Sparse and spurious: dictionary learning with noise and outliers

Rémi Gribonval, Rodolphe Jenatton, Francis Bach

Sparse and spurious: dictionary learning with noise and outliers

Rémi Gribonval, *IEEE Fellow*, Rodolphe Jenatton, Francis Bach

***Abstract*—A popular approach within the signal processing and machine learning communities consists in modelling signals as sparse linear combinations of atoms selected from a *learned* dictionary. While this paradigm has led to numerous empirical successes in various fields ranging from image to audio processing, there have only been a few theoretical arguments supporting these evidences. In particular, sparse coding, or sparse dictionary learning, relies on a non-convex procedure whose local minima have not been fully analyzed yet. In this paper, we consider a probabilistic model of sparse signals, and show that, with high probability, sparse coding admits a local minimum around the reference dictionary generating the signals. Our study takes into account the case of over-complete dictionaries, noisy signals, and possible outliers, thus extending previous work limited to noiseless settings and/or under-complete dictionaries. The analysis we conduct is non-asymptotic and makes it possible to understand how the key quantities of the problem, such as the coherence or the level of noise, can scale with respect to the dimension of the signals, the number of atoms, the sparsity and the number of observations.**

I. Introduction

Modelling signals as sparse linear combinations of atoms selected from a dictionary has become a popular paradigm in many fields, including signal processing, statistics, and machine learning. This line of research has witnessed the development of several well-founded theoretical frameworks (see, e.g., [44, 45]) and efficient algorithmic tools (see, e.g., [7] and references therein).

However, the performance of such approaches hinges on the representation of the signals, which makes the question of designing "good" dictionaries prominent. A great deal of effort has been dedicated to come up with efficient *predefined* dictionaries, e.g., the various types of wavelets [29]. These representations have notably contributed to many successful image processing applications such as compression, denoising and deblurring. More recently, the idea of simultaneously *learning* the dictionary and the sparse decompositions of the signals—also known as *sparse dictionary learning*, or simply, *sparse coding*—has emerged as a powerful framework, with state-of-the-art performance in many tasks, including inpainting and image classification (see, e.g., [28] and references therein).

Although sparse dictionary learning can sometimes be formulated as convex [6, 9], non-parametric Bayesian [47] and submodular [27] problems, the most popular and widely used definition of sparse coding brings into play a non-convex optimization problem. Despite its empirical and practical success, the theoretical analysis of the properties of sparse dictionary learning is still in its infancy. A recent line of work [31, 41, 32] establishes generalization bounds which quantify how much the *expected* signal-reconstruction error differs from the *empirical* one, computed from a random and finite-size sample of signals. In particular, the bounds obtained by Maurer and Pontil [31], Vainsencher et al. [41], Gribonval et al. [21] are non-asymptotic, and uniform with respect to the whole class of dictionaries considered (e.g., those with normalized atoms).

a) Dictionary identifiability.: This paper focuses on a complementary theoretical aspect of dictionary learning: the characterization of local minima of an optimization problem associated to sparse coding, in spite of the non-convexity of its formulation. This problem is closely related to the question of *identifiability*, that is, whether it is possible to *recover* a reference dictionary that is assumed to generate the observed signals. Identifying such a dictionary is important when the interpretation of the learned atoms matters, e.g., in source localization [12], where the dictionary corresponds to the so-called mixing matrix indicating directions of arrival, in topic modelling [24], where the atoms of the dictionary are expected to carry semantic information, or in neurosciences, where learned atoms have been related to the properties of the visual cortex in the pioneering work of Field and Olshausen [14].

In fact, characterizing how accurately one can estimate a dictionary through a given learning scheme also matters beyond such obvious scenarii where the dictionary intrinsically carries information of interest. For example, when learning a dictionary for coding or denoising, two dictionaries are considered as perfectly equivalent if they lead to the same distortion-rate curve, or the same denoising performance. In such contexts, learning an ideal dictionary through the direct optimization of the idealized performance measure is likely to be intractable, and it is routinely replaced by heuristics involving the minimization of proxy, i.e., a better behaved

This is a substantially revised version of a first draft that appeared as a preprint titled "Local stability and robustness of sparse dictionary learning in the presence of noise", [25].

This work was supported in part by the EU FET- Open programme through the SMALL Project under Grant 225913 and in part by the European Research Council through the PLEASE Project (ERC-StG-2011-277906) and the SIERRA project (ERC-StG-2011-239993).

R. Gribonval is with the Institut de Recherche en Systèmes Aléatoires (Inria & CNRS UMR 6074), Rennes 35042, France (email: remi.gribonval@inria.fr).

R. Jenatton was with the Laboratoire d'Informatique, École Normale Supérieure, Paris 75005, France. He is now the Amazon Development Center Germany, Berlin 10178, Germany (e-mail: jenatton@amazon.com).

F. Bach is with the Laboratoire d'Informatique, École Normale Supérieure, Paris 75005, France (e-mail: francis.bach@ens.fr).

cost function. Characterizing (local) minima of the proxy is likely to help in providing guarantees that such minima exist close to those of the idealized performance measure and, more importantly, that they also achieve near-optimal performance.

b) Contributions and related work.: In contrast to early identifiability results in this direction by Georgiev et al. [18], Aharon et al. [4], which focused on deterministic but combinatorial identifiability conditions with combinatorial algorithms, Gribonval and Schnass [19] pioneered the analysis of identifiability using a non-convex objective involving an ℓ^1 criterion, in the spirit of the cost function initially proposed by Zibulevsky and Pearlmutter [48] in the context of blind signal separation. In the case where the reference dictionary forms a basis, they obtained local identifiability results with noiseless random k-sparse signals, possibly corrupted by some "mild" outliers naturally arising with the considered Bernoulli-Gaussian model. Still in a noiseless setting and without outliers, with a k-sparse Gaussian signal model, the analysis was extended by Geng et al. [17] to *over-complete* dictionaries, *i.e.*, dictionaries composed of more atoms than the dimension of the signals. Following these pioneering results, a number of authors have established theoretical guarantees on sparse coding that we summarize in Table I. Most of the existing results do not handle noise, and none handles outliers. In particular, the structure of the proofs of Gribonval and Schnass [19], Geng et al. [17], hinges on the absence of noise and cannot be straightforwardly transposed to take into account some noise.

In this paper, we analyze the local minima of sparse coding *in the presence of noise and outliers*. For that, we consider sparse coding with a regularized least-square cost function involving an ℓ^1 penalty, under certain incoherence assumptions on the underlying ground truth dictionary and appropriate statistical assumptions on the distribution of the training samples. To the best of our knowledge, this is the first analysis which relates to the widely used sparse coding objective function associated to the online learning approach of Mairal et al. [28]. In contrast, most of the emerging work on dictionary identifiability considers either an objective function based on ℓ^1 minimization under equality constraints [19, 17], for which there is no known efficient heuristic implementation, or on an ℓ^0 criterion [35] *à la* K-SVD [4]. More algorithmic approaches have also recently emerged [37, 5] demonstrating the existence of provably good (sometimes randomized) algorithms of polynomial complexity for dictionary learning. Agarwal et al. [2] combine the best of both worlds by providing a polynomial complexity algorithm based on a clever randomized clustering initialization [3, 5] followed by alternate optimization based on an ℓ^1 minimization principle with equality constraints. While this is a definite theoretical breakthrough, these algorithms are yet to be tested on practical problems, while on open source implementation (SPAMS[1]) of the online learning approach of Mairal et al. [28] is freely available and has been extensively exploited on practical datasets over a range of applications.

c) Main contributions.: Our main contributions can be summarized as follows:

1) We consider the recovery of a dictionary with p atoms $\mathbf{D}^o \in \mathbb{R}^{m\times p}$ using ℓ_1-penalized formulations with penalty factor $\lambda > 0$, given a training set of n signals gathered in a data matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$. This is detailed in Section II-A.
2) We assume a general probabilistic model of sparse signals, where the data matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ is generated as $\mathbf{D}^o\mathbf{A}^o$ plus additive noise $\boldsymbol{\varepsilon}$. Our model, described in Section II-C, corresponds to a k-sparse support with loose decorrelation assumptions on the nonzero coefficients. It is closely connected to the $\Gamma_{k,C}$ model of Arora et al. [5, Definition 1.2]. In particular, unlike in independent component analysis (ICA) and in most related work, *no independence is assumed between nonzero coefficients*.
3) We show that under deterministic (cumulative) coherence-based sparsity assumptions (see Section II-D) the minimized cost function has a guaranteed local minimum around the generating dictionary $\mathbf{D}^o$ with high probability.
4) We also prove support and coefficient recovery, which is important for blind source separation.
5) Our work makes it possible to better understand:
   a) how small the neighborhood around the reference dictionary can be, i.e., tending to zero as the noise variance goes to zero.
   b) how many signals n are sufficient to hope for the existence of such a controlled local minimum, i.e., $n = \Omega(mp^3)$. In contrast to several recent results [35, 5, 36] where the sample complexity depends on the targeted resolution r such that $\|\hat{\mathbf{D}} - \mathbf{D}^o\| \leqslant r$, our main sample complexity estimates are *resolution-independent* in the noiseless case. This is similar in nature to the better sample complexity results $n = \Omega(p^2 \log p)$ obtained by Agarwal et al. [2] for a polynomial algorithm in a noiseless context, or $n = \Omega(p \log mp)$ obtained by Agarwal et al. [3] for Rademacher coefficients. This is achieved through a precise sample complexity analysis using Rademacher averages and Slepian's lemma. In the presence of noise, a factor $1/r^2$ seems unavoidable [35, 5, 36].
   c) what sparsity levels are admissible. Our main result is based on the cumulative coherence (see Section II-D) $\mu_k(\mathbf{D}^o) \leqslant 1/4$. It also involves a condition that restricts our analysis to overcomplete dictionaries where $p \lesssim m^2$, where previous works seemingly apply to very overcomplete settings. Intermediate results only involve restricted isometry properties. This may allow for much larger values of the sparsity level k, and more overcompleteness, but this is left to future work.
   d) what level of noise and outliers appear as manageable, with a precise control of the admissible "energy" of these outliers. While a first naive analysis would suggest a tradeoff between the presence of outliers and the targeted resolution r, we conduct a tailored analysis that demonstrates the existence of a *resolution-independent* threshold on the relative amount of outliers to which the approach is robust.

[1] http://spams-devel.gforge.inria.fr/

Reference	Overcomplete	Noise	Outliers	Global min / algorithm	Polynomial algorithm	Exact (no noise, no outlier, n finite)	Sample complexity (no noise)	Admissible sparsity for *exact* recovery	Coefficient model (main characteristics)
Georgiev et al. [18] *Combinatorial approach*	✓	✗	✗	✓	✗	✓	$m\binom{p}{m-1}$	$k = m-1$, $\underline{\delta}_m(\mathbf{D}^o) < 1$	Combinatorial
Aharon et al. [4] *Combinatorial approach*	✓	✗	✗	✓	✗	✓	$(k+1)\binom{p}{k}$	$\underline{\delta}_{2k}(\mathbf{D}^o) < 1$	Combinatorial
Gribonval and Schnass [19] ℓ^1 *criterion*	✗	✗	✗	✗	✗	✓	$\frac{m^2 \log m}{k}$	$\frac{k}{m} <$ $1 - \|\mathbf{D}^\top \mathbf{D} - \mathbf{I}\|_{2,\infty}$	Bernoulli(k/p) -Gaussian
Geng et al. [17] ℓ^1 *criterion*	✓	✗	✗	✗	✗	✓	kp^3	$O(1/\mu_1(\mathbf{D}^o))$	k-sparse -Gaussian
Spielman et al. [37] ℓ^0 *criterion* *ER-SpUD (***randomized***)*	✗	✗	✗	✓ $P(✓)$	✗ ✓	✓ ✓	$m \log m$ $m^2 \log^2 m$	$O(m)$ $O(\sqrt{m})$	Bernoulli(k/p) -Gaussian or -Rademacher
Schnass [35] *K-SVD criterion* (unit norm tight frames only)	✓	✓	✗	✗	✗	$\|\hat{\mathbf{D}} - \mathbf{D}^o\|_{2,\infty}$ $\leqslant r = O(pn^{-1/4})$	mp^3	$O(1/\mu_1(\mathbf{D}^o))$	"Symmetric decaying": $\boldsymbol{\alpha}_j = \epsilon_j \mathbf{a}_{\sigma(j)}$
Arora et al. [5] *Clustering (***randomized***)*	✓	✓	✗	$P(✓)$	✓	$\|\hat{\mathbf{D}} - \mathbf{D}^o\|_{2,\infty}$ $\leqslant r$	$\frac{p^2 \log p}{k^2}$ $+p \log p \cdot$ $\left(k^2 + \log \frac{1}{r}\right)$	$O\big(\min(\frac{1}{\mu_1(\mathbf{D}^o) \log m}$, $p^{2/5}\big)$	k-sparse $1 \leqslant \|\alpha_j\| \leqslant C$
Agarwal et al. [3] *Clustering (***randomized***) &* ℓ^1	✓	✗	✗	$P(✓)$	✓	✓	$p \log mp$	$O\big(\min(1/\sqrt{\mu_1(\mathbf{D}^o)}$, $m^{1/5}, p^{1/6})\big)$ (+ dynamic range)	k-sparse -Rademacher
Agarwal et al. [2] ℓ^1 *optim with AltMinDict &* **randomized** *clustering init.*	✓	✗	✗	$P(✓)$	✓	✓	$p^2 \log p$	$O\big(\min(1/\sqrt{\mu_1(\mathbf{D}^o)}$, $m^{1/9}, p^{1/8})\big)$	k-sparse - i.i.d. $\underline{\alpha} \leqslant \|\alpha_j\| \leqslant M$
Schnass [36] *Response maxim. criterion*	✓	✓	✗	✗	✗	$\|\hat{\mathbf{D}} - \mathbf{D}^o\|_{2,\infty}$ $\leqslant r$	$\frac{mp^3 k}{r^2}$	$O(1/\mu_1(\mathbf{D}^o))$	"Symmetric decaying"
This contribution *Regularized* ℓ^1 *criterion with penalty factor* λ	✓	✓	✓	✗	✗	$\|\hat{\mathbf{D}} - \mathbf{D}^o\|_F$ $\leqslant r = O(\lambda)$ ✓for $\lambda \to 0$	mp^3	$\mu_k(\mathbf{D}^o) \leqslant 1/4$	k-sparse, $\underline{\alpha} \leqslant \|\alpha_j\|$, $\|\boldsymbol{\alpha}\|_2 \leqslant M_{\boldsymbol{\alpha}}$

TABLE I: Overview of recent results in the field. For each approach, the table indicates (notations in Section II-0a):
1) whether the analysis can handle overcomplete dictionaries / the presence of noise / that of outliers;
2) when an optimization criterion is considered: whether its global minima are characterized (in contrast to characterizing the presence of a local minimum close to the ground truth dictionary $\mathbf{D}^o$); alternatively, whether a (randomized) algorithm with success guarantees is provided; the notation $P(✓)$ indicates success with high probability of a randomized algorithm;
3) whether a (randomized) algorithm with proved polynomial complexity is exhibited;
4) whether the output $\hat{\mathbf{D}}$ of the algorithm (resp. the characterized minimum of the criterion) is (with high probability) *exactly* the ground truth dictionary, in the absence of noise and outliers and with finitely many samples. Alternatively the guaranteed upper bound on the distance between $\hat{\mathbf{D}}$ and $\mathbf{D}^o$ is provided;
5) the sample complexity $n = \Omega(\cdot)$, under the scaling $\|\mathbf{D}^o\|_2 = O(1)$, in the absence of noise;
6) the sparsity levels k allowing "exact recovery";
7) a brief description of the models underlying the corresponding analyses. Most models are determined by: i) how the support is selected (a k-sparse support, or one selected through a $Bernoulli(k/p)$ distribution, i.e., each entry is nonzero with probability k/p); and ii) how the nonzero coefficients are drawn: Gaussian, Rademacher (± 1 entries with equal probability), i.i.d. with certain bound and variance constraints. The symmetric and decaying model of Schnass [35, Definitions 2.1,2.2] first generates a coefficient decay profile $\mathbf{a} \in \mathbb{R}^p$, then the coefficient vector $\boldsymbol{\alpha}$ using a random permutation σ of indices and i.i.d. signs ϵ_i.

II. Problem statement

We introduce in this section the material required to define our problem and state our results.

a) Notations.: For any integer p, we define the set $[\![1;p]\!] \triangleq \{1,\dots,p\}$. For all vectors $\mathbf{v} \in \mathbb{R}^p$, we denote by $\mathrm{sign}(\mathbf{v}) \in \{-1,0,1\}^p$ the vector such that its j-th entry $[\mathrm{sign}(\mathbf{v})]_j$ is equal to zero if $\mathbf{v}_j = 0$, and to one (respectively, minus one) if $\mathbf{v}_j > 0$ (respectively, $\mathbf{v}_j < 0$). The notations $\mathbf{A}^\top$ and $\mathbf{A}^+$ denote the transpose and the Moore-Penrose pseudo-inverse of a matrix $\mathbf{A}$. We extensively manipulate matrix norms in the sequel. For any matrix $\mathbf{A} \in \mathbb{R}^{m\times p}$, we define its Frobenius norm by $\|\mathbf{A}\|_F \triangleq [\sum_{i=1}^m \sum_{j=1}^p \mathbf{A}_{ij}^2]^{1/2}$; similarly, we denote the spectral norm of $\mathbf{A}$ by $|\!|\!|\mathbf{A}|\!|\!|_2 \triangleq \max_{\|\mathbf{x}\|_2\leqslant 1} \|\mathbf{A}\mathbf{x}\|_2$, we refer to the operator ℓ_∞-norm as $|\!|\!|\mathbf{A}|\!|\!|_\infty \triangleq \max_{\|\mathbf{x}\|_\infty\leqslant 1} \|\mathbf{A}\mathbf{x}\|_\infty = \max_{i\in[\![1;m]\!]} \sum_{j=1}^p |\mathbf{A}_{ij}|$, and we denote $\|\mathbf{A}\|_{1,2} \triangleq \sum_{j\in[\![1;p]\!]} \|\mathbf{a}^j\|_2$ with $\mathbf{a}^j$ the j-th column of $\mathbf{A}$. In several places we will exploit the fact that for any matrix $\mathbf{A}$ we have

$$|\!|\!|\mathbf{A}|\!|\!|_2 \leqslant \|\mathbf{A}\|_F.$$

For any square matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$, we denote by $\mathrm{diag}(\mathbf{B}) \in \mathbb{R}^n$ the vector formed by extracting the diagonal terms of $\mathbf{B}$, and conversely, for any $\mathbf{b} \in \mathbb{R}^n$, we use $\mathrm{Diag}(\mathbf{b}) \in \mathbb{R}^{n\times n}$ to represent the (square) diagonal matrix whose diagonal elements are built from the vector $\mathbf{b}$. Denote $\mathrm{off}(\mathbf{A}) \triangleq \mathbf{A} - \mathrm{Diag}(\mathrm{diag}(\mathbf{A}))$ the off-diagonal part of $\mathbf{A}$, which matches $\mathbf{A}$ except on the diagonal where it is zero. The identity matrix is denoted $\mathbf{I}$.

For any $m\times p$ matrix $\mathbf{A}$ and index set $\mathrm{J} \subset [\![1;p]\!]$ we denote by $\mathbf{A}_\mathrm{J}$ the matrix obtained by concatenating the columns of $\mathbf{A}$ indexed by J. The number of elements or size of J is denoted $|\mathrm{J}|$, and its complement in $[\![1;p]\!]$ is denoted J^c. Given a matrix $\mathbf{D} \in \mathbb{R}^{m\times p}$ and a support set J such that $\mathbf{D}_\mathrm{J}$ has linearly independent columns, we define the shorthands

$$\begin{aligned}
\mathbf{G}_\mathrm{J} &\triangleq \mathbf{G}_\mathrm{J}(\mathbf{D}) \triangleq \mathbf{D}_\mathrm{J}^\top \mathbf{D}_\mathrm{J} \\
\mathbf{H}_\mathrm{J} &\triangleq \mathbf{H}_\mathrm{J}(\mathbf{D}) \triangleq \mathbf{G}_\mathrm{J}^{-1} \\
\mathbf{P}_\mathrm{J} &\triangleq \mathbf{P}_\mathrm{J}(\mathbf{D}) \triangleq \mathbf{D}_\mathrm{J}\mathbf{D}_\mathrm{J}^+ = \mathbf{D}_\mathrm{J}\mathbf{H}_\mathrm{J}\mathbf{D}_\mathrm{J}^\top,
\end{aligned}$$

respectively the Gram matrix of $\mathbf{D}_\mathrm{J}$ and its inverse, and the orthogonal projector onto the span of the columns of $\mathbf{D}$ indexed by J.

For any function $h(\mathbf{D})$ we define $\Delta h(\mathbf{D}';\mathbf{D}) \triangleq h(\mathbf{D}') - h(\mathbf{D})$. Finally, the ball (resp. the sphere) of radius $r > 0$ centered on $\mathbf{D}$ in $\mathbb{R}^{m\times p}$ with respect to the Frobenius norm is denoted $\mathcal{B}(\mathbf{D};r)$ (resp. $\mathcal{S}(\mathbf{D};r)$).

The notation $a = O(b)$, or $a \lesssim b$, indicates the existence of a finite constant C such that $a \leqslant Cb$. Vice-versa, $a = \Omega(b)$, or $a \gtrsim b$, means $b = O(a)$, and $a \asymp b$ means that $a = O(b)$ and $b = O(a)$ hold simultaneously.

A. Background material on sparse coding

Let us consider a set of n signals $\mathbf{X} \triangleq [\mathbf{x}^1,\dots,\mathbf{x}^n] \in \mathbb{R}^{m\times n}$ each of dimension m, along with a dictionary $\mathbf{D} \triangleq [\mathbf{d}^1,\dots,\mathbf{d}^p] \in \mathbb{R}^{m\times p}$ formed of p columns called atoms—also known as dictionary elements. Sparse coding simultaneously learns $\mathbf{D}$ and a set of n sparse p-dimensional vectors $\mathbf{A} \triangleq [\boldsymbol{\alpha}^1,\dots,\boldsymbol{\alpha}^n] \in \mathbb{R}^{p\times n}$, such that each signal $\mathbf{x}^i$ can be well approximated by $\mathbf{x}^i \approx \mathbf{D}\boldsymbol{\alpha}^i$ for i in $[\![1;n]\!]$. By sparse, we mean that the vector $\boldsymbol{\alpha}^i$ has $k \ll p$ non-zero coefficients, so that we aim at reconstructing $\mathbf{x}^i$ from only a few atoms. Before introducing the sparse coding formulation [33, 48, 28], we need some definitions. We denote by $g\colon \mathbb{R}^p \to \mathbb{R}^+$ a penalty function that will typically promote sparsity.

Definition 1. *For any dictionary $\mathbf{D} \in \mathbb{R}^{m\times p}$ and signal $\mathbf{x} \in \mathbb{R}^m$, we define*

$$\begin{aligned}
\mathcal{L}_\mathbf{x}(\mathbf{D},\boldsymbol{\alpha}) &\triangleq \tfrac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + g(\boldsymbol{\alpha}) &(1)\\
f_\mathbf{x}(\mathbf{D}) &\triangleq \inf_{\boldsymbol{\alpha}\in\mathbb{R}^p} \mathcal{L}_\mathbf{x}(\mathbf{D},\boldsymbol{\alpha}). &(2)
\end{aligned}$$

Similarly for any set of n signals $\mathbf{X} \triangleq [\mathbf{x}^1,\dots,\mathbf{x}^n] \in \mathbb{R}^{m\times n}$, we introduce

$$F_\mathbf{X}(\mathbf{D}) \triangleq \tfrac{1}{n}\sum_{i=1}^n f_{\mathbf{x}^i}(\mathbf{D}). \quad (3)$$

Based on problem (2) with the ℓ^1 penalty,

$$g(\boldsymbol{\alpha}) \triangleq \lambda\|\boldsymbol{\alpha}\|_1, \quad (4)$$

refered to as Lasso in statistics [38], and basis pursuit in signal processing [11], the standard approach to perform sparse coding [33, 48, 28] solves the minimization problem

$$\min_{\mathbf{D}\in\mathcal{D}} F_\mathbf{X}(\mathbf{D}), \quad (5)$$

where the regularization parameter λ in (4) controls the tradeoff between sparsity and approximation quality, while $\mathcal{D} \subseteq \mathbb{R}^{m\times p}$ is a compact constraint set; in this paper, $\mathcal{D}$ denotes the set of dictionaries with unit ℓ_2-norm atoms, also called the *oblique manifold* [1], which is a natural choice in signal and image processing [28, 19, 34, 39]. Note however that other choices for the set $\mathcal{D}$ may also be relevant depending on the application at hand (see, e.g., Jenatton et al. [24] where in the context of topic models, the atoms in $\mathcal{D}$ belong to the unit simplex). The sample complexity of dictionary learning with general constraint sets is studied by Maurer and Pontil [31], Gribonval et al. [21] for various families of penalties $g(\boldsymbol{\alpha})$.

B. Main objectives

The goal of the paper is to characterize some local minima of the function $F_\mathbf{X}$ with the ℓ^1 penalty, under a generative model for the signals $\mathbf{x}^i$. Throughout the paper, the main model we consider is that of observed signals generated *independently* according to a specified probabilistic model. The signals are typically drawn as $\mathbf{x}^i \triangleq \mathbf{D}^o\boldsymbol{\alpha}^i + \boldsymbol{\varepsilon}^i$ where $\mathbf{D}^o$ is a fixed reference dictionary, $\boldsymbol{\alpha}^i$ is a sparse coefficient vector, and $\boldsymbol{\varepsilon}^i$ is a noise term. The specifics of the underlying probabilistic model, and its possible contamination with *outliers* are considered in Section II-C. Under this model, we can state more precisely our objective: we want to show that, for large enough n,

$$\mathbb{P}\big(F_\mathbf{X} \text{ has a local minimum in a "neighborhood" of } \mathbf{D}^o\big) \approx 1.$$

We loosely refer to a "neighborhood" since in our regularized formulation, a local minimum is not necessarily expected to

appear exactly at $\mathbf{D}^o$. The proper meaning of this neighborhood is in the sense of the Frobenius distance $\|\mathbf{D} - \mathbf{D}^o\|_F$. Other metrics can be envisioned and are left as future work. How large n should be for the results to hold is related to the notion of sample complexity.

a) Intrinsic ambiguities of sparse coding.: Importantly, we so far referred to $\mathbf{D}^o$ as *the* reference dictionary generating the signals. However, and as already discussed by Gribonval and Schnass [19], Geng et al. [17] and more generally in the related literature on blind source separation and independent component analysis [see, e.g., 12], it is known that the objective of (5) is invariant by sign flips and permutations of the atoms. As a result, while solving (5), we cannot hope to identify the specific $\mathbf{D}^o$. We focus instead on the local identifiability of the whole *equivalence class* defined by the transformations described above. From now on, we simply refer to $\mathbf{D}^o$ to denote one element of this equivalence class. Also, since these transformations are *discrete*, our local analysis is not affected by invariance issues, as soon as we are sufficiently close to some representant of $\mathbf{D}^o$.

C. The sparse and the spurious

The considered training set is composed of two types of vectors: *the sparse*, drawn i.i.d. from a distribution generating (noisy) signals that are sparse in the dictionary $\mathbf{D}^o$; and *the spurious*, corresponding to *outliers*.

*1) The sparse: probabilistic model of sparse signals (*inliers*):* Given a reference dictionary $\mathbf{D}^o \in \mathcal{D}$, each (*inlier*) signal $\mathbf{x} \in \mathbb{R}^m$ is built *independently* in three steps:

- **Support generation:** Draw uniformly without replacement k atoms out of the p available in $\mathbf{D}^o$. This procedure thus defines a support $\mathrm{J} \subset [\![1; p]\!]$ whose size is $|\mathrm{J}| = k$.
- **Coefficient vector:** Draw a sparse vector $\boldsymbol{\alpha}^o \in \mathbb{R}^p$ supported on J (i.e., with $\boldsymbol{\alpha}^o_{\mathrm{J}^c} = 0$).
- **Noise:** Eventually generate the signal $\mathbf{x} = \mathbf{D}^o\boldsymbol{\alpha}^o + \boldsymbol{\varepsilon}$.

The random vectors $\boldsymbol{\alpha}^o_{\mathrm{J}}$ and $\boldsymbol{\varepsilon}$ satisfy the following assumptions, where we denote $\mathbf{s}^o = \mathrm{sign}(\boldsymbol{\alpha}^o)$.

Assumption A (Basic signal model)**.**

$$\mathbb{E}\left\{\boldsymbol{\alpha}^o_{\mathrm{J}}[\boldsymbol{\alpha}^o_{\mathrm{J}}]^\top \mid \mathrm{J}\right\} = \mathbb{E}\{\alpha^2\} \cdot \mathbf{I} \tag{6}$$

coefficient whiteness

$$\mathbb{E}\left\{\mathbf{s}^o_J[\mathbf{s}^o_J]^\top \mid \mathrm{J}\right\} = \mathbf{I} \tag{7}$$

sign whiteness

$$\mathbb{E}\left\{\boldsymbol{\alpha}^o_{\mathrm{J}}[\mathbf{s}^o_J]^\top \mid \mathrm{J}\right\} = \mathbb{E}\{|\alpha|\} \cdot \mathbf{I} \tag{8}$$

sign/coefficient decorrelation

$$\mathbb{E}\left\{\boldsymbol{\varepsilon}[\boldsymbol{\alpha}^o_{\mathrm{J}}]^\top \mid \mathrm{J}\right\} = \mathbb{E}\left\{\boldsymbol{\varepsilon}[\mathbf{s}^o_J]^\top \mid \mathrm{J}\right\} = 0 \tag{9}$$

noise/coefficient decorrelation

$$\mathbb{E}\left\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top|\mathrm{J}\right\} = \mathbb{E}\{\epsilon^2\} \cdot \mathbf{I} \tag{10}$$

noise whiteness

In light of these assumptions we define the shorthand

$$\kappa_\alpha \triangleq \frac{\mathbb{E}|\alpha|}{\sqrt{\mathbb{E}\alpha^2}}. \tag{11}$$

By Jensen's inequality, we have $\kappa_\alpha \leqslant 1$, with $\kappa_\alpha = 1$ corresponding to the degenerate situation where $\boldsymbol{\alpha}_{\mathrm{J}}$ almost surely has all its entries of the same magnitude, i.e., with the smallest possible dynamic range. Conversely, $\kappa_\alpha \ll 1$ corresponds to marginal distributions of the coefficients with a wide dynamic range. In a way, κ_α measures the typical "flatness" of $\boldsymbol{\alpha}$ (the larger κ_α, the flatter the typical $\boldsymbol{\alpha}$)

A *boundedness* assumption will complete Assumption A to handle sparse recovery in our proofs.

Assumption B (Bounded signal model)**.**

$$\mathbb{P}(\min_{j\in\mathrm{J}} |\boldsymbol{\alpha}^o_j| < \underline{\alpha} \mid \mathrm{J}) = 0, \textit{ for some } \underline{\alpha} > 0 \tag{12}$$

coefficient threshold

$$\mathbb{P}(\|\boldsymbol{\alpha}^o\|_2 > M_{\boldsymbol{\alpha}}) = 0, \textit{ for some } M_{\boldsymbol{\alpha}} \tag{13}$$

coefficient boundedness

$$\mathbb{P}(\|\boldsymbol{\varepsilon}\|_2 > M_{\boldsymbol{\varepsilon}}) = 0, \textit{ for some } M_{\boldsymbol{\varepsilon}}. \tag{14}$$

noise boundedness

Remark 1. *Note that neither Assumption A nor Assumption B requires that the entries of $\boldsymbol{\alpha}^o$ indexed by* J *be i.i.d. In fact, the stable and robust identifiability of $\mathbf{D}^o$ from the training set $\mathbf{X}$ rather stems from geometric properties of the training set (its concentration close to a union of low-dimensional subspaces spanned by few columns of $\mathbf{D}^o$) than from traditional independent component analysis (ICA). This will be illustrated by a specific coefficient model (inspired by the symmetric decaying coefficient model of Schnass [35]) in Example 1.*

To summarize, the signal model is parameterized by the sparsity k, the expected coefficient energy $\mathbb{E}\,\alpha^2$, the minimum coefficient magnitude $\underline{\alpha}$, maximum norm $M_{\boldsymbol{\alpha}}$, and the flatness κ_α. These parameters are interrelated, e.g., $\underline{\alpha}\sqrt{k} \leqslant M_{\boldsymbol{\alpha}}$.

a) Related models: The Bounded model above is related to the $\Gamma_{k,C}$ model of Arora et al. [5] (which also covers [3, 2]): in the latter, our assumptions (12)-(13) are replaced by $1 \leqslant |\alpha_j| \leqslant C$. Note that the $\Gamma_{k,C}$ model of Arora et al. [5] does not assume that the support is chosen uniformly at random (among all k-sparse sets) and some mild dependencies are allowed. Alternatives to (13) with a control on $\|\boldsymbol{\alpha}\|_q$ for some $0 < q \leqslant \infty$ can easily be dealt with through appropriate changes in the proofs, but we chose to focus on $q = 2$ for the sake of simplicity. Compared to early work in the field considering a Bernoulli-Gaussian model [19] or a k-sparse Gaussian model [17], Assumptions A & B are rather generic and do not assume a specific shape of the distribution $\mathbb{P}(\boldsymbol{\alpha})$. In particular, the conditional distribution of $\boldsymbol{\alpha}_{\mathrm{J}}$ given J *may depend on* J, provided its "marginal moments" $\mathbb{E}\,\alpha^2$ and $\mathbb{E}\,|\alpha|$ satisfy the expressed assumptions.

2) The spurious: outliers*:* In addition to a set of n_{in} *inliers* drawn i.i.d. as above, the training set may contain n_{out} *outliers*, i.e., training vectors that may have completely distinct properties and may not relate in any manner to the reference dictionary $\mathbf{D}^o$. Since the considered cost function $F_{\mathbf{X}}(\mathbf{D})$ is not altered when we permute the columns of the matrix $\mathbf{X}$ representing the training set, without loss of generality we will consider that $\mathbf{X} = [\mathbf{X}_{\text{in}}, \mathbf{X}_{\text{out}}]$. As we will see, controlling the ratio $\|\mathbf{X}_{\text{out}}\|_F^2/n_{\text{in}}$ of the total energy of outliers to the number of inliers will be enough to ensure that the local

minimum of the sparse coding objective function is robust to outliers. While this control does not require any additional assumptions, the ratio $\|\mathbf{X}_{\text{out}}\|_F^2/n_{\text{in}}$ directly impacts the error in estimating the dictionary (i.e., the local minimum in $\mathbf{D}$ is further away from $\mathbf{D}^o$). With additional assumptions (namely that the reference dictionary is complete), we show that if $\|\mathbf{X}_{\text{out}}\|_{1,2}/n_{\text{in}}$ is sufficiently small, then our upper bound on the distance from the local minimum to $\mathbf{D}^o$ remains valid.

D. The dictionary: cumulative coherence and restricted isometry properties

Part of the technical analysis relies on the notion of *sparse recovery*. A standard sufficient support recovery condition is referred to as the *exact recovery condition* in signal processing [16, 40] or the *irrepresentability condition* (IC) in the machine learning and statistics communities [44, 46]. It is a key element to almost surely control the supports of the solutions of ℓ_1-regularized least-squares problems. To keep our analysis reasonably simple, we will impose the irrepresentability condition *via* a condition on the *cumulative coherence* of the reference dictionary $\mathbf{D}^o \in \mathcal{D}$, which is a stronger requirement [43, 15]. This quantity is defined (see, e.g., [16, 13]) for unit-norm columns (i.e., on the oblique manifold $\mathcal{D}$) as

$$\mu_k(\mathbf{D}) \triangleq \sup_{|\mathrm{J}|\leqslant k} \sup_{j\notin \mathrm{J}} \|\mathbf{D}_{\mathrm{J}}^\top \mathbf{d}^j\|_1. \tag{15}$$

The term $\mu_k(\mathbf{D})$ gives a measure of the level of correlation between columns of $\mathbf{D}$. It is for instance equal to zero in the case of an orthogonal dictionary, and exceeds one if $\mathbf{D}$ contains two colinear columns. For a given dictionary $\mathbf{D}$, the cumulative coherence of $\mu_k(\mathbf{D})$ increases with k, and $\mu_k(\mathbf{D}) \leqslant k\mu_1(\mathbf{D})$ where $\mu_1(\mathbf{D}) = \max_{i\neq j} |\langle \mathbf{d}^i, \mathbf{d}^j\rangle|$ is the plain coherence of $\mathbf{D}$.

For the theoretical analysis we conduct, we consider a deterministic assumption based on the cumulative coherence, slightly weakening the coherence-based assumption considered for instance in previous work on dictionary learning [19, 17]. Assuming that $\mu_k(\mathbf{D}^o) < 1/2$ where k is the level of sparsity of the coefficient vectors $\boldsymbol{\alpha}^i$, an important step will be to show that such an upper bound on $\mu_k(\mathbf{D}^o)$ loosely transfers to $\mu_k(\mathbf{D})$ provided that $\mathbf{D}$ is close enough to $\mathbf{D}^o$, leading to *locally stable exact recovery results in the presence of bounded noise* (Proposition 3).

Many elements of our proofs rely on a restricted isometry property (RIP), which is known to be weaker than the coherence assumption [43]. By definition the *restricted isometry constant* of order k of a dictionary $\mathbf{D}$, $\delta_k(\mathbf{D})$ is the smallest number δ_k such that for any support set J of size $|\mathrm{J}| = k$ and $\mathbf{z} \in \mathbb{R}^k$,

$$(1-\delta_k)\,\|\mathbf{z}\|_2^2 \leqslant \|\mathbf{D}_{\mathrm{J}}\mathbf{z}\|_2^2 \leqslant (1+\delta_k)\,\|\mathbf{z}\|_2^2. \tag{16}$$

In our context, *the best lower bound and best upper bound will play significantly different roles*, so we define them separately as $\underline{\delta}_k(\mathbf{D})$ and $\overline{\delta}_k(\mathbf{D})$, so that $\delta_k(\mathbf{D}) = \max(\underline{\delta}_k(\mathbf{D}), \overline{\delta}_k(\mathbf{D}))$. Both can be estimated by the cumulative coherence as $\delta_k(\mathbf{D}) \leqslant \mu_{k-1}(\mathbf{D})$ by Gersgorin's disc theorem [40]. Possible extensions of this work that would fully relax the incoherence assumption and only rely on the RIP are discussed in Section V.

III. Main results

Our main results, described below, show that under appropriate scalings of the dictionary dimensions m, p, number of training samples n, and model parameters, the sparse coding problem (5) admits a local minimum in a neighborhood of $\mathbf{D}^o$ of controlled size, for appropriate choices of the regularization parameter λ. The main building blocks of the results (Propositions 1-2-3) and the high-level structure of their proofs are given in Section IV. The most technical lemmata are postponed to the Appendix.

A. Stable local identifiability

We begin with asymptotic results (n being infinite), in the absence of outliers.

Theorem 1 (Asymptotic results, bounded model, no outlier)**.** *Consider the following assumptions:*

- **Coherence and sparsity level:** *consider $\mathbf{D}^o \in \mathcal{D}$ and k such that*

$$\mu_k(\mathbf{D}^o) \;\leqslant\; 1/4 \tag{17}$$

$$k \;\leqslant\; \frac{p}{16(|||\mathbf{D}^o|||_2+1)^2}. \tag{18}$$

- **Coefficient distribution:** *assume the Basic & Bounded signal model (Assumptions A & B) and*

$$\frac{\mathbb{E}\,\alpha^2}{M_{\boldsymbol{\alpha}}\mathbb{E}\,|\alpha|} \;>\; 84\cdot(|||\mathbf{D}^o|||_2+1)\cdot\frac{\frac{k}{p}\cdot\|[\mathbf{D}^o]^\top\mathbf{D}^o-\mathbf{I}\|_F}{1-2\mu_k(\mathbf{D}^o)}. \tag{19}$$

This implies $C_{\min} < C_{\max}$ where we define

$$C_{\min} \;\triangleq\; 24\kappa_\alpha^2\cdot(|||\mathbf{D}^o|||_2+1)\cdot\frac{k}{p}\cdot\|[\mathbf{D}^o]^\top\mathbf{D}^o-\mathbf{I}\|_F, \tag{20}$$

$$C_{\max} \;\triangleq\; \frac{2}{7}\cdot\frac{\mathbb{E}\,|\alpha|}{M_{\boldsymbol{\alpha}}}\cdot(1-2\mu_k(\mathbf{D}^o)). \tag{21}$$

- **Regularization parameter:** *consider a small enough regularization parameter,*

$$\lambda \leqslant \frac{\underline{\alpha}}{4}. \tag{22}$$

Denoting $\bar{\lambda} \triangleq \frac{\lambda}{\mathbb{E}\,|\alpha|}$, this implies $C_{\max}\cdot\bar{\lambda} \leqslant 0.15$.

- **Noise level:** *assume a small enough relative noise level,*

$$\frac{M_{\boldsymbol{\varepsilon}}}{M_{\boldsymbol{\alpha}}} < \frac{7}{2}\cdot(C_{\max}-C_{\min})\cdot\bar{\lambda}. \tag{23}$$

Then, for any resolution $r > 0$ such that

$$C_{\min}\cdot\bar{\lambda} < r < C_{\max}\cdot\bar{\lambda}, \tag{24}$$

and

$$\frac{M_{\boldsymbol{\varepsilon}}}{M_{\boldsymbol{\alpha}}} < \frac{7}{2}\left(C_{\max}\cdot\bar{\lambda}-r\right), \tag{25}$$

the function $\mathbf{D}\in\mathcal{D}\mapsto\mathbb{E}\,F_{\mathbf{X}}(\mathbf{D})$ admits a local minimum $\hat{\mathbf{D}}$ such that $\|\hat{\mathbf{D}}-\mathbf{D}^o\|_F < r$.

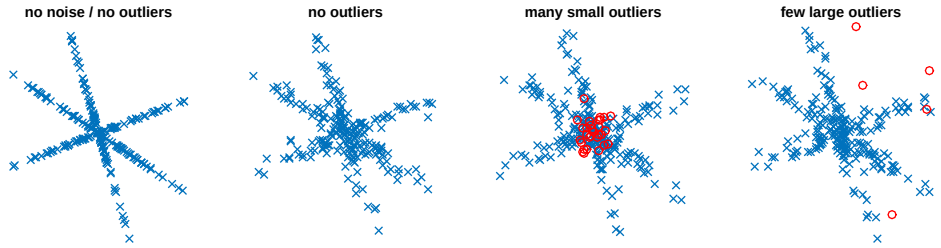

Fig. 1: Noise and outliers: illustration with three atoms in two dimensions (blue crosses: inliers, red circles:outliers).

Remark 2 (Limited over-completeness of $\mathbf{D}^o$)**.** *It is perhaps not obvious how strong a requirement is assumption* (19)*. On the one hand, its left hand side is easily seen to be less than one (and as seen above can be made arbitrarily close to one with appropriate coefficient distribution). On the other hand by the Welsh bound* $\|[\mathbf{D}^o]^\top \mathbf{D}^o - \mathbf{Id}\|_F \geq \sqrt{p(p-m)/m}$*, the bound* $|||\mathbf{D}^o|||_2 \geq \|\mathbf{D}^o\|_F/\sqrt{m} = \sqrt{p/m}$*, and the assumption* $\mu_k(\mathbf{D}^o) \leqslant 1/4$*, its right hand side is bounded from below by* $\Omega(k\sqrt{(p-m)/m^2})$*. Hence, a consequence of assumption* (19) *is that Theorem 1 only applies to dictionaries with limited over-completeness, with* $p \lesssim m^2$*. This is likely to be an artifact from the use of coherence in our proof, and a degree of overcompleteness* $p = O(m^2)$ *covers already interesting practical settings: for example [28] consider* $m = (8 \times 8)$ *patches with* $p = 256$ *atoms* $< m^2 = 64^2 = 4096$*)*

Since $\frac{k}{p} \cdot \|[\mathbf{D}^o]^\top \mathbf{D}^o - \mathbf{I}\|_F \leqslant k\mu_1(\mathbf{D}^o)$ and $\mu_k(\mathbf{D}^o) \leqslant k\mu_1(\mathbf{D}^o)$, a crude upper bound on the rightmost factor in (19) is $k\mu_1(\mathbf{D}^o)/(1-2k\mu_1(\mathbf{D}^o))$, which appears in many coherence-based sparse-recovery results.

1) Examples: Instantiating Theorem 1 on a few examples highlights the strength of its main assumptions.

Example 1 (Incoherent pair of orthonormal bases)**.** *When* $\mathbf{D}^o$ *is an incoherent dictionary in* $\mathbb{R}^{m\times p}$*, i.e., a dictionary with (plain) coherence* $\mu = \mu_1(\mathbf{D}^o) \ll 1$*, we have the estimates [40]* $\mu_k(\mathbf{D}^o) \leqslant k\mu$ *and*

$$\|[\mathbf{D}^o]^\top \mathbf{D}^o - \mathbf{I}\|_F \leqslant \sqrt{p(p-1)\mu^2} \leqslant p\mu.$$

Assumption (17) *therefore holds as soon as* $k \leqslant 1/(4\mu)$*. In the case where* $p = 2m$ *and* $\mathbf{D}^o$ *is not only incoherent but also a union of two orthonormal bases, we further have* $|||\mathbf{D}^o|||_2 = \sqrt{2}$ *hence assumption* (18) *is fulfilled as soon as* $k \leqslant p/100 = m/50$*. Moreover, the right hand side in* (19) *reads*

$$84 \cdot (|||\mathbf{D}^o||| + 1) \cdot \frac{\frac{k}{p} \cdot \|[\mathbf{D}^o]^\top \mathbf{D}^o - \mathbf{I}\|_F}{1 - 2\mu_k(\mathbf{D}^o)} \leqslant \frac{203k\mu}{1-2k\mu} \leqslant 406k\mu,$$

and assumption (19) *holds provided that* $\mathbb{E}\alpha^2/(M_{\boldsymbol{\alpha}}\mathbb{E}|\alpha|)$ *exceeds this threshold. We discuss below concrete signal settings where this condition can be satisfied:*

- **i.i.d. bounded coefficient model:** *on the one hand, consider nonzero coefficients drawn i.i.d. with* $\mathbb{P}(|\boldsymbol{\alpha}_j| < \underline{\alpha} | j \in J) = 0$*. The almost-sure upper-bound* $M_{\boldsymbol{\alpha}}$ *on* $\|\boldsymbol{\alpha}\|_2$ *implies the existence of* $\overline{\alpha} \geq \underline{\alpha}$ *such that* $\mathbb{P}(|\boldsymbol{\alpha}_j| > \overline{\alpha} | j \in J) = 0$*. As an example, consider coefficients drawn i.i.d. with* $\mathbb{P}(\boldsymbol{\alpha}_j = \pm\overline{\alpha} | j \in \mathrm{J}) = \pi \in (0,1)$ *and* $\mathbb{P}(\boldsymbol{\alpha}_j = \pm\underline{\alpha} | j \in \mathrm{J}) = 1-\pi$*. For large* $\overline{\alpha}$ *we have* $\mathbb{E}\alpha^2 = \pi\overline{\alpha}^2 + (1-\pi)\underline{\alpha}^2 \asymp \pi\overline{\alpha}^2$*,* $\mathbb{E}|\alpha| \asymp \pi\overline{\alpha}$*, and* $M_{\boldsymbol{\alpha}} = \sqrt{k}\overline{\alpha}$*. This yields*

$$\lim_{\overline{\alpha}\to\infty} \mathbb{E}\alpha^2/(M_{\boldsymbol{\alpha}}\mathbb{E}|\alpha|) = 1/\sqrt{k},$$

This shows the existence of a coefficient distribution satisfying (19) *as soon as* $406k\mu < 1/\sqrt{k}$*, that is to say* $k < 1/(406\mu)^{2/3}$*. In the maximally incoherent case, for large* p*, we have* $\mu = 1/\sqrt{m} \asymp p^{-1/2}$*, and conditions* (17)-(18)-(19) *read* $k = O(p^{1/3})$*.*
- **fixed amplitude profile coefficient model:** *on the other hand, completely relax the independence assumption and consider essentially the coefficient model introduced by Schnass [35] where* $\boldsymbol{\alpha}_j = \epsilon_j \mathbf{a}_{\sigma(j)}$ *with i.i.d. signs* ϵ_j *such that* $\mathbb{P}(\epsilon_j = \pm 1) = 1/2$*, a random permutation* σ *of the index set* J*, and* $\mathbf{a}$ *a given vector with entries* $\mathbf{a}_j \geq \underline{\alpha}, j \in \mathrm{J}$*. This yields*

$$\mathbb{E}\alpha^2/(M_{\boldsymbol{\alpha}}\mathbb{E}|\alpha|) = \tfrac{1}{k}\|\mathbf{a}\|_2^2/(\|\mathbf{a}\|_2 \cdot \tfrac{1}{k}\|\mathbf{a}\|_1) = \|\mathbf{a}\|_2/\|\mathbf{a}\|_1,$$

which can be made arbitrarily close to one even with the constraint $\mathbf{a}_j \geq \underline{\alpha}, j \in \mathrm{J}$*. This shows the existence of a coefficient distribution satisfying* (19) *as soon as* $406k\mu < 1$*, a much less restrictive condition leading to* $k = O(p^{1/2})$*. The reader may notice that such distributions concentrate most of the energy of* $\boldsymbol{\alpha}$ *on just a few coordinates, so in a sense such vectors are much sparser than* k*-sparse.*

Example 2 (Spherical ensemble)**.** *Consider* $\mathbf{D}^o \in \mathbb{R}^{m\times p}$ *a typical draw from the spherical ensemble, that is a dictionary obtained by normalizing a matrix with standard independent Gaussian entries. As discussed above, condition* (19) *imposes overall dimensionality constraints* $p \lesssim m^2$*. Moreover, using usual results for such dictionaries [see, e.g., 10], the condition in* (17) *is satisfied as soon as* $\mu_k \leqslant k\mu_1 \approx k\sqrt{\log p}/\sqrt{m} = O(1)$*, i.e.,* $k = O(\sqrt{m/\log p})$*, while the condition in* (18) *is satisfied as long as* $k = O(m)$ *(which is weaker).*

2) Noiseless case: exact recovery: In the noiseless case ($M_{\boldsymbol{\varepsilon}} = 0$), (23) imposes no lower bound on admissible regularization parameter. Hence, we deduce from Theorem 1 that a local minimum of $\mathbb{E}\, F_{\mathbf{X}}(\cdot)$ can be found arbitrarily close to $\mathbf{D}^o$, provided that the regularization parameter λ is small enough. This shows that the reference dictionary $\mathbf{D}^o$ itself is in fact a local minimum of the problem considered by Gribonval and Schnass [19], Geng et al. [17],

$$\min_{\mathbf{D}\in\mathcal{D}} F^0_{\mathbf{X}}(\mathbf{D}) \text{ where } F^0_{\mathbf{X}}(\mathbf{D}) \triangleq \min_{\mathbf{A}:\mathbf{DA}=\mathbf{X}} \|\mathbf{A}\|_1. \tag{26}$$

Note that here we consider a different random sparse signal model, and yet recover the same results together with a new extension to the noisy case.

3) Stability to noise: In the presence of noise, conditions (22) and (23) respectively impose an upper and a lower limit on admissible regularization parameters, which are only compatible for small enough levels of noise

$$M_{\boldsymbol{\varepsilon}} \lesssim \underline{\alpha}(1-2\mu_k(\mathbf{D}^o)).$$

In scenarios where $C_{\min} \ll C_{\max}$ (i.e., when the left hand side in (19) is large enough compared to its right hand side), admissible regularization parameters are bounded from below given (23) as $\bar{\lambda} \gtrsim \frac{M_{\boldsymbol{\varepsilon}}}{M_{\boldsymbol{\alpha}} C_{\max}}$, therefore limiting the achievable "resolution" r to

$$\begin{aligned} r > C_{\min}\bar{\lambda} &\gtrsim \frac{M_{\boldsymbol{\varepsilon}}}{M_{\boldsymbol{\alpha}}} \cdot \frac{C_{\min}}{C_{\max}} \\ &\asymp \frac{M_{\boldsymbol{\varepsilon}}}{\sqrt{\mathbb{E}\,\alpha^2}} \cdot \kappa_\alpha \cdot |\!|\!|\mathbf{D}^o|\!|\!|_2 \cdot \frac{\frac{k}{p}\cdot \|[\mathbf{D}^o]^\top \mathbf{D}^o - \mathbf{I}\|_F}{1-2\mu_k(\mathbf{D}^o)}. \end{aligned} \tag{27}$$

Hence, with enough training signals and in the absence of outliers, the main resolution-limiting factors are

- the relative noise level $M_{\boldsymbol{\varepsilon}}/\sqrt{\mathbb{E}\,\alpha^2}$: the smaller the better;
- the level of typical "flatness" of $\boldsymbol{\alpha}$ as measured by κ_α: the peakier (the smaller κ_α) the better;
- the coherence of the dictionary as measured jointly by $\mu_k(\mathbf{D}^o)$ and $\frac{k}{p}\cdot\|[\mathbf{D}^o]^\top\mathbf{D}^o-\mathbf{I}\|_F$: the least coherent the better.

Two other resolution-limiting factors are the finite number of training samples n and the presence of outliers, which we now discuss.

B. Robust finite sample results

We now trade off precision for concision and express finite sample results with two non-explicit constants C_0 and C_1. Their explicit expression in terms of the dictionary and signal model parameters can be tracked back by the interested reader in the proof of Theorem 2 (Section IV-G), but they are left aside for the sake of concision.

Theorem 2 (Robust finite sample results, bounded model)**.** *Consider a dictionary $\mathbf{D}^o \in \mathcal{D}$ and a sparsity level k satisfying the assumptions* (17)-(18) *of Theorem 1, and the Basic & Bounded signal model (Assumptions A & B) with parameters satisfying the assumption* (19)*. There are two constants $C_0, C_1 > 0$ independent of all considered parameters with the following property.*

Given a reduced regularization parameter $\bar{\lambda}$ and a noise level satisfying assumptions (22) *and* (23)*, a radius r satisfying* (24) *and* (25)*, and a confidence level $x > 0$, when n_{in} training samples are drawn according to the Basic & Bounded signal model with*

$$\begin{aligned} n_{in} > C_0 \cdot (mp+x)\cdot p^2 \cdot \left(\frac{M_{\boldsymbol{\alpha}}^2}{\mathbb{E}\|\boldsymbol{\alpha}\|_2^2}\right)^2 \\ \cdot \left(\frac{r+\left(\frac{M_{\boldsymbol{\varepsilon}}}{M_{\boldsymbol{\alpha}}}+\bar{\lambda}\right)+\left(\frac{M_{\boldsymbol{\varepsilon}}}{M_{\boldsymbol{\alpha}}}+\bar{\lambda}\right)^2}{r-C_{\min}\cdot\bar{\lambda}}\right)^2, \end{aligned} \tag{28}$$

we have: with probability at least $1-2e^{-x}$, the function $\mathbf{D} \in \mathcal{D} \mapsto F_{\mathbf{X}}(\mathbf{D})$ admits a local minimum $\hat{\mathbf{D}}$ such that $\|\mathbf{D}-\mathbf{D}^o\|_F < r$. Moreover, this is robust to the addition of outliers $\mathbf{X}_{out}$ provided that

$$\frac{\|\mathbf{X}_{out}\|_F^2}{n_{in}} \leqslant \mathbb{E}\|\boldsymbol{\alpha}\|_2^2 \cdot \left[\frac{1}{4p}\cdot\left(1-\frac{C_{\min}\cdot\bar{\lambda}}{r}\right) - C_1\sqrt{\frac{(mp+x)}{n_{in}}}\right]\cdot r^2. \tag{29}$$

As soon as the dictionary is coherent, we have $C_{\min} \neq 0$, hence the constraint (24) *implies that the right hand side of* (29) *scales as $O(r^2) = O(\lambda^2)$. In the noiseless case, this imposes a tradeoff between the seeked resolution r, the tolerable total energy of outliers, and the number of inliers. With a more refined argument, we obtain the alternative condition*

$$\begin{aligned} \frac{\|\mathbf{X}_{out}\|_{1,2}}{n_{in}} \leqslant 3\frac{\sqrt{k}\mathbb{E}\,\|\boldsymbol{\alpha}\|_2^2}{\mathbb{E}|\alpha|}\cdot\left[\frac{1}{p}\cdot\left(1-\frac{C_{\min}\cdot\bar{\lambda}}{r}\right) - C_1\sqrt{\frac{(mp+x)}{n_{in}}}\right] \\ \cdot\frac{r}{\lambda}\cdot\frac{(A^o)^{3/2}}{18p^{3/2}}, \end{aligned} \tag{30}$$

where A^o is the lower frame bound of $\mathbf{D}^o$, i.e., such that $A^o\|\mathbf{x}\|_2^2 \leqslant \|(\mathbf{D}^o)^\top\mathbf{x}\|_2^2$ for any signal $\mathbf{x}$.

The factor $M_{\boldsymbol{\alpha}}^2/\mathbb{E}\,\|\boldsymbol{\alpha}\|_2^2 = \text{“}\sup\|\boldsymbol{\alpha}\|_2^2\text{''}/\mathbb{E}\,\|\boldsymbol{\alpha}\|_2^2$ in the right hand side of (28) is always greater than 1, but typically remains bounded (note that if the distribution of $\boldsymbol{\alpha}$ allows outliers, they could be treated within the outlier model). In the symmetric decaying model of Schnass [35] where $\boldsymbol{\alpha}$ is a randomly permuted and signed flipped version of a given vector, this factor is equal to one.

Even though the robustness to outliers is expressed in (29) as a control of $\|\mathbf{X}_{\text{out}}\|_F^2/n_{\text{in}}$, it should really be considered as a control of an *outlier to inlier energy ratio*: $\|\mathbf{X}_{\text{out}}\|_F^2/[n_{\text{in}}\mathbb{E}\,\|\boldsymbol{\alpha}\|_2^2]$, and similarly with a proper adaptation in (30). One may notice that the robustness to outliers expressed in Theorem 2 is somehow a "free" side-effect of the conditions that hold on inliers with high probability, rather than the result of a specific design of the cost function $F_{\mathbf{X}}(\mathbf{D})$.

1) Example: orthonormal dictionary: Consider $p = m$ and $\mathbf{D}^o$ an orthonormal dictionary in $\mathbb{R}^{m\times p}$. Since $\mu_k(\mathbf{D}^o) = 0$, $|\!|\!|\mathbf{D}^o|\!|\!|_2 = 1$ and $\|[\mathbf{D}^o]^\top\mathbf{D}^o-\mathbf{I}\|_F = 0$, assumption (18) reads[2] $k \leqslant p/64$, assumptions (17) and (19) impose no constraint, and $C_{\min} = 0$. Moreover, the reader can check that if $M_{\boldsymbol{\varepsilon}} < \lambda \leqslant \underline{\alpha}/4$, then (22)-(25) hold for $0 < r < \frac{2(\lambda - M_{\boldsymbol{\varepsilon}})}{7M_{\boldsymbol{\alpha}}}$.

- **Low-noise regime:** if $M_{\boldsymbol{\varepsilon}} < \underline{\alpha}/4$ and $k \leqslant p/64$, then choosing $M_{\boldsymbol{\varepsilon}} < \lambda \leqslant \underline{\alpha}/4$ yields:
  - by Theorem 1 (the limit of large n), $\mathbb{E}F_{\mathbf{X}}(\mathbf{D})$ admits a local minimum *exactly at* $\mathbf{D}^o$;
  - by Theorem 2, *even though the regularization parameter cannot be made arbitrarily small*, we obtain that for any confidence level $x > 0$ and *arbitrary small precision* $r > 0$, $F_{\mathbf{X}}(\mathbf{D})$ admits a local minimum within radius r around $\mathbf{D}^o$ with probability at least $1-2e^{-x}$ provided that

$$n = \Omega\left((mp^3+xp^2)\left(\frac{M_{\boldsymbol{\varepsilon}}/M_{\boldsymbol{\alpha}}}{r}\right)^2\right).$$

[2]Improved constants in Theorem 1 are achievable when specializing to orthonormal dictionaries, they are left to the reader.

While the orthogonality of the dictionary remarkably allows to achieve an arbitrary precision despite the presence of noise, we still have to pay a price for the presence of noise through a resolution-dependent sample complexity.

- **Noiseless regime ($M_{\boldsymbol{\varepsilon}} = 0$):** with $\lambda \asymp r$, an arbitrary resolution r is reached with a *resolution independent* number of training samples

$$n = \Omega(mp^3 + xp^2).$$

This is robust to outliers provided $\|\mathbf{X}_{\text{out}}\|_{1,2}/n_{\text{in}}$ does not exceed a *resolution independent* threshold.

The case of orthonormal dictionaries is somewhat special in the sense that orthonormality yields $C_{\min} = 0$ and breaks the forced scaling $r \asymp \bar{\lambda}$ otherwise imposed by (24). Below we discuss in more details the more generic case of non-orthonormal dictionaries in the noiseless case.

2) Noiseless case: exact recovery and resolution independent sample complexity: Consider now the noiseless case ($M_{\boldsymbol{\varepsilon}} = 0$) without outlier ($\mathbf{X}_{\text{out}} = 0$). In general we have $C_{\min} > 0$ hence the best resolution $r > 0$ guaranteed by Theorem 1 in the asymptotic regime is $r = r_{\min} \triangleq C_{\min} \cdot \bar{\lambda} > 0$. When $C_{\max} > 2C_{\min}$, Theorem 2 establishes that the only slightly worse resolution $r = 2r_{\min}$ can be achieved with high probability with a number of training samples n which is *resolution independent*. More precisely (28) indicates that when $M_{\boldsymbol{\alpha}}^2/\mathbb{E}\,\|\boldsymbol{\alpha}\|_2^2 \approx 1$, it is sufficient to have a number of training samples

$$n = \Omega(mp^3)$$

to ensure the existence of a local minimum within a radius r around the ground truth dictionary $\mathbf{D}^o$, where *the resolution r can be made arbitrarily fine by choosing λ small enough*. This recovers the known fact that, with high probability, the function $F_{\mathbf{X}}^0(\mathbf{D})$ defined in (26) has a local minimum *exactly* at $\mathbf{D}^o$, as soon as $n = \Omega(mp^3)$. Given our boundedness assumption, the probabilistic decay as e^{-x} is expected and show that as soon as $n \geq \Omega(mp^3)$, the infinite sample result is reached quickly.

In terms of outliers, both (29) and (30) provide a control of the admissible "energy" of outliers. Without additional assumption on $\mathbf{D}^o$, the allowed energy of outliers in (29) has a leading term in r^2, i.e., to guarantee a high precision, we can only tolerate a small amount of outliers as measured by the ratio $\|\mathbf{X}_{\text{out}}\|_F^2/n_{\text{in}}$. However, when the dictionary $\mathbf{D}^o$ is complete –a rather mild assumption– the alternate ratio $\|\mathbf{X}_{\text{out}}\|_{1,2}/n_{\text{in}}$ does not need to scale with the targeted resolution r for $r = 2C_{\min}\bar{\lambda}$. In the proof, this corresponds to replacing the control of the minimized objective function by that of its variations.

The above described resolution-independent results are of course specific to the noiseless setting. In fact, as described in Section III-A3, the presence of noise when the dictionary is not orthonormal imposes an absolute limit to the resolution $r > r_{\min}$ we can guarantee with the techniques established in this paper. When there is noise, [5] discuss why it is in fact impossible to get a sample complexity with better than $1/r^2$ dependency.

IV. Main steps of the analysis

For many classical penalty functions g, including the considered ℓ^1 penalty $g(\boldsymbol{\alpha}) = \lambda\|\boldsymbol{\alpha}\|_1$, the function $\mathbf{D} \mapsto F_{\mathbf{X}}(\mathbf{D})$ is continuous, and in fact Lipschitz [21] with respect to the Frobenius metric $\rho(\mathbf{D}', \mathbf{D}) \triangleq \|\mathbf{D}' - \mathbf{D}\|_F$ on all $\mathbb{R}^{m\times p}$, hence in particular on the compact constraint set $\mathcal{D} \subset \mathbb{R}^{m\times p}$. Given a dictionary $\mathbf{D} \in \mathcal{D}$, we have $\|\mathbf{D}\|_F = \sqrt{p}$, and for any radius $0 < r \leqslant 2\sqrt{p}$ the sphere

$$\mathcal{S}(r) \triangleq \mathcal{S}(\mathbf{D}^o; r) = \{\mathbf{D} \in \mathcal{D} : \|\mathbf{D} - \mathbf{D}^o\|_F = r\}$$

is non-empty (for $r = 2\sqrt{p}$ it is reduced to $\mathbf{D} = -\mathbf{D}^o$). We define

$$\Delta F_{\mathbf{X}}(r) \quad \triangleq \quad \inf_{\mathbf{D}\in\mathcal{S}(r)} \Delta F_{\mathbf{X}}(\mathbf{D}; \mathbf{D}^o). \tag{31}$$

where we recall that for any function $h(\mathbf{D})$ we define $\Delta h(\mathbf{D}; \mathbf{D}') \triangleq h(\mathbf{D}) - h(\mathbf{D}')$. Our proof technique will consist in choosing the radius r to ensure that $\Delta F_{\mathbf{X}}(r) > 0$ (with high probability on the draw of $\mathbf{X}$): the compactness of the closed balls

$$\mathcal{B}(r) \triangleq \mathcal{B}(\mathbf{D}^o; r) = \{\mathbf{D} \in \mathcal{D} : \|\mathbf{D} - \mathbf{D}^o\|_F \leqslant r\} \tag{32}$$

will then imply the existence of a local minimum $\hat{\mathbf{D}}$ of $\mathbf{D} \mapsto F_{\mathbf{X}}(\mathbf{D})$ such that $\|\hat{\mathbf{D}} - \mathbf{D}^o\|_F < r$.

A. The need for a precise finite-sample (vs. asymptotic) analysis

Under common assumptions on the penalty function g and the distribution of "clean" training vectors $\mathbf{x} \sim \mathbb{P}$, the empirical cost function $F_{\mathbf{X}}(\mathbf{D})$ converges uniformly to its expectation $\mathbb{E}_{\mathbf{x}\sim\mathbb{P}} f_{\mathbf{x}}(\mathbf{D})$: except with probability at most $2e^{-x}$ [31, 41, 21], we have

$$\sup_{\mathbf{D}\in\mathcal{D}} |F_{\mathbf{X}}(\mathbf{D}) - \mathbb{E}_{\mathbf{x}\sim\mathbb{P}} f_{\mathbf{x}}(\mathbf{D})| \leqslant \eta_n. \tag{33}$$

where η_n depends on the penalty g, the data distribution $\mathbb{P}$, the set $\mathcal{S}(r)$ (via its covering number) and the targeted probability level $1 - 2e^{-x}$. Thus, with high probability,

$$\Delta F_{\mathbf{X}}(r) \geq \Delta f_{\mathbb{P}}(r) - 2\eta_n$$

with

$$\Delta f_{\mathbb{P}}(r) \quad \triangleq \quad \inf_{\mathbf{D}\in\mathcal{S}(r)} \Delta f_{\mathbb{P}}(\mathbf{D}; \mathbf{D}^o) \tag{34}$$

$$\text{where } f_{\mathbb{P}}(\mathbf{D}) \quad \triangleq \quad \mathbb{E}_{\mathbf{x}\sim\mathbb{P}} f_{\mathbf{x}}(\mathbf{D}). \tag{35}$$

As a result, showing that $\Delta f_{\mathbb{P}}(r) > 0$ will imply that, with high probability, the function $\mathbf{D} \mapsto F_{\mathbf{X}}(\mathbf{D})$ admits a local minimum $\hat{\mathbf{D}}$ such that $\|\hat{\mathbf{D}} - \mathbf{D}^o\|_F < r$, provided that the number of training samples n satisfies $\eta_n < \Delta f_{\mathbb{P}}(r)/2$. For the ℓ^1 penalty $g(\boldsymbol{\alpha}) = \lambda\|\boldsymbol{\alpha}\|_1$, the generative model considered in Section II-C1, and the oblique manifold $\mathcal{D}$, a direct application of the results of [21] yields $\eta_n \leqslant c\sqrt{\frac{(mp+x)\cdot\log n}{n}}$ for some explicit constant c. The desired result follows when the number of training samples satisfies

$$\frac{n}{\log n} \geq (mp + x) \cdot \frac{4c^2}{[\Delta f_{\mathbb{P}}(r)]^2}.$$

This is slightly too weak in our context where the interesting regime is when $\Delta f_{\mathbb{P}}(r)$ is non-negative but small. Typically, in the noiseless regime, we target an arbitrary small radius $r > 0$ through a penalty factor $\lambda \asymp r$ and get $\Delta f_{\mathbb{P}}(r) = O(r^2)$. Since c is a fixed constant, the above direct sample complexity estimates apparently suggests $n/\log n = \Omega(mpr^{-2})$, a number of training sample that grow arbitrarily large when the targeted resolution r is arbitrarily small. Even though this is the behavior displayed in recent related work [35, 5, 36], this is not fully satisfactory, and we get more satisfactory *resolution independent* sample complexity estimates $n = \Omega(mp)$ through more refined Rademacher averages and Slepian's lemma in Section IV-G. Incidentally we also gain a $\log n$ factor.

B. Robustness to outliers

Training collections are sometimes contaminated by *outliers*, i.e., training samples somehow irrelevant to the considered training task in the sense that they do not share the "dominant" properties of the training set. Considering a collection $\mathbf{X}$ of n_{in} *inliers* and n_{out} outliers, and $\mathbf{X}_{\text{in}}$ (resp. $\mathbf{X}_{\text{out}}$) the matrix extracted from $\mathbf{X}$ by keeping only its columns associated to inliers (resp. outliers), we have

$$(n_{\text{in}} + n_{\text{out}}) \cdot \Delta F_{\mathbf{X}}(r) \geq n_{\text{in}} \cdot \Delta F_{\mathbf{X}_{\text{in}}}(r) + n_{\text{out}} \cdot \Delta F_{\mathbf{X}_{\text{out}}}(r).$$

As a result, the robustness of the learning process with respect to the contamination of a "clean" training set $\mathbf{X}_{\text{in}}$ with outliers will follow from two quantitative bounds: a lower bound $\Delta F_{\mathbf{X}_{\text{in}}}(r) > 0$ for the contribution of inliers, together with an upper bound on the perturbating effects $n_{\text{out}} \cdot |\Delta F_{\mathbf{X}_{\text{out}}}(r)|$ of outliers.

For classical penalty functions g with $g(\mathbf{0}) = 0$, such as sparsity-inducing norms, one easily checks that for any $\mathbf{D}$ we have $0 \leqslant n_{\text{out}} \cdot F_{\mathbf{X}_{\text{out}}}(\mathbf{D}) \leqslant \frac{1}{2}\|\mathbf{X}_{\text{out}}\|_F^2$ [see, e.g., 21] hence the upper bound

$$n_{\text{out}} \cdot |\Delta F_{\mathbf{X}_{\text{out}}}(r)| \leqslant \tfrac{1}{2}\|\mathbf{X}_{\text{out}}\|_F^2. \tag{36}$$

This implies the robustness to outliers provided that:

$$\|\mathbf{X}_{\text{out}}\|_F^2 < 2n_{\text{in}} \cdot \Delta F_{\mathbf{X}_{\text{in}}}(r).$$

In our context, in the interesting regime we have (with high probability) $\Delta F_{\mathbf{X}_{\text{in}}}(r) = O(r^2)$ with r arbitrarily small and $\lambda \asymp r$. Hence, the above analysis suggests that $\|\mathbf{X}_{\text{out}}\|_F^2/n_{\text{in}}$ should scale as $O(r^2)$: the more "precision" we require (the smaller r), the least robust with respect to outliers.

In fact, the considered learning approach is much more robust to outliers that it would seem at first sight: in Section IV-G4, we establish an improved bound on $n_{\text{out}} \cdot |\Delta F_{\mathbf{X}_{\text{out}}}(r)|$: under the assumption that $\mathbf{D}^o$ is complete (i.e., $\mathbf{D}^o$ is a frame with lower frame bound A^o), we obtain when $\lambda \asymp r$

$$n_{\text{out}} \cdot |\Delta F_{\mathbf{X}_{\text{out}}}(r)| \leqslant \frac{18p^{3/2}}{\sqrt{k}}\|\mathbf{X}_{\text{out}}\|_{1,2}\Big(\mathbb{E}|\alpha|\frac{r\bar{\lambda}}{(A^o)^{3/2}}\Big), \tag{37}$$

where $\|\mathbf{X}_{\text{out}}\|_{1,2} \triangleq \sum_{i \in \text{out}} \|\mathbf{x}^i\|_2$. The upper bound on $n_{\text{out}} \cdot |\Delta F_{\mathbf{X}_{\text{out}}}(r)|$ now scales as $O(r^2)$ when $\lambda \asymp r$, and we have robustness to outliers provided that

$$\|\mathbf{X}_{\text{out}}\|_{1,2} < n_{\text{in}} \cdot \frac{\Delta F_{\mathbf{X}_{\text{in}}}(r)}{r^2} \cdot \frac{r}{\bar{\lambda}} \cdot \big[\frac{\sqrt{k}}{18p^{3/2}}\frac{(A^o)^{3/2}}{\mathbb{E}|\alpha|}\big].$$

This is now resolution-independent in the regime $\bar{\lambda} \asymp r$.

C. Closed-form expression

As the reader may have guessed, lower-bounding $\Delta f_{\mathbb{P}}(r)$ is the key technical objective of this paper. One of the main difficulties arises from the fact that $f_{\mathbf{x}}(\mathbf{D})$ is only implicitly defined through the minimization of $\mathcal{L}_{\mathbf{x}}(\mathbf{D}, \boldsymbol{\alpha})$ with respect to the coefficients $\boldsymbol{\alpha}$.

From now on we concentrate on the ℓ^1 penalty, $g(\boldsymbol{\alpha}) = \lambda\|\boldsymbol{\alpha}\|_1$. We leverage here a key property of the function $f_{\mathbf{x}}$. Denote by $\boldsymbol{\alpha}^\star = \boldsymbol{\alpha}^\star_{\mathbf{x}}(\mathbf{D}) \in \mathbb{R}^p$ a solution of problem (2), that is, the minimization defining $f_{\mathbf{x}}$. By the convexity of the problem, there always exists such a solution such that, denoting $\mathrm{J} \triangleq \{j \in [\![1; p]\!];\, \boldsymbol{\alpha}^\star_j \neq 0\}$ its support, the dictionary $\mathbf{D}_{\mathrm{J}} \in \mathbb{R}^{m \times |\mathrm{J}|}$ restricted to the atoms indexed by J has linearly independent columns (hence $\mathbf{D}_{\mathrm{J}}^\top \mathbf{D}_{\mathrm{J}}$ is invertible) [16]. Denoting $\mathbf{s}^\star = \mathbf{s}^\star_{\mathbf{x}}(\mathbf{D}) \in \{-1, 0, 1\}^p$ the sign of $\boldsymbol{\alpha}^\star$ and J its support, $\boldsymbol{\alpha}^\star$ has a closed-form expression in terms of $\mathbf{D}_{\mathrm{J}}$, $\mathbf{x}$ and $\mathbf{s}^\star$ [see, e.g., 44, 16]. This property is appealing in that it makes it possible to obtain a closed-form expression for $f_{\mathbf{x}}$, *provided that we can control the sign pattern of $\boldsymbol{\alpha}^\star$*. In light of this remark, it is natural to define:

Definition 2. *Let $\mathbf{s} \in \{-1, 0, 1\}^p$ be an arbitrary sign vector and $\mathrm{J} = \mathrm{J}(\mathbf{s})$ be its support. For $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{D} \in \mathbb{R}^{m \times p}$, we define*

$$\phi_{\mathbf{x}}(\mathbf{D}|\mathbf{s}) \triangleq \inf_{\boldsymbol{\alpha} \in \mathbb{R}^p,\ \text{support}(\boldsymbol{\alpha}) \subset \mathrm{J}} \tfrac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda \mathbf{s}^\top \boldsymbol{\alpha}. \tag{38}$$

Whenever $\mathbf{D}_{\mathrm{J}}^\top \mathbf{D}_{\mathrm{J}}$ is invertible, the minimum is achieved at $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}_{\mathbf{x}}(\mathbf{D}|\mathbf{s})$ defined by

$$\hat{\boldsymbol{\alpha}}_{\mathrm{J}} = \mathbf{D}_{\mathrm{J}}^+ \mathbf{x} - \lambda\big(\mathbf{D}_{\mathrm{J}}^\top \mathbf{D}_{\mathrm{J}}\big)^{-1}\mathbf{s}_{\mathrm{J}} \in \mathbb{R}^{\mathrm{J}} \quad \textit{and} \quad \hat{\boldsymbol{\alpha}}_{\mathrm{J}^c} = \mathbf{0}, \tag{39}$$

and we have

$$\phi_{\mathbf{x}}(\mathbf{D}|\mathbf{s}) = \frac{1}{2}\big[\|\mathbf{x}\|_2^2 - (\mathbf{D}_{\mathrm{J}}^\top \mathbf{x} - \lambda \mathbf{s}_{\mathrm{J}})^\top (\mathbf{D}_{\mathrm{J}}^\top \mathbf{D}_{\mathrm{J}})^{-1}(\mathbf{D}_{\mathrm{J}}^\top \mathbf{x} - \lambda \mathbf{s}_{\mathrm{J}})\big]. \tag{40}$$

Moreover, if $\operatorname{sign}(\hat{\boldsymbol{\alpha}}) = \mathbf{s}$, then

$$\begin{aligned} \phi_{\mathbf{x}}(\mathbf{D}|\mathbf{s}) &= \min_{\boldsymbol{\alpha} \in \mathbb{R}^p,\ \operatorname{sign}(\boldsymbol{\alpha}) = \mathbf{s}} \tfrac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda \mathbf{s}^\top \boldsymbol{\alpha} \\ &= \min_{\boldsymbol{\alpha} \in \mathbb{R}^p,\ \operatorname{sign}(\boldsymbol{\alpha}) = \mathbf{s}} \mathcal{L}_{\mathbf{x}}(\mathbf{D}, \boldsymbol{\alpha}) = \mathcal{L}_{\mathbf{x}}(\mathbf{D}, \hat{\boldsymbol{\alpha}}). \end{aligned} \tag{41}$$

Hence, with $\mathbf{s}^\star$ the sign of a minimizer $\boldsymbol{\alpha}^\star$, we have $f_{\mathbf{x}}(\mathbf{D}) = \phi_{\mathbf{x}}(\mathbf{D}|\mathbf{s}^\star)$. While $\boldsymbol{\alpha}^\star$ is unknown, in light of the generative model $\mathbf{x} = \mathbf{D}^o\boldsymbol{\alpha}^o + \boldsymbol{\varepsilon}$ for inliers (see Section II-C1), a natural guess for $\mathbf{s}^\star$ is $\mathbf{s}^o = \operatorname{sign}(\boldsymbol{\alpha}^o)$.

D. Closed form expectation and its lower bound

Under decorrelation assumptions, one can compute

$$\Delta\phi_{\mathbb{P}}(\mathbf{D}; \mathbf{D}^o|\mathbf{s}^o) \triangleq \mathbb{E}\ \Delta\phi_{\mathbf{x}}(\mathbf{D}; \mathbf{D}^o|\mathbf{s}^o). \tag{42}$$

We use the shorthands $\mathbf{G}_{\mathrm{J}}^o = \mathbf{G}_{\mathrm{J}}(\mathbf{D}^o)$, $\mathbf{H}_{\mathrm{J}}^o \triangleq \mathbf{H}_{\mathrm{J}}(\mathbf{D}^o)$, and $\mathbf{P}_{\mathrm{J}}^o \triangleq \mathbf{P}_{\mathrm{J}}(\mathbf{D}^o)$.

Proposition 1. *Assume that both $\underline{\delta}_k(\mathbf{D}^o) < 1$ and $\underline{\delta}_k(\mathbf{D}) < 1$ so that $\mathbf{D}_\mathrm{J}$ and $\mathbf{D}^o_\mathrm{J}$ have linearly independent columns for any* J *of size k. Under Assumption A we have*

$$\begin{aligned}\Delta\phi_{\mathbb{P}}(\mathbf{D};\mathbf{D}^o|\mathbf{s}^o) &= \tfrac{\mathbb{E}\{\alpha^2\}}{2}\cdot\ \mathbb{E}_J\mathrm{Tr}[\mathbf{D}^o_\mathrm{J}]^\top(\mathbf{I}-\mathbf{P}_\mathrm{J})\mathbf{D}^o_\mathrm{J}\\ &\quad -\lambda\cdot\mathbb{E}\{|\alpha|\}\cdot\ \mathbb{E}_J\mathrm{Tr}\big([\mathbf{D}^o_\mathrm{J}]^+-\mathbf{D}^+_\mathrm{J}\big)\mathbf{D}^o_\mathrm{J}\\ &\quad +\tfrac{\lambda^2}{2}\cdot\ \mathbb{E}_J\mathrm{Tr}\left(\mathbf{H}^o_\mathrm{J}-\mathbf{H}_\mathrm{J}\right). \qquad (43)\end{aligned}$$

The proof is in Appendix B. In light of this result we switch to the reduced regularization parameter $\bar{\lambda} \triangleq \frac{\lambda}{\mathbb{E}\ |\alpha|}$. Our main bound leverages Proposition 1 and Lemma 7 (Appendix C).

Proposition 2. *Consider a dictionary $\mathbf{D}^o \in \mathbb{R}^{m\times p}$ such that*

$$\begin{aligned}\underline{\delta}_k(\mathbf{D}^o) &\leqslant \frac{1}{4} \qquad (44)\\ k &\leqslant \frac{p}{16(\|\mathbf{D}^o\|_2+1)^2}. \qquad (45)\end{aligned}$$

Under the basic signal model (Assumption A):

- *when $\bar{\lambda} \leqslant 1/4$, for any $r \leqslant 0.15$ we have, uniformly for all $\mathbf{D} \in \mathcal{S}(r;\mathbf{D}^o)$:*

$$\Delta\phi_{\mathbb{P}}(\mathbf{D};\mathbf{D}^o|\mathbf{s}^o) \geq \frac{\mathbb{E}\ \alpha^2}{8}\cdot\frac{k}{p}\cdot r\Big(r-r_{\min}(\bar{\lambda})\Big). \qquad (46)$$

  with $r_{\min}(\bar{\lambda}) \triangleq \frac{2}{3}C_{\min}\cdot\bar{\lambda}\cdot\big(1+2\bar{\lambda}\big)$.
- *if in addition $\bar{\lambda} < \frac{3}{20C_{\min}}$, then $r_{\min}(\bar{\lambda}) < 0.15$ and the lower bound in (46) is non-negative for all $r \in (r_{\min}(\bar{\lambda}), 0.15]$.*

The proof is in Appendix C.

E. Exact recovery

The analysis of $\Delta\phi_{\mathbb{P}}(\mathbf{D};\mathbf{D}^o|\mathbf{s}^o)$ would suffice for our needs if the sign of the minimizer $\hat{\boldsymbol{\alpha}}_\mathbf{x}(\mathbf{D})$ was guaranteed to always match the ground truth sign $\mathbf{s}^o$. In fact, if the equality $\mathrm{sign}(\hat{\boldsymbol{\alpha}}_\mathbf{x}(\mathbf{D})) = \mathbf{s}^o$ held unconditionally on the radius r, then the analysis conducted up to Proposition 2 would show (assuming a large enough number of training samples) the existence of a local minimum of $F_\mathbf{X}(\cdot)$ within a ball $\mathcal{B}((1+o(1))r_{\min})$. Moreover, given the lower bound provided by Proposition 2, the *global minimum of $F_\mathbf{X}(\cdot)$ restricted over the ball $\mathcal{B}((1+o(1))r_{\min})$* would in fact be *global over the potentially much larger ball $\mathcal{B}(0.15)$*.

However, with the basic signal model (Assumption A), the equality $\Delta f_\mathbf{x}(\mathbf{D};\mathbf{D}^o) = \Delta\phi_\mathbf{x}(\mathbf{D};\mathbf{D}^o|\mathbf{s}^o)$ has no reason to hold in general. This motivates the introduction of stronger assumptions involving the cumulative coherence of $\mathbf{D}^o$ and the bounded signal model (Assumption B).

Proposition 3 (Exact recovery; bounded model)**.** *Let $\mathbf{D}^o$ be a dictionary in $\mathbb{R}^{m\times p}$ such that*

$$\mu^o_k \triangleq \mu_k(\mathbf{D}^o) < \frac{1}{2}. \qquad (47)$$

Consider the bounded signal model (Assumption B), $\bar{\lambda} \leqslant \frac{\alpha}{2\cdot\mathbb{E}\ |\alpha|}$ and $r < C_{\max}\cdot\bar{\lambda}$ where

$$C_{\max} \triangleq \frac{2}{7}\cdot\frac{\mathbb{E}\ |\alpha|}{M_{\boldsymbol{\alpha}}}\cdot(1-2\mu^o_k). \qquad (48)$$

If the relative noise level satisfies

$$\frac{M_{\boldsymbol{\varepsilon}}}{M_{\boldsymbol{\alpha}}} < \frac{7}{2}\left(C_{\max}\cdot\bar{\lambda}-r\right), \qquad (49)$$

then, for $\mathbf{D} \in \mathcal{D}$ such that $\|\mathbf{D}-\mathbf{D}^o\|_F = r$, $\hat{\boldsymbol{\alpha}}_\mathbf{x}(\mathbf{D}|\mathbf{s}^o)$ is almost surely the unique minimizer in $\mathbb{R}^p$ of $\boldsymbol{\alpha} \mapsto \frac{1}{2}\|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}\|_2^2+\lambda\|\boldsymbol{\alpha}\|_1$, and we have

$$\begin{aligned}\mathrm{sign}(\hat{\boldsymbol{\alpha}}_\mathbf{x}(\mathbf{D}|\mathbf{s}^o)) &= \mathbf{s}^o \qquad (50)\\ f_\mathbf{x}(\mathbf{D}) &= \phi_\mathbf{x}(\mathbf{D}|\mathbf{s}^o) \qquad (51)\\ \Delta f_\mathbf{x}(\mathbf{D};\mathbf{D}^o) &= \Delta\phi_\mathbf{x}(\mathbf{D};\mathbf{D}^o|\mathbf{s}^o). \qquad (52)\end{aligned}$$

F. Proof of Theorem 1

Noticing that $\underline{\alpha} \leqslant M_{\boldsymbol{\alpha}}$, we let the reader check that assumption (19) implies $\frac{\underline{\alpha}}{4\mathbb{E}\ |\alpha|} \leqslant \frac{3}{20C_{\min}}$. Hence, by (22) we have

$$\bar{\lambda} < \tfrac{\underline{\alpha}}{4\mathbb{E}\ |\alpha|} \leqslant \min\left(\frac{1}{4}, \frac{3}{20C_{\min}}, \frac{\underline{\alpha}}{2\cdot\mathbb{E}\ |\alpha|}\right),$$

where we use the inequality $\underline{\alpha} \leqslant \mathbb{E}\ |\alpha|$. Assumptions (17) and (18) imply (44) and (45), and we have $\bar{\lambda} \leqslant \min(\frac{1}{4}, \frac{3}{20C_{\min}})$, hence we can leverage Proposition 2. Similarly, assumption (17) implies (47), and we have $\bar{\lambda} \leqslant \frac{\underline{\alpha}}{2\cdot\mathbb{E}\ |\alpha|}$, hence we can also apply Proposition 3. Furthermore, assumption (19) implies $C_{\min} < C_{\max}$, and we have $\bar{\lambda} \leqslant \frac{1}{4}$, hence $\frac{2}{3}C_{\min}\cdot\bar{\lambda}\cdot(1+2\bar{\lambda}) \leqslant C_{\min}\cdot\bar{\lambda} < C_{\max}\cdot\bar{\lambda}$. Finally, the fact that $\bar{\lambda} \leqslant \frac{\underline{\alpha}}{2\cdot\mathbb{E}\ |\alpha|}$ further implies $C_{\max}\cdot\bar{\lambda} \leqslant 0.15$. Putting the pieces together, we have $\frac{2}{3}C_{\min}\cdot\bar{\lambda}\cdot(1+2\bar{\lambda}) \leqslant C_{\min}\cdot\bar{\lambda} < C_{\max}\cdot\bar{\lambda} \leqslant 0.15$, and for any $r \in \big(C_{\min}\cdot\bar{\lambda}, C_{\max}\cdot\bar{\lambda}\big)$ we obtain

$$\Delta f_{\mathbb{P}}(r) \geq \frac{\mathbb{E}\ \alpha^2}{8}\cdot\frac{k}{p}\cdot r\Big(r-C_{\min}\cdot\bar{\lambda}\Big) > 0. \qquad (53)$$

as soon as the relative noise level satisfies (49).

G. Proof of Theorem 2

In order to prove Theorem 2, we need to control the deviation of the average of functions $\Delta\phi_{\mathbf{x}^i}(\mathbf{D};\mathbf{D}^o|\mathbf{s}^o)$ around its expectation, uniformly in the ball $\{\mathbf{D},\ \|\mathbf{D}-\mathbf{D}^o\|_F \leqslant r\}$.

1) Review of Rademacher averages.: We first review results on Rademacher averages. Let $\mathcal{F}$ be a set of measurable functions on a measurable set $\mathcal{X}$, and n i.i.d. random variables $X_1,\ldots,X_n$, in $\mathcal{X}$. We assume that all functions are bounded by B (i.e., $|f(X)| \leqslant B$ almost surely). Using usual symmetrisation arguments [8, Sec. 9.3], we get

$$\begin{aligned}&\mathbb{E}_X\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n f(X_i)-\mathbb{E}_X f(X)\right)\\ &\qquad\leqslant 2\mathbb{E}_{X,\varepsilon}\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n\varepsilon_i f(X_i)\right),\end{aligned}$$

where $\varepsilon_i, 1 \leqslant i \leqslant n$ are independent Rademacher random variables, i.e., with values 1 and -1 with equal probability $\frac{1}{2}$. Conditioning on the data $X_1,\ldots,X_n$, the function $\varepsilon \in \mathbb{R}^n \mapsto \sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n\varepsilon_i f(X_i)\right)$ is convex. Therefore, if η is an independent standard normal vector, by Jensen's inequality,

using that the normal distribution is symmetric and $\mathbb{E}|\eta_i| = \sqrt{2/\pi}$, we get

$$\begin{aligned}\mathbb{E}_{X,\varepsilon} \sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right) &= \sqrt{\pi/2}\cdot \mathbb{E}_{X,\varepsilon}\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{E}|\eta_i| f(X_i)\right)\\ &\leqslant \sqrt{\pi/2}\cdot \mathbb{E}_{X,\eta}\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n \eta_i f(X_i)\right).\end{aligned}$$

Moreover, the random variable $Z = \sup_{f\in\mathcal{F}}\big(\frac{1}{n}\sum_{i=1}^n (f(X_i) - \mathbb{E}f(X))\big)$ only changes by at most $2B/n$ when changing a single of the n random variables. Therefore, by Mac Diarmid's inequality, we obtain with probability at least $1-e^{-x}$: $Z \leqslant \mathbb{E}Z + B\sqrt{\frac{2x}{n}}$. We may thus combine all of the above, to get, with probability at least $1-e^{-x}$:

$$\begin{aligned}&\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X)\right)\\ &\qquad\leqslant 2\sqrt{\pi/2}\cdot \mathbb{E}_{X,\eta}\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n \eta_i f(X_i)\right) + B\sqrt{\frac{2x}{n}}.\end{aligned} \tag{54}$$

Note that in the equation above, we may also consider the absolute value of the deviation by redefining $\mathcal{F}$ as $\mathcal{F}\cup(-\mathcal{F})$.

We may now prove two lemmas that will prove useful in our uniform deviation bound.

Lemma 1 (Concentration of a real-valued function on matrices $\mathbf{D}$)**.** *If $h_1,\dots,h_n$ are random real-valued i.i.d. functions on $\{\mathbf{D},\ \|\mathbf{D}-\mathbf{D}^o\|_F \leqslant r\}$, such that they are almost surely bounded by B on this set, as well as, R-Lipschitz-continuous (with respect to the Frobenius norm). Then, with probability greater than $1-e^{-x}$:*

$$\sup_{\|\mathbf{D}-\mathbf{D}^o\|_F\leqslant r}\left|\frac{1}{n}\sum_{i=1}^n h_i(\mathbf{D}) - \mathbb{E}h(\mathbf{D})\right| \leqslant 4\sqrt{\tfrac{\pi}{2}}\frac{Rr\sqrt{mp}}{\sqrt{n}} + B\sqrt{\frac{2x}{n}}.$$

Proof. Given Eq. (54), we only need to provide an upper-bound on $\mathbb{E}\sup_{\|\mathbf{D}-\mathbf{D}^o\|_F\leqslant r}\big|\frac{1}{n}\sum_{i=1}^n \eta_i h_i(\mathbf{D})\big|$ for η a standard normal vector. Conditioning on the draw of functions $h_1,\dots,h_n$, consider two Gaussian processes, indexed by $\mathbf{D}$, $A_{\mathbf{D}} = \frac{1}{n}\sum_{i=1}^n \eta_i h_i(\mathbf{D})$ and $C_{\mathbf{D}} = \frac{R}{\sqrt{n}}\sum_{i=1}^m\sum_{j=1}^p \zeta_{ij}(\mathbf{D}-\mathbf{D}^o)_{ij}$, where η and ζ are standard Gaussian vectors. We have, for all $\mathbf{D}$ and $\mathbf{D}'$, $\mathbb{E}|A_{\mathbf{D}} - A_{\mathbf{D}'}|^2 \leqslant \frac{R^2}{n}\|\mathbf{D}-\mathbf{D}'\|_F^2 = \mathbb{E}|C_{\mathbf{D}} - C_{\mathbf{D}'}|^2$.

Hence, by Slepian's lemma [30, Sec. 3.3], $\mathbb{E}\sup_{\|\mathbf{D}-\mathbf{D}^o\|_F\leqslant r} A_{\mathbf{D}} \leqslant \mathbb{E}\sup_{\|\mathbf{D}-\mathbf{D}^o\|_F\leqslant r} C_{\mathbf{D}} = \frac{Rr}{\sqrt{n}}\mathbb{E}\|\zeta\|_F \leqslant \frac{Rr\sqrt{mp}}{\sqrt{n}}$. Thus, by applying the above reasoning to the functions h_i and $-h_i$ and taking the expectation with respect to the draw of $h_1,\dots,h_n$, we get: $\mathbb{E}\sup_{\|\mathbf{D}-\mathbf{D}^o\|_F\leqslant r}\big|\frac{1}{n}\sum_{i=1}^n \eta_i h_i(\mathbf{D})\big| \leqslant 2\frac{Rr\sqrt{mp}}{\sqrt{n}}$, hence the result. □

Lemma 2 (Concentration of matrix-valued function on matrices $\mathbf{D}$)**.** *Consider $g_1,\dots,g_n$ random i.i.d. functions on $\{\mathbf{D},\ \|\mathbf{D}-\mathbf{D}^o\|_F \leqslant r\}$, with values in real symmetric matrices of size s. Assume that these functions are almost surely bounded by B (in operator norm) on this set, as well as, R-Lipschitz-continuous (with respect to the Frobenius norm, i.e., $|||g_i(\mathbf{D})|||_2 \leqslant B$ and $|||g_i(\mathbf{D}) - g_i(\mathbf{D}')|||_2 \leqslant R\|\mathbf{D}-\mathbf{D}'\|_F$). Then, with probability greater than $1-e^{-x}$:*

$$\sup_{\|\mathbf{D}-\mathbf{D}^o\|_F\leqslant r}\left|\left|\left|\frac{1}{n}\sum_{i=1}^n g_i(\mathbf{D}) - \mathbb{E}g(\mathbf{D})\right|\right|\right|_2 \leqslant 4\sqrt{\pi/2}\left(\frac{\sqrt{2mp}Rr}{\sqrt{n}} + \frac{B\sqrt{8s}}{\sqrt{n}}\right) + B\sqrt{\frac{2x}{n}}.$$

Proof. For any symmetric matrix $\mathbf{M}$, $|||\mathbf{M}|||_2 = \sup_{\|\mathbf{z}\|_2\leqslant 1}|\mathbf{z}^\top\mathbf{M}\mathbf{z}|$. Given Eq. (54), we only need to upper-bound

$$\begin{aligned}&\mathbb{E}\sup_{\|\mathbf{D}-\mathbf{D}^o\|_F\leqslant r}\left|\left|\left|\frac{1}{n}\sum_{i=1}^n \eta_i g_i(\mathbf{D})\right|\right|\right|_2\\ &\qquad= \mathbb{E}\sup_{\|\mathbf{D}-\mathbf{D}^o\|_F\leqslant r,\|\mathbf{z}\|_2\leqslant 1}\Big|\frac{1}{n}\sum_{i=1}^n \eta_i \mathbf{z}^\top g_i(\mathbf{D})\mathbf{z}\Big|,\end{aligned}$$

for η a standard normal vector. We thus consider two Gaussian processes, indexed by $\mathbf{D}$ and $\|\mathbf{z}\|_2 \leqslant 1$, $A_{\mathbf{D},\mathbf{z}} = \frac{1}{n}\sum_{i=1}^n \eta_i \mathbf{z}^\top g_i(\mathbf{D})\mathbf{z}$ and $C_{\mathbf{D},\mathbf{z}} = \frac{\sqrt{2}R}{\sqrt{n}}\sum_{i=1}^m\sum_{j=1}^p \zeta_{ij}(\mathbf{D}-\mathbf{D}^o)_{ij} + \frac{2B\sqrt{2}}{\sqrt{n}}\sum_{i=1}^s \xi_i \mathbf{z}_i$, where η and ζ are, again, standard normal vectors. We have, for all $(\mathbf{D},\mathbf{z})$ and $(\mathbf{D}',\mathbf{z}')$,

$$\begin{aligned}&\mathbb{E}|A_{\mathbf{D},\mathbf{z}} - A_{\mathbf{D}',\mathbf{z}'}|^2\\ &\leqslant \tfrac{1}{n}\big(R\|\mathbf{D}-\mathbf{D}'\|_F + \big|\mathbf{z}^\top g_i(\mathbf{D})\mathbf{z} - (\mathbf{z}')^\top g_i(\mathbf{D})\mathbf{z}'\big|\big)^2\\ &\leqslant \tfrac{1}{n}\big(R\|\mathbf{D}-\mathbf{D}'\|_F + 2B\|\mathbf{z}-\mathbf{z}'\|_2\big)^2\\ &\leqslant \tfrac{2}{n}R^2\|\mathbf{D}-\mathbf{D}'\|_F^2 + \tfrac{8B^2}{n}\|\mathbf{z}-\mathbf{z}'\|_2^2 = \mathbb{E}|C_{\mathbf{D},\mathbf{z}} - C_{\mathbf{D}',\mathbf{z}'}|^2.\end{aligned}$$

Applying Slepian's lemma to $A_{\mathbf{D},\mathbf{z}}$ and to $-A_{\mathbf{D},\mathbf{z}}$, we get

$$\begin{aligned}&\mathbb{E}\sup_{\|\mathbf{D}-\mathbf{D}^o\|_F\leqslant r}\left|\left|\left|\tfrac{1}{n}\sum_{i=1}^n \eta_i g_i(\mathbf{D})\right|\right|\right|_2\\ &\leqslant 2\tfrac{\sqrt{2}Rr}{\sqrt{n}}\mathbb{E}\|\zeta\|_F + 2\tfrac{2B\sqrt{2}}{\sqrt{n}}\mathbb{E}\|\xi\|_2\\ &\leqslant \tfrac{\sqrt{8mp}Rr}{\sqrt{n}} + \tfrac{B\sqrt{32s}}{\sqrt{n}},\end{aligned}$$

hence the result. □

2) Decomposition of $\Delta\phi_{\mathbf{x}}(\mathbf{D};\mathbf{D}^o|\mathbf{s}^o)$.: Our goal is to uniformly bound the deviations of $\mathbf{D}\mapsto\Delta\phi_{\mathbf{x}}(\mathbf{D};\mathbf{D}^o|\mathbf{s}^o)$ from its expectation on $\mathcal{S}(\mathbf{D}^o;r)$. With the notations of Appendix D, we use the following decomposition

$$\begin{aligned}\Delta\phi_{\mathbf{x}}(\mathbf{D};\mathbf{D}^o|\mathbf{s}^o) &= \big[\Delta\phi_{\mathbf{x}}(\mathbf{D};\mathbf{D}^o|\mathbf{s}^o) - \Delta\phi_{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\mathbf{D};\mathbf{D}^o)\big]\\ &\quad + \Delta\phi_{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\mathbf{D};\mathbf{D}^o)\\ &= h(\mathbf{D}) + \Delta\phi_{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\mathbf{D};\mathbf{D}^o),\end{aligned}$$

with $\Delta\phi_{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\mathbf{D};\mathbf{D}^o) := \frac{1}{2}[\boldsymbol{\alpha}^o]^\top[\mathbf{D}^o]^\top(\mathbf{I}-\mathbf{P}_{\mathrm{J}})\mathbf{D}^o\boldsymbol{\alpha}^o$ and $h(\mathbf{D}) := \big(\Delta\phi_{\mathbf{x}}(\mathbf{D};\mathbf{D}^o|\mathbf{s}^o) - \Delta\phi_{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\mathbf{D};\mathbf{D}^o)\big)$.

For the first term, by Lemma 9 in Appendix D, the function h on $\mathcal{B}(\mathbf{D}^o;r)$ is almost surely L-Lipschitz-continuous with respect to the Frobenius metric and almost surely bounded by $c = Lr$, where we denote

$$\sqrt{1-\underline{\delta}} \triangleq \sqrt{1-\underline{\delta}_k(\mathbf{D}^o)} - r > 0$$

and

$$L \triangleq \frac{1}{\sqrt{1-\underline{\delta}}} \cdot \left(M_{\boldsymbol{\varepsilon}} + \frac{\lambda\sqrt{k}}{\sqrt{1-\underline{\delta}}} \right) \cdot \left(2\sqrt{1+\overline{\delta}_k(\mathbf{D}^o)} M_{\boldsymbol{\alpha}} + M_{\boldsymbol{\varepsilon}} + \frac{\lambda\sqrt{k}}{\sqrt{1-\underline{\delta}}} \right).$$

We can thus apply Lemma 1, with $B = c = Lr$ and $R = L$.

Regarding the second term, since $(\mathbf{I} - \mathbf{P}_{\mathrm{J}})\mathbf{D}\boldsymbol{\alpha}^o = (\mathbf{I} - \mathbf{P}_{\mathrm{J}})\mathbf{D}_{\mathrm{J}}\boldsymbol{\alpha}^o_{\mathrm{J}} = 0$, one can rewrite it as

$$\begin{aligned}
&\Delta\phi_{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\mathbf{D};\mathbf{D}^o) \\
&= \frac{1}{2}[\boldsymbol{\alpha}^o]^\top (\mathbf{D}-\mathbf{D}^o)^\top (\mathbf{I}-\mathbf{P}_{\mathrm{J}})(\mathbf{D}-\mathbf{D}^o)\boldsymbol{\alpha}^o \\
&= \frac{1}{2}\mathrm{vec}(\mathbf{D}-\mathbf{D}^o)^\top \left\{ \boldsymbol{\alpha}^o[\boldsymbol{\alpha}^o]^\top \otimes (\mathbf{I}-\mathbf{P}_{\mathrm{J}}) \right\} \mathrm{vec}(\mathbf{D}-\mathbf{D}^o).
\end{aligned}$$

where $\mathbf{A} \otimes \mathbf{B}$ denotes the Kronecker product between two matrices (see, e.g., [22]). Thus, with $g(\mathbf{D}) = \boldsymbol{\alpha}^o[\boldsymbol{\alpha}^o]^\top \otimes (\mathbf{I} - \mathbf{P}_{\mathrm{J}})$ a random matrix-valued function with $s = mp$, we have an upper-bound of $B' = M_{\boldsymbol{\alpha}}^2$ (as the eigenvalues of $\mathbf{A} \otimes \mathbf{B}$ are products of eigenvalues of $\mathbf{A}$ and eigenvalues of $\mathbf{B}$ [22]) and, by Lemma 4 and Lemma 5 in Appendix D, a Lipschitz-constant $R' = M_{\boldsymbol{\alpha}}^2(1-\underline{\delta})^{-1/2}$. We may thus apply Lemma 2 to show that uniformly, the deviation of $\Delta\phi_{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\mathbf{D};\mathbf{D}^o)$ are bounded by $\|\mathrm{vec}(\mathbf{D}-\mathbf{D}^o)\|_2^2 = r^2$ times the deviations of $g(\mathbf{D})$ in operator norm.

We thus get, with probability at least $1 - 2e^{-x}$, deviations from the expectations upper-bounded by:

$$\begin{aligned}
&4\sqrt{\tfrac{\pi}{2}}\tfrac{Lr\sqrt{mp}}{\sqrt{n}} + Lr\sqrt{\tfrac{2x}{n}} \\
&\quad + r^2\left(4\sqrt{\tfrac{\pi}{2}}\left(\tfrac{\sqrt{2mp}}{\sqrt{n}} \tfrac{M_{\boldsymbol{\alpha}}^2 r}{\sqrt{1-\underline{\delta}}} + \tfrac{M_{\boldsymbol{\alpha}}^2\sqrt{8mp}}{\sqrt{n}} \right) + M_{\boldsymbol{\alpha}}^2\sqrt{\tfrac{2x}{n}} \right),
\end{aligned}$$

We notice that $R'r = M_{\boldsymbol{\alpha}}^2 r/\sqrt{1-\underline{\delta}} < B'$ since $r < \sqrt{1-\underline{\delta}}$, hence this is less than $\beta r\sqrt{\frac{2x}{n}} + \beta' r\sqrt{\frac{mp}{n}}$ with

$$\beta \triangleq L + rM_{\boldsymbol{\alpha}}^2, \qquad \beta' \triangleq 4\sqrt{\frac{\pi}{2}}\left(L + 3\sqrt{2}rM_{\boldsymbol{\alpha}}^2 \right) \leqslant 12\sqrt{\pi}\beta$$

Overall, with probability at least $1 - 2e^{-x}$, the deviations of $\mathbf{D} \mapsto \Delta\phi_{\mathbf{x}}(\mathbf{D};\mathbf{D}^o|\mathbf{s}^o)$ from its expectation on $\mathcal{S}(\mathbf{D}^o;r)$ are uniformly bounded by $\eta_n \triangleq r(L + M_{\boldsymbol{\alpha}}^2 r)\left(\sqrt{2x/n} + 12\sqrt{\pi mp/n} \right)$.

3) Sample complexity.: As briefly outlined in Section IV-A, with n_{in} inliers, the existence of a local minimum of $F_{\mathbf{X}}(\cdot)$ within a radius r around $\mathbf{D}^o$ is guaranteed with probability at least $1-2e^{-x}$ as soon as $2\eta_{n_{\mathrm{in}}} < \Delta f_{\mathbb{P}}(r)$. Combining with the asymptotic lower bound (53) and the above refined uniform control over $\mathcal{S}(\mathbf{D}^o;r)$, η_n , it is sufficient to have

$$2r(L+M_{\boldsymbol{\alpha}}^2 r)\cdot\left(\sqrt{\tfrac{2x}{n_{\mathrm{in}}}} + 12\sqrt{\tfrac{\pi mp}{n_{\mathrm{in}}}} \right) < \tfrac{\mathbb{E}\,\alpha^2}{8}\cdot\tfrac{k}{p}\cdot r\left(r - C_{\min}\cdot\bar{\lambda} \right).$$

i.e.

$$n_{\mathrm{in}} \geq \left(\sqrt{2x} + 12\sqrt{\pi mp} \right)^2 \cdot \left(\frac{16}{\mathbb{E}\,\alpha^2}\cdot\frac{p}{k}\cdot\frac{L+M_{\boldsymbol{\alpha}}^2 r}{\left(r - C_{\min}\cdot\bar{\lambda} \right)} \right)^2 \quad (55)$$

By (17) we have $\max(\underline{\delta}_k(\mathbf{D}^o), \overline{\delta}_k(\mathbf{D}^o)) \leqslant 1/4$, hence $2\sqrt{1+\overline{\delta}_k(\mathbf{D}^o)} \leqslant \sqrt{5}$. Moreover, since $r < C_{\max}\cdot\bar{\lambda} \leqslant 0.15$, we have $\sqrt{1-\underline{\delta}} = \sqrt{1-\underline{\delta}_k(\mathbf{D}^o)} - r \geq \sqrt{3/4} - 0.15 \geq \sqrt{1/2}$. As a result

$$\begin{aligned}
L &\leqslant \sqrt{2}(M_{\boldsymbol{\varepsilon}} + \lambda\sqrt{2k})\cdot\left(\sqrt{5}M_{\boldsymbol{\alpha}} + M_{\boldsymbol{\varepsilon}} + \lambda\sqrt{2k} \right) \\
&= \sqrt{10}M_{\boldsymbol{\alpha}}(M_{\boldsymbol{\varepsilon}} + \lambda\sqrt{2k}) + \sqrt{2}(M_{\boldsymbol{\varepsilon}} + \lambda\sqrt{2k})^2
\end{aligned}$$

Further, since $\lambda\sqrt{2k} = \bar{\lambda}\mathbb{E}|\alpha|\sqrt{2k} = \bar{\lambda}\sqrt{2/k}\mathbb{E}\|\boldsymbol{\alpha}\|_1 \leqslant \bar{\lambda}\sqrt{2/k}\mathbb{E}\sqrt{k}\|\boldsymbol{\alpha}\|_2 \leqslant \bar{\lambda}\sqrt{2}M_{\boldsymbol{\alpha}}$, we have

$$L + M_{\boldsymbol{\alpha}}^2 r \leqslant \sqrt{20}M_{\boldsymbol{\alpha}}^2\cdot\left(r + \frac{M_{\boldsymbol{\varepsilon}}}{M_{\boldsymbol{\alpha}}} + \bar{\lambda} + \left(\frac{M_{\boldsymbol{\varepsilon}}}{M_{\boldsymbol{\alpha}}} + \bar{\lambda} \right)^2 \right). \quad (56)$$

Eqs (55) and (56) with the bound $(\sqrt{2x}+12\sqrt{\pi mp})^2 \lesssim mp+x$ yield our main sample complexity result (28).

4) Robustness to outliers.: In the presence of outliers, we obtain the naive robustness to outliers (29) in Theorem 2 using the reasoning sketched in Section IV-B. with the naive bound (36). Obtaining the "resolution independent" robustness result (30) requires refining the estimate of the impact of outliers on the cost function $F_{\mathbf{X}}(\mathbf{D})$ by gaining two factors: one factor $O(r)$ (thanks to a Lipschitz property), and one factor $O(\lambda)$ (thanks to the completeness of the dictionary).

a) Gaining a first factor $O(r)$ using a Lipschitz property.: The arguments of [21, Lemma 3 and Corollary 2] can be straightforwardly adapted to show that for any signal $\mathbf{x}$, the function $\mathbf{D} \mapsto f_{\mathbf{x}}(\mathbf{D})$ is uniformly locally Lipschitz on the convex ball $\{\mathbf{D} \in \mathbb{R}^{m\times p} : \|\mathbf{D}-\mathbf{D}^o\|_F \leqslant r\}$ (not restricted to normalized dictionaries). Its Lipschitz constant is bounded by $L_{\mathbf{x}}(r) \triangleq \sup_{\|\mathbf{D}-\mathbf{D}^o\|_F \leqslant r} L_{\mathbf{x}}(\mathbf{D})$ with $L_{\mathbf{x}}(\mathbf{D}) \triangleq \|\boldsymbol{\alpha}\|_2 \cdot \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2$. where we denote $\boldsymbol{\alpha} = \boldsymbol{\alpha}_{\mathbf{x}}(\mathbf{D})$ a coefficient vector minimizing $\mathcal{L}_{\mathbf{x}}(\mathbf{D},\boldsymbol{\alpha})$. It follows that

$$n_{\mathrm{out}}|\Delta F_{\mathbf{X}_{\mathrm{out}}}(r)| \leqslant \left(\sum_{i\in\mathrm{out}} L_{\mathbf{x}^i}(r) \right)\cdot r.$$

Compared to the naive bound (36), we already gained a first factor r, provided we uniformly bound the Lipschitz constants $L_{\mathbf{x}}(r)$.

b) Gaining a second factor $O(\lambda)$ under a completeness assumption.: Introducing

$$C(\mathbf{D}) \triangleq \sup_{\mathbf{u}\neq 0, \mathbf{u}\in\mathrm{span}(\mathbf{D})} \inf_{\boldsymbol{\beta}:\mathbf{D}\boldsymbol{\beta}=\mathbf{u}} \frac{\|\boldsymbol{\beta}\|_1}{\|\mathbf{u}\|_2},$$

we first show that $\|\boldsymbol{\alpha}\|_2 \leqslant C(\mathbf{D})\cdot\|\mathbf{x}\|_2$. Indeed, denoting P the orthonormal projection onto $\mathrm{span}(\mathbf{D})$, by definition of $\boldsymbol{\alpha}$ we have, for any signal $\mathbf{x}$ and any coefficient vector $\boldsymbol{\beta}$,

$$\begin{aligned}
\tfrac{1}{2}\|\mathbf{x}-P\mathbf{x}\|_2^2 \quad &+ \quad \tfrac{1}{2}\|P\mathbf{x}-\mathbf{D}\alpha\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 \\
&= \quad \tfrac{1}{2}\|\mathbf{x}-\mathbf{D}\alpha\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 \\
&\leqslant \quad \tfrac{1}{2}\|\mathbf{x}-\mathbf{D}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \qquad (57) \\
&= \quad \tfrac{1}{2}\|\mathbf{x}-P\mathbf{x}\|_2^2 + \tfrac{1}{2}\|P\mathbf{x}-\mathbf{D}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1.
\end{aligned}$$

Specializing to the minimum ℓ^1 norm vector $\boldsymbol{\beta}$ such that $\mathbf{D}\boldsymbol{\beta} = P\mathbf{x}$ yields

$$\|\boldsymbol{\alpha}\|_2 \leqslant \|\boldsymbol{\alpha}\|_1 \leqslant \|\boldsymbol{\beta}\|_1 \leqslant C(\mathbf{D})\cdot\|P\mathbf{x}\|_2 \leqslant C(\mathbf{D})\cdot\|\mathbf{x}\|_2.$$

To complete the control of $L_{\mathbf{x}}(\mathbf{D}) = \|\boldsymbol{\alpha}\|_2\cdot\|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}\|_2$ we now bound $\|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}\|_2$. A first approach that does not require

any further assumption on $\mathbf{D}$ consists in specializing (57) to $\boldsymbol{\beta}=0$, yielding $\|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}\|_2 \leqslant \|\mathbf{x}\|_2$, $L_{\mathbf{x}}(\mathbf{D}) \leqslant C(\mathbf{D})\cdot\|\mathbf{x}\|_2^2$, and finally

$$n_{\text{out}}|\Delta F_{\mathbf{X}_{\text{out}}}(r)| \leqslant C(r)\cdot\|\mathbf{X}_{\text{out}}\|_F^2\cdot r,$$

with

$$C(r) \triangleq \sup_{\|\mathbf{D}-\mathbf{D}^o\|_F\leqslant r} C(\mathbf{D}).$$

However, as the reader may have noticed, this still lacks one $O(r)$ factor for our needs. This is obtained in the regime of interest $\lambda \asymp r$ under the assumption that $\mathbf{D}$ is *complete* ($\text{span}(\mathbf{D})=\mathbb{R}^m$). In this case we introduce

$$C'(\mathbf{D}) \triangleq \sup_{\mathbf{u}\neq 0}\frac{\|\mathbf{u}\|_2}{\|\mathbf{D}^\top\mathbf{u}\|_\infty} < \infty$$

and

$$C'(r) \triangleq \sup_{\|\mathbf{D}-\mathbf{D}^o\|_F\leqslant r} C'(\mathbf{D}).$$

By the well known optimality conditions for the ℓ^1 regression problem, $\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}_{\mathbf{x}}(\mathbf{D})$ satisfies $\|\mathbf{D}^\top(\mathbf{x}-\mathbf{D}\boldsymbol{\alpha})\|_\infty = \lambda$, hence

$$\|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}\|_2 \quad\leqslant\quad C'(\mathbf{D})\cdot\|\mathbf{D}^\top(\mathbf{x}-\mathbf{D}\boldsymbol{\alpha})\|_\infty \leqslant C'(\mathbf{D})\cdot\lambda.$$

Overall we get $L_{\mathbf{x}}(\mathbf{D}) \leqslant \lambda\cdot C(\mathbf{D})\cdot C'(\mathbf{D})\cdot\|\mathbf{x}\|_2$ and eventually

$$n_{\text{out}}|\Delta F_{\mathbf{X}_{\text{out}}}(r)| \leqslant \sum_{i\in\text{out}}\|\mathbf{x}^i\|_2\cdot\mathbb{E}|\alpha|\cdot C(r)\cdot C'(r)\cdot r\bar{\lambda}. \quad (58)$$

To conclude, we now bound $C(r)$ and $C'(r)$. Note that as soon as $\mathbf{D}\boldsymbol{\beta}=\mathbf{u}$, since $\|\mathbf{u}\|_2^2 = \langle\boldsymbol{\beta},\mathbf{D}^\top\mathbf{u}\rangle \leqslant \|\boldsymbol{\beta}\|_1\|\mathbf{D}^\top\mathbf{u}\|_\infty$, we have $C'(\mathbf{D}) \leqslant C(\mathbf{D})$.

Lemma 3. *Assume $\mathbf{D}\in\mathbb{R}^{m\times p}$ is a frame with lower frame bound A such that $A\|\mathbf{x}\|_2^2 \leqslant \|\mathbf{D}^\top\mathbf{x}\|_2^2$ for any signal $\mathbf{x}$. Then $C'(\mathbf{D}) \leqslant \sqrt{p/A}$. If in addition, $\underline{\delta}_k(\mathbf{D}) < 1$ then*

$$C(\mathbf{D}) \leqslant \frac{2}{A}\cdot\frac{p}{\sqrt{k}}\cdot\frac{1+\overline{\delta}_k(\mathbf{D})}{\sqrt{1-\underline{\delta}_k(\mathbf{D})}}.$$

Proof. For any $\mathbf{x}$ we have $\|\mathbf{D}^\top\mathbf{x}\|_\infty^2 \geq \|\mathbf{D}^\top\mathbf{x}\|_2^2/p \geq A\|\mathbf{x}\|_2^2/p$ hence the bound on $C'(\mathbf{D})$. To bound $C(\mathbf{D})$ we define P_T the orthoprojector onto $\text{span}(\mathbf{D}_T)$ where $T\subset[\![1;p]\!]$, $\mathbf{r}_0=\mathbf{x}$ and for $i\geq 1$

$$\begin{aligned} T_i &= \arg\max_{|T|\leqslant k}\|P_T\mathbf{r}_{i-1}\|_2 \\ \mathbf{r}_i &= \mathbf{r}_{i-1}-P_{T_i}\mathbf{r}_{i-1} \\ \boldsymbol{\alpha}_i &\quad s.t.\quad P_{T_i}\mathbf{r}_{i-1}=\mathbf{D}_{T_i}\boldsymbol{\alpha}_i. \end{aligned}$$

We notice that for any $\mathbf{r}$

$$\begin{aligned}\sup_{|T|\leqslant k}\|P_T\mathbf{r}\|_2^2 &\geq \sup_{|T|\leqslant k}\frac{\|\mathbf{D}_T^\top\mathbf{r}\|_2^2}{1+\overline{\delta}_k} \geq \frac{1}{1+\overline{\delta}_k}\cdot\frac{k}{p}\|\mathbf{D}^\top\mathbf{r}\|_2^2 \\ &\geq \frac{A\ell}{(1+\overline{\delta}_k)p}\|\mathbf{r}\|_2^2 =: \gamma^2\|\mathbf{r}\|_2^2.\end{aligned}$$

As a result for any $i\geq 1$, $\|\mathbf{r}_i\|_2^2 = \|\mathbf{r}_{i-1}\|_2^2 - \|P_{T_i}\mathbf{r}_{i-1}\|_2^2 \leqslant (1-\gamma^2)\|\mathbf{r}_{i-1}\|_2^2$ hence by induction $\|\mathbf{r}_i\|_2^2 \leqslant (1-\gamma^2)^i\cdot\|\mathbf{x}\|_2^2$. This implies

$$\begin{aligned}\|\boldsymbol{\alpha}_i\|_1 \leqslant \sqrt{k}\|\boldsymbol{\alpha}_i\|_2 &\leqslant \sqrt{\tfrac{k}{1-\underline{\delta}_k}}\|\mathbf{D}_{T_i}\boldsymbol{\alpha}_i\|_2 \leqslant \sqrt{\tfrac{k}{1-\underline{\delta}_k}}\|\mathbf{r}_{i-1}\|_2 \\ &\leqslant \sqrt{\tfrac{k}{1-\underline{\delta}_k}}(\sqrt{1-\gamma^2})^{i-1}\|\mathbf{x}\|_2\end{aligned}$$

Denoting $\boldsymbol{\alpha}=\sum_{i\geq 1}\boldsymbol{\alpha}_i$ we have $\mathbf{x}=\mathbf{D}\boldsymbol{\alpha}$ and

$$\begin{aligned}\|\boldsymbol{\alpha}\|_1 \leqslant \sum_{i\geq 1}\|\boldsymbol{\alpha}_i\|_1 &\leqslant \sqrt{\tfrac{k}{1-\underline{\delta}_k}}\cdot\|\mathbf{x}\|_2\cdot\sum_{i\geq 1}(\sqrt{1-\gamma^2})^{i-1} \\ &= \sqrt{\tfrac{k}{1-\underline{\delta}_k}}\cdot\|\mathbf{x}\|_2\cdot\frac{1}{1-\sqrt{1-\gamma^2}} \\ &= \sqrt{\tfrac{k}{1-\underline{\delta}_k}}\cdot\|\mathbf{x}\|_2\cdot\frac{1+\sqrt{1-\gamma^2}}{\gamma^2} \\ &\leqslant \sqrt{\tfrac{k}{1-\underline{\delta}_k}}\cdot\|\mathbf{x}\|_2\cdot\frac{2}{\gamma^2}.\end{aligned}$$

□

We may now provide a control on both $C'(r)$ and $C(r)$.

Corollary 1. *Assume $\mathbf{D}^o\in\mathbb{R}^{m\times p}$ is a frame with lower frame bound A^o such that $A^o\|\mathbf{x}\|_2^2 \leqslant \|(\mathbf{D}^o)^\top\mathbf{x}\|_2^2$ for any signal $\mathbf{x}$, and $\max\{\underline{\delta}_k(\mathbf{D}^o),\overline{\delta}_k(\mathbf{D}^o)\} \leqslant \frac{1}{4}$. Consider $r \leqslant \min\{\sqrt{A^o}/2, \sqrt{1-\underline{\delta}_k(\mathbf{D}^o)}\}$ and let $\overline{\delta} = \overline{\delta}(r) \triangleq (\sqrt{1+\overline{\delta}_k(\mathbf{D}^o)}+r)^2-1$ and $\underline{\delta}=\underline{\delta}(r) \triangleq 1-(\sqrt{1-\underline{\delta}_k(\mathbf{D}^o)}-r)^2$. Then, for any $\mathbf{D}$ such that $\|\mathbf{D}-\mathbf{D}^o\|_F \leqslant r$, $C'(\mathbf{D}) \leqslant \frac{8}{A^o}\cdot\frac{p}{\sqrt{k}}\cdot\frac{1+\overline{\delta}}{\sqrt{1-\underline{\delta}}}$ and $C(\mathbf{D}) \leqslant \sqrt{4p/A^o}$.*

Proof. From the proof of Lemma 4, for any $\mathbf{D}$ such that $\|\mathbf{D}-\mathbf{D}^o\|_F \leqslant r$, we have $\underline{\delta}_k(\mathbf{D}) \leqslant \underline{\delta}$. Using a similar reasoning, we get: $\overline{\delta}_k(\mathbf{D}) \leqslant \overline{\delta}$. Moreover, using the triangular inequality, we have

$$\|\mathbf{D}^\top\mathbf{x}\|_2 \geq \|[\mathbf{D}^o]^\top\mathbf{x}\|_2 - \|(\mathbf{D}^o-\mathbf{D})^\top\mathbf{x}\|_2 \geq \sqrt{A^o}\|\mathbf{x}\|_2 - r\|\mathbf{x}\|_2,$$

and thus with $A=A^o/4$ and $r\leqslant\sqrt{A^o}/2$, $\mathbf{D}^o$ is a frame with lower frame bound A. We may thus apply the lemma above, to obtain the desired results. □

c) Summary.: With the assumptions of Theorem 2, we thus obtain from (58) and Corollary 1 the following bound:

$$n_{\text{out}}|\Delta F_{\mathbf{X}_{\text{out}}}(r)| \leqslant \|\mathbf{X}_{\text{out}}\|_{1,2}\cdot\mathbb{E}|\alpha|\cdot\frac{8}{A^o}\cdot\frac{p}{\sqrt{k}}\cdot\frac{1+\overline{\delta}}{\sqrt{1-\underline{\delta}}}\cdot\sqrt{\frac{4p}{A^o}}\cdot r\bar{\lambda}, \quad (59)$$

where $\|\mathbf{X}_{\text{out}}\|_{1,2} \triangleq \sum_{i\in\text{out}}\|\mathbf{x}^i\|_2$. Assumption (17) implies $\delta_k(\mathbf{D}^o) \leqslant 1/4$, and the other assumptions of Theorem 2 imply $r<0.15$. It follows that

$$8\frac{1+\overline{\delta}}{\sqrt{1-\underline{\delta}}} \leqslant 8\frac{(\sqrt{1+1/4}+0.15)^2}{\sqrt{1-1/4}-0.15} \leqslant 18.$$

With this refined bound, we obtain the "resolution independent" robustness result (30) in Theorem 2 using the same reasoning sketched in Section IV-B with the naive bound (36).

V. Conclusion and discussion

We conducted an asymptotic as well as precise finite-sample analysis of the local minima of sparse coding in the presence of noise, thus extending prior work which focused on noiseless settings [19, 17]. Given a probabilistic model of sparse signals that only combines assumptions on certain first and second order moments, and almost sure boundedness, we have shown that a local minimum exists with high probability around the reference dictionary, under cumulative-coherence assumptions on the ground truth dictionary. We have shown the robustness

of the approach to the presence of outliers, provided a certain "outlier to inlier energy ratio" remains small enough. In contrast to related prior work, the sample complexity estimates we obtained are independent of the precision of the predicted recovery. Similarly, the admissible level of outliers under some additional completeness assumption has been shown to be harmless to the targeted resolution.

Our study could be further developed in multiple ways. First, we may target more realistic of widely accepted generative models for $\boldsymbol{\alpha}^o$ such as the spike and slab models of Ishwaran and Rao [23], or signals with compressible priors [20]. Second, one may want to deal with other constraint sets $\mathcal{D}$ on the dictionary to deal with related problems such as structured dictionary learning [21] or blind calibration. This may yield improved sample complexity estimates where, e.g., a factor mp could be replaced with the upper box-counting dimension of $\mathcal{D}$. Moreover, more refined estimates in the spirit of [31] could possibly provide sample complexity estimates that no longer depend on the signal dimension m, or fast rates $\eta_n = O(1/n)$, rather than $\eta_n = O(1/\sqrt{n})$ which would both translate into better sample complexity estimates (e.g., mp^2 rather than mp^3 with fast rates). Note here that the lower-bound recently proved by Jung et al. [26] leads to a sample complexity of at least p^2, which still leaves room for improvement (either for the lower or upper bounds).

Third, the analysis could potentially be extended to other penalties than ℓ^1, e.g., with mixed norms promoting group sparsity. A related problem is that of considering complex-valued rather than only real-valued dictionary learning problems. The recent results of Vaiter et al. [42] establishing the stable recovery of a generalized notion of "support" through a generalized irrepresentability condition might be instrumental with this respect.

a) Beyond exact recovery, and beyond coherence ?: The spirit of our analysis, as described in Section IV, is that one can approximate the empirical cost function $\mathbf{D} \mapsto \Delta F_{\mathbf{X}}(\mathbf{D}; \mathbf{D}^o)$ by the expectation of the idealized cost function $\mathbf{D} \mapsto \mathbb{E}_{\mathbf{x}} \Delta \phi_{\mathbf{x}}(\mathbf{D}; \mathbf{D}^o | \mathbf{s}^o)$ (Proposition 1). A simple restricted isometry property is enough to show the existence of a local minimum of the latter which is both close to $\mathbf{D}^o$ (Proposition 2) and global on a large ball around $\mathbf{D}^o$. However, we use more heavy artillery to control how closely $\Delta F_{\mathbf{X}}$ is approximated by $\mathbb{E}_{\mathbf{x}} \Delta \phi_{\mathbf{x}}$: a cumulative coherence assumption coupled with the assumption that nonzero coefficients are bounded from below. Using exact recovery arguments (Proposition 3), this implies that in a neighborhood of $\mathbf{D}^o$ of controlled (but small) size, we have almost surely equality between $\phi_{\mathbf{x}}(\mathbf{D})$ and $f_{\mathbf{x}}(\mathbf{D})$.

While this route has the merit of a relative simplicity[3], it is also introduces several limitations:

- *limited sparsity*: the cumulative coherence assumption restricts much more the admissible sparsity levels than a simple restricted isometry property assumption.
- *local vs global*: Proposition 3 controls the quality of the approximation of $\mathbb{E}\Delta F_{\mathbf{X}}$ by $\mathbb{E}_{\mathbf{x}} \Delta \phi_{\mathbf{x}}$ on a neighborhood whose size r cannot exceed $O(\lambda)$. In contrast, using only a RIP assumption, Proposition 2 provides a lower bound (46) of $\mathbf{D} \mapsto \mathbb{E}_{\mathbf{x}} \Delta \phi_{\mathbf{x}}(\mathbf{D}; \mathbf{D}^o | \mathbf{s}^o)$ which is valid on a large neighborhood of $\mathbf{D}^o$ of radius $r = O(1)$.
  Even though dictionaries in $\mathbb{R}^{m \times p}$ can be at much higher mutual Frobenius distances that $O(1)$, one cannot envision to significantly improve over the radius $r = O(1)$ for which $\mathbb{E}\Delta F_{\mathbf{X}}(r) > 0$. To see why, consider $\mathbf{D}$ a dictionary of coherence $\mu_1(\mathbf{D})$, and i, j a pair of distinct atoms such that $|[\mathbf{d}^i]^\top \mathbf{d}^j| = \mu_1(\mathbf{D})$. Consider $\mathbf{D}'$ obtained by permuting these two atoms and possibly flipping the sign of one of them: then $F_{\mathbf{X}}\mathbf{D}') = F_{\mathbf{X}}(\mathbf{D})$, and $\|\mathbf{D}' - \mathbf{D}\|_F^2 = 2\|\mathbf{d}^i \pm \mathbf{d}^j\|_2^2 = 2 - 2\mu_1(\mathbf{D})$. Hence, $\mathbf{D}'$ is within radius $r \leqslant \sqrt{2(1-\mu_1(\mathbf{D}))} = O(1)$ of $\mathbf{D}$ (in the Frobenius distance) but $\Delta F_{\mathbf{X}}(\mathbf{D}'; \mathbf{D}) = 0$.
  Of course, Proposition 3 is sufficient to prove the desired existence of a local minimum $\hat{\mathbf{D}}$ of $\mathbf{D} \mapsto \mathbb{E}\Delta F_{\mathbf{X}}(\mathbf{D})$ (Theorem 1). However, controlling the quality of the approximation of $\mathbb{E}\Delta F_{\mathbf{X}}$ by $\mathbb{E}_{\mathbf{x}} \Delta \phi_{\mathbf{x}}$ on a much larger neighborhood would seem desirable, since it would show that $\hat{\mathbf{D}}$ is not only a local minimum, but also that it is *global over a ball of large radius* $r = O(1)$ *around* $\mathbf{D}^o$. This has the potential of opening the way to algorithmic results in terms of the practical optimization of $F_{\mathbf{X}}(\mathbf{D})$ rather than just properties of this cost function, in the spirit of the recent results [2] etc. establishing the size of the basin of convergence of an alternate minimization approach based on exact ℓ^1 minimization.

To address the above limitations, one can envision an analysis that would replace the assumption on $\mu_k(\mathbf{D}^o)$ by an assumption on $\underline{\delta}_k(\mathbf{D}^o)$. This would imply, e.g., to replace Lemma 11 and Lemma 12 to obtain recovery results *with high probability* rather than *almost surely*, through an explicit expression of $\hat{\boldsymbol{\alpha}}_{\mathbf{x}}(\mathbf{D}|\mathbf{s}^o) - \boldsymbol{\alpha}^o$ and a control of its ℓ^∞ norm with high probability, in the spirit of Candès and Plan [10]. As a by-product of such improvements, one can expect to remove the unnecessarily conservative assumption (12) involving $\underline{\alpha}$, but also replacing $\underline{\alpha}$ with $\mathbb{E}\,|\alpha|$ in Theorem 1 (assumption (22)) and Theorem 2, as well as replacing $M_{\boldsymbol{\alpha}}$ and $M_{\boldsymbol{\varepsilon}}$ with expected values rather than worst case quantities. To support these improvements, a promising approach consists in exploiting convex duality to directly lower bound $F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}^o)$ without resorting to exact recovery. This also has the potential to yield guarantees where assumption (19) is relaxed, thus encompassing very overcomplete dictionaries beyond the $p \lesssim m^2$ barrier faced in this paper.

[3] From a certain point of view ...

ACKNOWLEDGEMENTS

Many thanks to Karin Schnass for suggesting to make our life much easier with a boundedness rather than sub-Gaussian assumption in the signal model, to Martin Kleinsteuber for helping to disentangle sample complexity from local stability, and to Nancy Bertin for suggesting the cinematographic reference in the title.

REFERENCES

[1] P. A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

[2] Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon. Learning Sparsely Used Overcomplete Dictionaries via Alternating Minimization. Technical Report 1310.7991v1, arXiv, October 2013.

[3] Alekh Agarwal, Animashree Anandkumar, and Praneeth Netrapalli. Exact Recovery of Sparsely Used Overcomplete Dictionaries. Technical Report 1309.1952v1, arXiv, September 2013.

[4] Michal Aharon, Michael Elad, and Alfred M Bruckstein. On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them. *Linear Algebra and its Applications*, 416(1):48–67, July 2006.

[5] Sanjeev Arora, Rong Ge, and Ankur Moitra. New Algorithms for Learning Incoherent and Overcomplete Dictionaries. Technical Report 1308.6273v1, arXiv, August 2013.

[6] F. Bach, J. Mairal, and J. Ponce. Convex sparse matrix factorizations. Technical Report 0812.1869, arXiv, 2008.

[7] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2011.

[8] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[9] D. M. Bradley and J. A. Bagnell. Convex coding. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009.

[10] E. J. Candès and Y. Plan. Near-ideal model selection by ℓ_1 minimization. *Annals of Statistics*, 37(5A):2145–2177, 2009.

[11] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[12] P. Comon and C. Jutten, editors. *Handbook of Blind Source Separation, Independent Component Analysis and Applications*. Academic Press, 2010.

[13] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.

[14] D J Field and B A Olshausen. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[15] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer, May 2012.

[16] J. J. Fuchs. Recovery of exact sparse representations in the presence of bounded noise. *IEEE Transactions on Information Theory*, 51(10):3601–3608, 2005.

[17] Q. Geng, H. Wang, and J. Wright. On the Local Correctness of L1 Minimization for Dictionary Learning. Technical Report 1101.5672, arXiv, 2011.

[18] P Georgiev, F J Theis, and Andrzej Cichocki. Sparse Component Analysis and Blind Source Separation of Underdetermined Mixtures. *Neural Networks, IEEE Transactions on*, 16(4):992–996, 2005.

[19] R. Gribonval and K. Schnass. Dictionary identification—sparse matrix-factorization via ℓ_1-minimization. *IEEE Transactions on Information Theory*, 56(7):3523–3539, 2010.

[20] Remi Gribonval, Volkan Cevher, and Michael E Davies. Compressible Distributions for High-Dimensional Statistics. *IEEE Trans. Information Theory*, 58(8):5016–5034, 2012.

[21] Remi Gribonval, R Jenatton, F Bach, M Kleinsteuber, and M Seibert. Sample Complexity of Dictionary Learning and Other Matrix Factorizations. *IEEE Trans. Inf. Theor.*, 61(6):3469–3486, 2015.

[22] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, 1990.

[23] H. Ishwaran and J. S. Rao. Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics*, 33(2):730–773, 2005.

[24] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12:2297–2334, 2011.

[25] Rodolphe Jenatton, Remi Gribonval, and Francis Bach. Local stability and robustness of sparse dictionary learning in the presence of noise. Technical Report arXiv:1202.3778, CMAP, 2012.

[26] Alexander Jung, Yonina C. Eldar, and Norbert Görtz. Performance limits of dictionary learning for sparse coding. Technical Report 1402.4078, arXiv, 2014.

[27] A. Krause and V. Cevher. Submodular dictionary selection for sparse representation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

[28] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(1):19–60, 2010.

[29] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 3rd edition, December 2008.

[30] P. Massart. *Concentration Inequalities and Model Selection: Ecole d'été de Probabilités de Saint-Flour 23*. Springer, 2003.

[31] A. Maurer and M. Pontil. k-dimensional coding schemes in hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, 2010.

[32] N. A. Mehta and A. G. Gray. On the sample complexity of predictive sparse coding. Technical report, preprint arXiv:1202.4050, 2012.

[33] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325, 1997.

[34] Ron Rubinstein, A M Bruckstein, and Michael Elad. Dictionaries for Sparse Representation Modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.

[35] Karin Schnass. On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD. *Appl. Comp. Harm. Anal.*, 37(3):464–491, November 2014.

[36] Karin Schnass. Local Identification of Overcomplete Dictionaries. Technical Report 1401.6354v1, arXiv, January 2014.

[37] Daniel A Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. *Journal of*

Machine Learning Research: Workshop and Conference Proceeedings, 23:1–18, June 2012.

[38] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, pages 267–288, 1996.

[39] I. Tosic and P. Frossard. Dictionary learning. *Signal Processing Magazine*, 28(2):27–38, 2011.

[40] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

[41] D Vainsencher, S Mannor, and A M Bruckstein. The Sample Complexity of Dictionary Learning. *Journal of Machine Learning Research*, 12:3259–3281, 2011.

[42] Samuel Vaiter, Mohammad Golbabaee, Jalal M Fadili, and Gabriel Peyré. Model Selection with Low Complexity Priors. *Information and Inference: A Journal of the IMA*, page 52 p., 2015.

[43] S. Van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[44] M. J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1- constrained quadratic programming. *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.

[45] T. Zhang. Some sharp performance bounds for least squares regression with l1 regularization. *Annals of Statistics*, 37(5A):2109–2144, 2009.

[46] P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

[47] M. Zhou, H. Chen, J. Paisley, L. Ren, G. Sapiro, and L. Carin. Non-parametric Bayesian dictionary learning for sparse image representations. In *Advances in Neural Information Processing Systems*, 2009.

[48] Michael Zibulevsky and Barak A Pearlmutter. Blind Source Separation by Sparse Decomposition in a Signal Dictionary. *Neural Computation*, 13(4):863–882, 2001.

APPENDIX

A. Expression of $\Delta\phi_{\mathbf{x}}(\mathbf{D};\mathbf{D}'|\mathbf{s}^o)$

By Definition 2 we have

$$\begin{aligned}
&\phi_{\mathbf{x}}(\mathbf{D}|\mathbf{s}^o)\\
&\quad=\frac{1}{2}\left[\|\mathbf{x}\|_2^2-(\mathbf{D}_{\mathrm{J}}^\top\mathbf{x}-\lambda\mathbf{s}_{\mathrm{J}})^\top(\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}})^{-1}(\mathbf{D}_{\mathrm{J}}^\top\mathbf{x}-\lambda\mathbf{s}_{\mathrm{J}})\right]\\
&\quad=\frac{1}{2}\|\mathbf{x}\|_2^2-\frac{1}{2}\mathbf{x}^\top\mathbf{P}_{\mathrm{J}}\mathbf{x}+\lambda\mathbf{s}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}^+\mathbf{x}-\frac{\lambda^2}{2}\mathbf{s}_{\mathrm{J}}^\top\mathbf{H}_{\mathrm{J}}\mathbf{s}_{\mathrm{J}}.
\end{aligned}$$

Since $\mathbf{x}=\mathbf{D}^o\boldsymbol{\alpha}^o+\boldsymbol{\varepsilon}=\mathbf{D}^o_{\mathrm{J}}[\boldsymbol{\alpha}^o]_{\mathrm{J}}+\boldsymbol{\varepsilon}$, it follows that for any pair $\mathbf{D},\mathbf{D}'$

$$\begin{aligned}
&\Delta\phi_{\mathbf{x}}(\mathbf{D};\mathbf{D}'|\mathbf{s}^o)\\
&\quad=\frac{1}{2}\mathbf{x}^\top[\mathbf{P}'_{\mathrm{J}}-\mathbf{P}_{\mathrm{J}}]\mathbf{x}-\lambda\mathbf{s}_{\mathrm{J}}^\top\left[[\mathbf{D}'_{\mathrm{J}}]^+-\mathbf{D}_{\mathrm{J}}^+\right]\mathbf{x}\\
&\qquad+\frac{\lambda^2}{2}\mathbf{s}_{\mathrm{J}}^\top[\mathbf{H}'_{\mathrm{J}}-\mathbf{H}_{\mathrm{J}}]\mathbf{s}_{\mathrm{J}}\\
&\quad=\Delta\phi_{\boldsymbol{\alpha},\boldsymbol{\alpha}}+\Delta\phi_{\boldsymbol{\alpha},\boldsymbol{\varepsilon}}+\Delta\phi_{\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}}+\Delta\phi_{\mathbf{s},\boldsymbol{\alpha}}+\Delta\phi_{\mathbf{s},\boldsymbol{\varepsilon}}+\Delta\phi_{\mathbf{s},\mathbf{s}}
\end{aligned}$$

with the following shorthands

$$\begin{aligned}
\Delta\phi_{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\mathbf{D};\mathbf{D}') &\triangleq \frac{1}{2}[\boldsymbol{\alpha}^o]^\top[\mathbf{D}^o]^\top(\mathbf{P}'_{\mathrm{J}}-\mathbf{P}_{\mathrm{J}})\mathbf{D}^o\boldsymbol{\alpha}^o\\
\Delta\phi_{\boldsymbol{\alpha},\boldsymbol{\varepsilon}}(\mathbf{D};\mathbf{D}') &\triangleq \boldsymbol{\varepsilon}^\top(\mathbf{P}'_{\mathrm{J}}-\mathbf{P}_{\mathrm{J}})\mathbf{D}^o\boldsymbol{\alpha}^o\\
\Delta\phi_{\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}}(\mathbf{D};\mathbf{D}') &\triangleq \frac{1}{2}\boldsymbol{\varepsilon}^\top(\mathbf{P}'_{\mathrm{J}}-\mathbf{P}_{\mathrm{J}})\boldsymbol{\varepsilon}\\
\Delta\phi_{\mathbf{s},\boldsymbol{\alpha}}(\mathbf{D};\mathbf{D}') &\triangleq -\lambda\mathbf{s}_{\mathrm{J}}^\top([\mathbf{D}'_{\mathrm{J}}]^+-\mathbf{D}_{\mathrm{J}}^+)\mathbf{D}^o\boldsymbol{\alpha}^o\\
\Delta\phi_{\mathbf{s},\boldsymbol{\varepsilon}}(\mathbf{D};\mathbf{D}') &\triangleq -\lambda\mathbf{s}_{\mathrm{J}}^\top([\mathbf{D}'_{\mathrm{J}}]^+-\mathbf{D}_{\mathrm{J}}^+)\boldsymbol{\varepsilon}\\
\Delta\phi_{\mathbf{s},\mathbf{s}}(\mathbf{D};\mathbf{D}') &\triangleq \frac{\lambda^2}{2}\mathbf{s}_{\mathrm{J}}^\top(\mathbf{H}'_{\mathrm{J}}-\mathbf{H}_{\mathrm{J}})\mathbf{s}_{\mathrm{J}}
\end{aligned}$$

B. Expectation of $\Delta\phi_{\mathbf{x}}(\mathbf{D};\mathbf{D}^o|\mathbf{s}^o)$

Specializing to $\mathbf{D}'=\mathbf{D}^o$ we have

$$\begin{aligned}
\Delta\phi_{\boldsymbol{\alpha},\boldsymbol{\alpha}} &\triangleq \frac{1}{2}[\boldsymbol{\alpha}^o]^\top[\mathbf{D}^o]^\top(\mathbf{I}-\mathbf{P}_{\mathrm{J}})\mathbf{D}^o\boldsymbol{\alpha}^o\\
\Delta\phi_{\boldsymbol{\alpha},\boldsymbol{\varepsilon}} &\triangleq \boldsymbol{\varepsilon}^\top(\mathbf{I}-\mathbf{P}_{\mathrm{J}})\mathbf{D}^o\boldsymbol{\alpha}^o\\
\Delta\phi_{\mathbf{s},\boldsymbol{\alpha}} &\triangleq -\lambda\mathbf{s}_{\mathrm{J}}^\top(\mathbf{I}-\mathbf{D}_{\mathrm{J}}^+\mathbf{D}^o)\boldsymbol{\alpha}^o
\end{aligned}\tag{60}$$

Moreover, under the basic signal model (Assumption A), by the decorrelation between $\boldsymbol{\alpha}$ and $\boldsymbol{\varepsilon}$ we have

$$\mathbb{E}\{\Delta\phi_{\boldsymbol{\alpha},\boldsymbol{\varepsilon}}\}=\mathbb{E}\{\Delta\phi_{\mathbf{s},\boldsymbol{\varepsilon}}\}=0.$$

Moreover, we can rewrite

$$\begin{aligned}
\Delta\phi_{\boldsymbol{\alpha},\boldsymbol{\alpha}} &= \frac{1}{2}\cdot\mathrm{Tr}\Big(\boldsymbol{\alpha}^o_{\mathrm{J}}[\boldsymbol{\alpha}^o]^\top_{\mathrm{J}}\cdot[\mathbf{D}^o_{\mathrm{J}}]^\top(\mathbf{I}-\mathbf{P}_{\mathrm{J}})\mathbf{D}^o_{\mathrm{J}}\Big)\\
\Delta\phi_{\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}} &= \frac{1}{2}\cdot\mathrm{Tr}\Big(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\cdot(\mathbf{P}^o_{\mathrm{J}}-\mathbf{P}_{\mathrm{J}})\Big)\\
\Delta\phi_{\mathbf{s},\boldsymbol{\alpha}} &= -\lambda\cdot\mathrm{Tr}\Big(\boldsymbol{\alpha}^o_{\mathrm{J}}\,\mathbf{s}_{\mathrm{J}}^\top\cdot(\mathbf{I}-\mathbf{D}_{\mathrm{J}}^+\mathbf{D}^o_{\mathrm{J}})\Big)\\
\Delta\phi_{\mathbf{s},\mathbf{s}} &= \frac{\lambda^2}{2}\cdot\mathrm{Tr}\left(\mathbf{H}^o_{\mathrm{J}}-\mathbf{H}_{\mathrm{J}}\right).
\end{aligned}$$

Since $\mathbf{P}_{\mathrm{J}}$ is an orthoprojector onto a subspace of dimension k, $\mathrm{Tr}(\mathbf{P}^o_{\mathrm{J}}-\mathbf{P}_{\mathrm{J}})=k-k=0$, hence

$$\begin{aligned}
\mathbb{E}\{\Delta\phi_{\boldsymbol{\alpha},\boldsymbol{\alpha}}\} &= \tfrac{\mathbb{E}\{\alpha^2\}}{2}\cdot\mathbb{E}_J\Big\{\mathrm{Tr}\big([\mathbf{D}^o_{\mathrm{J}}]^\top(\mathbf{I}-\mathbf{P}_{\mathrm{J}})\mathbf{D}^o_{\mathrm{J}}\big)\Big\}\\
\mathbb{E}\{\Delta\phi_{\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}}\} &= \tfrac{\mathbb{E}\{\varepsilon^2\}}{2}\cdot\mathbb{E}_J\Big\{\mathrm{Tr}(\mathbf{P}^o_{\mathrm{J}}-\mathbf{P}_{\mathrm{J}})\Big\}=0\\
\mathbb{E}\{\Delta\phi_{\mathbf{s},\boldsymbol{\alpha}}\} &= -\lambda\cdot\mathbb{E}\{|\alpha|\}\cdot\mathbb{E}_J\Big\{\mathrm{Tr}\left(\mathbf{I}-\mathbf{D}_{\mathrm{J}}^+\mathbf{D}^o_{\mathrm{J}}\right)\Big\}\\
\mathbb{E}\{\Delta\phi_{\mathbf{s},\mathbf{s}}\} &= \tfrac{\lambda^2}{2}\cdot\mathbb{E}_J\left\{\mathrm{Tr}\left(\mathbf{H}^o_{\mathrm{J}}-\mathbf{H}_{\mathrm{J}}\right)\right\}.
\end{aligned}$$

C. Proof of Proposition 2

The lower bound for $\Delta\phi_{\mathbb{P}}(\mathbf{D};\mathbf{D}^o|\mathbf{s}^o)$ relies on a series of lemmatas whose proof is postponed to Appendix D.

Lemma 4. *Let $\mathbf{D}\in\mathbb{R}^{m\times p}$ be a dictionary such that $\underline{\delta}_k(\mathbf{D})<1$. Then, for any J of size k, the $\mathrm{J}\times\mathrm{J}$ matrix $\mathbf{H}_{\mathrm{J}}$ is well defined and we have*

$$\|\mathbf{H}_{\mathrm{J}}\|_2 \leqslant \frac{1}{1-\underline{\delta}_k(\mathbf{D})}.\tag{61}$$

$$\|\mathbf{D}_{\mathrm{J}}^+\|_2 \leqslant \frac{1}{\sqrt{1-\underline{\delta}_k(\mathbf{D})}}.\tag{62}$$

Moreover, for any $\mathbf{D}'$ such that $\|\mathbf{D}'-\mathbf{D}\|_F\leqslant r<\sqrt{1-\underline{\delta}_k(\mathbf{D})}$ we have

$$1-\underline{\delta}_k(\mathbf{D}')\geq(\sqrt{1-\underline{\delta}_k(\mathbf{D})}-r)^2\triangleq 1-\underline{\delta}\tag{63}$$

Lemma 5. *For any $\underline{\delta} < 1$, $\mathbf{D}, \mathbf{D}'$ such that $\max(\underline{\delta}_k(\mathbf{D}), \underline{\delta}_k(\mathbf{D}')) \leqslant \underline{\delta}$, and J of size k, we have*

$$\begin{aligned}
|||\mathbf{I} - \mathbf{D}_{\mathrm{J}}^{+}\mathbf{D}'_{\mathrm{J}}|||_2 &\leqslant (1-\underline{\delta})^{-1/2} \|\mathbf{D}-\mathbf{D}'\|_F \\
|||\mathbf{H}'_{\mathrm{J}} - \mathbf{H}_{\mathrm{J}}|||_2 &\leqslant 2(1-\underline{\delta})^{-3/2} \|\mathbf{D}-\mathbf{D}'\|_F \\
|||[\mathbf{D}'_{\mathrm{J}}]^{+} - \mathbf{D}_{\mathrm{J}}^{+}||| &\leqslant 2(1-\underline{\delta})^{-1} \|\mathbf{D}-\mathbf{D}'\|_F \\
|||\mathbf{P}'_{\mathrm{J}} - \mathbf{P}_{\mathrm{J}}|||_2 &\leqslant 2(1-\underline{\delta})^{-1/2} \|\mathbf{D}-\mathbf{D}'\|_F.
\end{aligned}$$

Lemma 6. *Denote $\mathcal{D}$ the oblique manifold. Given any $\mathbf{D}_1, \mathbf{D}_2 \in \mathcal{D}$, there exists a matrix $\mathbf{W} \in \mathbb{R}^{m\times p}$ with $\mathrm{diag}(\mathbf{D}^\top\mathbf{W}) = 0$ and $\mathrm{diag}(\mathbf{W}^\top\mathbf{W}) = \mathbf{I}$, i.e., $\mathbf{W} = [\mathbf{w}^1, \ldots, \mathbf{w}^p]$, $\mathbf{w}^j \perp \mathbf{d}_1^j$, $\|\mathbf{w}^j\|_2 = 1$, $j \in [\![1; p]\!]$, and a vector $\boldsymbol{\theta} \triangleq \boldsymbol{\theta}(\mathbf{D}_1, \mathbf{D}_2) \in [0,\ \pi]^p$ such that*

$$\begin{aligned}
\mathbf{D}_2 &= \mathbf{D}_1\mathbf{C}(\boldsymbol{\theta}) + \mathbf{W}\mathbf{S}(\boldsymbol{\theta}) && (64)\\
\mathbf{C}(\boldsymbol{\theta}) &\triangleq \mathrm{Diag}(\cos\boldsymbol{\theta}) && (65)\\
\mathbf{S}(\boldsymbol{\theta}) &\triangleq \mathrm{Diag}(\sin\boldsymbol{\theta}) && (66)
\end{aligned}$$

where $\cos\boldsymbol{\theta}$ (resp. $\sin\boldsymbol{\theta}$) is the vector with entries $\cos\boldsymbol{\theta}_j$ (resp. $\sin\boldsymbol{\theta}_j$). Moreover, we have

$$\begin{aligned}
\frac{2}{\pi}\boldsymbol{\theta}_j &\leqslant \|\mathbf{d}_2^j - \mathbf{d}_1^j\|_2 = 2\sin\left(\frac{\boldsymbol{\theta}_j}{2}\right) \leqslant \boldsymbol{\theta}_j, \quad \forall j, && (67)\\
\frac{2}{\pi}\|\boldsymbol{\theta}\|_2 &\leqslant \|\mathbf{D}_2 - \mathbf{D}_1\|_F \leqslant \|\boldsymbol{\theta}\|_2 && (68)
\end{aligned}$$

Vice-versa, $\mathbf{D}_1 = \mathbf{D}_2\mathbf{C}(\boldsymbol{\theta}) + \mathbf{W}'\mathbf{S}(\boldsymbol{\theta})$ where $\mathbf{W}'$ has its unit columns orthogonal to those of $\mathbf{D}_2$.

The above lemma involves $\boldsymbol{\theta}(\cdot,\cdot)$, with is related to the geodesic distance $\|\boldsymbol{\theta}\|_2$ on the oblique manifold $\mathcal{D}$ [1]. Our main technical bounds exploit this distance.

Lemma 7. *Consider two dictionaries $\mathbf{D}, \mathbf{D}^o \in \mathbb{R}^{m\times p}$ and scalars $\underline{\delta}, A, B$ such that*

$$\begin{aligned}
A &\geq \max\left\{\|\mathbf{D}^\top\mathbf{D} - \mathbf{I}\|_F, \|[\mathbf{D}^o]^\top\mathbf{D}^o - \mathbf{I}\|_F)\right\} && (69)\\
B &\geq \max\left\{|||\mathbf{D}|||_2, |||\mathbf{D}^o|||_2\right\} && (70)\\
\underline{\delta} &\geq \max\{\underline{\delta}_k(\mathbf{D}), \underline{\delta}_k(\mathbf{D}^o)\}. && (71)
\end{aligned}$$

Then, with $\boldsymbol{\theta} = \boldsymbol{\theta}(\mathbf{D}^o, \mathbf{D})$:

$$\begin{aligned}
\mathbb{E}_{\mathrm{J}}\mathrm{Tr}[\mathbf{D}_{\mathrm{J}}^o]^\top(\mathbf{I} - \mathbf{P}_{\mathrm{J}})\mathbf{D}_{\mathrm{J}}^o &\geq \tfrac{k}{p}\|\boldsymbol{\theta}\|_2^2(1 - \tfrac{k}{p}\tfrac{B^2}{1-\underline{\delta}}) && (72)\\
\left|\mathbb{E}_{\mathrm{J}}\mathrm{Tr}\left(\mathbf{I} - \mathbf{D}_{\mathrm{J}}^{+}\mathbf{D}_{\mathrm{J}}^o\right)\right| &\leqslant \tfrac{k}{p}\tfrac{\|\boldsymbol{\theta}\|_2^2}{2} + \tfrac{k^2}{p^2}\tfrac{AB}{1-\underline{\delta}}\|\boldsymbol{\theta}\|_2 && (73)\\
\left|\mathbb{E}_{\mathrm{J}}\mathrm{Tr}\left(\mathbf{H}_{\mathrm{J}}^o - \mathbf{H}_{\mathrm{J}}\right)\right| &\leqslant \tfrac{k^2}{p^2}\tfrac{4AB}{(1-\underline{\delta})^2}\|\boldsymbol{\theta}\|_2. && (74)
\end{aligned}$$

Equipped with these lemmatas we first establish a lower bound on $\Delta\phi_{\mathbb{P}}(\mathbf{D}; \mathbf{D}^o)$ for a fixed pair $\mathbf{D}, \mathbf{D}^o$.

Lemma 8. *Consider two dictionaries $\mathbf{D}, \mathbf{D}^o \in \mathbb{R}^{m\times p}$ and scalars $\underline{\delta}, A, B$ such that (69)-(70)-(71) hold. Consider the basic signal model (Assumption A) and assume that the reduced regularization parameter satisfies*

$$\frac{k}{p}\frac{B^2}{1-\underline{\delta}} + \bar{\lambda}\kappa_\alpha^2 \leqslant \frac{1}{2}. \qquad (75)$$

Then, we have the lower bound

$$\Delta\phi_{\mathbb{P}}(\mathbf{D}; \mathbf{D}^o) \geq \frac{\mathbb{E}\ \alpha^2}{4}\cdot\frac{k}{p}\cdot\|\mathbf{D}-\mathbf{D}^o\|_F\cdot\left[\|\mathbf{D}-\mathbf{D}^o\|_F - r_0\right]. \qquad (76)$$

where

$$r_0 \triangleq (1+2\bar{\lambda})\cdot\bar{\lambda}\kappa_\alpha^2\cdot\frac{k}{p}\cdot\frac{2AB}{(1-\underline{\delta})^2}. \qquad (77)$$

Proof. Under the basic signal model, applying Proposition 1 and Lemma 7, yields the bound

$$\begin{aligned}
\Delta\phi_{\mathbb{P}}(\mathbf{D}; \mathbf{D}^o) \geq \tfrac{\mathbb{E}\ \alpha^2}{2}\tfrac{k}{p}\|\boldsymbol{\theta}\|_2\Big[&\|\boldsymbol{\theta}\|_2\left(1 - \tfrac{k}{p}\tfrac{B^2}{1-\underline{\delta}} - \bar{\lambda}\kappa_\alpha^2\right)\\
&- (1+2\bar{\lambda})\cdot\bar{\lambda}\kappa_\alpha^2\cdot\tfrac{k}{p}\cdot\frac{2AB}{(1-\underline{\delta})^2}\Big].
\end{aligned}$$

By assumption (75) it follows that

$$\begin{aligned}
\Delta\phi_{\mathbb{P}}(\mathbf{D}; \mathbf{D}^o) \geq \tfrac{\mathbb{E}\ \alpha^2}{4}\tfrac{k}{p}\|\boldsymbol{\theta}\|_2&\\
\cdot\Big[\|\boldsymbol{\theta}\|_2 - (1+2\bar{\lambda})\cdot\bar{\lambda}\kappa_\alpha^2\cdot\tfrac{k}{p}\cdot\tfrac{2AB}{(1-\underline{\delta})^2}\Big]&.
\end{aligned}$$

We conclude using the fact that $\|\boldsymbol{\theta}\|_2 \geq \|\mathbf{D}-\mathbf{D}^o\|_F$. □

We now show that the lower bound does not only hold for a given pair $\mathbf{D}, \mathbf{D}^o$: given $\mathbf{D}^o$, we identify a radius r such that $\Delta\phi_{\mathbb{P}}(\mathbf{D}; \mathbf{D}^o) > 0$ for any $\mathbf{D} \in \mathcal{S}(r)$. This establishes Proposition 2.

Proof of Proposition 2. Define the shorthands $A^o \triangleq \|[\mathbf{D}^o]^\top\mathbf{D}^o - \mathbf{I}\|_F$ and $B^o \triangleq |||\mathbf{D}^o|||_2$. Consider $r \leqslant 1$ and $\mathbf{D} \in \mathcal{S}(r)$, we have

$$\begin{aligned}
|||\mathbf{D}|||_2 &\leqslant |||\mathbf{D}^o|||_2 + |||\mathbf{D}-\mathbf{D}^o|||_2\\
&\leqslant B^o + \|\mathbf{D}-\mathbf{D}^o\|_F = B^o + r \leqslant B^o + 1\\
\|\mathbf{D}^\top\mathbf{D} - \mathbf{I}\|_F &\leqslant \|\mathbf{D}^\top\mathbf{D} - \mathbf{D}^\top\mathbf{D}^o\|_F\\
&\quad + \|\mathbf{D}^\top - [\mathbf{D}^o]^\top\|_F\cdot|||\mathbf{D}^o|||_2\\
&\quad + \|[\mathbf{D}^o]^\top\mathbf{D}^o - \mathbf{I}\|_F\\
&\leqslant |||\mathbf{D}^\top|||_2\cdot\|\mathbf{D}-\mathbf{D}^o\|_F\\
&\quad + \|\mathbf{D}^\top\mathbf{D}^o - [\mathbf{D}^o]^\top\mathbf{D}^o\|_F + A^o\\
&\leqslant A^o + 2Br = A^o + 2(B^o+1)r
\end{aligned}$$

By construction, assumptions (69)-(70) of Lemma 9 therefore hold with $B \triangleq B^o + 1$ and $A \triangleq A^o + 2Br$. Denoting $\underline{\delta} \triangleq 1/2$ and $\underline{\delta}^o \triangleq \underline{\delta}_k(\mathbf{D}^o)$ we have $\underline{\delta}^o \leqslant \underline{\delta} < 1$. Since $\bar{\lambda} \leqslant 1/4$ and $\kappa_\alpha \leqslant 1$, assumptions (44) and (45) imply that

$$\frac{k}{p}\frac{B^2}{1-\underline{\delta}} + \bar{\lambda}\kappa_\alpha^2 \leqslant \frac{k}{p}2B^2 + \bar{\lambda} \leqslant \frac{1}{2}.$$

showing that assumption (75) of Lemma 9 also holds. Now we observe that

$$0.15 \leqslant \sqrt{3/4} - \sqrt{1/2} \leqslant \sqrt{1-\underline{\delta}^o} - \sqrt{1-\underline{\delta}}.$$

Hence, by Lemma 4, when $r \leqslant 0.15$ we have $\sqrt{1-\underline{\delta}_k(\mathbf{D})} \geq \sqrt{1-\underline{\delta}^o} - r \geq \sqrt{1-\underline{\delta}}$, that is to say the remaining assumption (71) of Lemma 9 holds with $\underline{\delta}$, and we can leverage Lemma 9. This yields

$$\Delta\phi_{\mathbb{P}}(\mathbf{D}; \mathbf{D}^o) \geq \tfrac{\mathbb{E}\ \alpha^2}{4}\cdot\tfrac{k}{p}\cdot\|\mathbf{D}-\mathbf{D}^o\|_F\cdot\left[\|\mathbf{D}-\mathbf{D}^o\|_F - \tfrac{\gamma}{2}\tfrac{A}{B}\right]$$

with

$$\begin{aligned}
\gamma &\triangleq (1+2\bar{\lambda})\cdot\bar{\lambda}\kappa_\alpha^2\cdot\tfrac{k}{p}\tfrac{4B^2}{(1-\underline{\delta})^2}\\
&= (1+2\bar{\lambda})\cdot\bar{\lambda}\kappa_\alpha^2\cdot\tfrac{k}{p}\cdot 16B^2 \leqslant \left(1+2\bar{\lambda}\right)\cdot\bar{\lambda}\kappa_\alpha^2 \leqslant \tfrac{3}{8}
\end{aligned}$$

where we used (45) once more. Since $\gamma \leqslant 3/8 \leqslant 1/2$ and $\|\mathbf{D}-\mathbf{D}^o\|_F = r$ we have

$$\begin{aligned}\|\mathbf{D}-\mathbf{D}^o\|_F - \tfrac{\gamma}{2}\tfrac{A}{B} &= r - \gamma\tfrac{(A^o+2Br)}{2B} = r(1-\gamma) - \gamma\tfrac{A^o}{2B} \\ &\geq \tfrac{1}{2}\left(r - \gamma\tfrac{A^o}{B}\right).\end{aligned}$$

To summarize, noticing that

$$\gamma\frac{A^o}{B} = (1+2\bar{\lambda})\bar{\lambda}\cdot 16\kappa_\alpha^2\frac{k}{p}BA^o = \tfrac{2}{3}C_{\min}(1+2\bar{\lambda})\bar{\lambda} = r_{\min}(\bar{\lambda})$$

we have shown that holds (46) for any $r \leqslant 0.15$ and $\mathbf{D} \in \mathcal{S}(r;\mathbf{D}^o)$. Finally, since $1+2\bar{\lambda} \leqslant 3/2$, the assumption that $\bar{\lambda} < \frac{3}{20C_{\min}}$ implies that $r_{\min}(\bar{\lambda}) = \left(1+2\bar{\lambda}\right)\cdot\bar{\lambda}\cdot\frac{2}{3}C_{\min} < 0.15$. □

D. Control of $h(\mathbf{D})$

To obtain finite sample results in Section IV-G, we need to control $h(\mathbf{D}) = \Delta\phi_{\mathbf{x}}(\mathbf{D};\mathbf{D}^o|\mathbf{s}^o) - \Delta\phi_{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\mathbf{D};\mathbf{D}^o)$.

Lemma 9. *Consider a dictionary $\mathbf{D}^o \in \mathbb{R}^{m\times p}$, k and $r > 0$ such that $r < \sqrt{1-\underline{\delta}_k(\mathbf{D}^o)}$, and define*

$$\sqrt{1-\underline{\delta}} \triangleq \sqrt{1-\underline{\delta}_k(\mathbf{D}^o)} - r > 0.$$

Under the Bounded signal model (Assumption B), the function $h(\mathbf{D})$ is almost surely Lipschitz on $\mathcal{B}(\mathbf{D}^o;r)$ with respect to the Frobenius metric $\rho(\mathbf{D}',\mathbf{D}) \triangleq \|\mathbf{D}'-\mathbf{D}\|_F$, with Lipschitz constant upper bounded by

$$\begin{aligned}L \triangleq \tfrac{1}{\sqrt{1-\underline{\delta}}}&\left(M_{\boldsymbol{\varepsilon}} + \tfrac{\lambda\sqrt{k}}{\sqrt{1-\underline{\delta}}}\right) \\ &\cdot\left(2\sqrt{1+\overline{\delta}_k(\mathbf{D}^o)}M_{\boldsymbol{\alpha}} + M_{\boldsymbol{\varepsilon}} + \tfrac{\lambda\sqrt{k}}{\sqrt{1-\underline{\delta}}}\right)\end{aligned} \tag{78}$$

As a consequence, $|h(\mathbf{D})| = |h(\mathbf{D}) - h(\mathbf{D}^o)|$ is almost surely bounded on $\mathcal{B}(\mathbf{D}^o;r)$ by $c \triangleq Lr$.

Proof. Denote $\rho = \|\mathbf{D}'-\mathbf{D}\|_F$. Combining Lemma 4 and Lemma 5 we bound the operator norms of the matrix differences appearing in the terms $\Delta\phi_{\boldsymbol{\alpha},\boldsymbol{\varepsilon}}(\mathbf{D};\mathbf{D}')$ to $\Delta\phi_{\mathbf{s},\mathbf{s}}(\mathbf{D};\mathbf{D}')$ in Equation (60):

$$\begin{aligned}
|\Delta\phi_{\boldsymbol{\alpha},\boldsymbol{\varepsilon}}(\mathbf{D};\mathbf{D}')| &\leqslant M_{\boldsymbol{\varepsilon}}\cdot\|\mathbf{P}'_{\mathrm{J}}-\mathbf{P}_{\mathrm{J}}\|_2\cdot\sqrt{1+\overline{\delta}_k(\mathbf{D}^o)}M_{\boldsymbol{\alpha}} \\
&\leqslant \rho\cdot 2M_{\boldsymbol{\varepsilon}}\sqrt{1+\overline{\delta}_k(\mathbf{D}^o)}M_{\boldsymbol{\alpha}}\cdot(1-\underline{\delta})^{-1/2} \\
|\Delta\phi_{\mathbf{s},\boldsymbol{\alpha}}(\mathbf{D};\mathbf{D}')| &\leqslant \lambda\sqrt{k}\|[\mathbf{D}'_{\mathrm{J}}]^+ - \mathbf{D}_{\mathrm{J}}^+\|_2\sqrt{1+\overline{\delta}_k(\mathbf{D}^o)}M_{\boldsymbol{\alpha}} \\
&\leqslant \rho\cdot 2\lambda\sqrt{k}\sqrt{1+\overline{\delta}_k(\mathbf{D}^o)}M_{\boldsymbol{\alpha}}\cdot(1-\underline{\delta})^{-1} \\
|\Delta\phi_{\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}}(\mathbf{D};\mathbf{D}')| &\leqslant \tfrac{1}{2}\|\mathbf{P}'_{\mathrm{J}}-\mathbf{P}_{\mathrm{J}}\|_2\cdot M_{\boldsymbol{\varepsilon}}^2 \\
&\leqslant \rho\cdot M_{\boldsymbol{\varepsilon}}^2(1-\underline{\delta})^{-1/2} \\
|\Delta\phi_{\mathbf{s},\boldsymbol{\varepsilon}}(\mathbf{D};\mathbf{D}')| &\leqslant \lambda\sqrt{k}\cdot\|[\mathbf{D}'_{\mathrm{J}}]^+ - \mathbf{D}_{\mathrm{J}}^+\|_2\cdot M_{\boldsymbol{\varepsilon}} \\
&\leqslant \rho\cdot 2\lambda\sqrt{k}M_{\boldsymbol{\varepsilon}}\cdot(1-\underline{\delta})^{-1} \\
|\Delta\phi_{\mathbf{s},\mathbf{s}}(\mathbf{D};\mathbf{D}')| &\leqslant \tfrac{\lambda^2}{2}\cdot\|\mathbf{H}^o_{\mathrm{J}} - \mathbf{H}_{\mathrm{J}}\|_2\cdot k \\
&\leqslant \rho\cdot\lambda^2 k\cdot(1-\underline{\delta})^{-3/2}
\end{aligned}$$

Since $h(\mathbf{D}) - h(\mathbf{D}') = \Delta\phi_{\mathbf{x}}(\mathbf{D};\mathbf{D}'|\mathbf{s}^o) - \Delta\phi_{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\mathbf{D};\mathbf{D}')$, we obtain the desired bound on the Lipschitz constant by summing the right hand side of the above inequalities. To conclude, observe that $h(\mathbf{D}^o) = 0$. □

Lemma 10. *Consider $\mathbf{D} \in \mathbb{R}^{m\times p}$ with normalized columns and $\mathrm{J} \subseteq [\![1;p]\!]$ with $|\mathrm{J}| \leqslant k$. We have $\delta_k(\mathbf{D}) \leqslant \mu_{k-1}(\mathbf{D})$ hence $\|\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}} - \mathbf{I}\|_2 \leqslant \mu_{k-1}(\mathbf{D})$, and $\|\mathbf{D}_{\mathrm{J}}\mathbf{D}_{\mathrm{J}}^\top\|_2 = \|\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}\|_2 \leqslant 1+\mu_{k-1}(\mathbf{D})$. Similarly, it holds $\|\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}} - \mathbf{I}\|_\infty \leqslant \mu_{k-1}(\mathbf{D})$, $\|\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}\|_\infty \leqslant 1+\mu_{k-1}(\mathbf{D})$ and $\|\mathbf{D}_{\mathrm{J}^c}^\top\mathbf{D}_{\mathrm{J}}\|_\infty \leqslant \mu_k(\mathbf{D})$. If we further assume $\mu_{k-1}(\mathbf{D}) < 1$, then $\mathbf{H}_{\mathrm{J}} = (\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}})^{-1}$ is well-defined and*

$$\begin{aligned}\max\Big\{&\|\mathbf{H}_{\mathrm{J}} - \mathbf{I}\|_\infty, \|\mathbf{H}_{\mathrm{J}} - \mathbf{I}\|_2, \|\mathbf{D}_{\mathrm{J}^c}^\top\mathbf{D}_{\mathrm{J}}\left(\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}\right)^{-1}\|_\infty\Big\} \\ &\leqslant \frac{\mu_k}{1-\mu_{k-1}},\end{aligned}$$

along with $\max\left\{\|\mathbf{H}_{\mathrm{J}}\|_\infty, \|\mathbf{H}_{\mathrm{J}}\|_2\right\} \leqslant \frac{1}{1-\mu_{k-1}}$.

Proof. These properties are already well-known [see, e.g. 40, 16]. □

Lemma 11. *Consider a dictionary $\mathbf{D} \in \mathbb{R}^{m\times p}$, a support set $\mathrm{J} \subseteq [\![1;p]\!]$ such that $\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}$ is invertible, a sign vector $\mathbf{s} \in \{-1,1\}^{\mathrm{J}}$, and $\mathbf{x} \in \mathbb{R}^m$ a signal. If the following two conditions hold*

$$\begin{cases}\operatorname{sign}\Big(\mathbf{D}_{\mathrm{J}}^+\mathbf{x} - \lambda(\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}})^{-1}\mathbf{s}\Big) = \mathbf{s}, \\ \|\mathbf{D}_{\mathrm{J}^c}^\top(\mathbf{I}-\mathbf{P}_{\mathrm{J}})\mathbf{x}\|_\infty + \lambda\|\mathbf{D}_{\mathrm{J}^c}^\top\mathbf{D}_{\mathrm{J}}(\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}})^{-1}\|_\infty < \lambda,\end{cases}$$

then $\hat{\boldsymbol{\alpha}}_{\mathbf{x}}(\mathbf{D}|\mathbf{s})$ is the unique solution of $\min_{\boldsymbol{\alpha}\in\mathbb{R}^p}[\frac{1}{2}\|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1]$ and we have $\operatorname{sign}(\hat{\boldsymbol{\alpha}}_{\mathrm{J}}) = \mathbf{s}$.

Proof. We first check that $\hat{\boldsymbol{\alpha}}$ is a solution of the Lasso program. It is well-known [e.g., see 16, 44] that this statement is equivalent to the existence of a subgradient $\mathbf{z} \in \partial\|\hat{\boldsymbol{\alpha}}\|_1$ such that $-\mathbf{D}^\top(\mathbf{x}-\mathbf{D}\hat{\boldsymbol{\alpha}}) + \lambda\mathbf{z} = 0$, where $\mathbf{z}_j = \operatorname{sign}(\hat{\boldsymbol{\alpha}}_j)$ if $\hat{\boldsymbol{\alpha}}_j \neq 0$, and $|\mathbf{z}_j| \leqslant 1$ otherwise. We now build from $\mathbf{s}$ such a subgradient. Given the definition of $\hat{\boldsymbol{\alpha}}$ and the assumption made on its sign, we can take $\mathbf{z}_{\mathrm{J}} \triangleq \mathbf{s}$. It now remains to find a subgradient on J^c that agrees with the fact that $\hat{\boldsymbol{\alpha}}_{\mathrm{J}^c} = \mathbf{0}$. More precisely, we define $\mathbf{z}_{\mathrm{J}^c}$ by

$$\lambda\mathbf{z}_{\mathrm{J}^c} \triangleq \mathbf{D}_{\mathrm{J}^c}^\top(\mathbf{x}-\mathbf{D}\hat{\boldsymbol{\alpha}}) = \mathbf{D}_{\mathrm{J}^c}^\top(\mathbf{I}-\mathbf{P}_{\mathrm{J}})\mathbf{x} + \lambda\mathbf{D}_{\mathrm{J}^c}^\top\mathbf{D}_{\mathrm{J}}(\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}})^{-1}\mathbf{s}. \tag{79}$$

Using our assumption, we have $\|\mathbf{z}_{\mathrm{J}^c}\|_\infty < 1$. We have therefore proved that $\hat{\boldsymbol{\alpha}}$ is a solution of the Lasso program. The uniqueness comes from [44, Lemma 1]. □

Lemma 12. *Consider $\mathbf{x} = \mathbf{D}_{\mathrm{J}}^o\boldsymbol{\alpha}^o{}_{\mathrm{J}} + \boldsymbol{\varepsilon}$ for some $(\mathbf{D}^o, \boldsymbol{\alpha}^o, \boldsymbol{\varepsilon}) \in \mathbb{R}^{m\times p}\times\mathbb{R}^p\times\mathbb{R}^m$, $\mathbf{s}^o$ the sign of $\boldsymbol{\alpha}^o$ and J its support. Consider a dictionary $\mathbf{D} \in \mathbb{R}^{m\times p}$ such that $\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}$ is invertible. We have*

$$\begin{aligned}\|[\hat{\boldsymbol{\alpha}}_{\mathbf{x}}(\mathbf{D}|\mathbf{s}^o) - \boldsymbol{\alpha}^o]_{\mathrm{J}}\|_\infty &\\ \leqslant \|[\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}]^{-1}\|_\infty&\Big[\lambda + \|\mathbf{D}_{\mathrm{J}}^\top\left(\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}^o\right)\|_\infty\Big].\end{aligned}$$

Proof. The proof consists of simple algebraic manipulations. We plug the expression of $\mathbf{x}$ into that of $\hat{\boldsymbol{\alpha}}$, then use the triangle inequality for $\|.\|_\infty$, along with the definition and the sub-multiplicativity of $\|.\|_\infty$. □

Lemma 13. *Assume that $\mu_k(\mathbf{D}) \leqslant \mu_k < 1/2$. If*

$$\min_{j\in\mathrm{J}}|[\boldsymbol{\alpha}^o]_j| \geq 2\lambda \tag{80}$$

$$\|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}^o\|_2 < \lambda(1-2\mu_k). \tag{81}$$

then $\hat{\boldsymbol{\alpha}}_{\mathbf{x}}(\mathbf{D}|\mathbf{s}^o)$ *is the unique solution of* $\min_{\boldsymbol{\alpha}\in\mathbb{R}^p}[\frac{1}{2}\|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}\|_2^2+\lambda\|\boldsymbol{\alpha}\|_1]$.

Proof. Since $\|\mathbf{d}^j\|_2=1$ for all j, we have by assumption (81):

$$\begin{aligned}\|\mathbf{D}_{\mathrm{J}}^\top(\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}^o)\|_\infty &\leqslant \|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}^o\|_2<\lambda(1-2\mu_k) &(82)\\ \|\mathbf{D}_{\mathrm{J}^c}^\top(\mathbf{I}-\mathbf{P}_{\mathrm{J}})\mathbf{x}\|_\infty &\leqslant \|(\mathbf{I}-\mathbf{P}_{\mathrm{J}})\mathbf{x}\|_2\\ &\leqslant \|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}^o\|_2<\lambda(1-2\mu_k) &(83)\end{aligned}$$

where we use the fact that by definition of the orthogonal projector $\mathbf{P}_{\mathrm{J}}$ on the span of $\mathbf{D}_{\mathrm{J}}$, the vector $\mathbf{P}_{\mathrm{J}}\mathbf{x}$ is a better approximation to $\mathbf{x}$ than $\mathbf{D}\boldsymbol{\alpha}^o=\mathbf{D}_{\mathrm{J}}\boldsymbol{\alpha}^o_{\mathrm{J}}$. By Lemma 10 we have

$$\begin{aligned}|||\big(\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}\big)^{-1}|||_\infty &\leqslant \frac{1}{1-\mu_{k-1}(\mathbf{D})}\leqslant\frac{1}{1-\mu_k}\\ |||\mathbf{D}_{\mathrm{J}^c}^\top\mathbf{D}_{\mathrm{J}}\big(\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}\big)^{-1}|||_\infty &\leqslant \frac{\mu_k}{1-\mu_{k-1}(\mathbf{D})}\leqslant\frac{\mu_k}{1-\mu_k}<1\end{aligned}$$

Exploiting Lemma 12 and the bounds (80) and (82) we have

$$\begin{aligned}&\|[\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^o]_{\mathrm{J}}\|_\infty\\ &\quad\leqslant |||\big(\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}\big)^{-1}|||_\infty\Big[\lambda+\|\mathbf{D}_{\mathrm{J}}^\top(\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}^o)\|_\infty\Big]\\ &\quad<\frac{1}{1-\mu_k}\cdot\lambda\cdot[1+(1-2\mu_k)]=2\lambda\leqslant\min_{j\in\mathrm{J}}\big|[\boldsymbol{\alpha}^o]_j\big|,\end{aligned}$$

We conclude that $\mathrm{sign}(\hat{\boldsymbol{\alpha}})=\mathrm{sign}(\boldsymbol{\alpha}^o)$. There remains to prove that $\hat{\boldsymbol{\alpha}}$ is the unique solution of the Lasso program, using Lemma 11. We recall the quantity which needs to be smaller than λ

$$\|\mathbf{D}_{\mathrm{J}^c}^\top(\mathbf{I}-\mathbf{P}_{\mathrm{J}})\mathbf{x}\|_\infty+\lambda|||\mathbf{D}_{\mathrm{J}^c}^\top\mathbf{D}_{\mathrm{J}}\big(\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}\big)^{-1}|||_\infty.$$

The quantity above is first upper bounded by

$$\|\mathbf{D}_{\mathrm{J}^c}^\top(\mathbf{I}-\mathbf{P}_{\mathrm{J}})\mathbf{x}\|_\infty+\lambda\mu_k/(1-\mu_k),$$

and then, exploiting the bound (83), upper bounded by $\lambda(1-2\mu_k)+\lambda\mu_k/(1-\mu_k)<\lambda$. Putting together the pieces with $\mathrm{sign}(\hat{\boldsymbol{\alpha}})=\mathrm{sign}(\boldsymbol{\alpha}^o)$, Lemma 11 leads to the desired conclusion. □

Proof of Proposition 3. First observe that almost surely

$$\begin{aligned}\|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}^o\|_2 &= \|[\mathbf{D}^o-\mathbf{D}]_{\mathrm{J}}[\boldsymbol{\alpha}^o]_{\mathrm{J}}\|_2+\|\boldsymbol{\varepsilon}\|_2\\ &\leqslant |||[\mathbf{D}-\mathbf{D}^o]_{\mathrm{J}}|||_2\cdot\|[\boldsymbol{\alpha}^o]_{\mathrm{J}}\|_2+M_{\boldsymbol{\varepsilon}}\\ &\leqslant \|\mathbf{D}-\mathbf{D}^o\|_F\cdot M_{\boldsymbol{\alpha}}+M_{\boldsymbol{\varepsilon}}.\end{aligned}$$

Now, since $\mu_{k-1}(\mathbf{D}^o)\leqslant\mu_k^o\leqslant 1/2$, using Lemma 14 below and the shorthand $r=\|\mathbf{D}-\mathbf{D}^o\|_F$, we have

$$\begin{aligned}&\lambda(1-2\mu_k(\mathbf{D}))-\|\mathbf{D}-\mathbf{D}^o\|_F M_{\boldsymbol{\alpha}}-M_{\boldsymbol{\varepsilon}}\\ &\quad\geq\lambda(1-2\mu_k^o)-M_{\boldsymbol{\alpha}}r-2\lambda\sqrt{k}(2+1/2)r-M_{\boldsymbol{\varepsilon}}\\ &\quad\geq\lambda(1-2\mu_k^o)-\Big(M_{\boldsymbol{\alpha}}+5\lambda\sqrt{k}\Big)r-M_{\boldsymbol{\varepsilon}}\\ &\quad\geq\lambda(1-2\mu_k^o)-\tfrac{7}{2}M_{\boldsymbol{\alpha}}r-M_{\boldsymbol{\varepsilon}}=\tfrac{7}{2}M_{\boldsymbol{\alpha}}\left(C_{\max}\bar{\lambda}-r\right)-M_{\boldsymbol{\varepsilon}}\end{aligned}$$

where we used $\lambda\sqrt{k}\leqslant\frac{\alpha}{2}\sqrt{k}\leqslant\frac{M_{\boldsymbol{\alpha}}}{2}$. For $r<C_{\max}\cdot\bar{\lambda}$, the assumption on the noise level implies that $\|\mathbf{D}-\mathbf{D}^o\|_F M_{\boldsymbol{\alpha}}+M_{\boldsymbol{\varepsilon}}<\lambda(1-2\mu_k(\mathbf{D}))$, hence we can apply Lemma 13. We conclude by observing that the result applies in particular to $\mathbf{D}=\mathbf{D}^o$. □

Lemma 14. *Consider* $\mathbf{D},\mathbf{D}^o\in\mathbb{R}^{m\times p}$ *with normalized columns such that* $\|\mathbf{D}-\mathbf{D}^o\|_F\leqslant r$. *For* $k\leqslant p$ *we have*

$$\mu_k(\mathbf{D})\leqslant\mu_k(\mathbf{D}^o)+\sqrt{k}\cdot r\cdot[2+\mu_{k-1}(\mathbf{D}^o)]. \tag{84}$$

Proof of Lemma 14. Consider $\mathrm{J}\subseteq[\![1;p]\!]$ with $|\mathrm{J}|\leqslant k$ and $j\notin J$. By the triangle inequality

$$\begin{aligned}&\|\mathbf{D}_{\mathrm{J}}^\top\mathbf{d}^j\|_1\\ &\quad\leqslant\|[\mathbf{D}_{\mathrm{J}}^o]^\top[\mathbf{d}_0]^j\|_1+\|[\mathbf{D}_{\mathrm{J}}^o]^\top(\mathbf{d}^j-[\mathbf{d}_0]^j)\|_1\\ &\qquad+\|(\mathbf{D}_{\mathrm{J}}-\mathbf{D}_{\mathrm{J}}^o)^\top\mathbf{d}^j\|_1\\ &\quad\leqslant\mu_k(\mathbf{D}^o)+\sqrt{k}\|[\mathbf{D}_{\mathrm{J}}^o]^\top(\mathbf{d}^j-[\mathbf{d}_0]^j)\|_2\\ &\qquad+\sqrt{k}\|(\mathbf{D}_{\mathrm{J}}-\mathbf{D}_{\mathrm{J}}^o)^\top\mathbf{d}^j\|_2\\ &\quad\leqslant\mu_k(\mathbf{D}^o)+\sqrt{k}\sqrt{1+\mu_{k-1}(\mathbf{D}^o)}\cdot\|\mathbf{d}^j-[\mathbf{d}_0]^j\|_2\\ &\qquad+\sqrt{k}|||(\mathbf{D}_{\mathrm{J}}-\mathbf{D}_{\mathrm{J}}^o)^\top|||_2\\ &\quad\leqslant\mu_k(\mathbf{D}^o)+\sqrt{k}r\left[1+\mu_{k-1}(\mathbf{D}^o)+1\right].\end{aligned}$$

□

The final section of this appendix gathers technical lemmas required by the main results of the paper.

E. Proof of Lemma 4

By definition of $\underline{\delta}_k(\mathbf{D})$ we have, in the sense of symmetric positive definite matrices: $\big(1-\underline{\delta}_k(\mathbf{D})\big)\cdot\mathbf{I}\preceq\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}$. As a result, $\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}$ is invertible so $\mathbf{H}_{\mathrm{J}}$ is indeed well defined, and $|||\mathbf{H}_{\mathrm{J}}|||_2=|||(\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}})^{-1}|||_2\leqslant 1/(1-\underline{\delta}_k(\mathbf{D}))$. Moreover $|||\mathbf{D}_{\mathrm{J}}^+|||_2=|||\mathbf{H}_{\mathrm{J}}\mathbf{D}_{\mathrm{J}}^\top|||_2=\sqrt{|||\mathbf{H}_{\mathrm{J}}\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}}\mathbf{H}_{\mathrm{J}}|||_2}=\sqrt{|||\mathbf{H}_{\mathrm{J}}|||_2}\leqslant\frac{1}{\sqrt{1-\underline{\delta}_k(\mathbf{D})}}$.

Consider now $\mathbf{D}'$. By the triangle inequality for any J of size k and $\mathbf{z}\in\mathbb{R}^{\mathrm{J}}$ we have

$$\begin{aligned}\|\mathbf{D}'_{\mathrm{J}}\mathbf{z}\|_2 &\geq \|\mathbf{D}_{\mathrm{J}}\mathbf{z}\|_2-\|[\mathbf{D}'_{\mathrm{J}}-\mathbf{D}_{\mathrm{J}}]\mathbf{z}\|_2\\ &\geq \left(\sqrt{1-\underline{\delta}_k(\mathbf{D})}-r\right)\cdot\|\mathbf{z}\|_2=\sqrt{1-\underline{\delta}}\|\mathbf{z}\|_2.\end{aligned}$$

where we used the fact that $|||\mathbf{D}'_{\mathrm{J}}-\mathbf{D}_{\mathrm{J}}|||_2\leqslant\|\mathbf{D}'_{\mathrm{J}}-\mathbf{D}_{\mathrm{J}}\|_F\leqslant\|\mathbf{D}'-\mathbf{D}\|_F$.

F. Proof of Lemma 5

The assumptions combined with Lemma 4 yield

$$\begin{aligned}\max(|||\mathbf{H}_{\mathrm{J}}|||_2,|||\mathbf{H}'_{\mathrm{J}}|||_2) &\leqslant (1-\underline{\delta})^{-1}\\ \max(|||\mathbf{D}_{\mathrm{J}}^+|||_2,|||[\mathbf{D}'_{\mathrm{J}}]^+|||_2) &\leqslant (1-\underline{\delta})^{-1/2}.\end{aligned}$$

Moreover, denoting $r = \|\mathbf{D} - \mathbf{D}'\|_F$, we have $\|\mathbf{D}_\mathrm{J} - \mathbf{D}'_\mathrm{J}\|_2 \leqslant \|\mathbf{D}_\mathrm{J} - \mathbf{D}'_\mathrm{J}\|_F \leqslant r$. It follows that

$$\begin{aligned}
\|\mathbf{I} - \mathbf{D}_\mathrm{J}^+\mathbf{D}'_\mathrm{J}\|_2 &= \|\mathbf{D}_\mathrm{J}^+(\mathbf{D}_\mathrm{J} - \mathbf{D}'_\mathrm{J})\|_2 \leqslant \|\mathbf{D}_\mathrm{J}^+\|_2 r \\
&\leqslant r(1-\underline{\delta})^{-1/2} \\
\mathbf{H}'_\mathrm{J} - \mathbf{H}_\mathrm{J} &= \mathbf{H}'_\mathrm{J}(\mathbf{D}_\mathrm{J}^\top\mathbf{D}_\mathrm{J} - [\mathbf{D}'_\mathrm{J}]^\top\mathbf{D}_\mathrm{J} \\
&\quad +[\mathbf{D}'_\mathrm{J}]^\top\mathbf{D}_\mathrm{J} - [\mathbf{D}'_\mathrm{J}]^\top\mathbf{D}'_\mathrm{J})\mathbf{H}_\mathrm{J} \\
&= \mathbf{H}'_\mathrm{J}(\mathbf{D}_\mathrm{J}^\top - [\mathbf{D}'_\mathrm{J}]^\top)\mathbf{D}_\mathrm{J}\mathbf{H}_\mathrm{J} \\
&\quad +\mathbf{H}'_\mathrm{J}[\mathbf{D}'_\mathrm{J}]^\top(\mathbf{D}_\mathrm{J} - \mathbf{D}'_\mathrm{J})\mathbf{H}_\mathrm{J} \\
\|\mathbf{H}'_\mathrm{J} - \mathbf{H}_\mathrm{J}\|_2 &\leqslant \|\mathbf{H}'_\mathrm{J}\|_2 r\|[\mathbf{D}_\mathrm{J}^\top]^+\|_2 \\
&\quad +\|[\mathbf{D}'_\mathrm{J}]^+\|_2 r\|\mathbf{H}_\mathrm{J}\|_2 \\
&\leqslant 2r(1-\underline{\delta})^{-3/2} \\
(\mathbf{H}'_\mathrm{J} - \mathbf{H}_\mathrm{J})\mathbf{D}_\mathrm{J}^\top &= \mathbf{H}'_\mathrm{J}(\mathbf{D}_\mathrm{J}^\top - [\mathbf{D}'_\mathrm{J}]^\top)\mathbf{D}_\mathrm{J}\mathbf{H}_\mathrm{J}\mathbf{D}_\mathrm{J}^\top \\
&\quad +\mathbf{H}'_\mathrm{J}[\mathbf{D}'_\mathrm{J}]^\top(\mathbf{D}_\mathrm{J} - \mathbf{D}'_\mathrm{J})\mathbf{H}_\mathrm{J}\mathbf{D}_\mathrm{J}^\top \\
&= \mathbf{H}'_\mathrm{J}(\mathbf{D}_\mathrm{J}^\top - [\mathbf{D}'_\mathrm{J}]^\top)\mathbf{P}_\mathrm{J} \\
&\quad +[\mathbf{D}'_\mathrm{J}]^+(\mathbf{D}_\mathrm{J} - \mathbf{D}'_\mathrm{J})\mathbf{D}_\mathrm{J}^+ \\
[\mathbf{D}'_\mathrm{J}]^+ - \mathbf{D}_\mathrm{J}^+ &= \mathbf{H}'_\mathrm{J}([\mathbf{D}'_\mathrm{J}]^\top - \mathbf{D}_\mathrm{J}^\top) \\
&\quad +(\mathbf{H}'_\mathrm{J} - \mathbf{H}_\mathrm{J})\mathbf{D}_\mathrm{J}^\top \\
&= \mathbf{H}'_\mathrm{J}([\mathbf{D}'_\mathrm{J}]^\top - \mathbf{D}_\mathrm{J}^\top)(\mathbf{I} - \mathbf{P}_\mathrm{J}) \\
&\quad +[\mathbf{D}'_\mathrm{J}]^+(\mathbf{D}_\mathrm{J} - \mathbf{D}'_\mathrm{J})\mathbf{D}_\mathrm{J}^+ \\
\|[\mathbf{D}'_\mathrm{J}]^+ - \mathbf{D}_\mathrm{J}^+\|_2 &\leqslant \|\mathbf{H}'_\mathrm{J}\|_2 r + \|[\mathbf{D}'_\mathrm{J}]^+\|_2 r\|\mathbf{D}_\mathrm{J}^+\|_2 \\
&\leqslant 2r(1-\underline{\delta})^{-1} \\
\mathbf{P}'_\mathrm{J} - \mathbf{P}_\mathrm{J} &= \mathbf{D}'_\mathrm{J}([\mathbf{D}'_\mathrm{J}]^+ - \mathbf{D}_\mathrm{J}^+) + (\mathbf{D}'_\mathrm{J} - \mathbf{D}_\mathrm{J})\mathbf{D}_\mathrm{J}^+ \\
&= \mathbf{D}'_\mathrm{J}\mathbf{H}'_\mathrm{J}([\mathbf{D}'_\mathrm{J}]^\top - \mathbf{D}_\mathrm{J}^\top)(\mathbf{I} - \mathbf{P}_\mathrm{J}) \\
&\quad +\mathbf{D}'_\mathrm{J}[\mathbf{D}'_\mathrm{J}]^+(\mathbf{D}_\mathrm{J} - \mathbf{D}'_\mathrm{J})\mathbf{D}_\mathrm{J}^+ \\
&\quad +(\mathbf{D}'_\mathrm{J} - \mathbf{D}_\mathrm{J})\mathbf{D}_\mathrm{J}^+ \\
&= ([\mathbf{D}'_\mathrm{J}]^+)^\top([\mathbf{D}'_\mathrm{J}]^\top - \mathbf{D}_\mathrm{J}^\top)(\mathbf{I} - \mathbf{P}_\mathrm{J}) \\
&\quad +(\mathbf{I} - \mathbf{P}'_\mathrm{J})(\mathbf{D}'_\mathrm{J} - \mathbf{D}_\mathrm{J})\mathbf{D}_\mathrm{J}^+ \\
\|\mathbf{P}'_\mathrm{J} - \mathbf{P}_\mathrm{J}\|_2 &\leqslant \|[\mathbf{D}'_\mathrm{J}]^+\|_2 r + r\|\mathbf{D}_\mathrm{J}^+\|_2 \\
&\leqslant 2r(1-\underline{\delta})^{-1/2}.
\end{aligned}$$

G. Proof of Lemma 6

Each column $\mathbf{d}_2^j$ of $\mathbf{D}_2$ can be uniquely expressed as $\mathbf{d}_2^j = \mathbf{u} + \mathbf{z}$, with $\mathbf{u} \in \mathrm{span}(\mathbf{d}_1^j)$ and $\mathbf{u}^\top\mathbf{z} = 0$. Since $\|\mathbf{d}_2^j\|_2 = 1$, the previous relation can be rewritten as $\mathbf{d}_2^j = \cos(\boldsymbol{\theta}_j)\mathbf{d}_1^j + \sin(\boldsymbol{\theta}_j)\mathbf{w}^j$, for some $\boldsymbol{\theta}_j \in [0,\pi]$ and some unit vector $\mathbf{w}^j$ orthogonal to $\mathbf{d}_1^j$ (except for the case $\boldsymbol{\theta}_j \in \{0,\pi\}$, the vector $\mathbf{w}^j$ is unique). The sign indetermination in $\mathbf{w}^j$ is handled thanks to the convention $\sin(\boldsymbol{\theta}_j) \geq 0$. We have $\|\boldsymbol{\theta}\|_\infty \leqslant \pi$ and

$$\begin{aligned}
\|\mathbf{d}_2^j - \mathbf{d}_1^j\|_2^2 &= \|(1-\cos(\boldsymbol{\theta}_j))\mathbf{d}^j - \sin(\boldsymbol{\theta}_j)\mathbf{w}^j\|_2^2 \\
&= (1-\cos(\boldsymbol{\theta}_j))^2 + \sin^2(\boldsymbol{\theta}_j) \\
&= 2(1-\cos(\boldsymbol{\theta}_j)) = 4\sin^2(\boldsymbol{\theta}_j/2).
\end{aligned}$$

We conclude using the inequalities $\frac{2}{\pi} \leqslant \frac{\sin u}{u} \leqslant 1$ for $0 \leqslant u \leqslant \pi/2$. The result when we interchange $\mathbf{D}_1$ and $\mathbf{D}_2$ is obvious, and $\boldsymbol{\theta}(\mathbf{D}_1,\mathbf{D}_2) = \boldsymbol{\theta}(\mathbf{D}_2,\mathbf{D}_1)$ since $\|\mathbf{d}_1^j - \mathbf{d}_2^j\|_2 = \|\mathbf{d}_2^j - \mathbf{d}_1^j\|_2$ for all j.

H. Proof of Lemma 7

The proof of Lemma 7 will exploit the following lemmata.

Lemma 15. *Let $\mathrm{J} \subset p$ be a random support and denote by $\delta(i) \triangleq \mathbb{1}_\mathrm{J}(i)$ the indicator function of J. Assume that for all $i \neq j \in [\![1;p]\!]$*

$$\begin{aligned}
\mathbb{E}\{\delta(i)\} &= \frac{k}{p} \\
\mathbb{E}\{\delta(i)\delta(j)\} &= \frac{k(k-1)}{p(p-1)}.
\end{aligned}$$

Then for any integer m and matrices $\mathbf{A},\mathbf{B} \in \mathbb{R}^{m\times p}$ such that $\mathrm{diag}(\mathbf{A}^\top\mathbf{B}) = 0$, we have

$$\mathbb{E}_\mathrm{J}\{\|\mathbf{A}_\mathrm{J}\|_F^2\} = \frac{k}{p}\|\mathbf{A}\|_F^2 \tag{85}$$

$$\mathbb{E}_\mathrm{J}\{\|\mathbf{A}_\mathrm{J}^\top\mathbf{B}_\mathrm{J}\|_F^2\} = \frac{k(k-1)}{p(p-1)} \cdot \|\mathbf{A}^\top\mathbf{B}\|_F^2 \tag{86}$$

Proof. We simply expand

$$\begin{aligned}
\mathbb{E}_\mathrm{J}\|\mathbf{A}_\mathrm{J}\|_F^2 &= \mathbb{E}\sum_{j\in[\![1;p]\!]} \delta(j)\cdot\|\mathbf{A}_{\{j\}}\|_2^2 \\
&= \sum_{j\in[\![1;p]\!]} \tfrac{k}{p}\|\mathbf{A}_{\{j\}}\|_2^2 \\
\mathbb{E}_\mathrm{J}\|\mathbf{A}_\mathrm{J}^\top\mathbf{B}_\mathrm{J}\|_F^2 &= \mathbb{E}\sum_{i\in[\![1;p]\!]}\sum_{j\in[\![1;p]\!],j\neq i} \delta(i)\delta(j)\cdot\mathbf{A}_{\{i\}}^\top\mathbf{B}_{\{j\}} \\
&= \sum_{i\in[\![1;p]\!]}\sum_{j\in[\![1;p]\!],j\neq i} \tfrac{k(k-1)}{p(p-1)}\cdot\mathbf{A}_{\{i\}}^\top\mathbf{B}_{\{j\}}.
\end{aligned}$$

□

Lemma 16. *Assume that*

$$\begin{aligned}
\underline{\delta} &\geq \max\{\underline{\delta}_k(\mathbf{D}), \underline{\delta}_k(\mathbf{D}^o)\} \\
A &\geq \max\left\{\|\mathbf{D}^\top\mathbf{D} - \mathbf{I}\|_F, \|[\mathbf{D}^o]^\top\mathbf{D}^o - \mathbf{I}\|_F)\right\}
\end{aligned}$$

with $\underline{\delta} < 1$, and define $\mathbf{U}_k \triangleq \mathbb{E}_\mathrm{J}\left[\mathbf{I}_\mathrm{J}\mathbf{H}_\mathrm{J}\mathbf{I}_\mathrm{J}^\top\right]$ and $\mathbf{V}_k \triangleq \mathbb{E}_\mathrm{J}\left[\mathbf{I}_\mathrm{J}\mathbf{H}_\mathrm{J}\mathbf{H}_\mathrm{J}^o\mathbf{I}_\mathrm{J}^\top\right]$ where the expectation is taken over all supports J of size k drawn uniformly at random. Then we have

$$\|\mathrm{off}(\mathbf{U}_k)\|_F \leqslant \frac{k(k-1)}{p(p-1)}\frac{A}{1-\underline{\delta}} \tag{87}$$

$$\|\mathrm{off}(\mathbf{V}_k)\|_F \leqslant \frac{k(k-1)}{p(p-1)}\frac{2A}{(1-\underline{\delta})^2}. \tag{88}$$

Proof. Since $\mathbf{H}_\mathrm{J}\mathbf{G}_\mathrm{J} = \mathbf{I}$, using the RIP assumption and Lemma 4 we obtain

$$\begin{aligned}
\|\mathrm{off}(\mathbf{H}_\mathrm{J})\|_F &\leqslant \|\mathbf{H}_\mathrm{J} - \mathbf{I}\|_F = \|\mathbf{H}_\mathrm{J}(\mathbf{I} - \mathbf{G}_\mathrm{J})\|_F \\
&\leqslant \frac{1}{1-\underline{\delta}}\|\mathbf{I} - \mathbf{G}_\mathrm{J}\|_F \\
\|\mathrm{off}(\mathbf{H}_\mathrm{J}\mathbf{H}_\mathrm{J}^o)\|_F &\leqslant \|\mathbf{H}_\mathrm{J}\mathbf{H}_\mathrm{J}^o - \mathbf{I}\|_F \\
&= \|(\mathbf{H}_\mathrm{J} - \mathbf{I})\mathbf{H}_\mathrm{J}^o + (\mathbf{H}_\mathrm{J}^o - \mathbf{I})\|_F \\
&\leqslant \frac{1}{1-\underline{\delta}}\|\mathbf{H}_\mathrm{J} - \mathbf{I}\|_F + \|\mathbf{H}_\mathrm{J}^o - \mathbf{I}\|_F \\
&\leqslant \frac{1}{(1-\underline{\delta})^2}\|\mathbf{I} - \mathbf{G}_\mathrm{J}\|_F + \frac{1}{1-\underline{\delta}}\|\mathbf{I} - \mathbf{G}_\mathrm{J}^o\|_F
\end{aligned}$$

In the following, $\mathbf{K}_\mathrm{J}$ denotes either $\mathbf{H}_\mathrm{J}$ or $\mathbf{H}_\mathrm{J}\mathbf{H}_\mathrm{J}^o$, and $\mathbf{W}_k$ either $\mathbf{U}_k$ or $\mathbf{V}_k$. For any J, J' of size k, we denote $\mathbf{K}_\mathrm{J}^{J\cap J'}$

the restriction of $\mathbf{K}_{\mathrm{J}}$ to the pairs of indices in $J \cap J'$, i.e. $\mathbf{K}_{\mathrm{J}}^{J\cap J'} = \mathbf{I}_{J\cap J'}^\top \mathbf{K}_{\mathrm{J}} \mathbf{I}_{J\cap J'}$, where we recall that $\mathbf{I}_{J\cap J'}$ is the restriction of the $p\times p$ identity matrix $\mathbf{I}$ to its columns indexed by $J \cap J'$. We obtain

$$\begin{aligned}
&\|\mathrm{off}(\mathbf{W}_k)\|_F^2 \\
&\quad = \|\mathbb{E}_J \mathrm{off}(\mathbf{K}_{\mathrm{J}})\|_F^2 = \langle \mathbb{E}_J \mathrm{off}(\mathbf{K}_{\mathrm{J}}), \mathbb{E}_{J'} \mathrm{off}(\mathbf{K}_{\mathrm{J}'})\rangle_F \\
&\quad = \mathbb{E}_{J,J'} \langle \mathrm{off}(\mathbf{K}_{\mathrm{J}}), \mathrm{off}(\mathbf{K}_{\mathrm{J}'})\rangle_F \\
&\quad = \mathbb{E}_{J,J'} \langle \mathrm{off}(\mathbf{K}_{\mathrm{J}}^{J\cap J'}), \mathrm{off}(\mathbf{K}_{\mathrm{J}'}^{J\cap J'})\rangle \\
&\quad \leqslant \mathbb{E}_{J,J'} \|\mathrm{off}(\mathbf{K}_{\mathrm{J}}^{J\cap J'})\|_F \cdot \|\mathrm{off}(\mathbf{K}_{\mathrm{J}'}^{J\cap J'})\|_F \\
&\quad \leqslant \sqrt{\mathbb{E}_{J,J'} \|\mathrm{off}(\mathbf{K}_{\mathrm{J}}^{J\cap J'})\|_F^2} \cdot \sqrt{\mathbb{E}_{J,J'} \|\mathrm{off}(\mathbf{K}_{\mathrm{J}'}^{J\cap J'})\|_F^2} \\
&\quad = \mathbb{E}_J \mathbb{E}_{J'} \|\mathrm{off}(\mathbf{K}_{\mathrm{J}}^{J\cap J'})\|_F^2
\end{aligned}$$

Using Lemma 15 we obtain

$$\|\mathrm{off}(\mathbf{W}_k)\|_F^2 \leqslant \mathbb{E}_J \tfrac{k(k-1)}{p(p-1)} \|\mathrm{off}(\mathbf{K}_{\mathrm{J}})\|_F^2$$

Specializing to $\mathbf{U}_k$ and using again Lemma 15 we obtain

$$\begin{aligned}
\|\mathrm{off}(\mathbf{U}_k)\|_F^2 &\leqslant \mathbb{E}_J \tfrac{k(k-1)}{p(p-1)} \tfrac{1}{(1-\underline{\delta})^2} \|\mathbf{I} - \mathbf{G}_{\mathrm{J}}\|_F^2 \\
&= \tfrac{k(k-1)}{p(p-1)} \tfrac{1}{(1-\underline{\delta})^2} \tfrac{k(k-1)}{p(p-1)} \|\mathbf{I} - \mathbf{D}^\top \mathbf{D}\|_F^2
\end{aligned}$$

It follows that

$$\|\mathrm{off}(\mathbf{U}_k)\|_F \leqslant \tfrac{k(k-1)}{p(p-1)} \frac{A}{1-\underline{\delta}}.$$

Specializing now to $\mathbf{V}_k$ we obtain similarly

$$\begin{aligned}
&\|\mathrm{off}(\mathbf{V}_k)\|_F^2 \\
&\leqslant \tfrac{k(k-1)}{p(p-1)} \tfrac{2}{(1-\nu)^8} \\
&\quad \cdot \mathbb{E}_J \left\{ \|\mathbf{I} - \mathbf{G}_{\mathrm{J}}\|_F^2 + \|\mathbf{I} - \mathbf{G}_{\mathrm{J}}^o\|_F^2 \right\} \\
&= \tfrac{k(k-1)}{p(p-1)} \tfrac{2}{(1-\nu)^8} \tfrac{k(k-1)}{p(p-1)} \\
&\quad \cdot \left\{ \|\mathbf{I} - \mathbf{D}^\top \mathbf{D}\|_F^2 + \|\mathbf{I} - [\mathbf{D}^o]^\top \mathbf{D}^o\|_F^2 \right\}
\end{aligned}$$

and we finally obtain

$$\|\mathrm{off}(\mathbf{V}_k)\|_F \leqslant \tfrac{k(k-1)}{p(p-1)} \frac{2A}{(1-\underline{\delta})^2}.$$

□

We can now proceed to the proof of Lemma 7.

a) Proof of Equation (72)*:* We write $\mathbf{D} = \mathbf{D}^o \mathbf{C}(\boldsymbol{\theta}) + \mathbf{WS}(\boldsymbol{\theta})$ using Lemma 6. For simplicity we first assume that $\boldsymbol{\theta}_j \neq \pi/2$, for all $j \in [\![1; p]\!]$. Hence, the matrix $\mathbf{C}(\boldsymbol{\theta})$ is invertible and $\mathbf{D}^o = \mathbf{D}\mathbf{C}^{-1} - \mathbf{WT}$ with $\mathbf{T} = \mathrm{Diag}(\tan(\boldsymbol{\theta}_j))$. The columns of $[\mathbf{D}\mathbf{C}^{-1}]_{\mathrm{J}}$ belong to the span of $\mathbf{D}_{\mathrm{J}}$ hence

$$\begin{aligned}
\mathrm{Tr}\big([\mathbf{D}_{\mathrm{J}}^o]^\top (\mathbf{I} - \mathbf{P}_{\mathrm{J}}) \mathbf{D}_{\mathrm{J}}^o\big) &= \|(\mathbf{I} - \mathbf{P}_{\mathrm{J}}) \mathbf{D}_{\mathrm{J}}^o\|_F^2 \\
&= \|(\mathbf{I} - \mathbf{P}_{\mathrm{J}}) [\mathbf{WT}]_{\mathrm{J}}\|_F^2 \\
&= \|[\mathbf{WT}]_{\mathrm{J}}\|_F^2 - \|\mathbf{P}_{\mathrm{J}} [\mathbf{WT}]_{\mathrm{J}}\|_F^2.
\end{aligned}$$

For the first term, by Lemma 15, we have

$$\mathbb{E}_{\mathrm{J}} \, \|[\mathbf{WT}]_{\mathrm{J}}\|_F^2 = \tfrac{k}{p} \|\mathbf{WT}\|_F^2$$

For the second term, since $\mathbf{P}_{\mathrm{J}} = \mathbf{D}_{\mathrm{J}} \mathbf{H}_{\mathrm{J}} \mathbf{D}_{\mathrm{J}}^\top$, using Lemma 4, we have the bound

$$\|\mathbf{P}_{\mathrm{J}} [\mathbf{WT}]_{\mathrm{J}}\|_F^2 \leqslant \|\!|\mathbf{H}_{\mathrm{J}}|\!\|_2 \|\mathbf{D}_{\mathrm{J}}^\top [\mathbf{WT}]_{\mathrm{J}}\|_F^2 \leqslant \tfrac{1}{1-\underline{\delta}} \|\mathbf{D}_{\mathrm{J}}^\top [\mathbf{WT}]_{\mathrm{J}}\|_F^2,$$

Now, by Lemma 15,

$$\mathbb{E}_{\mathrm{J}} \, \|\mathbf{D}_{\mathrm{J}}^\top [\mathbf{WT}]_{\mathrm{J}}\|_F^2 = \tfrac{k(k-1)}{p(p-1)} \|\mathbf{D}^\top \mathbf{WT}\|_F^2 \leqslant \tfrac{k^2}{p^2} B^2 \|\mathbf{WT}\|_F^2$$

Putting the pieces together, we obtain the lower bound

$$\mathbb{E}_{\mathrm{J}} \, \mathrm{Tr}\big([\mathbf{D}_{\mathrm{J}}^o]^\top (\mathbf{I} - \mathbf{P}_{\mathrm{J}}) \mathbf{D}_{\mathrm{J}}^o\big) \quad \geq \quad \tfrac{k}{p} \|\mathbf{WT}\|_F^2 \left(1 - \tfrac{k}{p} \tfrac{B^2}{1-\underline{\delta}}\right).$$

To conclude, we observe that since $\|\mathbf{w}^j\|_2 = 1$ and $\tan u \geq u$ for $0 \leqslant u \leqslant \pi/2$,

$$\|\mathbf{WT}\|_F^2 = \sum_{j=1}^p \tan^2(\boldsymbol{\theta}_j)^2 \geq \sum_{j=1}^p \boldsymbol{\theta}_j^2 = \|\boldsymbol{\theta}\|_2^2.$$

Finally, by continuity the obtained bound also holds when $\boldsymbol{\theta}_j = \pi/2$ for some j.

b) Proof of Equation (73)*:* Applying Lemma 6 , we write $\mathbf{D}^o = \mathbf{D}\mathbf{C}(\boldsymbol{\theta}) + \mathbf{WS}(\boldsymbol{\theta})$, and obtain,

$$\begin{aligned}
\mathrm{Tr}\left(\mathbf{I} - \mathbf{D}_{\mathrm{J}}^+ \mathbf{D}_{\mathrm{J}}^o\right) &= k - \sum_{j\in \mathrm{J}} \cos(\boldsymbol{\theta}_j) - \mathrm{Tr}\left(\mathbf{D}_{\mathrm{J}}^+ [\mathbf{WS}(\boldsymbol{\theta})]_{\mathrm{J}}\right), \\
&= \sum_{j\in \mathrm{J}} (1 - \cos(\boldsymbol{\theta}_j)) - \mathrm{Tr}\left(\mathbf{H}_{\mathrm{J}} \mathbf{D}_{\mathrm{J}}^\top [\mathbf{WS}(\boldsymbol{\theta})]_{\mathrm{J}}\right).
\end{aligned}$$

The first term is simple to handle since we have, by Lemma 15 and the inequality $1 - \cos(u) \leqslant u^2/2$, $u \in \mathbb{R}$,

$$\mathbb{E}_{\mathrm{J}} \sum_{j\in \mathrm{J}} (1 - \cos(\boldsymbol{\theta}_j)) \leqslant \mathbb{E}_{\mathrm{J}} \, \frac{\|\boldsymbol{\theta}_{\mathrm{J}}\|_2^2}{2} = \frac{k}{p} \cdot \frac{\|\boldsymbol{\theta}\|_2^2}{2}.$$

We now turn to the second term

$$\begin{aligned}
\mathbb{E}_{\mathrm{J}} \, \mathrm{Tr}\left[\mathbf{H}_{\mathrm{J}} \mathbf{D}_{\mathrm{J}}^\top [\mathbf{WS}]_{\mathrm{J}}\right] &= \mathbb{E}_{\mathrm{J}} \, \mathrm{Tr}\left[\mathbf{H}_{\mathrm{J}} (\mathbf{D} \, \mathbf{I}_{\mathrm{J}})^\top \mathbf{WS} \mathbf{I}_{\mathrm{J}}\right] \\
&= \mathbb{E}_{\mathrm{J}} \, \mathrm{Tr}\left[\mathbf{I}_{\mathrm{J}} \mathbf{H}_{\mathrm{J}} \mathbf{I}_{\mathrm{J}}^\top \mathbf{D}^\top \mathbf{WS}\right] \\
&= \mathrm{Tr}\left[\mathbf{U}_k \mathbf{D}^\top \mathbf{WS}\right].
\end{aligned}$$

Since $\mathrm{diag}(\mathbf{D}^\top \mathbf{WS}) = 0$, it follows

$$\begin{aligned}
\left|\mathrm{Tr}\left[\mathbf{U}_k \mathbf{D}^\top \mathbf{WS}\right]\right| &\leqslant \|\mathrm{off}(\mathbf{U}_k)\|_F \cdot \|\mathbf{D}^\top \mathbf{WS}\|_F \\
&\leqslant \|\mathrm{off}(\mathbf{U}_k)\|_F \cdot B \cdot \|\mathbf{WS}\|_F.
\end{aligned}$$

Since $\|\mathbf{w}^j\|_2 = 1$ and $\sin u \leqslant u$ for $0 \leqslant u \leqslant \pi/2$, we have $\|\mathbf{WS}\|_F \leqslant \|\boldsymbol{\theta}\|_2$, and we conclude the proof using Lemma 16 and the fact that $(k-1)/(p-1) \leqslant k/p$.

c) Proof of Equation (74)*:* Since $\mathbf{H}_{\mathrm{J}} = (\mathbf{D}_{\mathrm{J}}^\top \mathbf{D}_{\mathrm{J}})^{-1}$ and similarly for $\mathbf{H}_{\mathrm{J}}^o$ we have

$$\begin{aligned}
\mathbf{H}_{\mathrm{J}}^o - \mathbf{H}_{\mathrm{J}} &= \mathbf{H}_{\mathrm{J}}^o (\mathbf{D}_{\mathrm{J}}^\top \mathbf{D}_{\mathrm{J}} - [\mathbf{D}_{\mathrm{J}}^o]^\top \mathbf{D}_{\mathrm{J}}^o) \mathbf{H}_{\mathrm{J}} \\
&= \mathbf{H}_{\mathrm{J}}^o \mathbf{I}_{\mathrm{J}}^\top (\mathbf{D}^\top \mathbf{D} - [\mathbf{D}^o]^\top \mathbf{D}^o) \mathbf{I}_{\mathrm{J}} \mathbf{H}_{\mathrm{J}} \quad (89) \\
\mathrm{Tr}\ [\mathbf{H}_{\mathrm{J}}^o - \mathbf{H}_{\mathrm{J}}] &= \mathrm{Tr}\ \left[\mathbf{I}_{\mathrm{J}} \mathbf{H}_{\mathrm{J}} \mathbf{H}_{\mathrm{J}}^o \mathbf{I}_{\mathrm{J}}^\top (\mathbf{D}^\top \mathbf{D} - [\mathbf{D}^o]^\top \mathbf{D}^o)\right] \quad (90) \\
\mathbb{E}_{\mathrm{J}}\ \mathrm{Tr}\ [\mathbf{H}_{\mathrm{J}}^o - \mathbf{H}_{\mathrm{J}}] &= \mathrm{Tr}\ \left[\mathbf{V}(\mathbf{D}^\top \mathbf{D} - [\mathbf{D}^o]^\top \mathbf{D}^o)\right] \quad (91)
\end{aligned}$$

Since $\mathrm{diag}(\mathbf{D}^\top \mathbf{D} - [\mathbf{D}^o]^\top \mathbf{D}^o) = 0$ we further have

$$|\mathbb{E}_{\mathrm{J}}\ \mathrm{Tr}\ [\mathbf{H}_{\mathrm{J}}^o - \mathbf{H}_{\mathrm{J}}]| \leqslant \|\mathrm{off}(\mathbf{V})\|_F \cdot \|\mathbf{D}^\top \mathbf{D} - [\mathbf{D}^o]^\top \mathbf{D}^o\|_F.$$

We conclude using Lemma 16 after noticing that

$$\begin{aligned}
\|\mathbf{D}^\top \mathbf{D} - [\mathbf{D}^o]^\top \mathbf{D}^o\|_F &\leqslant \|\mathbf{D}^\top (\mathbf{D} - \mathbf{D}^o)\|_F \\
&\quad + \|(\mathbf{D}^\top - [\mathbf{D}^o]^\top) \mathbf{D}^o\|_F \\
&\leqslant 2B \|\mathbf{D} - \mathbf{D}^o\|_F \leqslant 2B \|\boldsymbol{\theta}\|_2.
\end{aligned}$$

Rémi Gribonval (FM'14) is a Senior Researcher with Inria (Rennes, France), and the scientific leader of the PANAMA research group on sparse audio processing. A former student at École Normale Supérieure (Paris, France), he received the Ph. D. degree in applied mathematics from Université de Paris-IX Dauphine (Paris, France) in 1999, and his Habilitation à Diriger des Recherches in applied mathematics from Université de Rennes I (Rennes, France) in 2007. His research focuses on mathematical signal processing, machine learning, approximation theory and statistics, with an emphasis on sparse approximation, audio source separation and compressed sensing.

Rodolphe Jenatton received the PhD degree from the Ecole Normale Superieure, Cachan, France, in 2011 under the supervision of Francis Bach and Jean-Yves Audibert. He then joined the CMAP at Ecole Polytechnique, Palaiseau, France, as a postdoctoral researcher working with Alexandre d'Aspremont. From early 2013 until mid 2014, he worked for Criteo, Paris, France, where he was in charge of improving the statistical and optimization aspects of the ad prediction engine. He is now a machine learning scientist at Amazon Development Center Germany, Kurfürstenddamm, Berlin. His research interests revolve around machine learning, statistics, (convex) optimization, (structured) sparsity and unsupervised models based on latent factor representations.

Francis Bach graduated from the Ecole Polytechnique, Palaiseau, France, in 1997. He received the Ph.D. degree in 2005 from the Computer Science Division at the University of California, Berkeley. He is the leading researcher of the Sierra project-team of INRIA in the Computer Science Department of the Ecole Normale Suprieure, Paris, France. His research interests include machine learning, statistics, optimization, graphical models, kernel methods, and statistical signal processing. He is currently the action editor of the Journal of Machine Learning Research and associate editor of IEEE Transactions in Pattern Analysis and Machine Intelligence.