HAL open science

Compressible Distributions for High-dimensional Statistics

Rémi Gribonval, Volkan Cevher, Mike E. Davies

To cite this version:

Rémi Gribonval, Volkan Cevher, Mike E. Davies. Compressible Distributions for High-dimensional Statistics. 2011. inria-00563207v1

HAL Id: inria-00563207

https://inria.hal.science/inria-00563207v1

Preprint submitted on 4 Feb 2011 (v1), last revised 25 Apr 2012 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Submitted to the Annals of Statistics

COMPRESSIBLE PRIORS FOR HIGH-DIMENSIONAL STATISTICS

By Rémi Gribonval* and Volkan Cevher† and Mike E. Davies‡

*INRIA, Centre Inria Rennes - Bretagne Atlantique
†Laboratory for Information and Inference Systems,
IDIAP Research Institute,
Ecole Polytechnique Federale de Lausanne
‡Institute for Digital Communication,
Edinburgh University

We develop a principled way of identifying probability distributions whose independent and identically distributed (iid) realizations are compressible, i.e., can be approximated as sparse. We focus on the context of Gaussian random underdetermined linear regression (GULR) problems, where compressibility is known to ensure the success of estimators exploiting sparse regularization. We prove that many priors revolving around maximum a posteriori (MAP) interpretation of the ℓ^1 sparse regularization estimator and its variants are in fact incompressible, in the limit of large problem sizes. To show this, we identify non-trivial undersampling regions in GULR where the simple least squares solution almost surely outperforms an oracle sparse solution, when the data is generated from a prior such as the Laplace distribution. We provide rules of thumb to characterize large families of compressible (respectively incompressible) priors based on their second and fourth moments. Generalized Gaussians and generalized Pareto distributions serve as running examples for concreteness.

1. Introduction. High-dimensional data is shaping the current *modus operandi* of statistics. Surprisingly, while the ambient dimension is large in many problems, natural

*Rémi Gribonval is a member of the METISS project-team at IRISA, Rennes, France. This work was supported in part by the European Commission through the project SMALL (Sparse Models, Algorithms and Learning for Large-Scale data) under FET-Open, grant number 225913.

†This work was supported in part by part by the European Commission under Grant MIRG-268398 and DARPA KeCoM program #11-DARPA-1055. VC also would like to acknowledge Rice University for his Faculty Fellowship.

‡MED acknowledges support of his position from the Scottish Funding Council and their support of the Joint Research Institute in Signal and Image Processing with the Heriot-Watt University as a component of the Edinburgh Research Partnership. This work was supported in part by the European Commission through the SMALL project under FET-Open, grant number 225913 and by the UK Engineering and Physical Science Research Council, grant EP/F039697/1.

AMS 2000 subject classifications: Primary 62J07; secondary 62G30,62G20,62F12

Keywords and phrases: linear inverse problem; LASSO; sparsity; sparse regression; ridge regression; compressible prior; compressive sensing; instance optimality; maximum a posteriori; high-dimensional statistics; order statistics

constraints and parameterizations often cause data to cluster along low-dimensional structures. Identifying and exploiting such structures using probabilistic models is therefore quite important for statistical analysis, inference, and decision making.

In this paper, we consider *compressible priors*, whose independent and identically distributed (iid) realizations can be well-approximated as *sparse*:

Definition 1 (Compressible priors). *Define the* relative *best k-term approximation error $\bar{\sigma}_k(\mathbf{x})_q$ of a vector $\mathbf{x}$ as*

$$\bar{\sigma}_k(\mathbf{x})_q = \frac{\sigma_k(\mathbf{x})_q}{\|\mathbf{x}\|_q}, \tag{1}$$

where $\sigma_k(\mathbf{x})_q := \inf_{\|\mathbf{y}\|_0 \leq k} \|\mathbf{x} - \mathbf{y}\|_q$ is the best k-term approximation error of $\mathbf{x}$, and $\|\mathbf{x}\|_q$ is the ℓ^q-norm of $\mathbf{x}$, $q \in (0, \infty)$. We use the convention where $\|\mathbf{x}\|_0$ counts the non-zero coefficients of $\mathbf{x}$, and $\|\mathbf{x}\|_\infty$ selects its largest coefficient in magnitude.

*Let $X_n (n \in \mathbb{N})$ be iid samples from a probability distribution function (PDF) $p(x)$, and $\mathbf{x}_N = (X_1, \ldots, X_N) \in \mathbb{R}^N$. The PDF $p(x)$ is q-*compressible *with parameters (ϵ, κ) when*

$$\limsup_{N \to \infty} \bar{\sigma}_{k_N}(\mathbf{x}_N)_q \overset{a.s.}{\leq} \epsilon, (a.s.: \textit{almost surely}); \tag{2}$$

for any sequence k_N such that $\liminf_{N \to \infty} \frac{k_N}{N} \geq \kappa$.

The case of interest is when $\epsilon \ll 1$ and $\kappa \ll 1$: iid realizations of a q-compressible prior with parameters (ϵ, κ) live in ϵ-proximity to the union of κN-dimensional hyperplanes, where the closeness is measured in the ℓ^q-norm. These hyperplanes are aligned with the coordinate axes in N-dimensions. Compressible priors have many important applications, among which we highlight two here:

Statistics of natural images. Acquisition, compression, denoising, and analysis of natural images (similarly, medical, seismic, and hyperspectral images) draw high scientific and commercial interest.

Research to date in natural image modeling has had two distinct approaches, with one focusing on deterministic explanations and the other pursuing probabilistic models. Deterministic approaches operate under the assumption that the transform domain representations (e.g., wavelets, Fourier, curvelets, etc.) of images are compressible. Therefore, these approaches threshold the transform domain coefficients for sparse approximation, which can be used for compression or denoising.

Existing probabilistic approaches also exploit such power-law coefficient decays in transform domain representations, and learn probabilistic models by approximating the coefficient *histograms* or *moment matching*. For natural images, the canonical approach is to fit PDF's, such as generalized Gaussian distribution and the Gaussian scale mixtures, to the histograms of wavelet coefficients while trying to simultaneously capture the dependencies observed in their marginal and joint distributions.

Statistical regression. Underdetermined linear regression (ULR) is a fundamental problem in statistics, applied mathematics, and theoretical computer science with broad applications —from subset selection to compressive sensing and inverse problems (e.g., deblurring), and from data streaming to error corrective coding. In ULR, we seek an unknown vector $\mathbf{x} \in \mathbb{R}^N$, given its dimensionality reducing, linear projection $\mathbf{y} \in \mathbb{R}^m$ ($m < N$) obtained via a known *encoding* matrix $\mathbf{\Phi} \in \mathbb{R}^{m \times N}$, as

$$\mathbf{y} = \mathbf{\Phi}\mathbf{x} + \mathbf{n}, \tag{3}$$

where $\mathbf{n} \in \mathbb{R}^m$ accounts for the perturbations in the linear system, such as physical noise.

The core ULR challenge in *decoding* $\mathbf{x}$ from $\mathbf{y}$ stems from the simple fact that dimensionality reduction loses information in general: for any vector $v \in \text{kernel}(\mathbf{\Phi})$, it is impossible to distinguish $\mathbf{x}$ from $\mathbf{x}+v$ based on $\mathbf{y}$ alone. Prior information on $\mathbf{x}$ is therefore necessary to estimate the true $\mathbf{x}$ among the infinitely many possible solutions. It is now well-known that geometric *sparsity* models (associated to approximation of $\mathbf{x}$ from a finite union of low-dimensional subspaces in $\mathbb{R}^N$) play an important role in obtaining "good" solutions. A more probabilistic perspective considers $\mathbf{x}$ as drawn from a *prior*. As we will see, compressible (iid) priors countervail the ill-posed nature of ULR problems by generating vectors that, in high dimensions, typically fulfill the geometric sparsity model.

Main results. In this paper, we aim at bringing together the deterministic and probabilistic models of compressibility in a simple and general manner under the umbrella of compressible priors. To achieve our goal, we dovetail the concept of order statistics from probability theory with the deterministic models of compressibility from approximation theory. We focus on the incompressibility of the Laplace distribution as a running example for concreteness. Our contributions are summarized as follows:

1.1. *Relative sparse approximation error.* By using Wald's lemma on order statistics, we characterize the relative sparse approximation errors of iid PDF realizations, whereby providing solid mathematical ground to the earlier work of Cevher [5] on compressible priors. While Cevher exploits the decay of the expected order statistics, his approach is inconclusive in characterizing the "incompressibility" of priors. We close this gap with Proposition 2 (see Section 2) which introduces a function $G_q[p](\kappa)$ so that iid vectors as in Definition 1 satisfy

$$\lim_{N\to\infty} \bar{\sigma}_{k_N}(\mathbf{x}_N)_q^q \overset{a.s.}{=} G_q[p](\kappa) \tag{4}$$

when $\lim_{N\to\infty} \frac{k_N}{N} = \kappa \in (0,1)$. Proposition 2 provides a principled way of obtaining the compressibility parameters (ϵ, κ) of distributions in the high dimensional scaling of the vectors. An immediate application is the incompressibility of the Laplace distribution.

EXAMPLE 1. *As a stylized example, consider the Laplace distribution (also known as the double exponential) with scale parameter 1, whose PDF is given by*

$$p_1(x) := \frac{1}{2}\exp(-|x|). \tag{5}$$

We compute in Appendix I:

$$G_1[p_1](\kappa) = 1 - \kappa \cdot \Big(1 + \ln 1/\kappa\Big), \tag{6}$$

$$G_2[p_1](\kappa) = 1 - \kappa \cdot \Big(1 + \ln 1/\kappa + \frac{1}{2}(\ln 1/\kappa)^2\Big). \tag{7}$$

Therefore, it is straightforward to see that the Laplace distribution is not compressible *for* $q = 1$*: it is not possible to simultaneously have both* $\epsilon = G_1[p_1](\kappa)$ *and* κ *small.*

1.2. *Sparse modeling vs. sparsity promotion in ULR.* We show that the *maximum a posteriori* (MAP) interpretation of standard deterministic sparse recovery algorithms is, in some sense, inconsistent. To explain why, we consider the following *decoding* approaches to estimate a vector $\mathbf{x}$ from its encoding $\mathbf{y} = \mathbf{\Phi}\mathbf{x}$:

$$\Delta_1(\mathbf{y}) = \operatorname*{argmin}_{\tilde{\mathbf{x}}:\mathbf{y}=\mathbf{\Phi}\tilde{\mathbf{x}}} \|\tilde{\mathbf{x}}\|_1, \tag{8}$$

$$\Delta_{\mathrm{LS}}(\mathbf{y}) = \operatorname*{argmin}_{\tilde{\mathbf{x}}:\mathbf{y}=\mathbf{\Phi}\tilde{\mathbf{x}}} \|\tilde{\mathbf{x}}\|_2 = \mathbf{\Phi}^+\mathbf{y}, \tag{9}$$

$$\Delta_{\mathrm{oracle}}(\mathbf{y}, \Lambda) = \operatorname*{argmin}_{\tilde{\mathbf{x}}:\mathrm{support}(\tilde{\mathbf{x}})=\Lambda} \|\mathbf{y} - \mathbf{\Phi}\tilde{\mathbf{x}}\|_2 = \mathbf{\Phi}_\Lambda^+\mathbf{y}, \tag{10}$$

$$\Delta_{\mathrm{trivial}}(\mathbf{y}) = 0. \tag{11}$$

The decoder Δ_1 regularizes the solution space via the ℓ^1-norm. It is the *de facto* standard formulation for sparse recovery, and is tightly related to the least absolute shrinkage and selection operator (LASSO):

$$\Delta_{\mathrm{LASSO},1} = \operatorname*{argmin}_{\tilde{\mathbf{x}}} \lambda\|\tilde{\mathbf{x}}\|_1 + \frac{1}{2}\|\mathbf{y} - \mathbf{\Phi}\tilde{\mathbf{x}}\|_2^2, \text{ or } \Delta_{\mathrm{LASSO},2} = \operatorname*{argmin}_{\tilde{\mathbf{x}}} \|\mathbf{y} - \mathbf{\Phi}\tilde{\mathbf{x}}\|_2^2, \text{s.t. } \|\tilde{\mathbf{x}}\|_1 \leq \lambda,$$

where λ is a constant. Both Δ_1 and the LASSO formulations can be solved in polynomial time through convex optimization techniques. The decoder Δ_{LS} is the traditional minimum least-squares solution, which is related to Ridge Regression. It uses the Moore-Penrose pseudo-inverse $\mathbf{\Phi}^+ = \mathbf{\Phi}^T(\mathbf{\Phi}\mathbf{\Phi}^T)^{-1}$. The oracle sparse decoder Δ_{oracle} can be seen as an idealization of sparse decoders, which combine subset selection (the choice of Λ) with a form of linear regression. It is an "informed" decoder that has the side information of the index set Λ associated with the largest components in $\mathbf{x}$. The trivial decoder $\Delta_{\mathrm{trivial}}$ plays the devil's advocate for the performance guarantees of the other decoders.

1.2.1. *Guaranteed/expected performance of decoders ....* When the encoder $\mathbf{\Phi}$ provides near isometry to the set of sparse vectors [4], the decoder Δ_1 features an *instance optimality* property [6, 7]:

$$\|\Delta_1(\mathbf{\Phi}\mathbf{x}) - \mathbf{x}\|_1 \leq C_k \cdot \sigma_k(\mathbf{x})_1, \forall \mathbf{x}; \tag{12}$$

where C_k is a constant which depends on $\mathbf{\Phi}$. A similar result holds with the $\|\cdot\|_2$ norm on the left hand side. Unfortunately, it is impossible to have the same uniform guarantee for all $\mathbf{x}$ with $\sigma_k(\mathbf{x})_2$ on the right hand side [6], but for any given $\mathbf{x}$, it becomes possible *in probability* [6, 8]. For instance, when $\mathbf{\Phi}$ has iid Gaussian entries (i.e., $\varphi_{ij} \sim \mathcal{N}(0, \frac{1}{m})$ where $\varphi_{ij} = [\mathbf{\Phi}]_{i,j}$), Δ_1 recovers exact sparse vectors perfectly from as few as $m \approx 2ek \log N/k$ with high probability [9].

In the sequel, we consider a *Gaussian encoder*: we assume that $\mathbf{\Phi}$ is $m \times N$ with iid Gaussian entries, leading to Gaussian ULR (GULR) problems. We restrict our analysis to the noiseless setting ($\mathbf{n} = 0$). In Section 4, we theoretically characterize the expected performance (in terms of relative error) of the estimators Δ_{LS}, Δ_{oracle} for arbitrary high-dimensional vectors $\mathbf{x}$. The least squares decoder Δ_{LS} has expected performance $1 - \delta$, independent of the vector $\mathbf{x}$, where

$$\delta := m/N \tag{13}$$

is the *undersampling ratio* associated to the matrix $\mathbf{\Phi}$ (this terminology comes from compressive sensing, where $\mathbf{\Phi}$ is a sampling matrix). The expected performance of the oracle sparse decoder Δ_{oracle} satisfies

$$\frac{\mathbb{E}_{\mathbf{\Phi}}\|\Delta_{\text{oracle}}(\mathbf{\Phi}\mathbf{x}, \Lambda) - \mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} = \frac{1}{1 - \frac{k}{m-1}} \cdot \frac{\sigma_k(\mathbf{x})^2}{\|\mathbf{x}\|_2^2}$$

This error is the balance between two factors. The first factor grows with k (the size of the set Λ of largest entries of $\mathbf{x}$ used in the decoder) and reflects the (ill-)conditioning of the submatrix $\mathbf{\Phi}_\Lambda$, while the second factor is the best k-term relative approximation error, which shrinks as k increases. This highlights the inherent trade-off present in any sparse estimator, namely the level of sparsity k versus the conditioning of the submatrices of $\mathbf{\Phi}$.

1.2.2. *... and a few surprises.*

A crucial weakness in appealing to instance optimality. Although instance optimality (12) is usually considered as a strong property, it involves an implicit trade off: when k is small, the k-term error $\sigma_k(\mathbf{x})$ is large, while for larger k, the constant C_k is large. For instance, we have $C_k = \infty$, when $k \geq m$. In Section 3 we provide new key insights for instance optimality of algorithms by considering the *relative error* rather than the absolute error, and obtain the following result.

PROPOSITION 1. *Suppose that* $\mathbf{x} \in \mathbb{R}^N$ *is iid with respect to* $p(x)$ *as in Definition 1, and that* $p(x)$ *satisfies*

$$G_1[p](\kappa_0) \geq 1/2,$$

where $\kappa_0 \approx 0.18$ *is an absolute constant. In high dimension* N*, there is no undersampling ratio* $\delta = m/N$ *for which instance optimality for the decoder* Δ_1 *guarantees to outperform the trivial decoder* $\Delta_{trivial}$ *with* $\mathbf{\Phi}$ *an* $m \times N$ *Gaussian encoder.*

Proposition 1 is true for general PDF's; its conditions are easily verifiable for the Laplace distribution based on Example 1. This is discussed further in Section 3.

Fundamental limits of sparsity promoting decoders. The performance of the least-squares estimator Δ_{LS} degrades linearly with the undersampling factor $\delta := m/N$, and therefore does not provide good reconstruction at low sampling rates $\delta \ll 1$. It is therefore surprising that we can determine a large class of distributions for which the oracle sparse decoder Δ_{oracle} is outperformed by the least-squares decoder Δ_{LS}.

THEOREM 1. *Suppose that* $\mathbf{x} \in \mathbb{R}^N$ *is iid with respect to* $p(x)$ *as in Definition 1, and that* $p(x)$ *has a finite fourth-moment*

$$\mathbb{E}X^4 < \infty.$$

There exists a minimum undersampling ratio δ_0 *such that whenever* $\delta = m/N < \delta_0$*, for any* k*, the performance of the oracle* k*-sparse decoder* Δ_{oracle} *in high dimensions is almost surely worse than that of the least squares decoder* Δ_{LS}*.*

Thus if the data distribution $p(x)$ has a finite fourth moment, there exists a level of undersampling beyond which a simple least-squares reconstruction (typically a dense vector estimate) provides an estimate, which is closer to the true vector $\mathbf{x}$ than oracle sparse estimation!

Section 5 describes how to determine this undersampling boundary, e.g., for the generalized Gaussian distribution. For the Laplace distribution, $\delta_0 \approx 0.15$. In other words, when randomly sampling a high-dimensional Laplace vector, it is better to use least-squares reconstruction than minimum ℓ^1 norm reconstruction (or any other type of sparse estimator), unless the number of measures m is at least 15% of the original vector dimension N. To see how well Proposition 1 and Theorem 1 are grounded in practice, we provide the following example:

EXAMPLE 2. *Figure 1 examines in more detail the performance of the estimators for Laplace distributed data at various undersampling values. The horizontal lines indicate various signal-to-distortion-ratios (SDR) of* $3dB$*,* $10dB$ *and* $20dB$*. Thus for the oracle estimator to achieve* $10dB$*, the undersampling rate must be greater than 0.7, while to achieve a performance level of* $20dB$*, something that might reasonably be expected in many sensing applications, we can hardly afford any subsampling at all since this requires* $\delta > 0.9$*.*

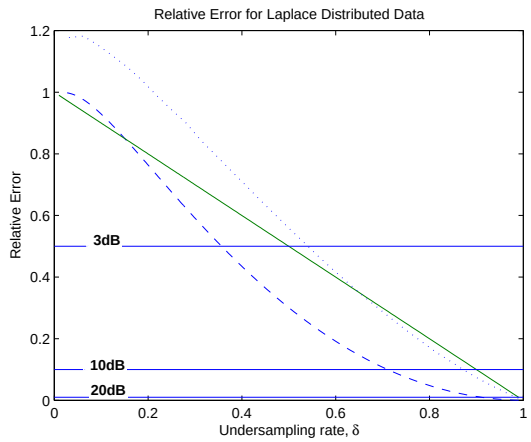


FIG 1. *The expected relative error as a function of the undersampling rates δ for data iid from a Laplace distribution using: (a) a linear least squares estimator (solid) and (b) the best oracle sparse estimator (dashed). Also plotted is the empirically observed average relative error over 5000 instances for the Δ_1 estimator (dotted). The horizontal lines indicate SDR values of $3dB$, $10dB$ and $20dB$, as marked.*

This may come as a surprise since, in Bayesian terminology, ℓ^1-norm minimization can be interpreted as the MAP estimator under the Laplace prior, while least squares is the MAP under the Gaussian prior. Such MAP interpretations of ULR decoders are further discussed below and contrasted to more geometric interpretations.

Pitfalls of MAP "interpretations" of decoders. Bayesian ULR methods employ probability measures as priors in the space of the unknown vector $\mathbf{x}$, and arbitrate the solution space by using the chosen measure. The decoder Δ_1 has a distinct probabilistic interpretation in the statistics literature: if we presume an iid probabilistic model for $\mathbf{x}$ as $p(X_n) \propto \exp(-c|X_n|)$ $(n = 1, \ldots, N)$, then $\Delta_{\text{LASSO},1}$ can be viewed as the MAP estimator

$$\Delta_{\text{MAP}}(\mathbf{y}) := \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}) = \arg\min_{\mathbf{x}} \{-\log p(\mathbf{x}|\mathbf{y})\},$$

when the noise $\mathbf{n}$ is iid Gaussian. However, as illustrated by Example 2, the decoder Δ_{MAP} performs quite poorly for iid Laplace vectors. The possible inconsistency of MAP estimators is a known phenomenon [18]. Yet, the fact that Δ_{MAP} is outperformed by Δ_{LS} –which is the MAP under the Gaussian prior– should remain somewhat of a surprise to many readers.

It is now not uncommon to stumble upon new proposals in the literature for the modification of Δ_1 or the LASSO with diverse thresholding or re-weighting rules based on different hierarchial probabilistic models—many of which correspond to a special Bayesian "sparsity" prior $p(\mathbf{x}) \propto \exp(-\phi(\mathbf{x}))$ [19], associated to the minimization of new cost func-

tions

$$\Delta_\phi(\mathbf{y}) := \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \phi(\mathbf{x}).$$

It has been shown in the context of Additive White Gaussian Noise denoising that the MAP interpretation of such penalized least-squares regression can be misleading [12]. Just as illustrated above with $\phi(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$, *while the geometric interpretations of the cost functions associated to such priors are useful for sparse recovery, the priors themselves do not necessarily constitute a relevant "generative model" for the vectors.* Hence, such proposals are losing a key strength of the Bayesian approach: the ability to evaluate the "goodness" or "confidence" of the estimates due to the probabilistic model itself or its conjugate prior mechanics.

In fact, the empirical success of Δ_1 (or Δ_{LASSO}) results from a combination of two properties:

1. the *sparsity enforcing* nature of the cost function, associated to the non-differentiability at zero of the ℓ^1 cost function;
2. the *compressible nature* of the unknown vector $\mathbf{x}$ to be estimated.

Geometrically speaking, the objective $\|\mathbf{x}\|_1$ is related to the ℓ^1-ball, which intersects with the constraints (e.g., a randomly oriented hyperplane, as defined by $\mathbf{y} = \mathbf{\Phi}\mathbf{x}$) along or near the k-dimensional hyperplanes ($k \ll N$) that are aligned with the canonical coordinate axes in $\mathbb{R}^N$. The geometric interplay of the objective and the constraints in high-dimensions inherently *promotes sparsity*. An important practical consequence is the ability to design efficient optimization algorithms for large-scale problems, using thresholding operations. Therefore, the decoding process of Δ_1 automatically sifts smaller subsets that best explain the observations, unlike the traditional least-squares Δ_{LS} in ULR.

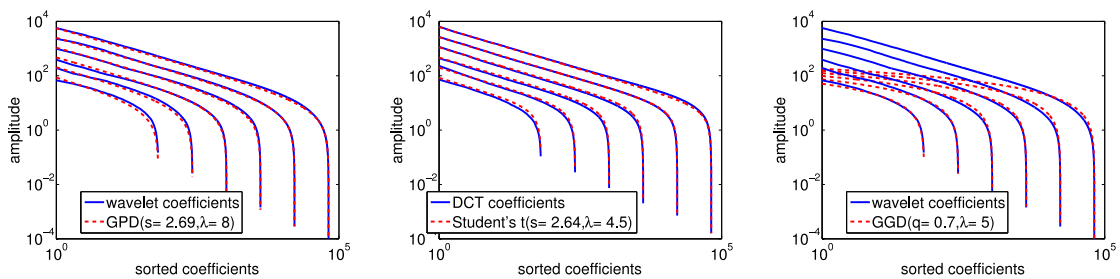
When $\mathbf{x}$ has iid coordinates as in Definition 1, compressibility is not so much related to the behavior (differentiable or not) of $p(x)$ around zero but, as we shall see, rather to the thickness of its tails, e.g., through the necessary property $\mathbb{E}X^4 = \infty$ (*cf* Theorem 1). We further show with Theorem 2 (in Section 3) that priors with infinite variance ($\mathbb{E}X^2 = \infty$) almost surely generate vectors which are sufficiently compressible to guarantee that the decoder Δ_1 with a Gaussian encoder $\mathbf{\Phi}$ of arbitrary (fixed) small sampling ratio $\delta = m/N$ has ideal performance in dimensions N growing to infinity:

$$\frac{\|\Delta_1(\mathbf{\Phi}_N\mathbf{x}_N) - \mathbf{x}_N\|_2}{\|\mathbf{x}_N\|_2} \overset{a.s.}{\to} 0$$

As shown in Section 6 there exist priors $p(x)$, which combine heavy tails with a non-smooth behaviour at zero, such that the associated MAP estimator is sparsity promoting. It is likely that the MAP with such priors can be shown to perform ideally well in the asymptotic regime.

TABLE 1
Summary of the main results

Moment property	$\mathbb{E}X^2 = \infty$	$\mathbb{E}X^2 < \infty$ and $\mathbb{E}X^4 = \infty$	$\mathbb{E}X^4 = \infty$
General result	*Theorem 2* Δ_1 performs ideally for any δ	N/A depends on finer properties of $p(x)$	*Theorem 1* Δ_{LS} outperforms Δ_{oracle} for small $\delta < \delta_0$
Compressibility	**compressible** status	**compressible** *or* **incompressible**	**incompressible**
Examples		*Proposition 4:* $p_0(x) := 2\lvert x\rvert/(x^2+1)^3$ Δ_{oracle} performs just as Δ_{LS}	*Section 5.2:* $p_\tau(x) \propto \exp(-\lvert x\rvert^\tau)$ $0 < \tau < \infty$ Generalized Gaussian
	Example 4: $p_{\tau,s}(x) \propto (1+\lvert x\rvert^\tau)^{-s/\tau}$ Generalized Pareto ($\tau = 1$) / Student's t ($\tau = 2$)		
	Case $1 < s \leq 3$	Case $3 < s < 5$ Δ_{oracle} outperforms Δ_{LS} for small $\delta < \delta_0$	Case $s > 5$



(a) Wavelet/GPD (b) DCT/Student's t distribution (c) Wavelet/GGD

FIG 2. *Statistics of natural images.*

1.3. *Statistics of natural images: compressible or incompressible ?.* Theorems 1 and 2 provide easy to check conditions for (in)compressibility of a prior $p(x)$ based on its second of fourth moments. These rules of thumb are summarized in Table 1, providing an overview at a glance of the main results obtained in this paper. We conclude with stylized application of these rules of thumb to wavelet and discrete cosine transform (DCT) coefficients of the natural images from the Berkeley database [16].

Figure 2 illustrates, in log-log scale, the average of the magnitude ordered wavelet coefficients (Figures 2-(a)-(c)), and of the DCT coefficients (Figure 2-(b)). They are obtained by randomly sampling 100 image patches of varying sizes $N = 2^j \times 2^j$ ($j = 3, \ldots, 8$), and taking their transforms (scaling filter for wavelets: Daubechies4). For comparison, we also plot the expected order statistics (dashed lines), as described in [5], of the following priors

- GPD: the scaled generalized Pareto distribution $\frac{1}{\lambda} p_{\tau,s}(x/\lambda)$, $\tau = 1$, with parameters

$s = 2.69$ and $\lambda = 8$ (Figure 2-(a));
- Student's t: the scaled Student's t distribution $\frac{1}{\lambda}p_{\tau,s}(x/\lambda)$, $\tau = 2$, with parameters $s = 2.64$ and $\lambda = 4.5$ (Figure 2-(b));
- GGD: the scaled generalized Gaussian distribution $\frac{1}{\lambda}p_{\tau}(x/\lambda)$, with parameters $\tau = 0.7$ and $\lambda = 5$ (Figure 2-(c)).

The GGD parameters were obtained by approximating the histogram of the wavelet coefficients at $N = 8 \times 8$, as it is the common practice in the signal processing community [10]. The GPD and Student's t parameters were tuned manually.

One should note that image transform coefficients are certainly not iid, for instance: nearby wavelets have correlated coefficients; wavelet coding schemes exploit well-known zero-trees indicating correlation across scales; the energy across wavelet scales often follows a power law decay.

Yet, the empirical goodness-of-fits in Figure 2 (a), (b) seem to indicate that the distribution of the coefficients of natural images, marginalized across all scales (in wavelets) or frequencies (DCT) can be well approximated by a distribution of the type $p_{\tau,s}$ (*cf* Table 1) with "compressibility parameter" $s \approx 2.67 < 3$. Interestingly, this corresponds to a regime where the results of [5] are inconclusive regarding the (in)compressibility, since the distribution is not sufficiently compressible to guarantee the performance of the ℓ^1 decoder Δ_1 using instance optimality. However, this does correspond to the regime where $\mathbb{E}X^2 = \infty$ (*cf* Example 4), indicating that in the limit of very high resolutions $N \to \infty$, such images are sufficiently compressible to be acquired using compressive sampling with both *arbitrary good relative precision* and *arbitrary small undersampling* factor $\delta = m/N \ll 1$.

Considering the GGD with parameter $\tau = 0.7$, the results of Section 5.2 (*cf* Figure 6) indicate that it is associated to a critical undersampling ratio $\delta_0(0.7) \approx 0.04$. Below this undersampling ratio, the oracle sparse decoder is outperformed by the least square decoder, which has the very poor expected relative error $1 - \delta \geq 0.96$. Should the GGD be an accurate model for coefficients of natural images, this would imply that compressive sensing of natural images requires a number of measures at least 4% of the target number of image pixels. However, while the generalized Gaussian approximation of the coefficients appear quite accurate at $N = 8 \times 8$, the empirical goodness-of-fits quickly deteriorate at higher resolution. For instance, the initial decay rate of the GGD coefficients varies with the dimension. Surprisingly, the GGD coefficients approximate the small coefficients (i.e., the histogram) rather well irrespective of the dimension. This phenomenon could be deceiving while predicting the compressibility of the images.

2. Asymptotics of best k-term relative error. In this section we estimate the relative best k-term approximation error $\sigma_k(\mathbf{x})_q/\|\mathbf{x}\|_q$ for random vectors $\mathbf{x}$ with iid entries, where we recall that $\sigma_k(\mathbf{x})_q := \inf_{\|\mathbf{y}\|_0 \leq k} \|\mathbf{x} - \mathbf{y}\|_q$. We postpone all proofs to the Appendix.

PROPOSITION 2. *Suppose $\mathbf{x}_N \in \mathbb{R}^N$ is iid with respect to $p(x)$ as in Definition 1. Denote $\bar{p}(x) := 0$ for $x < 0$, and $\bar{p}(x) := p(x) + p(-x)$ for $x \geq 0$ as the PDF of $|X_n|$, and $\bar{F}(t) :=$*

$\mathbb{P}(|X| \leq t)$ as its cumulative distribution function. Assume that $\bar{F}$ is continuous and strictly increasing on some interval $[a\ b]$, with $\bar{F}(a) = 0$ and $\bar{F}(b) = 1$, where $0 \leq a < b \leq \infty$. For any $0 < \kappa \leq 1$, define the following function:

$$G_q[p](\kappa) := \frac{\int_0^{\bar{F}^{-1}(1-\kappa)} x^q \bar{p}(x)dx}{\int_0^{\infty} x^q \bar{p}(x)dx}. \tag{14}$$

1. ***Bounded moments:*** *assume $\mathbb{E}|X|^q < \infty$ for some $q \in (0, \infty)$. Then, $G_q[p](\kappa)$ is also well defined for $\kappa = 0$, and given any sequence k_N such that $\lim_{N\to\infty} \frac{k_N}{N} = \kappa \in [0, 1]$, the following holds almost surely*

$$\lim_{N\to\infty} \bar{\sigma}_{k_N}(\mathbf{x}_N)_q^q \overset{a.s.}{=} G_q[p](\kappa). \tag{15}$$

2. ***Unbounded moments:*** *assume $\mathbb{E}|X|^q = \infty$ for some $q \in (0, \infty)$. Then, for $0 < \kappa \leq 1$ and any sequence k_N such that $\lim_{N\to\infty} \frac{k_N}{N} = \kappa$, the following holds almost surely*

$$\lim_{N\to\infty} \bar{\sigma}_{k_N}(\mathbf{x}_N)_q^q \overset{a.s.}{=} G_q[p](\kappa) = 0. \tag{16}$$

Remark 1. *To further characterize the typical asymptotic behaviour of the relative error when $\mathbb{E}_p(|X|^q) = \infty$ and $k_N/N \to 0$ appears to require a more detailed characterization of the probability density function, such as decay bounds on the tails of the distribution.*

3. Instance optimality, ℓ^r-balls and compressibility in ULR. Well-known results indicate that for certain matrices, $\mathbf{\Phi}$, and for certain types of sparse estimators of $\mathbf{x}$, such as the minimum ℓ^1 norm solution $\Delta_1(\mathbf{y})$:

$$\Delta_1(\mathbf{x}) = \underset{\mathbf{x}}{\operatorname{argmin}}\, ||\mathbf{x}||_1 \text{ such that } \mathbf{y} = \mathbf{\Phi}\mathbf{x},$$

an instance optimality property holds [6]. In the simplest case of noiseless observations, this reads: the pair $\{\mathbf{\Phi}, \Delta\}$ is instance optimal to order k with constant C_k if for all $\mathbf{x}$:

$$\|\Delta(\mathbf{\Phi}\mathbf{x}) - \mathbf{x}\| \leq C_k \cdot \sigma_k(\mathbf{x}) \tag{17}$$

where $\sigma_k(\mathbf{x})$ is the error of best approximation of $\mathbf{x}$ with k-sparse vectors, while C_k is a constant which depends on k. Various flavours of instance optimality are possible [4, 6]. We will initially focus on ℓ^1 instance optimality. For the ℓ^1 estimator (8) it is known that instance optimality in the ℓ^1 norm (i.e. ℓ^1 norms are used on both hand sides of (17)) is related to the following robust null space property. The matrix $\mathbf{\Phi}$ satisfies the robust null space property of order k with constant $\eta \leq 1$ if:

$$||\mathbf{z}_\Omega||_1 < \eta ||\mathbf{z}_{\bar{\Omega}}||_1 \tag{18}$$

for all nonzero $\mathbf{z}$ belonging to the null space kernel$(\mathbf{\Phi}) := \{\mathbf{z}, \mathbf{\Phi z} = 0\}$ and all index sets Ω of size k, where the notations $\mathbf{z}_\Omega$ stands for the vector matching $\mathbf{z}$ for indices in Ω and zero elsewhere. It has further been shown [7, 20] that the robust null space property of order k with constant η_k is a necessary and sufficient condition for ℓ^1-instance optimality with the constant C_k given by:

$$C_k = 2\frac{(1+\eta_k)}{(1-\eta_k)} \tag{19}$$

Instance optimality is commonly considered as a strong property, since it controls the *absolute* error in terms of the "compressibility" of $\mathbf{x}$, expressed through $\sigma_k(\mathbf{x})$. For instance optimality to be meaningful we therefore require that $\sigma_k(\mathbf{x})$ be small in some sense. This idea has been encapsulated in a deterministic notion of compressible vectors [6]. From an approximation theoretic point of view, it is usual to consider a vector $\mathbf{x}$ as compressible if it is contained within some *weak* ℓ^r ball where the *weak* ℓ^r ball of radius R contains all vectors $\mathbf{x}$ for which

$$\|\mathbf{x}\|_{w\ell^r} := \sup_n \left\{ |\mathbf{x}|_n^* \cdot n^{1/r} \right\} \leq R, \tag{20}$$

with $|x|_n^*$ the n-th largest absolute value of elements of $\mathbf{x}$.

For instance, if $\mathbf{x}$ lies inside an ℓ^p ball it will also be within a weak ℓ^p ball of the same radius, see Figure 3(a) which shows a weak ℓ^r ball together with the ℓ^r ball of the same radius. The motivation for such a definition of compressibility comes from the fact that we can then bound $\sigma_k(\mathbf{x})_q$ for $q > r$, as

$$\sigma_k(\mathbf{x})_q \leq R \left(\frac{r}{q-r} \right)^{1/q} k^{-(1/r-1/q)}, \tag{21}$$

therefore guaranteeing that the k-term approximation error is vanishingly small for large enough k.

A naive way to interpret the ℓ^r balls within the statistical data models is as follows. Let us assume that $\mathbf{x}_N = (X_1, \ldots, X_N)$ is a vector of iid samples drawn from some probability distribution $p(x)$. If $\mathbb{E}|X|^r = C < \infty$ then by the strong law of large numbers, the quantity $\|\mathbf{x}_N\|_r^r/N$, $N \in \mathbb{N}$, converges almost surely to C, i.e. the distance from $\mathbf{x}_N/N^{1/r}$ to the surface of the ℓ^r ball of radius $C^{1/r}$ converges almost surely to zero. This often leads to the assertion that a vector drawn from certain probability distributions is "compressible" since (when normalized) it lives in a finite radius ℓ^r ball.

Unfortunately, this is a common misconception. Finite dimensional ℓ^r balls also contain 'flat' vectors with entries of similar magnitude, that have very small k-term approximation error ... only because the vectors are very small themselves.

For example, if $\mathbf{x}_N$ has entries drawn from the Laplace distribution then $\mathbf{x}_N/N$ will have with high probability an ℓ^1 norm close to 1. However the Laplace distribution also has a

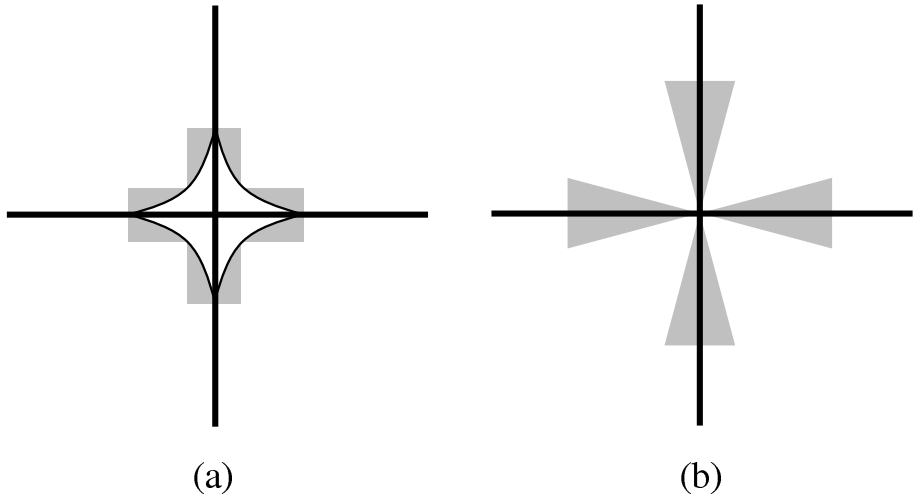


FIG 3. *(a) A cartoon view of an ℓ^r ball (white) and the weak ℓ^r ball of the same radius (grey); (b) A cartoon view of the notion of the compressible rays model.*

finite second moment $\mathbb{E}X^2 = 2$, hence with high probability $\mathbf{x}_N/N^{1/2}$ has ℓ^2 norm close to $\sqrt{2}$, i.e. $\mathbf{x}_N/N$ has ℓ^2 norm close to $\sqrt{2/N}$. This is not far from the ℓ^2 norm of the largest flat vectors that live in the unit ℓ^1 ball, which have the form $|\mathbf{x}|_n = 1/N$, $1 \leq n \leq N$, suggesting that the typical iid Laplace distributed vector is a small and relatively flat vector. This is illustrated on Figure 4.

One could argue that the above normalization by $1/N$ was incorrect and that there is a normalization that can define a weak ℓ^r ball that truely captures the decay behaviour of $|\mathbf{x}|_n^\star$. This basically forms the basis of the approach in [5], where specific values of R and r in the upper bound (20) are calculated for various distributions, in relation with order statistics. Now, we instead consider a more natural normalization of $\sigma_k(\mathbf{x})_q$ with respect to the size of the original vector $\mathbf{x}$ measured in the same norm. This is, of course, the best k-term relative error $\bar{\sigma}_k(\mathbf{x})_q$ that we investigated in Section 2. Note that the class defined by a bounded k-term relative error does not have the shape of an ℓ^r ball or weak ℓ^r ball. Instead it forms a set of *compressible 'rays'* as depicted in Figure 3 (b).

3.1. *Limits of ULR performance guarantees using instance optimality.* In terms of the relative best k-term approximation error, the instance optimality implies the following inequality:

$$\frac{\|\Delta(\mathbf{\Phi}\mathbf{x}) - \mathbf{x}\|}{\|\mathbf{x}\|} \leq \min_k \{C_k \cdot \bar{\sigma}_k(\mathbf{x})\}$$

Note that if we have the following inequality satisfied for the particular realization of $\mathbf{x}$

$$\frac{\sigma_k(\mathbf{x})}{\|\mathbf{x}\|} \geq C_k^{-1}, \forall k,$$

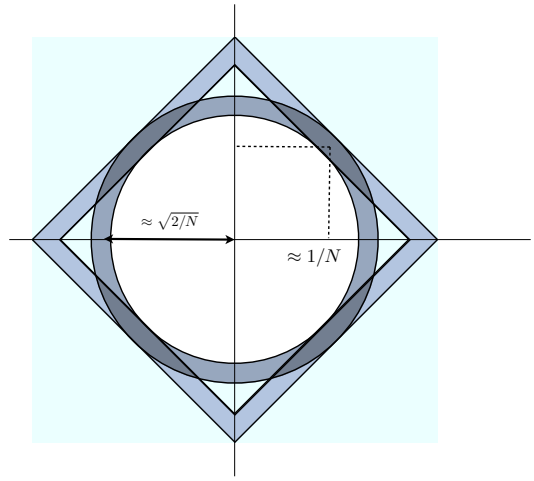


FIG 4. *A cartoon view of the ℓ^1 and ℓ^2 "rings" where vectors with iid Laplace-distributed entries concentrate. The radius of the ℓ^2 ring is of the order of $\sqrt{2/N}$ while that of the ℓ^1 ring is one, corresponding to vectors with flat entries $|\mathbf{x}|_n \approx 1/N$.*

then the only consequence of instance optimality is that $\|\Delta(\mathbf{\Phi x}) - \mathbf{x}\| \leq \|\mathbf{x}\|$. In other words, the performance guarantee for the considered vector $\mathbf{x}$ is no better than for the trivial zero estimator: $\Delta_{\text{trivial}}(\mathbf{y}) = 0$, for any $\mathbf{y}$.

This simple observation illustrates that one should be careful in the interpretation of instance optimality. In particular, ULR decoding algorithms with instance optimality guarantees may not universally perform better than other simple or more standard estimators.

To understand what this implies for specific priors, consider the case of ℓ^1 decoding with a Gaussian random sensing matrix Φ_N. For this coder, decoder pair, $\{\Phi_N, \Delta_1\}$, we know there is a strong phase transition associated with the robust null space property and hence the instance optimality property in terms of the undersampling factor $\delta := m/N$ and the factor $\rho := k/m$ as $k, m, N \to \infty$ [20]. This is a generalization of the ℓ^1 exact recovery phase transition of Donoho and Tanner [9] which corresponds to $\eta = 1$. We can therefore identify the smallest instance optimality constant asymptotically possible as a function of ρ and δ which we will term $C(\rho, \delta)$.

To check whether instance optimality guarantees can beat the zero estimator Δ_{trivial} for a given undersampling ratio δ, and a given probability model $p(x)$, we need to consider the product of $\bar{\sigma}_k(\mathbf{x})_1 \overset{a.s.}{\to} G_1[p](\kappa)$ and $C(\frac{\kappa}{\delta}, \delta)$. If

$$G_1[p](\kappa) > \frac{1}{C\left(\frac{\kappa}{\delta}, \delta\right)}, \quad \forall \kappa \in [0, 1] \tag{22}$$

then the instance optimality offers no guarantee to outperform the trivial zero estimator.

In order to bound the value of instance optimality we make the following observations:

- $C(\frac{\kappa}{\delta}, \delta) \geq 2$ for all κ and δ;
- $C(\frac{\kappa}{\delta}, \delta) = \infty$ for all δ if $\kappa > \kappa_0 \approx 0.18$.

The first observation comes from minimising C_k in (19) with respect to $0 \leq \eta \leq 1$. The second observation stems from the fact that $\kappa_0 := \max_{\{\eta,\delta\}} \rho_\eta(\delta) \approx 0.18$ [9] (where $\rho_\eta(\delta)$ is the strong threshold associated to the null space property with constant $\eta \leq 1$) therefore we have $\kappa = \delta\rho \leq \kappa_0 \approx 0.18$ for any finite C. From these observations we obtain:

PROPOSITION 3. *Suppose that the distribution $p(x)$ satisfies $G_1[p](\kappa_0) \geq 1/2$. Then, there is no undersampling ratio $\delta = m/N$ for which instance optimality for the ℓ^1 decoder Δ_1 guarantees to outperform the trivial decoder $\Delta_{trivial}$ (for large vectors $\mathbf{x}_N$ with iid entries drawn according to $p(x)$).*

One might try to weaken the analysis by considering typical joint behaviour of Φ_N and $\mathbf{x}_N$. This corresponds to the 'weak' phase transitions [9, 20]. For this scenario there is a modified ℓ^1 instance optimality property [20], however the constant still satisfies $C(\frac{\kappa}{\delta}, \delta) \geq 2$. Furthermore since $\kappa \leq \delta$ we can define an undersampling ratio δ_0 by $G_1[p](\delta_0) = 1/2$, such that weak instance optimality provides no guarantee that Δ_1 will outperform the trivial decoder Δ_{trivial} in the region $0 < \delta \leq \delta_0$. More careful analysis will only increase the size of this region.

EXAMPLE 3. *The Laplace distribution Suppose that $\mathbf{x}_N = (X_1, \ldots, X_N)$ has iid entries X_n that follow the Laplace distribution $p_1(x)$. Then for large N, as noted in Example 1, the relative best k-term error is given by:*

$$G_1[p_1](\kappa) = 1 - \kappa \cdot \Big(1 + \ln 1/\kappa\Big)$$

Figure 5 shows that unfortunately this function exceeds $1/2$ on the interval $\kappa \in [0,\ \kappa_0]$ indicating there are no performance guarantees from instance optimality. Even exploiting weak instance optimality we can have no non-trivial guarantees below $\delta_0 \approx 0.18$.

3.2. *GULR performance guarantees for random variables with unbounded second moment.* We end this section with a positive result showing that random variables with infinite second moment, which are highly compressible (cf Proposition 2), are almost perfectly estimated by the ℓ^1 decoder Δ_1.

This result is based upon a variant of instance optimality: ℓ^2 instance optimality *in probability* [6] which can be shown to hold for a large class of random matrices [8]. This can be combined with the fact that when $\mathbb{E}X^2 = \infty$, from Proposition 2, we have $G_2[p](\kappa) = 0$ for all $0 < \kappa \leq 1$ to give the following.

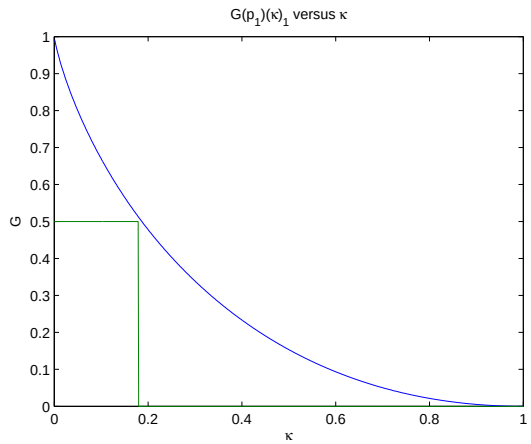


FIG 5. *The ℓ^1-norm best k-term approximation relative error $G_1[p_1](\kappa)$ as a function of $\kappa = k/N$ (top curve) along with a rectangular shaped function (bottom curve) that upper bounds $\inf_\delta C^{-1}(\kappa/\delta, \delta)$.*

THEOREM 2 (Asymptotic performance of the ℓ^1 decoder under infinite second moment). *Let $X_n, n \in \mathbb{N}$ be iid samples from a distribution with PDF $p(x)$ satisfying the hypotheses of Proposition 2. Assume that $\mathbb{E}X^2 = \infty$, and define the coefficient vector $\mathbf{x}_N = (X_1, \ldots, X_N) \in \mathbb{R}^N$. Similarly let $\phi_{i,j}$, $i, j \in \mathbb{N}$ be iid Gaussian variables $\mathcal{N}(0,1)$ and define the $m_N \times N$ Gaussian random matrix $\mathbf{\Phi}_N = \left[\phi_{ij}/\sqrt{m_N}\right]_{1 \le i \le m_N, 1 \le j \le N}$.*

Consider a sequence of integers m_N such that $\lim_{N\to\infty} m_N/N = \delta$ then

$$\frac{\|\Delta_1(\mathbf{\Phi}_N \mathbf{x}_N) - \mathbf{x}_N\|_2}{\|\mathbf{x}_N\|_2} \overset{a.s.}{\to} 0 \tag{23}$$

REMARK 2. *A similar but weaker result can be derived based on ℓ^1 instance optimality that shows that when $\mathbb{E}|X| = \infty$, (23) holds for the ℓ^1 decoder with a Gaussian encoder.*

We can therefore conclude that a random variable with infinite variance is not only compressible (in the sense of Proposition 2): it can also be accurately approximated from undersampled measurements within a compressive sensing scenario. In contrast, instance optimality provides no guarantees of compressibility when the variance is finite. At this juncture it is not clear where the blame for this result lies. Is it in the strength of the instance optimality theory, or are distributions with finite variance simply not able to generate sufficiently compressible vectors for sparse recovery to be succesful at all? We will explore this latter question further in subsequent sections.

4. GULR performance of oracle sparse reconstruction *vs* least squares. Consider $\mathbf{x}$ an arbitrary vector in $\mathbb{R}^N$ and $\mathbf{\Phi}$ be an $m \times N$ random Gaussian matrix, and let

$\mathbf{y} := \mathbf{\Phi x}$. Besides the trivial zero estimator Δ_{trivial} (11) and the ℓ^1 minimization estimator Δ_1 (8), the Least Squares (LS) estimator Δ_{LS} (9) is a commonly used alternative. Due to the Gaussianity of $\mathbf{\Phi}$ and its independence from $\mathbf{x}$, it is well known that the resulting relative expected performance is

$$\frac{\mathbb{E}_{\mathbf{\Phi}}\|\Delta_{\text{LS}}(\mathbf{\Phi x}) - \mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} = 1 - \frac{m}{N}. \tag{24}$$

Moreover, there is indeed a concentration around the expected value, as expressed by the inequality below:

$$(1-\epsilon)\left(1 - \frac{m}{N}\right) \leq \frac{\|\Delta_{\text{LS}}(\mathbf{\Phi x}) - \mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} \leq (1-\epsilon)^{-1}\left(1 - \frac{m}{N}\right), \tag{25}$$

for any $\epsilon > 0$ and $\mathbf{x} \in \mathbb{R}^N$, except with probability at most $2 \cdot e^{-(N-m)\epsilon^2/4} + 2 \cdot e^{-N\epsilon^2/4}$.

The result is independent of the vector $\mathbf{x}$, which should be no surprise since the Gaussian distribution is isotropic. The expected performance is directly governed by the *undersampling factor*, i.e. the ratio between the number of measures m and the dimension N of the vector $\mathbf{x}$, $\delta := m/N$.

In order to understand which statistical distributions $p(x)$ lead to "compressible enough" vectors $\mathbf{x}$, we wish to compare the performance of LS with that of estimators Δ that exploit the sparsity of $\mathbf{x}$ to estimate it. Instead of choosing a particular estimator (such as Δ_1), we consider the *oracle sparse estimator* Δ_{oracle} (10), which is likely to upper bound the performance of most sparsity based estimators. While in practice $\mathbf{x}$ must be estimated from $\mathbf{y} = \mathbf{\Phi x}$, the oracle is given a precious side information : the index set Λ associated to the k largest components in $\mathbf{x}$, where $k < m$. Given this information, the oracle computes

$$\Delta_{\text{oracle}}(\mathbf{y}, \Lambda) := \underset{\text{support}(\mathbf{x})=\Lambda}{\text{argmin}} \|\mathbf{y} - \mathbf{\Phi x}\|_2^2 = \mathbf{\Phi}_{\Lambda_k}^{+}\mathbf{y},$$

where, since $k < m$, the pseudo-inverse is $\mathbf{\Phi}_{\Lambda}^{+} = (\mathbf{\Phi}_{\Lambda}^T\mathbf{\Phi}_{\Lambda})^{-1}\mathbf{\Phi}_{\Lambda}^T$. Unlike LS, the expected performance of the oracle estimators drastically depend on the shape of the best k-term approximation relative error of $\mathbf{x}$. Denoting $\mathbf{x}_I$ the vector whose entries match those of $\mathbf{x}$ on an index set I and are zero elsewhere, and $\bar{I}$ the complement of an index set, we have the following result.

THEOREM 3 (Expected performance of Oracle sparse estimation). *Let $\mathbf{x} \in \mathbb{R}^N$ be an arbitrary vector, $\mathbf{\Phi}$ be an $m \times N$ random Gaussian matrix, and $\mathbf{y} := \mathbf{\Phi x}$. Let Λ be an index set of size $k < m-1$, either deterministic, or random but* statistically independent *from $\mathbf{\Phi}$. We have*

$$\frac{\mathbb{E}_{\mathbf{\Phi}}\|\Delta_{oracle}(\mathbf{\Phi x}, \Lambda) - \mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} = \frac{1}{1 - \frac{k}{m-1}} \cdot \frac{\|\mathbf{x}_{\bar{\Lambda}}\|_2^2}{\|\mathbf{x}\|_2^2} \geq \frac{1}{1 - \frac{k}{m-1}} \cdot \frac{\sigma_k(\mathbf{x})_2^2}{\|\mathbf{x}\|_2^2}. \tag{26}$$

If Λ is chosen to be the k largest components of $\mathbf{x}$, then the last inequality is an equality. Moreover, we can characterize the concentration around the expected value as

$$1 + \frac{k(1-\epsilon)^3}{m-k+1} \leq \frac{\|\Delta_{oracle}(\mathbf{\Phi}\mathbf{x}, \Lambda) - \mathbf{x}\|_2^2}{\|\mathbf{x}_{\bar{\Lambda}}\|_2^2} \leq 1 + \frac{k(1-\epsilon)^{-3}}{m-k+1} \tag{27}$$

except with probability at most

$$8 \cdot e^{-\min(k, m-k+1)\cdot c_l(\epsilon)/2}, \quad \text{where } c_l(\epsilon) := -\ln(1-\epsilon) - \epsilon \geq \epsilon^2/2. \tag{28}$$

REMARK 3. *Note that this result assumes that Λ is* statistically independent *from $\mathbf{\Phi}$. Interestingly, for practical decoders such as the ℓ^1 decoder, Δ_1, the selected Λ might not satisfy this assumption, unless the decoder succesfully identifies the support of the largest components of $\mathbf{x}$.*

4.1. *Compromise between approximation and conditioning.* We observe that the expected performance of both Δ_{LS} and Δ_{oracle} is essentially governed by the quantities $\delta = m/N$ and $\rho = k/m$, which are reminiscent of the parameters in the phase transition diagrams of Donoho and Tanner [9]. However, while in the work of Donoho and Tanner the quantity ρ parameterizes a *model* on the vector $\mathbf{x}_N$, which is assumed to be $\rho\delta N$-sparse, here ρ rather indicates the order of k-term approximation of $\mathbf{x}_N$ that is *chosen* in the oracle estimator. In a sense, it is more related to a stopping criterion that one would use in a greedy algorithm. The quantity that actually models $\mathbf{x}_N$ is the function $G_2[p]$, provided that $\mathbf{x}_N \in \mathbb{R}^N$ has iid entries X_n with PDF $p(x)$ and finite second moment $\mathbb{E}X^2 < \infty$. Indeed, combining Proposition 2 and Theorem 3 we obtain:

THEOREM 4. *Let $X_n, n \in \mathbb{N}$ be iid samples from a distribution with PDF $p(x)$. Assume that the hypotheses of Proposition 2 hold and that $\mathbb{E}X^2 < \infty$. Let $\phi_{i,j}$, $i, j \in \mathbb{N}$ be iid Gaussian variables $\mathcal{N}(0,1)$. Consider two sequences k_N, m_N of integers and assume that*

$$\lim_{N\to\infty} k_N/m_N = \rho \quad \text{and} \quad \lim_{N\to\infty} m_N/N = \delta. \tag{29}$$

Define the $m_N \times N$ Gaussian random matrix $\mathbf{\Phi}_N = \left[\phi_{ij}/\sqrt{m_N}\right]_{1\leq i\leq m_N, 1\leq j\leq N}$, the coefficient vector $\mathbf{x}_N = (X_1, \ldots, X_N) \in \mathbb{R}^N$, and the observation $\mathbf{y}_N = \mathbf{\Phi}_N\mathbf{x}_N$. Let Λ_N be the index of the k_N largest magnitude coordinates of $\mathbf{x}_N$. We have the almost sure convergence

$$\lim_{N\to\infty} \frac{\|\Delta_{oracle}(\mathbf{y}_N, \Lambda_N) - \mathbf{x}_N\|_2^2}{\|\mathbf{x}_N\|_2^2} \overset{a.s.}{=} \frac{G_2[p](\rho\delta)}{1-\rho}; \tag{30}$$

$$\lim_{N\to\infty} \frac{\|\Delta_{LS}(\mathbf{y}_N) - \mathbf{x}_N\|_2^2}{\|\mathbf{x}_N\|_2^2} \overset{a.s.}{=} 1 - \delta. \tag{31}$$

For a given undersampling ratio $\delta = m/N$, the asymptotic expected performance of the oracle therefore depends on the relative number of components that are kept $\rho = k/m$, and we observe the same tradeoff as discussed in Section 3:

- For large k, close to the number of measures m (ρ close to one), the ill-conditioning of the pseudo-inverse matrix $\mathbf{\Phi}_\Lambda$ (associated to the factor $1/(1-\rho)$) adversely impacts the expected performance;
- For smaller k, the pseudo-inversion of this matrix is better conditioned, but the k-term approximation error governed by $G_2[p](\rho\delta)$ is increased.

Overall, for some intermediate size $k \approx \rho^\star m$ of the oracle support set Λ_k, the best tradeoff between good approximation and good conditioning is achieved, leading at best to the asymptotic expected performance

$$H[p](\delta) := \inf_{\rho \in (0,1)} \frac{G_2[p](\rho\delta)}{1-\rho}. \tag{32}$$

5. Comparison between least squares and oracle sparse methods. The question that we will now investigate is how the expected performance of oracle sparse methods compares to that of least squares, i.e., how large is $H[p](\delta)$ compared to $1-\delta$? We are particularly interested in understanding how they compare for small δ. Indeed, large δ values are associated with scenarii that are quite irrelevant to, for example, compressive sensing since the projection $\mathbf{\Phi x}$ cannot significantly compress the dimension of $\mathbf{x}$. Moreover, it is in the regime where δ is small that the expected performance of least squares is very poor, and we would like to understand for which distributions p sparse approximation is an inappropriate tool. The answer will of course depend on the PDF p through the function $G[p](\cdot)$. To characterize this we will say that a PDF p is *incompressible at a subsampling rate of* δ if

$$H[p](\delta) > 1 - \delta$$

In practice, there is often a minimal undersampling rate, δ_0, such that for $\delta \in (0, \delta_0)$ least squares estimation dominates the oracle sparse estimator. Specifically we will show below that priors $p(x)$ with a finite fourth moment $\mathbb{E}X^4 < \infty$, such as generalized Gaussians, always have some minimal undersampling rate $\delta_0 \in (0,1)$ below which they are incompressible. As a result, unless we perform at least $m \geq \delta_0 \cdot N$ random Gaussian measurement of an associated $\mathbf{x}_N$, it is not worth relying on sparse methods for reconstruction since least squares can do as good a job.

When the fourth moment of the distribution is infinite, one might hope that the converse is true, i.e. that no such minimal undersampling rate δ_0 exists. However, this is not the case. We will show that there is a prior p_0, with infinite fourth moment and finite second moment, such that

$$H[p_0](\delta) = 1 - \delta, \qquad \forall \delta \in (0,1).$$

Up to a scaling factor, this prior is associated to the symmetric prior

$$p_0(x) := \frac{2|x|}{(x^2+1)^3} \tag{33}$$

and illustrates that least squares can be competitive with oracle sparse reconstruction even when the fourth moment is infinite.

5.1. *Priors incompatible with high levels of undersampling.* In this section we show that when a distribution $p(x)$ has a finite fourth moment, $\mathbb{E}X^4 < \infty$, then it will generate vectors which are not sufficiently compressible to be compatible with compressive sensing at high level of undersampling. We begin by showing that the comparison of $H[p](\delta)$ to $1-\delta$ is related to that of $G_2[p](\kappa)$ with $(1-\sqrt{\kappa})^2$.

LEMMA 1. *Consider a function $G(\kappa)$ defined on $(0,1)$ and define*

$$H(\delta) := \inf_{\rho\in(0,1)} \frac{G(\delta\rho)}{1-\rho}. \tag{34}$$

1. *If $G(\delta^2) \le (1-\delta)^2$, then $H(\delta) \le 1-\delta$.*
2. *If $G(\kappa) \le (1-\sqrt{\kappa})^2$ for all $\kappa \in (0,\sqrt{\delta_0})$, then $H(\delta) \le 1-\delta$ for all $\delta \in (0,\delta_0)$.*
3. *If $G(\kappa) \ge (1-\sqrt{\kappa})^2$ for all $\kappa \in (0,\delta_0)$, then $H(\delta) \ge 1-\delta$ for all $\delta \in (0,\delta_0)$.*

This Lemma allows us to deal directly with $G_2[p](\kappa)$ instead of $H[p](\delta)$. Furthermore the $(1-\sqrt{\kappa})^2$ term can be related to the fourth moment of the distribution giving the following result:

THEOREM 5. *If $\mathbb{E}_{p(x)}X^4 < \infty$, then there exists a minimum undersampling $\delta_0 = \delta_0[p] > 0$ such that for $\delta < \delta_0$,*

$$H[p](\delta) \ge 1-\delta, \forall\ \delta \in (0,\delta_0). \tag{35}$$

and the performance of the oracle k-sparse estimation as described in Lemma 3 is asymptotically almost surely worse than that of least squares estimation as $N \to \infty$.

Roughly speaking, if $p(x)$ has a finite fourth moment, then in the regime where the relative number of measurement is (too) small we obtain a better reconstruction with least squares than with the oracle sparse reconstruction!

Note that this is rather strong, since the oracle is allowed to know not only the support of the k largest components of the unknown vector, but also the best choice of k to balance approximation error against numerical conditioning. A striking example is the case of Generalized Gaussian distributions discussed below.

One might also hope that, reciprocally, having an infinite fourth moment would suffice for a distribution to be sparse-compatible. The following result disproves this hope.

PROPOSITION 4. *With the distribution $p_0(x)$ defined in (33), we have*

$$H[p_0](\delta) = 1 - \delta, \forall\ \delta \in (0,1). \tag{36}$$

On reflection this should not be that surprising. The distribution $p_0(x)$ has no probability mass at $x = 0$ and resembles a smoothed Benoulli distribution with heavy tails.

5.2. *Worked example: the Generalized Gaussian.* Theorem 5 applies in particular whenever $\mathbf{x}_N$ is drawn from a Generalized Gaussian distribution,

$$p_\tau(x) \propto \exp\left(-c|x|^\tau\right), \tag{37}$$

where $0 < \tau < \infty$. The shape parameter, τ controls how heavy or light the tails of the distribution are. When $\tau = 2$ the distribution reduces to the standard Gaussian, while for $\tau < 2$ it gives a family of heavy tailed distributions with positive kurtosis. When $\tau = 1$ we have the Laplace distribution and for $\tau \leq 1$ it is often considered that the prior is in some way "sparsity-promoting". However, the Generalized Gaussian always has a finite fourth moment for all $\tau > 0$. Thus Theorem 5 informs us that there is always a critical undersampling value below which the Generalized Gaussian is incompressible.

While Theorem 5 indiciates the existence of a critical δ_0 is does not provide us with a useful bound. Fortunately, although in general we are unable to derive explicit expressions for $G[p](\cdot)$ and $H[p](\delta)$ (with the exceptions of $\tau = 1, 2$ - see appendix I), the generalized Gaussian has a closed form expression for its cdf in terms of the incomplete gamma function.

$$F(x) = \frac{1}{2} + \mathrm{sgn}(x)\frac{\gamma\left(1/\tau, c|x|^\tau\right)}{2\Gamma(1/\tau)}$$

where $\Gamma(\cdot)$ and $\gamma(\cdot,\cdot)$ are respectively the gamma function and the lower incomplete gamma function. We are therefore able to numerically compute the value of δ_0 as a function of τ with relative ease. This is shown in Figure 6. We see that, unsurprisingly, when τ is around 2 there is little to be gained even with an oracle sparse estimator over standard least squares estimation. When $\tau = 1$ (Laplace distribution) the value of $\delta_0 \approx 0.15$, indicating that when subsampling by a factor of roughly 7 the least squares estimator will be superior. At this level of undersampling the relative error is a very poor: 0.85 - that is a performance of 0.7dB in terms of traditional Signal to Distortion Ratio (SDR).

The critical undersampling value steadily drops as τ tends towards zero and the distribution becomes increasingly leptokurtic. Thus data distributed according to the Generalized Gaussian for small $\tau \ll 1$ may still be a reasonable candidate for compressive sensing priors as long as the undersampling rate is kept significantly above the associated δ_0.

5.3. *Expected Relative Error for the Laplace distribution.* We conclude this section by examining in more detail the performance of the estimators for Laplace distributed data at various undersampling values. We have already seem from Figure 6 that the oracle

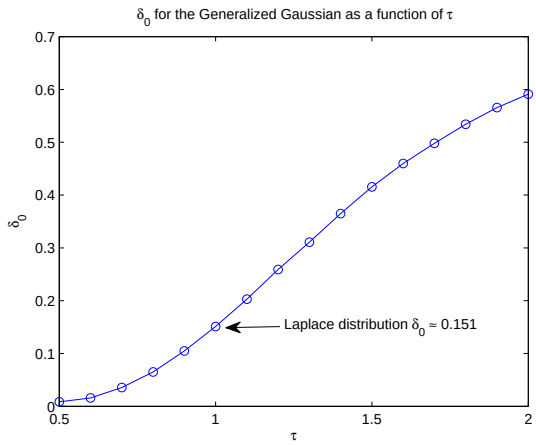


FIG 6. *A plot of the critical subsampling rate, δ_0 below which the Generalized Gaussian distribution is incompressible as a function of the shape parameter, τ.*

performance is poor when subsampling by roughly a factor of 7. What about more modest subsampling factors? Figure 1 plots the relative error as a function of undersampling rate, δ. The horizontal lines indicate SDR values of 3dB, 10dB and 20dB. Thus for the oracle estimator to achieve 10dB the undersampling rate must be greater than 0.7, while to achieve a performance level of 20dB, something that might reasonably be expected in many sensing applications, we can hardly afford any subsampling at all since this requires $\delta > 0.9$.

At this point we should remind the reader that these performance results are for the comparison between the *oracle* sparse estimator and linear least squares. For practically implementable reconstruction algorithms we would expect that the critical undersampling rate at which least squares wins would be significantly higher. Indeed, as shown in Figure 1, this is what is empirically observed for the average performance of the ℓ^1 estimator (8) applied to Laplace distributed data. This curve was calculated at various values of δ by averaging the relative error of 5000 ℓ^1 reconstructions of independent Laplace distributed realizations of $\mathbf{x}_N$ with $N = 256$. In particular note that the ℓ^1 estimator only outperforms least squares for undersampling δ above approximately 0.65!

6. Concluding discussion. As we have just seen, Generalized Gaussian distributions are incompressible at low subsampling rates because their fourth moment is always finite. This confirms the results of Cevher obtained with a different approach [5], but may come as a surprise: for $0 < \tau \leq 1$ the minimum ℓ^τ norm solution to $\mathbf{y} = \mathbf{\Phi x}$, which is also the MAP estimator under the Generalized Gaussian prior, is known to be a good estimator of $\mathbf{x}_0$ when $\mathbf{y} = \mathbf{\Phi x}_0$ and $\mathbf{x}_0$ is compressible [7]. This highlights the need to distinguish between an estimator and its MAP interpretation. In contrast, we describe below a family of statistical distributions $p_{\tau,s}$ which, for certain values of the parameters τ, s, combines:

- superior asymptotic almost sure performance of oracle sparse estimation over least squares reconstruction Δ_{oracle}, even in the largely undersampled scenarios $\delta \to 0$;
- connections between oracle sparse estimation and MAP estimation.

EXAMPLE 4. *For $0 < \tau < \infty$, $1 < s < \infty$ consider the probability density function*

$$p_{\tau,s}(x) \propto (1 + |x|^{\tau})^{-s/\tau}. \tag{38}$$

1. **When $1 < s \leq 3$, the distribution is compressible.**
   Since $\mathbb{E}_{p_{\tau,s}} X^2 = \infty$, Theorem 2 is applicable: the ℓ^1 decoder with a Gaussian encoder has ideal asymptotic performance, even at arbitrary small undersampling $\delta = m/N$;
2. **When $s > 5$, the distribution is incompressible.**
   Since $\mathbb{E}_{p_{\tau,s}} X^4 < \infty$, Theorem 5 is applicable: with a Gaussian encoder, there is an undersampling ratio δ_0 such that whenever $\delta < \delta_0$, the asymptotic almost sure performance of oracle sparse estimation is worse than that of least-squares estimation;
3. **When $3 < s < 5$, the distribution remains somewhat compressible.**
   On the one hand $\mathbb{E}_{p_{\tau,s}} X^2 < \infty$, on the other hand $\mathbb{E}_{p_{\tau,s}} X^4 = \infty$.
   A detailed examination of the $G_1[p_{\tau,s}]$ function shows that there exists a relative number of measures $\delta_0(\tau, s) > 0$ such that in the low measurement regime $\delta < \delta_0$, the asymptotic almost sure performance of oracle of k-sparse estimation, as described in Theorem 4, with the best choice of k, is better *than that of least squares estimation:*

$$H[p_{\tau,s}](\delta) < 1 - \delta, \forall \delta \in (0, \delta_0). \tag{39}$$

Comparing Proposition 4 with the above Example, one observes that both the PDF $p_0(x)$ (Equation (33)) and the PDFs $p_{\tau,s}$, $3 < s < 5$ satisfy $\mathbb{E}_{p_{\tau,s}} X^2 < \infty$ and $\mathbb{E}_{p_{\tau,s}} X^4 = \infty$. Yet, while p_0 is essentially incompressible, the PDFs $p_{\tau,s}$ in this range are compressible. This indicates that, for priors with finite second moment and infinite fourth moment, compressibility depends not only on the tail of the distribution *but also on their mass around zero.*

For $\tau = 2$, the PDF $p_{2,s}$ is a Student-t distribution. For $\tau = 1$, it is called a generalized Pareto distribution. These have been considered in [5, 1] as examples of "compressible" distributions, with the added condition that $s \leq 2$. Such a restriction results from the use of $\ell^2 - \ell^1$ instance optimality in [5, 1], which implies that sufficient compressibility conditions can only be satisfied when $\mathbb{E}_p|X| = \infty$. Here instead we exploit $\ell^2 - \ell^2$ instance optimality *in probability*, making it possible to obtain compressibility when $\mathbb{E}X^2 = \infty$. In other words, [5, 1] provides *sufficient* conditions on a PDF p to check its compressibility, but is inconclusive in characterizing their incompressibility.

The family of PDFs, $p_{\tau,s}$ in the range $0 < \tau \leq 1$, can also be linked with a sparsity inducing MAP estimate. Specifically for an observation $\mathbf{y} = \mathbf{\Phi x}$ of a given vector $\mathbf{x} \in \mathbb{R}^N$, one can define the MAP estimate under the probabilistic *model* where all entries of $\mathbf{x}$ are

considered as iid distributed according to $p_{\tau,s}$:

$$\Delta_{\mathrm{MAP}}(\mathbf{y}) := \arg \max_{\mathbf{x}|\mathbf{\Phi}\mathbf{x}=\mathbf{y}} \prod_{n=1}^{N} p_{\tau,s}(x_n) = \operatorname*{argmin}_{\mathbf{x}|\mathbf{\Phi}\mathbf{x}=\mathbf{y}} \sum_{n=1}^{N} f_\tau(|x_n|).$$

where for $t \in \mathbb{R}^+$ we define $f_\tau(t) := \log(1+t^\tau) = a_{\tau,s} - b_{\tau,s} \log p_{\tau,s}(|t|)$. One can check that the function f_τ is associated to an admissible f-norm as described in [14, 15]: $f(0) = 0$, $f(t)$ is non-decreasing, $f(t)/t$ is non-increasing (in addition, we have $f(t) \sim_{t\to 0} \cdot t^\tau$). Observing that the MAP estimate is a "minimum f-norm" solution to the linear problem $\mathbf{y} = \mathbf{\Phi}\mathbf{x}$, we can conclude that whenever $\mathbf{x}$ is a "sufficiently (exact) sparse" vector, we have in fact [14, 15] $\Delta_{\mathrm{MAP}}(\mathbf{\Phi}\mathbf{x}) = \mathbf{x}$, and $\Delta_{\mathrm{MAP}}(\mathbf{\Phi}\mathbf{x}) = \Delta_1(\mathbf{\Phi}\mathbf{x})$ is also the minimum ℓ^1 norm solution to $\mathbf{y} = \mathbf{\Phi}\mathbf{x}$, which can in turn be "interpreted" as the MAP estimate under the iid Laplace model. However, unlike the Laplace interpretation of ℓ^1 minimization, here Example 4 indicates that such densities are better aligned to sparse reconstruction techniques. Thus the MAP estimate interpretation here may be more valid.

It would be interesting to determine whether the MAP estimator $\Delta_{\mathrm{MAP}}(\mathbf{\Phi}\mathbf{x})$ for such distributions is in some way close to optimal (i.e. close to the minimum mean squared error solution for $\mathbf{x}$). This would give such estimators a degree of legitimacy from a Bayesian perspective. However, we have *not* shown that the estimator $\Delta_{\mathrm{MAP}}(\mathbf{\Phi}\mathbf{x})$ provides a good estimate for data that is distributed according to $p_{\tau,s}$ since, if $\mathbf{x}$ is a large dimensional typical instance with entries drawn iid from the PDF $p_{\tau,s}(x)$, it is typically *not* exactly sparse, hence the uniqueness results of [14, 15] do not directly apply. One would need to resort to a more detailed robustness analysis in the spirit of [13] to get more precise statements relating $\Delta_{\mathrm{MAP}}(\mathbf{\Phi}\mathbf{x})$ to $\mathbf{x}$.

APPENDIX A: PROOF OF PROPOSITION ??

To prove Proposition 2 we will rely on the following theorem [3][Theorem 2.2].

THEOREM 6. *Suppose that F_Y is a continuous and strictly increasing distribution function on $[a,b]$ where $0 \le a < b \le \infty$, with $F_Y(a) = 0$, $F_Y(b) = 1$. For $\sigma \in (0,\mu)$ where $\mu = \int_a^b y dF(y)$, let $\tau \in (a,b)$ be defined by the equation $\sigma = \int_a^\tau y dF(y)$. Let $s_1, s_2, \ldots$ be a sequence such that $\lim_{N\to\infty} s_N/N = \sigma$, and let $Y_1, Y_2 \ldots$ be iid random variables with distribution function F_Y. Let $Y_{1,N} \le \ldots \le Y_{N,N}$ be the increasing order statistics of $Y_1, \ldots, Y_N$ and let L_N be defined as*

$$L_N = L(N, s_n) := (\max\{\ell \le N, Y_{1,N} + \ldots + Y_{\ell,N} \le s_N\}; L_N = 0 \text{ if } Y_{1,N} > s_N. \tag{40}$$

Then

$$\lim_{N\to\infty} \frac{Y_{L_N,N}}{N} \overset{a.s.}{=} \tau, \qquad \lim_{N\to\infty} \frac{L_N}{N} \overset{a.s.}{=} F_Y(\tau), \qquad \lim_{N\to\infty} \frac{\mathbb{E}(L_N)}{N} = F(\tau). \tag{41}$$

PROOF 1 (Proof of Proposition 2). *We begin by the case where* $\mathbb{E}|X|^q < \infty$. *We consider random variables* X_n *drawn according the PDF* $p(x)$, *and we define the iid non-negative random variables* $Y_n = |X_n|^q$. *They have the distribution function* $F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(|X| \leq y^{1/q}) = \bar{F}(y^{1/q})$, *and we have* $\mu = \mathbb{E}Y = \mathbb{E}|X|^q = \int_0^\infty |x|^q d\bar{F}(x) \in (0, \infty)$. *We define* $\mathbf{x}_N = (X_i)_{n=1}^N$, *and we consider a sequence* k_N *such that* $\lim_{N\to\infty} k_N/N = \kappa \in (0,1)$. *By the assumptions on* F_Y *there is a unique* $\tau_0 \in (0,\infty)$ *such that* $\kappa = 1 - F_Y(\tau_0)$, *and we will prove that*

$$\liminf_{N\to\infty} \frac{\sigma_{k_N}(\mathbf{x}_N)_q^q}{N\mu} \overset{a.s.}{\geq} \frac{\int_0^{\tau_0} y dF_Y(y)}{\mu}, \tag{42}$$

$$\limsup_{N\to\infty} \frac{\sigma_{k_N}(\mathbf{x}_N)_q^q}{N\mu} \overset{a.s.}{\leq} \frac{\int_0^{\tau_0} y dF_Y(y)}{\mu}. \tag{43}$$

The proof of the two bounds is identical, hence we only detail the first one. Fix $0 < \epsilon < \tau_0$ *and define* $\tau = \tau(\epsilon) := \tau_0 - \epsilon$, $\sigma = \sigma(\epsilon) := \int_0^\tau y dF_Y(y)$, *and* $s_N = N\sigma$. *Defining* L_N *as in* (40), *we can apply Theorem 6 and obtain* $\lim_{N\to\infty} \frac{L_N}{N} \overset{a.s.}{=} F_Y(\tau)$. *Since* $\lim_{N\to\infty} \frac{k_N}{N} = 1 - F_Y(\tau_0)$, *it follows that*

$$\lim_{N\to\infty} \frac{N - k_N}{L_N} \overset{a.s.}{=} \frac{F_Y(\tau_0)}{F_Y(\tau)} > 1$$

where we used the fact that F_Y *is strictly increasing and* $\tau < \tau_0$. *In other words, almost surely, we have* $N - k_N > L_N$ *for all large enough* N. *Now remember that by definition*

$$L_N = \max\left\{\ell \leq N, \sigma_{N-\ell}(\mathbf{x}_N)_q^q \leq N\sigma\right\}.$$

As a result, almost surely, for all large enough N, *we have*

$$\sigma_{k_N}(\mathbf{x}_N)_q^q = \sigma_{N-(N-k_N)}(\mathbf{x}_N)_q^q > N\sigma.$$

Now, by the strong law of large number, we also have

$$\lim_{N\to\infty} \frac{\|\mathbf{x}_N\|_q^q}{N\mu} \overset{a.s.}{=} 1,$$

hence we obtain

$$\liminf_{N\to\infty} \frac{\sigma_{k_N}(\mathbf{x}_N)_q^q}{\|\mathbf{x}_N\|_q^q} \overset{a.s.}{\geq} \frac{\sigma}{\mu} = \frac{\int_0^{\tau_0-\epsilon} y dF_Y(y)}{\mu}.$$

Since this holds for any $\epsilon > 0$ *and* F_Y *is continuous, this implies* (42). *The other bound* (43) *is obtained similarly. Since the two match, we get*

$$\lim_{N\to\infty} \frac{\sigma_{k_N}(\mathbf{x}_N)_q^q}{\|\mathbf{x}_N\|_q^q} \overset{a.s.}{=} \frac{\int_0^{\tau_0} y dF_Y(y)}{\mu} = \frac{\int_0^{\tau_0} y dF_Y(y)}{\int_0^\infty y dF_Y(y)}.$$

Since $\kappa = 1 - F_Y(\tau_0) = 1 - \bar{F}(\tau_0^{1/q})$ *we have* $\tau_0 = \left[\bar{F}^{-1}(1-\kappa)\right]^q$. *Since* $F_Y(y) = \bar{F}(y^{1/q})$ *we have* $dF_Y(y) = \frac{1}{q} y^{1/q-1}\bar{p}(y^{1/q})dy$. *As a result*

$$\begin{aligned}\frac{\int_0^{\tau_0} y dF_Y(y)}{\int_0^{\infty} y dF_Y(y)} &= \frac{\int_0^{\left[\bar{F}^{-1}(1-\kappa)\right]^q} y^{1/q}\bar{p}(y^{1/q})dy}{\int_0^{\infty} y^{1/q}\bar{p}(y^{1/q})dy} \\ &\overset{(a)}{=} \frac{\int_0^{\bar{F}^{-1}(1-\kappa)} x\bar{p}(x)x^{q-1}dx}{\int_0^{\infty} x\bar{p}(x)x^{q-1}dx} = \frac{\int_0^{\bar{F}^{-1}(1-\kappa)} x^q\bar{p}(x)dx}{\int_0^{\infty} x^q\bar{p}(x)dx}\end{aligned}$$

where in (a) we used the change of variable $y = x^q$, $x = y^{1/q}$, $dy = qx^{q-1}dx$. *We have proved the result for* $0 < \kappa < 1$, *and we let the reader check that minor modifications yield the results for* $\kappa = 0$ *and* $\kappa = 1$.

Now we consider the case $\mathbb{E}|X|^q = +\infty$. *The idea is to use a "saturated" version* $\tilde{X}$ *of the random variable* X, *such that* $\mathbb{E}|\tilde{X}|^q < \infty$, *so as to use the results proven just above.*

One can easily build a family of smooth saturation functions $f_\eta : [0\ \ +\infty) \to [0\ 2\eta)$, $0 < \eta < \infty$ *with* $f_\eta(t) = t$, *for* $t \in [0, \eta]$, $f_\eta(t) \le t$, *for* $t > \eta$, *and two additional properties:*

1. *each function* $t \mapsto f_\eta(t)$ *is bijective from* $[0, \infty)$ *onto* $[0, 2\eta)$, *with* $f'_\eta(t) > 0$ *for all* t;
2. *each function* $t \mapsto f_\eta(t)/t$ *is monotonically decreasing;*

Denoting $f_\eta(\mathbf{x}) := (f_\eta(x_i))_{i=1}^N$, *by [15, Theorem 5], the first two properties ensure that*

$$\frac{\sigma_k(\mathbf{x})^q}{\|\mathbf{x}\|_q^q} \le \frac{\sigma_k(f_\eta(\mathbf{x}))^q}{\|f_\eta(\mathbf{x})\|_q^q}, \quad \forall 1 \le k \le N,\ \forall \mathbf{x} \in \mathbb{R}^N,\ \forall 0 < \eta, q < \infty. \tag{44}$$

Consider a fixed η *and the sequence of "saturated" random variables* $\tilde{X}_i = f_\eta(|X_i|)$. *They are iid with* $\mathbb{E}|\tilde{X}|^q < \infty$. *Moreover, the first property of* f_η *above ensures that their cdf* $t \mapsto \bar{F}_\eta(t) := \mathbb{P}(f_\eta(|X|) \le t)$ *is continuous and strictly increasing on* $[0\ 2\eta]$, *with* $\bar{F}_\eta(0) = 0$ *and* $\bar{F}_\eta(\infty) = 1$. *Hence, by the first part of Proposition 1 just proven above, we have*

$$\lim_{N\to\infty} \frac{\sigma_{k_N}(f_\eta(\mathbf{x}_N))^q}{\|f_\eta(\mathbf{x}_N)\|_q^q} \overset{a.s.}{=} G_q[\bar{p}_\eta](\kappa) = \frac{\int_0^{\bar{F}_\eta^{-1}(1-\kappa)} x^q\bar{p}_\eta(x)dx}{\int_0^{\infty} x^q\bar{p}_\eta(x)dx} \le \frac{|\bar{F}_\eta^{-1}(1-\kappa)|^q}{\mathbb{E}|f_\eta(X)|^q}. \tag{45}$$

Since $f_\eta(t) \le t$ *for all* t, *we have* $\bar{F}_\eta(t) = \mathbb{P}(f_\eta(|X|) \le t) \ge \mathbb{P}(|X| \le t) = \bar{F}(t)$ *for all* t, *hence* $\bar{F}_\eta^{-1}(1-\kappa) \le \bar{F}^{-1}(1-\kappa)$. *Moreover, since* $f_\eta(t) = t$ *for* $0 \le t \le \eta$, *we obtain* $\mathbb{E}|f_\eta(X)|^q \ge \int_0^\eta x^q\bar{F}(x)dx$. *Combining* (44) *and* (45) *with the above observations we obtain*

$$\limsup_{N\to\infty} \frac{\sigma_{k_N}(\mathbf{x}_N)^q}{\|\mathbf{x}_N\|_q^q} \le \lim_{N\to\infty} \frac{\sigma_{k_N}(f_\eta(\mathbf{x}_N))^q}{\|f_\eta(\mathbf{x}_N)\|_q^q} \overset{a.s.}{\le} \frac{|\bar{F}^{-1}(1-\kappa)|^q}{\int_0^\eta x^q\bar{p}(x)dx} \qquad \forall 0 < \eta < \infty.$$

Since $\mathbb{E}|X|^q = \int_0^\infty x^q\bar{p}(x)dx = \infty$, *the infimum over* η *of the right hand side is zero.*

APPENDIX B: PROOF OF THEOREM ??

We will need concentration bounds for several distributions. For the Chi-square distribution with n degrees of freedom χ^2_n, we will use the following standard result (see, e.g., [2, Proposition 2.2], and the intermediate estimates in the proof of [2, Corollary 2.3]):

PROPOSITION 5. *Let $X \in \mathbb{R}^n$ a standard Gaussian random variable. Then, for any $0 < \epsilon < 1$*

$$\mathbb{P}\big(\|X\|_2^2 \geq n(1-\epsilon)^{-1}\big) \leq e^{-n\cdot c_u(\epsilon)/2}, \qquad c_u(\epsilon) := \frac{\epsilon}{1-\epsilon} + \ln(1-\epsilon) \tag{46}$$

$$\mathbb{P}\big(\|X\|_2^2 \leq n(1-\epsilon)\big) \leq e^{-n\cdot c_l(\epsilon)/2}, \qquad c_l(\epsilon) := -\ln(1-\epsilon) - \epsilon. \tag{47}$$

Note that

$$\epsilon^2/2 \leq c_l(\epsilon) \leq c_u(\epsilon), \quad 0 < \epsilon < 1. \tag{48}$$

Its corollary, which provides concentration for projections of random variables from the unit sphere, will also be useful. The statement is obtained by adjusting [2, Lemma 3.2] and [2, Corollary 3.4] keeping the sharper estimate from above.

COROLLARY 1. *Let X be a random vector uniformly distributed on the unit sphere in $\mathbb{R}^n$, and let X_L be its orthogonal projection on a k-dimensional subspace L (alternatively, let X be an arbitrary random vector and L be a random k-dimensional subspace uniformly distributed on the Grassmannian manifold). For any $0 < \epsilon < 1$ we have*

$$\mathbb{P}\big(\sqrt{\frac{n}{k}}\|X_L\|_2 \geq \|X\|_2(1-\epsilon)^{-1}\big) \leq e^{-k\cdot c_u(\epsilon)/2} + e^{-n\cdot c_l(\epsilon)/2}, \tag{49}$$

$$\mathbb{P}\big(\sqrt{\frac{n}{k}}\|X_L\|_2 \leq \|X\|_2(1-\epsilon)\big) \leq e^{-k\cdot c_l(\epsilon)/2} + e^{-n\cdot c_u(\epsilon)/2}. \tag{50}$$

The above result directly implies the concentration inequality (25) for the LS estimator mentioned in Section 4. We will also need a result about Wishart matrices. The Wishart distribution [17] $\mathcal{W}_\ell(n,\Sigma)$ is the distribution of $\ell \times \ell$ matrices $A = Z^T Z$ where Z is an $n \times \ell$ matrix whose columns have the normal distribution $\mathcal{N}(0,\Sigma)$.

THEOREM 7 ([17] [Theorem 3.2.12 and consequence, p. 97-98]). *If A is $\mathcal{W}_\ell(n,\Sigma)$ wheren $-\ell+1 > 0$, and if $Z \in \mathbb{R}^\ell$ is a random vector distributed independently of A and with $P(Z=0)=0$, then the ratio $Z^T\Sigma^{-1}Z/Z^TA^{-1}Z$ follows a Chi-square distribution with $n-\ell+1$ degrees of freedom $\chi^2_{n-\ell+1}$, and is independent of Z. Moreover, if $n-\ell-1>0$ then*

$$\mathbb{E}A^{-1} = \Sigma^{-1}\cdot(n-\ell-1)^{-1}; \quad \mathbb{E}\left\{\operatorname{Trace} A^{-1}\right\} = \operatorname{Trace}\Sigma^{-1}\cdot(n-\ell-1)^{-1}. \tag{51}$$

Finally, for convenience we formalize below some useful but simple facts that we let the reader check.

LEMMA 2. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be two independent* $m \times k$ *and* $m \times \ell$ *random Gaussian matrices with iid entries* $\mathcal{N}(0, 1/m)$*, and let* $x \in \mathbb{R}^\ell$ *be a random vector independent from* $\mathbf{B}$*. Consider a singular value decomposition (SVD)* $\mathbf{A} = U\Sigma V$ *and let* u_ℓ *be the columns of* U*. Define* $w := \mathbf{B}x/\|\mathbf{B}x\|_2 \in \mathbb{R}^m$*,* $w_1 := (\langle u_\ell, w\rangle)_{\ell=1}^k \in \mathbb{R}^k$*,* $w_2 := w_1/\|w_1\|_2 \in \mathbb{R}^k$ *and* $w_3 := V^T w_2 \in \mathbb{R}^k$*. We have*

1. *w is uniformly distributed on the sphere in $\mathbb{R}^m$, and statistically independent from $\mathbf{A}$;*
2. *the distribution of w_1 is rotationally invariant in $\mathbb{R}^k$, and it is statistically independent from $\mathbf{A}$;*
3. *w_2 is uniformly distributed on the sphere in $\mathbb{R}^k$, and statistically independent from $\mathbf{A}$;*
4. *w_3 is uniformly distributed on the sphere in $\mathbb{R}^k$, and statistically independent from $\mathbf{A}$.*

We can now start the proof. For any index set J, we denote $\mathbf{x}_J$ the vector which is zero out of J. For matrices, the notation $\mathbf{\Phi}_J$ indicates the sub-matrix of $\mathbf{\Phi}$ made of the columns indexed by J. The notation $\bar{J}$ stands for the complement of the set J. For any index set Λ associated to linearly independent columns of $\mathbf{\Phi}_\Lambda$ we can write $\mathbf{y} = \mathbf{\Phi}_\Lambda \mathbf{x}_\Lambda + \mathbf{\Phi}_{\bar{\Lambda}} \mathbf{x}_{\bar{\Lambda}}$ hence

$$\begin{aligned} \Delta_{\text{oracle}}(\mathbf{y}, \Lambda) &:= \mathbf{\Phi}_\Lambda^+ \mathbf{y} = \mathbf{x}_\Lambda + \mathbf{\Phi}_\Lambda^+ \mathbf{\Phi}_{\bar{\Lambda}} \mathbf{x}_{\bar{\Lambda}} \\ \|\Delta_{\text{oracle}}(\mathbf{y}, \Lambda) - \mathbf{x}\|_2^2 &= \|\mathbf{\Phi}_\Lambda^+ \mathbf{\Phi}_{\bar{\Lambda}} \mathbf{x}_{\bar{\Lambda}}\|_2^2 + \|\mathbf{x}_{\bar{\Lambda}}\|_2^2 \end{aligned} \tag{52}$$

The last equality comes from the fact that the restriction of $\Delta_{\text{oracle}}(\mathbf{y}, \Lambda) - \mathbf{x}$ to the indices in Λ is $\mathbf{\Phi}_\Lambda^+ \mathbf{\Phi}_{\bar{\Lambda}} \mathbf{x}_{\bar{\Lambda}}$, while its restriction to $\bar{\Lambda}$ is $\mathbf{x}_{\bar{\Lambda}}$. Denoting

$$w := \frac{\mathbf{\Phi}_{\bar{\Lambda}} \mathbf{x}_{\bar{\Lambda}}}{\|\mathbf{\Phi}_{\bar{\Lambda}} \mathbf{x}_{\bar{\Lambda}}\|_2} \in \mathbb{R}^m \tag{53}$$

we obtain the relation

$$\frac{\|\Delta_{\text{oracle}}(\mathbf{y}, \Lambda) - \mathbf{x}\|_2^2}{\|\mathbf{x}_{\bar{\Lambda}}\|_2^2} = \underbrace{\|\mathbf{\Phi}_\Lambda^+ w\|_2^2}_{A} \cdot \underbrace{\frac{\|\mathbf{\Phi}_{\bar{\Lambda}} \mathbf{x}_{\bar{\Lambda}}\|_2^2}{\|\mathbf{x}_{\bar{\Lambda}}\|_2^2}}_{B} + 1. \tag{54}$$

From the singular value decomposition

$$\mathbf{\Phi}_\Lambda = U_m \cdot \begin{bmatrix} \Sigma_k \\ 0_{(m-k)\times k} \end{bmatrix} \cdot V_k,$$

where U_m is an $m \times m$ unitary matrix with columns u_ℓ, and V_k is a $k \times k$ unitary matrix, we deduce that $\mathbf{\Phi}_\Lambda^+ = V_k^T[\Sigma_k^{-1}, 0_{k\times(m-k)}]U_m^T$ and

$$\|\mathbf{\Phi}_\Lambda^+ w\|_2^2 = \|[\Sigma_k^{-1} 0_{k\times(m-k)}]U_m^T w\|_2^2 = \sum_{\ell=1}^{k} \sigma_\ell^{-2} |\langle u_\ell, w\rangle|^2. \tag{55}$$

Since $\mathbf{\Phi}_{\bar\Lambda}$ and $\mathbf{x}_{\bar\Lambda}$ are statistically independent, the random vector $\mathbf{\Phi}_{\bar\Lambda}\mathbf{x}_{\bar\Lambda} \in \mathbb{R}^m$ is Gaussian with zero-mean and covariance $m^{-1} \cdot \|\mathbf{x}_{\bar\Lambda}\|_2^2 \cdot \mathbf{Id}_m$. Therefore,

$$\mathbb{E}\left\{\|\mathbf{\Phi}_{\bar\Lambda}\mathbf{x}_{\bar\Lambda}\|_2^2 / \|\mathbf{x}_{\bar\Lambda}\|_2^2\right\} = 1 \tag{56}$$

and by Proposition 5, for any $0 < \epsilon_0 < 1$

$$\mathbb{P}\left(1 - \epsilon_0 \le \|\mathbf{\Phi}_{\bar\Lambda}\mathbf{x}_{\bar\Lambda}\|_2^2 / \|\mathbf{x}_{\bar\Lambda}\|_2^2 \le (1-\epsilon_0)^{-1}\right) \ge 1 - 2 \cdot e^{-m \cdot c_l(\epsilon_0)/2}. \tag{57}$$

Moreover, by Lemma 2-item 2, the random variables $\langle u_\ell, w\rangle$, $1 \le \ell \le k$ are identically distributed and independent from the random singular values σ_ℓ. Therefore,

$$\mathbb{E}\|\mathbf{\Phi}_\Lambda^+ w\|_2^2 = \mathbb{E}\left\{\sum_{\ell=1}^{k} \sigma_\ell^{-2}\right\} \cdot \mathbb{E}\{|\langle u, w\rangle|^2\} = \mathbb{E}\left\{\mathrm{Trace}(\mathbf{\Phi}_\Lambda^T \mathbf{\Phi}_\Lambda)^{-1}\right\} \cdot \frac{1}{m}.$$

The matrix $\mathbf{\Phi}_\Lambda^T \mathbf{\Phi}_\Lambda$ is $\mathcal{W}_k(m, \frac{1}{m}\mathbf{Id}_k)$ hence, by Theorem 7, when $m - k - 1 > 0$ we have

$$\mathbb{E}\|\mathbf{\Phi}_\Lambda^+ w\|_2^2 = \mathrm{Trace}(m\mathbf{Id}_k) \cdot (m-k-1)^{-1} \cdot m^{-1} = k(m-k-1)^{-1}. \tag{58}$$

Now, considering $w_1 := (\langle u_\ell, w\rangle)_{\ell=1}^k \in \mathbb{R}^k$, $w_2 := w_1/\|w_1\|_2$ and $w_3 := V_k^T w_2$, we obtain

$$\begin{aligned}\|\mathbf{\Phi}^+ w\|_2^2 &= \|\Sigma_k^{-1} w_1\|_2^2 = \|w_1\|_2^2 \cdot \|\Sigma_k^{-1} w_2\|_2^2 = \|w_1\|_2^2 \cdot \|\Sigma_k^{-1} V_k w_3\|_2^2 \\ &= \|w_1\|_2^2 \cdot w_3^T (\mathbf{\Phi}_\Lambda^T \mathbf{\Phi}_\Lambda)^{-1} w_3 = m\|w_1\|_2^2 / R(w_3),\end{aligned}$$

where $R(w_3) := m\|w_3\|_2^2 / w_3^T(\mathbf{\Phi}_\Lambda^T\mathbf{\Phi}_\Lambda)^{-1}w_3 = w_3^T(m^{-1}\mathbf{Id}_k)^{-1}w_3 / w_3^T(\mathbf{\Phi}_\Lambda^T\mathbf{\Phi}_\Lambda)^{-1}w_3$. By Lemma 2-item 4, w_3 is statistically independent from $\mathbf{\Phi}_\Lambda$. As a result, by Theorem 7, the random variable $R(w_3)$ follows a Chi-square distribution with $m-k+1$ degrees of freedom χ^2_{m-k+1}, and by Proposition 5, for any $0 < \epsilon_1 < 1$,

$$\mathbb{P}\Big(1 - \epsilon_1 \le R(w_3)^{-1} \cdot (m-k+1) \le (1-\epsilon_1)^{-1}\Big) \ge 1 - 2e^{-(m-k+1)\cdot c_l(\epsilon_1)/2}. \tag{59}$$

Moreover, since w_1 is a random k-dimensional orthogonal projection of the unit vector w, by Corollary 1, for any $0 < \epsilon_2 < 1$

$$\mathbb{P}\Big(1 - \epsilon_2 \le m\|w_1\|_2^2/k \le (1-\epsilon_2)^{-1}\Big) \ge 1 - 4e^{-k\cdot c_l(\epsilon_2)/2}. \tag{60}$$

To conclude, since $\mathbf{\Phi}_{\bar{\Lambda}}\mathbf{x}_{\bar{\Lambda}}$ is Gaussian, its ℓ^2-norm $\|\mathbf{\Phi}_{\bar{\Lambda}}\mathbf{x}_{\bar{\Lambda}}\|_2^2$ and direction w are mutually independent, hence $\|\mathbf{\Phi}_{\Lambda}^{+}w\|_2^2$ and $\|\mathbf{\Phi}_{\bar{\Lambda}}\mathbf{x}_{\bar{\Lambda}}\|_2^2$ are also mutually independent. Therefore, we can combine the decomposition (54) with the expected values (56) and (58) to obtain

$$\frac{\mathbb{E}\|\Delta_{\text{oracle}}(\mathbf{y},\Lambda)-\mathbf{x}\|_2^2}{\|\mathbf{x}_{\bar{\Lambda}}\|_2^2} = \mathbb{E}\|\mathbf{\Phi}_{\Lambda}^{+}w\|_2^2 \cdot \frac{\mathbb{E}\|\mathbf{\Phi}_{\bar{\Lambda}}\mathbf{x}_{\bar{\Lambda}}\|_2^2}{\|\mathbf{x}_{\bar{\Lambda}}\|_2^2} + 1 = \frac{k}{m-k-1} + 1 = \frac{1}{1-\frac{k}{m-1}}.$$

We conclude that: for any index set Λ of size at most k, with $k < m-1$, in expectation

$$\begin{aligned}\frac{\mathbb{E}\|\Delta_{\text{oracle}}(\mathbf{y},\Lambda)-\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} &= \frac{\mathbb{E}\|\Delta_{\text{oracle}}(\mathbf{y},\Lambda)-\mathbf{x}\|_2^2}{\|\mathbf{x}_{\bar{\Lambda}}\|_2^2} \cdot \frac{\|\mathbf{x}_{\bar{\Lambda}}\|_2^2}{\|\mathbf{x}\|_2^2} = \frac{1}{1-\frac{k}{m-1}} \cdot \frac{\|\mathbf{x}_{\bar{\Lambda}}\|_2^2}{\|\mathbf{x}\|_2^2} \\ &\geq \frac{1}{1-\frac{k}{m-1}} \cdot \frac{\sigma_k(\mathbf{x})_2^2}{\|\mathbf{x}\|_2^2}.\end{aligned}$$

In terms of concentration, combining (57), (59), and (60), we get that for $0 < \epsilon_0, \epsilon_1, \epsilon_2 < 1$:

$$(1-\epsilon_0)(1-\epsilon_1)(1-\epsilon_2) \leq \|\mathbf{\Phi}_{\Lambda}^{+}w\|_2^2 \cdot \frac{\|\mathbf{\Phi}_{\bar{\Lambda}}\mathbf{x}_{\bar{\Lambda}}\|_2^2}{\|\mathbf{x}_{\bar{\Lambda}}\|_2^2} \cdot \frac{m-k+1}{k} \leq (1-\epsilon_0)^{-1}(1-\epsilon_1)^{-1}(1-\epsilon_2)^{-1}$$

except with probability at most (setting $\epsilon_i = \epsilon$, $i = 0,1,2$)

$$2 \cdot e^{-m\cdot c_l(\epsilon_0)/2} + 4 \cdot e^{-k\cdot c_l(\epsilon_2)/2} + 2 \cdot e^{-(m-k+1)\cdot c_l(\epsilon_1)/2} \leq 8 \cdot e^{-\min(k,m-k+1)\cdot c_l(\epsilon)/2}.$$

APPENDIX C: PROOF OF THEOREM ??

Remember that we are considering sequences $k_N, m_N, \mathbf{\Phi}_N, \Lambda_N, \mathbf{x}_N$. Denoting $\rho_N = k_N/m_N$ and $\delta_N = m_N/N$, we observe that the probability (27) can be expressed as $1 - 8e^{-N\cdot c_N(\epsilon)/2}$ where $c_N(\epsilon) = c_l(\epsilon)\cdot\delta_N\cdot\min(\rho_N, 1-\rho_N)$. For any choice of ϵ, we have

$$\lim_{N\to\infty} c_N(\epsilon) = c_l(\epsilon)\cdot\delta\cdot\min(\rho, 1-\rho) > 0,$$

hence $\sum_N e^{-N\cdot c_N(\epsilon)/2} < \infty$ and we obtain that for any $\eta > 0$

$$\sum_N \mathbb{P}\left(\left|\left(\frac{\|\Delta_{\text{oracle}}(\mathbf{y}_N,\Lambda_N)-\mathbf{x}_N\|_2^2}{\sigma_{k_N}(\mathbf{x}_N)_2^2} - 1\right)\cdot\frac{m_N-k_N+1}{k_N} - 1\right| \geq \eta\right) < \infty.$$

This implies [11, Corollary 4.6.1] the almost sure convergence

$$\lim_{N\to\infty}\left(\frac{\|\Delta_{\text{oracle}}(\mathbf{y}_N,\Lambda_N)-\mathbf{x}_N\|_2^2}{\sigma_{k_N}(\mathbf{x}_N)_2^2} - 1\right)\cdot\frac{m_N-k_N+1}{k_N} \overset{a.s.}{=} 1.$$

Finally, since $k_N/m_N = \rho_N \to \rho$ and $\delta_N \to \delta$, we also have

$$\lim_{N\to\infty}\frac{k_N}{m_N-k_N+1} = \frac{\rho}{1-\rho}$$

and we conclude that

$$\lim_{N\to\infty} \frac{\|\Delta_{\text{oracle}}(\mathbf{y}_N, \Lambda_N) - \mathbf{x}_N\|_2^2}{\|\mathbf{x}_N\|_2^2} \overset{a.s.}{=} \frac{1}{1-\rho} \lim_{N\to\infty} \frac{\sigma_{k_N}(\mathbf{x}_N)_2^2}{\|\mathbf{x}_N\|_2^2} \overset{a.s.}{=} \frac{G_2[p](\delta\rho)}{1-\rho}.$$

We obtain the result for the least squares decoder by copying the above arguments and starting from (25).

APPENDIX D: PROOF OF THEOREM ??

The proof is based upon the following version of Theorem 5.1 from [8]:

THEOREM 8 (DeVore *et al.* [8]). *Let $\mathbf{\Phi}(\omega) \in \mathbb{R}^{m\times N}$ be a random matrix whose entries are iid and drawn from $\mathcal{N}(0, 1/m)$. There are some absolute constants $C_0, \ldots, C_6$, and C_7 depending on $C_1, \ldots, C_6$ such that, given any $\mathbf{x} \in \mathbb{R}^N$ and any $k \le C_0 m/\log(N/m)$, there is a set $\Omega(\mathbf{x}, k)$ with*

$$\mathbb{P}(\Omega^c(\mathbf{x}, k)) \le C_1 e^{-C_2 m} - e^{-C_3\sqrt{Nm}} - C_4 e^{-C_5 m} - 2me^{-\frac{\sqrt{m}}{C_6 \log(N/m)}} \tag{61}$$

such that

$$\|\mathbf{x} - \Delta_1(\mathbf{\Phi}(\omega)\mathbf{x})\|_2 \le C_7 \sigma_k(\mathbf{x})_2, \quad \textit{for each } \omega \in \Omega(x, k). \tag{62}$$

In this version of the theorem we have specialized to the case where the random matrices are Gaussian distributed. We have also removed the rather perculiar requirement in the original version that $N \ge [\ln 6]^2 m$ as careful scrutiny of the proofs (in particular the proof of Theorem 3.5 [8]) indicates that the effect of this term can be absorbed into the constant C_3 as long as $m/N \le [\frac{2}{\ln 6}]^2 \approx 1.2$, which is trivially satisfied.

We now proceed to prove Theorem 2. By assumption the undersampling ratio $\delta = \lim_{N\to\infty} \frac{m_N}{N} > 0$, therefore there exists a $0 < \kappa < 1$ such that

$$\delta > C_0 \kappa \log \frac{1}{\delta}.$$

Now choosing a sequence $k_N/N \to \kappa$ we have, for large enough N,

$$m_N \ge C_0 k_N \log(N/m_N).$$

Hence, applying Theorem 8, for all N large enough, there exist a set $\Omega_N(\mathbf{x}_N, k_N)$ with

$$\mathbb{P}(\Omega_N^c(\mathbf{x}_N, k_N)) \le C_8 m e^{-C_9\sqrt{m}} \tag{63}$$

such that (62) holds for all $\mathbf{\Phi}_N(\omega) \in \Omega(\mathbf{x}_N, k_N)$, i.e.,

$$\frac{\|\mathbf{x}_N - \Delta_1(\Phi_N(\omega)\mathbf{x}_N)\|_2}{\|\mathbf{x}_N\|_2} \le C_7 \bar{\sigma}_{k_N}(\mathbf{x}_N)_2. \tag{64}$$

A union bound argument similar to the one used in the proof of Theorem 4 (see Appendix C) gives:

$$\limsup_{N\to\infty} \frac{\|\mathbf{x}_N - \Delta_1(\Phi_N \mathbf{x}_N)\|_2}{\|\mathbf{x}_N\|_2} \overset{a.s.}{\leq} \limsup_{N\to\infty} C_7 \bar{\sigma}_{k_N}(\mathbf{x}_N)_2 \overset{a.s.}{=} C_7 G_2[p](\kappa) = 0. \tag{65}$$

APPENDIX E: PROOF OF LEMMA ??

For the first result we assume that $G(\delta^2) \leq (1-\delta)^2$. We take $\rho = \delta$ and obtain by definition

$$H(\delta) \leq \frac{G(\delta\rho)}{1-\rho} = \frac{G(\delta^2)}{1-\delta} \leq (1-\delta).$$

The second result is a straightforward consequence of the first one. For the last one, we consider $\delta \in (0, \delta_0)$. For any $\rho \in (0,1)$ we set $\kappa := \delta\rho \in (0, \delta_0)$. Since for any pair $a, b \in (0,1)$ we have $(1-a)(1-b) \leq (1-\sqrt{ab})^2$, we have

$$G(\kappa) \geq (1-\sqrt{\kappa})^2 \geq (1-\delta)(1-\rho)$$

and we conclude that

$$\forall\ \rho \in (0,1),\ \frac{G(\delta\rho)}{1-\rho} \geq 1-\delta.$$

APPENDIX F: PROOF OF THEOREM ??

Theorem 1 can be proved from Lemma 1 along with the following result.

LEMMA 3. *Let $p(x)$ be a distribution with finite fourth moment $\mathbb{E}X^4 < \infty$. Then there exists some $\delta_0 \in (0,1)$ such that the function $G_2[p](\kappa)$ as defined in Proposition 2 satisfies*

$$G_2[p](\kappa) \geq (1-\sqrt{\kappa})^2, \qquad \forall \kappa \in (0, \delta_0). \tag{66}$$

PROOF 2 (Proof of Lemma 3). *Without loss of generality we can assume that $p(x)$ has unit second moment, hence*

$$G_2[p](\kappa) := \frac{\int_0^{\bar{F}^{-1}(1-\kappa)} u^2 \bar{p}(u) du}{\int_0^\infty u^2 \bar{p}(u) du} = 1 - \int_\alpha^\infty u^2 \bar{p}(u) du,$$

where we denote $\alpha = \bar{F}^{-1}(1-\kappa)$, which is equivalent to $\kappa = 1 - \bar{F}(\alpha) = \int_\alpha^\infty \bar{p}(u)du$. The inequality $G(\kappa) \geq (1-\sqrt{\kappa})^2$ is equivalent to $2\sqrt{\kappa} \geq 1 + \kappa - G(\kappa)$, that is to say

$$2\sqrt{\int_\alpha^\infty \bar{p}(u)du} \geq \int_\alpha^\infty (u^2+1)\bar{p}(u)du \tag{67}$$

By the Cauchy-Schwarz inequality

$$\int_\alpha^\infty (u^2+1)\bar{p}(u)du \le \sqrt{\int_\alpha^\infty (u^2+1)^2\bar{p}(u)du} \cdot \sqrt{\int_\alpha^\infty \bar{p}(u)du}.$$

Since $\mathbb{E}X^4 < \infty$*, for all small enough* κ *(i.e., large enough* α*), the right hand side is arbitrarily smaller than* $2\sqrt{\int_\alpha^\infty \bar{p}(u)du}$ *hence the inequality* $G(\kappa) \ge (1-\sqrt{\kappa})^2$ *holds true.*

Proof 3 (Proof of Theorem 5). *Theorem 5 now follows by combining Lemma 3 and Lemma 1 to show that for a distribution with finite fourth moment there exists a* $\delta_0 \in (0,1)$ *such that* $H(\delta) \ge 1-\delta$ *for all* $\delta \in (0,\delta_0)$*. The asymptotic almost sure comparative performance of the estimators then follows from the concentration bounds in Theorem 3 and for the least squares estimator.*

APPENDIX G: PROOF OF PROPOSITION ??

Just as in the proof of Lemma 3 above, we denote $\alpha = \bar{F}^{-1}(1-\kappa)$, which is equivalent to $\kappa = 1-\bar{F}(\alpha) = \int_\alpha^\infty \bar{p}(u)du$. We know from Lemma 1 that the identity $H[p](\rho) = 1-\rho$ for all $0<\rho<1$ is equivalent to $G_2[p](\kappa) = (1-\sqrt{\kappa})^2$ for all $0<\kappa<1$. By the same computations as in the proof of Lemma 3, under the unit second moment constraint $\mathbb{E}_{p(x)}X^2 = 1$, the latter is equivalent to

$$2\sqrt{\int_\alpha^\infty \bar{p}(u)du} = \int_\alpha^\infty (u^2+1)\bar{p}(u)du \tag{68}$$

Denote $K(\alpha) := \int_\alpha^\infty (u^2+1)\bar{p}(u)du$. The constraint is $K(\alpha)\cdot K(\alpha) = 4\int_\alpha^\infty \bar{p}(u)du$. Taking the derivative and negating we must have $2K(\alpha)\cdot[(\alpha^2+1)\cdot\bar{p}(\alpha)] = 4\bar{p}(\alpha)$. If $\bar{p}(\alpha)\neq 0$ it follows that $K(\alpha) = 2/(\alpha^2+1)$ hence $(\alpha^2+1)\cdot\bar{p}(\alpha) = -K'(\alpha) = 4\alpha/(\alpha^2+1)^2$ that is to say $\bar{p}(\alpha) = 4\alpha/(\alpha^2+1)^3$ which is satisfied for $p(x) = p_0(x)$. One can check that

$$\int_0^\infty \frac{4\alpha}{(\alpha^2+1)^3}d\alpha = \left[-\frac{1}{(\alpha^2+1)^2}\right]_0^\infty = 1$$

and, since $\bar{p}(\alpha) \asymp 4\alpha^{-5}$, $\mathbb{E}_{p_0(x)}(X^4) = \infty$.

APPENDIX H: PROOF OF THE STATEMENTS IN EXAMPLE ??

Without loss of generality we rescale $p_{\tau,s}(x)$ in the form $p(x) = (1/a)\cdot p_{\tau,s}(x/a)$ so that $p_{\tau,s}$ is a proper PDF with unit variance $\mathbb{E}X^2 = 1$. Observing that $p_{\tau,s}(x) \asymp_{x\to\infty} x^{-s}$, we have: $\mathbb{E}X^2 < \infty$ if, and only if $s>3$; $\mathbb{E}X^4 < \infty$ if, and only if, $s>5$. For large α, $n = 0, 2$, $3 < s < 5$, we obtain

$$\int_\alpha^\infty x^n p(x)dx \asymp \int_\alpha^\infty x^{n-s}dx \asymp \left[\frac{x^{n+1-s}}{n+1-s}\right]_\alpha^\infty \asymp \alpha^{n+1-s}$$

hence, from the relation between κ and α, we obtain

$$\frac{1+\kappa-G_2[p](\kappa)}{2\sqrt{\kappa}} = \frac{\int_\alpha^\infty (u^2+1)p(u)du}{2\sqrt{\int_\alpha^\infty p(u)du}} \asymp \frac{\left(\alpha^{3-s}+\alpha^{1-s}\right)}{\sqrt{\alpha^{1-s}}} \asymp \alpha^{\frac{5-s}{2}}$$

For $3 < s < 5$ we get

$$\lim_{\kappa\to 0} \frac{1+\kappa-G_2[p](\kappa)}{2\sqrt{\kappa}} = \infty$$

hence there exists $\delta_0 > 0$ such that for $\kappa < \sqrt{\delta_0}$

$$G_2[p](\kappa) < 1+\kappa-2\sqrt{\kappa} = (1-\sqrt{\kappa})^2.$$

We conclude using Lemma 1.

APPENDIX I: THE LAPLACE DISTRIBUTION

First we compute $\bar{p}_1(x) = \exp(-x)$ for $x \geq 0$, $\bar{F}_1(z) = 1-e^{-z}$, $z \geq 0$ hence $\bar{F}_1^{-1}(1-\kappa) = -\ln\kappa$. For all integers $q \geq 1$ and $x > 0$, we obtain by integration by parts the recurrence relation

$$\int_0^x u^q e^{-u} du = q\int_0^x u^{q-1}e^{-u}du - x^q e^{-x}, \forall q \geq 1.$$

$\int_0^x e^{-u}du = 1-e^{-x}$, hence for $q=1$ we obtain $\int_0^x ue^{-u}du = 1-e^{-x}-xe^{-x} = 1-(1+x)e^{-x}$, and for $q=2$ it is easy to compute

$$\int_0^x u^2 e^{-u}du = 2-(2+2x+x^2)e^{-x}$$

(7) and (6) follow from substituting these expressions into:

$$G_q[p_1](\kappa) = \frac{\int_0^{-\ln\kappa} u^q \bar{p}_1(u)du}{\int_0^\infty u^q \bar{p}_1(u)du}.$$

REFERENCES

[1] R.G. Baraniuk, V. Cevher, and M.B. Wakin. Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective. *Proceedings of the IEEE*, 98(6):959 –971, June 2010.

[2] Alexander Barvinok. Math 710: Measure concentration. Lecture Notes, 2005.

[3] F. Thomas Bruss and James B. Robertson. 'Wald's lemma' for sums of order statistics of i.i.d. random variables. *Advances in Applied Probability*, 23(3):612–623, sep 1991.

[4] E. J. Candès, J. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math*, 59:1207–1223, 2006.

[5] V. Cevher. Learning with compressible priors. In *NIPS*, Vancouver, B.C., Canada, 7–12 December 2008.

[6] Albert Cohen, Wolfgang Dahmen, and Ronald A. DeVore. Compressed sensing and best k-term approximation. *J. Amer. Math. Soc.*, 22:211–231, 2009.

[7] M. E. Davies and Rémi Gribonval. On lp minimisation, instance optimality, and restricted isometry constants for sparse approximation. In *Proc. SAMPTA'09 (Sampling Theory and Applications)*, Marseille, France, may 2009.

[8] Ronald DeVore, Guergana Petrova, and Przemyslaw Wojtaszczyk. Instance-optimality in probability with an l1-minimization decoder. *Applied and Computational Harmonic Analysis*, 27(3):275 – 288, 2009.

[9] David Donoho and Jared Tanner. Counting faces of randomly-projected polytopes when the projection radically lowers dimension. *Journal of the AMS*, 22(1):1–53, January 2009.

[10] B. Yu G. Chang and M. Vetterli. Adaptive wavelet thresholding for image denoising and compression. *IEEE Trans. Image Proc.*, 9:1532–1546, 2000.

[11] Robert M. Gray. *Probability, Random Processes, and Ergodic Properties*. Springer Publishing Company, Incorporated, 2009.

[12] Rémi Gribonval. Should penalized least squares regression be interpreted as Maximum A Posteriori estimation? Technical Report 7484, INRIA, May 2010. to appear in IEEE Transactions on Signal Processing.

[13] Rémi Gribonval, Rosa Maria Figueras i Ventura, and Pierre Vandergheynst. A simple test to check the optimality of sparse signal approximations. *EURASIP Signal Processing, special issue on Sparse Approximations in Signal and Image Processing*, 86(3):496–510, March 2006.

[14] Rémi Gribonval and Morten Nielsen. On the strong uniqueness of highly sparse expansions from redundant dictionaries. In *Proc. Int Conf. Independent Component Analysis (ICA'04)*, LNCS, Granada, Spain, September 2004. Springer-Verlag.

[15] Rémi Gribonval and Morten Nielsen. Highly sparse representations from dictionaries are unique and independent of the sparseness measure. *Appl. Comput. Harm. Anal.*, 22(3):335–355, May 2007.

[16] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423, July 2001.

[17] Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, 2008.

[18] Mila Nikolova. Model distortions in Bayesian MAP reconstruction. *Inverse Problems and Imaging*, 1(2):399–422, 2007.

[19] David Wipf and Srikantan Nagarajan. A new view of automatic relevance determination. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20 (NIPS)*. MIT Press, 2008.

[20] W. Xu and B. Hassibi. Compressive sensing over the grassmann manifold: a unified geometric framework. Preprint. arXiv:1005.3729v1, 2010.

Rémi Gribonval,
Equipe-projet METISS,
Centre INRIA Rennes – Bretagne Atlantique,
Campus de Beaulieu,
F-35042 Rennes Cedex,
France
E-mail: remi.gribonval@inria.fr

Volkan Cevher,
EPFL STI IEL LIONS
ELE 233 (Bâtiment ELE)
Station 11
CH-1015 Lausanne
Switzerland
Phone number: +41216931101
Fax number: +41 216937600
E-mail: volkan.cevher@epfl.ch

Mike Davies,
Institute for Digital Communications (IDCOM)
& Joint Research Institute for Signal and Image Processing,
School of Engineering and Electronics,
University of Edinburgh,
The King's Buildings,Mayfield Road,
Edinburgh EH9 3JL
UK
E-mail: Mike.Davies@ed.ac.uk