INTERACTIONS BETWEEN COMPRESSED SENSING RANDOM MATRICES AND HIGH DIMENSIONAL GEOMETRY

D. Chafaï, O. Guédon
G. Lecué, A. Pajor

Panoramas et Synthèses

Numéro 37

SOCIÉTÉ MATHÉMATIQUE DE FRANCE
Publié avec le concours du Centre national de la recherche scientifique

PANORAMAS ET SYNTHÈSES 37

INTERACTIONS BETWEEN COMPRESSED SENSING RANDOM MATRICES AND HIGH DIMENSIONAL GEOMETRY

Djalil Chafaï
Olivier Guédon
Guillaume Lecué
Alain Pajor

Société Mathématique de France 2012
Publié avec le concours du Centre National de la Recherche Scientifique

Djalil CHAFAÏ

Laboratoire d'Analyse et de Mathématiques Appliquées, UMR CNRS 8050, 5 Bd Descartes, Champs-sur-Marne, F-77454 Marne-la-Vallée, Cedex 2, France.

E-mail : `djalil.chafai(at)univ-mlv.fr`

Url : `http://perso-math.univ-mlv.fr/users/chafai.djalil/`

Olivier GUÉDON

Laboratoire d'Analyse et de Mathématiques Appliquées, UMR CNRS 8050, 5 Bd Descartes, Champs-sur-Marne, F-77454 Marne-la-Vallée, Cedex 2, France.

E-mail : `olivier.guedon(at)univ-mlv.fr`

Url : `http://perso-math.univ-mlv.fr/users/guedon.olivier/`

Guillaume LECUÉ

Laboratoire d'Analyse et de Mathématiques Appliquées, UMR CNRS 8050, 5 Bd Descartes, Champs-sur-Marne, F-77454 Marne-la-Vallée, Cedex 2, France.

E-mail : `guillaume.lecue(at)univ-mlv.fr`

Url : `http://perso-math.univ-mlv.fr/users/lecue.guillaume/`

Alain PAJOR

Laboratoire d'Analyse et de Mathématiques Appliquées, UMR CNRS 8050, 5 Bd Descartes, Champs-sur-Marne, F-77454 Marne-la-Vallée, Cedex 2, France.

E-mail : `Alain.Pajor(at)univ-mlv.fr`

Url : `http://perso-math.univ-mlv.fr/users/pajor.alain/`

2000 ***Mathematics Subject Classification.*** **—** 15B52; 15A18 ; 41A46; 47B06; 46B07; 46B09; 42A16; 42A61; 52A23; 60B20; 60F05.

Key words and phrases. **—** Phenomena in high dimension; asymptotic geometric analysis; empirical methods; concentration of measure; compressed sensing; Gelfand widths; chaining methods; singular values; random matrices; selection of characters.

INTERACTIONS BETWEEN COMPRESSED SENSING RANDOM MATRICES AND HIGH DIMENSIONAL GEOMETRY

Djalil Chafaï, Olivier Guédon, Guillaume Lecué,
Alain Pajor

Abstract. — This book is based on a series of post-doctoral level lectures given at Université Paris-Est Marne-la-Vallée in November 2009, by Djalil Chafaï, Olivier Guédon, Guillaume Lecué, Shahar Mendelson, and Alain Pajor. It aims to bridge several actively developed domains of research around high dimensional phenomena and asymptotic geometric analysis. The covered topics include empirical methods and high dimensional geometry, concentration of measure, compressed sensing, Gelfand widths, chaining methods, singular values, Wishart matrices, and problems of selection of characters. This book focuses on methods and concepts. Chapters are mostly self-contained. An index is provided.

Résumé **(Interactions entre échantillonnage comprimé, matrices aléatoires, et géométrie de grande dimension)**

Ce livre est basé sur une série de cours de niveau post-doctoral donnés à l'Université Paris-Est Marne-la-Vallée en novembre 2009, par Djalil Chafaï , Olivier Guédon, Guillaume Lecué, Shahar Mendelson et Alain Pajor. Ce livre tente de faire le lien entre plusieurs domaines de recherche activement développés autour des phénomènes en grandes dimensions et de l'analyse géométrique asymptotique. Les thèmes abordés comprennent les méthodes empiriques en géométrie de grande dimension, la concentration de la mesure, l'échantillonnage comprimé, les épaisseurs de Gelfand, les méthodes de chaînage, les valeurs singulières, les matrices de Wishart et les problèmes de sélection de caractères. Ce livre met l'accent sur lesméthodes et les concepts. Les chapitres sont autonomes. Un index est fourni.

CONTENTS

INTRODUCTION

Compressed sensing, also referred to in the literature as compressive sensing or compressive sampling, is a framework that enables one to recover approximate or exact reconstruction of sparse signals from incomplete measurements. The existence of efficient algorithms for this reconstruction, such as the ℓ_1-minimization algorithm, and the potential for applications in signal processing and imaging, led to a rapid and extensive development of the theory after the seminal articles by D. Donoho [**26**], E. Candes, J. Romberg and T. Tao [**15**] and E. Candes and T. Tao [**17**].

The principles underlying the discoveries of these phenomena in high dimensions are related to more general problems and their solutions in Approximation Theory. One significant example of such a relation is the study of Gelfand and Kolmogorov widths of classical Banach spaces. There is already a huge literature on both the theoretical and numerical aspects of compressed sensing. Our aim is not to survey the state of the art in this rapidly developing field, but to highlight and study its interactions with other fields of mathematics, in particular with asymptotic geometric analysis, random matrices and empirical processes.

To introduce the subject, let T be a subset of $\mathbb{R}^N$ and let A be an $n \times N$ real matrix with rows $Y_1, \ldots, Y_n \in \mathbb{R}^N$. Consider the general problem of reconstructing a vector $x \in T$ from the *data* $Ax \in \mathbb{R}^n$: that is, from the known *measurements*

$$\langle Y_1, x\rangle, \ldots, \langle Y_n, x\rangle$$

of an unknown x. Classical linear algebra suggests that the number n of measurements should be at least as large as the dimension N in order to ensure reconstruction. Compressed sensing provides a way of reconstructing the original signal x from its compression Ax that uses only a small number of linear measurements: that is with $n \ll N$. Clearly one needs some a priori hypothesis on the subset T of signals that we want to reconstruct, and of course the matrix A should be suitably chosen in order to allow the reconstruction of every vector of T.

The first point concerns the subset T and is a matter of *complexity.* Many tools within this framework were developed in Approximation Theory and in the Geometry of Banach Spaces. One of our goals is to present these tools.

The second point concerns the design of the measurement matrix A. To date the only good matrices are random *sampling* matrices and the key is to sample $Y_1, \dots, Y_n \in \mathbb{R}^N$ in a suitable way. For this reason probability theory plays a central role in our exposition. These random sampling matrices will usually be of Gaussian or Bernoulli (± 1) type or be random sub-matrices of the discrete Fourier $N \times N$ matrix (partial Fourier matrices). There is a huge technical difference between the study of unstructured compressive matrices (with i.i.d entries) and structured matrices such as partial Fourier matrices. Another goal of this work is to describe the main tools from probability theory that are needed within this framework. These tools range from classical probabilistic inequalities and concentration of measure to the study of empirical processes and random matrix theory.

The purpose of Chapter 1 is to present some basic tools and preliminary background. We will look briefly at elementary properties of Orlicz spaces in relation to tail inequalities for random variables. An important connection between high dimensional geometry and the study of empirical processes comes from the behavior of the sum of independent centered random variables with sub-exponential tails. An important step in the study of empirical processes is discretization: in which we replace an infinite space by an approximating net. It is essential to estimate the size of the discrete net and such estimates depend upon the study of covering numbers. Several upper estimates for covering numbers, such as Sudakov's inequality, are presented in the last part of Chapter 1.

Chapter 2 is devoted to compressed sensing. The purpose is to provide some of the key mathematical insights underlying this new sampling method. We present first the exact reconstruction problem informally introduced above. The *a priori* hypothesis on the subset of signals T that we investigate is *sparsity.* A vector in $\mathbb{R}^N$ is said to be m-sparse ($m \leqslant N$) if it has at most m non-zero coordinates. An important feature of this subset is its peculiar structure: its intersection with the Euclidean unit sphere is a union of unit spheres supported on m-dimensional coordinate subspaces. This set is highly compact when the degree of compactness is measured in terms of covering numbers. As long as $m \ll N$ the sparse vectors form a *very small* subset of the sphere.

A fundamental feature of compressive sensing is that practical reconstruction can be performed by using efficient algorithms such as the ℓ_1-minimization method which consists, for given data $y = Ax$, to solve the "linear program"

$$\min_{t\in\mathbb{R}^N} \sum_{i=1}^{N} |t_i| \quad \text{subject to} \quad At = y.$$

At this step, the problem becomes that of finding matrices for which the algorithm reconstructs any m-sparse vector with m relatively large. A study of the cone of constraints that ensures that every m-sparse vector can be reconstructed by the ℓ_1-minimization method leads to a necessary and sufficient condition known as the *null space property* of order m:

$$\forall h \in \ker A,\ h \neq 0,\ \forall I \subset [N],\ |I| \leq m, \quad \sum_{i\in I} |h_i| < \sum_{i\in I^c} |h_i|.$$

This property has a nice geometric interpretation in terms of the structure of faces of polytopes called *neighborliness*. Indeed, if P is the polytope obtained by taking the symmetric convex hull of the columns of A, the *null space property* of order m for A is equivalent to the *neighborliness* property of order m for P: that the matrix A which maps the vertices of the cross-polytope

$$B_1^N = \Big\{t \in \mathbb{R}^N \ : \ \sum_{i=1}^{N} |t_i| \leq 1\Big\}$$

onto the vertices of P preserves the structure of k-dimensional faces up to the dimension $k = m$. This remarkable connection between compressed sensing and high dimensional geometry is due to D. Donoho [**25**].

Unfortunately, the null space property is not easy to verify nor is the neighborliness. An ingenious sufficient condition is the so-called *Restricted Isometry Property* (RIP) of order m that requires that all sub-matrices of size $n \times m$ of the matrix A are uniformly well-conditioned. More precisely, we say that A satisfies the RIP of order $p \leqslant N$ with parameter $\delta = \delta_p$ if the inequalities

$$1 - \delta_p \leqslant |Ax|_2^2 \leqslant 1 + \delta_p$$

hold for all p-sparse unit vectors $x \in \mathbb{R}^N$. An important feature of this concept is that if A satisfies the RIP of order $2m$ with a parameter δ small enough, then every m-sparse vector can be reconstructed by the ℓ_1-minimization method. Even if this RIP condition is difficult to check on a given matrix, it actually holds true with high probability for certain models of random matrices and can be easily checked for some of them.

Here probabilistic methods come into play. Among good unstructured sampling matrices we shall study the case of Gaussian and Bernoulli random matrices. The case of partial Fourier matrices, which is more delicate, will be studied in Chapter 5. Checking the RIP for the first two models may be treated with a simple scheme: the ε-net argument presented in Chapter 2.

Another way to tackle the problem of reconstruction by ℓ_1-minimization is to analyse the Euclidean diameter of the intersection of the cross-polytope B_1^N with the kernel of A. This study leads to the notion of Gelfand widths, particularly for the cross-polytope B_1^N. Its Gelfand widths are defined by the numbers

$$d^n(B_1^N, \ell_2^N) = \inf_{\operatorname{codim} S \leqslant n} \operatorname{rad}(S \cap B_1^N), \quad n = 1, \dots, N,$$

where $\operatorname{rad}(S \cap B_1^N) = \max\{|x|_2 : x \in S \cap B_1^N\}$ denotes the half Euclidean diameter of the section of B_1^N and the infimum is over all subspaces S of $\mathbb{R}^N$ of dimension less than or equal to n.

A great deal of work was done in this direction in the seventies. These Approximation Theory and Asymptotic Geometric Analysis standpoints shed light on a new aspect of the problem and are based on a celebrated result of B. Kashin [**65**] stating that

$$d^n(B_1^N, \ell_2^N) \leqslant \frac{C}{\sqrt{n}} \log^{O(1)}\left(\frac{N}{n}\right)$$

for some numerical constant C. The relevance of this result to compressed sensing is highlighted by the following fact.

> *Let $1 \leq m \leq n$, if $\operatorname{rad}(\ker A \cap B_1^N) < \frac{1}{2\sqrt{m}}$ then every m-sparse vector can be reconstructed by ℓ_1-minimization.*

From this perspective, the goal is to estimate the diameter $\operatorname{rad}(\ker A \cap B_1^N)$ from above. We discussed this in detail for several models of random matrices. The connection with the RIP is clarified by the following result.

> *Assume that A satisfies the RIP of order p with parameter δ. Then*
>
> $$\operatorname{rad}(\ker A \cap B_1^N) \leq \frac{C}{\sqrt{p}} \cdot \frac{1}{1-\delta}$$
>
> *where C is a numerical constant and so $\operatorname{rad}(\ker A \cap B_1^N) < \frac{1}{2\sqrt{m}}$ is satisfied with $m = O(p)$.*

The ℓ_1-minimization method extends to the study of approximate reconstruction of vectors which are not too far from being sparse. Let $x \in \mathbb{R}^N$ and let $x^\sharp$ be a minimizer of

$$\min_{t \in \mathbb{R}^N} \sum_{i=1}^N |t_i| \quad \text{subject to } At = Ax.$$

Again the notion of width is very useful. We prove the following:

Assume that $\mathrm{rad}\,(\ker A \cap B_1^N) < \frac{1}{4\sqrt{m}}$. *Then for any* $I \subset [N]$ *such that* $|I| \leqslant m$ *and any* $x \in \mathbb{R}^N$, *we have*

$$|x - x^\sharp|_2 \leq \frac{1}{\sqrt{m}} \sum_{i \notin I} |x_i|.$$

This applies in particular to unit vectors of the space $\ell_{p,\infty}^N$, $0 < p < 1$ for which $\min_{|I| \leqslant m} \sum_{i \notin I} |x_i| = O(m^{1-1/p})$.

In the last section of Chapter 2 we introduce a measure $\ell_*(T)$ of complexity of a subset $T \subset \mathbb{R}^N$ defined by

$$\ell_*(T) = \mathbb{E} \sup_{t \in T} \sum_{i=1}^N g_i t_i,$$

where $g_1, \ldots, g_N$ are independent $\mathcal{N}(0,1)$ Gaussian random variables. This kind of parameter plays an important role in the theory of empirical processes and in the geometry of Banach spaces (see [**87**], [**96**] and [**119**]). It allows to control the size of $\mathrm{rad}\,(\ker A \cap T)$ which as we have seen is a crucial issue in approximate reconstruction.

This line of investigation goes deeper in Chapter 3 where we first present classical results from the theory of Gaussian processes. To make the link with compressed sensing, observe that if A is a $n \times N$ matrix with row vectors $Y_1, \ldots, Y_n$, then the RIP of order p with parameter δ_p can be rewritten in terms of an empirical process property since

$$\delta_p = \sup_{x \in S_2(\Sigma_p)} \left| \frac{1}{n} \sum_{i=1}^n \langle Y_i, x \rangle^2 - 1 \right|$$

where $S_2(\Sigma_p)$ is the set of norm one p-sparse vectors of $\mathbb{R}^N$. While Chapter 2 makes use of a simple ε-net argument to study such processes, we present in Chapter 3 the chaining and generic chaining techniques based on measures of metric complexity such as the γ_2 functional. The γ_2 functional is equivalent to the parameter $\ell_*(T)$ in consequence of the majorizing measure theorem of M. Talagrand [**119**]. This technique enables to provide a criterion that implies the RIP for unstructured models of random matrices, which include the Bernoulli and Gaussian models.

It is worth noticing that the ε-net argument, the chaining argument and the generic chaining argument all share two ideas: the classical trade-off between complexity and concentration on the one hand and an approximation principle on the other. For instance, consider a Gaussian matrix

$$A = \frac{1}{\sqrt{n}} (g_{ij})_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant N}$$

where the g_{ij}'s are i.i.d. standard Gaussian variables. Let T be a subset of the unit sphere S^{N-1} of $\mathbb{R}^N$. A classical problem is to understand how A acts on T. In particular, does A preserve the Euclidean norm on T? In the Compressed Sensing setup, the "input" dimension N is much larger than the number of measurements n, because A is used as a compression matrix. So clearly A cannot preserve the Euclidean norm on the whole sphere S^{N-1}. Hence, it is natural to identify the subsets T of S^{N-1} for which A acts on T in a norm preserving way. Let's start with a single point $x \in T$. Then for any $\varepsilon \in (0,1)$, with probability greater than $1 - 2\exp(-c_0 n\varepsilon^2)$, one has

$$1 - \varepsilon \leqslant |Ax|_2^2 \leqslant 1 + \varepsilon.$$

This result is the one expected since $\mathbb{E}|Ax|_2^2 = |x|_2^2$ (we say that the standard Gaussian measure is isotropic) and the Gaussian measure on $\mathbb{R}^N$ has strong concentration properties. Thus proving that A acts in a norm preserving way on a single vector is only a matter of isotropicity and concentration. Now we want to see how many points in T may share this property simultaneously. This is where the trade-off between complexity and concentration is at stake. A simple union bound argument tells us that if $\Lambda \subset T$ has a cardinality less than $\exp(\frac{1}{2}c_0 n\varepsilon^2)$, then, with probability greater than $1 - 2\exp(-\frac{1}{2}c_0 n\varepsilon^2)$, one has

$$\forall x \in \Lambda, \quad 1 - \varepsilon \leqslant |Ax|_2^2 \leqslant 1 + \varepsilon.$$

This means that A preserves the norm of all the vectors of Λ at the same time, as long as $|\Lambda| \leqslant \exp(\frac{1}{2}c_0 n\varepsilon^2)$. If the entries in A had different concentration properties, we would have ended up with a different cardinality for $|\Lambda|$. As a consequence, it is possible to control the norm of the images by A of $\exp(\frac{1}{2}c_0 n\varepsilon^2)$ points in T simultaneously. The first way of choosing Λ that may come to mind is to use an ε-net of T with respect to ℓ_2^N and then to ask if the norm preserving property of A on Λ extends to T? Indeed, if $m \leq C(\varepsilon) n \log^{-1}\left(N/n\right)$, there exists an ε-net Λ of size $\exp(\frac{1}{2}c_0 n\varepsilon^2)$ in $S_2(\Sigma_m)$ for the Euclidean metric. And, by what is now called the ε-net argument, we can describe all the points in $S_2(\Sigma_m)$ using only the points in Λ:

$$\Lambda \subset S_2(\Sigma_m) \subset (1-\varepsilon)^{-1}\mathrm{conv}(\Lambda).$$

This allows to extend the norm preserving property of A on Λ to the entire set $S_2(\Sigma_m)$ and was the scheme used in Chapter 2.

But this scheme does not apply to several important sets T in S^{N-1}. That is why we present the chaining and generic chaining methods in Chapter 3. Unlike the ε-net argument which demanded only to know how A acts on a single ε-net of T, these two methods require to study the action of A on a sequence (T_s) of subsets of T with exponentially increasing cardinality. In the case of the chaining argument, T_s can be chosen as an ε_s-net of T where ε_s is chosen so that $|T_s| = 2^s$ and for the generic chaining argument, the choice of (T_s) is recursive: for large values of s, the set T_s is a maximal separated set in T of cardinality 2^{2^s} and for small values of s, the construction of T_s depends on the sequence $(T_r)_{r\geqslant s+1}$. For these methods, the approximation argument follows from the fact that $d_{\ell_2^N}(t, T_s)$ tends to zero when s tends to infinity for any $t \in T$ and the trade-off between complexity and concentration is used at every stage s of the approximation of T by T_s. The metric complexity parameter coming from the chaining method is called the Dudley entropy integral

$$\int_0^\infty \sqrt{\log N(T, d, \varepsilon)}\,\mathrm{d}\varepsilon$$

while the one given by the generic chaining mechanism is the γ_2 functional

$$\gamma_2(T, \ell_2^N) = \inf_{(T_s)_s} \sup_{t\in T} \sum_{s=0}^{\infty} 2^{\frac{1}{2}s} d_{\ell_2^N}(t, T_s)$$

where the infimum is taken over all sequences (T_s) of subsets of T such that $|T_0| \leqslant 1$ and $|T_s| \leqslant 2^{2^s}$ for every $s \geqslant 1$. In Chapter 3, we prove that A acts in a norm preserving way on T with probability exponentially in n close to 1 as long as

$$\gamma_2(T, \ell_2^N) = O(\sqrt{n}\,).$$

In the case $T = S_2(\Sigma_m)$ treated in Compressed Sensing, this condition implies that $m = O\big(n \log^{-1}\big(N/n\big)\big)$ which is the same as the condition obtained using the ε-net argument in Chapter 2. So, as far as norm preserving properties of random operators are concerned, the results of Chapter 3 generalize those of Chapter 2. Nevertheless, the norm preserving property of A on a set T implies an exact reconstruction property of A of all m-sparse vectors by the ℓ_1-minimization method only when $T = S_2(\Sigma_m)$. In this case, the norm preserving property is the RIP of order m.

On the other hand, the RIP constitutes a control on the largest and smallest singular values of all sub-matrices of a certain size. Understanding the singular values of matrices is precisely the subject of Chapter 4. An $m \times n$ matrix A with $m \leqslant n$ maps the unit sphere to an ellipsoid, and the half lengths of the principle axes of this ellipsoid are precisely the singular values $s_1(A) \geqslant \cdots \geqslant s_m(A)$ of A. In particular,

$$s_1(A) = \max_{|x|_2=1} |Ax|_2 = \|A\|_{2\to 2} \quad \text{and} \quad s_n(A) = \min_{|x|_2=1} |Ax|_2.$$

Geometrically, A is seen as a correspondence-dilation between two orthonormal bases. In matrix form $UAV^* = \mathrm{diag}(s_1(A), \ldots, s_m(A))$ for a pair of unitary matrices U and V of respective sizes $m \times m$ and $n \times n$. This *singular value decomposition* – SVD for short – has tremendous importance in numerical analysis. One can read off from the singular values the rank and the norm of the inverse of the matrix: the singular values are the eigenvalues of the

Hermitian matrix $\sqrt{AA^*}$: and the largest and smallest singular values appear in the definition of the condition number s_1/s_m which allows to control the behavior of linear systems under perturbations of small norm.

The first part of Chapter 4 is a compendium of results on the singular values of deterministic matrices, including the most useful perturbation inequalities. The Gram-Schmidt algorithm applied to the rows and the columns of A allows to construct a bidiagonal matrix which is unitarily equivalent to A. This structural fact is at the heart of most numerical algorithms for the actual computation of singular values.

The second part of Chapter 4 deals with random matrices with i.i.d. entries and their singular values. The aim is to offer a cultural tour in this vast and growing subject. The tour begins with Gaussian random matrices with i.i.d. entries forming the Ginibre Ensemble. The probability density of this Ensemble is proportional to $G \mapsto \exp(-\mathrm{Tr}(GG^*))$. The matrix $W = GG^*$ follows a Wishart law, a sort of multivariate χ^2. The unitary bidiagonalization allows to compute the density of the singular values of these Gaussian random matrices, which turns out to be proportional to a function of the form

$$s \longmapsto \prod_k s_k^\alpha \mathrm{e}^{-s_k^2} \prod_{i \neq j} |s_i^2 - s_j^2|^\beta .$$

The change of variable $s_k \mapsto s_k^2$ reveals Laguerre weights in front of the Vandermonde determinant, the starting point of a story involving orthogonal polynomials. As for most random matrix ensembles, the determinant measures a logarithmic repulsion between eigenvalues. Here it comes from the Jacobian of the SVD. Such Gaussian models can be analysed with explicit but cumbersome computations. Many large dimensional aspects of random matrices depend only on the first two moments of the entries, and this makes the Gaussian case universal. The most well known universal asymptotic result is indubitably the Marchenko-Pastur theorem. More precisely if M is an $m \times n$ random matrix with i.i.d. entries of variance $n^{-\frac{1}{2}}$, the empirical counting probability measure of the singular values of M

$$\frac{1}{m} \sum_{k=1}^{m} \delta_{s_k(M)}$$

tends weakly, when $n, m \to \infty$ with $m/n \to \rho \in (0, 1]$, to the Marchenko-Pastur law

$$\frac{1}{\rho \pi x} \sqrt{((x+1)^2 - \rho)(\rho - (x-1)^2)}\, \mathbf{1}_{[1-\sqrt{\rho}, 1+\sqrt{\rho}]}(x) \mathrm{d}x .$$

We provide a proof of the Marchenko-Pastur theorem by using the methods of moments. When the entries of M have zero mean and finite fourth moment, Bai-Yin theorem furnishes the convergence at the edge of the support, in the