Local stability and robustness of sparse dictionary learning in the presence of noise

Rodolphe Jenatton, Rémi Gribonval, Francis Bach

Local stability and robustness of sparse dictionary learning in the presence of noise

Rodolphe Jenatton*,⋆ jenatton@cmap.polytechnique.fr
Rémi Gribonval† remi.gribonval@inria.fr
Francis Bach° francis.bach@inria.fr

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 and noisy signals, 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.

1 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., Wainwright [2009], Zhang [2009]) and efficient algorithmic tools (see, e.g., Bach et al. [2011] 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 [Mallat, 2008]. 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., Mairal et al. [2010] and references therein).

Although sparse dictionary learning can sometimes be formulated as convex [Bach et al., 2008, Bradley and Bagnell, 2009], non-parametric Bayesian [Zhou et al., 2009] and submodular [Krause and Cevher, 2010] 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, there has only been little theoretical analysis of the properties of sparse dictionary learning. For instance, Maurer and Pontil [2010], Vainsencher et al. [2010], Mehta and Gray [2012] derive generalization bounds which quantify how much the *expected*

*CMAP, Ecole Polytechnique (UMR CNRS 7641), 91128 Palaiseau, France.
†INRIA Rennes, Campus de Beaulieu, 35042 Rennes, France.
°INRIA - SIERRA project, LIENS (INRIA/ENS/CNRS UMR 8548), 23, avenue d'Italie 75214 Paris, France.
⋆Most of the work was done while affiliated with°.

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 [2010], Vainsencher et al. [2010] are non-asymptotic and uniform with respect to the whole class of dictionaries considered (e.g., those with normalized atoms). As discussed later, the questions raised in this paper explore a different and complementary direction.

Another theoretical aspect of interest consists in characterizing the local minima of the 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 [Comon and Jutten, 2010] or in topic modelling [Jenatton et al., 2011]. The authors of Gribonval and Schnass [2010] pioneered research in this direction by considering noiseless sparse signals, possibly corrupted by some outliers, in the case where the reference dictionary forms a basis. Still in a noiseless setting, and without outliers, Geng et al. [2011] extended the analysis to *over-complete* dictionaries, i.e., these composed of more atoms than the dimension of the signals. To the best of our knowledge, comparable analysis have not been carried out yet for noisy signals. In particular, the structure of the proofs of Gribonval and Schnass [2010], Geng et al. [2011] hinges on the absence of noise and cannot be straightforwardly transposed to take into account some noise; this point will be discussed subsequently.

In this paper, we therefore analyze the local minima of sparse coding *in the presence of noise* and make the following contributions:

- Within a probabilistic model of sparse signals, we derive a *non-asymptotic* lower bound of the probability of finding a local minimum in a neighborhood of the reference dictionary.
- Our work makes it possible to better understand (a) how small the neighborhood around the reference dictionary can be, (b) how many signals are required to hope for identifiability, (c) what the impact of the degree of over-completeness is, and (d) what level of noise appears as manageable.
- We show that under deterministic coherence-based assumptions, such a local minimum is guaranteed to exist with high probability.

2 Problem statement

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

Notation. 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 $\text{sign}(\mathbf{v}) \in \{-1,0,1\}^p$ the vector such that its j-th entry $[\text{sign}(\mathbf{v})]_j$ is equal to zero if $\mathbf{v}_j = 0$, and to one (respectively, minus one) if $\mathbf{v}_j > 0$ (respectively, $\mathbf{v}_j < 0$). We extensively manipulate matrix norms in the sequel. For any matrix $\mathbf{A} \in \mathbb{R}^{n\times p}$, we define its Frobenius norm by $\|\mathbf{A}\|_{\text{F}} \triangleq [\sum_{i=1}^n \sum_{j=1}^p \mathbf{A}_{ij}^2]^{1/2}$; similarly, we denote the spectral norm of $\mathbf{A}$ by $|\!|\!|\mathbf{A}|\!|\!|_2 \triangleq \max_{\|\mathbf{x}\|_2\leq 1} \|\mathbf{A}\mathbf{x}\|_2$, and refer to the operator ℓ_∞-norm as $|\!|\!|\mathbf{A}|\!|\!|_\infty \triangleq \max_{\|\mathbf{x}\|_\infty\leq 1} \|\mathbf{A}\mathbf{x}\|_\infty = \max_{i\in[\![1;n]\!]} \sum_{j=1}^p |\mathbf{A}_{ij}|$.

For any square matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$, we denote by $\text{diag}(\mathbf{B}) \in \mathbb{R}^n$ the vector formed by extracting the diagonal terms of $\mathbf{B}$, and conversely, for any $\mathbf{b} \in \mathbb{R}^n$, we use $\text{Diag}(\mathbf{b}) \in \mathbb{R}^{n\times n}$ to represent the (square) diagonal matrix whose diagonal elements are built from the vector $\mathbf{b}$. For any $m \times p$ matrix $\mathbf{A}$ and index set $\text{J} \subset [\![1;p]\!]$ we denote by $\mathbf{A}_\text{J}$ the matrix obtained by concatenating the columns of $\mathbf{A}$ indexed by J. Finally, the sphere in $\mathbb{R}^p$ is denoted $\mathcal{S}^p \triangleq \{\mathbf{v} \in \mathbb{R}^p;\ \|\mathbf{v}\|_2 = 1\}$ and $\mathcal{S}_+^p \triangleq \mathcal{S}^p \cap \mathbb{R}_+^p$.

2.1 Background material on sparse coding

Let us consider a set of n signals $\mathbf{X} \triangleq [\mathbf{x}^1,\dots,\mathbf{x}^n] \in \mathbb{R}^{m\times n}$ 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 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 [Mairal et al., 2010, Olshausen and Field, 1997], we need some definitions:

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

$$\begin{aligned} \mathcal{L}_{\mathbf{x}}(\mathbf{D}, \boldsymbol{\alpha}) &\triangleq \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 \\ f_{\mathbf{x}}(\mathbf{D}) &\triangleq \min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \mathcal{L}_{\mathbf{x}}(\mathbf{D}, \boldsymbol{\alpha}). \end{aligned} \tag{1}$$

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

$$F_n(\mathbf{D}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_{\mathbf{x}^i}(\mathbf{D}).$$

Based on problem (1), refered to as Lasso in statistics [Tibshirani, 1996], and basis pursuit in signal processing [Chen et al., 1998], the standard approach to perform sparse coding [Olshausen and Field, 1997, Mairal et al., 2010] solves the minimization problem

$$\min_{\mathbf{D}\in\mathcal{D}} F_n(\mathbf{D}), \tag{2}$$

where the regularization parameter λ in (1) controls the level of sparsity, while $\mathcal{D} \subseteq \mathbb{R}^{m\times p}$ is a compact set; in this paper, $\mathcal{D}$ denotes the set of dictionaries with unit ℓ_2-norm atoms, which is a natural choice in image processing [Mairal et al., 2010, Gribonval and Schnass, 2010]. 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. [2011] where in the context of topic models, the atoms in $\mathcal{D}$ belong to the unit simplex).

2.2 Main objectives

The goal of the paper is to characterize some local minima of the function F_n under a generative model for the signals $\mathbf{x}^i$. Throughout the paper, we assume the observed signals are generated *independently* according to a specified probabilistic model. The considered signals are typically drawn as $\mathbf{x}^i \triangleq \mathbf{D}_0\boldsymbol{\alpha}_0^i + \boldsymbol{\varepsilon}^i$ where $\mathbf{D}_0$ is a fixed reference dictionary, $\boldsymbol{\alpha}_0^i$ is a sparse coefficient vector, and $\boldsymbol{\varepsilon}^i$ is a noise term. The specifics of the underlying probabilistic model are given in Sec. 2.6. Under this model, we can state more precisely our objective: we want to show that

$$\Pr\big(F_n \text{ has a local minimum in a "neighborhood" of } \mathbf{D}_0\big) \approx 1.$$

We loosely refer to a certain "neighborhood" since in our regularized formulation, a local minimum cannot appear exactly at $\mathbf{D}_0$. The proper meaning of this neighborhood is the subject of Sec. 2.3.

Intrinsic ambiguities of sparse coding. Importantly, we have so far referred to $\mathbf{D}_0$ as *the* reference dictionary generating the signals. However, and as already discussed in Gribonval and Schnass [2010], Geng et al. [2011] and more generally the related literature on blind source separation and independent component analysis [Comon and Jutten, 2010], it is known that the objective of (2) is invariant by sign flips and atoms permutations. As a result, while solving (2), we cannot hope to identify the specific $\mathbf{D}_0$. 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 $\mathbf{D}_0$ 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 $\mathbf{D}_0$.

2.3 Local minima on the oblique manifold

The minimization of F_n is carried out over $\mathcal{D}$, which is the set of dictionaries with unit ℓ_2-norm atoms. This set turns out to be a manifold, known as the *oblique manifold* [Absil et al., 2008]. Since $\mathbf{D}_0$ is assumed to belong to $\mathcal{D}$, it is therefore natural to consider the behavior of F_n according to the geometry and topology of $\mathcal{D}$. To this end, we consider a specific (local) parametrization of $\mathcal{D}$.

Parametrization of the oblique manifold. Specifically, let us consider the set of matrices

$$\mathcal{W}_{\mathbf{D}_0} \triangleq \big\{\mathbf{W} \in \mathbb{R}^{m\times p};\ \mathrm{diag}(\mathbf{W}^\top \mathbf{D}_0) = \mathbf{0} \text{ and } \mathrm{diag}(\mathbf{W}^\top \mathbf{W}) = \mathbf{1}\big\}.$$

In words, a matrix $\mathbf{W} \in \mathcal{W}_{\mathbf{D}_0}$ has unit norm columns $\|\mathbf{w}^j\|_2 = 1$ that are orthogonal to the corresponding columns of $\mathbf{D}_0$: $[\mathbf{w}^j]^\top \mathbf{d}^j = 0$, for any $j \in [\![1; p]\!]$. Now, for any matrix $\mathbf{W} \in \mathcal{W}_{\mathbf{D}_0}$, for any unit norm *velocity* vector $\mathbf{v} \in \mathcal{S}^p$, and for all $t \in \mathbb{R}$, we introduce the parameterized dictionary:

$$\mathbf{D}(\mathbf{D}_0, \mathbf{W}, \mathbf{v}, t) \triangleq \mathbf{D}_0 \mathrm{Diag}[\cos(\mathbf{v}t)] + \mathbf{W}\mathrm{Diag}[\sin(\mathbf{v}t)], \tag{3}$$

where $\mathrm{Diag}[\cos(\mathbf{v}t)]$ and $\mathrm{Diag}[\sin(\mathbf{v}t)] \in \mathbb{R}^{p\times p}$ stand for the diagonal matrices with diagonal terms equal to $\{\cos(\mathbf{v}_j t)\}_{j\in[\![1;p]\!]}$ and $\{\sin(\mathbf{v}_j t)\}_{j\in[\![1;p]\!]}$ respectively. By construction, we have $\mathbf{D}(\mathbf{D}_0, \mathbf{W}, \mathbf{v}, t) \in \mathcal{D}$ for all $t \in \mathbb{R}$ and $\mathbf{D}(\mathbf{D}_0, \mathbf{W}, \mathbf{v}, 0) = \mathbf{D}_0$. To ease notation, we will denote $\mathbf{D}(\mathbf{W}, \mathbf{v}, t)$, leaving the dependence on the reference dictionary $\mathbf{D}_0$ implicit. Also, when it will be made clear from the context, we will drop the dependence on $\mathbf{W}, \mathbf{v}$ in $\mathbf{D}$. Note that the set of matrices given by $\mathbf{W}\mathrm{Diag}(\mathbf{v})$ corresponds to the tangent space of $\mathcal{D}$ at $\mathbf{D}_0$, intersected with the set of matrices in $\mathbb{R}^{m\times p}$ with unit Frobenius norm (since we have $\|\mathbf{W}\mathrm{Diag}(\mathbf{v})\|_{\mathrm{F}} = 1$).

Characterization of local minima on the oblique manifold. We can exploit the above parametrization of the manifold $\mathcal{D}$ to characterize the existence of a local minimum as follows:

Proposition 1 (Local minimum characterization)**.** *Let $t > 0$ be some fixed scalar and define*

$$\Delta F_n(\mathbf{W}, \mathbf{v}, t) \triangleq F_n(\mathbf{D}(\mathbf{W}, \mathbf{v}, t)) - F_n(\mathbf{D}_0). \tag{4}$$

If we have

$$\inf_{\mathbf{W}\in\mathcal{W}_{\mathbf{D}_0},\ \mathbf{v}\in\mathcal{S}^p_+} \Delta F_n(\mathbf{W}, \mathbf{v}, t) > 0,$$

then $F_n : \mathcal{D} \to \mathbb{R}_+$ admits a local minimum in $\big\{\mathbf{D} \in \mathcal{D};\ \|\mathbf{D}_0 - \mathbf{D}\|_{\mathrm{F}} < t\big\}$.

The detailed proof of this result is given in Sec. A of the appendix. It relies on the continuity of F_n and the fact that the curves $\mathbf{D}(\mathbf{W}, \mathbf{v}, t)$ define a surjective mapping onto $\mathcal{D}$ (see Lemma 1 in the appendix). We next describe some other ingredients required to state our results.

2.4 Closed-form expression for F_n?

Although the function F_n is Lipschitz-continuous [Mairal et al., 2010], its minimization is challenging since it is non-convex and subject to the non-linear constraints of $\mathcal{D}$. Moreover, F_n is defined through the minimization over the vectors $\mathbf{A}$, which, at first sight, does not lead to a simple and convenient expression. However, it is known that F_n has a simple closed-form in some favorable scenarios.

Closed-form expression for $f_{\mathbf{x}}$. We leverage here a key property of the function $f_{\mathbf{x}}$. Denote by $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^p$ a solution of problem (1), that is, the minimization defining $f_{\mathbf{x}}$. By the convexity of the problem, there always exists such a solution such that, denoting $\mathrm{J} \triangleq \{j \in [\![1; p]\!];\ \hat{\boldsymbol{\alpha}}_j \neq 0\}$ 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). Denoting $\hat{\mathbf{s}} \in \{-1, 0, 1\}^p$ the sign of $\hat{\boldsymbol{\alpha}}$ and J its support, $\hat{\boldsymbol{\alpha}}$ has a closed-form expression in terms of $\mathbf{D}_{\mathrm{J}}$, $\mathbf{x}$ and $\hat{\mathbf{s}}$ (see, e.g., Wainwright [2009], Fuchs [2005]). This property is appealing in that it makes it possible to obtain a closed-form expression for $f_{\mathbf{x}}$ (and hence, F_n), *provided that we can control the sign patterns of $\hat{\boldsymbol{\alpha}}$*. 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* J *be its support. For $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{D} \in \mathbb{R}^{m\times p}$, we define*

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

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

$$\tilde{\boldsymbol{\alpha}}_{\mathrm{J}} = \left[\mathbf{D}_{\mathrm{J}}^\top \mathbf{D}_{\mathrm{J}}\right]^{-1}\left[\mathbf{D}_{\mathrm{J}}^\top \mathbf{x} - \lambda \mathbf{s}_{\mathrm{J}}\right] \in \mathbb{R}^{|\mathrm{J}|} \quad \textit{and} \quad \tilde{\boldsymbol{\alpha}}_{\mathrm{J}^c} = \mathbf{0},$$

and we have

$$\phi_{\mathbf{x}}(\mathbf{D}|\mathbf{s}) = \frac{1}{2}\left[\|\mathbf{x}\|_2^2 - (\mathbf{D}_{\mathrm{J}}^\top \mathbf{x} - \lambda \mathbf{s}_{\mathrm{J}})^\top (\mathbf{D}_{\mathrm{J}}^\top \mathbf{D}_{\mathrm{J}})^{-1}(\mathbf{D}_{\mathrm{J}}^\top \mathbf{x} - \lambda \mathbf{s}_{\mathrm{J}})\right]. \tag{5}$$

Moreover, if $\mathrm{sign}(\tilde{\boldsymbol{\alpha}}) = \mathbf{s}$*, then*

$$\phi_{\mathbf{x}}(\mathbf{D}|\mathbf{s}) = \min_{\boldsymbol{\alpha}\in\mathbb{R}^p,\ \mathrm{sign}(\boldsymbol{\alpha})=\mathbf{s}} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda \mathbf{s}^\top \boldsymbol{\alpha} = \min_{\boldsymbol{\alpha}\in\mathbb{R}^p,\ \mathrm{sign}(\boldsymbol{\alpha})=\mathbf{s}} \mathcal{L}_{\mathbf{x}}(\mathbf{D}, \boldsymbol{\alpha}) = \mathcal{L}_{\mathbf{x}}(\mathbf{D}, \tilde{\boldsymbol{\alpha}}).$$

We define $\Phi_n(\mathbf{D}|\mathbf{S})$ *analogously to* $F_n(\mathbf{D})$*, for a sign matrix* $\mathbf{S} \in \{-1, 0, 1\}^{p\times n}$*.*

Hence, with $\hat{\mathbf{s}}$ the sign of the (unknown) minimizer $\hat{\boldsymbol{\alpha}}$, we have $f_{\mathbf{x}}(\mathbf{D}) = \mathcal{L}_{\mathbf{x}}(\mathbf{D}, \hat{\boldsymbol{\alpha}}) = \phi_{\mathbf{x}}(\mathbf{D}|\hat{\mathbf{s}})$.

Showing that the function F_n is accurately approximated by $\Phi_n(\cdot|\mathbf{S})$ for a controlled $\mathbf{S}$ will be a key ingredient of our approach. This will exploit sign recovery properties of ℓ_1-regularized least-squares problems, a topic which is already well-understood (see, e.g., Wainwright [2009], Fuchs [2005] and references therein).

2.5 Coherence assumption on the reference dictionary $\mathbf{D}_0$

We consider a standard sufficient support recovery condition referred to as the *exact recovery condition* in signal processing [Fuchs, 2005, Tropp, 2004] or the *irrepresentability condition* (IC) in the machine learning and statistics communities [Wainwright, 2009, Zhao and Yu, 2006]. It is a key element to 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 *mutual coherence* of the reference dictionary $\mathbf{D}_0$, which is a stronger requirement Van de Geer and Bühlmann [2009]. This quantity is defined (see, e.g., Fuchs [2005], Donoho and Huo [2001]) as

$$\mu_0 \triangleq \max_{i,j\in[\![1;p]\!],i\neq j} |[\mathbf{d}_0^i]^\top [\mathbf{d}_0^j]| \in [0, 1].$$

The term μ_0 gives a measure of the level of correlation between columns of $\mathbf{D}_0$. It is for instance equal to zero in the case of an orthogonal dictionary, and to one if $\mathbf{D}_0$ contains two colinear columns. Similarly, we introduce $\mu(\mathbf{W}, \mathbf{v}, t)$ for the dictionary $\mathbf{D}(\mathbf{W}, \mathbf{v}, t)$ defined in (3). For any $\mathbf{W}, \mathbf{v}$, $t \geq 0$, we have the simple inequality:

$$\mu(\mathbf{W}, \mathbf{v}, t) \triangleq \max_{i,j\in[\![1;p]\!],i\neq j} |[\mathbf{d}^i(\mathbf{W}, \mathbf{v}, t)]^\top [\mathbf{d}^j(\mathbf{W}, \mathbf{v}, t)]| \leq \mu(t) \triangleq \mu_0 + 3t. \tag{6}$$

In particular, we have $\mu(\mathbf{W}, \mathbf{v}, 0) = \mu_0$. For the theoretical analysis we conduct, we consider a deterministic coherence-based assumption, as considered for instance in the previous work on dictionary learning by Geng et al. [2011], such that the coherence μ_0 and the level of sparsity k of the coefficient vectors $\boldsymbol{\alpha}^i$ should be inversely proportional, i.e., $k\mu_0 = O(1)$. In light of (6), such an upper bound on μ_0 will loosely transfer to $\mu(t)$ provided that t is small enough. In fact, and as further developed in the appendix, most of the elements of our proofs work based on a restricted isometry property (RIP), which is known to be weaker than the coherence assumption [Van de Geer and Bühlmann, 2009]. However, since we still face a problem related to IC when using RIP, we keep the coherence in our analysis. Unifying our proofs under a RIP criterion is the object of future work.

2.6 Probabilistic model of sparse signals

Equipped with the main concepts, we now present our signal model. Given a *fixed* reference dictionary $\mathbf{D}_0 \in \mathcal{D}$, each noisy sparse signal $\mathbf{x} \in \mathbb{R}^m$ is built *independently* from the following steps:

(1) **Support generation:** Draw uniformly without replacement k atoms out of the p available in $\mathbf{D}_0$. This procedure thus defines a support $\mathrm{J} \triangleq \{j \in [\![1;p]\!];\ \delta(j)=1\}$ whose size is $|\mathrm{J}|=k$, and where $\delta(j)$ denotes the indicator function equal to one if the j-th atom is selected, zero otherwise, so that

$$\mathbb{E}[\delta(j)] = \tfrac{k}{p}, \text{ and for } i \neq j, \text{ we further have } \mathbb{E}[\delta(j)\delta(i)] = \tfrac{k(k-1)}{p(p-1)}.$$

Our result holds for any support generation scheme yielding the above expectations.
(2) **Coefficient generation:** Define a sparse vector $\boldsymbol{\alpha}_0 \in \mathbb{R}^p$ supported on J whose entries in J are generated i.i.d. according to a *sub-Gaussian distribution*: for j not in J, $[\boldsymbol{\alpha}_0]_j$ is set to zero; on the other hand, we assume there exists some $c>0$ such that for $j \in \mathrm{J}$ we have, for all $t \in \mathbb{R}$, $\mathbb{E}\{\exp(t[\boldsymbol{\alpha}_0]_j)\} \leq \exp(c^2t^2/2)$. We denote σ_α the smallest value of c such that this property holds. For background about sub-Gaussian random variables, see, e.g., Buldygin and Kozachenko [2000]. For simplicity of the analysis we restrict to the case where the distribution also has all its mass *bounded away from zero*. Formally, there exist $\underline{\alpha} > 0$ such that $\Pr(|[\boldsymbol{\alpha}_0]_j| < \underline{\alpha} \mid j \in \mathrm{J}) = 0$.
(3) **Noise:** Eventually generate the signal $\mathbf{x} = \mathbf{D}_0\boldsymbol{\alpha}_0 + \boldsymbol{\varepsilon}$, where the entries of the additive noise $\boldsymbol{\varepsilon} \in \mathbb{R}^m$ are assumed i.i.d. sub-Gaussian with parameter σ.

3 Main results

This section describes the main results of this paper which show that under appropriate scalings of the dimensions (m,p), number of samples n, and model parameters $k, \underline{\alpha}, \sigma_\alpha, \sigma, \mu_0$, it is possible to prove that, with high probability, the problem (2) admits a local minimum in a neighborhood of $\mathbf{D}_0$ of controlled size, for appropriate choices of the regularization parameter λ. The detailed proofs of the following results may be found in the appendix, but we provide their main outlines in Sec. B.

Theorem 1 (Local minimum of sparse coding)**.** *Let us consider our generative model of signals for some reference dictionary $\mathbf{D}_0 \in \mathbb{R}^{m\times p}$ with coherence μ_0, and define $1/\gamma_{\mathbf{D}_0} \triangleq \|\mathbf{D}_0\|_2 \cdot k\mu_0$, where $\|\mathbf{D}_0\|_2$ refers to the spectral norm of $\mathbf{D}_0$. If the following conditions hold:*

$$\textit{(Coherence)} \qquad \Omega\big(\sqrt{\log(p)}\big) = \gamma_{\mathbf{D}_0} = O\big(\sqrt{\log(n)}\big),$$

$$\textit{(Sample complexity)} \qquad \frac{\log(n)}{n} = O\Big(\frac{\mu_0^2}{m\cdot p^3 \cdot \gamma_{\mathbf{D}_0}^2}\Big),$$

then, with probability exceeding $1-[\frac{mpn}{9}]^{-\frac{mp}{2}} - e^{-4\sqrt{n}}$, problem (2) admits a local minimum in

$$\left\{\mathbf{D}\in\mathcal{D};\ \|\mathbf{D}_0-\mathbf{D}\|_{\mathrm{F}} = O\Big(\max\Big\{p\cdot\gamma_{\mathbf{D}_0}\cdot\Big[e^{-\frac{\gamma_{\mathbf{D}_0}^2}{2}} + \sqrt{mp\log(n)/n}\Big],\ \frac{\sigma}{\sigma_\alpha}\cdot\sqrt{m}\Big\}\Big)\right\}.$$

First, it is worth noting that this theorem is presented on purpose in a simplified form, in order to highlight its message. In particular, all quantities related to the distribution of $\boldsymbol{\alpha}_0$ (e.g., σ_α) are assumed to be $O(1)$ and are therefore kept "hidden" in the big-O notation. A detailed statement of this theorem is however available in the appendix (see Theorem 3).

In words, the main message of Theorem 1 is that provided (a) the reference dictionary is incoherent enough, and (b) we observe enough signals, we can guarantee the existence of a local minimum for problem (2) in a ball centered at $\mathbf{D}_0$. We can see that the radius of this ball decomposes according to three different contributions: (1) the coherence of $\mathbf{D}_0$, via the term $\gamma_{\mathbf{D}_0}$, (2) the number of signals, and (3) the level of noise. These three factors limit the possible resolution we can guarantee.

While a coherence condition scaling in $k\mu_0 = O(1)$ is standard for sparse models (see, e.g., Fuchs [2005]), we impose a slightly more conservative constraint in $O(1/\sqrt{\log(p)})$. A typical example for which our result applies is the Hadamard-Dirac dictionary built as the concatenaton of a Hadamard matrix and the identity matrix. In this case, we have $p=2m$, $\|\mathbf{D}_0\|_2 = \sqrt{2}$, and $\mu_0 = 1/\sqrt{m}$ with $k = O(\sqrt{m/\log(2m)})$. In Sec. 5, we use such over-complete dictionaries for our simulations. In addition, observe that because of the

upperbound on $\gamma_{\mathbf{D}_0}$, Theorem 1 does not handle per se the case of orthogonal dictionary, which we remedy in Theorem 2.

Perhaps surprisingly (and disappointingly), our result indicates that, even in a low-noise setting with sufficiently many signals (i.e., the asymptotic regime in n), we cannot arbitrarily lower the resolution of the local minimum because of the coherence μ_0. In fact, the term $e^{-\gamma^2_{\mathbf{D}_0}/2}$ is a direct consequence of our proof technique which relies on exact recovery. It is however worth noting that, since $e^{-\gamma^2_{\mathbf{D}_0}/2}$ decreases exponentially fast in $\gamma_{\mathbf{D}_0}$, the dependence on μ_0 is quite mild (e.g., for a radius τ, we have a constraint scaling in $\|\mathbf{D}_0\|_2 \cdot k\mu_0 = O(1/\sqrt{\log(1/\tau)})$). We next state a complementary theorem for orthogonal dictionaries where the radius is not constrained anymore by the coherence:

Local correctness of sparse coding with orthogonal dictionaries: If we now assume that $\mathbf{D}_0$ is orthogonal (i.e., $\mu_0 = 0$ and $p = m$ with $\|\mathbf{D}_0\|_2 = 1$), we obtain the following result:

Theorem 2 (Local minimum of sparse coding—Orthogonal dictionary)**.** *Let us consider our generative model of signals for some reference, orthogonal dictionary $\mathbf{D}_0 \in \mathbb{R}^{m\times m}$. If we have:*

$$\textit{(Sample complexity)} \quad \frac{\log^3(n)}{n} = O\Big(\frac{1}{k^2 \cdot m^4}\Big),$$

then, with probability exceeding $1 - [\frac{m^2 n}{9}]^{-\frac{m^2}{2}} - e^{-4\sqrt{n}}$, problem (2) admits a local minimum in

$$\left\{\mathbf{D} \in \mathcal{D};\ \|\mathbf{D}_0 - \mathbf{D}\|_{\mathrm{F}} = O\Big(\max\Big\{m \cdot \log(n) \cdot (\sqrt{\log(n)} + m)/\sqrt{n},\ \frac{\sigma}{\sigma_\alpha} \cdot \sqrt{m}\Big\}\Big)\right\}.$$

Interestingly, we observe in this case that, given sufficiently many signals, we can localize arbitrarily well (up to the noise level) the local minimum around $\mathbf{D}_0$. We now discuss relations with previous work in the noiseless setting.

Local correctness of sparse coding without noise: If we remove the noise from our signal model, i.e., $\sigma = 0$, the result of Theorems 1-2 remains unchanged, except that the radius is not limited anymore by $\frac{\sigma}{\sigma_\alpha}\sqrt{m}$. We mention that Gribonval and Schnass [2010] obtain a sample complexity in $O(p^2 \log(p))$ in the noiseless and *square* dictionary setting, while the result of Geng et al. [2011] leads to a scaling in $O(p^3)$ (assuming both $k = O(1)$ and $\|\mathbf{D}_0\|_2 = O(1)$) in the noiseless, over-complete case. In comparison, our analysis suggests a sample complexity in $O(mp^3)$.

These discrepancies are due to the fact that we want to handle the noisy setting; this has led us to consider a scheme of proof radically different from those proposed in the related work Gribonval and Schnass [2010], Geng et al. [2011]. In particular, our formulation in problem (2) differs from that of Gribonval and Schnass [2010], Geng et al. [2011] where the ℓ_1-norm of $\mathbf{A}$ is minimized over the *equality* constraint $\mathbf{DA} = \mathbf{X}$ and the dictionary normalization $\mathbf{D} \in \mathcal{D}$. Optimality is then characterized through the linearization of the equality constraint, a technique that could not be easily extended to the noisy case. We next discuss the main building blocks of the results and give a high-level structure of the proof.

4 Architecture of the proof of Theorem 1

Our proof strategy consists in using Proposition 1, that is, controlling the sign of $\Delta F_n(\mathbf{W}, \mathbf{v}, t)$ defined in (8). In fact, since we expect to have for many training samples the equality $f_{\mathbf{x}}(\mathbf{D}(\mathbf{W}, \mathbf{v}, t)) = \phi_{\mathbf{x}}(\mathbf{D}(\mathbf{W}, \mathbf{v}, t)|\mathrm{sign}(\boldsymbol{\alpha}_0))$ uniformly for all $(\mathbf{W}, \mathbf{v})$, the main idea is to first concentrate on the study of the smooth function

$$\Delta\Phi_n(\mathbf{W}, \mathbf{v}, t) \triangleq \Phi_n(\mathbf{D}(\mathbf{W}, \mathbf{v}, t)|\mathrm{sign}(\mathbf{A}_0)) - \Phi_n(\mathbf{D}_0|\mathrm{sign}(\mathbf{A}_0)), \tag{7}$$

instead of the original function $\Delta F_n(\mathbf{W}, \mathbf{v}, t)$.

Control of $\Delta\Phi_n$: This first step consists in uniformly lower bounding $\Delta\Phi_n$ with high probability. As opposed to ΔF_n, the function $\Delta\Phi_n$ is available explicitly, see (2) and (18), and corresponds to bilinear/quadratic forms in $(\boldsymbol{\alpha}_0, \mathrm{sign}(\boldsymbol{\alpha}_0), \boldsymbol{\varepsilon})$ which we can concentrate around their expectations. Finally, the uniformity with respect to $(\mathbf{W}, \mathbf{v})$ is obtained by a standard ϵ-net argument.

Control of ΔF_n via $\Delta\Phi_n$: The second step consists in lower bounding ΔF_n in terms of $\Delta\Phi_n$ uniformly for all parameters $(\mathbf{W}, \mathbf{v}) \in \mathcal{W}_{\mathbf{D}_0} \times \mathcal{S}^p$. For a given $t \geq 0$, consider the independent events $\{\mathcal{E}^i_{\mathrm{coincide}}(t)\}_{i \in [\![1;n]\!]}$ defined by

$$\mathcal{E}^i_{\mathrm{coincide}}(t) \triangleq \left\{ \omega \;\middle|\; f_{\mathbf{x}^i(\omega)}(\mathbf{D}(\mathbf{W}, \mathbf{v}, t)) = \phi_{\mathbf{x}^i(\omega)}(\mathbf{D}(\mathbf{W}, \mathbf{v}, t)|\mathbf{s}_0), \quad \forall (\mathbf{W}, \mathbf{v}) \in \mathcal{W}_{\mathbf{D}_0} \times \mathcal{S}^p \right\},$$

with $\mathbf{s}_0 = \mathrm{sign}(\boldsymbol{\alpha}_0)$. In words, the event $\mathcal{E}^i_{\mathrm{coincide}}(t)$ corresponds to the fact that target function $f_{\mathbf{x}^i(\omega)}(\mathbf{D}(\cdot, \cdot, t))$ coincides with the idealized one $\phi_{\mathbf{x}^i(\omega)}(\mathbf{D}(\cdot, \cdot, t)|\mathbf{s}_0)$ for the "radius" t.

Importantly, the event $\mathcal{E}^i_{\mathrm{coincide}}(t)$ only involves a *single* signal; when we consider a collection of n independent signals, we should instead study the event $\bigcap_{i=1}^n \mathcal{E}^i_{\mathrm{coincide}}(t)$ to guarantee that Φ_n and F_n (and therefore, $\Delta\Phi_n$ and ΔF_n) do coincide. However, as the number of observations n becomes large, it is unrealistic and not possible to ensure exact recovery both *simultaneously* for the n signals and *with high probability*. To get around this issue, we seek to prove that ΔF_n is well approximated by $\Delta\Phi_n$ (rather than equal to it) uniformly for all $(\mathbf{W}, \mathbf{v})$. We show that, when $f_{\mathbf{x}^i}(\mathbf{D}(t))$ and $\phi_{\mathbf{x}^i}(\mathbf{D}(t)|\mathbf{s}_0)$ *do not* coincide, their difference can be bounded, and we obtain:

$$\Delta F_n(\mathbf{W}, \mathbf{v}, t) \geq \Delta\Phi_n(\mathbf{W}, \mathbf{v}, t) - r_n.$$

where we detail the definition of the residual term

$$r_n(\omega) \triangleq \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{[\mathcal{E}^i_{\mathrm{coincide}}(t) \cap \mathcal{E}^i_{\mathrm{coincide}}(0)]^c}(\omega) \cdot \left\{ \mathcal{L}_{\mathbf{x}^i}(\mathbf{D}, \boldsymbol{\alpha}^i_0) + \mathcal{L}_{\mathbf{x}^i}(\mathbf{D}_0, \boldsymbol{\alpha}^i_0) \right\}.$$

In the appendix, we show that with high probability: $r_n = O([t^2 \cdot \sigma_\alpha^2 + 2m \cdot \sigma^2 + 2\lambda k \sigma_\alpha] \cdot (3 - \log \kappa)\kappa)$ with $\kappa \triangleq \max_{i \in [\![1;n]\!]} \Pr(\left[\mathcal{E}^i_{\mathrm{coincide}}(t) \cap \mathcal{E}^i_{\mathrm{coincide}}(0)\right]^c)$. To bound the size of r_n, we now control κ.

Control of κ, exact sign recovery for *perturbed* dictionaries: We need to determine sufficient conditions under which $\phi_{\mathbf{x}}(\mathbf{D}(\mathbf{W}, \mathbf{v}, t)|\mathrm{sign}(\boldsymbol{\alpha}_0))$ and $f_{\mathbf{x}}(\mathbf{D}(\mathbf{W}, \mathbf{v}, t))$ coincide for all $(\mathbf{W}, \mathbf{v})$, and control the probability of this event. As briefly exposed in Sec. 2.1, it turns out that this question comes down to studying exact recovery for some ℓ_1-regularized least-squares problems. Exact sign recovery in the problem associated with $f_{\mathbf{x}}(\mathbf{D}_0)$ has already been well-studied (see, e.g., Wainwright [2009], Fuchs [2005], Zhao and Yu [2006]). However, in our context, we need the same conclusion to hold *not only at the dictionary* $\mathbf{D}_0$, *but also at* $\mathbf{D}(\mathbf{W}, \mathbf{v}, t) \neq \mathbf{D}_0$ *uniformly* for all parameters $(\mathbf{W}, \mathbf{v})$. It turns out that going away from the reference dictionary $\mathbf{D}_0$ acts as a second source of noise whose variance depends on the radius t. We make this statement precise in Propositions 2-3 in the supplementary material. These results are in the same line as Theorem 1 in Mehta and Gray [2012].

Discussing when the lower-bound on ΔF_n is positive: With all the previous elements in place, we have a lower-bound for $\inf_{\mathbf{W} \in \mathcal{W}_{\mathbf{D}_0}, \mathbf{v} \in \mathcal{S}^p} \Delta F_n(\mathbf{W}, \mathbf{v}, t)$, valid with high probability. It finally suffices to discuss when it is stricly positive to conclude with Proposition 1.

5 Experiments

We illustrate the results from Sec. 3. Although we do not manage to highlight the exact scalings in (p, m) which we proved in Theorems 1-2, our experiments still underline the main interesting trends put forward by our results, such as the dependencies with respect to n and σ.

Throughout this section, the non-zero coefficients of $\boldsymbol{\alpha}_0$ are uniformly drawn with $|[\boldsymbol{\alpha}_0]_j| \in [0.1, 10]$ and the noise follows a standard Gaussian distribution with variance σ. We detail two important aspects of the experiments, namely, the choice of λ, and how we deal with the invariance of problem (2) (see Sec. 2.2). Since our analysis relies on exact recovery, we first tune λ over a logarithmic grid to match the oracle sparsity level. Note that this tuning step is performed over an auxiliary set of signals. On the other hand, we know that the dictionary $\hat{\mathbf{D}}$ that we learn by minimizing problem (2) may differ from $\mathbf{D}_0$ up to sign flips and atom permutations. Since both $\hat{\mathbf{D}}$ and $\mathbf{D}_0$ have normalized atoms, finding the closest dictionary (in Frobenius norm) up to these transformations is equivalent to an assignment problem based on the absolute correlation matrix $\hat{\mathbf{D}}^\top \mathbf{D}_0$, which can be efficiently solved using the Hungarian algorithm [Kuhn, 1955].

To solve problem (2), we use the stochastic algorithm from Mairal et al. [2010][1] where the batch size is fixed to 512, while the number of epochs is chosen so as to pass over each signal 25 times (on average). We consider two types of initialization, i.e., either from (1) a random dictionary, or (2) the correct $\mathbf{D}_0$.

To begin with, we illustrate Theorem 1 with $\mathbf{D}_0$ a Hadamard-Dirac (over-complete) dictionary. The sparsity level is fixed such that $\|\mathbf{D}_0\|_2 \cdot k\mu_0 = O(1/\sqrt{\log(p)})$, and we consider a small enough noise level, so that the radius is primarily limited by the number n of signals. The normalized error $\|\mathbf{D}_0 - \hat{\mathbf{D}}\|_{\mathrm{F}}/\sqrt{mp^3}$ versus n is plotted in Fig. 1. We then focus on Theorem 2, with $\mathbf{D}_0$ a Hadamard (orthogonal) dictionary. We consider sufficiently many signals ($n = 75,000$) so that the radius is only limited by $\sqrt{m} \cdot \sigma/\sigma_\alpha$. The normalized error $\|\mathbf{D}_0 - \hat{\mathbf{D}}\|_{\mathrm{F}}/\sqrt{m}$ versus the level of noise is displayed in Fig. 1.

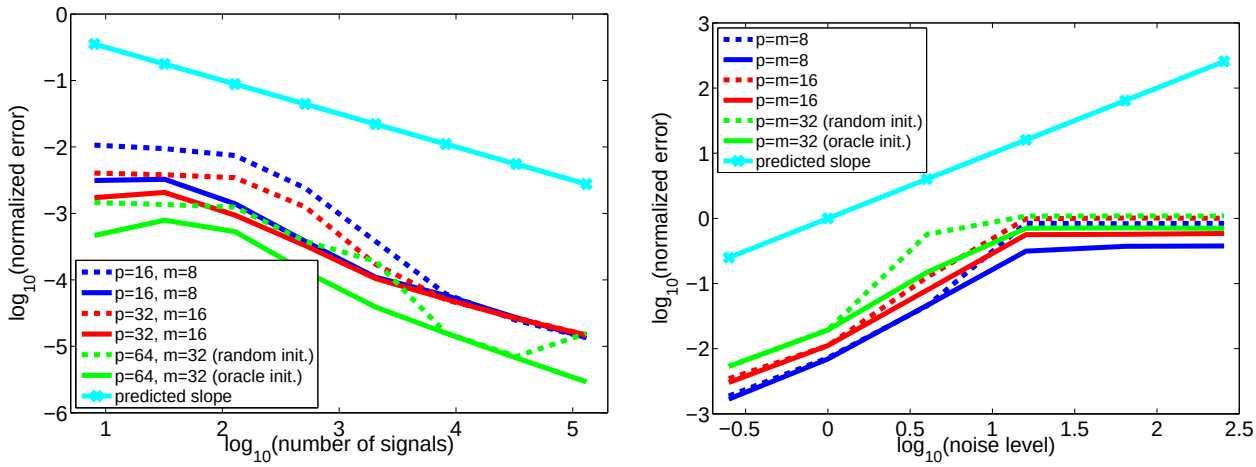

Figure 1: Normalized error between $\mathbf{D}_0$ and the solution of problem (2), versus the number of signals (left) and the noise level (right). The curves represent the median error based on 5 runs, for random and oracle initializations. More details can be found in the text; best seen in color.

The curves represented in Fig. 1 do not perfectly superimposed, thus implying that our results do not capture the exact scalings in (p, m) (our bounds appear in fact as too pessimistic). However, our theory seems to account for the main dependencies with respect to n and σ, as the good agreement with the predicted slopes proves it. Interestingly, while we would expect the curves in the left plot of Fig. 1 to tail off at some point because of the coherence (term $e^{-\gamma^2_{\mathbf{D}_0}/2}$ in the bound of the radius), it seems that there is in practice a much milder dependency with respect to the coherence. Finally, we can observe that both the random and oracle initializations seem to lead to the same behavior, thus raising the questions of the potential *global* characterization of these local minima.

[1]The code is available at `http://www.di.ens.fr/willow/SPAMS/`.

6 Conclusion

We have conducted a non-asymptotic analysis of the local minima of sparse coding in the presence of noise, thus extending prior work which focused on noiseless settings [Gribonval and Schnass, 2010, Geng et al., 2011]. Within a probabilistic model of sparse signals, we have shown that a local minimum exists with high probability around the reference dictionary.

Our study can be further developed in multiple ways. On the one hand, while we have assumed *deterministic* coherence-based conditions scaling in $O(1/k)$, it may interesting to consider non-deterministic assumptions [Candès and Plan, 2009], which are likely to lead to improved scalings. On the other hand, we may also use more realistic generative models for $\boldsymbol{\alpha}_0$, for instance, spike and slab models [Ishwaran and Rao, 2005], or signals with compressible priors [Gribonval et al., 2011].

Also, we believe that our approach can handle the presence of outliers, provided their total energy remains small enough; we plan to make this argument formal in future work.

Finally, it remains challenging to extend our local properties to global ones due to the intrinsic non-convexity of the problem; an appropriate use of convex relaxation techniques [Bach et al., 2008] may prove useful in this context.

Acknowledgements

This work was supported by the European Research Council (SIERRA and SIPA Projects) and by the EU FP7, SMALL project, FET-Open grant number 225913.

References

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

F. Bach, J. Mairal, and J. Ponce. Convex sparse matrix factorizations. Technical report, Preprint arXiv:0812.1869, 2008.

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.

D. M. Bradley and J. A. Bagnell. Convex coding. In *Proc. UAI*, 2009.

V. V. Buldygin and I. U. V. Kozachenko. *Metric characterization of random variables and random processes*, volume 188. American Mathematical Society, 2000.

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

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.

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

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2002.

V. De la Peña and E. Giné. *Decoupling: from dependence to independence*. Springer Verlag, 1999.

D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE T. Inform. Theory*, 47(7):2845–2862, 2001.

H. Dym. *Linear algebra in action*. 2007.

J. J. Fuchs. Recovery of exact sparse representations in the presence of bounded noise. *IEEE T. Inform. Theory*, 51(10):3601–3608, 2005.

W. Gautschi. The incomplete Gamma functions since Tricomi. In *In Tricomi's Ideas and Contemporary Applied Mathematics, Atti dei Convegni Lincei, n.147, Accademia Nazionale dei Lincei*, 1998.

Q. Geng, H. Wang, and J. Wright. On the Local Correctness of L1 Minimization for Dictionary Learning. Technical report, Preprint arXiv:1101.5672, 2011.

R. Gribonval and K. Schnass. Dictionary identification—sparse matrix-factorization via ℓ_1-minimization. *IEEE T. Inform. Theory*, 56(7):3523–3539, 2010.

R. Gribonval, V. Cevher, and M. E. Davies. Compressible distributions for high-dimensional statistics. Technical report, preprint arXiv:1102.1249, 2011.

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

D. Hsu, S. M. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. Technical report, Preprint arXiv:1110.2842, 2011.

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

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

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

H. W. Kuhn. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97, 1955.

J. R. Magnus and H. Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons, 1988.

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.

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

A. Maurer and M. Pontil. k-dimensional coding schemes in hilbert spaces. *IEEE T. Inform. Theory*, 56(11): 5839–5846, 2010.

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

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.

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

J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE T. Inform. Theory*, 50(10): 2231–2242, 2004.

D. Vainsencher, S. Mannor, and A. M. Bruckstein. The sample complexity of dictionary learning. Technical report, Preprint arXiv:1011.5395, 2010.

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.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. Technical report, Preprint arXiv:1011.3027, 2010.

M. J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1- constrained quadratic programming. *IEEE T. Inform. Theory*, 55:2183–2202, 2009.

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

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

M. Zhou, H. Chen, J. Paisley, L. Ren, G. Sapiro, and L. Carin. Non-parametric Bayesian dictionary learning for sparse image representations. In *Adv. NIPS*, 2009.

A Detailed Statements of the Main results

We gather in this appendix the detailed statements and the proofs of the simplified results presented in the core of the paper. In particular, we show in this section that under appropriate scalings of the problem dimensions (m, p), number of training samples n, and model parameters $k, \underline{\alpha}, \sigma_\alpha, \sigma, \mu_0$, it is possible to prove that, with high probability, the problem of sparse coding admits a local minimum in a certain neighborhood of $\mathbf{D}_0$ of controlled size, for appropriate choices of the regularization parameter λ.

A.1 Minimum local of sparse coding

We present here a complete and detailed version of our result upon which the theorems presented in the paper are built.

Theorem 3 (Local minimum of sparse coding)**.** *Let us consider our generative model of signals for some reference dictionary $\mathbf{D}_0 \in \mathbb{R}^{m\times p}$ with coherence μ_0. Introduce the parameters $q_\alpha \triangleq \frac{\mathbb{E}[\alpha^2]}{\sigma_\alpha^2}$ and $\mathcal{Q}_\alpha \triangleq \frac{\mathbb{E}[\alpha^2]}{\sigma_\alpha \cdot \mathbb{E}[|\alpha|]}$ which depend on the distribution of $\boldsymbol{\alpha}_0$ only. Consider the following quantities:*

$$\begin{aligned}\tau = \tau(\mathbf{D}_0, \boldsymbol{\alpha}_0) &\triangleq \min\left\{\frac{\underline{\alpha}}{\sigma_\alpha}, \frac{1}{3c_0}\cdot\frac{\mathcal{Q}_\alpha}{k\|\mathbf{D}_0\|_2}\right\}\\ \gamma = \gamma(n, \mathbf{D}_0, \boldsymbol{\alpha}_0) &\triangleq \frac{1}{2}\min\left\{\sqrt{2\log(n)}, \frac{1}{2\sqrt{2}c_0c_\gamma}\cdot\frac{\mathcal{Q}_\alpha}{\|\mathbf{D}_0\|_2\cdot k\mu_0}\right\},\end{aligned}$$

and let us define the radius $t \in \mathbb{R}_+$ by

$$t \triangleq \max\left\{\frac{4\sqrt{2}c_\gamma}{q_\alpha}\cdot p\cdot\left\{c_1\gamma^3 e^{-\gamma^2} + 2c_2\cdot\gamma\cdot\left[mp\frac{\log(n)}{n}\right]^{1/2}\right\}, \frac{\sigma}{\sigma_\alpha}\cdot\sqrt{m}\right\}$$

for some universal constants c_*. Provided the following conditions are satisfied:

(Coherence) $$\|\mathbf{D}_0\|_2\cdot k\mu_0 \leq \frac{1}{4\sqrt{2}c_0c_\gamma}\cdot\frac{\mathcal{Q}_\alpha}{\sqrt{\log(69c_1c_\gamma^2\cdot\frac{1}{q_\alpha}\cdot\frac{p}{\tau})}},$$

(Sample complexity) $$\frac{\log(n)}{n} \leq \frac{q_\alpha^2}{c_3}\cdot\frac{1}{m\cdot p^3}\cdot\frac{\tau^2}{\gamma^4},$$

one can find a regularization parameter λ *proportional to* $\gamma \cdot \sigma_\alpha \cdot t$*, and with probability exceeding*

$$1 - \Big(\frac{mpn}{9}\Big)^{-\frac{mp}{2}} - e^{-4\sqrt{n}},$$

there exists a local minimum in $\Big\{\mathbf{D} \in \mathcal{D};\ \|\mathbf{D}_0 - \mathbf{D}\|_{\mathrm{F}} < t\Big\}$.

As it will discussed at greater length in Sec. B, we can see that the probability of success of Theorem 3 can be decomposed into the contributions of the concentration of the surrogate function and the residual term. We next present a second result which assumes a more constrained signal model:

Theorem 4 (Local minimum of sparse coding with noiseless/bounded signals)**.** *Let us consider our generative model of signals for some reference dictionary* $\mathbf{D}_0 \in \mathbb{R}^{m\times p}$ *with coherence* μ_0*. Further assume that* $\boldsymbol{\alpha}_0$ *is almost surely upper bounded by* $\overline{\alpha}$ *and that there is no noise, that is,* $\sigma = 0$*. Introduce the parameters* $q_\alpha \triangleq \frac{\mathbb{E}[\alpha^2]}{\overline{\alpha}\cdot\mathbb{E}[|\alpha|]}$ *and* $\mathcal{Q}_\alpha \triangleq \frac{\mathbb{E}[\alpha^2]}{\sigma_\alpha\cdot\mathbb{E}[|\alpha|]}$ *which depend on the distribution of* $\boldsymbol{\alpha}_0$ *only. Consider the radius* $t \in \mathbb{R}_+$*:*

$$t \triangleq \frac{8c_1c_\lambda}{q_\alpha}\left[kmp^3 \cdot \frac{\log(n)}{n}\right]^{1/2}$$

for some universal constants c_**. Provided the following conditions are satisfied:*

(Coherence) $\qquad \|\mathbf{D}_0\|_2 \cdot k^{3/2}\mu_0 \leq \dfrac{1}{c_0c_\lambda}\cdot q_\alpha$

(Sample complexity) $\dfrac{\log(n)}{n} \leq \dfrac{1}{k^2mp^3}\cdot\left[\dfrac{\mathbb{E}[\alpha^2]}{\overline{\alpha}^2}\cdot\dfrac{1}{9c_1c_\lambda^2}\cdot\min\left\{\dfrac{\underline{\alpha}}{\sigma_\alpha}, \dfrac{1}{5c_0}\cdot\dfrac{\mathcal{Q}_\alpha}{k\cdot\|\mathbf{D}_0\|_2}\right\}\right]^2,$

one can find a regularization parameter λ *proportional to* $\sqrt{k}\cdot\overline{\alpha}\cdot t$*, and with probability exceeding*

$$1 - \Big(\frac{mpn}{9}\Big)^{-mp/2},$$

there exists a local minimum in $\Big\{\mathbf{D} \in \mathcal{D};\ \|\mathbf{D}_0 - \mathbf{D}\|_{\mathrm{F}} < t\Big\}$.

These two theorems, which are proved in Section B, heavily relies on the following central result.

A.2 The backbone of the analysis

We concentrate on the result which constitutes the backbone of our analysis. Indeed, we next show how the difference

$$\Delta F_n(\mathbf{W}, \mathbf{v}, t) \triangleq F_n(\mathbf{D}(\mathbf{W}, \mathbf{v}, t)) - F_n(\mathbf{D}_0). \tag{8}$$

is lower bounded with high probability and uniformly with respect to all possible choices of the parameters $(\mathbf{W}, \mathbf{v})$. The theorem and corollaries displayed in the core of the paper are consequences of this general theorem, discussing under which conditions/scalings this lower bound can be proved to be sufficient (i.e., strictly positive) to exploit Proposition 1 and conclude to the existence of a local minimum for t appropriately chosen. We define

$$Q_t \triangleq \frac{1}{\sqrt{1-k\mu(t)}} \quad \text{and} \quad C_t \triangleq \frac{1}{\sqrt{1-\delta_k(\mathbf{D}_0)} - t}, \tag{9}$$

where the quantity $\delta_k(\mathbf{D}_0)$ is the RIP constant itself defined in Section C.

Theorem 5. *Let* $\underline{\alpha}, \sigma_\alpha$ *be the parameters of the coefficient model. Consider* $\mathbf{D}_0$ *a dictionary in* $\mathbb{R}^{m\times p}$ *with* $\mu_0 < 1/2$ *and let* k*,* $t > 0$ *be such that*

$$k\mu(t) \quad < \quad 1/2 \tag{10}$$

$$\frac{3t}{2-Q_t^2} \quad < \quad \frac{4\underline{\alpha}}{9\sigma_\alpha} \tag{11}$$

Then for small enough noise levels σ one can find a regularization parameter $\lambda > 0$ such that

$$\frac{3}{2-Q_t^2}\cdot\sqrt{t^2\sigma_\alpha^2+m\sigma^2}\leq\lambda\leq\frac{4}{9}\underline{\alpha}. \tag{12}$$

Given σ and λ satisfying (12), we define

$$\gamma\triangleq\frac{\lambda(2-Q_t^2)}{\sqrt{5}\cdot\sqrt{t^2\sigma_\alpha^2+m\sigma^2}}\geq\sqrt{2\log 2}. \tag{13}$$

Let $\mathbf{x}^i\in\mathbb{R}^m$, $i\in[\![1;n]\!]$, where $n/\log n\geq mp$, be generated according to the signal model. Then, except with probability at most $\left(\frac{mpn}{9}\right)^{-mp/2}+\exp(-4n\cdot e^{-\gamma^2})$ we have

$$\begin{aligned}\inf_{\mathbf{W}\in\mathcal{W}_{\mathbf{D}_0},\mathbf{v}\in\mathcal{S}^p}\Delta F_n(\mathbf{W},\mathbf{v},t) \quad\geq\quad & (1-\mathcal{K}^2)\cdot\frac{\mathbb{E}[\alpha_0^2]}{2}\cdot\frac{k}{p}\cdot t^2\\ & -Q_t^2\left(\frac{16}{9}Q_t^2+3\right)\cdot\mathbb{E}\{|\alpha_0|\}\cdot t\cdot\frac{k}{p}\cdot\|\mathbf{D}_0\|_2\cdot k\mu(t)\cdot\lambda\\ & -A\cdot\gamma^2\cdot e^{-\gamma^2}\\ & -B\cdot\sqrt{mp\frac{\log n}{n}},\end{aligned} \tag{14}$$

where

$$\mathcal{K}\triangleq C_t\cdot\left(\|\mathbf{D}_0\|_2\cdot\sqrt{k/p}+t\right) \tag{15}$$

$$A\triangleq 367\cdot\left(t^2\sigma_\alpha^2+2m\sigma^2+2\lambda k\sigma_\alpha\right) \tag{16}$$

$$B\triangleq 3045\left(k\sigma_\alpha^2\cdot t+2m\sigma^2+2\lambda k\sigma_\alpha\right). \tag{17}$$

Roughly speaking, the lower bound we obtain can be decomposed into three terms: (1) the expected value of our surrogate function valid uniformly for all parameters $(\mathbf{v},\mathbf{W})$, (2) the contributions of the residual term (discussed in the next section) introducing the quantity γ, and (3) the probabilisitc concentrations over the n signals of the surrogate function and the residual term.

The proof of the theorem and its main building blocks are detailed in the next section.

B Architecture of the proof of Theorem 5

Since we expect to have for many training samples the equality $f_{\mathbf{x}}(\mathbf{D}(\mathbf{W},\mathbf{v},t))=\phi_{\mathbf{x}}(\mathbf{D}(\mathbf{W},\mathbf{v},t)|\mathrm{sign}(\boldsymbol{\alpha}_0))$ uniformly for all $(\mathbf{W},\mathbf{v})$, the main idea is to first concentrate on the study of the smooth function

$$\Delta\Phi_n(\mathbf{W},\mathbf{v},t)\triangleq\Phi_n(\mathbf{D}(\mathbf{W},\mathbf{v},t)|\mathrm{sign}(\mathbf{A}_0))-\Phi_n(\mathbf{D}_0|\mathrm{sign}(\mathbf{A}_0)), \tag{18}$$

instead of the original function $\Delta F_n(\mathbf{W},\mathbf{v},t)$.

B.1 Control of $\Delta\Phi_n$

The first step consists in uniformly lower bounding $\Delta\Phi_n$ with high probability.

Proposition 2. *Assume that $k\mu(t)\leq 1/2$ then for any n such that*

$$\frac{n}{\log n}\geq mp, \tag{19}$$

except with probability at most $\left(\frac{mpn}{9}\right)^{-mp/2}$*, we have*

$$\begin{aligned}\inf_{\mathbf{W}\in\mathcal{W}_{\mathbf{D}_0},\mathbf{v}\in\mathcal{S}^p}\Delta\Phi_n(\mathbf{W},\mathbf{v},t) \quad &\geq \quad (1-\mathcal{K}^2)\cdot\frac{\mathbb{E}[\alpha_0^2]}{2}\cdot\frac{k}{p}\cdot t^2 \\ &\quad -Q_t^2\cdot t\cdot\frac{k}{p}\cdot\left\|\mathbf{D}_0\right\|_2\cdot k\mu(t)\cdot\lambda\cdot\left(4Q_t^2\lambda+3\mathbb{E}\{|\alpha_0|\}\right) \\ &\quad -B\cdot\sqrt{mp\frac{\log n}{n}}, \end{aligned} \tag{20}$$

where

$$\begin{aligned}\mathcal{K} \quad &\triangleq \quad C_t\cdot\left(\left\|\mathbf{D}_0\right\|_2\cdot\sqrt{k/p}+t\right) \\ B \quad &\triangleq \quad 3045\left(k\sigma_\alpha^2\cdot t+2m\sigma^2+\lambda k\sigma_\alpha+\lambda^2 k\cdot t\right).\end{aligned}$$

The proof of this proposition is given in Section E.

B.2 Control of ΔF_n in terms of $\Delta\Phi_n$

The second step consists in lower bounding ΔF_n in terms of $\Delta\Phi_n$ uniformly for all $(\mathbf{W},\mathbf{v})\in\mathcal{W}_{\mathbf{D}_0}\times\mathcal{S}^p$. For a given $t\geq 0$, consider the independent events $\{\mathcal{E}^i_{\text{coincide}}(t)\}_{i\in[\![1;n]\!]}$ defined by

$$\mathcal{E}^i_{\text{coincide}}(t)\triangleq\left\{\omega\;\middle|\; f_{\mathbf{x}^i(\omega)}(\mathbf{D}(\mathbf{W},\mathbf{v},t))=\phi_{\mathbf{x}^i(\omega)}(\mathbf{D}(\mathbf{W},\mathbf{v},t)|\mathbf{s}_0),\quad\forall(\mathbf{W},\mathbf{v})\in\mathcal{W}_{\mathbf{D}_0}\times\mathcal{S}^p\right\},$$

with $\mathbf{s}_0=\text{sign}(\boldsymbol{\alpha}_0)$. In words, the event $\mathcal{E}^i_{\text{coincide}}(t)$ corresponds to the fact that target function $f_{\mathbf{x}^i(\omega)}(\mathbf{D}(\cdot,\cdot,t))$ coincides with the idealized one $\phi_{\mathbf{x}^i(\omega)}(\mathbf{D}(\cdot,\cdot,t)|\mathbf{s}_0)$ for the "radius" t.

Importantly, the event $\mathcal{E}^i_{\text{coincide}}(t)$ only involves a *single* signal; when we consider a collection of n independent signals, we should instead study the event $\bigcap_{i=1}^n\mathcal{E}^i_{\text{coincide}}(t)$ to guarantee that Φ_n and F_n (and therefore, $\Delta\Phi_n$ and ΔF_n) do coincide. However, as the number of observations n becomes large, it is unrealistic and not possible to ensure exact recovery both *simultaneously* for the n signals and *with high probability*.

To get around this issue, we will relax our expectations and seek to prove that ΔF_n is well approximated by $\Delta\Phi_n$ (rather than equal to it) uniformly for all $(\mathbf{W},\mathbf{v})$. This will be achieved by showing that, when $f_{\mathbf{x}^i}(\mathbf{D}(t))$ and $\phi_{\mathbf{x}^i}(\mathbf{D}(t)|\mathbf{s}_0)$ *do not* coincide, their difference can be bounded. For any $\mathbf{D}\in\mathbb{R}^{m\times p}$, we have by the very definition (1), $0\leq f_{\mathbf{x}}(\mathbf{D})\leq\mathcal{L}_{\mathbf{x}}(\mathbf{D},\boldsymbol{\alpha}_0)$. We have as well by the definition (2):

$$0\leq\phi_{\mathbf{x}}(\mathbf{D}|\text{sign}(\boldsymbol{\alpha}_0))\leq\min_{\boldsymbol{\alpha}\in\mathbb{R}^p,\ \text{sign}(\boldsymbol{\alpha})=\text{sign}(\boldsymbol{\alpha}_0)}\frac{1}{2}\|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}\|_2^2+\lambda\cdot\text{sign}(\boldsymbol{\alpha}_0)^\top\boldsymbol{\alpha}\leq\mathcal{L}_{\mathbf{x}}(\mathbf{D},\boldsymbol{\alpha}_0).$$

It follows that for all $(\mathbf{W},\mathbf{v})\in\mathcal{W}_{\mathbf{D}_0}\times\mathcal{S}^p$ we have, with $\mathbf{D}=\mathbf{D}(\mathbf{W},\mathbf{v},t)$,

$$\phi_{\mathbf{x}}(\mathbf{D}_0|\mathbf{s}_0)-\phi_{\mathbf{x}}(\mathbf{D}|\mathbf{s}_0)+f_{\mathbf{x}}(\mathbf{D})-f_{\mathbf{x}}(\mathbf{D}_0)\quad\geq\quad-\phi_{\mathbf{x}}(\mathbf{D}|\mathbf{s}_0)-f_{\mathbf{x}}(\mathbf{D}_0)\geq-\left\{\mathcal{L}_{\mathbf{x}}(\mathbf{D},\boldsymbol{\alpha}_0)+\mathcal{L}_{\mathbf{x}}(\mathbf{D}_0,\boldsymbol{\alpha}_0)\right\}.$$

When both functions coincide uniformly at radius t (the event $\mathcal{E}_{\text{coincide}}(t)$ holds) *and* at radius zero ($\phi_{\mathbf{x}}(\mathbf{D}_0|\mathbf{s}_0)=f_{\mathbf{x}}(\mathbf{D}_0)$, i.e., the event $\mathcal{E}_{\text{coincide}}(0)$ holds), the left hand side is indeed zero. As a result we have, uniformly for all $(\mathbf{W},\mathbf{v})\in\mathcal{W}_{\mathbf{D}_0}\times\mathcal{S}^p$:

$$\begin{aligned}f_{\mathbf{x}^i}(\mathbf{D})-f_{\mathbf{x}^i}(\mathbf{D}_0) \quad &\geq \quad \phi_{\mathbf{x}}(\mathbf{D}|\mathbf{s}_0)-\phi_{\mathbf{x}}(\mathbf{D}_0|\mathbf{s}_0)-r_{\mathbf{x}^i}, \\ \text{with } r_{\mathbf{x}^i} \quad &\triangleq \quad \mathbb{1}_{\left[\mathcal{E}^i_{\text{coincide}}(t)\cap\mathcal{E}^i_{\text{coincide}}(0)\right]^c}(\omega)\cdot\left\{\mathcal{L}_{\mathbf{x}^i}(\mathbf{D},\boldsymbol{\alpha}_0^i)+\mathcal{L}_{\mathbf{x}^i}(\mathbf{D}_0,\boldsymbol{\alpha}_0^i)\right\}.\end{aligned}$$

Averaging the above inequality over a set of n signals, we obtain a similar uniform lower bound for ΔF_n:

$$\Delta F_n(\mathbf{W},\mathbf{v},t)\geq\Delta\Phi_n(\mathbf{W},\mathbf{v},t)-r_n. \tag{21}$$

where we detail the definition

$$r_n(\omega) \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[\mathcal{E}^i_{\text{coincide}}(t)\cap\mathcal{E}^i_{\text{coincide}}(0)]^c}(\omega)\cdot\left\{\mathcal{L}_{\mathbf{x}^i}(\mathbf{D},\boldsymbol{\alpha}^i_0)+\mathcal{L}_{\mathbf{x}^i}(\mathbf{D}_0,\boldsymbol{\alpha}^i_0)\right\}. \tag{22}$$

Using Lemma 23 and Corollary 4 in the Appendix, one can show that with high probability:

$$r_n \leq 25\left(t^2\cdot\sigma_\alpha^2+2m\cdot\sigma^2+2\lambda k\sigma_\alpha\right)(1+\log 2)\cdot(3-\log\kappa)\kappa$$

with $\kappa \triangleq \max_{i\in[\![1;n]\!]}\Pr\left(\left[\mathcal{E}^i_{\text{coincide}}(t)\cap\mathcal{E}^i_{\text{coincide}}(0)\right]^c\right)$. To bound the size of the residual r_n, we now control κ.

B.2.1 Control of κ: exact sign recovery for *perturbed* dictionaries

The objective of this section is to determine sufficient conditions under which $\phi_{\mathbf{x}}(\mathbf{D}(\mathbf{W},\mathbf{v},t)|\text{sign}(\boldsymbol{\alpha}_0))$ and $f_{\mathbf{x}}(\mathbf{D}(\mathbf{W},\mathbf{v},t))$ coincide for all $(\mathbf{W},\mathbf{v})$, and control the probability of this event. We make this statement precise in the following proposition, proved in Appendix F.

Proposition 3 (Exact recovery for perturbed dictionaries and one training sample)**.** *Condider $\mathbf{D}_0$ a dictionary in $\mathbb{R}^{m\times p}$ and let k,t such that $k\mu(t)<1/2$. Let $\underline{\alpha},\sigma_\alpha,\sigma$ be the remaining parameters of our signal model, and let $\mathbf{x}\in\mathbb{R}^m$ be generated according to this model. Assume that the regularization parameter λ satisfies*

$$0<\lambda\leq\frac{4}{9}\underline{\alpha}.$$

Consider $0\leq t'\leq t$. Except with probability at most

$$\Pr(\mathcal{E}^c_{\text{coincide}}(t'))\leq 2\cdot\exp\left(-\frac{\lambda^2(2-Q_t^2)^2}{5(t'^2\cdot\sigma_\alpha^2+m\sigma^2)}\right)$$

we have, uniformly for all $(\mathbf{W},\mathbf{v})\in\mathcal{W}_{\mathbf{D}_0}\times\mathcal{S}^p$, the vector $\hat{\boldsymbol{\alpha}}(t')\in\mathbb{R}^p$ defined by

$$\hat{\boldsymbol{\alpha}}(t')=\begin{pmatrix}\left[[\mathbf{D}(t')]_{\mathrm{J}}^\top[\mathbf{D}(t')]_{\mathrm{J}}\right]^{-1}\left[[\mathbf{D}(t')]_{\mathrm{J}}^\top\mathbf{x}-\lambda\text{sign}([\boldsymbol{\alpha}_0]_{\mathrm{J}})\right]\\ \mathbf{0}\end{pmatrix},$$

is the unique solution of $\min_{\boldsymbol{\alpha}\in\mathbb{R}^p}[\frac{1}{2}\|\mathbf{x}-\mathbf{D}(t')\boldsymbol{\alpha}\|_2^2+\lambda\|\boldsymbol{\alpha}\|_1]$, and $\text{sign}(\hat{\boldsymbol{\alpha}}(t'))=\text{sign}(\boldsymbol{\alpha}_0)$.

We also need a modified version of this proposition to handle a simplified, noiseless setting where the coefficients $\boldsymbol{\alpha}_0$ are almost surely upper bounded. Its proof can be found in Section F as well.

Proposition 4 (Exact recovery for perturbed dictionaries and one training sample; noiseless and bounded $\boldsymbol{\alpha}_0$)**.** *Condider $\mathbf{D}_0$ a dictionary in $\mathbb{R}^{m\times p}$ and let k,t such that $k\mu(t)<1/2$. Consider our signal model with the following additional assumptions:*

$$\begin{aligned}\sigma&=0 && (\textbf{\textit{no noise}})\\ \Pr(|[\boldsymbol{\alpha}_0]_j|>\overline{\alpha}|j\in J)&=0,\quad \textit{for some } \overline{\alpha}\geq\underline{\alpha}>0 && (\textbf{\textit{signal boundedness}}).\end{aligned}$$

Let $\mathbf{x}\in\mathbb{R}^m$ be generated according to this model. Assume that the regularization parameter λ satisfies

$$\frac{\sqrt{k}\overline{\alpha}}{2-Q_t^2}t<\lambda\leq\frac{4}{9}\underline{\alpha}.$$

Consider $0\leq t'\leq t$. Almost surely, we have, uniformly for all $(\mathbf{W},\mathbf{v})\in\mathcal{W}_{\mathbf{D}_0}\times\mathcal{S}^p$, the vector $\hat{\boldsymbol{\alpha}}(t')\in\mathbb{R}^p$ defined by

$$\hat{\boldsymbol{\alpha}}(t')=\begin{pmatrix}\left[[\mathbf{D}(t')]_{\mathrm{J}}^\top[\mathbf{D}(t')]_{\mathrm{J}}\right]^{-1}\left[[\mathbf{D}(t')]_{\mathrm{J}}^\top\mathbf{x}-\lambda\text{sign}([\boldsymbol{\alpha}_0]_{\mathrm{J}})\right]\\ \mathbf{0}\end{pmatrix},$$

is the unique solution of $\min_{\boldsymbol{\alpha}\in\mathbb{R}^p}[\frac{1}{2}\|\mathbf{x}-\mathbf{D}(t')\boldsymbol{\alpha}\|_2^2+\lambda\|\boldsymbol{\alpha}\|_1]$, and $\text{sign}(\hat{\boldsymbol{\alpha}}(t'))=\text{sign}(\boldsymbol{\alpha}_0)$. In other words, it holds that $\Pr(\mathcal{E}^c_{\text{coincide}}(t'))=0$.

B.2.2 Control of the residual

The last step of the proof of Theorem 5 consists in controlling the residual term (22). Its proof can be found in Section H.

Proposition 5. *Let $\underline{\alpha}, \sigma_\alpha$ be the parameters of the coefficient model. Consider $\mathbf{D}_0$ a dictionary in $\mathbb{R}^{m\times p}$ with $\mu_0 < 1/2$ and let k, t be such that*

$$k\mu(t) < 1/2 \tag{23}$$

$$\frac{3t}{2-Q_t^2} < \frac{4\underline{\alpha}}{9\sigma_\alpha} \tag{24}$$

Then for small enough noise levels σ one can find a regularization parameter $\lambda > 0$ such that

$$\frac{3}{2-Q_t^2}\cdot\sqrt{t^2\sigma_\alpha^2 + m\sigma^2} \leq \lambda \leq \frac{4}{9}\underline{\alpha}. \tag{25}$$

Given σ and λ satisfying (25), we define

$$\gamma \triangleq \frac{\lambda(2-Q_t^2)}{\sqrt{5}\cdot\sqrt{t^2\sigma_\alpha^2 + m\sigma^2}} \geq \sqrt{2\log 2}. \tag{26}$$

Let $\mathbf{x}^i \in \mathbb{R}^m$, $i \in [\![1; n]\!]$ be generated according to our noisy signal model. Then,

$$r_n \leq \left(t^2\sigma_\alpha^2 + 2m\sigma^2 + 2\lambda k\sigma_\alpha\right)\cdot 367\cdot\gamma^2\cdot e^{-\gamma^2}. \tag{27}$$

except with probability at most $\exp(-4n\cdot e^{-\gamma^2})$.

We have stated the main results and showed how they are structured in key propositions, which we now prove.

A Proof of Proposition 1

The topology we consider on $\mathcal{D}$ is the one induced by its natural embedding in $\mathbb{R}^{m\times p}$: the open sets are the intersection of open sets of $\mathbb{R}^{m\times p}$ with $\mathcal{D}$. Recall that all norms are equivalent on $\mathbb{R}^{m\times p}$ and induce the same topology. For convenience we will consider the balls associated to the Froebenius norm. To prove the existence of a local minimum for F_n, say at $\mathbf{D}^\star$, we will show the existence of a ball centered at $\mathbf{D}^\star$, $\mathcal{B}_h \triangleq \left\{\mathbf{D}\in\mathcal{D};\ \|\mathbf{D}^\star - \mathbf{D}\|_{\text{F}} \leq h\right\}$ such that for any $\mathbf{D}\in\mathcal{B}_h$, we have $F_n(\mathbf{D}^\star) \leq F_n(\mathbf{D})$.

First step: We recall the notation $\mathcal{S}_+^p \triangleq \mathcal{S}^p \cap \mathbb{R}_+^p$ for the sphere intersected with the positive orthant. Moreover, we introduce

$$\mathcal{Z}_t \triangleq \left\{\mathbf{D}(\mathbf{W},\mathbf{v},t');\ \mathbf{W}\in\mathcal{W}_{\mathbf{D}_0}, \mathbf{v}\in\mathcal{S}_+^p, t'\in[0,t],\ \text{and } t'\|\mathbf{v}\|_\infty \leq \pi\right\}.$$

The set $\mathcal{Z}_t$ is compact as the image of a compact set by the continuous function $(\mathbf{W},\mathbf{v},t')\mapsto\mathbf{D}(\mathbf{W},\mathbf{v},t')$. As a result, the continuous function F_n admits a global minimum in $\mathcal{Z}_t$ which we denote by $\mathbf{D}^\star = \mathbf{D}(\mathbf{W}^\star,\mathbf{v}^\star,t^\star)$. Moreover, and according to the assumption of Proposition 1, we have $t^\star < t$.

Second step: We will now prove the existence of $h > 0$ such that $\mathcal{B}_h \subseteq \mathcal{Z}_t$. This will imply that $F_n(\mathbf{D}^\star) \leq F_n(\mathbf{D})$ for $\mathbf{D}\in\mathcal{B}_h$, hence that $\mathbf{D}^\star$ is a local mimimum. First, we formalize the following lemma.

Lemma 1. *Given any matrix $\mathbf{D}_1 \in \mathcal{D}$, any matrix $\mathbf{D}_2 \in \mathcal{D}$ can be described as $\mathbf{D}_2 = \mathbf{D}(\mathbf{D}_1, \mathbf{W}, \mathbf{v}, \tau)$, with $\mathbf{W} \in \mathcal{W}_{\mathbf{D}_1}$, $\mathbf{v} \in \mathcal{S}_+^p$ and $\tau \geq 0$ such that $\tau\|\mathbf{v}\|_\infty \leq \pi$. Moreover, we have*

$$\begin{aligned} \frac{2}{\pi}\tau\mathbf{v}_j &\leq \|\mathbf{d}_2^j - \mathbf{d}_1^j\|_2 = 2\sin\left(\frac{\tau\mathbf{v}_j}{2}\right) \leq \tau\mathbf{v}_j, \quad \forall j, & (28)\\ \frac{2}{\pi}\tau &\leq \|\mathbf{D}_2 - \mathbf{D}_1\|_{\text{F}} \leq \tau. & (29) \end{aligned}$$

Vice-versa, $\mathbf{D}_1 = \mathbf{D}(\mathbf{D}_2, \mathbf{W}', \mathbf{v}', \tau')$ for some $\mathbf{W}' \in \mathcal{W}_{\mathbf{D}_2}$, and with the same $\mathbf{v}' = \mathbf{v} \in \mathcal{S}_+^p$, $\tau' = \tau \geq 0$.

Proof. The result is trivial if $\mathbf{D}_2 = \mathbf{D}_1$, hence we focus on the case $\mathbf{D}_2 \neq \mathbf{D}_1$. Each column $\mathbf{d}_2^j$ of $\mathbf{D}_2$ can be uniquely expressed as

$$\mathbf{d}_2^j = \mathbf{u} + \mathbf{z}, \text{ with } \mathbf{u} \in \text{span}(\mathbf{d}_1^j) \text{ and } \mathbf{u}^\top\mathbf{z} = 0.$$

Since $\|\mathbf{d}_2^j\|_2 = 1$, the previous relation can be rewritten as

$$\mathbf{d}_2^j = \cos(\theta_j)\mathbf{d}_1^j + \sin(\theta_j)\mathbf{w}^j,$$

for some $\theta_j \in [0, \pi]$ and some unit vector $\mathbf{w}^j$ orthogonal to $\mathbf{d}_1^j$ (except for the case $\theta_j = 0$, the vector $\mathbf{w}^j$ is unique). The sign indetermination in $\mathbf{w}^j$ is handled thanks to the convention $\sin(\theta_j) \geq 0$. One can define a matrix $\mathbf{W} \in \mathcal{W}_{\mathbf{D}_1}$ which j-th column is $\mathbf{w}^j$. Denote $\boldsymbol{\theta} \triangleq (\theta_1, \ldots, \theta_p)$ and $\tau \triangleq \|\boldsymbol{\theta}\|_2$. Since $\mathbf{D}_2 \neq \mathbf{D}_1$ we have $\tau > 0$ and we can define $\mathbf{v} \in \mathcal{S}_+^p$ with coordinates

$$\mathbf{v}_j \triangleq \frac{\theta_j}{\tau}.$$

Next we notice that $\tau\|\mathbf{v}\|_\infty = \|\boldsymbol{\theta}\|_\infty \leq \pi$ and

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

We conclude using the inequalities $\frac{2}{\pi} \leq \frac{\sin u}{u} \leq 1$ for $0 \leq u \leq \pi/2$ and the fact that $\|\mathbf{v}\|_2 = 1$. The reciprocal $\mathbf{D}_1 = \mathbf{D}(\mathbf{D}_2, \mathbf{W}', \mathbf{v}', \tau')$ is obvious, and the fact that $\mathbf{v}' = \mathbf{v}$, $\tau' = \tau$ follows from the equality $\|\mathbf{d}_1^j - \mathbf{d}_2^j\|_2 = 2\sin\mathbf{v}_j\tau = 2\sin\mathbf{v}'_j\tau'$ for all j. □

Using the parameterization built in Lemma 1 for $\mathbf{D} \in \mathcal{B}_h$, there remains to prove that $\mathbf{D} = \mathbf{D}(\mathbf{D}_0, \mathbf{W}, \mathbf{v}, \tau)$ belongs to $\mathcal{Z}_t$ provided that h is small enough. For that, we need to show that $\tau < t$ (we will need of course to assume that h is small enough). To this end, notice that

$$\begin{aligned} \|\mathbf{D}^\star - \mathbf{D}\|_{\text{F}}^2 &= \sum_{j=1}^p \|(\cos(\mathbf{v}_j^\star t^\star) - \cos(\mathbf{v}_j\tau))\mathbf{d}_0^j + \sin(\mathbf{v}_j^\star t^\star)\mathbf{w}^{\star,j} - \sin(\mathbf{v}_j\tau)\mathbf{w}^j\|_2^2\\ &= 2\sum_{j=1}^p (1 - \cos(\mathbf{v}_j^\star t^\star)\cos(\mathbf{v}_j\tau) - \sin(\mathbf{v}_j^\star t^\star)\sin(\mathbf{v}_j\tau)[\mathbf{w}^j]^\top\mathbf{w}^{\star,j}) \end{aligned}$$

where the simplifications in the second equality come from the fact that both $\mathbf{W}$ and $\mathbf{W}^\star$ have their columns normalized and orthogonal to the corresponding columns of $\mathbf{D}_0$. Since $t^\star\|\mathbf{v}^\star\|_\infty \leq \pi$ and $\tau\|\mathbf{v}\|_\infty \leq \pi$, the product of sine terms is positive, so that with $|[\mathbf{w}^j]^\top\mathbf{w}^{\star,j}| \leq 1$, we obtain

$$\|\mathbf{D}^\star - \mathbf{D}\|_{\text{F}}^2 \geq 2\sum_{j=1}^p (1 - \cos(\mathbf{v}_j^\star t^\star)\cos(\mathbf{v}_j\tau) - \sin(\mathbf{v}_j^\star t^\star)\sin(\mathbf{v}_j\tau)) = 2\sum_{j=1}^p (1 - \cos(\Delta_j)) = 4\sum_{j=1}^p \sin^2(\Delta_j/2)$$

where $\Delta_j \triangleq \mathbf{v}_j^\star t^\star - \mathbf{v}_j \tau$. Now, since $0 \leq t^\star \mathbf{v}_j^\star, \tau \mathbf{v}_j \leq \pi$, we have $\Delta_j/2 \in [-\pi/2\ ,\ \pi/2]$, hence using that $\sin^2(u) \geq \frac{4}{\pi^2}u^2$ for $|u| \leq \frac{\pi}{2}$, we finally have

$$h^2 \geq \|\mathbf{D}^\star - \mathbf{D}\|_{\mathrm{F}}^2 \quad \geq \quad \frac{4}{\pi^2}\sum_{j=1}^p \Delta_j^2 = \frac{4}{\pi^2}\sum_{j=1}^p ([\mathbf{v}_j^\star t^\star]^2 + [\mathbf{v}_j \tau]^2 - 2t^\star \tau \mathbf{v}_j^\star \mathbf{v}_j) \geq \frac{4}{\pi^2}(t^\star - \tau)^2,$$

where we have exploited that both $\mathbf{v}^\star$ and $\mathbf{v}$ are normalized. As a consequence, we have $\tau \leq t^\star + \frac{\pi}{2}h$ hence for $h < \frac{2}{\pi}(t - t^\star)$ we guarantee $\tau < t$, so that $\mathbf{D} \in \mathcal{Z}_t$. We conclude that $\mathcal{B}_h \subseteq \mathcal{Z}_t$ for $h < \frac{2}{\pi}(t - t^\star)$.

Third and last step: To recapitulate, we have shown that there exists a ball $\mathcal{B}_h$ in $\mathcal{D}$, such that $\mathcal{B}_h \subseteq \mathcal{Z}_t$ and for any $\mathbf{D} \in \mathcal{B}_h$, we have

$$F_n(\mathbf{D}) \geq F_n(\mathbf{D}^\star),$$

since the previous inequality is true over the entire set $\mathcal{Z}_t$. We can finally observe using Lemma 1 that

$$\|\mathbf{D}_0 - \mathbf{D}^\star\|_{\mathrm{F}}^2 = 2\sum_{j=1}^p \|\mathbf{d}^{\star,j} - \mathbf{d}_0\|_2^2 \leq \sum_{j=1}^p [\mathbf{v}_j^\star t^\star]^2 \leq [t^\star]^2 < t^2,$$

which leads to the advertised conclusion.

B Proof of Theorem 3 and 4

We start with the more general theorem:

B.1 Proof of Theorem 3

We recall that we assume in Theorem 5 that $c_\lambda \cdot t < \frac{4\underline{\alpha}}{9\sigma_\alpha}$ and for small enough noise levels σ one can find a regularization parameter $\lambda > 0$ such that

$$c_\lambda \cdot \sqrt{t^2\sigma_\alpha^2 + m\sigma^2} \leq \lambda \leq \frac{4}{9}\underline{\alpha}.$$

Given such σ and λ, we define

$$\gamma \triangleq \frac{\lambda}{c_\gamma \sqrt{t^2\sigma_\alpha^2 + m\sigma^2}} \geq \sqrt{2\log(2)}.$$

Here, c_λ and $c_\gamma \triangleq \frac{\sqrt{5}}{3}c_\lambda$ stand for some universal constants which can be made explicit thanks to Theorem 5.

Goal: To determine when the lower bound proved in Theorem 5 is stricly positive, it is sufficient to consider when it holds that

$$\begin{aligned}
\mathbb{E}[\alpha^2] \cdot \frac{k}{p} \cdot t^2 \quad &- \quad c_0 \lambda \cdot \mathbb{E}[|\alpha|] \cdot \frac{k}{p} \cdot t \cdot \|\mathbf{D}_0\|_2 \cdot k\mu(t) \\
&- \quad c_1(t^2\sigma_\alpha^2 + 2m\sigma^2 + 2\lambda k \sigma_\alpha^2) \cdot \gamma^2 e^{-\gamma^2} \\
&- \quad c_2(tk\sigma_\alpha^2 + 2m\sigma^2 + 2\lambda k\sigma_\alpha^2) \cdot \Lambda_n \quad \text{with} \quad \Lambda_n \triangleq \left[mp\frac{\log(n)}{n}\right]^{\frac{1}{2}} \\
&\geq \quad t \cdot (-a_2 t^2 + a_1 t - a_0) > 0,
\end{aligned}$$

for some universal constants c_j which we can make explicit based on Theorem 5, but which we keep hidden for clarity.

Probability of success: The probability of success of Theorem 5 is given by

$$1 - \Big(\frac{mpn}{9}\Big)^{-mp/2} - \exp(-4ne^{-\gamma^2}).$$

This induces a first condition over γ (a upperbound), namely

$$ne^{-\gamma^2} \geq \epsilon_n \;\Rightarrow\; \gamma^2 \leq \log(n) - \log(\epsilon_n), \text{ for some } \epsilon_n \to \infty.$$

From now on, we make the choice $\epsilon_n = \sqrt{n}$, so that $\exp(-4ne^{-\gamma^2}) \leq \exp(-4\sqrt{n})$, along with the condition

$$\gamma^2 \leq \frac{1}{2}\log(n). \tag{30}$$

.

Noiseless/low-noise regime: Even though they are conceptually two different regimes, the treatment of the noisy and noiseless regimes follow the very same reasoning. From now on, we therefore assume that

$$m\sigma^2 \leq t^2\sigma_\alpha^2, \tag{31}$$

which determines the upper level of noise we will be able to handle.

Second-order polynomial function in t: By simply using (31), $\lambda \leq \sqrt{2}c_\gamma \cdot \gamma \cdot \sigma_\alpha t$ and $3+2\sqrt{2}c_\gamma \leq 4\sqrt{2}c_\gamma$, we now make explicit the a_j, $j \in \{0,1,2\}$, which define the second-order polynomial function in t:

$$\begin{aligned}
a_2 &\triangleq 3\sqrt{2}c_0c_\gamma \cdot \sigma_\alpha \mathbb{E}[|\alpha|] \cdot \frac{k}{p} \cdot |||\mathbf{D}_0|||_2 \cdot k \cdot \gamma \\
a_1 &\triangleq \mathbb{E}[\alpha^2] \cdot \frac{k}{p} - \sqrt{2}c_0c_\gamma \cdot \sigma_\alpha \mathbb{E}[|\alpha|] \cdot \frac{k}{p} \cdot |||\mathbf{D}_0|||_2 \cdot k\mu_0 \cdot \gamma - 3c_1\sigma_\alpha^2 \cdot \gamma^2 e^{-\gamma^2} \\
a_0 &\triangleq 2\sqrt{2}c_\gamma \cdot k\sigma_\alpha^2 \cdot [c_1\gamma^3 e^{-\gamma^2} + 2c_2 \cdot \gamma \cdot \Lambda_n].
\end{aligned}$$

We will make use of the following simple lemma to discuss the sign of this polynomial function:

Lemma 2. *Let $(a_0, a_1, a_2) \in \mathbb{R}_+^3$. If $4a_0a_2 < a_1^2$, and $t \in \big[\frac{2a_0}{a_1}, \frac{a_1}{2a_2}\big]$, then $-a_2t^2 + a_1t - a_0 > 0$.*

Some key definitions: Let θ be defined as

$$\theta \triangleq \min\left\{\frac{\underline{\alpha}}{\sigma_\alpha}, \frac{1}{3c_0} \cdot \frac{\mathcal{Q}_\alpha}{k \cdot |||\mathbf{D}_0|||_2}\right\}.$$

We also define $\gamma_{\min} > 1$ the unique number such that

$$\gamma_{\min}^4 e^{-\gamma_{\min}^2} \triangleq \frac{q_\alpha}{69c_1c_\gamma^2} \cdot \frac{1}{p} \cdot \theta, \tag{32}$$

and

$$\gamma_{\max} \triangleq \frac{1}{2}\min\left\{\sqrt{2\log(n)}, \frac{1}{2\sqrt{2}c_0c_\gamma} \cdot \frac{\mathcal{Q}_\alpha}{k\mu_0 \cdot |||\mathbf{D}_0|||_2}\right\}. \tag{33}$$

Moreover, we consider

$$\Lambda_{n,\max} \triangleq \frac{q_\alpha}{138c_2c_\gamma^2} \cdot \frac{1}{p \cdot \gamma^2} \cdot \theta, \tag{34}$$

First step, non-emptiness of $[\gamma_{\min}, \gamma_{\max}]$: We first check that the interval $[\gamma_{\min}, \gamma_{\max}]$ is not empty. On the one hand, if the value of $\gamma_{\max}$ is obtained by $\sqrt{1/2\log(n)}$, we use the fact that $\gamma_{\min} < \gamma_{\max}$ is equivalent to $\gamma_{\min}^4 e^{-\gamma_{\min}^2} > \gamma_{\max}^4 e^{-\gamma_{\max}^2}$. In particular, we have

$$\gamma_{\max}^4 e^{-\gamma_{\max}^2} = \frac{\log^2(n)}{4\sqrt{n}} < \gamma_{\min}^4 e^{-\gamma_{\min}^2},$$

a condition that will be implied by the more stringent condition $\Lambda_n \leq \Lambda_{n,\max}$.

On the other hand, and in the second scenario for $\gamma_{\max}$, we conclude based on the following lemma:

Lemma 3. *Let $a > 1$ and $b \in (0, 1/5]$. If $a^4 e^{-a^2} = b$, then $\sqrt{\log(1/b)} \leq a \leq 2\sqrt{\log(1/b)}$.*

The sufficient condition which stems form this lemma reads

$$k\mu_0 \cdot |\!|\!| \mathbf{D}_0 |\!|\!|_2 \leq \frac{1}{4\sqrt{2}c_0 c_\gamma} \cdot \frac{\mathcal{Q}_\alpha}{\sqrt{\log\left(69 c_1 c_\gamma^2 \cdot \frac{1}{q_\alpha} \cdot \frac{p}{\theta}\right)}}.$$

Second step, lower bound on a_1: For any $\gamma \in [\gamma_{\min}, \gamma_{\max}]$, it is first easy to check that

$$\sqrt{2}c_0 c_\gamma \cdot \sigma_\alpha \mathbb{E}[|\alpha|] \cdot \frac{k}{p} \cdot |\!|\!| \mathbf{D}_0 |\!|\!|_2 \cdot k\mu_0 \cdot \gamma \leq \frac{1}{4}\mathbb{E}[\alpha^2] \cdot \frac{k}{p}.$$

Moreover, since $\gamma^2 e^{-\gamma^2} \leq \gamma^4 e^{-\gamma^2}$ and $\frac{1}{12c_1} q_\alpha \cdot \frac{k}{p} > \gamma_{\min}^4 e^{-\gamma_{\min}^2}$, we therefore obtain that

$$a_1 \geq \mathbb{E}[\alpha^2] \cdot \frac{k}{p} - \frac{1}{4} \cdot \mathbb{E}[\alpha^2] \cdot \frac{k}{p} - \frac{1}{4} \cdot \mathbb{E}[\alpha^2] \cdot \frac{k}{p} \geq \frac{1}{2} \cdot \mathbb{E}[\alpha^2] \cdot \frac{k}{p}.$$

Third step, the condition $4a_0 a_2 < a_1^2$: Since we have $a_1 > \frac{1}{2} \cdot \mathbb{E}[\alpha^2] \cdot \frac{k}{p}$, and

$$a_2 \leq 4\sqrt{2}c_\gamma \cdot k\sigma_\alpha^2 \cdot \max\left\{c_1 \gamma^3 e^{-\gamma^2}, 2c_2 \cdot \gamma \cdot \Lambda_n\right\},$$

simple computations show that $\gamma \geq \gamma_{\min}$ and $\Lambda_n \leq \Lambda_{n,\max}$, as defined in (32) and (34), lead to $4a_0 a_2 < a_1^2$.

Conclusions: We have proved that for $\gamma \in [\gamma_{\min}, \gamma_{\max}]$, $\Lambda_n \leq \Lambda_{n,\max}$, and

$$k\mu_0 \cdot |\!|\!| \mathbf{D}_0 |\!|\!|_2 \leq \frac{1}{4\sqrt{2}c_0 c_\gamma} \cdot \frac{\mathcal{Q}_\alpha}{\sqrt{\log\left(69 c_1 c_\gamma^2 \cdot \frac{1}{q_\alpha} \cdot \frac{p}{\theta}\right)}},$$

the lower bound provided by Theorem 5 is stricly positive for a radius $t \in \left[\frac{2a_0}{a_1}, \frac{a_1}{2a_2}\right]$ (see Lemma 2) and a noise $\sigma \leq \sigma_\alpha \sqrt{m} t$. Taking the smallest allowed radius (i.e., $t = \frac{2a_0}{a_1}$ with $\gamma = \gamma_{\max}$) leads to the displayed result.

B.2 Proof of Theorem 4

We now discuss the version of Theorem 3 in the simpler setting where there is no noise (i.e., $\sigma = 0$) and $\boldsymbol{\alpha}_0$ is almost surely bounded by $\overline{\alpha} \geq \underline{\alpha} > 0$. The main consequence of these simplifying assumptions is that there is no residual term to consider anymore and our surrogate function coincide almost surely with the true sparse coding function, provided the radius t is small enough, as proved in Proposition 4. As a result, the terms depending on γ in Theorem 5 disappear, and the probability of success simplifies to

$$1 - \left(\frac{mpn}{9}\right)^{-mp/2}.$$

Moreover, in light of Proposition 4, we now ask for

$$\frac{1}{3} c_\lambda \sqrt{k}\overline{\alpha} t \leq \lambda \leq \frac{4}{9}\underline{\alpha}.$$

The backbone of the proof remains identical, we adapt the discussion about the polynomial function in t.

Goal: To determine when the lower bound proved in Theorem 5 is stricly positive, it is sufficient to consider when it holds that

$$\begin{aligned}\mathbb{E}[\alpha^2]\cdot\frac{k}{p}\cdot t^2 &- c_0\lambda\cdot\mathbb{E}[|\alpha|]\cdot\frac{k}{p}\cdot t\cdot\|\mathbf{D}_0\|_2\cdot k\mu(t)\\ &- c_1(tk\sigma_\alpha^2+2\lambda k\sigma_\alpha^2)\cdot\Lambda_n \quad\text{with}\quad \Lambda_n\triangleq\Big[mp\frac{\log(n)}{n}\Big]^{\frac{1}{2}}\\ &\geq t\cdot(-a_2t^2+a_1t-a_0)>0,\end{aligned}$$

for some universal constants c_j which we can make explicit based on Theorem 5, but which we keep hidden for clarity.

Second-order polynomial function in t: By making the choice $\lambda\triangleq\frac{1}{2}c_\lambda\cdot\overline{\alpha}\cdot\sqrt{k}\cdot t$, we now make explicit the a_j, $j\in\{0,1,2\}$, which define the second-order polynomial function in t:

$$\begin{aligned}a_2 &\triangleq \frac{3}{2}c_0c_\lambda\cdot\overline{\alpha}\cdot\mathbb{E}[|\alpha|]\cdot\frac{k}{p}\cdot\|\mathbf{D}_0\|_2\cdot k^{3/2}\\ a_1 &\triangleq \mathbb{E}[\alpha^2]\cdot\frac{k}{p}-\frac{1}{2}c_0c_\lambda\cdot\overline{\alpha}\cdot\mathbb{E}[|\alpha|]\cdot\frac{k}{p}\cdot\|\mathbf{D}_0\|_2\cdot k^{3/2}\mu_0\\ a_0 &\triangleq 2c_1c_\lambda\cdot k^{3/2}\sigma_\alpha\overline{\alpha}\cdot\Lambda_n.\end{aligned}$$

Conclusions: Consider the condition

$$\|\mathbf{D}_0\|_2\cdot k^{3/2}\mu_0\leq\frac{1}{c_0c_\lambda}\frac{\mathbb{E}[\alpha^2]}{\overline{\alpha}\cdot\sigma_\alpha},$$

so that $a_1\geq\frac{1}{2}\cdot\mathbb{E}[\alpha^2]\cdot\frac{k}{p}$. By using again Lemma 2, and by defining

$$\Lambda_{n,\max}\triangleq\frac{1}{9c_1c_\lambda^2}\cdot\frac{\mathbb{E}[\alpha^2]}{\overline{\alpha}^2}\cdot\frac{1}{kp}\cdot\min\left\{\frac{\underline{\alpha}}{\sigma_\alpha},\frac{1}{5c_0}\cdot\frac{\mathcal{Q}_\alpha}{k\cdot\|\mathbf{D}_0\|_2}\right\},$$

it is easy to check that $\Lambda_n\leq\Lambda_{n,\max}$ implies that $4a_0a_2<a_1^2$ along with

$$2\frac{a_0}{a_1}<\frac{8}{9c_\lambda}\frac{\underline{\alpha}}{\overline{\alpha}}\cdot\frac{1}{\sqrt{k}},$$

as required by our choice of λ and the fact that $\lambda\leq\frac{4}{9}\underline{\alpha}$.

C Uniform restricted isometry and coherence properties

First, we introduce $\mathbf{P}_{\mathrm{J}}(t)\in\mathbb{R}^{m\times m}$ the orthogonal projector which projects onto the span of $[\mathbf{D}(t)]_{\mathrm{J}}$ and establish a result that holds without any assumption on $\mathbf{D}_0$.

Lemma 4. *For any $\mathbf{W}\in\mathcal{W}_{\mathbf{D}_0},\mathbf{v}\in\mathcal{S}^p$, $t\geq0$ and* J,

$$\|[\mathbf{D}(t)-\mathbf{D}_0]_{\mathrm{J}}\|_2^2\leq\|[\mathbf{D}(t)-\mathbf{D}_0]_{\mathrm{J}}\|_{\mathrm{F}}^2 \leq t^2\cdot\|\mathbf{v}_{\mathrm{J}}\|_2^2 \tag{35}$$

$$\|(\mathbf{I}-\mathbf{P}_{\mathrm{J}}(t))[\mathbf{D}_0]_{\mathrm{J}}\|_2^2\leq\|(\mathbf{I}-\mathbf{P}_{\mathrm{J}}(t))[\mathbf{D}_0]_{\mathrm{J}}\|_{\mathrm{F}}^2 \leq t^2\cdot\|\mathbf{v}_{\mathrm{J}}\|_2^2. \tag{36}$$

Proof. For the first result we observe

$$\|[\mathbf{D}(t)-\mathbf{D}_0]_{\mathrm{J}}\|_{\mathrm{F}}^2=\sum_{j\in\mathrm{J}}\|\mathbf{d}_0^j(t)-\mathbf{d}_0^j\|_2^2=4\sum_{j\in\mathrm{J}}\sin^2(\mathbf{v}_jt/2)\leq4\sum_{j\in\mathrm{J}}\mathbf{v}_j^2\frac{t^2}{4}\leq t^2\cdot\|\mathbf{v}_{\mathrm{J}}\|_2^2.$$

For the second one, using Lemma 1 with $\mathbf{D}_1 = \mathbf{D}(t) = \mathbf{D}(\mathbf{D}_0, \mathbf{W}, \mathbf{v}, t)$, $\mathbf{D}_2 = \mathbf{D}_0$, there exists $\mathbf{W}' \in \mathcal{W}_{\mathbf{D}(t)}$ such that for each j, $\mathbf{d}_0^j = \mathbf{d}^j(t)\cos(\mathbf{v}_j t) + \mathbf{w}'^j \sin(\mathbf{v}_j t)$. Hence, denoting $\mathbf{C} = \mathrm{Diag}(\cos(\mathbf{v}_j t))$ and $\mathbf{S} = \mathrm{Diag}(\sin(\mathbf{v}_j t))$ we have $[\mathbf{D}_0]_{\mathrm{J}} = [\mathbf{D}(t)\mathbf{C}]_{\mathrm{J}} + [\mathbf{W}'\mathbf{S}]_{\mathrm{J}}$. Each column of $[\mathbf{D}(t)\mathbf{C}]_{\mathrm{J}}$ belongs to the span of the columns of $[\mathbf{D}(t)]_{\mathrm{J}}$, so that

$$(\mathbf{I} - \mathbf{P}_{\mathrm{J}}(t))[\mathbf{D}_0]_{\mathrm{J}} = (\mathbf{I} - \mathbf{P}_{\mathrm{J}}(t))[\mathbf{W}'\mathbf{S}]_{\mathrm{J}}. \tag{37}$$

As a result,

$$\|(\mathbf{I} - \mathbf{P}_{\mathrm{J}}(t))[\mathbf{D}_0]_{\mathrm{J}}\|_{\mathrm{F}}^2 \;=\; \|(\mathbf{I} - \mathbf{P}_{\mathrm{J}}(t))[\mathbf{W}'\mathbf{S}]_{\mathrm{J}}\|_{\mathrm{F}}^2 \leq \|[\mathbf{W}'\mathbf{S}]_{\mathrm{J}}\|_{\mathrm{F}}^2 = \sum_{j\in\mathrm{J}} \sin^2(\mathbf{v}_j t) \leq \|\mathbf{v}_{\mathrm{J}}\|_2^2 \cdot t^2.$$

□

Next, we control the norms of $\mathbf{\Theta}_{\mathrm{J}}(t') \triangleq \left[\mathbf{D}_{\mathrm{J}}^\top(t')\mathbf{D}_{\mathrm{J}}(t')\right]^{-1}$ when this is a well-defined matrix. For that, we first recall the definition of the restricted isometry constant of order k of a dictionary $\mathbf{D}$, $\delta_k(\mathbf{D})$, as the smallest number δ_k such that for any support set J of size $|\mathrm{J}| = k$ and $\mathbf{z} \in \mathbb{R}^k$,

$$(1 - \delta_k)\,\|\mathbf{z}\|_2^2 \leq \|\mathbf{D}\mathbf{z}\|_2^2 \leq (1 + \delta_k)\,\|\mathbf{z}\|_2^2. \tag{38}$$

Lemma 5. *Let $\mathbf{D}_0 \in \mathbb{R}^{m\times p}$ be a dictionary and k such that $\delta_k(\mathbf{D}_0) < 1$. For any $t < \sqrt{1-\delta_k(\mathbf{D}_0)}$ define*

$$C_t \triangleq \frac{1}{\sqrt{1-\delta_k(\mathbf{D}_0)} - t}. \tag{39}$$

For any $\mathbf{W} \in \mathcal{W}_{\mathbf{D}_0}$, $\mathbf{v} \in \mathcal{S}^p$, $0 \leq t' \leq t$ and J of size k, the $\mathrm{J}\times\mathrm{J}$ matrix

$$\mathbf{\Theta}_{\mathrm{J}}(t') \triangleq \left[\mathbf{D}_{\mathrm{J}}^\top(t')\mathbf{D}_{\mathrm{J}}(t')\right]^{-1} \tag{40}$$

is well defined and we have

$$\begin{aligned}
\|\mathbf{D}_{\mathrm{J}}(t')\|_2 = \|\mathbf{D}_{\mathrm{J}}^\top(t')\|_2 &\leq C_t &(41)\\
\|\mathbf{\Theta}_{\mathrm{J}}(t')\|_2 &\leq C_t^2 &(42)\\
\|\mathbf{D}_{\mathrm{J}}(t')\mathbf{\Theta}_{\mathrm{J}}(t')\|_2 &\leq C_t. &(43)
\end{aligned}$$

Proof. By the triangle inequality and Lemma 4-Equation (35), for any J of size k and $\mathbf{z} \in \mathbb{R}^k$ we have

$$\begin{aligned}
\|\mathbf{D}_{\mathrm{J}}(t')\mathbf{z}\|_2 &\geq \|[\mathbf{D}_0]_{\mathrm{J}}\mathbf{z}\|_2 - \|[\mathbf{D}(t') - \mathbf{D}_0]_{\mathrm{J}}\mathbf{z}\|_2 \geq \left(\sqrt{1-\delta_k(\mathbf{D}_0)} - t'\|\mathbf{v}_{\mathrm{J}}\|_2\right)\cdot\|\mathbf{z}\|_2 \geq \left(\sqrt{1-\delta_k(\mathbf{D}_0)} - t\right)\cdot\|\mathbf{z}\|_2\\
\|\mathbf{D}_{\mathrm{J}}(t)\mathbf{z}\|_2 &\leq \|[\mathbf{D}_0]_{\mathrm{J}}\mathbf{z}\|_2 + \|[\mathbf{D}(t) - \mathbf{D}_0]_{\mathrm{J}}\mathbf{z}\|_2 \leq \left(\sqrt{1+\delta_k(\mathbf{D}_0)} + t'\|\mathbf{v}_{\mathrm{J}}\|_2\right)\cdot\|\mathbf{z}\|_2 \leq \left(\sqrt{1+\delta_k(\mathbf{D}_0)} + t\right)\cdot\|\mathbf{z}\|_2.
\end{aligned}$$

Hence, in the sense of symmetric positive definite matrices

$$\left(\sqrt{1-\delta_k(\mathbf{D}_0)} - t\right)^2 \cdot \mathbf{I} \preceq \mathbf{D}_{\mathrm{J}}^\top(t')\mathbf{D}_{\mathrm{J}}(t') \preceq \left(\sqrt{1+\delta_k(\mathbf{D}_0)} + t\right)^2 \cdot \mathbf{I}.$$

As a result, $\mathbf{D}_{\mathrm{J}}^\top(t')\mathbf{D}_{\mathrm{J}}(t')$ is invertible so $\mathbf{\Theta}_{\mathrm{J}}(t')$ is indeed well defined, and

$$\begin{aligned}
\|\mathbf{D}_{\mathrm{J}}(t')\|_2 &= \|\mathbf{D}_{\mathrm{J}}^\top(t')\|_2 = \sqrt{\|\mathbf{D}_{\mathrm{J}}^\top(t')\mathbf{D}_{\mathrm{J}}(t')\|_2} \leq \sqrt{1+\delta_k(\mathbf{D}_0)} + t \leq \frac{1}{\sqrt{1-\delta_k(\mathbf{D}_0)} - t}\\
\|\mathbf{\Theta}_{\mathrm{J}}(t')\|_2 &= \|(\mathbf{D}_{\mathrm{J}}^\top(t')\mathbf{D}_{\mathrm{J}}(t'))^{-1}\|_2 \leq \frac{1}{\left(\sqrt{1-\delta_k(\mathbf{D}_0)} - t\right)^2}\\
\|\mathbf{D}_{\mathrm{J}}(t')\mathbf{\Theta}_{\mathrm{J}}(t')\|_2 &= \sqrt{\|\mathbf{\Theta}_{\mathrm{J}}(t')\mathbf{D}_{\mathrm{J}}^\top(t')\mathbf{D}_{\mathrm{J}}(t')\mathbf{\Theta}_{\mathrm{J}}(t')\|_2} = \sqrt{\|\mathbf{\Theta}_{\mathrm{J}}(t')\|_2} \leq \frac{1}{\sqrt{1-\delta_k(\mathbf{D}_0)} - t}.
\end{aligned}$$

□

To continue, we control certain norms of the dictionary when it has low coherence:

Lemma 6. *Let $\mathbf{D}_0 \in \mathbb{R}^{m\times p}$ be a dictionary with coherence μ and normalized columns (i.e., with unit ℓ_2-norm). For any $\mathrm{J} \subseteq [\![1;p]\!]$ with $|\mathrm{J}| \leq k$, We have*

$$\|\mathbf{D}_\mathrm{J}^\top \mathbf{D}_\mathrm{J} - \mathbf{I}\|_2 \leq \|\mathbf{D}_\mathrm{J}^\top \mathbf{D}_\mathrm{J} - \mathbf{I}\|_\mathrm{F} \leq k\mu,$$

along with

$$\|\mathbf{D}_\mathrm{J}\mathbf{D}_\mathrm{J}^\top\|_2 = \|\mathbf{D}_\mathrm{J}^\top \mathbf{D}_\mathrm{J}\|_2 \leq 1 + k\mu \quad \textit{and} \quad \delta_k(\mathbf{D}) \leq k\mu.$$

Similarly, it holds

$$\|\mathbf{D}_\mathrm{J}^\top \mathbf{D}_\mathrm{J}\|_\infty \leq 1 + k\mu \quad \textit{and} \quad \|\mathbf{D}_{\mathrm{J}^\mathrm{c}}^\top \mathbf{D}_\mathrm{J}\|_\infty \leq k\mu.$$

Moreover, introduce for any $\mathbf{A} \in \mathbb{R}^{k\times k}$ the matrix norm

$$N(\mathbf{A}) \triangleq k \cdot \max_{i,j\in[\![1;k]\!]} |\mathbf{A}_{i,j}|$$

and consider

$$\mathbf{\Theta}_\mathrm{J} \triangleq \left[\mathbf{D}_\mathrm{J}^\top \mathbf{D}_\mathrm{J}\right]^{-1}.$$

If we further assume $k\mu < 1$, then $\mathbf{\Theta}_\mathrm{J}$ is well-defined and

$$\max\left\{\|\mathbf{\Theta}_\mathrm{J} - \mathbf{I}\|_\infty, \|\mathbf{\Theta}_\mathrm{J} - \mathbf{I}\|_2, \|\mathbf{\Theta}_\mathrm{J} - \mathbf{I}\|_\mathrm{F}, N(\mathbf{\Theta}_\mathrm{J} - \mathbf{I})\right\} \leq \frac{k\mu}{1-k\mu},$$

along with

$$\max\left\{\|\mathbf{\Theta}_\mathrm{J}\|_\infty, \|\mathbf{\Theta}_\mathrm{J}\|_2\right\} \leq \frac{1}{1-k\mu}.$$

Proof. These properties are already well-known [see, e.g. Tropp, 2004, Fuchs, 2005]. We briefly prove them. First, we introduce $\mathbf{H} = \mathbf{D}_\mathrm{J}^\top \mathbf{D}_\mathrm{J} - \mathbf{I}$. A straightforward elementwise upper bound leads to

$$\|\mathbf{H}\|_2 \leq \|\mathbf{H}\|_\mathrm{F} = \sum_{i\in\mathrm{J}} \sum_{j\in\mathrm{J}\setminus\{i\}} ([\mathbf{d}^i]^\top \mathbf{d}^j)^2 \leq k(k-1)\mu^2 \leq k^2\mu^2.$$

This proves that in the sense of positive definite matrices, $(1-k\mu)\mathbf{I} \preceq \mathbf{D}_\mathrm{J}^\top \mathbf{D}_\mathrm{J} \preceq (1+k\mu)\mathbf{I}$, which shows in turn the bound on $\delta_k(\mathbf{D})$. Moreover, and since $\|\mathbf{I}\|_2 = 1$ with $\|\mathbf{A}^\top\mathbf{A}\|_2 = \|\mathbf{A}\mathbf{A}^\top\|_2$ for any matrix $\mathbf{A}$, we have

$$\|\mathbf{D}_\mathrm{J}^\top \mathbf{D}_\mathrm{J}\|_2 = \|\mathbf{D}_\mathrm{J}\mathbf{D}_\mathrm{J}^\top\|_2 \leq 1 + k\mu.$$

By definition of $\|.\|_\infty$, we also have

$$\|\mathbf{D}_\mathrm{J}^\top \mathbf{D}_\mathrm{J}\|_\infty \leq 1 + \|\mathbf{H}\|_\infty = 1 + \max_{i\in\mathrm{J}} \sum_{j\in\mathrm{J}, j\neq i} |[\mathbf{d}^i]^\top \mathbf{d}^j| \leq 1 + k\mu.$$

Note that for $\|\mathbf{D}_{\mathrm{J}^\mathrm{c}}^\top \mathbf{D}_\mathrm{J}\|_\infty$, there are no diagonal terms to take into account.

Now, if $k\mu < 1$ holds, then we have $\max\{\|\mathbf{H}\|_\infty, \|\mathbf{H}\|_2, \|\mathbf{H}\|_\mathrm{F}, N(\mathbf{H})\} \leq k\mu < 1$ and there are convergent series expansion of $[\mathbf{I}+\mathbf{H}]^{-1}$ in each of these norms [Horn and Johnson, 1990]. By sub-multiplicativity, we obtain

$$\|\mathbf{\Theta}_\mathrm{J} - \mathbf{I}\| = \|\sum_{t=1}^{\infty} (-1)^t \mathbf{H}^t\| \leq k\mu/(1-k\mu)$$

where $\|.\|$ stands for one the four aforementioned matrix norms. The last result lies in the fact that for the norms $\|\cdot\|_\infty$, $\|\cdot\|_2$, we have $\|\mathbf{I}\| = 1$ and

$$\|\mathbf{\Theta}_\mathrm{J}\| \leq \|\mathbf{\Theta}_\mathrm{J} - \mathbf{I}\| + \|\mathbf{I}\| \leq 1 + k\mu/(1-k\mu) = 1/(1-k\mu).$$

□

We now derive a simple corollary which will be useful for the computation of expectations:

Corollary 1. *Let $\mathbf{D} \in \mathbb{R}^{m\times p}$ be a dictionary with normalized columns and coherence μ. With the notation from Lemma 6, if $k\mu < 1$, we have for any $a \in \{1,2\}$ and for any $\mathrm{J} \subseteq [\![1;p]\!]$ with $|\mathrm{J}| \leq k$,*

$$\max_{i,j\in[\![1;k]\!],i\neq j} |[\mathbf{\Theta}_{\mathrm{J}}^a]_{i,j}| \leq \frac{a\mu}{(1-k\mu)^a}.$$

Proof. We first make use of Lemma 6 which gives

$$N(\mathbf{\Theta}_{\mathrm{J}} - \mathbf{I}) = k \cdot \max_{i,j\in[\![1;k]\!]} |[\mathbf{\Theta}_{\mathrm{J}} - \mathbf{I}]_{i,j}| \leq \frac{k\mu}{1-k\mu},$$

which notably implies that

$$\max_{i,j\in[\![1;k]\!],i\neq j} |[\mathbf{\Theta}_{\mathrm{J}}]_{i,j}| \leq \frac{\mu}{(1-k\mu)}.$$

We continue by noticing that $[\mathbf{\Theta}_{\mathrm{J}} - \mathbf{I}]^2 = \mathbf{\Theta}_{\mathrm{J}}^2 - \mathbf{I} + 2(\mathbf{I} - \mathbf{\Theta}_{\mathrm{J}})$ and by sub-multiplicativity of N

$$N([\mathbf{\Theta}_{\mathrm{J}} - \mathbf{I}]^2) \leq [N(\mathbf{\Theta}_{\mathrm{J}} - \mathbf{I})]^2 \leq \frac{(k\mu)^2}{(1-k\mu)^2}.$$

Applying the triangle inequality, we obtain

$$N(\mathbf{\Theta}_{\mathrm{J}}^2 - \mathbf{I}) \leq 2N(\mathbf{\Theta}_{\mathrm{J}} - \mathbf{I}) + \frac{(k\mu)^2}{(1-k\mu)^2} \leq \frac{2k\mu(1-k\mu)+(k\mu)^2}{(1-k\mu)^2} = \frac{2k\mu-(k\mu)^2}{(1-k\mu)^2} \leq \frac{2k\mu}{(1-k\mu)^2}.$$

As a result, we finally get

$$\max_{i,j\in[\![1;k]\!],i\neq j} |[\mathbf{\Theta}_{\mathrm{J}}^2]_{i,j}| \leq \frac{2\mu}{(1-k\mu)^2},$$

hence the advertised conclusion. □

Corollary 2. *Let $\mathbf{D}_0 \in \mathbb{R}^{m\times p}$ be a dictionary with normalized columns. If $k\mu(t) < 1/2$ then, for any $\mathbf{W} \in \mathcal{W}_{\mathbf{D}_0}$, $\mathbf{v} \in \mathcal{S}^p$ and $0 \leq t' \leq t$ we have*

$$\left\| \left[\mathbf{D}_{\mathrm{J}^c}^\top(t')\mathbf{D}_{\mathrm{J}}(t')\right] \left[\mathbf{D}_{\mathrm{J}}^\top(t')\mathbf{D}_{\mathrm{J}}(t')\right]^{-1} \right\|_\infty \leq \frac{k\mu(t)}{1-k\mu(t)} = k\mu(t)Q_t^2 = Q_t^2 - 1 < 1, \tag{44}$$

where we introduce

$$Q_t \triangleq \frac{1}{\sqrt{1-k\mu(t)}} \geq C_t.$$

D Expectation over J

Lemma 7. *Let $\mathbf{D}_0 \in \mathbb{R}^{m\times p}$ be any dictionary and J a random support. Denoting by $\delta(i) \triangleq \mathbb{1}_{\mathrm{J}}(i)$ the indicator function of J, we 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 we have for any $\mathbf{v} \in \mathcal{S}^p$ and $0 \leq t' \leq t$,

$$\mathbb{E}\{\|[\mathbf{D}_0]_{\mathrm{J}}^\top [\mathbf{D}_0]_{\mathrm{J}} - \mathbf{I}\|_{\mathrm{F}}^2\} = \|\mathbf{D}_0^\top\mathbf{D}_0 - \mathbf{I}\|_{\mathrm{F}}^2 \cdot \frac{k(k-1)}{p(p-1)} \tag{45}$$

$$\mathbb{E}\{\|\mathbf{v}_{\mathrm{J}}\|_2^2\} = \frac{k}{p} \tag{46}$$

$$\mathbb{E}\{\|\mathbf{D}_{\mathrm{J}}^\top(t')\mathbf{D}_{\mathrm{J}}(t') - \mathbf{I}\|_{\mathrm{F}} \cdot \|\mathbf{v}_{\mathrm{J}}\|_2\} \leq \left(\|\mathbf{D}_0^\top\mathbf{D}_0 - \mathbf{I}\|_{\mathrm{F}} \cdot \sqrt{\frac{k-1}{p-1}}\right) \cdot \frac{k}{p} + 2 \cdot C_t \cdot t \cdot \frac{k}{p}. \tag{47}$$

Proof. To obtain (45) and (46) we simply expand

$$\begin{aligned}\mathbb{E}\{\|[\mathbf{D}_0]_{\mathrm{J}}^\top[\mathbf{D}_0]_{\mathrm{J}}-\mathbf{I}\|_{\mathrm{F}}^2\} &= \mathbb{E}\Big\{\sum_{i\in[\![1;p]\!]}\sum_{j\in[\![1;p]\!],j\neq i}\delta(i)\delta(j)\cdot[\mathbf{d}_0^i]^\top\mathbf{d}_0^j\Big\}=\sum_{i\in[\![1;p]\!]}\sum_{j\in[\![1;p]\!],j\neq i}\frac{k(k-1)}{p(p-1)}\cdot[\mathbf{d}_0^i]^\top\mathbf{d}_0^j\\ \mathbb{E}\{\|\mathbf{v}_{\mathrm{J}}\|_2^2\} &= \mathbb{E}\Big\{\sum_{i\in[\![1;p]\!]}\delta(i)\cdot\mathbf{v}_i^2\Big\}=\sum_{i\in[\![1;p]\!]}\frac{k}{p}\mathbf{v}_i^2=\frac{k}{p}\cdot\|\mathbf{v}\|_2^2=\frac{k}{p}.\end{aligned}$$

Now, by Lemma 5 and the Cauchy-Schwartz inequality for random variables

$$\begin{aligned}\mathbb{E}\{\|\mathbf{D}_{\mathrm{J}}^\top(t')\mathbf{D}_{\mathrm{J}}(t')-\mathbf{I}\|_{\mathrm{F}}\cdot\|\mathbf{v}_{\mathrm{J}}\|_2\} &\leq \mathbb{E}\{\|[\mathbf{D}_0]_{\mathrm{J}}^\top[\mathbf{D}_0]_{\mathrm{J}}-\mathbf{I}\|_{\mathrm{F}}\cdot\|\mathbf{v}_{\mathrm{J}}\|_2\}+2\cdot C_t\cdot t\cdot\mathbb{E}\{\|\mathbf{v}_{\mathrm{J}}\|_2^2\}\\ &\leq \sqrt{\mathbb{E}\{\|[\mathbf{D}_0]_{\mathrm{J}}^\top[\mathbf{D}_0]_{\mathrm{J}}-\mathbf{I}\|_{\mathrm{F}}^2\}}\cdot\sqrt{\mathbb{E}\{\|\mathbf{v}_{\mathrm{J}}\|_2^2\}}+2\cdot C_t\cdot t\cdot\frac{k}{p}\\ &\leq \|\mathbf{D}_0^\top\mathbf{D}_0-\mathbf{I}\|_{\mathrm{F}}\cdot\sqrt{\frac{k(k-1)}{p(p-1)}}\cdot\sqrt{\frac{k}{p}}+2\cdot C_t\cdot t\cdot\frac{k}{p}\end{aligned}$$

□

E Proof of Proposition 2

In this section, we establish the results required to lower bound $\Delta\Phi_n(\mathbf{W},\mathbf{v},t)$. We denote

$$\Delta\phi_{\mathbf{x}^i}(\mathbf{W},\mathbf{v},t)\triangleq\phi_{\mathbf{x}^i}(\mathbf{D}(\mathbf{W},\mathbf{v},t)|\mathbf{s}_0^i)-\phi_{\mathbf{x}^i}(\mathbf{D}_0|\mathbf{s}_0^i). \tag{48}$$

The overall approach consists of the following steps:

1. **Concentration around the expectation:**

   Lemma 8. *Under our signal model, for any $\mathbf{W}\in\mathcal{W}_{\mathbf{D}_0}$, $\mathbf{v}\in\mathcal{S}^p$, $\tau\in[0,\sqrt{n}]$, we have*

   $$\Pr\Big(\Delta\Phi_n(\mathbf{W},\mathbf{v},t)<\mathbb{E}\{\Delta\phi_{\mathbf{x}}(\mathbf{W},\mathbf{v},t)\}-c(t)\frac{\tau}{\sqrt{n}}\Big)\leq 2\cdot\exp(-\tau^2) \tag{49}$$

   with

   $$c(t)\triangleq 102\cdot\big(t^2\sigma_\alpha^2+2m\sigma^2+2\lambda k\sigma_\alpha\big) \tag{50}$$

2. **Control of the Lipschitz constant:** the second step consists in showing that $(\mathbf{W},\mathbf{v})\mapsto\Delta\Phi_n(\mathbf{W},\mathbf{v},t)$ is Lipschitz with controlled constant with respect to the metric

   $$d\big((\mathbf{W},\mathbf{v}),(\mathbf{W}',\mathbf{v}')\big)\triangleq\max\left\{\max_{j\in[\![1;p]\!]}\|\mathbf{w}^j-\mathbf{w}'^j\|_2\ ,\ \|\mathbf{v}-\mathbf{v}'\|_2\right\}. \tag{51}$$

   Lemma 9. *Assume that $t<\sqrt{1-\delta_k(\mathbf{D}_0)}$. Under our signal model we have for any $\tau\in[0,\sqrt{n}]$, except with probability at most $2\exp(-\tau^2)$: for all $(\mathbf{W},\mathbf{v})$ and $(\mathbf{W}',\mathbf{v}')$*

   $$\big|\Delta\Phi_n(\mathbf{W},\mathbf{v},t)-\Delta\Phi_n(\mathbf{W}',\mathbf{v}',t)\big|\leq L\cdot\left(1+\frac{4\tau}{\sqrt{n}}\right)\cdot d\big((\mathbf{W},\mathbf{v}),(\mathbf{W}',\mathbf{v}')\big).$$

   where

   $$L\triangleq 30C_t^3\cdot t\cdot\big[5(k\sigma_\alpha^2+m\sigma^2)+\lambda^2k\big] \tag{52}$$

3. **ϵ-net argument:** combining Lemmata 8-9 together with an estimate of the size of an ϵ-net of $\mathcal{W}\times\mathcal{S}^p$ with respect to the considered metric, we obtain

Lemma 10. *Assume that* $t < \sqrt{1-\delta_k(\mathbf{D}_0)}$ *and that* $C_t \leq 1.5$. *Under our signal model, and assuming that*

$$\frac{n}{\log n} \geq mp$$

we have, except with probability at most $\left(\frac{mpn}{9}\right)^{-mp/2}$,

$$\inf_{\mathbf{W}\in\mathcal{W}_{\mathbf{D}_0},\mathbf{v}\in\mathcal{S}^p} \Delta\Phi_n(\mathbf{W},\mathbf{v},t) \geq \inf_{\mathbf{W}\in\mathcal{W}_{\mathbf{D}_0},\mathbf{v}\in\mathcal{S}^p} \mathbb{E}\{\Delta\phi_{\mathbf{x}}(\mathbf{W},\mathbf{v},t)\} - B\cdot\sqrt{mp\frac{\log n}{n}}.$$

with

$$B \triangleq 3045\left(k\sigma_\alpha^2\cdot t + 2m\sigma^2 + \lambda k\sigma_\alpha + \lambda^2 k\cdot t\right). \tag{53}$$

4. **Control in expectation:**

Lemma 11. *Assume that* $k\mu(t) < 1/2$. *Under our signal model, we have*

$$\begin{aligned}\inf_{\mathbf{W}\in\mathcal{W}_{\mathbf{D}_0},\mathbf{v}\in\mathcal{S}^p} \mathbb{E}\{\Delta\phi_{\mathbf{x}}(\mathbf{W},\mathbf{v},t)\} \quad &\geq \quad (1-\mathcal{K}^2)\cdot\frac{\mathbb{E}[\alpha_0^2]}{2}\cdot\frac{k}{p}\cdot t^2 \\ &\quad -Q_t^2\cdot t\cdot\frac{k}{p}\cdot\left\|\mathbf{D}_0\right\|_2\cdot k\mu(t)\cdot\lambda\cdot\left(4Q_t^2\lambda + 3\mathbb{E}\{|\alpha_0|\}\right).\end{aligned}$$

with $\mathcal{K} \triangleq C_t\cdot(\left\|\mathbf{D}_0\right\|_2\cdot\sqrt{k/p}+t)$.

We obtain Proposition 2 by combining Lemmata 10-11. We now proceed to the proof of these lemmata.

E.1 Expansion of $\Delta\phi_{\mathbf{x}}$

We expand $\Delta\phi_{\mathbf{x}}$ into the sum of six terms.

Lemma 12. *We have*

$$\begin{aligned}\Delta\phi_{\mathbf{x}}(t) \quad &\triangleq \quad \phi_{\mathbf{x}}(\mathbf{D}(t)|\mathbf{s}) - \phi_{\mathbf{x}}(\mathbf{D}_0|\mathbf{s}) \\ &= \quad \frac{1}{2}\mathbf{x}^\top\left[\mathbf{P}_{\mathrm{J}}(0)-\mathbf{P}_{\mathrm{J}}(t)\right]\mathbf{x} - \lambda\mathbf{s}_{\mathrm{J}}^\top\left[\mathbf{\Theta}_{\mathrm{J}}(0)[\mathbf{D}_0]_{\mathrm{J}}^\top - \mathbf{\Theta}_{\mathrm{J}}(t)[\mathbf{D}(t)]_{\mathrm{J}}^\top\right]\mathbf{x} \\ &\quad +\frac{\lambda^2}{2}\mathbf{s}_{\mathrm{J}}^\top\left[\mathbf{\Theta}_{\mathrm{J}}(0)-\mathbf{\Theta}_{\mathrm{J}}(t)\right]\mathbf{s}_{\mathrm{J}} \qquad (54) \\ &= \quad \zeta_{\boldsymbol{\alpha},\boldsymbol{\alpha}}(t)+\zeta_{\boldsymbol{\alpha},\boldsymbol{\varepsilon}}(t)+\zeta_{\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}}(t)+\zeta_{\mathbf{s},\boldsymbol{\alpha}}(t)+\zeta_{\mathbf{s},\boldsymbol{\varepsilon}}(t)+\zeta_{\mathbf{s},\mathbf{s}}(t) \qquad (55)\end{aligned}$$

where

$$\begin{aligned}\zeta_{\boldsymbol{\alpha},\boldsymbol{\alpha}}(t) \quad &\triangleq \quad \frac{1}{2}\boldsymbol{\alpha}_0^\top\mathbf{D}_0^\top\Big[\mathbf{P}_{\mathrm{J}}(0)-\mathbf{P}_{\mathrm{J}}(t)\Big]\mathbf{D}_0\boldsymbol{\alpha}_0 \qquad (56)\\ \zeta_{\boldsymbol{\alpha},\boldsymbol{\varepsilon}}(t) \quad &\triangleq \quad \boldsymbol{\alpha}_0^\top\mathbf{D}_0^\top\Big[\mathbf{P}_{\mathrm{J}}(0)-\mathbf{P}_{\mathrm{J}}(t)\Big]\boldsymbol{\varepsilon} \qquad (57)\\ \zeta_{\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}}(t) \quad &\triangleq \quad \frac{1}{2}\boldsymbol{\varepsilon}^\top\Big[\mathbf{P}_{\mathrm{J}}(0)-\mathbf{P}_{\mathrm{J}}(t)\Big]\boldsymbol{\varepsilon} \qquad (58)\\ \zeta_{\mathbf{s},\boldsymbol{\alpha}}(t) \quad &\triangleq \quad -\lambda[\mathrm{sign}(\boldsymbol{\alpha}_0)]_{\mathrm{J}}^\top\Big[\mathbf{\Theta}_{\mathrm{J}}(0)[\mathbf{D}_0]_{\mathrm{J}}^\top - \mathbf{\Theta}_{\mathrm{J}}(t)[\mathbf{D}(t)]_{\mathrm{J}}^\top\Big]\mathbf{D}_0\boldsymbol{\alpha}_0 \qquad (59)\\ \zeta_{\mathbf{s},\boldsymbol{\varepsilon}}(t) \quad &\triangleq \quad -\lambda[\mathrm{sign}(\boldsymbol{\alpha}_0)]_{\mathrm{J}}^\top\Big[\mathbf{\Theta}_{\mathrm{J}}(0)[\mathbf{D}_0]_{\mathrm{J}}^\top - \mathbf{\Theta}_{\mathrm{J}}(t)[\mathbf{D}(t)]_{\mathrm{J}}^\top\Big]\boldsymbol{\varepsilon} \qquad (60)\\ \zeta_{\mathbf{s},\mathbf{s}}(t) \quad &\triangleq \quad \frac{\lambda^2}{2}[\mathrm{sign}(\boldsymbol{\alpha}_0)]_{\mathrm{J}}^\top\Big[\mathbf{\Theta}_{\mathrm{J}}(0)-\mathbf{\Theta}_{\mathrm{J}}(t)\Big]\mathrm{sign}(\boldsymbol{\alpha}_0)_{\mathrm{J}}. \qquad (61)\end{aligned}$$

Proof. Denoting $\mathbf{s} = \mathrm{sign}(\boldsymbol{\alpha}_0)$ and $\mathrm{J} \subseteq [\![1;p]\!]$ the support of $\boldsymbol{\alpha}_0$, we have by definition (see Equation (5)):

$$\begin{aligned}\phi_{\mathbf{x}}(\mathbf{D}(t)|\mathbf{s}) &= \frac{1}{2}\big[\|\mathbf{x}\|_2^2 - ([\mathbf{D}(t)]_{\mathrm{J}}^\top \mathbf{x} - \lambda \mathbf{s}_{\mathrm{J}})^\top ([\mathbf{D}(t)]_{\mathrm{J}}^\top [\mathbf{D}(t)]_{\mathrm{J}})^{-1} ([\mathbf{D}(t)]_{\mathrm{J}}^\top \mathbf{x} - \lambda \mathbf{s}_{\mathrm{J}})\big] \\ &= \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{1}{2}\mathbf{x}^\top \mathbf{P}_{\mathrm{J}}(t)\mathbf{x} + \lambda \mathbf{s}_{\mathrm{J}}^\top \boldsymbol{\Theta}_{\mathrm{J}}(t)[\mathbf{D}(t)]_{\mathrm{J}}^\top \mathbf{x} - \frac{\lambda^2}{2}\mathbf{s}_{\mathrm{J}}^\top \boldsymbol{\Theta}_{\mathrm{J}}(t)\mathbf{s}_{\mathrm{J}}. \end{aligned} \tag{62}$$

This yields (54) and we conclude thanks to $\mathbf{x} = \mathbf{D}_0\boldsymbol{\alpha}_0 + \boldsymbol{\varepsilon} = [\mathbf{D}_0]_{\mathrm{J}}[\boldsymbol{\alpha}_0]_{\mathrm{J}} + \boldsymbol{\varepsilon}$. □

E.2 Proof of Lemma 8

Fix $\mathbf{W}$ and $\mathbf{v}$ and denote $y^i(t) \triangleq \phi_{\mathbf{x}^i}(\mathbf{D}(\mathbf{W},\mathbf{v},t)|\mathbf{s}_0^i)$. By definition of $\phi_{\mathbf{x}}$ we have $y^i(t) \leq \mathcal{L}_{\mathbf{x}^i}(\mathbf{D}(\mathbf{W},\mathbf{v},t), \boldsymbol{\alpha}_0^i)$ hence, using Lemma 23 we have for any $\tau \geq 1$

$$\Pr(y^i(t) \geq A_{\mathcal{L}}(t)\cdot\tau) \leq e^{-\tau}$$

where

$$A_{\mathcal{L}}(t) \triangleq \frac{5(1+\log 2)}{2}\cdot\left(t^2\sigma_\alpha^2 + m\sigma^2 + \lambda k\sigma_\alpha\right).$$

Hence, exploiting Corollary 4 with $\kappa = 1$ and $0 \leq \tau \leq \sqrt{n}$, we obtain,

$$\Pr\left(\Big|\frac{1}{n}\sum_{i=1}^n \big(y^i(t) - \mathbb{E}\{y^i(t)\}\big)\Big| \geq 24A_{\mathcal{L}}(t)\cdot\frac{\tau}{\sqrt{n}}\right) \leq \exp(-\tau^2).$$

Observing that $\Delta\Phi_n(\mathbf{W},\mathbf{v},t) = \frac{1}{n}\sum_{i=1}^n (y^i(t) - y^i(0))$ we obtain, by a union bound,

$$\Pr\left(\Big|\Delta\Phi_n(\mathbf{W},\mathbf{v},t) - \mathbb{E}\{\Delta\Phi_n(\mathbf{W},\mathbf{v},t)\}\Big| \geq 24(A_{\mathcal{L}}(t) + A_{\mathcal{L}}(0))\cdot\frac{\tau}{\sqrt{n}}\right) \leq 2\exp(-\tau^2).$$

We conclude by expliciting

$$24(A_{\mathcal{L}}(t) + A_{\mathcal{L}}(0)) = 60(1+\log 2)\left(t^2\sigma_\alpha^2 + 2m\sigma^2 + 2\lambda k\sigma_\alpha\right) \leq c(t).$$

E.3 Proof of Lemma 9

Given the expansion (54), using the shorthands $\mathbf{P}_{\mathrm{J}} = \mathbf{P}_{\mathrm{J}}(\mathbf{W},\mathbf{v},t)$ and $\mathbf{P}'_{\mathrm{J}} = \mathbf{P}_{\mathrm{J}}(\mathbf{W}',\mathbf{v}',t)$, as well as for other similar quantities, and averaging over n, we obtain

$$\begin{aligned}\left|\Delta\Phi_n - \Delta\Phi'_n\right| &\leq \frac{1}{2n}\sum_{i=1}^n \|\mathbf{x}^i\|_2^2 \cdot \max_{i\in[\![1;n]\!]} \|\mathbf{P}_{\mathrm{J}^i} - \mathbf{P}'_{\mathrm{J}^i}\|_2 + \frac{\lambda\sqrt{k}}{n}\sum_{i=1}^n \|\mathbf{x}^i\|_2 \cdot \max_{i\in[\![1;n]\!]} \|\boldsymbol{\Theta}_{\mathrm{J}^i}\mathbf{D} - \boldsymbol{\Theta}'_{\mathrm{J}^i}\mathbf{D}'\|_2 \\ &\quad + \frac{\lambda^2 k}{2}\cdot \max_{i\in[\![1;n]\!]} \|\boldsymbol{\Theta}_{\mathrm{J}^i} - \boldsymbol{\Theta}'_{\mathrm{J}^i}\|_2\end{aligned}$$

Using Lemma 19 this yields the Lipschitz bound $\left|\Delta\Phi_n - \Delta\Phi'_n\right| \leq L_n \cdot d\big((\mathbf{W},\mathbf{v}),(\mathbf{W}',\mathbf{v}')\big)$ with

$$L_n \leq \frac{5}{2}\cdot t\cdot C_t^3 \cdot \left\{\frac{1}{n}\sum_{i=1}^n \|\mathbf{x}^i\|_2^2 + 2\lambda\sqrt{k}\cdot\frac{1}{n}\sum_{i=1}^n \|\mathbf{x}^i\|_2 + \lambda^2 k\right\} \tag{63}$$

Using Lemma 22 we check that $y^i = \|\mathbf{x}^i\|_2^2$ satisfies the hypothesis (see Eq. (100)) of Lemma 24 with $A = 5(k\sigma_\alpha^2 + m\sigma^2)$. Hence, exploiting Corollary 4 with $\kappa = 1$ and $0 \leq \tau \leq \sqrt{n}$, we obtain, except with probability at most $2\exp(-\tau^2)$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \|\mathbf{x}^i\|_2^2 &\leq 6\cdot[5\cdot(k\sigma_\alpha^2 + m\sigma^2)]\cdot\left(1 + \frac{4\tau}{\sqrt{n}}\right) \\ \frac{1}{n}\sum_{i=1}^n \|\mathbf{x}^i\|_2 &\leq 6\cdot\sqrt{5\cdot(k\sigma_\alpha^2 + m\sigma^2)}\cdot\left(1 + \frac{4\tau}{\sqrt{n}}\right).\end{aligned}$$

Inserting the above estimates into (63) yields, except with probability at most $2\exp(-\tau^2)$,

$$\begin{aligned} L_n &\leq L' \cdot \left(1 + \frac{4\tau}{\sqrt{n}}\right) \\ L' &= \frac{30}{2} \cdot C_t^3 \cdot t \cdot \left[\sqrt{5 \cdot (k\sigma_\alpha^2 + m\sigma^2)} + \lambda\sqrt{k}\right]^2 \leq 30 C_t^3 \cdot t \cdot \left[5(k\sigma_\alpha^2 + m\sigma^2) + \lambda^2 k\right] \triangleq L. \end{aligned}$$

E.4 Proof of Lemma 10

The proof of Lemma 10 exploits the covering number $\mathcal{N}$ of $\mathcal{W}_{\mathbf{D}_0} \times \mathcal{S}^p$ with respect to the metric (51). For background about covering numbers, we refer the reader to Cucker and Smale [2002] and references therein.

Lemma 13 (ϵ-nets for $\mathcal{W}_{\mathbf{D}_0} \times \mathcal{S}^p$)**.** *For the Euclidean metric, and for any $\epsilon > 0$, we have*

$$\mathcal{N}(\mathcal{S}^p, \epsilon) \leq \left(1 + \frac{2}{\epsilon}\right)^p.$$

Moreover, define on $\mathbb{R}^{m\times p}$ the norm $\Omega(\mathbf{M}) \triangleq \max_{j\in[\![1;p]\!]} \|\mathbf{m}^j\|_2$. For the metric induced by Ω, and for any $\epsilon > 0$, we have

$$\mathcal{N}(\mathcal{W}_{\mathbf{D}_0}, \epsilon) \leq \left(1 + \frac{2}{\epsilon}\right)^{p(m-1)}.$$

Proof. We resort to Lemma 2 in Vershynin [2010], which gives the first conclusion for the sphere in $\mathbb{R}^p$. As for the second result, remember that the set $\mathcal{W}_{\mathbf{D}_0}$ is defined as a product of spheres in spaces of dimension $m-1$. Indeed, we have for any $\mathbf{W} \in \mathcal{W}_{\mathbf{D}_0}$ and for any $j \in [\![1;p]\!]$, $\|\mathbf{w}^j\|_2 = 1$ along with the constraint $[\mathbf{d}_0^j]^\top \mathbf{w}^j = 0$, which implies that $\mathbf{w}^j$ belongs to the orthogonal space of $\text{span}(\mathbf{d}_0^j)$ of dimension $m-1$. Considering a product of p nets such as that used for $\mathcal{S}^p$, the second conclusion follows from the definition of the metric based on Ω. □

From Lemma 13 we know that for any $0 < \epsilon \leq 1$ there exists ϵ-net of $\mathcal{W} \times \mathcal{S}^p$ with respect to the metric (51) with at most $(3/\epsilon)^{mp}$ elements. Combining this with Lemmata 8-9, we have for any $0 \leq \tau \leq \sqrt{n}$: except with probability at most $(3/\epsilon)^{mp} \cdot 2\exp(-\tau^2) + 2\exp(-\tau^2) \leq 4 \cdot (3/\epsilon)^{mp} \cdot \exp(-\tau^2)$

$$\inf_{\mathbf{W},\mathbf{v}} \Delta\Phi_n(\mathbf{W},\mathbf{v},t) \geq \inf_{\mathbf{W},\mathbf{v}} \mathbb{E}\{\Delta\Phi_n(\mathbf{W},\mathbf{v},t)\} - \left(c(t) \cdot \frac{\tau}{\sqrt{n}} + L \cdot \left(1 + \frac{4\tau}{\sqrt{n}}\right) \cdot \epsilon\right)$$

Now we set $\tau \triangleq \sqrt{mp \log n}$, and $\epsilon \triangleq \frac{\tau}{\sqrt{n}} = \sqrt{mp\frac{\log n}{n}}$. Under the assumption that

$$\frac{n}{\log n} \geq mp$$

we check that $\tau \leq \sqrt{n}$, $\epsilon \leq 1$, hence, the probability bound holds. We estimate the probability bound with :

$$\begin{aligned} \log\frac{3}{\epsilon} &= \log\frac{3}{\sqrt{mp}} + \log\sqrt{\frac{n}{\log n}} \leq \frac{1}{2}\log\frac{9}{mp} + \frac{1}{2}\log n \\ mp\log\frac{3}{\epsilon} - \tau^2 &\leq \frac{mp}{2}\log\frac{9}{mp} + \frac{mp}{2}\log n - mp\log n = \frac{mp}{2}\log\frac{9}{mp} - \frac{mp}{2}\log n \\ (3/\epsilon)^{mp}\exp(-\tau^2) &\leq \left(\frac{mpn}{9}\right)^{-mp/2}. \end{aligned}$$

Finally, recalling that

$$\begin{aligned} L &\triangleq 30 C_t^3 \cdot t \cdot \left[5(k\sigma_\alpha^2 + m\sigma^2) + \lambda^2 k\right] \\ c(t) &\triangleq 102 \cdot \left(t^2\sigma_\alpha^2 + 2m\sigma^2 + 2\lambda k\sigma_\alpha\right), \end{aligned}$$

and since the assumption $C_t \leq 1.5$ implies $150C_t^3 \leq 507$, we obtain

$$\begin{aligned} c(t) + L &\leq 609\left(k\sigma_\alpha^2 \cdot t + 2m\sigma^2 + \lambda k\sigma_\alpha + \lambda^2 k \cdot t\right) \triangleq B/5 \\ c(t) \cdot \frac{\tau}{\sqrt{n}} + L \cdot \left(1 + \frac{4\tau}{\sqrt{n}}\right) \cdot \epsilon &\leq (c(t)+L) \cdot \epsilon + 4(c(t)+L)\epsilon^2 = (c(t)+L)5\epsilon \leq B\epsilon. \end{aligned}$$

E.5 Proof of Lemma 11

First, we observe that by the statistical independence between $\boldsymbol{\alpha}$ and $\boldsymbol{\varepsilon}$ we have

$$\mathbb{E}\{\zeta_{\boldsymbol{\alpha},\boldsymbol{\varepsilon}}(t)\} = \mathbb{E}\{\zeta_{\mathbf{s},\boldsymbol{\varepsilon}}(t)\} = 0.$$

Moreover, we can rewrite

$$\begin{aligned} \zeta_{\boldsymbol{\alpha},\boldsymbol{\alpha}}(t) &= \frac{1}{2} \cdot \mathrm{Tr}\left([\boldsymbol{\alpha}_0]_{\mathrm{J}}[\boldsymbol{\alpha}_0]_{\mathrm{J}}^\top \cdot [\mathbf{D}_0]_{\mathrm{J}}^\top (\mathbf{P}_{\mathrm{J}}(0) - \mathbf{P}_{\mathrm{J}}(t))[\mathbf{D}_0]_{\mathrm{J}}\right) \\ \zeta_{\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}}(t) &= \frac{1}{2} \cdot \mathrm{Tr}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \cdot (\mathbf{P}_{\mathrm{J}}(0) - \mathbf{P}_{\mathrm{J}}(t))\right) \\ \zeta_{\mathbf{s},\boldsymbol{\alpha}}(t) &= -\lambda \cdot \mathrm{Tr}\left([\boldsymbol{\alpha}_0]_{\mathrm{J}}\,\mathrm{sign}(\boldsymbol{\alpha}_0)_{\mathrm{J}}^\top \cdot \left[\boldsymbol{\Theta}_{\mathrm{J}}(0)[\mathbf{D}_0]_{\mathrm{J}}^\top - \boldsymbol{\Theta}_{\mathrm{J}}(t)[\mathbf{D}(t)]_{\mathrm{J}}^\top\right][\mathbf{D}_0]_{\mathrm{J}}\right) \\ \zeta_{\mathbf{s},\mathbf{s}}(t) &= \frac{\lambda^2}{2} \cdot \mathrm{Tr}\left(\boldsymbol{\Theta}_{\mathrm{J}}(0) - \boldsymbol{\Theta}_{\mathrm{J}}(t)\right). \end{aligned}$$

Since the coefficients $\boldsymbol{\alpha}, \boldsymbol{\varepsilon}$ are independent from the support J we obtain

$$\mathbb{E}\{\zeta_{\boldsymbol{\alpha},\boldsymbol{\alpha}}(t)\} = \frac{\mathbb{E}\{\alpha^2\}}{2} \cdot \mathbb{E}_J\left\{\mathrm{Tr}\left([\mathbf{D}_0]_{\mathrm{J}}^\top (\mathbf{I} - \mathbf{P}_{\mathrm{J}}(t))[\mathbf{D}_0]_{\mathrm{J}}\right)\right\} \tag{64}$$

$$\mathbb{E}\{\zeta_{\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}}(t)\} = \frac{\mathbb{E}\{\varepsilon^2\}}{2} \cdot \mathbb{E}_J\left\{\mathrm{Tr}\left(\mathbf{P}_{\mathrm{J}}(0) - \mathbf{P}_{\mathrm{J}}(t)\right)\right\} = 0 \tag{65}$$

$$\mathbb{E}\{\zeta_{\mathbf{s},\boldsymbol{\alpha}}(t)\} = -\lambda \cdot \mathbb{E}\{|\alpha|\} \cdot \mathbb{E}_J\left\{\mathrm{Tr}\left(\left[\boldsymbol{\Theta}_{\mathrm{J}}(0)[\mathbf{D}_0]_{\mathrm{J}}^\top - \boldsymbol{\Theta}_{\mathrm{J}}(t)[\mathbf{D}(t)]_{\mathrm{J}}^\top\right][\mathbf{D}_0]_{\mathrm{J}}\right)\right\} \tag{66}$$

$$\mathbb{E}\{\zeta_{\mathbf{s},\mathbf{s}}(t)\} = \frac{\lambda^2}{2} \cdot \mathbb{E}_J\left\{\mathrm{Tr}\left(\boldsymbol{\Theta}_{\mathrm{J}}(0) - \boldsymbol{\Theta}_{\mathrm{J}}(t)\right)\right\} \tag{67}$$

where we used the fact that: (a) $\mathbf{P}_{\mathrm{J}}(0)\mathbf{D}_0 = \mathbf{D}_0$; (b) since $\mathbf{P}_{\mathrm{J}}(t)$ is an orthogonal projector onto a subspace of dimension k, $\mathrm{Tr}(\mathbf{P}_{\mathrm{J}}(0) - \mathbf{P}_{\mathrm{J}}(t)) = k - k = 0$.

The lemma below provide estimates of the remaining non-vanishing expectations which come up in the quadratic forms (56) and (61) and the bilinear form (59). They directly provide Lemma 11 as a corollary.

Lemma 14. *If $k\mu(t) < 1/2$ then for any $\mathbf{W} \in \mathcal{W}_{\mathbf{D}_0}$, $\mathbf{v} \in \mathcal{S}^p$ we have*

$$\mathbb{E}_{\mathrm{J}}\left\{\mathrm{Tr}\left([\mathbf{D}_0]_{\mathrm{J}}^\top (\mathbf{I} - \mathbf{P}_{\mathrm{J}}(t))[\mathbf{D}_0]_{\mathrm{J}}\right)\right\} \geq (1 - \mathcal{K}^2) \cdot \frac{k}{p} t^2 \tag{68}$$

$$\left|\mathbb{E}_{\mathrm{J}}\left\{\mathrm{Tr}\left(\left[\boldsymbol{\Theta}_{\mathrm{J}}(0)[\mathbf{D}_0]_{\mathrm{J}}^\top - \boldsymbol{\Theta}_{\mathrm{J}}(t)[\mathbf{D}(t)_{\mathrm{J}}]^\top\right][\mathbf{D}_0]_{\mathrm{J}}\right)\right\}\right| \leq 3Q_t^2 \cdot t \cdot \frac{k}{p} \cdot \|\mathbf{D}_0\|_2 \cdot k\mu(t) \tag{69}$$

$$\left|\mathbb{E}_{\mathrm{J}}\left\{\mathrm{Tr}\left(\boldsymbol{\Theta}_{\mathrm{J}}(0) - \boldsymbol{\Theta}_{\mathrm{J}}(t)\right)\right\}\right| \leq 8Q_t^4 \cdot t \cdot \frac{k}{p} \cdot \|\mathbf{D}_0\|_2 \cdot k\mu(t). \tag{70}$$

with $\mathcal{K} \triangleq C_t \cdot (\|\mathbf{D}_0\|_2 \cdot \sqrt{k/p} + t)$.

Proof of Lemma 14 - Equation (68). Since $k\mu(t) < 1/2$, we have $t < 1/(6k) \leq 1/6$ and $\delta_k(\mathbf{D}_0) \leq k\mu_0 < 1/2$, so that $t < \sqrt{1 - \delta_k(\mathbf{D}_0)}$. In particular, we have $t < \pi/2$ and the matrix $\mathbf{C} = \mathrm{Diag}(\cos(\mathbf{v}_j t))$ is invertible. From the equality $\mathbf{D}(t) = \mathbf{D}_0\mathbf{C} + \mathbf{W}\mathbf{S}$ with $\mathbf{S} = \mathrm{Diag}(\cos(\mathbf{v}_j t))$ we deduce $\mathbf{D}_0 = \mathbf{D}(t)\mathbf{C}^{-1} - \mathbf{W}\mathbf{T}$ with $\mathbf{T} = \mathrm{Diag}(\tan(\mathbf{v}_j t))$. Since the columns of $[\mathbf{D}(t)\mathbf{C}^{-1}]_{\mathrm{J}}$ belong to the span of $[\mathbf{D}(t)]_{\mathrm{J}}$ we obtain

$$\begin{aligned} \mathrm{Tr}\big([\mathbf{D}_0]_{\mathrm{J}}^\top (\mathbf{I} - \mathbf{P}_{\mathrm{J}}(t))[\mathbf{D}_0]_{\mathrm{J}}\big) &= \|(\mathbf{I} - \mathbf{P}_{\mathrm{J}}(t))[\mathbf{D}_0]_{\mathrm{J}}\|_{\mathrm{F}}^2 = \|(\mathbf{I} - \mathbf{P}_{\mathrm{J}}(t))[\mathbf{W}\mathbf{T}]_{\mathrm{J}}\|_{\mathrm{F}}^2 \\ &= \|[\mathbf{W}\mathbf{T}]_{\mathrm{J}}\|_{\mathrm{F}}^2 - \|\mathbf{P}_{\mathrm{J}}(t)[\mathbf{W}\mathbf{T}]_{\mathrm{J}}\|_{\mathrm{F}}^2. \end{aligned}$$

For the first term, since $\|\mathbf{w}^j\|_2 = 1$, we have

$$\|[\mathbf{WT}]_{\mathrm{J}}\|_{\mathrm{F}}^2 = \sum_{j=1}^{p} \delta(j) \cdot \|\mathbf{w}^j\|_2^2 \cdot \tan^2(\mathbf{v}_j t) = \sum_{j=1}^{p} \delta(j) \cdot \tan^2(\mathbf{v}_j t)$$

hence, since $\|\mathbf{v}\|_2 = 1$, and $\tan^2(u) \geq u^2$ for $|u| \leq 1$ we have

$$\mathbb{E}\ \|[\mathbf{WT}]_{\mathrm{J}}\|_{\mathrm{F}}^2 = \frac{k}{p} \cdot \|\mathbf{WT}\|_{\mathrm{F}}^2 = \frac{k}{p} \cdot \sum_{j=1}^{p} \tan^2(\mathbf{v}_j t)^2 \geq \frac{k}{p} \cdot \sum_{j=1}^{p} t^2 \mathbf{v}_j^2 = \frac{k}{p} \cdot t^2.$$

For the second term, since $\mathbf{P}_{\mathrm{J}}(t) = \mathbf{D}_{\mathrm{J}}(t)\mathbf{\Theta}_{\mathrm{J}}(t)\mathbf{D}_{\mathrm{J}}^{\top}(t)$, using Lemma 5, we have the bound

$$\|\mathbf{P}_{\mathrm{J}}(t)[\mathbf{WT}]_{\mathrm{J}}\|_{\mathrm{F}}^2 \quad \leq \quad C_t^2 \cdot \|\mathbf{D}_{\mathrm{J}}^{\top}(t)[\mathbf{WT}]_{\mathrm{J}}\|_{\mathrm{F}}^2,$$

Moreover, by Lemma 4, using the Cauchy-Schwarz inequality for random variables

$$\|\mathbf{D}_{\mathrm{J}}^{\top}(t)[\mathbf{WT}]_{\mathrm{J}}\|_{\mathrm{F}} \quad \leq \quad \|[\mathbf{D}_0]_{\mathrm{J}}^{\top}[\mathbf{WT}]_{\mathrm{J}}\|_{\mathrm{F}} + \|[\mathbf{D}(t) - \mathbf{D}_0]_{\mathrm{J}}^{\top}[\mathbf{WT}]_{\mathrm{J}}\|_{\mathrm{F}} \leq \|[\mathbf{D}_0]_{\mathrm{J}}^{\top}[\mathbf{WT}]_{\mathrm{J}}\|_{\mathrm{F}} + t \cdot \|[\mathbf{WT}]_{\mathrm{J}}\|_{\mathrm{F}},$$

$$\begin{aligned}
\mathbb{E}\{\|\mathbf{D}_{\mathrm{J}}^{\top}(t)[\mathbf{WT}]_{\mathrm{J}}\|_{\mathrm{F}}^2\} &\leq \mathbb{E}\left\{\|[\mathbf{D}_0]_{\mathrm{J}}^{\top}[\mathbf{WT}]_{\mathrm{J}}\|_{\mathrm{F}}^2\right\} + 2t \cdot \mathbb{E}\left\{\|[\mathbf{D}_0]_{\mathrm{J}}^{\top}[\mathbf{WT}]_{\mathrm{J}}\|_{\mathrm{F}} \cdot \|[\mathbf{WT}]_{\mathrm{J}}\|_{\mathrm{F}}\right\} + t^2 \cdot \mathbb{E}\left\{\|[\mathbf{WT}]_{\mathrm{J}}\|_{\mathrm{F}}^2\right\} \\
&\leq \mathbb{E}\left\{\|[\mathbf{D}_0]_{\mathrm{J}}^{\top}[\mathbf{WT}]_{\mathrm{J}}\|_{\mathrm{F}}^2\right\} + 2t \cdot \sqrt{\mathbb{E}\left\{\|[\mathbf{D}_0]_{\mathrm{J}}^{\top}[\mathbf{WT}]_{\mathrm{J}}\|_{\mathrm{F}}^2\right\}} \cdot \sqrt{\mathbb{E}\left\{\|[\mathbf{WT}]_{\mathrm{J}}\|_{\mathrm{F}}^2\right\}} \\
&\quad + t^2 \cdot \mathbb{E}\left\{\|[\mathbf{WT}]_{\mathrm{J}}\|_{\mathrm{F}}^2\right\} \\
&\leq \left(\sqrt{\mathbb{E}\left\{\|[\mathbf{D}_0]_{\mathrm{J}}^{\top}[\mathbf{WT}]_{\mathrm{J}}\|_{\mathrm{F}}^2\right\}} + t \cdot \sqrt{\frac{k}{p}} \cdot \|\mathbf{WT}\|_{\mathrm{F}}\right)^2
\end{aligned}$$

Now, proceeding as in Lemma 7, we compute

$$\mathbb{E}\Big[\|[\mathbf{D}_0]_{\mathrm{J}}^{\top}[\mathbf{WT}]_{\mathrm{J}}\|_{\mathrm{F}}^2\Big] \quad = \quad \frac{k(k-1)}{p(p-1)} \cdot \|\mathbf{D}_0^{\top}\mathbf{WT}\|_{\mathrm{F}}^2 \leq \frac{k(k-1)}{p(p-1)} \cdot \|\mathbf{D}_0^{\top}\|_2^2 \cdot \|\mathbf{WT}\|_{\mathrm{F}}^2$$

hence

$$\|\mathbf{D}_{\mathrm{J}}^{\top}(t)[\mathbf{WT}]_{\mathrm{J}}\|_{\mathrm{F}} \leq \frac{k}{p} \cdot \|\mathbf{WT}\|_{\mathrm{F}}^2 \cdot \left(\sqrt{\frac{k-1}{p-1}} \cdot \|\mathbf{D}_0^{\top}\|_2 + t\right)^2 \leq \frac{k}{p} \cdot \|\mathbf{WT}\|_{\mathrm{F}}^2 \cdot \left(\sqrt{\frac{k}{p}} \cdot \|\mathbf{D}_0\|_2 + t\right)^2.$$

Putting the pieces together, we obtain the lower bound

$$\begin{aligned}
\mathrm{Tr}\big([\mathbf{D}_0]_{\mathrm{J}}^{\top}(\mathbf{I} - \mathbf{P}_{\mathrm{J}}(t))[\mathbf{D}_0]_{\mathrm{J}}\big) &\geq \frac{k}{p} \cdot \|\mathbf{WT}\|_{\mathrm{F}}^2 \cdot \left(1 - \left\{C_t \cdot \left(\|\mathbf{D}_0\|_2 \cdot \sqrt{\frac{k}{p}} + t\right)\right\}^2\right) \\
&= \frac{k}{p} \cdot \|\mathbf{WT}\|_{\mathrm{F}}^2 \cdot \left(1 - \mathcal{K}^2\right).
\end{aligned}$$

□

Proof of Lemma 14 - Equation (69). We first develop Equation (69) and use that $\mathbf{\Theta}_{\mathrm{J}}(0)[\mathbf{D}_0]_{\mathrm{J}}^{\top}[\mathbf{D}_0]_{\mathrm{J}} = \mathbf{I}$ in order to obtain

$$\mathrm{Tr}\left(\left[\mathbf{\Theta}_{\mathrm{J}}(0)[\mathbf{D}_0]_{\mathrm{J}}^{\top} - \mathbf{\Theta}_{\mathrm{J}}(t)\mathbf{D}(t)_{\mathrm{J}}^{\top}(t)\right][\mathbf{D}_0]_{\mathrm{J}}\right) = k - \mathrm{Tr}\left(\mathbf{\Theta}_{\mathrm{J}}(t)[\mathbf{D}(t)]_{\mathrm{J}}^{\top}[\mathbf{D}_0]_{\mathrm{J}}\right).$$

Appyling Lemma 1, we know there exists $\mathbf{W}_t \in \mathcal{W}_{\mathbf{D}(t)}$ such that

$$\mathbf{D}_0 = \mathbf{D}(t)\mathrm{Diag}(\cos(\mathbf{v}_j t)) + \mathbf{W}_t\mathrm{Diag}(\sin(\mathbf{v}_j t)),$$

and the trace above further simplifies as

$$\begin{aligned}\mathrm{Tr}\left(\left[\boldsymbol{\Theta}_{\mathrm{J}}(0)[\mathbf{D}_0]_{\mathrm{J}}^\top - \boldsymbol{\Theta}_{\mathrm{J}}(t)[\mathbf{D}(t)]_{\mathrm{J}}^\top\right][\mathbf{D}_0]_{\mathrm{J}}\right) &= k - \sum_{j\in\mathrm{J}}\cos(\mathbf{v}_j t) - \mathrm{Tr}\left(\boldsymbol{\Theta}_{\mathrm{J}}(t)[\mathbf{D}(t)]_{\mathrm{J}}^\top[\mathbf{W}_t\mathbf{S}(t)]_{\mathrm{J}}\right),\\ &= \sum_{j\in\mathrm{J}}(1-\cos(\mathbf{v}_j t)) - \mathrm{Tr}\left(\boldsymbol{\Theta}_{\mathrm{J}}(t)[\mathbf{D}(t)]_{\mathrm{J}}^\top[\mathbf{W}_t\mathbf{S}(t)]_{\mathrm{J}}\right),\end{aligned}$$

where for short, we refer to $\mathrm{Diag}(\sin(\mathbf{v}_j t))$ as $\mathbf{S}(t)$.

The first term is simple to handle since we have

$$\mathbb{E}_{\mathrm{J}}\big[\sum_{j\in\mathrm{J}}(1-\cos(\mathbf{v}_j t))\big] \leq \frac{t^2}{2}\mathbb{E}_{\mathrm{J}}[\|\mathbf{v}_{\mathrm{J}}\|_2^2] = \frac{t^2}{2}\frac{k}{p}.$$

We now turn to the second term whose control is more involved. Following Geng et al. [2011], we introduce the self-adjoint operator $\Gamma_{\mathbf{D}(t)}$ defined for any $\mathbf{M}\in\mathbb{R}^{m\times p}$ by

$$\Gamma_{\mathbf{D}(t)}(\mathbf{M}) \triangleq \Big[\Gamma_1(t)\,\mathbf{m}^1,\ldots,\Gamma_p(t)\,\mathbf{m}^p\Big], \quad \text{with } \Gamma_j(t)\triangleq \mathbf{I}-\mathbf{d}(t)^j[\mathbf{d}(t)^j]^\top.$$

In words, $\Gamma_{\mathbf{D}(t)}(\mathbf{M})$ projects each column of $\mathbf{M}$ onto the orthogonal complement of the corresponding column of the dictionary $\mathbf{D}(t)$. In particular, note that for any $\mathbf{M}\in\mathcal{W}_{\mathbf{D}(t)}$, we therefore have $\Gamma_{\mathbf{D}(t)}(\mathbf{M})=\mathbf{M}$. Considering the symmetric matrix $\mathbf{U}(t)=\mathbb{E}_{\mathrm{J}}\big[\Pi_{\mathrm{J}}\boldsymbol{\Theta}_{\mathrm{J}}(t)\Pi_{\mathrm{J}}^\top\big]$, we next obtain

$$\begin{aligned}\mathbb{E}_{\mathrm{J}}\big[\mathrm{Tr}\left(\boldsymbol{\Theta}_{\mathrm{J}}(t)[\mathbf{D}(t)]_{\mathrm{J}}^\top[\mathbf{W}_t\mathbf{S}(t)]_{\mathrm{J}}\right)\big] &= \mathbb{E}_{\mathrm{J}}\big[\mathrm{Tr}\left(\Pi_{\mathrm{J}}\boldsymbol{\Theta}_{\mathrm{J}}(t)\Pi_{\mathrm{J}}^\top[\mathbf{D}(t)]^\top\mathbf{W}_t\mathbf{S}(t)\right)\big]\\ &= \mathrm{Tr}\left(\mathbf{U}(t)[\mathbf{D}(t)]^\top\mathbf{W}_t\mathbf{S}(t)\right)\\ &= \mathrm{Tr}\left((\mathbf{D}(t)\mathbf{U}(t))^\top\Gamma_{\mathbf{D}(t)}(\mathbf{W}_t\mathbf{S}(t))\right)\\ &= \mathrm{Tr}\left(\Gamma_{\mathbf{D}(t)}(\mathbf{D}(t)\mathbf{U}(t))\mathbf{W}_t\mathbf{S}(t)\right)\\ &\leq t\|\Gamma_{\mathbf{D}(t)}(\mathbf{D}(t)\mathbf{U}(t))\|_{\mathrm{F}},\end{aligned}$$

where we have successively used the fact that $\Gamma_{\mathbf{D}(t)}$ is self-adjoint and that for any $\mathbf{W}\in\mathcal{W}_{\mathbf{D}(t)}$, the norm $\|\mathbf{W}_t\mathbf{S}(t)\|_{\mathrm{F}}$ is upper bounded by t.

Observe that the j-th column of the matrix $\Gamma_j(t)\,\mathbf{D}$ is equal to zero. As a consequence, we have

$$\|\Gamma_{\mathbf{D}(t)}(\mathbf{D}(t)\mathbf{U}(t))\|_{\mathrm{F}} = \|\Gamma_{\mathbf{D}(t)}(\mathbf{D}(t)\mathbf{U}_{\mathrm{off}}(t))\|_{\mathrm{F}},$$

where $\mathbf{U}_{\mathrm{off}}(t)$ denotes the matrix $\mathbf{U}(t)$ with its diagonal terms set to zero. This leads to

$$\begin{aligned}\|\Gamma_{\mathbf{D}(t)}(\mathbf{D}(t)\mathbf{U}(t))\|_{\mathrm{F}}^2 &= \|\Gamma_{\mathbf{D}(t)}(\mathbf{D}(t)\mathbf{U}_{\mathrm{off}}(t))\|_{\mathrm{F}}^2 = \sum_{j=1}^p\|\Gamma_j(t)\,\mathbf{D}(t)\mathbf{u}_{\mathrm{off}}^j\|_2^2\\ &\leq \|\mathbf{D}(t)\|_2^2\sum_{j=1}^p\|\mathbf{u}_{\mathrm{off}}^j\|_2^2 = \|\mathbf{D}(t)\|_2^2\|\mathbb{E}_{\mathrm{J}}\big[\Pi_{\mathrm{J}}\boldsymbol{\Theta}_{\mathrm{J}}(t)\Pi_{\mathrm{J}}^\top\big]_{\mathrm{off}}\|_{\mathrm{F}}^2,\end{aligned}$$

where we have exploited the fact that projectors have their spectral norms bounded by one. Using Corollary 1, we have for $i\neq j$ with $i,j\in[\![1;p]\!]$

$$|\big[\Pi_{\mathrm{J}}\boldsymbol{\Theta}_{\mathrm{J}}(t)\Pi_{\mathrm{J}}^\top\big]_{i,j}| \leq \delta(i)\delta(j)\frac{\mu(t)}{1-k\mu(t)}$$

and

$$|\mathbb{E}_{\mathrm{J}}\big[(\Pi_{\mathrm{J}}\boldsymbol{\Theta}_{\mathrm{J}}(t)\Pi_{\mathrm{J}}^\top)_{i,j}\big]| \leq \frac{k(k-1)}{p(p-1)}\frac{\mu(t)}{1-k\mu(t)},$$

hence

$$\|\mathbb{E}_{\mathrm{J}}\big[\Pi_{\mathrm{J}}\boldsymbol{\Theta}_{\mathrm{J}}(t)\Pi_{\mathrm{J}}^{\top}\big]_{\mathrm{off}}\|_{\mathrm{F}}^2 \leq p(p-1)\Big(\frac{k(k-1)}{p(p-1)}\frac{\mu(t)}{1-k\mu(t)}\Big)^2 \leq \frac{k^2}{p^2}\frac{(k\mu(t))^2}{(1-k\mu(t))^2}.$$

To recapitulate and putting all the pieces together, we obtain the following upper bound

$$\begin{aligned}\big|\mathbb{E}_{\mathrm{J}}\left\{\mathrm{Tr}\left(\big[\boldsymbol{\Theta}_{\mathrm{J}}(0)[\mathbf{D}_0]_{\mathrm{J}}^{\top} - \boldsymbol{\Theta}_{\mathrm{J}}(t)[\mathbf{D}(t)_{\mathrm{J}}]^{\top}\big][\mathbf{D}_0]_{\mathrm{J}}\right)\right\}\big| &\leq \frac{t^2}{2}\frac{k}{p} + t|||\mathbf{D}(t)|||_2\frac{k}{p}\frac{k\mu(t)}{1-k\mu(t)} \\ &\leq t\frac{k}{p}\Big[\frac{t}{2} + |||\mathbf{D}(t)|||_2\frac{k\mu(t)}{1-k\mu(t)}\Big].\end{aligned}$$

To conclude, we use Lemma 4 to get $|||\mathbf{D}(t)|||_2 \leq 2|||\mathbf{D}_0|||_2$, and the fact that $|||\mathbf{D}_0|||_2 \geq 1$. □

Proof of Lemma 14 - Equation (70). We start by writting Equation (70) in the following integral form

$$\mathrm{Tr}(\boldsymbol{\Theta}_{\mathrm{J}}(t) - \boldsymbol{\Theta}_{\mathrm{J}}(0)) = \int_0^t \mathrm{Tr}(\nabla_t\boldsymbol{\Theta}_{\mathrm{J}}(\tau))d\tau,$$

where the derivative is computed in Lemma 17, namely,

$$\mathrm{Tr}(\nabla_t\boldsymbol{\Theta}_{\mathrm{J}}(t)) = -2\mathrm{Tr}\Big(\boldsymbol{\Theta}_{\mathrm{J}}(t)[\nabla_t\mathbf{D}(t)]_{\mathrm{J}}^{\top}\mathbf{D}_{\mathrm{J}}(t)\boldsymbol{\Theta}_{\mathrm{J}}(t)\Big).$$

Introducing the symmetric matrix $\mathbf{U}(t) = \mathbb{E}_{\mathrm{J}}\big[\Pi_{\mathrm{J}}[\boldsymbol{\Theta}_{\mathrm{J}}(t)]^2\Pi_{\mathrm{J}}^{\top}\big]$, we next obtain by linearity of the trace and the integral

$$\mathbb{E}_{\mathrm{J}}[\mathrm{Tr}(\boldsymbol{\Theta}_{\mathrm{J}}(t) - \boldsymbol{\Theta}_{\mathrm{J}}(0))] = -2\int_0^t \mathrm{Tr}\Big(\mathbf{D}(\tau)\mathbf{U}(\tau)[\nabla_t\mathbf{D}(\tau)]^{\top}\Big)d\tau \leq 2t\max_{\tau\in[0,t]}\Big|\mathrm{Tr}\Big(\mathbf{D}(\tau)\mathbf{U}(\tau)[\nabla_t\mathbf{D}(\tau)]^{\top}\Big)\Big|.$$

Noticing that we are (almost) in the same setting as that of the previous proof, we are going to make use again of the operator $\Gamma_{\mathbf{D}(t)}$ in order to control the off-diagonal terms of $\mathbf{U}(t)$. More precisely, since $\mathrm{diag}([\nabla_t\mathbf{D}(\tau)]^{\top}\mathbf{D}(\tau)) = \mathbf{0}$ and $\|\nabla_t\mathbf{D}(\tau)\|_{\mathrm{F}} = 1$, the same reasoning as that followed in the previous proof leads to

$$\mathbb{E}_{\mathrm{J}}[\mathrm{Tr}(\boldsymbol{\Theta}_{\mathrm{J}}(t) - \boldsymbol{\Theta}_{\mathrm{J}}(0))] \leq 2t\cdot\max_{\tau\in[0,t]}|||\mathbf{D}(\tau)|||_2\cdot\|\mathbb{E}_{\mathrm{J}}\big[\Pi_{\mathrm{J}}[\boldsymbol{\Theta}_{\mathrm{J}}(\tau)]^2\Pi_{\mathrm{J}}^{\top}\big]_{\mathrm{off}}\|_{\mathrm{F}}.$$

Invoking Corollary 1, we have for $i \neq j$ with $i, j \in [\![1; p]\!]$

$$|\mathbb{E}_{\mathrm{J}}\big[(\Pi_{\mathrm{J}}[\boldsymbol{\Theta}_{\mathrm{J}}(\tau)]^2\Pi_{\mathrm{J}}^{\top})_{i,j}\big]| \leq \frac{k(k-1)}{p(p-1)}\frac{2\mu(t)}{(1-k\mu(t))^2},$$

hence

$$\|\mathbb{E}_{\mathrm{J}}\big[\Pi_{\mathrm{J}}[\boldsymbol{\Theta}_{\mathrm{J}}(\tau)]^2\Pi_{\mathrm{J}}^{\top}\big]_{\mathrm{off}}\|_{\mathrm{F}}^2 \leq p(p-1)\Big(\frac{k(k-1)}{p(p-1)}\frac{2\mu(t)}{(1-k\mu(t))^2}\Big)^2 \leq \frac{k^2}{p^2}\frac{(2k\mu(t))^2}{(1-k\mu(t))^4},$$

which gives the advertised conclusion. □

F Proof of Proposition 3

We begin by a few lemmata related to the considered optimization problem.

Lemma 15. *Let* $\mathrm{J} \subseteq [\![1; p]\!]$ *and* $\mathbf{s} \in \{-1, 0, 1\}^{|\mathrm{J}|}$. *Consider a dictionary* $\mathbf{D} \in \mathbb{R}^{m\times p}$ *such that* $\mathbf{D}_{\mathrm{J}}^{\top}\mathbf{D}_{\mathrm{J}}$ *is invertible. Consider also the vector* $\boldsymbol{\alpha} \in \mathbb{R}^p$ *defined by*

$$\boldsymbol{\alpha} = \begin{pmatrix}[\mathbf{D}_{\mathrm{J}}^{\top}\mathbf{D}_{\mathrm{J}}]^{-1}[\mathbf{D}_{\mathrm{J}}^{\top}\mathbf{x} - \lambda\mathbf{s}] \\ \mathbf{0}_{\mathrm{J}^c}\end{pmatrix},$$

with $\mathbf{x} \in \mathbb{R}^m$ *and* λ *a nonnegative scalar. If* $\mathbf{x} = [\mathbf{D}_0]_{\mathrm{J}}[\boldsymbol{\alpha}_0]_{\mathrm{J}} + \boldsymbol{\varepsilon}$ *for some* $(\mathbf{D}_0, \boldsymbol{\alpha}_0, \boldsymbol{\varepsilon}) \in \mathbb{R}^{m\times p}\times\mathbb{R}^p\times\mathbb{R}^m$, *then we have*

$$\|[\boldsymbol{\alpha} - \boldsymbol{\alpha}_0]_{\mathrm{J}}\|_\infty \leq |||[\mathbf{D}_{\mathrm{J}}^{\top}\mathbf{D}_{\mathrm{J}}]^{-1}|||_\infty\Big[\lambda + \|\mathbf{D}_{\mathrm{J}}^{\top}\left(\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}_0\right)\|_\infty\Big].$$

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

Lemma 16. *Let $\mathbf{x} \in \mathbb{R}^m$ be a signal. Consider $\mathrm{J} \subseteq [\![1;p]\!]$ and a dictionary $\mathbf{D} \in \mathbb{R}^{m\times p}$ such that $\mathbf{D}_\mathrm{J}^\top \mathbf{D}_\mathrm{J}$ is invertible. Consider also a sign vector $\mathbf{s} \in \{-1,1\}^{|\mathrm{J}|}$ and define $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^p$ by*

$$\hat{\boldsymbol{\alpha}} = \begin{pmatrix} [\mathbf{D}_\mathrm{J}^\top \mathbf{D}_\mathrm{J}]^{-1}[\mathbf{D}_\mathrm{J}^\top \mathbf{x} - \lambda \mathbf{s}] \\ \mathbf{0}_{\mathrm{J}^c} \end{pmatrix},$$

for some regularization parameter $\lambda \geq 0$. If the following two conditions hold

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

then $\hat{\boldsymbol{\alpha}}$ 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 $\mathrm{sign}(\hat{\boldsymbol{\alpha}}_\mathrm{J}) = \mathbf{s}$.

Proof. We first check that $\hat{\boldsymbol{\alpha}}$ is a solution of the Lasso program. It is well-known [e.g., see Fuchs, 2005, Wainwright, 2009] 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 $\mathbf{z}_j = \mathrm{sign}(\hat{\boldsymbol{\alpha}}_j)$ if $\hat{\boldsymbol{\alpha}}_j \neq 0$, and $|\mathbf{z}_j| \leq 1$ 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

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

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 Lemma 1 in Wainwright [2009]. □

Corollary 3. *Assume that $k\mu(t) \leq 1/2$, $0 \leq t' \leq t$, $\frac{9}{4}\lambda \leq \underline{\alpha} \leq \min_{j\in\mathrm{J}} |[\boldsymbol{\alpha}_0]_j|$, and that*

$$\begin{aligned} \|[\mathbf{D}(t')]_\mathrm{J}^\top (\mathbf{x} - \mathbf{D}(t')\boldsymbol{\alpha}_0)\|_\infty &< \lambda(2 - Q_t^2) && (71) \\ \|[\mathbf{D}(t')]_{\mathrm{J}^c}^\top (\mathbf{I} - \mathbf{P}_\mathrm{J}(t'))\mathbf{x}\|_\infty &< \lambda(2 - Q_t^2) && (72) \end{aligned}$$

Then $\hat{\boldsymbol{\alpha}}(t')$ is the unique solution of $\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} [\frac{1}{2}\|\mathbf{x} - \mathbf{D}(t')\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1]$

Proof. Since $k\mu(t) \leq 1/2$, we have $Q_t^2 \leq 2$, and by Corollary 2 we have, uniformly for all $(\mathbf{W}, \mathbf{v})$ and $0 \leq t' \leq t$

$$\begin{aligned} |||\big[[\mathbf{D}(t')]_\mathrm{J}^\top [\mathbf{D}(t')]_\mathrm{J}\big]^{-1}|||_\infty &\leq Q_t^2 \\ |||[\mathbf{D}(t')]_{\mathrm{J}^c}^\top [\mathbf{D}(t')]_\mathrm{J} \big([\mathbf{D}(t')]_\mathrm{J}^\top [\mathbf{D}(t')]_\mathrm{J}\big)^{-1}|||_\infty &\leq Q_t^2 - 1 \leq 1 \end{aligned}$$

Exploiting Lemma 15 and the bound (71) we have

$$\begin{aligned} \|[\hat{\boldsymbol{\alpha}}(t') - \boldsymbol{\alpha}_0]_\mathrm{J}\|_\infty &\leq |||\big[[\mathbf{D}(t')]_\mathrm{J}^\top [\mathbf{D}(t')]_\mathrm{J}\big]^{-1}|||_\infty \Big[\lambda + \|[\mathbf{D}(t')]_\mathrm{J}^\top (\mathbf{x} - \mathbf{D}(t')\boldsymbol{\alpha}_0)\|_\infty\Big] \\ &< Q_t^2 \cdot \lambda \cdot \big[1 + (2 - Q_t^2)\big] = \lambda \cdot Q_t^2 \cdot (3 - Q_t^2) \leq \frac{9}{4}\lambda \leq \underline{\alpha} \leq \min_{j\in\mathrm{J}} \big|[\boldsymbol{\alpha}_0]_j\big|, \end{aligned}$$

where we used that $u(3-u) \leq 9/4$ for all $u \in \mathbb{R}$. We conclude that $\mathrm{sign}(\hat{\boldsymbol{\alpha}}(t')) = \mathrm{sign}(\boldsymbol{\alpha}_0)$.

It remains to prove that $\hat{\boldsymbol{\alpha}}(t')$ is the unique solution of the Lasso program. To this end, we take advantage of Lemma 16. We recall the quantity which needs to be smaller than λ

$$\|[\mathbf{D}(t')]_{\mathrm{J}^c}^\top (\mathbf{I} - \mathbf{P}_\mathrm{J}(t'))\mathbf{x}\|_\infty + \lambda |||[\mathbf{D}(t')_{\mathrm{J}^c}]^\top [\mathbf{D}(t')]_\mathrm{J} \big([\mathbf{D}(t')]_\mathrm{J}^\top [\mathbf{D}(t')]_\mathrm{J}\big)^{-1}|||_\infty.$$

The quantity above is first upper bounded by

$$\|[\mathbf{D}(t')]_{\mathrm{J}^c}^\top(\mathbf{I}-\mathbf{P}_{\mathrm{J}}(t'))\mathbf{x}\|_\infty+\lambda(Q_t^2-1),$$

and then, exploiting the bound (72), *strictly* upper bounded by $\lambda(2-Q_t^2)+\lambda(Q_t^2-1)=\lambda$. Putting together the pieces with $\mathrm{sign}(\hat{\boldsymbol{\alpha}}(t'))=\mathrm{sign}(\boldsymbol{\alpha}_0)$, Lemma 16 leads to the desired conclusion.

□

We can now proceed to the proof of Proposition 3. Since $\|\mathbf{d}^j(t')\|_2=1$ for all j, we have

$$\begin{aligned}\|[\mathbf{D}(t')]_{\mathrm{J}}^\top(\mathbf{x}-\mathbf{D}(t')\boldsymbol{\alpha}_0)\|_\infty &\leq \|\mathbf{x}-\mathbf{D}(t')\boldsymbol{\alpha}_0\|_2 &(73)\\ \|[\mathbf{D}(t')]_{\mathrm{J}^c}^\top(\mathbf{I}-\mathbf{P}_{\mathrm{J}}(t'))\mathbf{x}\|_\infty &\leq \|(\mathbf{I}-\mathbf{P}_{\mathrm{J}}(t'))\mathbf{x}\|_2 &(74)\end{aligned}$$

Using Lemma 22, provided that

$$\tau=\frac{\lambda^2(2-Q_t^2)^2}{5\cdot(t'^2\cdot\sigma_\alpha^2+m\cdot\sigma^2)}\geq 1$$

we have

$$\begin{aligned}\Pr\left(\|\mathbf{x}-\mathbf{D}(t')\boldsymbol{\alpha}_0\|_2\geq\lambda(2-Q_t^2)\right) &= \Pr\left(\|\mathbf{x}-\mathbf{D}(t')\boldsymbol{\alpha}_0\|_2^2\geq 5(t'^2\cdot\sigma_\alpha^2+m\cdot\sigma^2)\tau\right)\leq\exp(-\tau)\\ \Pr\left(\|(\mathbf{I}-\mathbf{P}_{\mathrm{J}}(t'))\mathbf{x}\|_2\geq\lambda(2-Q_t^2)\right) &= \Pr\left(\|(\mathbf{I}-\mathbf{P}_{\mathrm{J}}(t'))\mathbf{x}\|_2^2\geq 5(t'^2\cdot\sigma_\alpha^2+m\cdot\sigma^2)\tau\right)\leq\exp(-\tau)\end{aligned}$$

With a union bound, we conclude that $\|\mathbf{x}-\mathbf{D}(t')\boldsymbol{\alpha}_0\|_2<\lambda(2-Q_t^2)$ and $\|(\mathbf{I}-\mathbf{P}_{\mathrm{J}}(t'))\mathbf{x}\|_2\geq\lambda(2-Q_t^2)$, except with probability at most

$$\Pr\left(\mathcal{E}_{\mathrm{coincide}}^c(t')\right)\leq 2\cdot\exp(-\tau)=2\cdot\exp\left(-\frac{\lambda^2(2-Q_t^2)^2}{5\cdot(t'^2\cdot\sigma_\alpha^2+m\cdot\sigma^2)}\right).$$

G Proof of Proposition 4

We now consider the proof of Proposition 4 whose structure is identical to that of Proposition 3. We recall that we are in noiseless setting, i.e., $\sigma=0$, and we assume that the coefficients of $\boldsymbol{\alpha}_0$ are almost surely bounded by $\overline{\alpha}$.

In the light of Lemma 4, let us first observe that almost surely

$$\|[\mathbf{D}(t')]_{\mathrm{J}}^\top(\mathbf{x}-\mathbf{D}(t')\boldsymbol{\alpha}_0)\|_\infty\leq\|\mathbf{x}-\mathbf{D}(t')\boldsymbol{\alpha}_0\|_2\leq\|[\mathbf{D}_0-\mathbf{D}(t')]_{\mathrm{J}}[\boldsymbol{\alpha}_0]_{\mathrm{J}}\|_2\leq t\cdot\|[\boldsymbol{\alpha}_0]_{\mathrm{J}}\|_2\leq\sqrt{k}\overline{\alpha}t.$$

Similarly, it follows

$$\|[\mathbf{D}(t')]_{\mathrm{J}^c}^\top(\mathbf{I}-\mathbf{P}_{\mathrm{J}}(t'))\mathbf{x}\|_\infty\leq\|(\mathbf{I}-\mathbf{P}_{\mathrm{J}}(t'))\mathbf{x}\|_2\leq\sqrt{k}\overline{\alpha}t.$$

Now, we can apply Corollary 3 provided that $\sqrt{k}\overline{\alpha}t\leq\lambda(2-Q_t^2)$, as required by Proposition 4. This leads to the desired conclusion.

H Proof of Proposition 5

Exploiting Proposition 3 we have

$$\max_{i\in[\![1;n]\!]}\Pr([\mathcal{E}_{\mathrm{coincide}}^i(t)\cup\mathcal{E}_{\mathrm{coincide}}^i(0)]^c)\leq 4\exp\left(-\frac{[\lambda(2-Q_t^2)]^2}{5(t^2\sigma_\alpha^2+m\sigma^2)}\right)\triangleq\kappa. \quad (75)$$

The assumption (24) is equivalent to

$$\frac{3}{2-Q_t^2}\cdot t\cdot\sigma_\alpha<\frac{4}{9}\underline{\alpha}$$

hence there exists indeed $\sigma > 0$ and λ satisfying the assumption (25). Moreover, since $5\log 4 \approx 6.93 \leq 9$, the assumption (25) implies that

$$\frac{\gamma^2}{\log 4} = \frac{\lambda^2(2-Q_t^2)^2}{5\log 4 \cdot (t^2\sigma_\alpha^2 + m\sigma^2)} \geq 1,$$

hence $\gamma^2 \geq \log 4$ and $\kappa = 4 \cdot e^{-\gamma^2} \leq 1$. Therefore, we can exploit Lemma 23 and Corollary 4. Given (22), with

$$A_r \triangleq \left(t^2 \cdot \sigma_\alpha^2 + 2m \cdot \sigma^2 + 2\lambda k \cdot \sigma_\alpha\right) \cdot \frac{5(1+\log 2)}{2}$$

we have, except with probability at most $\exp(-n\kappa) = \exp(-4ne^{-\gamma^2})$,

$$\begin{aligned} r_n &\leq 10A_r \cdot (3 - \log\kappa) \cdot \kappa = A \cdot 10 \cdot (3 - \log 4 + \gamma^2) \cdot 4e^{-\gamma^2} \\ &\leq \left(t^2 \cdot \sigma_\alpha^2 + 2m \cdot \sigma^2 + 2\lambda k\sigma_\alpha\right) \cdot 10(1+\log 2) \cdot 10 \cdot \frac{3}{\log 4} \cdot \gamma^2 \cdot e^{-\gamma^2} \\ &\leq \left(t^2 \cdot \sigma_\alpha^2 + 2m \cdot \sigma^2 + 2\lambda k\sigma_\alpha\right) \cdot 367 \cdot \gamma^2 \cdot e^{-\gamma^2}. \end{aligned}$$

I Technical lemmas

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

I.1 Control on the differences of operators

We will now establish several lemmata regarding the difference of operators that appear in the paper.

The following result will exploit Taylor formula with remainder, based on simple matrix and vector derivative computations of $\mathbf{D}(\mathbf{W}, \mathbf{v}, t)$; we refer the interested reader to Magnus and Neudecker [1988] for details about such manipulations. For convenience, let us define

$$\begin{aligned} \mathbf{C}(t) &\triangleq \mathrm{Diag}(\cos(\mathbf{v}_j t)) && (76)\\ \mathbf{S}(t) &\triangleq \mathrm{Diag}(\sin(\mathbf{v}_j t)) && (77)\\ \mathbf{V} &\triangleq \mathrm{Diag}(\mathbf{v}_j) && (78)\\ \mathbf{R}_{\mathrm{J}}(t) &\triangleq \mathbf{D}_{\mathrm{J}}(t)\mathbf{\Theta}_{\mathrm{J}}(t)[\nabla_t \mathbf{D}(t)]_{\mathrm{J}}^\top. && (79) \end{aligned}$$

and denote the symmetric part of a square matrix $\mathbf{M}$ by $\mathrm{sym}(\mathbf{M}) \triangleq \frac{1}{2}(\mathbf{M} + \mathbf{M}^\top)$.

Lemma 17.

$$\begin{aligned} \nabla_t \mathbf{D}(t) &= (-\mathbf{D}_0\mathbf{S}(t) + \mathbf{W}\mathbf{C}(t))\mathbf{V} && (80)\\ \|[\nabla_t \mathbf{D}(t)]_{\mathrm{J}}\|_{\mathrm{F}} &= \|\mathbf{v}_{\mathrm{J}}\|_2 && (81)\\ \nabla_t \mathbf{P}_{\mathrm{J}}(t) &= 2\mathrm{sym}\big(\mathbf{R}_{\mathrm{J}}(t)(\mathbf{I} - \mathbf{P}_{\mathrm{J}}(t))\big) && (82)\\ \nabla_t[\mathbf{\Theta}_{\mathrm{J}}(t)\mathbf{D}_{\mathrm{J}}^\top(t)] &= \mathbf{\Theta}_{\mathrm{J}}(t)([\nabla_t \mathbf{D}(t)]_{\mathrm{J}}^\top(\mathbf{I} - \mathbf{P}_{\mathrm{J}}(t)) - [\mathbf{D}(t)]_{\mathrm{J}}^\top[\mathbf{R}_{\mathrm{J}}(t)]^\top) && (83)\\ \nabla_t[\mathbf{\Theta}_{\mathrm{J}}(t)] &= -2\mathrm{sym}\Big(\mathbf{\Theta}_{\mathrm{J}}(t)[\nabla_t \mathbf{D}(t)]_{\mathrm{J}}^\top \mathbf{D}_{\mathrm{J}}(t)\mathbf{\Theta}_{\mathrm{J}}(t)\Big). && (84) \end{aligned}$$

Lemma 18. *Assume $t < \sqrt{1 - \delta_k(\mathbf{D}_0)}$, then for any $\mathbf{W} \in \mathcal{W}_{\mathbf{D}_0}$, $\mathbf{v} \in \mathcal{S}^p$ and J with $|\mathrm{J}| \leq k$ we have*

$$\begin{aligned} \|\mathbf{P}_{\mathrm{J}}(t) - \mathbf{P}_{\mathrm{J}}(0)\|_2 \leq \|\mathbf{P}_{\mathrm{J}}(t) - \mathbf{P}_{\mathrm{J}}(0)\|_{\mathrm{F}} &\leq 2t \cdot C_t \cdot \|\mathbf{v}_{\mathrm{J}}\|_2, && (85)\\ \|\mathbf{\Theta}_{\mathrm{J}}(t)[\mathbf{D}]_{\mathrm{J}}^\top(t) - \mathbf{\Theta}_{\mathrm{J}}(0)[\mathbf{D}_0]_{\mathrm{J}}^\top\|_2 \leq \|\mathbf{\Theta}_{\mathrm{J}}(t)[\mathbf{D}]_{\mathrm{J}}^\top(t) - \mathbf{\Theta}_{\mathrm{J}}(0)[\mathbf{D}_0]_{\mathrm{J}}^\top\|_{\mathrm{F}} &\leq 2t \cdot C_t^2 \cdot \|\mathbf{v}_J\|_2, && (86)\\ \|\mathbf{\Theta}_{\mathrm{J}}(t) - \mathbf{\Theta}_{\mathrm{J}}(0)\|_2 \leq \|\mathbf{\Theta}_{\mathrm{J}}(t) - \mathbf{\Theta}_{\mathrm{J}}(0)\|_2 &\leq 2t \cdot C_t^3 \cdot \|\mathbf{v}_{\mathrm{J}}\|_2. && (87) \end{aligned}$$

Lemma 19. *Assume that* $t < \sqrt{1-\delta_k(\mathbf{D}_0)}$. *Denote* $\mathbf{P}_{\mathrm{J},1} = \mathbf{P}_{\mathrm{J}}(\mathbf{W}_1, \mathbf{v}_1, t)$ *and* $\mathbf{P}_{\mathrm{J},2} = \mathbf{P}_{\mathrm{J}}(\mathbf{W}_2, \mathbf{v}_2, t)$ *and similarly for the other considered quantities. For any* $\mathbf{W}_1, \mathbf{W}_2 \in \mathcal{W}_{\mathbf{D}_0}$, $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{S}_+^p$, *and* J *with* $|\mathrm{J}| \leq k$ *we have*

$$\begin{aligned}
\|\mathbf{D}_{\mathrm{J},1} - \mathbf{D}_{\mathrm{J},2}\|_2 \leq \|\mathbf{D}_{\mathrm{J},1} - \mathbf{D}_{\mathrm{J},2}\|_{\mathrm{F}} &\leq 2t \cdot C_t \cdot d\big((\mathbf{W}_1, \mathbf{v}_1), (\mathbf{W}_2, \mathbf{v}_2)\big) && (88)\\
\|\mathbf{P}_{\mathrm{J},1} - \mathbf{P}_{\mathrm{J},2}\|_2 \leq \|\mathbf{P}_{\mathrm{J},1} - \mathbf{P}_{\mathrm{J},2}\|_{\mathrm{F}} &\leq 5t \cdot C_t \cdot d\big((\mathbf{W}_1, \mathbf{v}_1), (\mathbf{W}_2, \mathbf{v}_2)\big) && (89)\\
\|\mathbf{\Theta}_{\mathrm{J},1}[\mathbf{D}_1]_{\mathrm{J}}^\top - \mathbf{\Theta}_{\mathrm{J},2}[\mathbf{D}_2]_{\mathrm{J}}^\top\|_2 \leq \|\mathbf{\Theta}_{\mathrm{J},1}[\mathbf{D}_1]_{\mathrm{J}}^\top - \mathbf{\Theta}_{\mathrm{J},2}[\mathbf{D}_2]_{\mathrm{J}}^\top\|_{\mathrm{F}} &\leq 5t \cdot C_t \cdot d\big((\mathbf{W}_1, \mathbf{v}_1), (\mathbf{W}_2, \mathbf{v}_2)\big) && (90)\\
\|\mathbf{\Theta}_{\mathrm{J},1} - \mathbf{\Theta}_{\mathrm{J},2}\|_2 \leq \|\mathbf{\Theta}_{\mathrm{J},1} - \mathbf{\Theta}_{\mathrm{J},2}\|_{\mathrm{F}} &\leq 5t \cdot C_t^3 \cdot d\big((\mathbf{W}_1, \mathbf{v}_1), (\mathbf{W}_2, \mathbf{v}_2)\big). && (91)
\end{aligned}$$

Proof of Lemma 18-Equation (85). We apply a Taylor formula with remainder [e.g., Theorem 14.4 in Dym, 2007] based on Lemma 17 (Equation (82)): for any $\mathbf{U} \in \mathbb{R}^{m\times m}$ there exists $0 \leq t' = t'(\mathbf{U}) \leq t$ such that

$$\mathrm{Tr}\Big(\mathbf{U} \cdot (\mathbf{P}_{\mathrm{J}}(t) - \mathbf{P}_{\mathrm{J}}(0))\Big) = 2t \cdot \mathrm{Tr}\Big(\mathbf{U} \cdot \mathrm{sym}\big(\mathbf{R}_{\mathrm{J}}(t')(\mathbf{I} - \mathbf{P}_{\mathrm{J}}(t'))\big)\Big) \leq 2t \cdot \|\mathbf{R}_{\mathrm{J}}(t')(\mathbf{I} - \mathbf{P}_{\mathrm{J}}(t'))\|_{\mathrm{F}} \cdot \|\mathbf{U}\|_{\mathrm{F}}.$$

Given that $\|[\nabla \mathbf{D}(t')]_{\mathrm{J}}\|_{\mathrm{F}} = \|\mathbf{v}_{\mathrm{J}}\|_2$, we have using the bound (43)

$$\|\mathbf{R}_{\mathrm{J}}(t')\|_{\mathrm{F}} \leq \|\mathbf{D}_{\mathrm{J}}(t')\mathbf{\Theta}_{\mathrm{J}}(t')\|_2 \cdot \|[\nabla \mathbf{D}(t')]_{\mathrm{J}}\|_{\mathrm{F}} \leq C_t \cdot \|\mathbf{v}_{\mathrm{J}}\|_2, \tag{92}$$

hence the upper bound

$$\mathrm{Tr}\Big(\mathbf{U} \cdot (\mathbf{P}_{\mathrm{J}}(t) - \mathbf{P}_{\mathrm{J}}(0))\Big) \leq 2t \cdot \|\mathbf{R}_{\mathrm{J}}(t')\|_{\mathrm{F}} \cdot \|\mathbf{U}\|_{\mathrm{F}} \leq 2t \cdot C_t \cdot \|\mathbf{v}_{\mathrm{J}}\|_2 \cdot \|\mathbf{U}\|_{\mathrm{F}}.$$

We conclude using the fact that $\|\mathbf{P}_{\mathrm{J}}(t) - \mathbf{P}_{\mathrm{J}}(0)\|_2 \leq \|\mathbf{P}_{\mathrm{J}}(t) - \mathbf{P}_{\mathrm{J}}(0)\|_{\mathrm{F}} = \max_{\|\mathbf{U}\|_{\mathrm{F}} \leq 1} \mathrm{Tr}(\mathbf{U}^\top(\mathbf{P}_{\mathrm{J}}(t) - \mathbf{P}_{\mathrm{J}}(0)))$, □

Proof of Lemma 18-Equation (86). Again, we apply a Taylor formula with remainder and Lemma 17 (Equation (83)): for any $\mathbf{U} \in \mathbb{R}^{m\times p}$, there exists some $0 \leq t' \leq t$ such that

$$\begin{aligned}
\mathrm{Tr}\left(\mathbf{U}(\mathbf{\Theta}_{\mathrm{J}}(t)\mathbf{D}_{\mathrm{J}}^\top(t) - \mathbf{\Theta}_{\mathrm{J}}(0)[\mathbf{D}_0]_{\mathrm{J}}^\top)\right) &= t \cdot \mathrm{Tr}\left(\mathbf{U}\Big[\mathbf{\Theta}_{\mathrm{J}}(t')\big([\nabla_t \mathbf{D}(t')]_{\mathrm{J}}^\top(\mathbf{I} - \mathbf{P}_{\mathrm{J}}(t')) - [\mathbf{D}(t')]_{\mathrm{J}}^\top[\mathbf{R}_{\mathrm{J}}(t')]^\top\big)\Big]\right)\\
&\leq t \cdot \|\mathbf{\Theta}_{\mathrm{J}}(t')\big([\nabla_t \mathbf{D}(t')]_{\mathrm{J}}^\top(\mathbf{I} - \mathbf{P}_{\mathrm{J}}(t')) - [\mathbf{D}(t')]_{\mathrm{J}}^\top[\mathbf{R}_{\mathrm{J}}(t')]^\top\big)\|_{\mathrm{F}} \cdot \|\mathbf{U}\|_{\mathrm{F}}
\end{aligned}$$

Now, using the bounds (42), (43) and (92) we have

$$\begin{aligned}
\|\mathbf{\Theta}_{\mathrm{J}}(t')\big([\nabla_t \mathbf{D}(t')]_{\mathrm{J}}^\top(\mathbf{I} - \mathbf{P}_{\mathrm{J}}(t'))\|_{\mathrm{F}} &\leq \|\mathbf{\Theta}_{\mathrm{J}}(t')\|_2 \cdot \|[\nabla_t \mathbf{D}(t')]_{\mathrm{J}}\|_{\mathrm{F}} \leq C_t^2 \cdot \|\mathbf{v}_{\mathrm{J}}\|_2,\\
\|\mathbf{\Theta}_{\mathrm{J}}(t')[\mathbf{D}(t')]_{\mathrm{J}}^\top[\mathbf{R}_{\mathrm{J}}(t')]^\top\|_{\mathrm{F}} &\leq \|\mathbf{\Theta}_{\mathrm{J}}(t')[\mathbf{D}(t')]_{\mathrm{J}}^\top\|_2 \cdot \|\mathbf{R}_{\mathrm{J}}(t')\|_{\mathrm{F}} \leq C_t \cdot (C_t \cdot \|\mathbf{v}_{\mathrm{J}}\|_2) \leq C_t^2 \cdot \|\mathbf{v}_{\mathrm{J}}\|_2
\end{aligned}$$

and we can conclude. □

Proof of Lemma 18-Equation (87). We follow the same line, using the intermediate result from Lemma 17 (Equation (84)). For any $\mathbf{U} \in \mathbb{R}^{p\times p}$ there is some $0 \leq t' = t'(\mathbf{U}) \leq t$ such that

$$\begin{aligned}
|\mathrm{Tr}\left(\mathbf{U} \cdot (\mathbf{\Theta}_{\mathrm{J}}(t) - \mathbf{\Theta}_{\mathrm{J}}(0))\right)| &= \big|2t \cdot \mathrm{Tr}\left(\mathbf{U} \cdot \mathrm{sym}\big(\mathbf{\Theta}_{\mathrm{J}}(t')[\nabla_t \mathbf{D}(t')]_{\mathrm{J}}^\top \mathbf{D}_{\mathrm{J}}(t')\mathbf{\Theta}_{\mathrm{J}}(t')\big)\right)\big|\\
&\leq 2t \cdot \|\mathbf{\Theta}_{\mathrm{J}}(t')[\nabla_t \mathbf{D}(t')]_{\mathrm{J}}^\top \mathbf{D}_{\mathrm{J}}(t')\mathbf{\Theta}_{\mathrm{J}}(t')\|_{\mathrm{F}} \cdot \|\mathbf{U}\|_{\mathrm{F}}.
\end{aligned}$$

Since $\|[\nabla \mathbf{D}(t')]_{\mathrm{J}}\|_{\mathrm{F}} = \|\mathbf{v}_{\mathrm{J}}\|_2$, using (43) and (42) we obtain the upper bound

$$\begin{aligned}
2t \cdot \|\mathbf{\Theta}_{\mathrm{J}}(t')[\nabla_t \mathbf{D}(t')]_{\mathrm{J}}^\top \mathbf{D}_{\mathrm{J}}(t')\mathbf{\Theta}_{\mathrm{J}}(t')\|_{\mathrm{F}} &\leq 2t \cdot \|\mathbf{D}_{\mathrm{J}}(t')\mathbf{\Theta}_{\mathrm{J}}(t')\|_2 \cdot \|\mathbf{\Theta}_{\mathrm{J}}(t')\|_2 \cdot \|[\nabla_t \mathbf{D}(t')]_{\mathrm{J}}\|_{\mathrm{F}}\\
&\leq 2t \cdot C_t \cdot C_t^2 \cdot \|\mathbf{v}_{\mathrm{J}}\|_2.
\end{aligned}$$

□

Proof of Lemma 19. Since $d\left((\mathbf{W}_1,\mathbf{v}_1),(\mathbf{W}_2,\mathbf{v}_2)\right) = \max[\max_j \|\mathbf{w}_1^j-\mathbf{w}_2^j\|_2, \|\mathbf{v}_1-\mathbf{v}_2\|_2] \leq \varepsilon$, we can bound the difference between the columns of $\mathbf{D}_i = \mathbf{D}(\mathbf{D}_0,\mathbf{W}_i,\mathbf{v}_i,t)$, $i=1,2$:

$$\begin{aligned}
\mathbf{d}_1^j - \mathbf{d}_2^j &= (\cos(\mathbf{v}_1^j t)-\cos(\mathbf{v}_2^j t))\cdot\mathbf{d}_0^j + \sin(\mathbf{v}_1^j t)\cdot\mathbf{w}_1^j - \sin(\mathbf{v}_2^j t)\cdot\mathbf{w}_2^j \\
\|\mathbf{d}_1^j - \mathbf{d}_2^j\|_2^2 &= (\cos(\mathbf{v}_1^j t)-\cos(\mathbf{v}_2^j t))^2 + \|\sin(\mathbf{v}_1^j t)\cdot\mathbf{w}_1^j - \sin(\mathbf{v}_2^j t)\cdot\mathbf{w}_2^j\|_2^2 \\
&= \cos^2(\mathbf{v}_1^j t)+\cos^2(\mathbf{v}_2^j t) - 2\cos(\mathbf{v}_1^j t)\cos(\mathbf{v}_2^j t) \\
&\quad + \sin^2(\mathbf{v}_1^j t)+\sin^2(\mathbf{v}_2^j t) - 2\sin(\mathbf{v}_1^j t)\sin(\mathbf{v}_2^j t)[\mathbf{w}_1^j]^\top\mathbf{w}_2^j \\
&= 2 - 2\cos(\mathbf{v}_1^j t)\cos(\mathbf{v}_2^j t) - [2-\|\mathbf{w}_1^j-\mathbf{w}_2^j\|_2^2]\sin(\mathbf{v}_1^j t)\sin(\mathbf{v}_2^j t) \\
&= \|\mathbf{w}_1^j-\mathbf{w}_2^j\|_2^2\cdot\sin(\mathbf{v}_1^j t)\sin(\mathbf{v}_2^j t) + 4\sin^2\frac{(\mathbf{v}_1^j-\mathbf{v}_2^j)t}{2} \leq \left(\varepsilon^2\mathbf{v}_1^j\mathbf{v}_2^j + (\mathbf{v}_1^j-\mathbf{v}_2^j)^2\right)t^2
\end{aligned}$$

As a result we obtain

$$\|\mathbf{D}_1-\mathbf{D}_2\|_{\text{F}}^2 = \sum_{j=1}^p \|\mathbf{d}_1^j-\mathbf{d}_2^j\|_2^2 \leq \left(\varepsilon^2\mathbf{v}_1^\top\mathbf{v}_2 + \|\mathbf{v}_1-\mathbf{v}_2\|_2^2\right)t^2 \leq 2\varepsilon^2t^2.$$

Exploiting Lemma 1, we can write $\mathbf{D}_2 = \mathbf{D}(\mathbf{D}_1,\mathbf{W},\mathbf{v},t')$ with $t' \leq \frac{\pi}{2}\|\mathbf{D}_1-\mathbf{D}_2\|_{\text{F}} \leq \frac{\pi}{\sqrt{2}}\varepsilon t$.

Now consider $\mathbf{D}(\tau) \triangleq \mathbf{D}(\mathbf{D}_1,\mathbf{W},\mathbf{v},\tau)$ with $0\leq\tau\leq t'$ and $\mathbf{d}^j(\tau)$ its columns. Noticing that $\tau\mapsto\mathbf{d}^j(\tau)$ is a geodesic on the unit sphere that joins $\mathbf{d}^j(0)=\mathbf{d}_1^j$ to $\mathbf{d}^j(t')=\mathbf{d}_2^j$, we obtain

$$\|\mathbf{d}^j(\tau)-\mathbf{d}_0^j\|_2 \leq \max(\|\mathbf{d}_1^j-\mathbf{d}_0^j\|_2, \|\mathbf{d}_2^j-\mathbf{d}_0^j\|_2) = 2\sin\left(\frac{t\mathbf{v}_j}{2}\right).$$

Hence, exploiting Lemma 1 again, we can also write $\mathbf{D}(\tau) = \mathbf{D}(\mathbf{D}_0,\mathbf{W}',\mathbf{v}',\tau')$, with $\tau'\leq t$. This implies that for every dictionary on the curve $\tau\mapsto\mathbf{D}(\tau)$, $0\leq\tau\leq t'$, the bounds of Lemma 5 with the constant C_t hold true. We can therefore repeat the Taylor argument of the proof of Lemma 18, noticing that since the considered end point is at $t'\leq\frac{\pi}{\sqrt{2}}\varepsilon t$ instead of t, the factor $2t$ in the resulting bounds is replaced by $2t'\leq\pi\sqrt{2}\varepsilon t\leq 5\varepsilon t$. □

I.2 Control of norms

In this section, we first recall some known concentration results.

Lemma 20 (From Hsu et al. [2011])**.** *Let us consider $\mathbf{z}\in\mathbb{R}^m$ a random vector of independent sub-Gaussian variables with parameters upper bounded by $\sigma>0$. Let $\mathbf{A}\in\mathbb{R}^{m\times p}$ be a fixed matrix. For all $\tau>0$, it holds*

$$\Pr\Big(\|\mathbf{A}\mathbf{z}\|_2^2 > \sigma^2(\|\mathbf{A}\|_{\text{F}}^2 + 2\sqrt{\text{Tr}[(\mathbf{A}^\top\mathbf{A})^2]\tau} + 2|\!|\!|\mathbf{A}^\top\mathbf{A}|\!|\!|_2\tau)\Big) \leq \exp(-\tau).$$

In particular, for any $\tau\geq 1$, we have

$$\Pr\Big(\|\mathbf{A}\mathbf{z}\|_2^2 > 5\sigma^2\|\mathbf{A}\|_{\text{F}}^2\tau\Big) \leq \exp(-\tau).$$

Lemma 21 (Bernstein's Inequality)**.** *Let $\{z_j\}_{j\in[\![1;n]\!]}$ be a collection of independent, zero-mean random variables. If there exist $M,\varsigma\in\mathbb{R}_+$ such that for any integer $k\geq 2$ and any $j\in[\![1;n]\!]$, it holds*

$$\mathbb{E}[|z_j|^k] \leq \frac{k!}{2}M^{k-2}\varsigma^2,$$

then we have for any $\tau\geq 0$,

$$\Pr\Big(\sum_{j=1}^n z_j > \tau\Big) \leq \exp\Big(-\frac{\tau^2}{2n\varsigma^2+2M\tau}\Big).$$

In particular, for any $\tau \leq \frac{\varsigma\sqrt{n}}{2M}$, *we have*

$$\Pr\Big(\frac{1}{n}\sum_{j=1}^{n} z_j > 2\varsigma\frac{\tau}{\sqrt{n}}\Big) \leq \exp\big(-\tau^2\big).$$

Proof. The displayed result is a straightforward adaptation of Lemma 4.1.9 in De la Peña and Giné [1999], where we use the term ς^2 in lieu of the true variance. □

Lemma 22 (Control of the ℓ_2-norm of a signal and its coefficients)**.** *Let* $\mathbf{x}$ *be a signal following our generative model, and* $\boldsymbol{\alpha}_0$ *be its coefficients. For any* $\tau \geq 1$ *and* $\mathbf{D} = \mathbf{D}(\mathbf{W}, \mathbf{v}, t)$, *we have*

$$\Pr\left(\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}_0\|_2^2 + \|\boldsymbol{\varepsilon}\|_2^2 > 5(t^2\sigma_\alpha^2 + 2m\sigma^2)\tau\right) \leq \exp(-\tau) \quad (93)$$

$$\Pr\left(\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}_0\|_2^2 > 5(t^2\sigma_\alpha^2 + m\sigma^2)\tau\right) \leq \exp(-\tau) \quad (94)$$

$$\Pr\left(\|(\mathbf{I} - \mathbf{P}_{\mathrm{J}}(t))\mathbf{x}\|_2^2 > 5(t^2\sigma_\alpha^2 + (m-k)\sigma^2)\tau\right) \leq \exp(-\tau) \quad (95)$$

$$\Pr\left(\|\boldsymbol{\alpha}_0\|_2^2 > 5k\sigma_\alpha^2\tau\right) \leq \exp(-\tau) \quad (96)$$

$$\Pr\left(\|\mathbf{x}\|_2^2 > 5(k\sigma_\alpha^2 + m\sigma^2)\tau\right) \leq \exp(-\tau) \quad (97)$$

Proof. We prove the result for $\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}_0\|_2^2 + \|\boldsymbol{\varepsilon}\|_2^2$. The same technique applies to the other quantities. We recall that $\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}_0 = [\mathbf{D}_0 - \mathbf{D}]_{\mathrm{J}}[\boldsymbol{\alpha}_0]_{\mathrm{J}} + \boldsymbol{\varepsilon}$, and that the considered norm can be expressed as follows

$$\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}_0\|_2^2 + \|\boldsymbol{\varepsilon}\|_2^2 = \left\| \begin{bmatrix} \sigma_\alpha[\mathbf{D}_0 - \mathbf{D}]_{\mathrm{J}} & \sigma\mathbf{I} \\ \mathbf{0} & \sigma\mathbf{I} \end{bmatrix} \begin{pmatrix} \frac{1}{\sigma_\alpha}[\boldsymbol{\alpha}_0]_{\mathrm{J}} \\ \frac{1}{\sigma}\boldsymbol{\varepsilon} \end{pmatrix} \right\|_2 .$$

The result is a direct application of Lemma 20 conditioned to the draw of J, using Lemma 4 to control

$$\left\| \begin{bmatrix} \sigma_\alpha[\mathbf{D}_0 - \mathbf{D}]_{\mathrm{J}} & \sigma\mathbf{I} \\ \mathbf{0} & \sigma\mathbf{I} \end{bmatrix} \right\|_{\mathrm{F}}^2 = \|[\mathbf{D}_0 - \mathbf{D}]_{\mathrm{J}}\|_{\mathrm{F}}^2 \cdot \sigma_\alpha^2 + 2m\sigma^2 \leq t^2\sigma_\alpha^2 + 2m\sigma^2.$$

The bound being independent of J, the result is also true without conditioning. Note that to control the behaviour of $\|(\mathbf{I} - \mathbf{P}_{\mathrm{J}}(t))\mathbf{x}\|_2^2$ we use the fact that since $\mathbf{P}_{\mathrm{J}}(t)$ is an orthogonal projector on a subspace of dimension k, we have $\|\mathbf{I} - \mathbf{P}_{\mathrm{J}}(t)\|_{\mathrm{F}}^2 = m - k$. □

Lemma 23. *Let* $\mathbf{x}$ *and* $\boldsymbol{\alpha}_0$ *be drawn according to our signal model. Define*

$$\begin{aligned} y &= \sup_{\mathbf{W},\mathbf{v}} \mathcal{L}_{\mathbf{x}}(\mathbf{D}(\mathbf{W}, \mathbf{v}, t), \boldsymbol{\alpha}_0) \\ y' &= \sup_{\mathbf{W},\mathbf{v}} \{\mathcal{L}_{\mathbf{x}}(\mathbf{D}(\mathbf{W}, \mathbf{v}, t), \boldsymbol{\alpha}_0) + \mathcal{L}_{\mathbf{x}}(\mathbf{D}_0, \boldsymbol{\alpha}_0)\} \end{aligned}$$

For any $\tau \geq 1$ *we have*

$$\Pr(y \geq A_{\mathcal{L}}(t) \cdot \tau) \leq e^{-\tau} \quad (98)$$

$$\Pr(y' \geq A_r(t) \cdot \tau) \leq e^{-\tau} \quad (99)$$

where

$$\begin{aligned} A_{\mathcal{L}}(t) &\triangleq \frac{5(1+\log 2)}{2} \cdot \left(t^2\sigma_\alpha^2 + m\sigma^2 + \lambda k\sigma_\alpha\right) \\ A_r(t) &\triangleq \frac{5(1+\log 2)}{2} \cdot \left(t^2\sigma_\alpha^2 + 2m\sigma^2 + 2\lambda k\sigma_\alpha\right). \end{aligned}$$

Proof. Using Lemma 4 we have, for $\mathbf{D} = \mathbf{D}(\mathbf{W}, \mathbf{v}, t)$, uniformly over $\mathbf{W}, \mathbf{v}$:

$$\begin{aligned}\mathcal{L}_{\mathbf{x}}(\mathbf{D}, \boldsymbol{\alpha}_0) &= \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}_0\|_2^2 + \lambda\|\boldsymbol{\alpha}_0\|_1 \leq \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}_0\|_2^2 + \lambda\sqrt{k}\|\boldsymbol{\alpha}_0\|_2 \\ \mathcal{L}_{\mathbf{x}}(\mathbf{D}, \boldsymbol{\alpha}_0) + \mathcal{L}_{\mathbf{x}}(\mathbf{D}_0, \boldsymbol{\alpha}_0) &\leq \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}_0\|_2^2 + \frac{1}{2}\|\boldsymbol{\varepsilon}\|_2^2 + 2\lambda\sqrt{k}\|\boldsymbol{\alpha}_0\|_2.\end{aligned}$$

Fix any $\tau \geq 1$ and define $\tau' = (1+\log 2)\tau \geq \tau + \log 2$. If $\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}_0\|_2^2 \leq 5(t^2\sigma_\alpha^2 + m\sigma^2)\tau'$ and $\|\boldsymbol{\alpha}_0\|_2^2 \leq 5k\sigma_\alpha^2\tau'$, then $\|\boldsymbol{\alpha}_0\|_2 \leq \sqrt{5k}\sigma_\alpha\sqrt{\tau'} \leq \sqrt{5k}\sigma_\alpha\tau'$, hence we have

$$y \leq \left(\frac{1}{2}\left(5t^2\sigma_\alpha^2 + 5m\sigma^2\right) + \lambda\sqrt{k}\sqrt{5k\sigma_\alpha^2}\right)\tau' \leq \frac{5}{2}\left(t^2\sigma_\alpha^2 + m\sigma^2 + \lambda k\sigma_\alpha\right)\tau' = A_{\mathcal{L}}(t)\cdot\tau.$$

Lemma 20 and a union bound yield $\Pr(y \geq A_{\mathcal{L}}(t)\cdot\tau) \leq 2\exp(-\tau') \leq \exp(-\tau)$. The proof for y' is similar. □

Lemma 24. *Let y be a random variable satisfying for any $\tau \geq 1$*

$$\Pr\left(|y| > A\tau\right) \leq \exp(-\tau). \tag{100}$$

for some positive constant $A > 0$. Consider an event $\mathcal{E}$ defined on the same probability space as that of y. For any $u \geq 1$, any integer $q \geq 1$, and $0 < p \leq 1$, we have

$$\mathbb{E}\Big[\mathbb{1}_{\mathcal{E}}|y|^{pq}\Big] \leq q!\Big[A^p u\Big]^q\Big[\Pr(\mathcal{E}) + \exp(3-u)\Big] \tag{101}$$

$$\mathbb{E}\Big[\Big|\mathbb{1}_{\mathcal{E}}|y|^p - \mathbb{E}\big[\mathbb{1}_{\mathcal{E}}|y|^p\big]\Big|^q\Big] \leq q!\Big[2A^p u\Big]^q\Big[\Pr(\mathcal{E}) + \exp(3-u)\Big]. \tag{102}$$

Proof. To begin with, let us notice that by invoking twice the triangle inequality, we have

$$\Big(\mathbb{E}\Big\{\Big|\mathbb{1}_{\mathcal{E}}|y|^p - \mathbb{E}\big\{\mathbb{1}_{\mathcal{E}}|y|^p\big\}\Big|^q\Big\}\Big)^{1/q} \leq \big(\mathbb{E}\big\{\mathbb{1}_{\mathcal{E}}|y|^{pq}\big\}\big)^{1/q} + \Big(\mathbb{E}\big\{\big(\mathbb{E}\big\{\mathbb{1}_{\mathcal{E}}|y|^p\big\}\big)^q\big\}\Big)^{1/q},$$

so that by using Jensen's inequality, we obtain

$$\mathbb{E}\Big\{\Big|\mathbb{1}_{\mathcal{E}}|y|^p - \mathbb{E}\big[\mathbb{1}_{\mathcal{E}}|y|^p\big]\Big|^q\Big\} \leq 2^q\mathbb{E}\big[\mathbb{1}_{\mathcal{E}}|y|^{pq}\big],$$

thus proving (102) provided that (101) holds. We now focus on these raw moments. Let fix some $u \geq 1$. We introduce the event

$$\mathcal{K} \triangleq \Big\{\omega;\ \frac{|y(\omega)|}{A} \leq u\Big\},$$

and define l_u as the largest integer such that $u \in [l_u, l_u + 1)$. We can then "discretize" the event $\mathcal{K}^c$ as

$$\mathcal{K}^c \subseteq \bigcup_{l=l_u}^{\infty}\mathcal{K}_l^c, \quad \text{with } \mathcal{K}_l^c = \Big\{\omega;\ \frac{|y(\omega)|}{A} \in [l, l+1)\Big\}.$$

We have

$$\begin{aligned}\mathbb{E}\{\mathbb{1}_{\mathcal{E}}|y|^{pq}\} &= \mathbb{E}\{\mathbb{1}_{\mathcal{E}\cap\mathcal{K}}|y|^{pq}\} + \mathbb{E}\{\mathbb{1}_{\mathcal{E}\cap\mathcal{K}^c}|y|^{pq}\} \leq \big(Au\big)^{pq}\cdot\Pr(\mathcal{E}) + \sum_{l=l_u}^{\infty}\mathbb{E}\{\mathbb{1}_{\mathcal{E}\cap\mathcal{K}_l^c}|y|^{pq}\} \\ &\leq A^{pq}\cdot\Big[u^{pq}\cdot\Pr(\mathcal{E}) + \sum_{l=l_u}^{\infty}(l+1)^{pq}\cdot\mathbb{E}\{\mathbb{1}_{\mathcal{E}\cap\mathcal{K}_l^c}\}\Big] \\ &\leq A^{pq}\cdot\Big[u^q\cdot\Pr(\mathcal{E}) + \sum_{l=l_u}^{\infty}(l+1)^{pq}\cdot\mathbb{E}[\mathbb{1}_{\{\omega;\ |y(\omega)|\geq Al\}}]\Big]\end{aligned}$$

where in the last line we used $u^p \leq u$ since $u \geq 1$ and $p \leq 1$. Using the hypothesis (100), we continue

$$\mathbb{E}\{\mathbb{1}_{\mathcal{E}}|y|^{pq}\} \leq A^{pq} \cdot \Big[u^q \cdot \Pr(\mathcal{E}) + \sum_{l=l_u}^{\infty}(l+1)^{pq}\exp(-l)\Big].$$

Upper bounding the discrete sum by a continuous integral, we recognize here the incomplete Gamma function [Gautschi, 1998],

$$\begin{aligned}
\sum_{l=l_u}^{\infty}(l+1)^{pq}e^{-l} &= \sum_{l=l_u}^{\infty}\int_l^{l+1}(l+1)^{pq}e^{-l}dt \leq \sum_{l=l_u}^{\infty}\int_l^{l+1}(t+1)^{pq}e^{-(t+1)+t+1-l}dt \\
&\leq e^2\sum_{l=l_u}^{\infty}\int_l^{l+1}(t+1)^{pq}e^{-(t+1)}dt = e^2\int_{l_u}^{\infty}(t+1)^{pq}e^{-(t+1)}dt \\
&= e^2\int_{l_u+1}^{\infty}t^{pq}e^{-t}dt \leq e^2\int_u^{\infty}t^q e^{-t}dt = e^2\Gamma\left(q+1,u\right)
\end{aligned}$$

where again we used $t^{pq} \leq t^q$ for $t \geq 1$. A standard formula [see equation (1.3) in Gautschi, 1998] leads to, for $u \geq 1$,

$$\Gamma(q+1,u) = q!\exp(-u)\sum_{j=0}^{q}\frac{u^j}{j!} \leq e\,q!\exp(-u)u^q.$$

Putting all the pieces together we thus reach the advertised conclusion. □

Corollary 4. *Consider n independent draws $\{y^i\}_{i\in[\![1;n]\!]}$ satisfying the hypothesis* (100)*. Consider also n independent events $\{\mathcal{E}^i\}_{i\in[\![1;n]\!]}$ defined on the same probability space, with $\max_{i\in[\![1;n]\!]}\Pr(\mathcal{E}^i) \leq \kappa \leq 1$. Then, for any $0 < p \leq 1$ and $0 \leq \tau \leq \sqrt{n\kappa}$, we have*

$$\mathbb{E}\{\mathbb{1}_{\mathcal{E}^i}|y^i|^p\} \leq 2A^p \cdot (3-\log\kappa)\cdot\kappa \tag{103}$$

$$\Pr\left(\Big|\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{\mathcal{E}^i}|y^i|^p - \mathbb{E}\{\mathbb{1}_{\mathcal{E}^i}|y^i|^p\}\right)\Big| \geq 8A^p\cdot(3-\log\kappa)\cdot\sqrt{\kappa}\cdot\frac{\tau}{\sqrt{n}}\right) \leq \exp(-\tau^2) \tag{104}$$

Proof. Applying Lemma 24-Equation (101) with $u = 3-\log\kappa$ for $q = 1$ we obtain (103) where we used that $\Pr(\mathcal{E}^i) + e^{3-u} \leq \kappa + e^{3-u} = 2\kappa$. Similarly, applying Lemma 24-Equation (102) for $q \geq 2$, we can apply Lemma 21 with $z_i = \mathbb{1}_{\mathcal{E}^i}|y^i|^p - \mathbb{E}\{\mathbb{1}_{\mathcal{E}^i}|y^i|^p\}$, $M = 2A^p u$ and $\varsigma = \sqrt{2}M\sqrt{\kappa + e^{3-u}} = 2M\sqrt{\kappa} = 4A^p(3-\log\kappa)\sqrt{\kappa}$. This shows that for $0 \leq \tau \leq \frac{\sqrt{n}}{2}\frac{\varsigma}{M} = \sqrt{n\kappa}$ we have (104). □