cdot\mathbf{B}_0^{-1}\mathbf{A}_0\,\mathbf{Z}_0 = -\mathbf{W}_0\cdot\mathbf{B}_0^{-1}\mathbf{A}_0\,\mathbf{W}_0\,,\\
&\mathbf{T}_0\cdot\mathbf{B}_0^{-1}\mathbf{A}_0\,\mathbf{T}_0 = 0\,,\qquad \mathbf{T}_0\cdot\mathbf{B}_0^{-1}\mathbf{A}_0\,\mathbf{Z}_0 = 0\,,\\
&\mathbf{e}_1\cdot\mathbf{V}_0 = 1\,,\qquad \mathbf{e}_1\cdot\mathbf{W}_0 = 0\,,\qquad \mathbf{e}_1\cdot\mathbf{Z}_0 = 0\,,\qquad \mathbf{e}_1\cdot\mathbf{T}_0 = 0\,,
\end{aligned}$$

and

$$\frac{X_x^{(s)}}{\pi}\,(\mathbf{B}_0^{-1}\mathbf{A}_0\mathbf{V}_0)\cdot\mathbf{X}_s = -\partial_{c_x}\Theta_{(s)}(c_x^{(0)},\rho^{(0)},k_\phi^{(0)})\,,$$

$$\frac{X_x^{(s)}}{\pi}\,(\mathbf{B}_0^{-1}\mathbf{A}_0\mathbf{V}_0)\cdot\mathbb{O}_s\,\mathbf{B}_0^{-1}\mathbf{A}_0\mathbf{V}_0 = \partial_{c_x}^2\Theta_{(s)}(c_x^{(0)},\rho^{(0)},k_\phi^{(0)})\,.$$

As a consequence, after a few straightforward but tedious computations, omitted here but detailed in the discussion preceding[32] [11, Theorem 8] stems that

$$\frac{\pi}{X_x^{(s)}}\begin{pmatrix}1&0&0\\0&\frac{\sqrt{1+\varepsilon}}{\varepsilon}&0\\0&0&\mathrm{I}_2\end{pmatrix}\mathbf{P}_0(-\mathbf{B}_0^{-1}\mathbf{A}_0)\,\mathrm{Hess}\,\Theta\,\mathbf{P}_0^{-1}\begin{pmatrix}1&0&0\\0&\frac{\varepsilon}{\sqrt{1+\varepsilon}}&0\\0&0&\mathrm{I}_2\end{pmatrix}$$

$$=\begin{pmatrix}-\mathfrak{c}_s\,w_0\ln(\varepsilon)+\mathcal{O}(1) & \frac{\sqrt{1+\varepsilon}}{\varepsilon}\,\mathfrak{h}_s+\mathcal{O}(\varepsilon\ln(\varepsilon)) & \mathcal{O}(\ln(\varepsilon))\\ \frac{\sqrt{1+\varepsilon}}{\varepsilon}\,\frac{X_x^{(s)}}{\pi}\partial_{c_x}^2\Theta_{(s)}+\mathcal{O}(\ln(\varepsilon)) & -\mathfrak{c}_s\,w_0\ln(\varepsilon)+\mathcal{O}(1) & \frac{\sqrt{1+\varepsilon}}{\varepsilon}\,\mathbf{y}_s{}^{\mathsf{T}}+\mathcal{O}(\ln(\varepsilon))\\ \mathcal{O}(1) & \mathcal{O}(\varepsilon\ln(\varepsilon)) & \Sigma_0^{-1}\ln(\varepsilon)+\mathcal{O}(1)\end{pmatrix}$$

with $\mathbf{y}_s$ some two-dimensional vector, and

$$\mathbf{P}_0 := \begin{pmatrix}-\,\mathbf{e}_2 & \mathbf{V}_0 & \mathbf{T}_0 & \mathbf{W}_0\end{pmatrix}^{\mathsf{T}}$$

$$\Sigma_0 := \begin{pmatrix}\sigma_0&0\\\frac{w_0}{2}&1\end{pmatrix}(\mathbf{B}\,\mathrm{Hess}\,H^{(0)}(\rho^{(0)},k_\phi^{(0)})+c_x\,\mathrm{I}_2)\begin{pmatrix}\sigma_0&0\\\frac{w_0}{2}&1\end{pmatrix}^{-1}, \tag{109}$$

where

$$\sigma_0 := -\mathbf{T}_0\cdot\mathbf{B}_0^{-1}\mathbf{A}_0\mathbf{W}_0 = \frac{1}{\sqrt{\kappa(2\,\rho^{(0)})\,2\,\rho^{(0)}}}$$

$$w_0 := -\mathbf{W}_0\cdot\mathbf{B}_0^{-1}\mathbf{A}_0\mathbf{W}_0 = 2\partial_\rho\nu(\rho^{(0)})\,.$$

32. In notation of [11], $\mathbf{P}_0 = \mathbb{P}_s{}^{\mathsf{T}}\mathrm{SS}$, $-\mathbf{B}_0^{-1}\mathbf{A}_0 = \mathrm{SS}^{-1}$, $\mathbf{y}_s = (\mathbf{D}_s^{-1})^{\mathsf{T}}\mathbf{y}_s$, $\Sigma_0^{-1} = \Sigma_s\mathbf{D}_s^{-1}$ and

$$\begin{pmatrix}\sigma_s&0\\\frac{w_s}{2}&1\end{pmatrix} = \mathbf{A}_s\,\mathbf{B}^{-1}\,.$$

Moreover,

$$\Sigma_0^{-1} = \begin{pmatrix} 0 & 2\mathfrak{c}_s\sigma_0 \\ \sigma_0 & 2\mathfrak{c}_s w_0 \end{pmatrix}.$$

We recall that $H^{(0)}$ is the zero dispersion limit of the Hamiltonian H_0 of the hydrodynamic formulation of the Schrödinger equation, and $\mathbf{B}$ is the self-adjoint matrix involved in this formulation; see (51).

Observe also that from the foregoing it follows that

$$\begin{aligned}
\delta_{(4,0,0)} &= \det(\operatorname{Hess}\Theta) \\
&= -\frac{(\ln(\varepsilon))^2}{\varepsilon^2}\,\frac{(X_x^{(s)})^5}{\pi^5}\,\frac{\mathfrak{h}_s\,\partial_{c_x}^2\Theta_{(s)}}{\det(\mathbf{B}\operatorname{Hess}H^{(0)}+c_x\,\mathrm{I}_2)}+\mathcal{O}\left(\frac{\ln(\varepsilon)}{\varepsilon^2}\right) \\
&= \frac{(\ln(\varepsilon))^2}{\varepsilon^2}\,\frac{(X_x^{(s)})^5}{\pi^5} 2\mathfrak{c}_s\,\sigma_0^2\,\mathfrak{h}_s\,\partial_{c_x}^2\Theta_{(s)}+\mathcal{O}\left(\frac{\ln(\varepsilon)}{\varepsilon^2}\right) \\
&= -\frac{(\ln(\varepsilon))^2}{\varepsilon^2}\,\frac{(X_x^{(s)})^5}{\pi^5}\,\frac{\mathfrak{h}_s\,\partial_{c_x}^2\Theta_{(s)}}{\kappa(2\,\rho^{(0)})\,2\,\rho^{(0)}\,\partial_\rho^2\mathcal{W}_\rho(\rho^{(0)})}+\mathcal{O}\left(\frac{\ln(\varepsilon)}{\varepsilon^2}\right).
\end{aligned}$$

Likewise,

$$\begin{aligned}
\delta_{(0,4,0)} &= 1\,, \qquad \delta_{(1,3,0)} = \mathcal{O}(1)\,, \\
\delta_{(2,2,0)} &= -\frac{1}{\varepsilon^2}\,\frac{(X_x^{(s)})^3}{\pi^3}\,\mathfrak{h}_s\,\partial_{c_x}^2\Theta_{(s)}+\mathcal{O}\left(\frac{\ln(\varepsilon)}{\varepsilon}\right), \\
\delta_{(3,1,0)} &= -\frac{\ln(\varepsilon)}{\varepsilon^2}\,\frac{(X_x^{(s)})^4}{\pi^4}\,2\mathfrak{c}_s\,w_0\,\mathfrak{h}_s\,\partial_{c_x}^2\Theta_{(s)}\;+\mathcal{O}\left(\frac{1}{\varepsilon^2}\right).
\end{aligned}$$

To go on we need to compute a similar expansion for

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{\sqrt{1+\varepsilon}}{\varepsilon} & 0 \\ 0 & 0 & \mathrm{I}_2 \end{pmatrix}\mathbf{P}_0(-\mathbf{B}_0^{-1}\mathbf{C}_0)\,\mathbf{P}_0^{-1}\begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{\varepsilon}{\sqrt{1+\varepsilon}} & 0 \\ 0 & 0 & \mathrm{I}_2 \end{pmatrix}.$$

A direct computation yields

$$-\mathbf{P}_0\mathbf{B}_0^{-1}\mathbf{C}_0 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \sigma_3+\nu(\rho^{(0)})\sigma_2 & -(\sigma_2+\nu(\rho^{(0)})\sigma_1) & 0 \\ 0 & \sigma_0\sigma_2 & -\sigma_0\sigma_1 & 0 \\ 0 & \frac{w_0}{2}\sigma_2 & -\frac{w_0}{2}\sigma_1 & 0 \end{pmatrix},$$

and

$$\mathbf{P}_0^{-1} = \begin{pmatrix} * & * & * & * \\ -1 & 0 & 0 & 0 \\ \nu(\rho^{(0)}) & 0 & -\frac{w_0}{2\sigma_0} & 1 \\ * & * & * & * \end{pmatrix},$$

so that

$$-\mathbf{P}_0\mathbf{B}_0^{-1}\mathbf{C}_0\mathbf{P}_0^{-1} =$$
$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ -(\sigma_3+2\nu(\rho^{(0)})\sigma_2+\nu(\rho^{(0)})^2\sigma_1) & 0 & \frac{w_0}{2\sigma_0}(\sigma_2+\nu(\rho^{(0)})\sigma_1) & -(\sigma_2+\nu(\rho^{(0)})\sigma_1) \\ -\sigma_0(\sigma_2+\nu(\rho^{(0)})\sigma_1) & 0 & \frac{w_0}{2}\sigma_1 & -\sigma_0\sigma_1 \\ -\frac{w_0}{2}(\sigma_2+\nu(\rho^{(0)})\sigma_1) & 0 & \frac{w_0^2}{4\sigma_0}\sigma_1 & -\frac{w_0}{2}\sigma_1 \end{pmatrix}.$$

To ease computations and materialize both symmetry and size we introduce

$$\delta_1^{(s)} := \frac{\sigma_1}{-\ln(\varepsilon)X_x^{(s)}/\pi}, \qquad \delta_2^{(s)} := \frac{\sigma_2+\nu(\rho^{(0)})\sigma_1}{-\ln(\varepsilon)X_x^{(s)}/\pi},$$
$$\delta_3^{(s)} := \frac{\sigma_3+2\nu(\rho^{(0)})\sigma_2+\nu(\rho^{(0)})^2\sigma_1}{-\ln(\varepsilon)X_x^{(s)}/\pi}.$$

Thus,

$$-\begin{pmatrix}1 & 0 & 0\\ 0 & \frac{\sqrt{1+\varepsilon}}{\varepsilon} & 0 \\ 0 & 0 & \mathrm{I}_2\end{pmatrix}\mathbf{P}_0\mathbf{B}_0^{-1}\mathbf{C}_0\mathbf{P}_0^{-1}\begin{pmatrix}1 & 0 & 0\\ 0 & \frac{\varepsilon}{\sqrt{1+\varepsilon}} & 0 \\ 0 & 0 & \mathrm{I}_2\end{pmatrix}$$
$$= \frac{X_x^{(s)}}{\pi}\begin{pmatrix} 0 & 0 & 0 & 0 \\ \delta_3^{(s)}\frac{\sqrt{1+\varepsilon}}{\varepsilon}\ln(\varepsilon) & 0 & -\frac{w_0}{2\sigma_0}\delta_2^{(s)}\frac{\sqrt{1+\varepsilon}}{\varepsilon}\ln(\varepsilon) & \delta_2^{(s)}\frac{\sqrt{1+\varepsilon}}{\varepsilon}\ln(\varepsilon) \\ \sigma_0\delta_2^{(s)}\ln(\varepsilon) & 0 & -\frac{w_0}{2}\delta_1^{(s)}\ln(\varepsilon) & \sigma_0\delta_1^{(s)}\ln(\varepsilon) \\ \frac{w_0}{2}\delta_2^{(s)}\ln(\varepsilon) & 0 & -\frac{w_0^2}{4\sigma_0}\delta_1^{(s)}\ln(\varepsilon) & \frac{w_0}{2}\delta_1^{(s)}\ln(\varepsilon) \end{pmatrix},$$

with

$$\delta_1^{(s)} = f_1(\rho^{(0)}) + \mathcal{O}\left(\frac{1}{\ln(\varepsilon)}\right),$$
$$\delta_2^{(s)} = 2\,\nu(\rho^{(0)})\,f_1(\rho^{(0)}) + \mathcal{O}\left(\frac{1}{\ln(\varepsilon)}\right),$$
$$\delta_3^{(s)} = 4\,\nu(\rho^{(0)})^2\,f_1(\rho^{(0)}) + \mathcal{O}\left(\frac{1}{\ln(\varepsilon)}\right).$$

At main order the matrix has rank 1, and with this observation we find

$$\delta_{(2,0,2)} = \mathcal{O}\left(\frac{(\ln(\varepsilon))^2}{\varepsilon^2}\right),$$
$$\delta_{(3,0,1)} = \frac{(\ln(\varepsilon))^3}{\varepsilon^2}\,\frac{(X_x^{(s)})^4}{\pi^4}2\,\mathfrak{c}_s\,\sigma_0^2\,\mathfrak{h}_s\delta_3^{(s)} + \mathcal{O}\left(\frac{(\ln(\varepsilon))^2}{\varepsilon^2}\right).$$

Note that this already yields the instability condition in the large period regime:

$$\partial_{c_x}^2\Theta_{(s)} > 0\,,$$

This may be derived, for instance, by examining the limit $\varepsilon \to 0$ of the rescaled

$$\begin{aligned}\Delta_0 &\left(\frac{\sqrt{\varepsilon}\,\lambda}{\sqrt{|\ln(\varepsilon)|}}, 0, \frac{\sqrt{\varepsilon}\,\zeta}{\ln(\varepsilon)}\right)\\ &= \left(-\frac{(X_x^{(s)})^5}{\pi^5}\frac{\mathfrak{h}_s\,\partial_{c_x}^2\Theta_{(s)}}{\kappa^{(0)}\,2\,\rho^{(0)}\,\partial_\rho^2\mathcal{W}_\rho^{(0)}} + \mathcal{O}\left(\frac{1}{\ln\varepsilon}\right)\right)\lambda^4 + \mathcal{O}\left(\frac{\zeta^4}{\ln^2\varepsilon}\right)\\ &\quad + \left(\frac{(X_x^{(s)})^4}{\pi^4} 2\,\mathfrak{c}_s\,\sigma_0^2\,\mathfrak{h}_s 4(\nu^{(0)})^2 2\rho^{(0)}\kappa^{(0)} + \mathcal{O}\left(\frac{1}{\ln\varepsilon}\right)\right)\lambda^2\zeta^2\,.\end{aligned}$$

According to both Theorem 4.3 and Theorem 4.6, $\partial_{c_x}^2\Theta_{(s)} < 0$ also yields instability. Applying Theorem 4.3 shows that the instability may be obtained with $\xi = 0$, whereas applying Theorem 4.6 shows that it also corresponds to a failure of weak hyperbolicity of the modulated system.

This achieves the proof of Theorem 5.6.

5.4. Small-amplitude regime. — We now turn to the small-amplitude limit. Our goal is to prove the following theorem.

THEOREM 5.8. — *In the small amplitude regime near a* $(\underline{c}_x^{(0)}, \underline{\rho}^{(0)}, \underline{k}_\phi^{(0)})$*, such that*

$$\partial_\rho \nu(\underline{\rho}^{(0)}; \underline{c}_x^{(0)}, \underline{\mu}_\phi^{(0)}) \neq 0\,,$$

and[33]

$$\delta_{hyp} \times \delta_{BF}(\underline{c}_x^{(0)}, \underline{\omega}_\phi^{(0)}, \underline{\mu}_\phi^{(0)}) \neq 0$$

waves are spectrally exponentially unstable to transversally slow, longitudinal sideband perturbations.

Let us stress that, as proved in Appendix B, the limiting constant state is spectrally stable if and only if $\delta_{hyp} < 0$, so that, when $\delta_{hyp} > 0$, the result is nontrivial from the point of view of spectral perturbation. The overall proof strategy is the same as in the large-period regime, but the final argument is significantly more cumbersome. In particular, it involves essentially all coefficients of Δ_0.

To begin with, we gather relevant expansions. The following is a straightforward extension of [10, Theorem 3.14 & Lemma 4.1], with notation taken from [11, Theorem 2 & Proposition 4].

33. Where δ_{hyp}, δ_{BF} are as in (71)–(72).

THEOREM 5.9 ([10]). — *In the small-amplitude regime, there exist a real number* $\mathfrak{b}_0$ *and a positive number* $\mathfrak{c}_0$ *– depending smoothly on the parameters* $(c_x, \omega_\phi, \mu_\phi)$ *– such that, with* $\mathfrak{a}_0$ *given by* (64) *and*[34]

$$\varepsilon(\mu_x) := \frac{\rho_{\max}(\mu_x) - \rho_{\min}(\mu_x)}{2} \frac{\rho_{\min}(\mu_x) - \rho_{\text{dual}}(\mu_x)}{\rho_{\max}(\mu_x) - \rho_{\min}(\mu_x)}, \quad \mathfrak{c}_0 := \frac{1}{2\partial_\rho^2 \mathcal{W}_\rho(\rho^{(0)})},$$

we have

$$\frac{4\mathfrak{c}_0}{X_x^{(0)}} \nabla\Theta = 4\mathfrak{c}_0 \, \mathbf{V}_0 + (\mathfrak{a}_0 \, \mathbf{V}_0 + \mathfrak{b}_0 \, \mathbf{W}_0 + \mathfrak{c}_0 \, \mathbf{Z}_0) \, \varepsilon^2 + \mathcal{O}(\varepsilon^4), \tag{110}$$

$$\begin{aligned} \frac{1}{X_x^{(0)}} \operatorname{Hess}\Theta = {}& \mathfrak{a}_0 \, \mathbf{V}_0 \otimes \mathbf{V}_0 + \mathfrak{b}_0 \, (\mathbf{V}_0 \otimes \mathbf{W}_0 + \mathbf{W}_0 \otimes \mathbf{V}_0) - \mathbf{T}_0 \otimes \mathbf{T}_0 \\ & + 2\,\mathfrak{c}_0 \, \mathbf{W}_0 \otimes \mathbf{W}_0 + \mathfrak{c}_0 \, (\mathbf{V}_0 \otimes \mathbf{Z}_0 + \mathbf{Z}_0 \otimes \mathbf{V}_0) + \mathcal{O}(\varepsilon^2), \end{aligned} \tag{111}$$

and, for $j = 1, 2, 3$,

$$\frac{4\mathfrak{c}_0}{X_x^{(0)}} \, \sigma_j = 4\mathfrak{c}_0 \, f_j(\rho^{(0)}) + \mathcal{O}(\varepsilon^2),$$

where $X_x^{(0)}$ *denotes the harmonic period* (48)*, and the other quantities are as in Theorem 5.7.*

Our starting point is

$$\begin{aligned} &\frac{1}{X_x^{(0)}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\varepsilon} & 0 \\ 0 & 0 & \mathrm{I}_2 \end{pmatrix} \widetilde{\mathbf{P}}_0^{-1} \mathbf{P}_0 (-\mathbf{B}_0^{-1} \mathbf{A}_0) \operatorname{Hess}\Theta \, \mathbf{P}_0^{-1} \widetilde{\mathbf{P}}_0 \begin{pmatrix} 1 & 0 & 0 \\ 0 & \varepsilon & 0 \\ 0 & 0 & \mathrm{I}_2 \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\lambda_\varepsilon} & \mathfrak{d}_0 \, \varepsilon + \mathcal{O}(\varepsilon^3) & \mathcal{O}(\varepsilon^2) \\ \mathfrak{e}_0 \, \varepsilon + \mathcal{O}(\varepsilon^2) & \frac{1}{\lambda_\varepsilon} & \mathcal{O}(\varepsilon) \\ \mathcal{O}(\varepsilon^2) & \mathcal{O}(\varepsilon^3) & \Sigma_\varepsilon^{-1} \end{pmatrix}, \end{aligned}$$

with

$$\lambda_\varepsilon = -\frac{1}{\mathfrak{c}_0 \, w_0} + \mathcal{O}(\varepsilon^2), \qquad \Sigma_\varepsilon = \Sigma_0 + \mathcal{O}(\varepsilon^2),$$

34. The parameter ε goes to zero as $\sqrt{\mu_x - \mu_x^{(0)}(c_x, \omega_\phi, \mu_\phi)}$.

$\mathfrak{d}_0$ and $\mathfrak{e}_0$ having the sign[35] respectively of δ_{BF} and of $2\rho^{(0)}\,\kappa(2\rho^{(0)})\,w_0^2 + 8\delta_{hyp}$,

$$\widetilde{\mathbf{P}}_0 = \begin{pmatrix} 1 & 0 & \ell^{\mathsf{T}} \\ 0 & 1 & 0 \\ 0 & r & \mathrm{I}_2 \end{pmatrix}, \qquad \widetilde{\mathbf{P}}_0^{-1} = \begin{pmatrix} 1 & \ell^{\mathsf{T}} r & -\ell^{\mathsf{T}} \\ 0 & 1 & 0 \\ 0 & -r & \mathrm{I}_2 \end{pmatrix},$$

where

$$r = -(\Sigma_0^{-1} + \mathfrak{c}_0\, w_0)^{-1} \begin{pmatrix} \mathfrak{b}_0\,\sigma_0 \\ \mathfrak{b}_0\, w_0 + \mathfrak{c}_0\,\zeta_0 \end{pmatrix}, \qquad \ell = -\begin{pmatrix} 0 & \sigma_0 \\ \sigma_0 & w_0 \end{pmatrix}^{-1} r\,.$$

Note that

$$\begin{aligned} \Sigma_0^{-1} &= \begin{pmatrix} 0 & -2\mathfrak{c}_0\sigma_0 \\ \sigma_0 & -2\mathfrak{c}_0 w_0 \end{pmatrix}, \qquad \delta_{hyp} = \frac{1}{4}\left(\frac{w_0^2}{4\sigma_0^2} - \frac{1}{2\mathfrak{c}_0}\right), \\ &\det(\Sigma_0^{-1} + \mathfrak{c}_0\, w_0) = -16\,\mathfrak{c}_0^2\,\sigma_0^2\,\delta_{hyp}\,, \\ &(\mathrm{Tr}(\Sigma_0^{-1}))^2 - 4\,\det(\Sigma_0^{-1}) = 64\,\mathfrak{c}_0^2\,\sigma_0^2\,\delta_{hyp}\,. \end{aligned}$$

This yields $\delta_{(0,4,0)} = 1$,

$$\begin{aligned} \delta_{(4,0,0)} &= \det(\mathrm{Hess}\,\Theta) = (X_x^{(0)})^4\left(\frac{1}{\lambda_\varepsilon^2} - \varepsilon^2\,\mathfrak{d}_0\,\mathfrak{e}_0\right)\det(\Sigma_\varepsilon^{-1}) + \mathcal{O}\left(\varepsilon^4\right), \\ \delta_{(3,1,0)} &= (X_x^{(0)})^3\left(\frac{2}{\lambda_\varepsilon}\,\det(\Sigma_\varepsilon^{-1}) + \left(\frac{1}{\lambda_\varepsilon^2} - \varepsilon^2\,\mathfrak{d}_0\,\mathfrak{e}_0\right)\mathrm{Tr}(\Sigma_\varepsilon^{-1})\right) + \mathcal{O}\left(\varepsilon^4\right), \\ \delta_{(2,2,0)} &= (X_x^{(0)})^2\left(\det(\Sigma_\varepsilon^{-1}) + \left(\frac{1}{\lambda_\varepsilon^2} - \varepsilon^2\,\mathfrak{d}_0\,\mathfrak{e}_0\right) + \frac{2}{\lambda_\varepsilon}\,\mathrm{Tr}(\Sigma_\varepsilon^{-1})\right) + \mathcal{O}\left(\varepsilon^4\right), \\ \delta_{(1,3,0)} &= X_x^{(0)}\left(\frac{2}{\lambda_\varepsilon} + \mathrm{Tr}(\Sigma_\varepsilon^{-1})\right), \end{aligned}$$

35. In notation of [11], with $(b,g) \to (1,\nu)$,

$$\begin{aligned} \delta_{BF} &= \frac{1}{16}\,\frac{4\tau(\partial_v g)^5}{b^3 k_0}\,\frac{\partial_v^2\mathcal{W} + 3\tau\,(\partial_v g)^2}{-\partial_v^2\mathcal{W} + 3\tau\,(\partial_v g)^2} \times \Delta_{MI}\,, \\ \delta_{hyp} &= \frac{1}{4}\partial_v^2\mathscr{H} = \frac{1}{4}(\tau\,(\partial_v g)^2 - \partial_v^2\mathcal{W})\,, \\ 2\rho^{(0)}\,\kappa(2\rho^{(0)})\,w_0^2 + 8\delta_{hyp} &= 2(-\partial_v^2\mathcal{W} + 3\tau\,(\partial_v g)^2)\,, \qquad w_0 = w_0 = \frac{2\partial_v g}{b}\,, \\ \mathfrak{d}_0 &= \frac{\tau(\partial_v g)^5}{b^3\,k_0\,(\partial_v^2\mathcal{W})^3}(-\partial_v^2\mathcal{W} + 3\tau\,(\partial_v g)^2)\,\Delta_{MI}\,, \qquad \mathfrak{d}_0\,\mathfrak{e}_0 = \frac{\mathfrak{c}_0^3\,w_0^5}{4\,k_0}\,\Delta_{MI}\,, \end{aligned}$$

which is equivalent to

$$\Delta_0\left(\frac{\lambda}{X_x^{(0)}}, z, 0\right) = \left(\left(\frac{\lambda}{\lambda_\varepsilon}+z\right)^2 - \lambda^2\varepsilon^2\mathfrak{d}_0\,\epsilon_0\right)\det(\lambda\Sigma_\varepsilon^{-1}+z\,\mathrm{I}_2)$$
$$+\,\mathcal{O}(\epsilon^4\,\lambda^2\,(|\lambda|^2+|z|^2))\,,$$
$$\Delta_0\left(\frac{\lambda}{X_x^{(0)}}, z-\frac{\lambda}{\lambda_\varepsilon}, 0\right) = \left(z^2-\lambda^2\varepsilon^2\mathfrak{d}_0\,\epsilon_0\right)\det\left(\lambda\left(\Sigma_\varepsilon^{-1}-\frac{1}{\lambda_\varepsilon}\mathrm{I}_2\right)+z\,\mathrm{I}_2\right)$$
$$+\,\mathcal{O}(\epsilon^4\,\lambda^2\,(|\lambda|^2+|z|^2))\,.$$

One recovers the instability criteria on δ_{BF} and δ_{hyp} in the form that instability stems both from $\mathfrak{d}_0\,\epsilon_0<0$ and from $\left(\mathrm{Tr}(\Sigma_0^{-1})\right)^2-4\det(\Sigma_0^{-1})<0$.

This motivates a first shift to

$$\tilde{\Delta}_0(\lambda,z,\zeta) := \Delta_0\left(\frac{\lambda}{X_x^{(0)}}, z-\frac{\lambda}{\lambda_\varepsilon}, \frac{\zeta}{X_x^{(0)}}\right).$$

Note that we still have an expansion in the form

$$\tilde{\Delta}_0(\lambda,z,\zeta) = \sum_{\substack{0\leq m,n,p\leq 4\\ m+n+p=4\\ p\leq m}} \tilde{\delta}_{(m,n,p)}\lambda^{m-p}\,z^n\,\zeta^{2p}\,,$$

but that expressing instability criteria in terms of $\tilde{\Delta}_0$ is not obvious and will require some care. We already know that $\tilde{\delta}_{(0,4,0)}=1$, and then

$$\tilde{\delta}_{(4,0,0)} = -\varepsilon^2\,\mathfrak{d}_0\,\mathfrak{e}_0\,\det\left(\Sigma_\varepsilon^{-1}-\frac{1}{\lambda_\varepsilon}\mathrm{I}_2\right)+\mathcal{O}\left(\varepsilon^4\right),$$
$$\tilde{\delta}_{(3,1,0)} = -\varepsilon^2\,\mathfrak{d}_0\,\mathfrak{e}_0\,\mathrm{Tr}\left(\Sigma_\varepsilon^{-1}-\frac{1}{\lambda_\varepsilon}\mathrm{I}_2\right)+\mathcal{O}\left(\varepsilon^4\right),$$
$$\tilde{\delta}_{(2,2,0)} = \det\left(\Sigma_\varepsilon^{-1}-\frac{1}{\lambda_\varepsilon}\mathrm{I}_2\right)-\varepsilon^2\,\mathfrak{d}_0\,\mathfrak{e}_0+\mathcal{O}\left(\varepsilon^4\right),$$
$$\tilde{\delta}_{(1,3,0)} = \mathrm{Tr}\left(\Sigma_\varepsilon^{-1}-\frac{1}{\lambda_\varepsilon}\mathrm{I}_2\right)=\mathcal{O}(\varepsilon^2)\,.$$

Now, as in the solitary-wave limit, we introduce

$$\delta_1^{(0)} := \frac{\sigma_1}{X_x^{(0)}}\,,\quad \delta_2^{(0)} := \frac{\sigma_2+\nu(\rho^{(0)})\sigma_1}{X_x^{(0)}}\,,\quad \delta_3^{(0)} := \frac{\sigma_3+2\nu(\rho^{(0)})\sigma_2+\nu(\rho^{(0)})^2\sigma_1}{X_x^{(0)}}\,,$$

so that

$$-\mathbf{P}_0\mathbf{B}_0^{-1}\mathbf{C}_0\mathbf{P}_0^{-1} = -X_x^{(0)}\begin{pmatrix} 0 & 0 & 0 & 0\\ \delta_3^{(0)} & 0 & -\frac{w_0}{2\sigma_0}\delta_2^{(0)} & \delta_2^{(0)}\\ \sigma_0\delta_2^{(0)} & 0 & -\frac{w_0}{2}\delta_1^{(0)} & \sigma_0\delta_1^{(0)}\\ \frac{w_0}{2}\delta_2^{(0)} & 0 & -\frac{w_0^2}{4\sigma_0}\delta_1^{(0)} & \frac{w_0}{2}\delta_1^{(0)}\end{pmatrix},$$

with

$$
\begin{aligned}
\delta_1^{(0)} &= f_1(\rho^{(0)}) + \mathcal{O}\left(\varepsilon^2\right),\\
\delta_2^{(0)} &= 2\,\nu(\rho^{(0)})\, f_1(\rho^{(0)}) + \mathcal{O}\left(\varepsilon^2\right),\\
\delta_3^{(0)} &= 4\,\nu(\rho^{(0)})^2\, f_1(\rho^{(0)}) + \mathcal{O}\left(\varepsilon^2\right).
\end{aligned}
$$

Then we compute that

$$
\begin{aligned}
&-\frac{1}{X_x^{(0)}}\begin{pmatrix}1&0&0\\0&\frac{1}{\varepsilon}&0\\0&0&\mathrm{I}_2\end{pmatrix}\widetilde{\mathbf{P}}_0^{-1}\mathbf{P}_0\mathbf{B}_0^{-1}\mathbf{C}_0\,\mathbf{P}_0^{-1}\widetilde{\mathbf{P}}_0\begin{pmatrix}1&0&0\\0&\varepsilon&0\\0&0&\mathrm{I}_2\end{pmatrix}\\
&\qquad=\begin{pmatrix}\mathcal{O}(1)&\mathcal{O}(\varepsilon)&\mathcal{O}(1)&\mathcal{O}(1)\\ \frac{\delta_3^{(0)}}{\varepsilon}&\mathcal{O}(1)&\mathcal{O}\left(\varepsilon^{-1}\right)&\mathcal{O}\left(\varepsilon^{-1}\right)\\ \mathcal{O}(1)&\mathcal{O}(\varepsilon)&\mathcal{O}(1)&\mathcal{O}(1)\\ \mathcal{O}(1)&\mathcal{O}(\varepsilon)&\mathcal{O}(1)&\mathcal{O}(1)\end{pmatrix}
\end{aligned}
$$

and observe that the latter matrix takes the form

$$
\text{matrix of rank } 1\times(\,\mathrm{I}_4+\mathcal{O}(\varepsilon^2)\,)\,.
$$

As a result,

$$
\begin{aligned}
\tilde{\delta}_{(2,0,2)} &= \mathcal{O}(\varepsilon^2)\,, &\qquad \tilde{\delta}_{(3,0,1)} &= -\delta_3^{(0)}\,\mathfrak{d}_0\,\det\left(\Sigma_\varepsilon^{-1}-\frac{1}{\lambda_\varepsilon}\mathrm{I}_2\right)+\mathcal{O}(\varepsilon^2)\,,\\
\tilde{\delta}_{(2,1,1)} &= \mathcal{O}(1)\,, & \tilde{\delta}_{(1,2,1)} &= \mathcal{O}(1)\,.
\end{aligned}
$$

Taking the limit $\varepsilon\to 0$ in

$$
\Delta_0\left(\frac{\lambda_\varepsilon}{X_x^{(0)}}\left(-\varepsilon^{-\frac14}Z+\varepsilon^{\frac14}\Lambda\right),\varepsilon^{-\frac14}Z,\varepsilon^{\frac14}\frac{\Gamma}{X_x^{(0)}}\right)
$$

yields the limiting

$$
\Lambda^2\,Z^2\,\det\left(\Sigma_0^{-1}-\frac{1}{\lambda_0}\mathrm{I}_2\right)-\delta_3^{(0)}\,\mathfrak{d}_0\,\det\left(\Sigma_0^{-1}-\frac{1}{\lambda_0}\mathrm{I}_2\right)\lambda_0^2\,Z^2\,\Gamma^2\;=0\,,
$$

where we recall $\det\left(\Sigma_0^{-1}-\frac{1}{\lambda_0}\mathrm{I}_2\right)=-16\mathfrak{c}_0^2\sigma_0^2\delta_{hyp}$. From this, one deduces that when $\delta_{hyp}\neq 0$, $\delta_{BF}>0$ gives instability since $\mathfrak{d}_0$ has the sign of δ_{BF}.

Since Theorem 4.6 already concludes instability from $\delta_{BF}>0$ this concludes the proof of Theorem 5.8.

Appendix A. Symmetries and conservation laws

In the present paper, including the current section, we only consider functional densities depending on derivatives up to order 1. In particular,

$$\begin{aligned} L\mathcal{A}[U](\mathbf{J}\delta\mathcal{B}[U]) &= -L\mathcal{B}[U](\mathbf{J}\delta\mathcal{A}[U]) \\ &\quad + \sum_\ell \partial_\ell \left(\nabla_{U_{x_\ell}}\mathcal{A}[U]\cdot\mathbf{J}\delta\mathcal{B}[U] + \nabla_{U_{x_\ell}}\mathcal{B}[U]\cdot\mathbf{J}\delta\mathcal{A}[U]\right). \end{aligned}$$

As a consequence, if $U_t = \mathbf{J}\delta\mathcal{H}[U]$, then

$$\begin{aligned} (\mathcal{G}[U])_t &= -L\mathcal{H}[U](\mathbf{J}\delta\mathcal{G}[U]) \\ &\quad + \sum_\ell \partial_\ell \left(\nabla_{U_{x_\ell}}\mathcal{G}[U]\cdot\mathbf{J}\delta\mathcal{H}[U] + \nabla_{U_{x_\ell}}\mathcal{H}[U]\cdot\mathbf{J}\delta\mathcal{G}[U]\right). \end{aligned} \tag{112}$$

The main point in concrete uses of the latter equality is that $U \mapsto L\mathcal{H}[U](\mathbf{J}\delta\mathcal{G}[U])$ encodes the variation of $\mathcal{H}$ under the action of the group generated by $\mathcal{G}$. Here, we consider two kinds of invariance by the action of a group generated by a functional density:

- stationarity of the density functional $\mathcal{H}$ under the action of the group generated by $\mathcal{G}$ encoded by
$$L\mathcal{H}[U](\mathbf{J}\delta\mathcal{G}[U]) \equiv 0\,,$$
in which case (112) reduces to
$$(\mathcal{G}[U])_t = \sum_\ell \partial_\ell \left(\nabla_{U_{x_\ell}}\mathcal{G}[U]\cdot\mathbf{J}\delta\mathcal{H}[U] + \nabla_{U_{x_\ell}}\mathcal{H}[U]\cdot\mathbf{J}\delta\mathcal{G}[U]\right)\;;$$
- commutation of the density functional $\mathcal{H}$ with the action of the group generated by $\mathcal{G}$ encoded by
$$L\mathcal{H}[U](\mathbf{J}\delta\mathcal{G}[U]) = \mathbf{J}\delta\mathcal{G}[\mathcal{H}[U]]\,,$$
in which case (112) reduces to
$$(\mathcal{G}[U])_t = -\mathbf{J}\delta\mathcal{G}[\mathcal{H}[U]] + \sum_\ell \partial_\ell \left(\nabla_{U_{x_\ell}}\mathcal{G}[U]\cdot\mathbf{J}\delta\mathcal{H}[U] + \nabla_{U_{x_\ell}}\mathcal{H}[U]\cdot\mathbf{J}\delta\mathcal{G}[U]\right)\,.$$

Note that in the latter case if the group is a group of translations, then the latter equation is still a conservation law.

Specializing the first case to $\mathcal{G} = \mathscr{M}$ gives

$$(\mathscr{M}[U])_t = \sum_\ell \partial_\ell \left(\nabla_{U_{x_\ell}}\mathcal{H}[U]\cdot\mathbf{J}U\right),$$

whereas a specialization of the second case to $\mathcal{G} = \mathbb{Q}_j$ and to $\mathcal{G} = \mathcal{H}$ gives, respectively,

$$(\mathbb{Q}_j[U])_t = \partial_j(\nabla_{U_{x_j}}\mathbb{Q}_j[U]\cdot\mathbf{J}\delta\mathcal{H}[U] - \mathcal{H}[U]) + \sum_\ell \partial_\ell \left(\nabla_{U_{x_\ell}}\mathcal{H}[U]\cdot U_{x_j}\right)$$

and

$$(\mathcal{H}[U])_t = \sum_\ell \partial_\ell \left(\nabla_{U_{x_\ell}} \mathcal{H}[U] \cdot \mathbf{J}\delta\mathcal{H}[U]\right) .$$

To compute how these conservation laws are transformed when going to uniformly moving frames we also record the following simple but useful relations

$$\begin{aligned}
L\mathscr{M}[U](\mathbf{J}\delta\mathscr{M}[U]) &= 0\,, & L\mathscr{M}[U](\mathbf{J}\delta\mathbb{Q}_\ell[U]) &= \partial_\ell(\mathscr{M}[U])\,,\\
L\mathbb{Q}_j[U](\mathbf{J}\delta\mathscr{M}[U]) &= -\partial_j(\mathscr{M}[U])\,, & L\mathbb{Q}_j[U](\mathbf{J}\delta\mathbb{Q}_\ell[U]) &= \partial_\ell(\mathbb{Q}_j[U])\,,\\
L\mathcal{H}[U](\mathbf{J}\delta\mathscr{M}[U]) &= 0\,, & L\mathcal{H}[U](\mathbf{J}\delta\mathbb{Q}_\ell[U]) &= \partial_\ell(\mathcal{H}[U])\,.
\end{aligned}$$

Among the foregoing identities only the third one is not a simple expression of invariances of $\mathscr{M}$, $\mathbb{Q}_j$ and $\mathcal{H}$, but it may be deduced from the second one.

For our purposes, it is also crucial to derive linearized versions of the algebraic relations expounded above. Let us define $\mathcal{F}_{\mathcal{G},\mathcal{H}}$ by $\mathcal{F}_{\mathcal{G},\mathcal{H}}[U] = L\mathcal{G}[U](\mathbf{J}\delta\mathcal{H}[U])$, so that $U_t = \mathbf{J}\delta\mathcal{H}[U]$ implies $(\mathcal{G}(U))_t = \mathcal{F}_{\mathcal{G},\mathcal{H}}[U]$. Now note that if $\underline{U}$ is such that $\delta\mathcal{H}[\underline{U}] = 0$, then $L\mathcal{F}_{\mathcal{G},\mathcal{H}}[\underline{U}](V) = L\mathcal{G}[\underline{U}](\mathbf{J}L\delta\mathcal{H}[\underline{U}](V))$. In particular, if $\underline{U}_t = 0$ and $\delta\mathcal{H}[\underline{U}] = 0$, then $V_t = \mathbf{J}L\delta\mathcal{H}[\underline{U}](V)$ implies

$$(L\mathcal{G}[\underline{U}]V)_t = L\mathcal{F}_{\mathcal{G},\mathcal{H}}[\underline{U}](V)\,.$$

The latter computations also provide similar conclusions for the associated spectral problems.

Appendix B. Spectral stability of constant states

In the present section, we study the spectral stability of constant solutions to (3). By constant solutions we mean solutions that are constant up to the symmetries, thus solutions in the form[36]

$$\mathbf{U}(t,\mathbf{x}) = \mathrm{e}^{(\mathbf{k}_\phi\cdot\mathbf{x}+\omega_\phi\, t)\mathbf{J}}\,\mathbf{U}^{(0)}\,, \tag{113}$$

with $\mathbf{U}^{(0)}$ a constant vector of $\mathbf{R}^2$, $\mathbf{k}_\phi \in \mathbf{R}^d$, and $\omega_\phi \in \mathbf{R}$. Since it is almost costless and will be useful in Appendix C, we consider an even more general class of Hamiltonian equations

$$\partial_t\mathbf{U} = \mathbf{J}\,\delta\mathscr{H}_0[\mathbf{U}]\,,\qquad \text{with}\quad \mathscr{H}_0\,[\mathbf{U}] = \tfrac{1}{2}\nabla_\mathbf{x}\mathbf{U}\cdot\mathbf{D}(\|\mathbf{U}\|^2)\nabla_\mathbf{x}\mathbf{U} + W(\|\mathbf{U}\|^2)\,, \tag{114}$$

where $\mathbf{D}$ is valued in real symmetric $d\times d$-matrices. That $\mathbf{U}$ from (113) solves (114) reduces to either $\mathbf{U}^{(0)}$ is zero or

$$\omega_\phi = 2\,W'(\|\mathbf{U}^{(0)}\|^2) + \mathbf{k}_\phi\cdot\mathbf{D}(\|\mathbf{U}^{(0)}\|^2)\,\mathbf{k}_\phi + \|\mathbf{U}^{(0)}\|^2\,\mathbf{k}_\phi\cdot\mathbf{D}'(\|\mathbf{U}^{(0)}\|^2)\,\mathbf{k}_\phi\,. \tag{115}$$

36. The action of spatial translations is redundant with the action of rotations for this class of solutions.

For the sake of concision and comparison, it is expedient to introduce the dispersionless hydrodynamic Hamiltonian

$$H^{(0)}(\rho,\mathbf{v}) := \rho\,\mathbf{v}\cdot\mathbf{D}(2\,\rho)\mathbf{v} + W(2\,\rho)\,.$$

As a first instance, note that with $\rho^{(0)} := \mathscr{M}(\mathbf{U}^{(0)})$, (115) takes the concise form $\omega_\phi = \partial_\rho H^{(0)}(\rho^{(0)},\mathbf{k}_\phi)$.

Changing frame through $\mathbf{U}(t,\mathbf{x}) = \mathrm{e}^{(\mathbf{k}_\phi\cdot\mathbf{x}+\omega_\phi\,t)\mathbf{J}}\,\mathbf{V}(t,\mathbf{x})$, linearizing, and using the Fourier transform brings the spectral stability question under consideration to the question of knowing whether for any $\boldsymbol{\xi}\in\mathbf{R}^d$, the linear operator on $\mathbf{C}^2$

$$\begin{aligned}\mathbf{V}\mapsto &\Big[\,\mathbf{U}^{(0)}\cdot\mathbf{V}\times\partial_\rho^2 H^{(0)}(\rho^{(0)},\mathbf{k}_\phi) + 2\,\mathbf{J}\mathbf{U}^{(0)}\cdot\mathbf{V}\times\mathbf{k}_\phi\cdot\mathbf{D}'(2\,\rho^{(0)})\,\mathrm{i}\,\boldsymbol{\xi}\Big]\;\mathbf{J}\mathbf{U}^{(0)}\\ &+\Big[2\,\mathbf{U}^{(0)}\cdot\mathbf{V}\times\mathbf{k}_\phi\cdot\mathbf{D}'(2\,\rho^{(0)})\,\mathrm{i}\,\boldsymbol{\xi}\Big]\;\mathbf{U}^{(0)}\\ &+\Big[\boldsymbol{\xi}\cdot\mathbf{D}(2\,\rho^{(0)})\boldsymbol{\xi}\Big]\;\mathbf{J}\mathbf{V} + \Big[2\,\mathbf{k}_\phi\cdot\mathbf{D}(2\,\rho^{(0)})\,\mathrm{i}\,\boldsymbol{\xi}\Big]\;\mathbf{V}\end{aligned}$$

has purely imaginary spectrum. If $\mathbf{U}^{(0)}=0$, then, by diagonalizing $\mathbf{J}$, one gets that the latter spectrum is

$$\mathrm{i}\,\Big(\pm\boldsymbol{\xi}\cdot\mathbf{D}(2\,\rho^{(0)})\boldsymbol{\xi} + 2\,\mathbf{k}_\phi\cdot\mathbf{D}(2\,\rho^{(0)})\boldsymbol{\xi}\Big)\in\mathrm{i}\,\mathbf{R}\,,$$

hence spectral stability holds. When $\mathbf{U}^{(0)}\neq 0$, we may use $\mathbf{V}\mapsto(\mathbf{U}^{(0)}\cdot\mathbf{V},\mathbf{J}\mathbf{U}^{(0)}\cdot\mathbf{V})$ as a coordinate map in which the above operator's matrix is

$$\begin{pmatrix} 2\,\partial_\rho\nabla_{\mathbf{v}}H^{(0)}(\rho^{(0)},\mathbf{k}_\phi)\cdot\mathrm{i}\,\boldsymbol{\xi} & -\boldsymbol{\xi}\cdot\mathbf{D}(2\,\rho^{(0)})\boldsymbol{\xi}\\ 2\,\rho^{(0)}\,\partial_\rho^2 H^{(0)}(\rho^{(0)},\mathbf{k}_\phi)+\boldsymbol{\xi}\cdot\mathbf{D}(2\,\rho^{(0)})\boldsymbol{\xi} & 2\,\partial_\rho\nabla_{\mathbf{v}}H^{(0)}(\rho^{(0)},\mathbf{k}_\phi)\cdot\mathrm{i}\,\boldsymbol{\xi}\end{pmatrix}.$$

Thus, when $\mathbf{U}^{(0)}\neq 0$, spectral stability holds if and only if for any $\boldsymbol{\xi}\in\mathbf{R}^d$, the solutions in λ of

$$\begin{aligned}&\Big(\lambda - 2\,\mathrm{i}\;\partial_\rho\nabla_{\mathbf{v}}H^{(0)}(\rho^{(0)},\mathbf{k}_\phi)\cdot\boldsymbol{\xi}\Big)^2\\ &+\boldsymbol{\xi}\cdot\mathbf{D}(2\,\rho^{(0)})\boldsymbol{\xi}\;\Big(2\,\rho^{(0)}\,\partial_\rho^2 H^{(0)}(\rho^{(0)},\mathbf{k}_\phi)+\boldsymbol{\xi}\cdot\mathbf{D}(2\,\rho^{(0)})\boldsymbol{\xi}\Big)=0\end{aligned}$$

are purely imaginary, that is, if and only if, for any $\boldsymbol{\xi}\in\mathbf{R}^d$,

$$\boldsymbol{\xi}\cdot\mathbf{D}(2\,\rho^{(0)})\boldsymbol{\xi}\;\Big(2\,\rho^{(0)}\,\partial_\rho^2 H^{(0)}(\rho^{(0)},\mathbf{k}_\phi)+\boldsymbol{\xi}\cdot\mathbf{D}(2\,\rho^{(0)})\boldsymbol{\xi}\Big)\;\geq\;0\,.$$

As a conclusion, spectral stability holds if and only if, for any unitary $\mathbf{e}\in\mathbf{R}^d$,

$$\mathbf{e}\cdot\mathbf{D}(2\,\rho^{(0)})\;\mathbf{e}\times\rho^{(0)}\,\partial_\rho^2 H^{(0)}(\rho^{(0)},\mathbf{k}_\phi)\;\geq\;0\,.$$

For comparison, let us point out that $\partial_\rho^2 H^{(0)}(\rho^{(0)},\mathbf{k}_\phi) = 4\,\delta_{hyp}$.

Lemma B.1. — *Let* $\mathbf{U}^{(0)}$ *be a constant profile for a solution to* (114) *in the sense of* (113). *Then, with* $\rho^{(0)} := \mathscr{M}(\mathbf{U}^{(0)})$, *the corresponding solution is* *spectrally exponentially unstable if and only if* $\mathbf{D}(2\,\rho^{(0)})$ *is not the zero matrix,* $\rho^{(0)}\,\delta_{hyp}\neq 0$*, and one of the two following possibilities hold*

1. *there exists* $\mathbf{e}_+$ *such that* $\mathbf{e}_+\cdot\mathbf{D}(2\,\rho^{(0)})\,\mathbf{e}_+ > 0$ *and* $\mathbf{e}_-$ *such that* $\mathbf{e}_-\cdot\mathbf{D}(2\,\rho^{(0)})\,\mathbf{e}_- < 0$*;*
2. $\mathbf{D}(2\,\rho^{(0)})$ *is nonnegative (respectively nonpositive) and* $\delta_{hyp}<0$ *(respectively,* $\delta_{hyp}>0$*).*

Remark B.2. — In the present contribution, we are interested in constant solutions only as far as they are reachable either as the constant limit in the small-amplitude regime or as the limiting solitary-wave end state in the large-period regime. Extending [10, Appendix A], let us point out that when $\mathbf{D}$ has a sign (either nonnegative or nonpositive), constant states associated with a large-period regime are always spectrally stable. We prove here this claim for equations of type (3). Let us recall that with ν and $\mathcal{W}_\rho$ defined through

$$\begin{aligned}\mathcal{W}_\rho(\rho)&=-H^{(0)}(\rho,\nu(\rho)\,\mathbf{e}_x+\widetilde{\mathbf{k}}_\phi)+\omega_\phi\,\rho+\mu_\phi\,\nu(\rho)-c_x\rho\,\nu(\rho)\,,\\ 0&=-\,\mathbf{e}_x\cdot\nabla_{\mathbf{v}}H^{(0)}(\rho,\nu(\rho)\,\mathbf{e}_x+\widetilde{\mathbf{k}}_\phi)+\mu_\phi-c_x\,\rho\,,\end{aligned}$$

this means that we focus on the case when $\partial_\rho^2\mathcal{W}_\rho(\rho^{(0)})<0$. By differentiating the foregoing identities, one deduces that

$$\begin{aligned}\partial_\rho\mathcal{W}_\rho(\rho)&=-\partial_\rho H^{(0)}(\rho,\nu(\rho)\,\mathbf{e}_x+\widetilde{\mathbf{k}}_\phi)+\omega_\phi-c_x\nu(\rho)\,,\\ 2\,\rho\,\kappa(2\,\rho)\,\nu'(\rho)&=-\,\mathbf{e}_x\cdot\partial_\rho\nabla_{\mathbf{v}}H^{(0)}(\rho,\nu(\rho)\,\mathbf{e}_x+\widetilde{\mathbf{k}}_\phi)-c_x\,,\\ \partial_\rho^2\mathcal{W}_\rho(\rho)&=-\partial_\rho^2H^{(0)}(\rho,\nu(\rho)\,\mathbf{e}_x+\widetilde{\mathbf{k}}_\phi)+\frac{(\nu'(\rho))^2}{2\,\rho\,\kappa(2\,\rho)}\,.\end{aligned}$$

From this stems that, for $(\rho^{(0)},\mathbf{k}_\phi)=(\rho^{(0)},\nu(\rho^{(0)})\,\mathbf{e}_x+\widetilde{\mathbf{k}}_\phi)$, the saddle condition $\partial_\rho^2\mathcal{W}_\rho(\rho^{(0)})<0$ implies $\partial_\rho^2H^{(0)}(\rho,\mathbf{k}_\phi)>0$ *i.e.*, $\delta_{hyp}>0$, as claimed.

Appendix C. Anisotropic equations

In the present section, we show how to generalize most of our results from systems of the form (3) to systems of the form (114), namely,

$$\partial_t\mathbf{U}=\mathbf{J}\,\delta\mathscr{H}_0[\mathbf{U}]\,,\qquad\text{with}\quad\mathscr{H}_0\,[\mathbf{U}]=\tfrac12\nabla_{\mathbf{x}}\mathbf{U}\cdot\mathbf{D}(\|\mathbf{U}\|^2)\nabla_{\mathbf{x}}\mathbf{U}+W(\|\mathbf{U}\|^2)\,,$$

where $\mathbf{D}$ is valued in real symmetric $d\times d$-matrices. As in Appendix D, our goal is not to transfer our methodology (with possibly different outcomes) but to point out what is readily accessible by simple changes in notation.

Consistently with the rest of the present paper, we shall discuss explicitly only waves in the form (7). Yet let us anticipate from Appendix D that even for System (114) as considered here, all longitudinal results apply equally well to waves of the form (15) and that instability results about general perturbations also generalize when either $\mathbf{D}$ is constant (semilinear case) or when $d \geq 3$, and, for any α, $\widetilde{\mathbf{k}}_\phi$ is an eigenvector of $\mathbf{D}(\alpha)$.

The restriction on generality that we make here is that we consider waves of type (7) propagating in a direction that is a principal direction for the dispersion of (114). We assume that, for any α, $\underline{\mathbf{e}}_x$ is an eigenvector of $\mathbf{D}(\alpha)$ for a nonzero eigenvalue. This includes the case, considered in [40], that when $d = 2$, waves propagate in the direction $\mathbf{e}_1$ and

$$\mathbf{D} \equiv \begin{pmatrix} 1 & 0 \\ 0 & \pm 1 \end{pmatrix}.$$

It follows readily from the principal-direction restriction that all longitudinal results still hold with

$$\kappa(\alpha) := \underline{\mathbf{e}}_x \cdot \mathbf{D}(\alpha)\, \underline{\mathbf{e}}_x\,, \tag{116}$$

and we recall that at this stage, there is no loss in generality in assuming κ positive valued. Unfortunately, in genuinely anisotropic cases, the principal-direction restriction is essentially incompatible with modulation of the direction $\mathbf{e}_x$ and, thus, rules out any hope for a modulational interpretation in the spirit of Section 5.1.

Therefore, under this assumption, we focus on extending the instability results of Section 5. As far as this goal is concerned, it is sufficient to deal with the case when $d = 2$, $\underline{\mathbf{e}}_x = \mathbf{e}_1$, and

$$\mathscr{H}_0\,[\mathbf{U}] = \tfrac{1}{2}\kappa(\|\mathbf{U}\|^2)\|\partial_x \mathbf{U}\|^2 + W(\|\mathbf{U}\|^2) + \tfrac{1}{2}\widetilde{\kappa}(\|\mathbf{U}\|^2)\|\partial_y \mathbf{U}\|^2\,, \tag{117}$$

with κ as in (116) and $\widetilde{\kappa}$ ranging over all the functions $\widetilde{\kappa}$ given by

$$\widetilde{\kappa}(\alpha) := \mathbf{e} \cdot \mathbf{D}(\alpha)\ \mathbf{e}\,,$$

where $\mathbf{e}$ is a unitary vector orthogonal to $\underline{\mathbf{e}}_x$. Note that this reduction hinges on the obvious facts that there is no loss in taking $\boldsymbol{\eta}$ under the form $\|\boldsymbol{\eta}\|\ \mathbf{e}$ with $\mathbf{e}$ as above, and that, for any α, the space of vectors orthogonal to $\underline{\mathbf{e}}_x$ is stable under the action of $\mathbf{D}(\alpha)$.

Up to minor changes that we detail below, Corollary 5.1 and results of Section 5.2 extend readily to the case (117) with $\underline{\mathbf{e}}_x = \mathbf{e}_1$. Indeed, the changes required in Theorem 3.2 and its proof are purely notational, and in the statement, the only place where κ should be replaced with $\widetilde{\kappa}$ is in the definition of $\Sigma_{\mathbf{y}}$, or, in other words, in the definition of σ_1, σ_2 and σ_3. Explicitly,

$$\sigma_1 := \int_0^{\underline{X}_x} \widetilde{\kappa}(\|\underline{\mathcal{V}}\|^2)\, \|\underline{\mathcal{V}}\|^2\,, \qquad \sigma_2 := \int_0^{\underline{X}_x} \widetilde{\kappa}(\|\underline{\mathcal{V}}\|^2)\, \mathbf{J}\underline{\mathcal{V}} \cdot \underline{\mathcal{V}}_x\,,$$
$$\sigma_3 := \int_0^{\underline{X}_x} \widetilde{\kappa}(\|\underline{\mathcal{V}}\|^2)\, \|\underline{\mathcal{V}}_x\|^2\,.$$

The proof of Proposition 3.1 requires more significant changes, but all of them are elementary. The upshot is that in Proposition 3.1, the ellipticity condition

$|\lambda| + \|\boldsymbol{\eta}\|^2 \geq R_0$ should be replaced with $|\lambda| \geq R_0\,(1 + \|\boldsymbol{\eta}\|^2)$. This weaker conclusion is still sufficient to derive Corollary 5.1. Once the above-mentioned change in $\Sigma_{\mathbf{y}}$ has been performed, all the results of Section 5.2 hold unchanged.

Note, for instance, that in Lemma 5.2, only the coefficients $\delta_{(m,n,p)}$ with $p \neq 0$ depend on the choice of the transverse coefficient $\widetilde{\kappa}$. This stems from the fact that the wave profiles are independent of this coefficient. Note, moreover, that the dependence of $\delta_{(m,n,p)}$ on $\widetilde{\kappa}$ has the parity of p. Thus, it follows from Lemma 5.2 that waves cannot be spectrally stable to perturbations that are longitudinally coperiodic for both $\widetilde{\kappa}$ and $-\widetilde{\kappa}$ except possibly if $\delta_{4,0,0} = \delta_{3,0,1} = \delta_{2,0,2} = 0$. Note that in the latter degenerate case, in particular, 0 has algebraic multiplicity larger than 4 as an eigenvalue of $\mathcal{L}_{0,0}$. Moreover, it follows from an inspection of the coefficients of Σ_t and $\Sigma_{\mathbf{y}}$ and a Cauchy–Schwarz argument that this latter degenerate case cannot occur when $\widetilde{\kappa}$ has a definite sign (either positive or negative).

Now we turn to the generalization of asymptotic results in Sections 5.3 and 5.4. It is important to track there how the replacement of κ with $\widetilde{\kappa}$ at some places impacts the proof. In the integral representations of σ_1, σ_2 and σ_3, κ should be replaced with $\widetilde{\kappa}$ in definitions of f_1, f_2, and f_3, and the formula for σ_3 should be modified as

$$\begin{aligned}\sigma_3 = &\int_{\rho_{\min}(\mu_x)}^{\rho_{\max}(\mu_x)} \frac{f_3(\rho)}{\sqrt{\mu_x - \mathcal{W}_\rho(\rho)}}\sqrt{\frac{2\,\kappa(2\,\rho)}{2\,\rho}}\,\mathrm{d}\,\rho\\ &+\int_{\rho_{\min}(\mu_x)}^{\rho_{\max}(\mu_x)} \frac{\widetilde{\kappa}(2\,\rho)}{\kappa(2\,\rho)}\sqrt{\mu_x - \mathcal{W}_\rho(\rho)}\sqrt{\frac{2\,\kappa(2\,\rho)}{2\,\rho}}\,\mathrm{d}\,\rho\,.\end{aligned}$$

These changes appear in proofs of Theorems 5.6 and 5.8 only through the value $f_1(2\,\rho^{(0)})$. When $\widetilde{\kappa}(2\,\rho^{(0)}) > 0$, the arguments still apply, so that instability still occurs.

Let us now focus on the case when $\widetilde{\kappa}(2\,\rho^{(0)}) < 0$. Recall that we have normalized signs to ensure $\kappa(2\,\rho^{(0)}) > 0$. Thus, it follows from Lemma B.1 that if $\delta_{hyp} \neq 0$, the limiting constant state is spectrally unstable. Note that, as pointed out in Remark B.2, the condition $\delta_{hyp} \neq 0$ holds systematically at the large-period limit. So we only need to explain how to transfer spectral instability from limiting constant states to nearby periodic waves.

In the small-amplitude regimes, the transfer follows from a direct standard perturbation argument for isolated eigenvalues of finite multiplicity, considering the constant-coefficient operators obtained by linearizing about the constant state as a periodic operator. To perform this comparison it is important, even at the constant limit, to choose a frame adapted to the harmonic limit. Indeed, let us observe that the choice of a frame – among those in which the reference solution is stationary – does impact the spectrum of the linearized operator, yet without altering the instable character of this spectrum. We omit details of the standard argument and, again, refer the reader to [38] for background on spectral perturbation theory. The large-period regime is trickier to analyze but is covered by [57].

To summarize, we have obtained that under the principal-direction assumption, when $\mathbf{D}(2\,\rho^{(0)})$ is nontrivial – in the sense that there exists $\mathbf{e}$ orthogonal to $\underline{\mathbf{e}}_x$ such that $\mathbf{D}(2\,\rho^{(0)})\,\mathbf{e}$ is not zero –, spectral instability holds

1. in the large-period regime when at the limiting solitary wave $\partial_{c_x}^2\Theta_{(s)} \neq 0$;
2. in the small-amplitude regime when at the limiting constant $\partial_\rho \nu \neq 0$ and $\delta_{hyp}\,\delta_{BF} \neq 0$.

Appendix D. General plane waves

In the present section, we show how our general analysis may be applied to more general plane waves in the form (15). Whereas we believe that our methodology can be applied to all these waves (with possibly different outcomes), our aim here is merely to point out what is readily accessible by a simple change in frame or notation.

As was already highlighted in Section 2.6, all our longitudinal results apply as they are, once one has replaced W with $W_{\widetilde{\mathbf{k}}_\phi}$ defined through

$$W_{\widetilde{\mathbf{k}}_\phi}(\alpha) := W(\alpha) + \frac{1}{2}\,\alpha\,\kappa(\alpha)\,\|\widetilde{\mathbf{k}}_\phi\|^2 = H^{(0)}\left(\frac{\alpha}{2}, \widetilde{\mathbf{k}}_\phi\right).$$

Thus, we only need to discuss our results on general perturbations, focused on proving spectral exponential instability.

Dimension larger than 2. — A simple but efficient observation is that when one restricts oneself to perturbations that are constant in the direction of $\widetilde{\mathbf{k}}_\phi$, all transverse contributions due to the fact that $\widetilde{\mathbf{k}}_\phi$ is nonzero disappear. As a direct consequence, when $\widetilde{\mathbf{k}}_\phi \neq 0$ but $d \geq 3$, all the spectral instability results hold as they are (up to the change $W \to W_{\widetilde{\mathbf{k}}_\phi}$), and a modulational interpretation is available for the spectral expansion of $D_\xi(\lambda, \boldsymbol{\eta})$ when $(\lambda, \xi, \boldsymbol{\eta}) \to (0,0,0)$ under the condition $\boldsymbol{\eta} \cdot \widetilde{\mathbf{k}}_\phi = 0$. In particular, when $d \geq 3$, in nondegenerate cases, spectral instability occurs in both small-amplitude and large-period regimes. Except for the generalization of the modulational interpretation, this argument also applies to the more general form of the equations considered in Appendix C under the assumption that, for any α, $\widetilde{\mathbf{k}}_\phi$ is an eigenvector of $\mathbf{D}(\alpha)$.

The semilinear case. — In the semilinear case, one may go further by using a form of Galilean invariance. Let us consider System (114) with $\mathbf{D} \equiv \mathbf{D}_0$. Then for any vector $\widetilde{\mathbf{k}}_\phi$, if $\mathbf{U}$ solves (114) so does

$$(t, \mathbf{x}) \mapsto \mathrm{e}^{\left(\widetilde{\mathbf{k}}_\phi \cdot \mathbf{D}_0 \widetilde{\mathbf{k}}_\phi\, t + \widetilde{\mathbf{k}}_\phi \cdot \mathbf{x}\right)\mathbf{J}}\,\mathbf{U}(t, \mathbf{x} + t\,2\,\mathbf{D}_0\,\widetilde{\mathbf{k}}_\phi)\,.$$

The foregoing transformation preserves (in)stability properties and brings waves of type (15) into waves of type (7). Thus, in the semilinear case, there is absolutely no loss in generality in assuming the form (7).

Appendix E. Table of symbols

Here we have gathered page numbers of main definitions for symbols that are used recurrently throughout the text. Pieces of notation specific to a section are not indexed here. For groups of symbols introduced simultaneously, the definitions may run over a few pages. We recall that underlining is used throughout to denote specialization at a specific background wave.

BIBLIOGRAPHY

[1] B. Barker – "Numerical proof of stability of roll waves in the small-amplitude limit for inclined thin film flow", *J. Differential Equations* **257** (2014), no. 8, p. 2950–2983.

[2] B. Barker, M. A. Johnson, P. Noble, L. M. Rodrigues & K. Zumbrun – "Nonlinear modulational stability of periodic traveling-wave solutions of the generalized Kuramoto–Sivashinsky equation", *Phys. D* **258** (2013), p. 11–46.

[3] ______, "Note on the stability of viscous roll-waves", *Comptes Rendus Mécanique* **345** (2017), no. 2, p. 125–129.

[4] ______, "Stability of viscous St. Venant roll waves: from onset to infinite Froude number limit", *J. Nonlinear Sci.* **27** (2017), no. 1, p. 285–342.

[5] T. B. Benjamin – "The stability of solitary waves", *Proc. Roy. Soc. (London) Ser. A* **328** (1972), p. 153–183.

[6] ———, "Impulse, flow force and variational principles", *IMA J. Appl. Math.* **32** (1984), no. 1-3, p. 3–68.

[7] S. Benzoni-Gavage – "Spectral transverse instability of solitary waves in Korteweg fluids", *J. Math. Anal. Appl.* **361** (2010), no. 2, p. 338–357.

[8] ———, "Planar traveling waves in capillary fluids", *Differential Integral Equations* **26** (2013), no. 3-4, p. 439–485.

[9] S. Benzoni-Gavage, C. Mietka & L. M. Rodrigues – "Co-periodic stability of periodic waves in some Hamiltonian pdes", *Nonlinearity* **29** (2016), no. 11, p. 3241–3308.

[10] ———, "Stability of periodic waves in Hamiltonian PDEs of either long wavelength or small amplitude", *Indiana Univ. Math. J.* **69** (2020), no. 2, p. 545–619.

[11] ———, "Modulated equations of Hamiltonian PDEs and dispersive shocks", *Nonlinearity* **34** (2021), no. 1, p. 578–641.

[12] S. Benzoni-Gavage, P. Noble & L. M. Rodrigues – "Stability of periodic waves in Hamiltonian PDEs", 2013, `http://jedp.cedram.org/cedram-bin/article/JEDP_2013____A2_0.pdf`, p. 1–22.

[13] ———, "Slow modulations of periodic waves in Hamiltonian PDEs, with application to capillary fluids", *J. Nonlinear Sci.* **24** (2014), no. 4, p. 711–768.

[14] S. D. Bièvre & S. R. Nodari – "Orbital stability via the energy-momentum method: the case of higher dimensional symmetry groups", *Arch. Ration. Mech. Anal.* **231** (2019), no. 1, p. 233–284.

[15] N. Bottman, B. Deconinck & M. Nivala – "Elliptic solutions of the defocusing NLS equation are stable", *J. Phys. A* **44** (2011), no. 28, p. 285201–285224.

[16] J. Boussinesq – "Théorie des ondes et des remous qui se propagent le long d'un canal rectangulaire horizontal, en communiquant au liquide contenu dans ce canal des vitesses sensiblement pareilles de la surface au fond", *J. Math. Pures Appl.* **17** (1872), no. 2, p. 55–108.

[17] J. C. Bronski, V. M. Hur & M. A. Johnson – "Modulational instability in equations of KdV type", Lecture Notes in Phys., no. 908, p. 83–133, Springer, Cham, 2016.

[18] R. Carles, R. Danchin & J.-C. Saut – "Madelung, Gross-Pitaevskii and Korteweg", *Nonlinearity* **25** (2012), no. 10, p. 2843–2873.

[19] T. Cazenave – *Semilinear schrödinger equations*, Courant Lecture Notes in Mathematics, no. 10, New York University Courant Institute of Mathematical Sciences, New York, 2003.

[20] W. A. Clarke & R. Marangell – "Rigorous justification of the Whitham modulation theory for equations of NLS type", *Stud. Appl. Math.* **147** (2021), no. 2, p. 577–621.

[21] E. B. Davies – *Linear operators and their spectra*, Cambridge Studies in Advanced Mathematics, no. 106, Cambridge University Press, Cambridge, 2007.

[22] V. Duchêne & L. M. Rodrigues – "Stability and instability in scalar balance laws: fronts and periodic waves", *Anal. PDE*, To appear.

[23] M. B. Erdoğan & N. Tzirakis – "Wellposedness and applications", London Mathematical Society Student Texts, no. 86, Cambridge University Press, Cambridge, 2016.

[24] T. Gallay & M. Hărăguş – "Stability of small periodic waves for the nonlinear Schrödinger equation", *J. Differential Equations* **234** (2007), no. 2, p. 544–581.

[25] R. A. Gardner – "On the structure of the spectra of periodic travelling waves", *J. Math. Pures Appl. (9)* **72** (1993), no. 5, p. 415–439.

[26] ———, "Spectral analysis of long wavelength periodic waves and applications", *J. Reine Angew. Math.* **491** (1997), p. 149–181.

[27] M. Grillakis, J. Shatah & W. Strauss – "Stability theory of solitary waves in the presence of symmetry. II", *J. Funct. Anal.* **94** (1990), no. 2, p. 308–348.

[28] S. Hakkaev, M. Stanislavova & A. Stefanov – "Transverse instability for periodic waves of KP-I and Schrödinger equations", *Indiana Univ. Math. J.* **61** (2012), no. 2, p. 461–492.

[29] D. Hoff & K. Zumbrun – "Multi-dimensional diffusion waves for the Navier–Stokes equations of compressible flow", *Indiana Univ. Math. J.* **44** (1995), no. 2, p. 603–676.

[30] M. Hărăguş & T. Kapitula – "On the spectra of periodic waves for infinite-dimensional Hamiltonian systems", *Phys. D* **237** (2008), no. 20, p. 2649–2671.

[31] J. Jin, S. Liao & Z. Lin – "Nonlinear modulational instability of dispersive PDE models", *Arch. Ration. Mech. Anal.* **231** (2019), no. 3, p. 1487–1530.

[32] M. A. Johnson, P. Noble, L. M. Rodrigues, Z. Yang & K. Zumbrun – "Spectral stability of inviscid roll waves", *Comm. Math. Phys.* **367** (2019), no. 1, p. 265–316.

[33] M. A. Johnson, P. Noble, L. M. Rodrigues & K. Zumbrun – "Nonlocalized modulation of periodic reaction diffusion waves: the Whitham equation", *Arch. Ration. Mech. Anal.* **207** (2013), no. 2, p. 669–692.

[34] ———, "Behavior of periodic solutions of viscous conservation laws under localized and nonlocalized perturbations", *Invent. Math.* **197** (2014), no. 1, p. 115–213.

[35] ———, "Spectral stability of periodic wave trains of the Korteweg–de Vries/Kuramoto–Sivashinsky equation in the Korteweg–de vries limit", *Trans. Amer. Math. Soc.* **367** (2015), no. 3, p. 2159–2212.

[36] B. Kabil & L. M. Rodrigues – "Spectral validation of the Whitham equations for periodic waves of lattice dynamical systems", *J. Differential Equations* **260** (2016), no. 3, p. 2994–3028.

[37] T. Kapitula & K. Promislow – *Spectral and dynamical stability of nonlinear waves*, Applied Mathematical Sciences, no. 185, Springer, New York, 2013.

[38] T. Kato – *Perturbation theory for linear operators*, 2 ed., Grundlehren der Mathematischen Wissenschaften, no. 132, Springer-Verlag, Berlin, 1976.

[39] P. Kuchment – *Floquet theory for partial differential equations*, Operator Theory: Advances and Applications, no. 60, Birkhäuser Verlag, Basel, 1993.

[40] K. P. Leisman, J. C. Bronski, M. A. Johnson & R. Marangell – "Stability of traveling wave solutions of nonlinear dispersive equations of NLS type", *Arch. Ration. Mech. Anal.* **240** (2021), no. 2, p. 927–969.

[41] Z. Lin – "Stability and instability of traveling solitonic bubbles", *Adv. Differential Equations* **7** (2002), no. 8, p. 897–918.

[42] F. Linares & G. Ponce – *Introduction to nonlinear dispersive equations*, 2 ed., Universitext, Springer, New York, 2015.

[43] E. Madelung – "Quantentheorie in hydrodynamischer Form", *Zeitschrift fur Physik* **40** (1927), no. 3, p. 322–326.

[44] P. Noble & L. M. Rodrigues – "Whitham's modulation equations and stability of periodic wave solutions of the Korteweg–de Vries–Kuramoto–Sivashinsky equation", *Indiana Univ. Math. J.* **62** (2013), no. 3, p. 753–783.

[45] M. Oh & K. Zumbrun – "Low-frequency stability analysis of periodic traveling-wave solutions of viscous conservation laws in several dimensions", *Z. Anal. Anwend.* **25** (2006), no. 1, p. 1–21.

[46] J. A. Pava – "Existence and stability of solitary and periodic travelling wave solutions", Mathematical Surveys and Monographs, no. 156, American Mathematical Society, Providence, RI, 2009.

[47] L. M. Rodrigues – "Vortex-like finite-energy asymptotic profiles for isentropic compressible flows", *Indiana Univ. Math. J.* **58** (2009), no. 4, p. 1747–1776.

[48] ———, "Asymptotic stability and modulation of periodic wavetrains, general theory & applications to thin film flows", Habilitation à diriger des recherches, Université Lyon 1, 2013.

[49] ———, "Space-modulated stability and averaged dynamics", *Journées Équations aux Dérivées Partielles* **2015** (2015), no. 8, p. 1–15, `http://www.numdam.org/item/JEDP_2015____A8_0`.

[50] ———, "Linear asymptotic stability and modulation behavior near periodic waves of the Korteweg–de Vries equation", *J. Funct. Anal.* **274** (2018), no. 9, p. 2553–2605.

[51] F. Rousset & N. Tzvetkov – "A simple criterion of transverse linear instability for solitary waves", *Math. Res. Lett.* **17** (2010), no. 1, p. 157–169.

[52] B. Sandstede & A. Scheel – "On the stability of periodic travelling waves with large spatial period", *J. Differential Equations* **172** (2001), no. 1, p. 134–188.

[53] D. Serre – "Spectral stability of periodic solutions of viscous conservation laws: large wavelength analysis", *Comm. Partial Differential Equations* **30** (2005), no. 1, p. 259–282.

[54] C. Sulem & P.-L. Sulem – "Self-focusing and wave collapse", Applied Mathematical Sciences, no. 139, Springer-Verlag, New York, 1999.

[55] N. G. Vakhitov & A. A. Kolokolov – "Stationary solutions of the wave equation in a medium with nonlinearity saturation", *Radiophysics and Quantum Electronics* **16** (1973), no. 7, p. 783–789.

[56] G. B. Whitham – *Linear and nonlinear waves*, Pure and Applied Mathematics, Wiley-Interscience [John Wiley & Sons], New York, 1974.

[57] Z. Yang & K. Zumbrun – "Convergence as period goes to infinity of spectra of periodic traveling waves toward essential spectra of a homoclinic limit", *J. Math. Pures Appl.* **9** (2019), no. 132, p. 27–40.

[58] P. E. Zhidkov – *Korteweg–de Vries and nonlinear Schrödinger equations: qualitative theory*, Lecture Notes in Mathematics, no. 1756, Springer-Verlag, Berlin, 2001.

Bull. Soc. Math. France
150 (1), 2022, p. 209-249

A STUDY OF NEFNESS IN HIGHER CODIMENSION

by Xiaojun Wu

Abstract. — In this work, following the fundamental work of Boucksom, we construct the nef cone of a compact complex manifold in higher codimension and give explicit examples for which these cones are different. In the third and fourth sections, we give different versions of Kawamata–Viehweg vanishing theorems regarding nefness in higher codimension and numerical dimensions. We also show through examples the optimality of the divisorial Zariski decomposition given in [5].

Résumé (*Une étude de l'effectivité numérique en codimension supérieure*). — Dans ce travail, à la suite des travaux fondamentaux de Boucksom, nous construisons le cône nef d'une variété complexe compacte de codimension supérieure et donnons des exemples explicites pour lesquels ces cônes sont différents. Dans les troisième et quatrième sections, nous donnons différentes versions des théorèmes d'annulation de Kawamata-Viehweg en termes de l'effectivité numérique en codimension supérieure et des dimensions numériques. Nous montrons aussi par des exemples l'optimalité de la décomposition divisoriale de Zariski donnée dans [5].

Texte reçu le 28 janvier 2021, modifié le 12 novembre 2021, accepté le 10 janvier 2022.

Xiaojun Wu, Universität Bayreuth, Universitätsstraße 30, 95447 Bayreuth • *E-mail :* xiaojun.wu@uni-bayreuth.de

Mathematical subject classification (2010). — 32F32, 32Q15, 32U40.

Key words and phrases. — Nefness in higher codimension, numerical dimension, Kawamata–Viehweg vanishing theorem.

This work is supported by the PhD program AMX of École Polytechnique and Ministère de l'Enseignement Supérieur et de la Recherche et de l'Innovation and the European Research Council grant ALKAGE number 670846 managed by J.-P. Demailly.

BULLETIN DE LA SOCIÉTÉ MATHÉMATIQUE DE FRANCE 0037-9484/2022/209/$ 5.00

doi:10.24033/bsmf.2847

1. Introduction

One of the reformulations of the Kodaira embedding theorem is that a compact complex manifold is projective if and only if the Kähler cone, i.e. the convex cone spanned by Kähler forms in $H^2(X,\mathbb{R})$, contains a rational point (i.e. an element in $H^2(X,\mathbb{Q})$).

As a general matter of fact, it is obviously interesting to study positive cones attached to compact complex manifolds and relate them with the geometry of the manifold. In classical algebraic or complex geometry, the emphasis is on two types of positive cones: the nef and psef cones, defined as the closed convex cones spanned by nef classes and psef classes, respectively. The nef cone is, of course, contained in the psef cone.

The work of Boucksom [5] defines and studies the so-called modified nef cone for an arbitrary compact complex manifold. Due to this definition, Boucksom was able to show the existence of a divisorial Zariski decomposition for any psef class (i.e. any cohomology class containing a positive current). The modified cone just sits between the nef and psef cones.

Inspired by Boucksom's definition, in Section 2, we introduce the concept of a nef cone in arbitrary codimension for any compact complex manifold, which is an interpolation between the above positive cones.

Definition A. — Let $\alpha \in H^{1,1}_{BC}(X,\mathbb{R})$ be a psef class. We say that α is nef in codimension k, if for any irreducible analytic subset $Z \subset X$ of codimension at most equal to k, we have the generic minimal multiplicity of α along Z as (defined in [5])

$$\nu(\alpha, Z) = 0.$$

With this terminology, the nef cone is the nef cone in codimension n, where n is the complex dimension of the manifold, while the psef cone is the nef cone in codimension 0, and the modified nef cone is the nef cone in codimension 1. We notice that the algebraic analogue in the projective case is introduced in [32]. In Section 4, we show that these cones are, in general, different and construct explicit examples where they are different.

Inspired by the work of [9] and using Guan–Zhou's solution of Demailly's strong openness conjecture, we get the following Kawamata–Viehweg vanishing theorem in Section 3. The proof follows Cao's proof closely:

Theorem A. — *Let (L,h) be a pseudo-effective line bundle on a compact Kähler n-dimensional manifold X with singular positive metric h. Then the morphism induced by the inclusion $K_X \otimes L \otimes \mathcal{I}(h) \to K_X \otimes L$*

$$H^q(X, K_X \otimes L \otimes \mathcal{I}(h)) \to H^q(X, K_X \otimes L)$$

vanishes for every $q \geq n - \mathrm{nd}(L) + 1$.

As an application, in Section 4, we obtain the following generalisation from the nef case to the psef case of a similar result stated in [16].

Theorem B. — *Let (X, ω) be a compact Kähler manifold of dimension n and L a line bundle on X that is nef in codimension 1. Assume that $\langle L^2 \rangle \neq 0$ where $\langle \cdot \rangle$ is the positive product defined in* [4]. *Assume that there exists an effective integral divisor D such that $c_1(L) = c_1(D)$. Then*

$$H^q(X, K_X + L) = 0,$$

for $q \geq n-1$.

The proof of the above theorem is an induction on the dimension, using Theorem A. A difference compared with the nef case treated in [16] is that instead of an intersection number we need to use a positive product (or movable intersection number), which is a non-linear operation. Nevertheless, under a condition of nefness in higher codimension, we get the following estimate.

Lemma A. — *Let α be a nef class in codimension p on a compact Kähler manifold (X, ω), then for any $k \leq p$ and Θ any positive closed $(n-k, n-k)$-form, we have*

$$(\alpha^k, \Theta) \geq \langle \alpha^k, \Theta \rangle.$$

With this inequality, the intersection number calculation in [16] is still valid, and thus the cohomology calculations can be recycled.

Observe that a current with minimal singularities need not have analytic singularities for every big class α that is nef in codimension 1 but not nef in codimension 2; such an example was given by [32], and also observed by Matsumura [30].

As a consequence of Matsumura's observation, the assumption of our Kawamata–Viehweg vanishing theorem that the line bundle is numerically equivalent to an effective integral divisor is actually required. In the nef case considered in [16], the authors deduce from their assumption that the line bundle L is nef with $(L^2) \neq 0$ that L is numerically equivalent to an effective integral divisor D, and that there exists a positive singular metric h on L, such that $\mathcal{I}(h) = \mathcal{O}(-D)$.

However, for a big line bundle L that is nef in codimension 1 but not nef in codimension 2 over an arbitrary compact Kähler manifold (X, ω), we have that $\langle L^2 \rangle \neq 0$ and $\frac{i}{2\pi}\Theta(L, h_{\min})$ need not be a current associated with an effective integral divisor.

Another by-product is the (probably already known) example of a projective manifold X with $-K_X$ psef, for which the Albanese morphism is not surjective. It was proven in [8], [34] (and [38] for the projective case) that the Albanese morphism of a compact Kähler manifold with $-K_X$ nef is always surjective. Thus, replacing nefness by pseudo-effectivity in the study of Albanese morphism seems to be a non-trivial problem.

2. Nefness in higher codimension

We first recall some technical preliminaries introduced in [5]. Throughout this paper, X is assumed to be a compact complex manifold equipped with some reference Hermitian metric ω (i.e. a smooth positive definite $(1,1)$-form); we usually take ω to be Kähler, if X possesses such metrics. The Bott–Chern cohomology group $H^{1,1}_{BC}(X,\mathbb{R})$ is the space of d-closed smooth (1,1)-forms modulo $i\partial\overline{\partial}$-exact ones. By the quasi-isomorphism induced by the inclusion of smooth forms into currents, $H^{1,1}_{BC}(X,\mathbb{R})$ can also be seen as the space of d-closed $(1,1)$-currents modulo $i\partial\overline{\partial}$-exact ones. A cohomology class $\alpha \in H^{1,1}_{BC}(X,\mathbb{R})$ is said to be pseudo-effective iff it contains a positive current; α is nef iff, for each $\varepsilon > 0$, α contains a smooth form α_ε, such that $\alpha_\varepsilon \geq -\varepsilon\omega$; α is big iff it contains a Kähler current, i.e. a closed $(1,1)$-current T, such that $T \geq \varepsilon\omega$ for $\varepsilon > 0$ small enough.

Definition 2.1 ([19]). — Let φ_1, φ_2 be two quasi-psh functions on X (i.e. $i\partial\overline{\partial}\varphi_i \geq -C\omega$ in the sense of currents for some $C \geq 0$). The function φ_1 is said to be less singular than φ_2 (one then writes $\varphi_1 \preceq \varphi_2$) if $\varphi_2 \leq \varphi_1 + C_1$ for some constant C_1. Let α be a fixed psef class in $H^{1,1}_{BC}(X,\mathbb{R})$. Given $T_1, T_2, \theta \in \alpha$ with θ smooth, and $T_i = \theta + i\partial\overline{\partial}\varphi_i$ with φ_i quasi-psh ($i = 1,2$), we write $T_1 \preceq T_2$ iff $\varphi_1 \preceq \varphi_2$ (notice that for any choice of θ, the potentials φ_i are defined up to smooth bounded functions, since X is compact). If γ is a smooth real $(1,1)$-form on X, the collection of all potentials φ, such that $\theta + i\partial\overline{\partial}\varphi \geq \gamma$ admits a minimal element $T_{min,\gamma}$ for the pre-order relation $\preceq$, constructed as the semi-continuous upper envelope of the sub-family of potentials $\varphi \leq 0$ in the collection.

Definition 2.2 (Minimal multiplicities). — The minimal multiplicity at $x \in X$ of the pseudo-effective class $\alpha \in H^{1,1}_{BC}(X,\mathbb{R})$ is defined as

$$\nu(\alpha, x) := \sup_{\varepsilon > 0} \nu(T_{\min,\varepsilon}, x),$$

where $T_{\min,\varepsilon}$ is the minimal element $T_{\min,-\varepsilon\omega}$ in the above definition, and $\nu(T_{\min,\varepsilon}, x)$ is the Lelong number of $T_{\min,\varepsilon}$ at x. When Z is an irreducible analytic subset, we define the generic minimal multiplicity of α along Z as

$$\nu(\alpha, Z) := \inf\{\nu(\alpha, x), x \in Z\}.$$

When Z is positive dimensional, there exists for each $\ell \in \mathbb{N}^*$ a countable union of proper analytic subsets of Z denoted by $Z_\ell = \bigcup_p Z_{\ell,p}$, such that $\nu(T_{\min,\frac{1}{\ell}}, Z) := \inf_{x\in Z} \nu(T_{\min,\frac{1}{\ell}}, x) = \nu(T_{\min,\frac{1}{\ell}}, x)$, for $x \in Z \smallsetminus Z_\ell$. By construction, when $\varepsilon_1 < \varepsilon_2$, $T_{\min,\varepsilon_1} \succeq T_{\min,\varepsilon_2}$. Hence, for a very general point $x \in Z \smallsetminus \bigcup_{\ell\in\mathbb{N}^*} Z_\ell$,

$$\nu(\alpha, Z) \leq \nu(\alpha, x) = \sup_{\ell} \nu(T_{\min,\frac{1}{\ell}}, Z).$$

On the other hand, for any $y \in Z$,

$$\sup_{\ell} \nu(T_{\min,\frac{1}{\ell}}, Z) \leq \sup_{\ell} \nu(T_{\min,\frac{1}{\ell}}, y) = \nu(\alpha, y).$$

In conclusion, $\nu(\alpha, Z) = \nu(\alpha, x)$, for a very general point $x \in Z \setminus \bigcup_{\ell \in \mathbb{N}^*} Z_\ell$, and $\nu(\alpha, Z) = \sup_\varepsilon \nu(T_{\min,\varepsilon}, Z)$.

Now we can define the concept of nefness in higher codimension implicitly used in [5]. It is the generalisation of the concept of "modified nefness" to the higher codimensional case.

Definition 2.3. — Let $\alpha \in H^{1,1}_{BC}(X, \mathbb{R})$ be a psef class. We say that α is nef in codimension k, if for every irreducible analytic subset $Z \subset X$ of codimension at most equal to k, we have

$$\nu(\alpha, Z) = 0.$$

We denote by $\mathcal{N}_k$ the cone generated by nef classes in codimension k. By Proposition 3.2 in [5], a psef class α is nef, iff for any $x \in X$, $\nu(\alpha, x) = 0$. By our definition, psef is equivalent to nef in codimension 0, and nef is equivalent to nef in codimension $n := \dim_{\mathbb{C}} X$. In this way, we get a bunch of positive cones on X, satisfying the inclusion relations

$$\mathcal{N} = \mathcal{N}_n \subset \cdots \subset \mathcal{N}_1 \subset \mathcal{N}_0 = \mathcal{E},$$

where $\mathcal{N}$ and $\mathcal{E}$ are cones of nef and psef classes, respectively. By a proof similar to those of Propositions 3.5 and 3.6 in [5], we get:

Proposition 2.4. — (1) *For every $x \in X$ and every irreducible analytic subset Z, the map $\mathcal{E} \to \mathbb{R}^+$ defined on the cone $\mathcal{E}$ of psef classes by $\alpha \mapsto \nu(\alpha, Z)$ is convex and homogeneous. It is continuous on the interior $\mathcal{E}^\circ$ and lower semi-continuous on the whole of $\mathcal{E}$.*

(2) *If $T_{\min} \in \alpha$ is a positive current with minimal singularities, we have $\nu(\alpha, Z) \leq \nu(T_{\min}, Z)$.*

(3) *If α is moreover big, we have $\nu(\alpha, Z) = \nu(T_{\min}, Z)$.*

The following lemma is a direct application of the proposition.

Lemma 2.5. — *Let $Y \subset X$ be a smooth sub-manifold of X and $\pi : \tilde{X} \to X$ be the blow-up of X along Y. We denote by E the exceptional divisor. If $\alpha \in H^{1,1}_{BC}(X, \mathbb{R})$ is a big class, we have*

$$\nu(\alpha, Y) = \nu(\pi^* \alpha, E).$$

For Z any irreducible analytic set not included in Y, we denote by $\tilde{Z}$ the strict transform of Z. Then

$$\nu(\alpha, Z) = \nu(\pi^* \alpha, \tilde{Z}).$$

For W any irreducible analytic set in Y, we have

$$\nu(\alpha, W) = \nu(\pi^*\alpha, \mathbb{P}(N_{Y/X}|_W)).$$

Proof. — Since α is big, we know that by taking a suitable regularisation, there exists a Kähler current $T \in \alpha$ with analytic singularities. The pull-back π^*T of this current is a smooth Kähler current on some dense open set U, where π is a biholomorphism. Hence, the volume of $\pi^*\alpha$ defined as $\int_{T\in\pi^*\alpha, T\geq 0,} T_{ac}^n$ (ac means the absolute part of the current) is larger than the mass of π^*T on U, which is strictly positive. By [4] $\pi^*\alpha$ is thus big.

By the proposition, we have

$$\nu(\alpha, Y) = \inf_{T\in\alpha} \nu(T, Y), \quad \nu(\pi^*\alpha, E) = \inf_{S\in\pi^*\alpha} \nu(S, E).$$

On the other hand, the push-forward and pull-back operators acting on positive $(1,1)$ currents induce bijections between positive currents in the class α and positive currents in the class $\pi^*\alpha$. Let $\theta \in \alpha$ be a smooth form such that $T = \theta + i\partial\overline{\partial}\varphi$. We recall that for any irreducible analytic set W with local generators $(g_1, \dots, g_r)$ near a regular point $w \in W$, the generic Lelong number along W is the largest γ, such that $\varphi \leq \gamma \log(\sum |g_i|^2) + O(1)$ near w. Since $\pi^*(g_1, \dots, g_r) \cdot \mathcal{O}_{\tilde{X}} = \mathcal{I}_E$, we have $\nu(T, Y) = \nu(\pi^*T, E)$. In particular, this implies that

$$\nu(\alpha, Y) = \nu(\pi^*\alpha, E).$$

For W any irreducible analytic set in the centre Y, since the exceptional divisor is isomorphic to $\mathbb{P}(N_{Y/X})$, the pre-image of W under the blow-up is isomorphic to $\mathbb{P}(N_{Y/X}|_W)$. In suitable local coordinates $(z_1, \dots, z_n)$ on X and $(w_1, \dots, w_n)$ on $\tilde{X}$, the blow-up map is given by

$$\pi(w_1, \dots, w_n) = (w_1, w_1w_2, \dots, w_1w_s; w_{s+1}, \dots, w_n).$$

In these coordinates, the centre Y is given by the zero variety $V(z_{s+1}, \dots, z_n)$. Assume that in this chart, $W = V(z_{s+1}, \dots, z_n; f_1, \dots, f_r)$ where f_i is a function of $z_1, \dots, z_s$ (as we can assume without loss of generality). Then

$$\begin{aligned}\pi^*(\mathcal{I}_W) \cdot \mathcal{O}_{\tilde{X}} &= (w_1, f_1(w_1, w_1w_2, \dots, w_1w_s), \dots, f_r(w_1, w_1w_2, \dots, w_1w_s)) \\ &= \mathcal{I}_{\mathbb{P}(N_{Y/X}|_W)}.\end{aligned}$$

In particular, this implies that

$$\nu(\alpha, W) = \nu(\pi^*\alpha, \mathbb{P}(N_{Y/X}|_W)).$$

For the second statement, we just observe that the generic Lelong number along Z (or $\tilde{Z}$) is equal to the Lelong number at some very general point. Since Z is not contained in Y, we can assume without loss of generality that the very general point is not in Y (or E). Since the Lelong number is a coordinate invariant local property, for such a very general point $x \in \tilde{Z}$ near which π

is a local biholomorphism and any $T \in \alpha, T \geq 0$, $\nu(T, Z) = \nu(T, \pi(x)) = \nu(\pi^* T, x) = \nu(\pi^* T, \tilde{Z})$. Hence, we have

$$\nu(\alpha, Z) = \nu(\pi^* \alpha, \tilde{Z}). \qquad \square$$

As a corollary, we find:

Corollary 2.6. — *Let $\mu : \tilde{X} \to X$ be a composition of finitely many blow-ups with smooth centres in X. If $\alpha \in H^{1,1}_{BC}(X, \mathbb{R})$ is a big class on X, such that $\mu^* \alpha$ is nef in codimension k, then α is a nef class in codimension k.*

Proof. — Without loss of generality, we can reduce ourselves to the case where μ is a blow-up of smooth centre Y in X. By Lemma 2.5, the generic minimal multiplicity of α along any irreducible analytic set of X of codimension at most equal to k is equal to the generic minimal multiplicity of $\mu^* \alpha$ along with a certain irreducible analytic set of $\tilde{X}$ of codimension at most equal to k. So, by the definition of nefness in codimension k, the fact $\mu^* \alpha$ is nef in codimension k implies that α is nef in codimension k. $\square$

Remark 2.7. — Let X be a compact complex manifold whose big cone is non-empty. Recall that by Proposition 2.3 of [5], a class α is modified Kähler (i.e. α is in the interior of nef cone in codimension 1) iff there exists a modification $\mu : \tilde{X} \to X$ and a Kähler class $\tilde{\alpha}$ on $\tilde{X}$, such that $\alpha = \mu_* \tilde{\alpha}$. As a consequence, for $\mu : \tilde{X} \to X$, a modification between compact Kähler manifolds and $\tilde{\alpha} \in H^{1,1}_{BC}(\tilde{X}, \mathbb{R})$ a big and nef class on $\tilde{X}$ in codimension k, it is false in general that $\mu_* \tilde{\alpha}$ is a nef class in codimension k.

To give an equivalent definition of nefness in higher codimension, we will need the following definition.

Definition 2.8 (Non-nef locus). — The non-nef locus of a pseudo-effective class $\alpha \in H^{1,1}_{BC}(X, \mathbb{R})$ is defined by

$$E_{nn}(\alpha) := \{x \in X, \nu(\alpha, x) > 0\}.$$

Proposition 2.9. — *A psef class α is nef in codimension k iff for any $\varepsilon > 0$, any $c > 0$, the codimension of any irreducible component of $E_c(T_{\min,\varepsilon})$ is larger than $k+1$.*

Proof. — By the definition of non-nef locus, we have

$$E_{nn}(\alpha) = \bigcup_{\varepsilon > 0} \bigcup_{c > 0} E_c(T_{\min,\varepsilon}) = \bigcup_{m \in \mathbb{N}^*} \bigcup_{n \in \mathbb{N}^*} E_{\frac{1}{n}}(T_{\min,\frac{1}{m}}).$$

We know by Siu's theorem [35] that $E_{\frac{1}{n}}(T_{\min,\frac{1}{m}})$ is an analytic set. Hence the non-nef locus is a countable union of irreducible analytic sets. If for any $\varepsilon > 0$, any $c > 0$, the codimension of any irreducible component of $E_c(T_{\min,\varepsilon})$ is larger than $k+1$, then for any irreducible analytic set Z of codimension k, $E_{nn}(\alpha) \cap Z$ is strictly contained in Z. Hence, $\nu(\alpha, Z) = 0$.

In the other direction, assume there exists an irreducible component Z of $E_{\frac{1}{n}}(T_{\min,\frac{1}{m}})$ with codimension at most equal to k. On each point x of this irreducible component, $\nu(\alpha,x)\geq\nu(T_{\min,\frac{1}{m}},x)\geq\frac{1}{n}$. In particular, $\nu(\alpha,Z)\geq\frac{1}{n}$, which contradicts the fact that α is nef in codimension k. □

Remark 2.10. — If the manifold X is projective, it is enough to test the minimal multiplicity along irreducible analytic subsets of codimension k to prove that the class is nef in codimension k. The argument is as follows:

For any irreducible analytic set Z of codimension strictly smaller than k, for any $z\in Z$, since X is projective, there exists some hypersurfaces H_i such that $z\in H_i$, and the irreducible component of $Z\cap\bigcap_i H_i$ containing z has codimension k. In other words, Z is covered by the irreducible analytic subsets of codimension exactly k. By assumption, the generic minimal multiplicity along any of these irreducible analytic subsets is 0. This implies that the generic minimal multiplicity along Z at most equal to the generic minimal multiplicity along any of these irreducible analytic sets is 0.

Remark 2.11. — In the general setting of compact complex manifolds, it is crucial to test the generic minimal multiplicity along any analytic set of codimension at most equal to k, instead of any analytic set of codimension k, to obtain the inclusion of the various positive cones. The problem is that there may exist too few analytic subsets in an arbitrary compact complex manifold.

A typical example can be taken as follows. For example, let X_1 be a compact manifold such that the nef cone is strictly contained in the psef cone (for example, we can take the projectivisation of an unstable rank 2 vector bundle over a curve of genus larger than 2, whose cones are explicitly calculated on page 70 of [29]). Let X_2 be a very general torus, such that the only analytic sets in X_2 are either a union of points or X_2. Let β be a psef but not nef class on X_1. Let $X:=X_1\times X_2$ with natural projections π_1,π_2 and $\alpha:=\pi_1^*\beta$. Assume that $\dim(X_1)<\dim(X_2)$. Fix ω_1,ω_2, two reference Hermitian metrics on X_1,X_2.

Now α is a psef but not nef class on X. The only analytic subsets of codimension $\dim(X_1)$ is the fibre of π_2; α has generic minimal multiplicity 0 along any fibre of π_2. The reason is as follows: The minimal current in α larger than $-\varepsilon(\pi_1^*\omega_1+\pi^*\omega_2)$ denoting $\min\{T\in\alpha,T\geq-\varepsilon(\pi_1^*\omega_1+\pi^*\omega_2)\}$ is less singular than the pull-back of the minimal current in β larger than $-\varepsilon\omega_1$ denoting $\min\{S\in\beta,S\geq-\varepsilon\omega_1\}$, and the restriction of these minimal currents on the fibre of π_2 is trivial. In other words, the generic Lelong number of $\min\{T\in\alpha,T\geq-\varepsilon(\pi_1^*\omega_1+\pi^*\omega_2)\}$ along the fibres is smaller than the generic Lelong number of the pull-back of $\min\{S\in\beta,S\geq-\varepsilon\omega_1\}$, which is 0. Hence, it is itself 0.

On the other hand, for any positive integers m,n, take Z a positive dimensional irreducible component of $E_{\frac{1}{n}}(T_{\min,\frac{1}{m}})$ in the non-nef locus of β. The existence of such an irreducible component will be shown in Lemma 2.13, which implies that α has to be nef in codimension at most equal to $n-2$. Now $Z \times X_2$ is an irreducible analytic set of codimension strictly smaller than $\dim(X_1)$. However, the generic minimal multiplicity along $Z \times X_2$ is larger than $\frac{1}{n}$. In particular, this shows that α is not nef in codimension $\dim(X_1) - \dim(Z)$.

Remark 2.12. — Let us mention that our definition of nefness in codimension 1 is equivalent to the definition of modified nefness. By definition, a psef class is modified nef iff its generic minimal multiplicity is 0 along any prime divisor. To prove the equivalence, we need to show that for any psef class α on X, we automatically have

$$\nu(\alpha, X) = 0.$$

This is because that $\nu(\alpha, X) \leq \nu(T_{\min}, X)$, where the latter is 0. We notice that by Siu's decomposition theorem [35], the set $E_{c>0}(T_{\min}) = \bigcup_{n\in\mathbb{N}^*} E_{\frac{1}{n}}(T_{\min})$ is a countable union of proper analytic sets.

By this observation, we can also say that the "nef in codimension 0" cone is exactly the psef cone.

In analogy to the case of surfaces for which the nef cone coincides with the modified nef cone, the nef cone in codimension $n-1$ coincides with the nef cone.

Lemma 2.13. — *Let α be a psef class, then α is nef in codimension $n-1$ iff α is nef.*

Proof. — If α is nef, by inclusion of different positive cones, it is nef in codimension $n-1$. In the other direction, we will need the following proposition 3.4 in [5], which is a reformulation of a result of Păun [33].

A pseudo-effective class α is nef iff $\alpha|_Y$ is pseudo-effective, for every irreducible analytic subset $Y \subset E_{nn}(\alpha)$.

Given a class α that is nef in codimension $n-1$, Proposition 2 implies that for any $\varepsilon > 0$ and any $c > 0$, the analytic set $E_c(T_{\min,\varepsilon})$ is a finite set. Therefore, the non-nef locus, which is a countable union of finite sets, has at most countably many points. In particular, this implies that the restriction of α on any $Y \subset E_{nn}(\alpha)$ is 0, hence psef. By the above proposition, α is nef. □

Remark 2.14. — Recall that a line bundle L over a projective manifold is nef iff its intersection number with any curve satisfies $(L \cdot C) \geq 0$. By the important work of [6], a class is psef iff its pairing with any movable curve is positive. Here, a curve C is said to be movable if $C = C_{t_0}$ is a member of an analytic family $(C_t)_{t\in S}$, such that $\bigcup_{t\in S} C_t = X$, and, as such, C is a reduced irreducible 1-cycle. Notice also that nef is equivalent to nef in codimension $n-1$, and psef is equivalent to nef in codimension 0.

Then it is natural to conjecture that a class over a projective manifold is nef in codimension k if and only if its pairing with any movable curve in codimension k is positive. Here, a curve C is said to be movable in codimension k, if $C = C_{t_0}$ is a member of an analytic family $(C_t)_{t\in S}$, such that $\bigcup_{t\in S} C_t$ is an analytic subset of X of codimension k, and, as such, C is a reduced irreducible 1-cycle.

REMARK 2.15. — Inspired by the result of Păun, it seems to be natural to conjecture that a psef class $\{T\}$ with T a positive current on X is nef in codimension k, if and only if that for any irreducible component of codimension at most k in $\bigcup_{c>0} E_c(T)$ $\{T\}|_Z$ is nef in codimension $k - \mathrm{codim}(Z, X)$. When $k = n$, this is exactly the result of Păun. When $k = 0$, it is trivial. The "only if" part is quite similar. The restriction of the potentials of $T_{\min,\varepsilon}$ on any irreducible analytic set of codimension at most k decreases to a potential on the sub-manifold. If we fix the maximum of the potentials on X to be 0, they form a compact family. The limit potential would be quasi-psh, and, thus, the restriction of the class on the analytic set is psef. The "if" part is, of course, true if the manifold is a Kähler surface by Păun's result.

The "if" part is also true for the case $k = 1$ if the manifold is hyperkähler. By Lemma 4.9 [5] (see also [27]) a psef class α on a hyperkähler manifold is modified nef, if and only if for any prime divisor D, one has $q(\alpha, D) \geq 0$. Here, we let σ be a non-trivial symplectic holomorphic form on X and define

$$q(\alpha, \beta) := \int_X \alpha \wedge \beta \wedge (\sigma \wedge \overline{\sigma})^{\frac{n}{2}-1}$$

to be the Beauville-–Bogomolov quadratic form for any $(1, 1)$-classes α, β. For a psef $(1, 1)$-class α, such that $\alpha|_D$ is psef for any prime divisor D, we have

$$q(\alpha, \{[D]\}) = \int_X \alpha \wedge \{[D]\} \wedge (\sigma \wedge \overline{\sigma})^{\frac{n}{2}-1} = \int_D \alpha \wedge (\sigma \wedge \overline{\sigma})^{\frac{n}{2}-1} \geq 0.$$

Thus, α is nef in codimension 1.

A natural idea to attack this question in general consists of extending the current on this sub-variety Z to X. If this is possible, the current with minimal singularity would have a potential larger than that of the extended current. In particular, the current with minimal singularity would have generic Lelong number 0 along Z.

In this direction, Collins and Tosatti proved the following results in [10] and [11], which we now recall.

THEOREM 2.16 (Theorem 3.2 in [11]). — *Let X be a compact Fujiki manifold and α a closed smooth real (1,1)-form on X with $\{\alpha\}$ nef and $\int_X \alpha^n > 0$. Let $E = V \cup \bigcup_{i=1}^{I} Y_i$ be an analytic sub-variety of X, with V, Y_i its irreducible components, and V a positive dimensional compact, complex sub-manifold of X. Let $R = \alpha + i\partial\overline{\partial}F$ be a Kähler current in the class $\{\alpha\}$ on X with analytic* *singularities precisely along E and let $T = \alpha|_V + i\partial\overline{\partial}\varphi$ be a Kähler current in the class $\{\alpha|_V\}$ on V with analytic singularities. Then there exists a Kähler current $\tilde{T} = \alpha + i\partial\overline{\partial}\Phi$ in the class $\{\alpha\}$ on X with $\tilde{T}|_V$ smooth in a neighbourhood of the very general point of V.*

Theorem 2.17 (Theorem 1.1 in [10]). — *Let (X, ω) be a compact Kähler manifold and let $V \subset X$ be a positive-dimensional compact, complex sub-manifold. Let T be a Kähler current with analytic singularities along V in the Kähler class $\{\omega|_V\}$. Then there exists a Kähler current $\tilde{T}$ on X in the class $\{\omega\}$ with $T = \tilde{T}|_V$.*

Using their results, in a given Kähler class, one can extend Kähler currents with analytic singularities defined in a smooth sub-variety. If the class is just nef and big on the Kähler manifold, one can only show the existence of a Kähler current whose potential is not identically infinity along the sub-manifold. Following Example 5.4 in [7], one can show that in a nef and big class (i.e., a nef class containing a Kähler current) on a Kähler manifold X, one cannot always extend a positive current along a sub-manifold into a positive current on X. In their example, the positive current on the sub-manifold can even be chosen to be smooth. More precisely, there exists C, a sub-manifold of a certain compact Kähler manifold X, $\{\alpha\}$, a nef and big class on X with a smooth representative α and $\varphi \in L^1_{\text{loc}}(C)$ with $\alpha|_C + i\partial\overline{\partial}\varphi \geq 0$, such that there does not exist $\psi \in L^1_{\text{loc}}(X)$ satisfying $\alpha + i\partial\overline{\partial}\psi \geq 0$ and $\psi|_C = \varphi$.

Let us start the construction of the example. Let C be an elliptic curve and let A be an ample divisor on C. Let V be the rank 2 vector bundle over C the unique non-trivial extension of $\mathcal{O}_C$. Define $X := \mathbb{P}(V \oplus A)$ and $\{\alpha\} := c_1(\mathcal{O}_X(1))$ with smooth representative α. Then $\mathcal{O}_X(1)$ is a big and nef line bundle over X. The quotient map $V \oplus A \to \mathcal{O}_C$ induces a closed immersion $C \to X$. In particular, we have $\mathcal{O}_X(1)|_C = \mathcal{O}_C$. Since $c_1(\mathcal{O}_X(1)|_C) = 0$, there exists a smooth function φ on C such that $\alpha|_C + i\partial\overline{\partial}\varphi = 0$. We prove by contradiction that there does not exist $\psi \in L^1_{\text{loc}}(X)$, such that $\alpha + i\partial\overline{\partial}\psi \geq 0$ and $\psi|_C = \varphi$. The quotient map $V \oplus A \to V$ induces a closed immersion $\mathbb{P}(V) \to X$. On the contrary, we would have $\alpha|_{\mathbb{P}(V)} + i\partial\overline{\partial}\psi|_{\mathbb{P}(V)} \geq 0$ in the class $c_1(\mathcal{O}_{\mathbb{P}(V)}(1))$. By the calculation made in Example 1.7 of [18], we know that

$$\alpha|_{\mathbb{P}(V)} + i\partial\overline{\partial}\psi|_{\mathbb{P}(V)} = [C],$$

where $[C]$ is the current associated with C. In particular, this shows that $\psi|_C \equiv -\infty$, which is a contradiction.

In other words, Theorem 1.1 of [10] cannot be strengthened to obtain an extension of an arbitrary closed positive current in a class that is merely nef and big. Similarly, one cannot drop the Kähler current condition in the theorem of [11].

Let us return to our previous question. In order to get an analogue of Păun's result, the above discussion shows that we need to generalise Theorem 3.2 of [11] to the class of a big class that is nef in codimension k by adding a small Kähler form to the class and by using the semi-continuity of the generic minimal multiplicity. Unfortunately, we do not know how to do this at this point.

3. Junyan Cao's and Guan–Zhou's vanishing theorem

Following the ideas of [16], we get a Kähler version of the Kawamata–Viehweg vanishing theorem in the next section. To prepare for the proof, we give a version of Junyan Cao's and Guan–Zhou's vanishing theorem in terms of numerical dimension of line bundle instead of the numerical dimension of singular metric. In [9], Junyan Cao proved the following Kawamata–Viehweg–Nadel-type vanishing theorem.

Theorem 3.1. — *Let (L, h) be a pseudo-effective line bundle on a compact Kähler n-dimensional manifold X. Then,*

$$H^q(X, K_X \otimes L \otimes \mathcal{I}(h)) = 0,$$

for every $q \geq n - \mathrm{nd}(L, h) + 1$.

The numerical dimension $\mathrm{nd}(L, h)$ used in Cao's theorem is the numerical dimension of the closed positive $(1,1)$-current $i\Theta_{L,h}$ defined in his paper. Since we will not need this definition, we refer to his paper for further information. We just recall the remark on page 22 of [9]. In Example 1.7 of [18], the nef line bundle $\mathcal{O}(1)$ over the projectivisation of a rank 2 vector bundle over the elliptic curve C was considered, which is the only non-trivial extension of $\mathcal{O}_C$. It was proven that there exists a unique positive singular metric h on $\mathcal{O}(1)$. For this metric, $\mathrm{nd}(\mathcal{O}(1), h) = 0$. However, the numerical dimension of $\mathcal{O}(1)$ is equal to 1. We recall that for a nef line bundle L the numerical dimension is defined as

$$\mathrm{nd}(L) := \max\{p; c_1(L)^p \neq 0\}.$$

We also remark that Cao's technique of proof actually yields the result for the upper semi-continuous regularisation of multiplier ideal sheaf defined as

$$\mathcal{I}_+(h) := \lim_{\varepsilon \to 0} \mathcal{I}(h^{1+\varepsilon}),$$

instead of $\mathcal{I}(h)$, but we can apply Guan–Zhou's theorem [21, 22, 23, 24] to see that the equality $\mathcal{I}_+(h) = \mathcal{I}(h)$ always holds. In particular, by the Noetherian property of ideal sheaves, we have

$$\mathcal{I}_+(h) = \mathcal{I}(h^{\lambda_0}) = \mathcal{I}(h),$$

for some $\lambda_0 > 1$. This fact will also be used in our result.

Here, we prove the following version of Junyan Cao's and Guan–Zhou's vanishing theorem, following closely the ideas of Junyan Cao [9] and the version that was a bit simplified in [12].

THEOREM 3.2. — *Let L be a pseudo-effective line bundle on a compact Kähler n-dimensional manifold X. Then the morphism induced by inclusion $K_X \otimes L \otimes \mathcal{I}(h_{\min}) \to K_X \otimes L$*

$$H^q(X, K_X \otimes L \otimes \mathcal{I}(h_{\min})) \to H^q(X, K_X \otimes L)$$

is a 0 map for every $q \geq n - \mathrm{nd}(L) + 1$.

REMARK 3.3. — In Example 1.7 of [18], since the rank 2 vector bundle is the only non-trivial extension of $\mathcal{O}_C$, there exists a surjective morphism from this vector bundle to $\mathcal{O}_C$, which induces a closed immersion C into the ruled surface. The only positive metric on $\mathcal{O}(1)$ has curvature $[C]$, the current associated to C. On the other hand, $\mathcal{O}(1) = \mathcal{O}(C)$. So we have $H^2(X, K_X \otimes \mathcal{O}(1)) = H^0(X, \mathcal{O}(-1)) = H^0(X, \mathcal{O}(-C)) = 0$, and $H^2(X, K_X \otimes \mathcal{O}(1) \otimes \mathcal{I}(h_{\min})) = H^2(X, K_X \otimes \mathcal{O}(1) \otimes \mathcal{O}(-C)) = H^0(X, \mathcal{O}_X) = \mathbb{C}$. This shows that to get a numerical dimension version of theorem, the best that we can hope for is that the morphism is a 0 map instead of that $H^q(X, K_X \otimes L \otimes \mathcal{I}(h_{\min})) = 0$. We notice that, in general, one would expect the vanishing result

$$H^q(X, K_X \otimes L) = 0,$$

for $q \geq n - \mathrm{nd}(L) + 1$, whenever L is a nef line bundle. Here, the difficulty is to prove a general Kähler version since the results follow easily from an inductive hyperplane section argument when X is projective (cf., e.g. Corollary (6.26) of [14]).

The Kähler version of the definition of numerical dimension is stated in [12]. For L a psef line bundle on a compact Kähler manifold (X, ω), we define

$$\mathrm{nd}(L) := \max \Big\{ p \in [0, n]; \exists c > 0, \forall \varepsilon > 0, \exists h_\varepsilon, i\Theta_{L,h_\varepsilon} \geq -\varepsilon\omega, \\ \text{such that } \int_{X \setminus Z_\varepsilon} (i\Theta_{L,h_\varepsilon} + \varepsilon\omega)^p \wedge \omega^{n-p} \geq c \Big\}.$$

Here, the metrics h_ε are supposed to have analytic singularities, and Z_ε is the singular set of the metric.

By the following remark, we can even assume that h_ε as stated in the definition of the numerical dimension is increasing to $h_{\min}$ as $\varepsilon \to 0$. What we need here is that the weight functions φ_ε has limit $\varphi_{\min}$ and is pointwise at least equal to $\varphi_{\min}$ with a universal upper bound on X.

REMARK 3.4. — $T_{\min,\varepsilon'\omega} \preceq T_{\min,\varepsilon\omega} + (\varepsilon' - \varepsilon)\omega$ for any $\varepsilon \leq \varepsilon'$. Denote $T_{\min,\varepsilon\omega} = \theta + \varepsilon\omega + i\partial\overline{\partial}\varphi_{\min,\varepsilon\omega}$. We can arrange that

$$\varphi_{\min,0} \leq \varphi_{\min,\varepsilon\omega} \leq \varphi_{\min,\varepsilon'\omega}.$$

The Bergman kernel regularisation preserves the ordering of potentials (cf. [12]), so we have

$$\varphi_{0,\delta} \leq \varphi_{\varepsilon,\delta} \leq \varphi_{\varepsilon',\delta}.$$

for any $\delta > 0$. If $\delta(\varepsilon)$ is increasing with respect to ε, we can choose the metric h_ε to be decreasing with respect to ε. The limit of $\varphi_{\varepsilon,\delta(\varepsilon)}$ as $\varepsilon \to 0$ is equal to $\varphi_{\min,0}$ corresponding to the metric with minimal singularities on L.

Before giving the proof of the vanishing theorem, we give the general lines of the ideas and compare them with Cao's and Guan–Zhou's theorem. The idea is using the L^2 resolution of the multiplier ideal sheaf and proving that every $\overline{\partial}$-closed $L^2(h_{\min})$ global section can be approximated by $\overline{\partial}$-exact $L^2(h_\infty)$ global sections with h_∞ some smooth reference metric on L. To prove it, we solve the $\overline{\partial}$-equation using a Bochner technique with the error term (as in [16]), and we prove that the error term tends to 0.

For this proposal, we need to estimate the curvature asymptotically by some special approximating Hermitian metrics constructed using the Calabi–Yau theorem. Cao tried to prove that the error term tends to be 0 in the topology induced by the L^2-norm with respect to the given singular metric. In this way, he tried to keep the multiplier ideal sheaf unchanged when approximating the singular metric utilizing suitable "equisingular approximation". For our proposal, we try to prove that the error term tends to be 0 in the topology induced by the L^2-norm with respect to some (hence, any) smooth metric. It would be enough for us that the multiplier ideal sheaf of $h_{\min}$ is included in the multiplier ideal sheaf of the approximating hermitian metric. In some sense, Cao's theorem is more precise in studying the singularity of the metric, which somehow explains why his approach works for any singular metric, while our approach applies only to the image of the natural inclusion.

We start the proof of the vanishing theorem by the following technical curvature and singularity estimate.

Proposition 3.5. — *Let $(L, h_{\min})$ be a pseudo-effective line bundle on a compact Kähler manifold (X, ω). Let us write $T_{\min} = \frac{i}{2\pi}\Theta_{L,h_{\min}} = \alpha + \frac{i}{2\pi}\partial\overline{\partial}\varphi_{\min}$, where α is the curvature of some smooth metric h_∞ on L, and $\varphi_{\min}$ is a quasi-psh potential. Let $p = \mathrm{nd}(L)$ be the numerical dimension of L. Then, for every $\gamma \in]0,1]$ and $\delta \in]0,1]$, there exists a quasi-psh potential $\Phi_{\gamma,\delta}$ on X satisfying the following properties:*

(a) *$\Phi_{\gamma,\delta}$ is smooth in the complement $X \setminus Z_\delta$ of an analytic set $Z_\delta \subset X$.*
(b) *$\alpha + \delta\omega + \frac{i}{2\pi}\partial\overline{\partial}\Phi_{\gamma,\delta} \geq \frac{\delta}{2}(1-\gamma)\omega$ on X.*
(c) *$(\alpha + \delta\omega + \frac{i}{2\pi}\partial\overline{\partial}\Phi_{\gamma,\delta})^n \geq a\,\gamma^n\delta^{n-p}\omega^n$ on $X \setminus Z_\delta$.*
(d) *$\sup_X \Phi_{1,\delta} = 0$, and for all $\gamma \in]0,1]$, there are estimates $\Phi_{\gamma,\delta} \leq A$ and*

$$\exp\big(-\Phi_{\gamma,\delta}\big) \leq e^{-(1+b\delta)\varphi_{\min}} \exp\big(A - \gamma\Phi_{1,\delta}\big)$$

(e) *For* γ_0, $\delta_0 > 0$ *small,* $\gamma \in]0, \gamma_0]$, $\delta \in]0, \delta_0]$, *we have*

$$\mathcal{I}_+(\varphi_{\min}) = \mathcal{I}(\varphi_{\min}) \subset \mathcal{I}(\Phi_{\gamma,\delta}).$$

Here a, b, A, γ_0, δ_0 *are suitable constants independent of* γ, δ.

Proof. — Denote by ψ_ε the (non-increasing) sequence of weight functions as stated in the definition of numerical dimension. We have $\psi_\varepsilon \geq \varphi_{\min}$, for all $\varepsilon > 0$, the ψ_ε have analytic singularities, and

$$\alpha + \frac{i}{2\pi}\partial\overline{\partial}\psi_\varepsilon \geq -\varepsilon\omega.$$

Then for $\varepsilon \leq \frac{\delta}{4}$, we have

$$\begin{aligned}\alpha + \delta\omega + \frac{i}{2\pi}\partial\overline{\partial}\big((1+b\delta)\psi_\varepsilon\big) &\geq \alpha + \delta\omega - (1+b\delta)(\alpha+\varepsilon\omega)\\ &\geq \delta\omega - (1+b\delta)\varepsilon\omega - b\delta\alpha \geq \tfrac{\delta}{2}\omega,\end{aligned}$$

for $b \in]0, \frac{1}{5}]$ small enough such that $\omega - b\alpha \geq 0$.

Let $\mu : \widehat{X} \to X$ be a log-resolution of ψ_ε, so that

$$\mu^*\big(\alpha + \delta\omega + \frac{i}{2\pi}\partial\overline{\partial}((1+b\delta)\psi_\varepsilon)\big) = [D_\varepsilon] + \beta_\varepsilon,$$

where $\beta_\varepsilon \geq \frac{\delta}{2}\mu^*\omega \geq 0$ is a smooth closed $(1,1)$-form on $\widehat{X}$ that is strictly positive in the complement $\widehat{X} \setminus E$ of the exceptional divisor, and D_ε is an effective $\mathbb{R}$-divisor that includes all components E_ℓ of E. The map μ can be obtained by Hironaka [26] as a composition of a sequence of blow-ups with smooth centres, and we can even achieve that D_ε and E are normal crossing divisors. For arbitrary small enough numbers $\eta_\ell > 0$, $\beta_\varepsilon - \sum \eta_\ell[E_\ell]$ is a Kähler class on $\widehat{X}$. Hence, we can find a quasi-psh potential $\widehat{\theta}_\varepsilon$ on $\widehat{X}$, such that $\widehat{\beta}_\varepsilon := \beta_\varepsilon - \sum \eta_\ell[E_\ell] + \frac{i}{2\pi}\partial\overline{\partial}\widehat{\theta}_\varepsilon$ is a Kähler metric on $\widehat{X}$. By taking the η_ℓ small enough, we may assume that

$$\int_{\widehat{X}} (\widehat{\beta}_\varepsilon)^n \geq \frac{1}{2}\int_{\widehat{X}} \beta_\varepsilon^n.$$

We will use Yau's theorem [37] to construct a form in the cohomology class of $\widehat{\beta}_\varepsilon$ with better volume estimate. We have

$$\begin{aligned}\alpha + \delta\omega + \frac{i}{2\pi}\partial\overline{\partial}\big((1+b\delta)\psi_\varepsilon\big) &\geq \alpha + \varepsilon\omega + \frac{i}{2\pi}\partial\overline{\partial}\psi_\varepsilon + (\delta-\varepsilon)\omega - b\delta(\alpha+\varepsilon\omega)\\ &\geq (\alpha + \varepsilon\omega + \frac{i}{2\pi}\partial\overline{\partial}\psi_\varepsilon) + \frac{\delta}{2}\omega.\end{aligned}$$

The assumption on the numerical dimension of L implies the existence of a constant $c>0$, such that, with $Z=\mu(E)\subset X$, we have

$$\begin{aligned}\int_{\widehat X}\beta_\varepsilon^n &= \int_{X\setminus Z}\big(\alpha+\delta\omega+\frac{i}{2\pi}\partial\overline\partial((1+b\delta)\psi_\varepsilon)\big)^n\\ &\geq \binom{n}{p}\Big(\frac{\delta}{2}\Big)^{n-p}\int_{X\setminus Z}\big(\alpha+\varepsilon\omega+\frac{i}{2\pi}\partial\overline\partial\psi_\varepsilon\big)^p\wedge\omega^{n-p}\geq c\delta^{n-p}\int_X\omega^n.\end{aligned}$$

Therefore, we may assume

$$\int_{\widehat X}(\widehat\beta_\varepsilon)^n\geq\frac{c}{2}\,\delta^{n-p}\int_X\omega^n.$$

We take $\widehat f$ a volume form on $\widehat X$, such that $\widehat f>\frac{c}{3}\delta^{n-p}\mu^*\omega^n$ everywhere on $\widehat X$ and such that $\int_{\widehat X}\widehat f=\int_{\widehat X}\widehat\beta_\varepsilon^n$. By Yau's theorem [37], there exists a quasi-psh potential $\widehat\tau_\varepsilon$ on $\widehat X$, such that $\widehat\beta_\varepsilon+\frac{i}{2\pi}\partial\overline\partial\widehat\tau_\varepsilon$ is a Kähler metric on $\widehat X$ with the prescribed volume form $\widehat f>0$.

Now push our focus back to X. Set $\theta_\varepsilon=\mu_*\widehat\theta_\varepsilon$ and $\tau_\varepsilon=\mu_*\widehat\tau_\varepsilon\in L^1_{\mathrm{loc}}(X)$. We define

$$\Phi_{\gamma,\delta}:=(1+b\delta)\psi_\varepsilon+\gamma(\theta_\varepsilon+\tau_\varepsilon).$$

By construction it is smooth in the complement $X\setminus Z_\delta$, i.e. property (a). It satisfies

$$\begin{aligned}&\mu^*\big(\alpha+\delta\omega+\frac{i}{2\pi}\partial\overline\partial((1+b\delta)\psi_\varepsilon+\gamma(\theta_\varepsilon+\tau_\varepsilon))\big)\\ &\quad=[D_\varepsilon]+(1-\gamma)\beta_\varepsilon+\gamma\Big(\sum_\ell\eta_\ell[E_\ell]+\widehat\beta_\varepsilon+\frac{i}{2\pi}\partial\overline\partial\widehat\tau_\varepsilon\Big),\\ &\quad\geq(1-\gamma)\beta_\varepsilon\geq\frac{\delta}{2}(1-\gamma)\,\mu^*\omega\end{aligned}$$

since $\widehat\beta_\varepsilon+\frac{i}{2\pi}\partial\overline\partial\widehat\tau_\varepsilon$ is a Kähler metric on $\widehat X$. Thus, the property (b) is satisfied. Putting $Z_\delta=\mu(|D_\varepsilon|)\supset\mu(E)=Z$, we have on $X\setminus Z_\delta$

$$\begin{aligned}\mu^*\big(\alpha+\delta\omega+\frac{i}{2\pi}\partial\overline\partial\Phi_{\gamma,\delta}\big)^n&\geq\big(\beta_\varepsilon+\gamma\frac{i}{2\pi}\partial\overline\partial(\widehat\theta_\varepsilon+\widehat\tau_\varepsilon)\big)^n\\ &\geq\gamma^n\,(\widehat\beta_\varepsilon+\frac{i}{2\pi}\partial\overline\partial\widehat\tau_\varepsilon)^n\geq\frac{c}{3}\,\gamma^n\delta^{n-p}\mu^*\omega^n.\end{aligned}$$

Since $\mu:\widehat X\setminus D_\varepsilon\to X\setminus Z_\delta$ is a biholomorphism, the condition (c) is satisfied, if we set $a=\frac{c}{3}$.

We adjust constants in $\widehat\theta_\varepsilon+\widehat\tau_\varepsilon$ so that $\sup_X\Phi_{1,\delta}=0$. Since $\varphi_{\min}\leq\psi_\varepsilon\leq\psi_{\varepsilon_0}\leq A_0:=\sup_X\psi_{\varepsilon_0}$ for $\varepsilon\leq\varepsilon_0$,

$$\Phi_{\gamma,\delta}=(1+b\delta)\psi_\varepsilon+\gamma\big(\Phi_{1,\delta}-\psi_\varepsilon\big)\geq(1+b\delta)\varphi_{\min}+\gamma\Phi_{1,\delta}-\gamma A_0,$$

and we have $\Phi_{\gamma,\delta}\leq(1-\gamma+b\delta)A_0$. Thus the property (d) is satisfied if we set $A:=(1+b)A_0$.

We observe that $\Phi_{1,\delta}$ satisfies $\alpha+\omega+dd^c\Phi_{1,\delta}\geq 0$ and $\sup_X \Phi_{1,\delta}=0$, hence $\Phi_{1,\delta}$ belongs to a compact family of quasi-psh functions. By Theorem 2.50, a uniform version of Skoda's integrability theorem in [25], there exists a uniform small constant $c_0>0$, such that $\int_X \exp(-c_0\Phi_{1,\delta})dV_\omega<+\infty$ for all $\delta\in\,]0,1]$. If $f\in\mathcal{O}_{X,x}$ is a germ of holomorphic function and U a small neighbourhood of x, the Hölder inequality combined with estimate (d) implies

$$\begin{aligned}&\int_U |f|^2\exp(-\Phi_{\gamma,\delta})dV_\omega\\ &\leq e^A\Big(\int_U |f|^2e^{-p(1+b\delta)\varphi_{\min}}dV_\omega\Big)^{\frac{1}{p}}\Big(\int_U |f|^2e^{-q\gamma\Phi_{1,\delta}}dV_\omega\Big)^{\frac{1}{q}}.\end{aligned}$$

Take $p\in\,]1,\lambda_0[$ (say $p=(1+\lambda_0)/2$) and take

$$\gamma\leq\gamma_0:=\frac{c_0}{q}=c_0\frac{\lambda_0-1}{\lambda_0+1}\quad\text{and}\quad\delta\leq\delta_0\in\,]0,1]\text{ so small that }p(1+b\delta_0)\leq\lambda_0.$$

Then $f\in\mathcal{I}_+(\varphi_{\min})=\mathcal{I}(\lambda_0\varphi_{\min})$ implies $f\in\mathcal{I}(\Phi_{\gamma,\delta})$, which proves the condition (e). □

The rest of the proof follows from the proof of [9] (cf. also [12], [16], [31]). We will just give an outline of the proof for completeness.

Let $\{f\}$ be a cohomology class in the group $H^q(X,K_X\otimes L\otimes\mathcal{I}(h_{\min}))$, $q\geq n-\mathrm{nd}(L)+1$. The sheaf $\mathcal{O}(K_X\otimes L)\otimes\mathcal{I}(h_{\min})$ can be resolved by the complex $(K^\bullet,\overline{\partial})$, where K^i is the sheaf of (n,i)-forms u, such that both u and $\overline{\partial}u$ are locally L^2 with respect to the weight $\varphi_{\min}$. So $\{f\}$ can be represented by a (n,q)-form f, such that both f and $\overline{\partial}f$ are L^2 with respect to the weight $\varphi_{\min}$, i.e. $\int_X|f|^2\exp(-\varphi_{\min})dV_\omega<+\infty$ and $\int_X|\overline{\partial}f|^2\exp(-\varphi_{\min})dV_\omega<+\infty$.

We can also equip L by the hermitian metric h_δ defined by the quasi-psh weight $\Phi_\delta=\Phi_{\gamma_0,\delta}$ obtained in Proposition 3, with $\delta\in\,]0,\delta_0]$. Since Φ_δ is smooth on $X\setminus Z_\delta$, the Bochner–Kodaira inequality shows that for every smooth (n,q)-form u with values in $K_X\otimes L$ that is compactly supported on $X\setminus Z_\delta$, we have

$$\|\overline{\partial}u\|_\delta^2+\|\overline{\partial}^*u\|_\delta^2\geq 2\pi\int_X(\lambda_{1,\delta}+\ldots+\lambda_{q,\delta}-q\delta)|u|^2e^{-\Phi_\delta}dV_\omega,$$

where $\|u\|_\delta^2:=\int_X|u|^2_{\omega,h_\delta}dV_\omega=\int_X|u|^2_{\omega,h_\infty}e^{-\Phi_\delta}dV_\omega$. Condition (b) of Proposition 3 shows that

$$0<\frac{\delta}{2}(1-\gamma_0)\leq\lambda_{1,\delta}(x)\leq\ldots\leq\lambda_{n,\delta}(x),$$

where $\lambda_{i,\delta}$ are at each point $x\in X$, the eigenvalues of $\alpha+\delta\omega+\frac{i}{2\pi}\partial\overline{\partial}\Phi_\delta$ with respect to the base Kähler metric ω. In other words, we have up to a multiple 2π

$$\|\overline{\partial}u\|_\delta^2+\|\overline{\partial}^*u\|_\delta^2+\delta\|u\|_\delta^2\geq\int_X(\lambda_{1,\delta}+\ldots+\lambda_{q,\delta})|u|^2_{\omega,h_\infty}e^{-\Phi_\delta}dV_\omega.$$

By the proof of Theorem 3.3 in [16], we have the following lemma:

Lemma 3.6. — *For every L^2 section of $\Lambda^{n,q}T_X^* \otimes L$ such that $\|f\|_\delta < +\infty$ and $\overline{\partial} f = 0$ in the sense of distributions, there exists a L^2 section $v = v_\delta$ of $\Lambda^{n,q-1}T_X^* \otimes L$ and a L^2 section $w = w_\delta$ of $\Lambda^{n,q}T_X^* \otimes L$ such that $f = \overline{\partial} v + w$ with*

$$\|v\|_\delta^2 + \frac{1}{\delta}\|w\|_\delta^2 \leq \int_X \frac{1}{\lambda_{1,\delta} + \ldots + \lambda_{q,\delta}} |f|^2 e^{-\Phi_\delta} dV_\omega.$$

By Lemma 3.6 and Condition (d) of Proposition 3, the error term w satisfies the L^2 bound,

$$\begin{aligned}\int_X |w|^2_{\omega,h_\infty} e^{-A} dV_\omega &\leq \int_X |w|^2_{\omega,h_\infty} e^{-\Phi_\delta} dV_\omega \\ &\leq \int_X \frac{\delta}{\lambda_{1,\delta} + \ldots + \lambda_{q,\delta}} |f|^2_{\omega,h_\infty} e^{-\Phi_\delta} dV_\omega.\end{aligned}$$

We will show that the right-hand term tends to 0 as $\delta \to 0$. To do it, we need to estimate the ratio function $\rho_\delta := \frac{\delta}{\lambda_{1,\delta}+\ldots+\lambda_{q,\delta}}$. The ratio function is first estimated in [31].

By Estimates (b,c) in Proposition 3, we have $\lambda_{j,\delta}(x) \geq \frac{\delta}{2}(1-\gamma_0)$ and $\lambda_{1,\delta}(x) \ldots \lambda_{n,\delta}(x) \geq a\gamma_0^n \delta^{n-p}$. Therefore, we already find $\rho_\delta(x) \leq 2/q(1-\gamma_0)$. On the other hand, we have

$$\begin{aligned}\int_{X \setminus Z_\delta} \lambda_{n,\delta}(x) dV_\omega \leq \int_X (\alpha + \delta\omega + dd^c \Phi_\delta) \wedge \omega^{n-1} &= \int_X (\alpha + \delta\omega) \wedge \omega^{n-1} \\ &\leq \text{Const},\end{aligned}$$

therefore, the "bad set" $S_\varepsilon \subset X \setminus Z_\delta$ of points x, where $\lambda_{n,\delta}(x) > \delta^{-\varepsilon}$ has a volume with respect to ω $\mathrm{Vol}(S_\varepsilon) \leq C\delta^\varepsilon$ converging to 0 as $\delta \to 0$. Outside of S_ε,

$$\lambda_{q,\delta}(x)^q \delta^{-\varepsilon(n-q)} \geq \lambda_{q,\delta}(x)^q \lambda_{n,\delta}(x)^{n-q} \geq a\gamma_0^n \delta^{n-p}.$$

Thus, we have $\rho_\delta(x) \leq C\delta^{1-\frac{n-p+(n-q)\varepsilon}{q}}$. If we take $q \geq n - \mathrm{nd}(L) + 1$ and $\varepsilon > 0$ small enough, the exponent of δ in the final estimate is strictly positive. Thus, there exists a sub-sequence (ρ_{δ_ℓ}), $\delta_\ell \to 0$ that tends almost everywhere to 0 on X.

Estimate (e) in Proposition 3 implies the Hölder inequality

$$\begin{aligned}&\int_X \rho_\delta |f|^2_{\omega,h_\infty} \exp(-\Phi_\delta) dV_\omega \\ &\quad\leq e^A \Big(\int_X \rho_\delta^p |f|^2_{\omega,h_\infty} e^{-p(1+b\delta)\varphi_{\min}} dV_\omega\Big)^{\frac{1}{p}} \Big(\int_X |f|^2_{\omega,h_\infty} e^{-q\gamma_0 \Phi_{1,\delta}} dV_\omega\Big)^{\frac{1}{q}},\end{aligned}$$

for suitable $p, q > 1$ as in the proposition. For some constant $C > 0$, $|f|^2_{\omega,h_\infty} \leq C$ since X is compact. Taking $\delta \to 0$ yields that $w_\delta \to 0$ in $L^2(h_\infty)$ by the Lebesgue dominating theorem.

$H^q(X, K_X \otimes L)$ is a finite-dimensional Hausdorff vector space whose topology is induced by the L^2 Hilbert space topology on the space of forms. In particular, the sub-space of coboundaries is closed in the space of cocycles. Hence, f is a co-boundary that completes the proof.

For any singular positive metric h on L, by definition, h is more singular than $h_{\min}$, which implies that $\mathcal{I}(h) \subset \mathcal{I}(h_{\min})$. A direct corollary of the above theorem is the following.

Corollary 3.7. — *Let (L, h) be a pseudo-effective line bundle on a compact Kähler n-dimensional manifold X. Then the morphism induced by inclusion $K_X \otimes L \otimes \mathcal{I}(h) \to K_X \otimes L$,*

$$H^q(X, K_X \otimes L \otimes \mathcal{I}(h)) \to H^q(X, K_X \otimes L)$$

is a 0 map, for every $q \geq n - \mathrm{nd}(L) + 1$.

4. Kawamata–Viehweg vanishing theorem

We first give a "numerical dimension version" of the Kawamata–Viehweg vanishing theorem in the projective case. In the following, we study various properties of nef classes in higher codimension. Then we end the section with a numerical version of the Kawamata–Viehweg vanishing theorem in the Kähler case.

To start with we need the relation between movable intersection defined in [6] and [3] and intersection number.

Lemma 4.1. — *Let α be a nef class in codimension p on a compact Kähler manifold (X, ω). Then, for any $k \leq p$ and Θ any positive closed $(n-k, n-k)$-form, we have*

$$(\alpha^k, \Theta) \geq \langle \alpha^k, \Theta \rangle.$$

Here, we use the definition of movable intersection defined in [3] and [6]. The movable intersection number $\langle \alpha^k, \Theta \rangle$ in [3] is defined as the limit for $\varepsilon > 0$ converging to 0 of the quantity:

$$\sup_{T_i} \int_{X \smallsetminus F} (T_1 + \varepsilon\omega) \wedge \cdots (T_k + \varepsilon\omega) \wedge \Theta \ ,$$

where T_i ranges all closed currents with analytic singularities in the class α, such that $T_i \geq -\varepsilon\omega$, and F is the union of all singular parts of T_i. (In [3], the movable intersection number is defined for any closed positive current Θ. In the following, we will take Θ to be ω^{n-k}. Thus, we consider only the case when Θ is a positive closed form.)

The proof of the boundedness of the quantity is a consequence of regularisation and the theory of Monge–Ampère operator. In the general case, we approximate the current T_i decreasingly by the smooth forms as shown in [13], with a uniform lower bound $-C\omega$ depending on (X,ω) and $\{T_i\}$. Now, on $X \smallsetminus F$, the current $(T_1 + C\omega) \wedge \cdots (T_k + C\omega) \wedge \Theta$ is the limit of corresponding terms changing T_i by its smooth approximation, using the continuity of Monge–Ampère operator with respect to decreasing sequence. However, the integral on $X \smallsetminus F$ obtained for the smooth approximation is bounded by its integral on X, which is the intersection number of cohomology classes $\{T_i + C\omega\}$ and $\{\Theta\}$.

Proof. — Our observation is that with better regularity on the cohomology class α, we can define directly the Monge–Ampère operator on X. So compared to the general case, we can skip the approximation process and get rid of the dependence of C, which only depends on (X,ω) and α, but not explicitly.

We recall the following Theorem (4.6) on the Monge–Ampère operators in Chapter 3 of [15].

Let $u_1, \dots, u_q$ be quasi-plurisubharmonic functions on X and T be a closed positive current of bidimension (p,p). The currents $u_1 i\partial\overline{\partial}u_2 \wedge \cdots \wedge i\partial\overline{\partial}u_q \wedge T$ and $i\partial\overline{\partial}u_1 \wedge i\partial\overline{\partial}u_2 \wedge \cdots \wedge i\partial\overline{\partial}u_q \wedge T$ are well defined and have locally finite mass in X as soon as $q \leq p$, and

$$H_{2p-2m+1}(L(u_{j_1}) \cap \cdots \cap L(u_{j_m}) \cap \mathrm{Supp}(T)) = 0 \ ,$$

for all choices of indices $j_1 < \cdots < j_m$ in $\{1, \dots, q\}$.

Here, $H_{2p-2m+1}$ means the $(2p-2m+1)$-dimensional Hausdorff content of the subset of X seen as a metric space induced by the Kähler metric. The unbounded locus $L(u)$ is defined to be the set of points $x \in X$, such that u is unbounded in every neighbourhood of x. When u has analytic singularities, it is the singular part of u (i.e. $\{u = -\infty\}$).

Now we return to the proof of the lemma. By definition, $T_{i,\min,-\varepsilon\omega}$ is less singular than T_i. Since for any $c > 0$, $E_c(T_{i,\min,-\varepsilon\omega})$ has codimension larger than $p+1$, the singular set of T_i, which has analytic singularities, is also of codimension larger than $p+1$. By Theorem (4.6) cited above, the current $(T_1 + \varepsilon\omega) \wedge \cdots (T_k + \varepsilon\omega) \wedge \Theta$ is well-defined on X. Thus, we have

$$\begin{aligned}\int_{X \smallsetminus F} (T_1 + \varepsilon\omega) \wedge \cdots (T_k + \varepsilon\omega) \wedge \Theta &\leq \int_X (T_1 + \varepsilon\omega) \wedge \cdots (T_k + \varepsilon\omega) \wedge \Theta \\ &= (\alpha + \varepsilon\{\omega\}) \cdot \cdots \cdot (\alpha + \varepsilon\{\omega\}) \cdot \{\Theta\}.\end{aligned}$$

Taking $\varepsilon \to 0$, we get $(\alpha^k, \Theta) \geq \langle \alpha^k, \Theta \rangle$. □

Remark 4.2. — From the definition of movable intersection, it is easy to see that the movable intersection is superadditive and that $\langle L^p \rangle \neq 0$ implies that $\mathrm{nd}(L) \geq p$. In fact, the inverse direction also holds. If (X,ω) is compact

Kähler, there exists $C > 0$, such that $C\{\omega\} - c_1(L)$ contains a closed positive current. Thus, we have that

$$\langle c_1(L)^p, C^{\mathrm{nd}(L)-p}\omega^{n-p}\rangle \geq \langle c_1(L)^{\mathrm{nd}(L)}, \omega^{n-\mathrm{nd}(L)}\rangle \neq 0.$$

In particular, $\mathrm{nd}(L)$ is the maximum p, such that $\langle c_1(L)^p\rangle \neq 0$.

In the projective case, we can now give the following version of the Kawamata–Viehweg theorem in terms of nefness in higher codimension. The simple proof given below was suggested to us by Demailly.

Theorem 4.3. — *Let X be a projective manifold and L a nef line bundle in codimension $p-1$. If $\langle c_1(L)^p\rangle \neq 0$, then for any $q \geq n-p+1$, we have*

$$H^q(X, K_X \otimes L) = 0.$$

Proof. — The proof is an induction on the dimension of X. Let A be an ample divisor on X and $\omega \in c_1(A)$ be a Kähler form. Let $Y \in |kA|$ be a generic smooth hypersurface. With the choice of k big enough, we can assume that $H^q(X, L^{-1} \otimes \mathcal{O}(-Y)) = 0$, for any $q < n$, by the Kodaira vanishing theorem. By Serre duality, the statement of the theorem is equivalent to prove that, for any $q \leq p-1$, we have

$$H^q(X, L^{-1}) = 0.$$

Consider the long exact sequence associated to the short exact sequence

$$0 \to L^{-1} \otimes \mathcal{O}(-Y) \to L^{-1} \to L^{-1}|_Y \to 0.$$

It turns out that it is enough to prove that $H^q(Y, L^{-1}) = 0$, for any $q \leq p-1$.

We check that conditions are preserved under the intersection with a generic hypersurface. Since α is nef in codimension $p-1$, we find that any irreducible component of

$$E_{nn}(\alpha) = \bigcup_{m\in\mathbb{N}^*}\bigcup_{n\in\mathbb{N}^*} E_{\frac{1}{n}}(T_{\min,\frac{1}{m}})$$

has codimension larger than p. By regularisation of $T_{\min,\frac{1}{m}}$, there exist currents T_m with analytic singularities in α larger than $-\frac{2}{m}\omega$. Any irreducible component of the singular set of these currents has a codimension larger than p. For generic Y, the restriction of these currents on Y is well defined for any m. Since the inclusion of analytic sets is a Zariski closed condition, for generic Y, we can also assume that the singular set of T_m is not contained in Y for any m.

On the other hand, in the class $\alpha|_Y$, the current with minimal singularities that admits a lower bound $-\frac{2}{m}\omega|_Y$ is certainly less singular than $T_m|_Y$. The upper-level set of the Lelong number of these minimal currents is included in the singular set of $T_m|_Y$, so it has codimension larger than p. This means that $\alpha|_Y$ is nef in codimension $p-1$.

The condition $\langle \alpha^p \rangle \neq 0$ implies that

$$\int_X \langle \alpha^p \rangle \wedge \omega^{n-p} > 0.$$

In other words, there exists a sequence of currents with analytic singularities $T_m \in \alpha$, such that $T_m \geq -\frac{1}{m}\omega$ and

$$\int_{X \smallsetminus F_m} (T_m + \frac{1}{m}\omega)^p \wedge \omega^{n-p} > c,$$

for some $c > 0$ independent of m, where F_m is the singular set of T_m.

With a generic choice of Y, we can still assume that the restriction of T_m is a current with analytic singularities. They satisfy the conditions $T_m|_Y \geq -\frac{1}{m}\omega|_Y$ and

$$\int_{Y \smallsetminus F_m} (T_m|_Y + \frac{1}{m}\omega|_Y)^p \wedge \omega^{n-p-1} > \frac{c}{k}\,.$$

In other words, $\langle \alpha|_Y^p \rangle \neq 0$.

By induction on the dimension, we are reduced to proving the case where X has dimension p, and L is nef in codimension $p-1$, in which case, L is (plainly) nef by Lemma 2.13. The condition of the movable intersection reduces to $\langle c_1(L)^p \rangle \neq 0$. By Lemma 4.1, this implies that $(L^p) > 0$. In particular, L is a nef and big line bundle. Now the vanishing of cohomology classes follows from the classical Kawamata–Viehweg theorem. □

As was pointed out to us by A. Höring, this can also be proven using the result of [28].

Remark 4.4. — When $p = n$, the above theorem is the classical Kawamata–Viehweg vanishing theorem for nef and big line bundle. We notice that $\langle c_1(L)^n \rangle = \mathrm{Vol}(L)$ by Theorem 3.5 of [6]. When $p = 1$, the theorem states that if L is a psef line bundle with $\langle c_1(L) \rangle \neq 0$, then $H^n(X, K_X \otimes L) = 0$. This case is trivial by the following easy lemma. The first interesting case is when L is nef in codimension 1 and $\langle c_1(L)^2 \rangle \neq 0$. In the following example, we show that we cannot weaken the condition to the case that L is only psef and $\langle c_1(L)^2 \rangle \neq 0$. On the other hand, by the divisorial Zariski decomposition, we can write any psef line bundle numerically as a sum of a nef class in codimension 1 and an effective class. This shows that, in some sense, this kind of theorem is the best we can hope for.

Now we begin our example. Let V be the unique non-trivial rank 2 extension of $\mathcal{O}_C$ over an elliptic curve C. Let X be the blow-up of a point of $\mathbb{P}(V) \times \mathbb{P}^1$ and L be the pull-back of $\mathcal{O}_{\mathbb{P}(V)}(1) \boxtimes \mathcal{O}_{\mathbb{P}^1}(1)$; $\mathcal{O}_{\mathbb{P}(V)}(1)$ is a nef line bundle. We also notice that $c_1(\mathcal{O}_{\mathbb{P}(V)}(1))^2 = 0$ and $c_1(\mathcal{O}_{\mathbb{P}(V)}(1)) \neq 0$. So, L is a nef line bundle over X, and $\mathrm{nd}(L) = 2$. By the above theorem, we have that

$H^2(X, K_X + L) = 0$. Let E be the exceptional divisor of the blow-up. The short exact sequence

$$0 \to K_X + L \to K_X + L + E \to K_X + L + E|_E \to 0$$

induces the long exact sequence

$$\begin{aligned} H^2(X, K_X + L) &\to H^2(X, K_X + L + E) \\ &\to H^2(E, K_X + L + E|_E) = H^0(E, -L) \to H^3(X, K_X + L). \end{aligned}$$

By Serre duality and the following lemma, $H^3(X, K_X + L) = H^0(X, -L) = 0$. Since $L|_E = \mathcal{O}_E$, $H^0(E, -L) \cong \mathbb{C}$. Thus, we have that

$$H^2(X, K_X + L + E) \cong \mathbb{C}.$$

Now $L + E$ is a psef line bundle over X, and $\mathrm{nd}(L + E) \geq 2$, but $H^2(X, K_X + L + E) \neq 0$. The reason for the numerical dimension is as follows. By the super-additivity of movable intersections, we have that

$$\langle (L + E)^2 \rangle \geq \langle L^2 \rangle + \langle E^2 \rangle + 2\langle L \cdot E \rangle \geq \langle L^2 \rangle.$$

Lemma 4.5. — *Let (L, h) be a non-trivial (i.e. $L \neq \mathcal{O}_X$) psef line bundle over a compact complex manifold X. Then we have*

$$H^0(X, L^{-1}) = 0.$$

Proof. — We argue by contradiction. Let s be a non-zero section in $H^0(X, L^{-1})$. Consider the function $\log |s|^2_{L^{-1}, h^{-1}}$. Let φ be the local weight of h, such that $h = e^{-\varphi}$ locally. Thus, the above function can be locally written as $\log |s|^2 + \varphi$. In particular, it is a psh function on X. Since X is compact, the only psh functions are the constant functions. On the other hand,

$$i\partial\overline{\partial} \log |s|^2_{L^{-1}, h^{-1}} = [s = 0] + i\Theta_{L,h} = 0,$$

where $[s = 0]$ is the current associated to the (possible trivial) divisor $s = 0$, and $i\Theta_{L,h}$ is the curvature of (L, h). Since both $[s = 0]$ and $i\Theta_{L,h}$ are positive currents, they are 0. In particular, s never vanishes on X, which contradicts the fact that L is a non-trivial line bundle. $\square$

A classical result for nef line bundles is the following. Let $A + B$ be a nef line bundle over a compact manifold X, where A, B are effective $\mathbb{R}$-divisors without intersection. Then A, B are both nef divisors. In the case of nefness in higher codimension, we have the following generalized version.

Lemma 4.6. — *Let $A + B$ be a line bundle that is nef in codimension k over a compact manifold X (by this we mean that $c_1(A + B)$ is nef in codimension k), where A, B are effective $\mathbb{R}$-divisors without intersection. Then the divisors A, B are both nef in codimension k.*

More generally, let $\alpha + c_1(E)$ be a class that is nef in codimension k over a compact manifold X, where E is an effective $\mathbb{R}$-divisor, and $E_{nn}(\alpha) \cap E = \emptyset$. Then α is nef in codimension k.

Proof. — Fix $\alpha_0 \in \alpha, \beta_0 \in c_1(E)$ as two smooth representatives. By assumption, for any $\varepsilon > 0$, there exists a quasi-psh function φ_ε on X with analytic singularities, such that

$$\alpha_0 + i\partial\overline{\partial}\varphi_\varepsilon \geq -\varepsilon\omega,$$

where ω is some Hermitian metric on X (not necessarily Kähler). (For example, we can take a regularisation of the minimal potential $\varphi_{\min,-\frac{\varepsilon}{2}}$.) We can assume that the singular set of φ_ε has an empty intersection with V_E. Here, V_E is some small tubular neighbourhood of E.

Let ψ_ε be a family of quasi-psh functions on X with analytic singularities, such that

$$\alpha_0 + \beta_0 + i\partial\overline{\partial}\psi_\varepsilon \geq -\varepsilon\omega.$$

We can assume that the singular set of ψ_ε has a codimension of at least $k+1$.

Let φ_E be a quasi-psh function on X, such that $\beta_0 + i\partial\overline{\partial}\varphi_E = [E]$, where $[E]$ is the current associated to E. By definition, the pole of φ_E is exactly the support of E. In particular, we have that $\psi_\varepsilon - \varphi_E$ is a well-defined quasi-psh function outside E, such that

$$\alpha_0 + i\partial\overline{\partial}(\psi_\varepsilon - \varphi_E) \geq -\varepsilon\omega,$$

on $X \smallsetminus E$.

Now we glue the potentials to get a quasi-psh function Φ_ε with analytic singularities on X, such that

$$\alpha_0 + i\partial\overline{\partial}\Phi_\varepsilon \geq -\varepsilon\omega.$$

We also demand that the singular set of Φ_ε be included in the singular set of ψ_ε. This will finish the proof of the lemma.

On $X \smallsetminus V_E$, we define that $\Phi_\varepsilon = \max(\psi_\varepsilon - \varphi_E, \varphi_\varepsilon + C_\varepsilon)$ where C_ε is a constant, which will be determined later. In particular, on $X \smallsetminus V_E$, we have

$$\alpha_0 + i\partial\overline{\partial}\Phi_\varepsilon \geq -\varepsilon\omega.$$

On V_E, we define $\Phi_\varepsilon = \varphi_\varepsilon + C_\varepsilon$. On $X \smallsetminus V_E$, φ_E is bounded, and ψ_ε is bounded from above. Near the boundary of V_E, φ_ε is also bounded since the singular set of φ_ε has empty intersection with V_E. Thus, for C_ε large enough near the boundary of V_E, $\psi_\varepsilon - \varphi_E < \varphi_\varepsilon + C_\varepsilon$. In particular, Φ_ε is a global well-defined quasi-psh function, such that $\alpha_0 + i\partial\overline{\partial}\Phi_\varepsilon \geq -\varepsilon\omega$. The singular set of Φ_ε in $X \smallsetminus V_E$ is included in the singular set of ψ_ε. On V_E, Φ_ε is smooth. This finishes our construction. □

REMARK 4.7. — The condition that the intersection is empty is necessary for the lemma. Otherwise, we have the following counterexample.

The construction uses Cutkosky's construction detailed in the next section. Let Y be a projective manifold, such that $\mathcal{N}_Y = \mathcal{E}_Y$. Let $\beta \in H^{1,1}(Y, \mathbb{R})$ be a non-psef class. Let A_1, A_2 be very ample divisors on Y. Define

$$t_0 := \min\{t | \beta + tc_1(A_1) \text{ nef}\}.$$

We can assume that $\beta + t_0 c_1(A_2)$ is nef. Define $X := \mathbb{P}(A_1 \oplus A_2)$ and denote by $\pi : \mathbb{P}(A_1 \oplus A_2) \to Y$ the natural projection. By Proposition 4 in the next section, $\pi^*\beta + t_0 c_1(\mathcal{O}(1))$ and $c_1(\mathcal{O}(1))$ are nef. Notice that $\mathcal{O}(1)$ is an effective divisor since $H^0(X, \mathcal{O}(1)) = H^0(Y, A_1 \oplus A_2) \neq 0$.

By Proposition 5, for any $t < t_0$, $\nu(\pi^*\beta + tc_1(\mathcal{O}(1)), \mathbb{P}(A_2)) > 0$ and $E_{nn}(\pi^*\beta + tc_1(\mathcal{O}(1)) = \mathbb{P}(A_2)$. This shows that for any $t < t_0$, $\pi^*\beta + tc_1(\mathcal{O}(1))$ is not nef in codimension 1. In other words, the nef class $\pi^*\beta + t_0 c_1(\mathcal{O}(1))$ is a sum of not nef in codimension 1 class $\pi^*\beta + tc_1(\mathcal{O}(1))$ and an effective divisor $(t_0 - t)\mathcal{O}(1)$. Let $(s_1, s_2) \in H^0(Y, A_1) \oplus H^0(Y, A_2) = H^0(X, \mathcal{O}(1))$ be a non-trivial section. Then we have

$$V(s_1, s_2) = \{(x, \xi^*) | \xi^* \in (A_1 \oplus A_2)^*,\ \xi^*(s_1, s_2) = 0\}.$$

Identifying $\mathbb{P}(A_2)$ as Y, we have $V(s_1, s_2) \cap \mathbb{P}(A_2) = V(s_2) \neq \emptyset$. A similar calculation shows that, for any $E \in |\mathcal{O}(m)|$ for any $m \in N^*$,

$$E_{nn}(\pi^*\beta + tc_1(\mathcal{O}(1)) \cap E \neq \emptyset.$$

Now we give a version of the Kawamata–Viehweg vanishing theorem over a compact Kähler manifold.

THEOREM 4.8. — *Let (X, ω) be a compact Kähler manifold of dimension n and L a nef in the codimension 1 line bundle on X. Assume that $\langle L^2 \rangle \neq 0$. Assume that there exists an effective $\mathbb{N}$-divisor D, such that $c_1(L) = c_1(D)$. Then*

$$H^q(X, K_X + L) = 0,$$

for $q \geq n - 1$.

Proof. — In the case $q = n$, we have $H^n(X, K_X + L) = H^0(X, -L)^*$ by Serre duality. For L psef, $-L$ has no section unless L is trivial by Lemma 4.5. Since $\langle L^2 \rangle \neq 0$, L is not trivial. Therefore, the only interesting case is $q = n - 1$. We divide the proof into two cases.

Case 1. We assume that $L = D$. Since the canonical section of D induces a positive singular metric on (L, h) with multiplier ideal sheaf $\mathcal{I}(h) \subset \mathcal{O}(-D)$. In fact, we have equality outside an analytic set whose all irreducible components have codimension larger than 2. Write $D = \sum_i n_i D_i$ where $n_i \geq 0$ and D_i are the irreducible components of D. Define

$$Y = (D_{\mathrm{red}})_{\mathrm{Sing}} = \bigcup_{i \neq j} (D_i \cap D_j) \cup \bigcup_i D_{i,\mathrm{Sing}},$$

where $D_{i,\mathrm{Sing}}$ means the singular part of D_i. It is easy to see that we have an equality outside Y and that each irreducible component of Y is of codimension larger than 2.

In particular, the short exact sequence

$$0 \to \mathcal{I}(h) \to \mathcal{O}(-D) \to \mathcal{O}(-D)/\mathcal{I}(h) \to 0$$

induces that

$$\begin{aligned} H^{n-1}(X, K_X + L \otimes \mathcal{I}(h)) &\to H^{n-1}(X, K_X + L - D) \\ &\to H^{n-1}(X, K_X + L \otimes \mathcal{O}(-D)/\mathcal{I}(h)) = 0 \end{aligned}$$

since the support of $\mathcal{O}(-D)/\mathcal{I}(h)$ is included in Y.

Denote by $h_{\min}$ the minimal metric on L, where we have a natural inclusion of $\mathcal{I}(h) \subset \mathcal{I}(h_{\min})$. Thus, we have the following commuting diagram

$$\begin{array}{ccc} H^{n-1}(X, K_X + L \otimes \mathcal{I}(h)) & \longrightarrow & H^{n-1}(X, K_X + L \otimes \mathcal{I}(h_{\min})), \\ \downarrow & & \downarrow \\ H^{n-1}(X, K_X + L - D) & \longrightarrow & H^{n-1}(X, K_X + L). \end{array}$$

By Corollary 3.7 proved in Section 3 and the condition that $\mathrm{nd}(L) \geq 2$, we know that the morphism

$$H^{n-1}(X, K_X + L \otimes \mathcal{I}(h_{\min})) \to H^{n-1}(X, K_X + L)$$

is the 0 map. Since the left vertical arrow is surjective in the above diagram, we conclude that the morphism

$$H^{n-1}(X, K_X + L - D) \to H^{n-1}(X, K_X + L)$$

is also the 0 map. Thus, the short exact sequence

$$0 \to K_X + L - D \to K_X + L \to K_X + L|_D = K_D \to 0$$

gives in cohomology

$$\begin{aligned} H^{n-1}(X, K_X + L - D) \to H^{n-1}(X, K_X + L) \to H^{n-1}(D, K_D) \simeq H^0(D, \mathcal{O}_D) \\ \to H^n(X, K_X + L - D) \to 0. \end{aligned}$$

On the other hand, $H^n(X, K_X + L - D) \simeq H^0(X, \mathcal{O}_X) \simeq \mathbb{C}$. Therefore, we need only show that

$$h^0(D, \mathcal{O}_D) = 1.$$

More precisely, D is an effective Cartier divisor in the manifold X. Notice that D is a (possibly non-reduced) Gorenstein variety. In this case, adjunction gives the dualising sheaf K_D as $(K_X+D)|_D$. Moreover, Serre duality holds in the same form as in the smooth case.

To calculate the dimension of global sections of D, first we show that D is connected. In fact, otherwise we would have $D=A+B$ with A and B effective non-trivial divisors, such that $A\cap B=\emptyset$. In particular, we have $(A\cdot B\cdot\omega^{n-2})=0$. However, A and B are necessarily nef in codimension 1 by Lemma 4.6.

We recall the Hodge index theorem on a compact Kähler manifold (X,ω) as Theorem 6.33 and 6.34 in [36]. By the Hard Lefschetz theorem, we have

$$H^2(X,\mathbb{C})=\mathbb{C}\{\omega\}\oplus H^2(X,\mathbb{C})_{\mathrm{prim}},$$

where $H^2(X,\mathbb{C})_{\mathrm{prim}}$ means primitive classes. The intersection form $(\alpha,\beta)\mapsto(\alpha\cdot\beta\cdot\omega^{n-2})$ has the signature $(1,h^{1,1}(X)-1)$ on $H^{1,1}(X)$ since $H^2(X,\mathbb{C})_{\mathrm{prim}}$ is orthogonal to ω, and the intersection form is negative definite on $H^2(X,\mathbb{C})_{\mathrm{prim}}$.

On the other hand, by Lemma 4.1, we have that

$$(A\cdot A\cdot\omega^{n-2})\geq\langle A\cdot A\cdot\omega^{n-2}\rangle\geq 0$$

and a similar inequality for B. We also notice that

$$(L\cdot L\cdot\omega^{n-2})\geq\langle L\cdot L\cdot\omega^{n-2}\rangle>0.$$

Since the intersection form (unlike the movable intersection) is bilinear, we have either $(A\cdot A\cdot\omega^{n-2})>0$ or $(B\cdot B\cdot\omega^{n-2})>0$. Without loss of generality, assume that $(A\cdot A\cdot\omega^{n-2})>0$. Thus, $B\in A^{\perp}$ and $(B\cdot B\cdot\omega^{n-2})\geq 0$. The Hodge index theorem implies that $B=0$, which is a contradiction to our assumption. Hence, D is connected, and if $h^0(D,\mathcal{O}_D)\geq 2$, then $\mathcal{O}_D$ contains a nilpotent section $t\neq 0$. In other words, the pull-back of t under the natural morphism $D_{\mathrm{red}}\to D$ is 0 but lies as a non-trivial section in $H^0(D_{\mathrm{red}},\mathcal{O}(-\sum_{j\in I}\mu_jD_j))$, for some $1\leq\mu_j\leq n_j$, for all j. Let

$$J:=\left\{j\in J\;\middle|\;\frac{n_j}{\mu_j}\text{ maximal}\right\}$$

and let $c=\frac{n_j}{\mu_j}$ be the maximal value. Notice that $\mathrm{div}(t)|_{D_i}=-\sum_{j\in I}\mu_jD_j|_{D_i}$ is effective (possibly 0) for all i. We claim that it is impossible that $c=\frac{n_j}{\mu_j}$ for all $j\in I$. Otherwise, $L|_{D_i}=c\sum\mu_jD_j|_{D_i}$ is psef. (L is nef in codimension 1, so its restriction to any prime divisor is psef.) Its dual is effective, hence $L|_{D_i}\equiv 0$, for all i. This implies that $(L\cdot L\cdot\omega^{n-2})=0$, which is a contradiction.

Thus, we find some j, such that

$$c>\frac{n_j}{\mu_j}.$$

By connectedness of D we can choose $i_0 \in J$ in such a way that there exists $j_1 \in I \setminus J$ with $D_{i_0} \cap D_{j_1} \neq \emptyset$. Now

$$\sum_{j \in I} (n_j - c\mu_j) D_j|_{D_{i_0}}$$

is pseudo-effective as a sum of a psef and an effective line bundle (this has nothing to do with the choice of i_0). Since the sum taken over I is the same as the sum taken over $I \setminus \{i_0\}$, we conclude that

$$\sum_{j \neq i_0} (n_j - c\mu_j) D_j|_{D_{i_0}}$$

is pseudo-effective, too. Now all $n_j - c\mu_j \leq 0$ and $n_{j_1} - c\mu_{j_1} < 0$ with $D_{j_1} \cap D_{i_0} \neq \emptyset$, hence the dual of

$$\sum_{j \neq i_0} (n_j - c\mu_j) D_j|D_{i_0}$$

is effective and non-zero, which is a contradiction. This finishes the proof of Case 1.

Case 2. General case. We can write

$$L = D + L_0,$$

where $L_0^m \in \mathrm{Pic}^0(X)$ (the exponent m is there because there might be torsion in $H^2(X, \mathbb{Z})$; we take m to kill the denominator of the torsion part). We may, in fact, assume that $m = 1$; otherwise, we pass to a finite étale cover $\tilde{X}$ of X and argue there (the vanishing on $\tilde{X}$ clearly implies the vanishing on X by Leray spectral sequence). In other words, we write L as a sum of D and a flat line bundle (L_0, h_0). Here, h_0 is the flat metric. Thus, there exists a bijection between singular positive metrics on L and those on D, via the tensor product by h_0. In particular, the minimal metric on L is the minimal metric on D, tensored by h_0.

The short exact sequence used above is modified into

$$0 \to K_X + L - D \to K_X + L \to (K_X + L)|_D = (K_D + L_0)|_D \to 0.$$

Taking cohomology as before and using a similar discussion the arguments come down to proving

$$H^0(D, -L_0|_D) = 0$$

since $H^n(X, K_X + L - D) \simeq H^0(X, -L_0) = 0$.

The argument on the connectedness of D still works since the arguments only involve the first Chern class and since L_0 has no contribution in the first Chern class. If $-L_0|_D \neq 0$, then we see, as above, that $-L_0|_D$ cannot have a nilpotent section. Since L_0 is flat adding a multiple of L_0 does not change the pseudo-effectiveness. By adding a suitable such multiple, the arguments on the non-existence of nilpotent section are still valid.

So, if $H^0(D, -L_0|_D) = 0$ fails, then $-L_0|_D$ has a section s, such that $s|_{D_{\mathrm{red}}}$ has no zeroes. In other words, $-L_0|_{D_{\mathrm{red}}}$ is trivial. However, then $-L_0|_D$ is trivial since the nowhere vanishing section of $H^0(X, -L_0 \otimes_{\mathcal{O}_X} \mathcal{O}_X/\mathcal{I}_{D_{\mathrm{red}}})$ is mapped to a nowhere vanishing section in $H^0(X, -L_0 \otimes_{\mathcal{O}_X} \mathcal{O}_X/\mathcal{I}_D)$ by passing to the quotient.

Now let $\alpha : X \to \mathrm{Alb}(X)$ be the Albanese map with image Y. Then $L_0 = \alpha^*(L'_0)$ with some line bundle L'_0 on $\mathrm{Alb}(X)$. (We observe that $\mathrm{Pic}^0(X) \cong \mathrm{Pic}^0(\mathrm{Alb}(X))$.) Notice that L'_0 is a non-trivial line bundle with $c_1(L'_0) = 0$. Since $L_0|_D$ is trivial and L_0 is non-trivial we conclude that $\alpha(D) \neq Y$. We claim that $\alpha(D)$ is contained in some proper subtorus B of $\mathrm{Alb}(X)$.

The reasoning is as follows. Let $\nu : \tilde{X} \to X$ be a modification, such that $\nu^*(D)$ is an SNC divisor. Denote by E_j the irreducible components of $\nu^*(D)$. Define as $S \subset \prod_i \mathrm{Pic}^0(E_i)$ the connected component containing $(\nu^* L_0|_{E_i})$ of

$$\{(L_i) \in \prod_i \mathrm{Pic}^0(E_i) | L_i|_{E_i \cap E_j} = L_j|_{E_i \cap E_j}\}.$$

By Proposition 1.5 of [1], S is a subtorus since S is a translation of the kernel of

$$\prod_i \mathrm{Pic}^0(E_i) \to \prod_{i,j,i\neq j} \mathrm{Pic}^0(E_i \cap E_j),$$
$$(L_i) \mapsto (L_i|_{E_i \cap E_j} - L_j|_{E_i \cap E_j}).$$

Notice that $\mathrm{Pic}^0(E_i)$ is a torus by Hodge theory since E_i is smooth. The natural group morphism of $\mathrm{Pic}^0(X) \to S$ given by $L \mapsto (\nu^* L|_{E_i})$ induces by duality the following commuting diagram

$$\begin{tikzcd}
\prod_i \mathrm{Pic}^0(E_i)^* \arrow[r] \arrow[dr] & S^* \arrow[d] \\
& (\mathrm{Pic}^0(X))^* \cong \mathrm{Alb}(X).
\end{tikzcd}$$

Since $L_0 \in S$ is non-trivial, the image of S^* as a complex torus is a proper subtorus in $\mathrm{Alb}(X)$. We denote its image by B. (Let us observe that by Proposition 1.5 of [1] that the image of a homomorphism of complex tori is a subtorus.)

Consider the induced map

$$\beta : X \to \mathrm{Alb}(X)/B$$

and denote its image by Z. (Z can be singular!) The image $\beta(D)$ is a point p by construction. Let U be a Stein neighbourhood of p in Z (or some coordinate chart of p). Denote by m_p the maximal ideal of p in Z. In particular, for any $k \in \mathbb{N}^*$, m_p^k is globally generated on U (by Cartan theorem A).

Let $D = \sum_i n_i D_i$ and define $n_{\max} := \max(n_i)$. Then we have the inclusion $\beta^* H^0(U, m_p^{n_{\max}}) \subset H^0(D, \mathcal{O}(-n_{\max} D_{\mathrm{red}})|_D) \subset H^0(D, \mathcal{O}(-D)|_D)$, where the second inclusion uses the fact that $n_{\max} D_{\mathrm{red}} - D$ is an effective divisor in X. In particular, for any i, $H^0(D_i, \mathcal{O}(-D)|_{D_i}) \neq 0$. On the other hand, $\mathcal{O}(D)|_{D_i}$ is psef since D is nef in codimension 1. (Observe that nefness is a numerical property. Since $c_1(L_0) = 0$, D is nef in codimension 1 as L is.) By Lemma 4.5, $D|_{D_i}$ is trivial.

Thus, we have, for any i,

$$(D \cdot D_i \cdot \omega^{n-2}) = \int_{D_i} c_1(D|_{D_i}) \wedge \omega^{n-2} = 0.$$

This implies that $(L^2 \cdot \omega^{n-2}) = (D^2 \cdot \omega^{n-2}) = 0$. On the other hand, since L is nef in codimension 1, $(L^2 \cdot \omega^{n-2}) \geq \langle L^2 \cdot \omega^{n-2} \rangle$. However, this is a contradiction with our assumption. □

Remark 4.9. — If D is a smooth reduced divisor, we can also argue as follows at the end of Case 2. We observe that L_0 is a non-trivial element in a translation of the kernel of $\mathrm{Pic}^0(X) \to \mathrm{Pic}^0(D)$. On the other hand, we have

$$H^{n-1}(X, K_X + D) = H^1(X, -D) = 0 \to H^1(X, \mathcal{O}_X) \to H^1(D, \mathcal{O}_D)$$

since by Case 1, $H^{n-1}(X, K_X + D) = 0$. However, $H^1(X, \mathcal{O}_X) \to H^1(D, \mathcal{O}_D)$ is the tangent map of $\mathrm{Pic}^0(X) \to \mathrm{Pic}^0(D)$. By Proposition 1.5 of [1], the kernel is discrete. Moreover, the connected component containing the zero point of the kernel is of finite index in the kernel. In particular, L_0 is a torsion element. This yields a contradiction.

5. Examples and counterexamples

In this section, we first give for each $k \in \mathbb{N}^*$ an example of a psef class α_k on some manifold X_k, such that α_k is nef in codimension k but not nef in codimension $k+1$. This shows, in particular, that the inclusion of the various types of nef cones can be strict.

For the convenience of the reader, we recall Cutkosky's construction described in [5], as well as all the material needed for our use.

Let $\mathcal{E}$ be a vector bundle of rank r over a manifold Y and L be a line bundle over Y. Since there exists a surjective bundle morphism given by projection $\mathcal{E} \oplus L \to \mathcal{E}$, we can view $D := \mathbb{P}(\mathcal{E})$ as a closed submanifold of $\mathbb{P}(\mathcal{E} \oplus L)$. Note that the restriction of $\mathcal{O}_{\mathbb{P}(\mathcal{E} \oplus L)}(1)$ on $\mathbb{P}(\mathcal{E})$ is the tautological line bundle $\mathcal{O}_{\mathbb{P}(\mathcal{E})}(1)$. We notice that the canonical line bundle of the projectivisation of a vector bundle $\mathbb{P}(\mathcal{E})$ is given by

$$K_{\mathbb{P}(\mathcal{E})} = \mathcal{O}_{\mathbb{P}(\mathcal{E})}(-(r+1)) + \pi^*(K_Y + \det\mathcal{E}),$$

where $\pi : \mathbb{P}(\mathcal{E}) \to Y$ is the projection. From the short exact sequence

$$0 \to T_{\mathbb{P}(\mathcal{E})} \to T_{\mathbb{P}(\mathcal{E}\oplus L)}|_{\mathbb{P}(\mathcal{E})} \to N_{\mathbb{P}(\mathcal{E})/\mathbb{P}(\mathcal{E}\oplus L)} = \mathcal{O}(D)|_{\mathbb{P}(\mathcal{E})} \to 0,$$

we have

$$K_{\mathbb{P}(\mathcal{E}\oplus L)}|_{\mathbb{P}(\mathcal{E})} = K_{\mathbb{P}(\mathcal{E})} \otimes \mathcal{O}(-D)|_{\mathbb{P}(\mathcal{E})}.$$

Using the formula for the canonical line bundle, we have

$$\mathcal{O}(1)|_{\mathbb{P}(\mathcal{E})} = (\mathcal{O}(D) \otimes \pi^* L)|_{\mathbb{P}(\mathcal{E})}.$$

We observe that by the Leray–Hirsh theorem for Bott–Chern cohomology,

$$H^{1,1}_{BC}(\mathbb{P}(\mathcal{E} \oplus L), \mathbb{R}) = \mathbb{R}c_1(\mathcal{O}(1)) \oplus \pi^* H^{1,1}_{BC}(Y, \mathbb{R}).$$

In particular, this implies that the inclusion $i : \mathbb{P}(\mathcal{E}) \to \mathbb{P}(\mathcal{E} \oplus L)$ induces an isomorphism on $H^{1,1}_{BC}$. Hence, we find that on $\mathbb{P}(\mathcal{E} \oplus L)$,

$$c_1(\mathcal{O}(1)) = c_1(\mathcal{O}(D)) + \pi^* c_1(L).$$

Now let Y be a compact complex manifold of dimension m and $L_0, \dots, L_r$ the line bundles over Y. We define

$$X := \mathbb{P}(L_0 \oplus \cdots \oplus L_r).$$

We denote $H := \mathcal{O}(1)$ the tautological line bundle over the projectivisation and $h := c_1(H)$. For any i, the projection $L_0 \oplus \cdots L_r \to L_0 \oplus \cdots \hat{L}_i \oplus \cdots \oplus L_r$ induces inclusions of hypersurfaces

$$D_i := \mathbb{P}(L_0 \oplus \cdots \oplus \hat{L}_i \oplus \cdots \oplus L_r).$$

By the above discussion

$$d_i + l_i = h,$$

where $d_i := c_i(\mathcal{O}(D_i))$ and $l_i := c_1(L_i)$. In fact, by calculating the transition function, we can show that $\mathcal{O}(1)$ is linearly equivalent to $L_i + D_i$. However, the relation of Chern classes is enough for our use here.

We have the following explicit description of nef cone and psef cone in this case. We denote by $\mathcal{C}$ the cone generated by the l_i.

Proposition 5.1. — *Let $\alpha \in H^{1,1}_{BC}(X, \mathbb{R})$ be a class that is decomposed as $\alpha = \pi^*\beta + \lambda h$. Then*

(1) *α is nef iff $\lambda \geq 0$ and $\beta + \lambda\mathcal{C}$ is contained in $\mathcal{N}_Y$.*
(2) *α is psef iff $\lambda \geq 0$ and $(\beta + \lambda\mathcal{C}) \cap \mathcal{E}_Y \neq \emptyset$.*

Proof. — We notice that if α contains a positive current $T = \theta + i\partial\overline{\partial}\varphi$ with θ smooth, then the pluripolar set $P(\varphi) = \{\varphi = -\infty\}$ is of Lebesgue measure 0. Hence, by the Fubini theorem, the set

$$\{y \in Y, \pi^{-1}(y) \subset P(\varphi)\}$$

is of Lebesgue measure 0. For y outside the measure 0 set, $\alpha|_{\pi^{-1}(z)}$ is the class of $T|_{\pi^{-1}(z)}$. It is also equal to the class of $\lambda c_1(\mathcal{O}_{\mathbb{P}^r}(1))$, and this implies that $\lambda \geq 0$. We always assume in the following that $\lambda \geq 0$.

(1) If α is nef, the restriction of α to $\mathbb{P}(L_i)$, for any i, is also nef, where $\mathbb{P}(L_i)$ is biholomorphic to Y given by π. Note that $\alpha|_{\mathbb{P}(L_i)} = \lambda l_i + \beta$ is nef as a restriction of nef class. So $\beta + \lambda\mathcal{C}$ is contained in $\mathcal{N}_Y$.

On the other hand, $\alpha = \pi^*\beta + h = \pi^*(\beta + \lambda l_i) + \lambda d_i$ for any i where $\beta + \lambda l_i$ is nef by assumption. Hence, the non-nef locus of α is contained in D_i. Since the intersection of all D_i is empty, we conclude that α is nef.

(2) Let $t_i \in [0,1]$ such that $\sum t_i = 1$ and $\beta + \sum_{i=0}^r t_i l_i \in \mathcal{E}_Y$. Hence, $h = \sum t_i h = \sum t_i \pi^* l_i + \sum t_i d_i$, and $\alpha = \pi^*(\beta + \lambda \sum t_i l_i) + \lambda \sum t_i d_i$; d_i is psef since it contains the positive current associated to D_i. As a sum of psef classes, α is psef.

For the other direction, we argue by induction. When $r = 0$, $X = Y$ and $\alpha = \beta + \lambda l_0$. By the assumption that α is psef, we have

$$\alpha \in (\beta + \lambda\mathcal{C}) \cap \mathcal{E}_Y.$$

Continue the induction on r. Letting T be a closed positive current in α, we have that $\alpha - \nu(T, D_0)d_0$ is psef containing the current $T - \nu(T, D_0)[D_0]$. Moreover, $(\alpha - \nu(T, D_0)d_0)|_{D_0}$ is psef since the restriction of the current $T - \nu(T, D_0)[D_0]$ on D_0 is well defined. Now D_0 is the projectivisation of a vector bundle of rank r over Y. As a cohomology class

$$\alpha - \nu(T, D_0)d_0 = \pi^*(\beta + \lambda l_0) + (\lambda - \nu(T, D_0))d_0.$$

Restricting α on some fibre of π as above we have that $\lambda \geq \nu(T, D_0)$. By induction, we see that the psef class $(\alpha - \nu(T, D_0)d_0)|_{D_0}$, which is also equal to $\pi^*(\beta + \nu(T, D_0)l_0) + (\lambda - \nu(T, D_0))h$, which satisfies

$$(\beta + \nu(T, D_0)l_0 + (\lambda - \nu(T, D_0))\mathcal{C}_0) \cap \mathcal{E}_Y \neq \emptyset,$$

where $\mathcal{C}_0$ is the cone generated by $l_1, \ldots, l_r$. In other words,

$$(\beta + \lambda\mathcal{C}) \cap \mathcal{E}_Y \neq \emptyset. \qquad \square$$

We will also need the following explicit calculation of the generic minimal multiplicity in this example. From now on, we choose Y such that the nef cone $\mathcal{N}_Y$ and the psef cone $\mathcal{E}_Y$ coincide (for example, we can take Y to be a Riemann surface).

We denote I a subset of $\{0, \cdots, r\}$ with complement J. We denote $V_I := \bigcap_{i\in I} D_i = \mathbb{P}(\oplus_{j\in J} L_j)$ and $\mathcal{C}_I$ the convex envelope of $l_i (i \in I)$.

We observe that the non-nef locus of any psef class is contained in the union of D_i. The reason is as follows: since $\alpha = \pi^*\beta + \lambda h$ is psef, by Proposition 4 we know that there exist $t_i \in [0,1]$ with $\sum t_i = 1$, such that $\beta + \lambda(\sum t_i l_i) \in$

$\mathcal{E}_Y = \mathcal{N}_Y$. Hence,

$$\alpha = \pi^*\Big(\beta + \lambda\Big(\sum t_i l_i\Big)\Big) + \lambda\Big(\sum t_i d_i\Big)$$

is a sum of nef divisor and effective divisor. (Since α is psef, $\lambda \geq 0$.) So the non-nef locus of α is contained in the union of D_i.

Proposition 5.2. — *Let α be a big class, such that $\alpha = \pi^*\beta + \lambda h$. The generic minimal multiplicity of α along V_I is equal to*

$$\nu(\alpha, V_I) = \min\{t \geq 0, (\beta + tC_I + (\lambda - t)C_J) \cap N_Y \neq \emptyset\}.$$

More precisely, we have $\nu(\alpha, V_I) = \nu(\alpha, x)$, for any $x \in V_I \setminus \bigcup_{j \in J} D_j$.

Proof. — Let $\mu : X_I \to X$ be the blow-up of X along V_I with the exceptional divisor E_I. Hence, we have $E_I = \mathbb{P}(N^*_{V_I/X})$ with $N^*_{V_I/X} = \oplus_{i \in I} \mathcal{O}_{V_i}(-D_i)$. By Lemma 2.5, we get

$$\nu(\alpha, V_I) = \nu(\mu^*\alpha, E_I).$$

Denote by H_I the tautological line bundle over E_I, where we have $\mathcal{O}_{E_I}(-E_I) = H_I$.

For $t \geq 0$, the restriction of $\mu^*\alpha - tc_1(\mathcal{O}(E_I))$ to E_I is psef and is, hence, equivalent to that $\mu^*\alpha + tc_1(H_I)$ is psef. By Proposition 4, the latter is equivalent to the fact that $\alpha + t\mathcal{C}(\pi^* l_i - h) = \alpha - th + t\pi^*\mathcal{C}(l_i)$ intersects $\mathcal{E}_{V_I}$, where $\mathcal{C}(l_i)$ is the convex envelope of l_i $(i \in I)$. Note also that

$$\alpha - th + t\pi^*\mathcal{C}(l_i) = \pi^*(\beta + t\mathcal{C}(l_i)) + (\lambda - t)h,$$

where we denote by the same notation π to be the projection from V_I to Y and h to be the first Chern class of the tautological line bundle over V_I. By Proposition 4, it is psef if and only if $\beta + t\mathcal{C}_I + (\lambda - t)C_J$ intersects the psef cone $\mathcal{E}_Y$.

Since the class $\mu^*\alpha - \nu(\alpha, V_I)c_1(\mathcal{O}(E_I))$ has positive current $\mu^* T_{\min} - \nu(T_{\min}, V_I)[E_I]$, whose restriction to E_I is well defined by Siu's decomposition theorem. By the last paragraph we have

$$\nu(\alpha, V_I) = \nu(T_{\min}, V_I) \geq \min\{t \geq 0, (\beta + tC_I + (\lambda - t)C_J) \cap N_Y \neq \emptyset\}.$$

In the other direction, let $\gamma := \beta + t\sum_{i \in I} a_i l_i + (\lambda - t)\sum_{j \in J} b_j l_j$ be a psef (equivalently nef by assumption) class on Y with $\sum a_i = \sum b_j = 1$. Hence, $\alpha = \pi^*\gamma + t\sum a_i d_i + (\lambda - t)\sum b_j d_j$. For $x \in V_I \setminus \bigcup_{j \in J} D_j$,

$$\nu(\alpha, x) \leq t\sum a_i \nu([D_i], x) + (\lambda - t)\sum b_j \nu([D_j], x) \leq t\sum a_i = t.$$

In particular, this shows that

$$\nu(\alpha, V_I) \leq \min\{t \geq 0, (\beta + tC_I + (\lambda - t)C_J) \cap N_Y \neq \emptyset\}.$$

By the proof, the equality is attained for $x \in V_I \setminus \bigcup_{j \in J} D_j$. □

We notice that if we use the algebraic analogue in the projective case as in [32], we can weaken the assumption to the case that α is just a psef class.

In particular, Proposition 5 shows that $\cup D_i$ is stratified by the sets $V_I \smallsetminus \bigcup_{j\in J} D_j$ with respect to the generic minimal multiplicity.

Now we are prepared to give our construction. Let Y as above be a projective manifold, such that the nef cone coincides with the psef cone. Define $X_k = \mathbb{P}(\mathcal{O}_Y \oplus \mathcal{O}_Y(A_1) \oplus \cdots \oplus \mathcal{O}_Y(A_{k+1}))$, where A_i are the ample line bundles over Y. Let $\beta \in H^{1,1}_{BC}(Y, \mathbb{R})$ be a not-nef class. Denote H the tautological line bundle over X_k and denote h its first Chern class. Define $\alpha = \pi^*\beta + h$. We assume that:

For any i, $\beta + c_1(A_i)$ is nef and big.

As above, $\mathbb{P}(\mathcal{O}_Y) \simeq Y$ is a closed submanifold of X_k of codimension $k+1$ via the projection of $\mathcal{O}_Y \oplus \mathcal{O}_Y(A_1) \oplus \cdots \oplus \mathcal{O}_Y(A_{k+1}) \to \mathcal{O}_Y$; α is psef but not nef on X_k by Proposition 4. In fact, if α is nef, its restriction to the submanifold Y (i.e. β) will be nef. For any subset $I \neq \{1, \cdots r\}$ (taking $L_0 := \mathcal{O}_Y$), by Proposition 5, $\nu(\alpha, V_I) = 0$ since $\beta + \sum_{j\in J} c_1(A_j)$ is nef, which means we can take $t = 0$ on the right-hand of the equation. By Proposition 5, $\nu(\alpha, x)$ is constant on $\mathbb{P}(\mathcal{O}_Y)$. The non-nef locus cannot be empty; otherwise, α would be nef. Nevertheless, the non-nef locus has to be contained in $\mathbb{P}(\mathcal{O}_Y)$. Hence, the constant cannot be zero.

In conclusion, we have $\nu(\alpha, \mathbb{P}(\mathcal{O}_Y)) > 0$, which in particular shows that α is not nef in codimension $k+1$. On the other hand, the non-nef locus is also $\mathbb{P}(\mathcal{O}(Y))$, which in particular shows that α is nef in codimension k.

With the explicit calculation of generic minimal multiplicity, we discuss the optimality of the divisorial Zariski decomposition. Take $k = 1$ in the above construction. Take β to be the first Chern class of some line bundle. Hence, by the above calculation, α is nef in codimension 1 but not nef in codimension 2. Its non-nef locus is $\mathbb{P}(\mathcal{O}_Y)$. For α, there does not exist a unique decomposition of this psef class $\alpha = c_1(L)$ into a nef in codimension 2 $\mathbb{R}$-divisor P and an effective $\mathbb{R}$-divisor N, such that the canonical inclusion $H^0(\lfloor kP \rfloor) \to H^0(kL)$ is an isomorphism for each $k > 0$. Here, the round-down of an $\mathbb{R}$-divisor is defined coefficient-wise. On the contrary, this decomposition will also be the divisorial Zariski decomposition. However, α is nef in codimension 1, and the uniqueness of the divisorial Zariski decomposition shows that the nef in codimension 2 part has to be α itself. This is a contradiction. In particular, when Y is a Riemann surface, it gives an example in dimension 3, where the classical Zariski decomposition does not exist (although it is always possible in dimension 2).

Consider a psef class α on some compact manifold X. In general, there does not always exist a composition of finite blow-up(s) of smooth centres $\mu : \tilde{X} \to X$, such that the nef in the codimension 1 part of $\mu^*\alpha$ is, in fact, nef. This example was first shown in [32].

Let α be a big class on a compact Kähler manifold X. Assume that there exists no finite composition of blow-up(s) with smooth centres, such that the the nef in the codimension 1 part of $\mu^*\alpha$ is, in fact, nef. For example, we can take the pull-back of the class constructed by Nakayama on X. We have the following arguments to conclude that, in fact, there exists no modification such that the the nef in the codimension 1 part of $\mu^*\alpha$ is, in fact, nef. In general, a modification is not necessarily a composition of blow-up(s) with smooth centres. However, by Hironaka's results, any modification is dominated by a finite composition of blow-up(s) with smooth centres. In other words, for $\nu : \tilde{X} \to X$ a modification, there exists a commutative diagram

$$\begin{tikzcd} Y \arrow[r, "f"] \arrow[dr, "g"'] & \tilde{X} \arrow[d, "\nu"] \\ & X \end{tikzcd}$$

where g is a finite composition of blow-up(s) with smooth centres, and f is holomorphic. To prove that there exists no modification, such that the nef in the codimension 1 part of the pull-back of some cohomology class is nef, we have to prove that if $Z(\nu^*\alpha)$ is nef, $Z(g^*\alpha)$ is also nef. This is done by the following proposition. It shows in particular that in the above example, if $Z(\nu^*\alpha)$ is nef, $Z(g^*\alpha) = f^*Z(\nu^*\alpha)$ is also nef.

Notice that the initial argument of Nakayama already proves the non-existence of the Zariski decomposition for any modification.

Proposition 5.3. — (1) *Let $f : Y \to X$ be a holomorphic map between two compact complex manifolds and α be a psef class on X. Assume that $Z(\alpha)$ is nef. Then $f^*N(\alpha) \geq N(f^*\alpha)$, where the inequality relation $\geq$ means that the difference is a psef class.*

(2) *Let $f : Y \to X$ be a modification between two compact, complex manifolds and α a big class on X. Then $N(f^*\alpha) \geq f^*N(\alpha)$.*

Proof. — (1) By the convexity of minimal multiplicity along the subvarieties,

$$N(f^*\alpha) \leq N(f^*N(\alpha)) + N(f^*Z(\alpha)) .$$

Since $Z(\alpha)$ is nef, $f^*Z(\alpha)$ is also nef, and thus $N(f^*Z(\alpha)) = 0$. The conclusion follows observing that $N(f^*N(\alpha)) \leq f^*N(\alpha)$.

(2) We claim that for any positive current $T \in f^*\alpha$, there exists a positive current $S \in \alpha$, such that $T = f^*S$. It is proven in Proposition 1.2.7 [5] in a more general setting. For the convenience of the reader, we give the proof in this special case.

Fix a smooth representative α_∞ in α. There exists a quasi-psh function φ, such that $T = f^*\alpha_\infty + i\partial\overline{\partial}\varphi$. Let U be an open set of X, such that $\alpha_\infty = i\partial\overline{\partial}v$ on U. The function $v \circ f + \varphi$ is psh on $f^{-1}(U)$. All the fibres are compact and connected (the limit of the general connected fibre, the points, is still connected); thus, $v \circ f + \varphi$ is constant along the fibres. Thus, there exists a function ψ on U, such that $\varphi = \psi \circ f$. Since φ is L^1_{loc}, and f is biholomorphic on a dense Zariski open set, ψ is also L^1_{loc}. It is easy to see that ψ is independent of the choice of v and is defined on X. Defining $S = \alpha_\infty + i\partial\overline{\partial}\psi$ we have $T = f^*S$.

In particular, the minimal current in $f^*\alpha$ is the pull-back of the minimal current in α $T_{\min}$. Thus,

$$N(f^*\alpha) = \Big\{ \sum \nu(f^*T_{\min}, E)[E] \Big\} \geq \Big\{ \sum_{\operatorname{codim}(f(E))=1} \nu(f^*T_{\min}, E)[E] \Big\}$$
$$= \Big\{ \sum_{\operatorname{codim}(f(E))=1} \nu(T_{\min}, f(E))[E] \Big\} = f^*N(\alpha),$$

where the sum is taken over all irreducible hypersurfaces of Y. □

Let us point out that a current with minimal singularities does not necessarily have analytic singularities for such a big class α that is nef in codimension 1 but not nef in codimension 2; this was observed by Matsumura [30]. The reason is as follows. In such a situation, there exists a modification $\nu : \tilde{X} \to X$, such that the pull-back of α has a minimal current of the form $\beta + [E]$, where β is a semi-positive smooth form, and $[E]$ is the current associated to an effective divisor supported in the exceptional divisor. In particular, the sum $\{\beta\} + \{[E]\}$ as a cohomology class gives the divisorial Zariski decomposition of the class $\nu^*\alpha$. We remember that for a big class, the Zariski projection of α is given by

$$\alpha - \sum_D \nu(T_{\min}, D)\{[D]\},$$

where D runs over all the irreducible divisors on X, and $T_{\min}$ is the current with minimal singularity in the class α (cf. Proposition 3.6 of [5]). On the other hand, the push-forward ν_* and pull-pack ν^* induce isomorphism between $\nu^*\alpha_\infty$-psh functions on $\tilde{X}$ and α_∞-psh functions on X, where α_∞ is a smooth element in α. In particular, the pull-back of the minimal current of α is the minimal current in $\nu^*\alpha$, which is also a big class. Hence, $\nu^*\alpha$ admits a divisorial Zariski decomposition where the Zariski projection is semi-positive (hence, nef). This contradicts the last paragraph.

Remark 5.4. — As a direct consequence of Matsumura's observation, the assumption in our Kawamata–Viehweg vanishing theorem that the line bundle is numerically equivalent to an effective integral divisor cannot be deduced from the other conditions. In the nef case as in [20], they reduce from the assumption that the line bundle L is nef with $(L^2) \neq 0$ that L is numerically equivalent to an effective integral divisor D and that there exists a positive singular metric h on L, such that $\mathcal{I}(h) = \mathcal{O}(-D)$.

Now we show that for a big line bundle L, which is nef in codimension 1 but not nef in codimension 2, over a compact Kähler manifold (X,ω), $\langle L^2\rangle \neq 0$, and $\frac{i}{2\pi}\Theta(L,h)$ is not a current associated to an effective integral divisor. In particular, Nakayama's example provides such an example.

By the observation of Matsumura, the curvature current of the minimal metric cannot even be a current associated with a real divisor. Since L is big, $\langle L^n\rangle = \mathrm{Vol}(L) \neq 0$. By the Teissier–Hovanskii inequalities, we get

$$\langle L^2 \cdot \omega^{n-2}\rangle = \langle L^2\rangle \cdot \omega^{n-2} \geq \mathrm{Vol}(L)^{2/n}\mathrm{Vol}(\omega)^{(n-2)/n} > 0 .$$

This shows, in particular, that $\langle L^2\rangle \neq 0$.

Remark 5.5. — Let us observe that this kind of construction can also be used to give an example of the manifold with a psef anti-canonical line bundle, for which the Albanese morphism is not surjective.

According to the knowledge of the author, this kind of question was first proposed in [17], where the authors ask whether the Albanese map of a compact Kähler manifold is surjective under the assumption that the anti-canonical line bundle is nef. The statement was first proven by Păun [34], who used the positivity of direct image, and then by Junyan Cao [8] via a different and simpler method. In the case that the manifold is projective, and the anti-canonical divisor is nef, this wad proven earlier by Qi Zhang [38].

Let us use the same notation as above. Take Y to be a complex curve of genus larger than 2. By a classical result, the Albanese map of Y is the embedding of the curve into its Jacobian variety $\mathrm{Jac}(Y)$. In particular, the Albanese map is not surjective. Fix A as an ample divisor on Y. Define $E = A^{\otimes p} \oplus A^{\otimes -q}$, where $p, q \in \mathbb{N}$ will be determined later. Denote $X = \mathbb{P}(E)$ with $\pi : X \to Y$.

We claim that the Albanese morphism of X is the composition of the natural projection π and the Albanese morphism of Y. The reason is the following (cf. Proposition 3.12 in [18]).

Since the fibres of π are $\mathbb{P}^1$, which is connected, and since π is differentially a locally trivial fibre bundle, we have $R^0\pi_*\mathbb{R}_X = \mathbb{R}_Y$, while $R^1\pi_*\mathbb{R}_X = 0$. We remark that $H^1(\mathbb{P}^1,\mathbb{R}) = 0$. The Leray spectral sequence of the constant sheaf $\mathbb{R}_X$ over X satisfies

$$E_2^{s,t} = H^s(Y, R^t\pi_*\mathbb{R}_X), E_r^{s,t} \Rightarrow H^{s+t}(X,\mathbb{R}).$$

Since $R^1\pi_*\mathbb{R}_X = 0$, the Leray spectral sequence always degenerates in E_2. (In fact, by [2], the Leray spectral sequence always degenerates in E_2 for Kähler fibrations.) Hence, we have

$$H^1(X,\mathbb{R}_X) \cong H^1(Y,\mathbb{R}_Y).$$

Since Y and X are compact Kähler, we have by the Hodge decomposition theorem that

$$H^0(X,\Omega_X^1) \cong H^0(Y,\Omega_Y^1).$$

Since $\pi^* : H^0(Y, \Omega_Y^1) \cong H^0(X, \Omega_X^1)$ is an injective morphism, it induces an isomorphism. Passing to the quotient it induces an isomorphism $\pi^* : \mathrm{Alb}(X) \cong \mathrm{Alb}(Y)$. The claim is proven by the universality of the Albanese morphism:

$$\begin{tikzcd} X \arrow[r, "\pi"] \arrow[d, "\alpha_X"] & Y \arrow[d, "\alpha_Y"] \\ \mathrm{Alb}(X) \arrow[r, "\pi^*"] & \mathrm{Alb}(Y). \end{tikzcd}$$

We also claim that for well chosen p, q, the anti-canonical line bundle $-K_X$ is big but not nef in codimension 1. In particular, this shows that there exists a compact Kähler manifold X, such that $-K_X$ is psef, but the Albanese morphism is not surjective. Recall that

$$K_X = \pi^*(K_Y \otimes \det E) \otimes \mathcal{O}_X(-2).$$

In particular, for $q \gg p$, $-(K_Y \otimes \det E) = (q-p)A - K_Y$ is ample. On the other hand, $\mathcal{O}_X(1)$ is big since one of the components in the direct sum bundle E is big. Thus, $-K_X$ is big for $q \gg p$. On the other hand, the surjective morphism $E \to A^{\otimes p}$ induces the closed immersion $\mathbb{P}(A^{\otimes p}) \cong Y \to X$. We have that $-K_X|_{\mathbb{P}(A^{\otimes p})} = -K_Y - pA$. For p big enough, we can assume that $-K_Y - pA$ is not psef. As a consequence, $-K_X$ is not nef in codimension 1.

In fact, we can calculate the generic minimal multiplicity as

$$\nu(c_1(-K_X), \mathbb{P}(A^{\otimes p})) = \min\{t, -K_Y + (q-p)A + tpA - (2-t)qA \text{ is nef}\}.$$

Since K_Y is ample, we know that the generic minimal multiplicity along $\mathbb{P}(A^{\otimes p})$ is strictly larger than 1. In particular, consider any singular metric h_ε on $-K_X$, such that its curvature satisfies $i\Theta(-K_X, h_\varepsilon) \geq -\varepsilon\omega$, where ω is some Kähler form on X. Then the multiplier ideal sheaf is not trivial. Near a point of $\mathbb{P}(A^{\otimes p})$, choose some local coordinate, such that $\mathbb{P}(A^{\otimes p}) = \{z_1 = 0\}$. By Siu's decomposition, the local weight of h_ε is more singular than $\log(|z_1|^2)$. This implies that $\mathcal{I}(h_\varepsilon) \subset \mathcal{I}_{\mathbb{P}(A^{\otimes p})}$, where $\mathcal{I}_{\mathbb{P}(A^{\otimes p})}$ is the ideal sheaf associated to $\mathbb{P}(A^{\otimes p})$.

Therefore, some additional condition is certainly needed to ensure surjectivity of the Albanese morphism.

Acknowledgement. — The author would like to thank Jean-Pierre Demailly, his PhD supervisor, for guidance, patience and generosity. He would also like to thank Junyan Cao, Henri Guenancia and Andreas Höring for some very useful suggestions on the previous draft of this work. He would also like to express his gratitude to colleagues at the Institut Fourier for all the interesting discussions. Thanks also go to the anonymous reviewer for the very careful reading of this paper and for insightful comments and suggestions.

BIBLIOGRAPHY

[1] C. Birkenhake & H. Lange – *Complex abelian varieties*, Grundlehren der mathematischen Wissenschaften, Springer-Verlag, Berlin and Heidelberg, 1992.

[2] A. Blanchard – "Sur les variétés analytiques complexes", *Ann. Sci. Ec. Norm. Sup.* **73** (1956), p. 151–202.

[3] S. Boucksom – "Cônes positifs des variétés complexes compactes", Ph.D. thesis, 2002.

[4] ———, "On the volume of a line bundle", *Internat. J. Math.* **13** (2002), no. 10, p. 1043–1063.

[5] ———, "Divisorial zariski decompositions on compact complex manifolds", *Ann. Sci. ENS (4)* **37** (2004), no. 1, p. 45–76.

[6] S. Boucksom, J.-P. Demailly, M. Păun & T. Peternell – "The pseudoeffective cone of a compact Kähler manifold and varieties of negative Kodaira dimension", *J. Algebraic Geom.* **22** (2013), no. 2, p. 201–248, arXiv:math/0405285.

[7] S. Boucksom, P. Eyssidieux, V. Guedj & A. Zeriahi – "Monge–Ampère equations in big cohomology classes", *Acta Math.* **205** (2010), no. 2, p. 199–262.

[8] J. Cao – "A remark on compact Kähler manifolds with nef anticanonical bundles and applicaions", arXiv:1305.4397.

[9] ———, "Numerical dimension and a kawamata-viehweg-nadel type vanishing theorem on compact Kähler manifolds", *Compos. Math.* **150** (2014), no. 11, p. 1869–1902.

[10] T. C. Collins & V. Tosatti – "An extension theorem for Kähler currents with analytic singularities", *Ann. Fac. Sci. Toulouse Math.* **23** (2014), no. 4, p. 893–905.

[11] ———, "Kähler currents and null loci", *Invent. Math.* **202** (2015), no. 3, p. 1167–1198.

[12] J.-P. Demailly – "On the cohomology of pseudoeffective line bundles", written exposition of a talk given at the Abel Symposium, Trondheim, July 2013; manuscript Institut Fourier, January 9, 2014.

[13] ———, "Estimations L^2 pour l'opérateur $\overline{\partial}$ d'un fibré vectoriel holomorphe semi-positif au dessus d'une variété kählerienne complète", *Ann. Sci. Ecole Norm. Sup.* **15** (1982), p. 457–511.

[14] ———, *Analytic methods in algebraic geometry*, International Press, 2012.

[15] ———, *Complex analytic and differential geometry*, 2012, `https://www-fourier.ujf-grenoble.fr/~demailly/manuscripts/agbook.pdf`.

[16] J.-P. Demailly & T. Peternell – *A Kawamata–Viehweg vanishing theorem on compact Kähler manifolds*, Surveys in Differential Geometry, no. VIII, Int. Press, Somerville, MA, 2003.

[17] J.-P. Demailly, T. Peternell & M. Schneider – "Kähler manifolds with numerically effective Ricci class", *Compositio Math.* **89** (1993), p. 217–240.
[18] ______, "Compact complex manifolds with numerically effective tangent bundles", *J. Algebraic Geom.* **3** (1994), no. 2, p. 295–345.
[19] ______, "Pseudoeffective line bundles on compact Kähler manifolds", *International Journal of Math* **6** (2001), p. 689–741, arXiv:math.AG/0006205.
[20] J.-P. Demailly & M. Păun – "Numerical characterization of the Kähler cone of a compact Kähler manifold", *Annals of Math.* **159** (2004), p. 1247–1274, arXiv:math.AG/0105176.
[21] Q. Guan & X. Zhou – "A solution of an L^2 extension problem with an optimal estimate and applications", *Annals of Mathematics* **181** (2015), no. 3, p. 1139–1208.
[22] Q. Guan & X. Zhou – "Strong openness conjecture and related problems for plurisubharmonic functions", 2014, arXiv:math.CV/1401.7158.
[23] ______, "Effectiveness of Demailly's strong openness conjecture and related problems", *Inventiones mathematicae* **202** (2015), p. 635–676, arXiv:math.CV/1403.7247.
[24] ______, "A proof of Demailly's strong openness conjecture", *Annals of Mathematics* **182** (2015), p. 605–616, arXiv:math.CV/1311.3781.
[25] V. Guedj & A. Zeriahi – *Degenerate complex Monge–Ampère equations*, EMS Tracts in Mathematics, no. 26, 2017, ISBN 978-3-03719-167-5.
[26] H. Hironaka – "Resolution of singularities of an algebraic variety over a field of characteristic zero. I, II", *Ann. of Math. (2)* **79** (1964), p. 109–326.
[27] D. Huybrechts – "The Kähler cone of a compact hyperkähler manifold", *Math. Ann.* **326** (2003), p. 499, https://doi.org/10.1007/s00208-003-0433.
[28] A. Küronya – "Positivity on subvarieties and vanishing of higher cohomology", *Annales de l'Institut Fourier* **63** (2013), no. 5, p. 1717–1737.
[29] R. Lazarsfeld – *Positivity in algebraic geometry – volume 1, classical setting: Line bundles and linear series*, Springer-Verlag, Berlin Heidelberg, 2004, `https://doi.org/10.1007/978-3-642-18808-4`, ISBN 978-3-540-22528-7.
[30] S. ichi Matsumura – "A Nadel vanishing theorem for metrics with minimal singularities on big line bundles", *Advances in Mathematics* **280** (2015), p. 188–207.
[31] C. Mourougane – "Versions kählériennes du théorème d'annulation de Bogomolov", *Collectanea Mathematica* **49** (1998), no. 2–3, p. 433–445.
[32] N. Nakayama – *Zariski decomposition and abundance*, MSJ Memoirs, no. 14, The Mathematical Society of Japan, Tokyo, Japan, 2004.
[33] M. Păun – "Sur l'effectivité numérique des images inverses de fibrés en droites", *Math. Ann.* **310** (1998), p. 411–421.

[34] ———, “Relative adjoint transcendental classes and Albanese maps of compact Kähler manifolds with nef Ricci curvature”, in *Higher Dimensional Algebraic Geometry: In honour of Professor Yujiro Kawamata’s sixtieth birthday* (K. Oguiso, C. Birkar, S. Ishii & S. Takayama, eds.), Adv. Stud. Pure Math., Mathematical Society of Japan, Tokyo, 2017, p. 335–356.

[35] Y.-T. Siu – “Analyticity of sets associated to Lelong numbers and the extension of closed positive currents”, *Invent. Math.* **27** (1974), p. 53–156.

[36] C. Voisin – *Hodge theory and complex algebraic geometry I*, Cambridge University Press, 2002.

[37] S. T. Yau – “On the Ricci curvature of a compact Kähler manifold and the complex Monge–Ampère equation. I”, *Comm. Pure Appl. Math.* **31** (1978), no. 3, p. 339–411.

[38] Q. Zhang – “On projective manifolds with nef anticanonical bundles”, *J. Reine Angew.Math.* **478** (1996), p. 57–60.