Sample Complexity of Dictionary Learning and other Matrix Factorizations

Rémi Gribonval, Rodolphe Jenatton, Francis Bach, Martin Kleinsteuber, Matthias Seibert

Sample Complexity of Dictionary Learning and other Matrix Factorizations

Rémi Gribonval, *IEEE Fellow*, Rodolphe Jenatton, Francis Bach, Martin Kleinsteuber, Matthias Seibert

***Abstract*—Many modern tools in machine learning and signal processing, such as sparse dictionary learning, principal component analysis (PCA), non-negative matrix factorization (NMF), K-means clustering, etc., rely on the factorization of a matrix obtained by concatenating high-dimensional vectors from a training collection. While the idealized task would be to optimize the expected quality of the factors over the underlying distribution of training vectors, it is achieved in practice by minimizing an empirical average over the considered collection. The focus of this paper is to provide sample complexity estimates to uniformly control how much the empirical average deviates from the expected cost function. Standard arguments imply that the performance of the empirical predictor also exhibit such guarantees. The level of genericity of the approach encompasses several possible constraints on the factors (tensor product structure, shift-invariance, sparsity ...), thus providing a unified perspective on the sample complexity of several widely used matrix factorization schemes. The derived generalization bounds behave proportional to $\sqrt{\log(n)/n}$ w.r.t. the number of samples n for the considered matrix factorization techniques.**

***Index Terms*—Dictionary learning, sparse coding, principal component analysis, K-means clustering, non-negative matrix factorization, structured learning, sample complexity.**

I. Introduction

THE fact that a signal $\mathbf{x} \in \mathbb{R}^m$ which belongs to a certain class has a representation over some class dependent dictionary $\mathbf{D} \in \mathbb{R}^{m \times d}$ is the backbone of many successful signal reconstruction and data analysis algorithms [1]–[3].

That is, $\mathbf{x}$ is the linear combination of columns of $\mathbf{D}$, referred to as *atoms*. Formally, this reads as

$$\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}, \tag{1}$$

where the coefficient vector $\boldsymbol{\alpha} \in \mathbb{R}^d$ as well as the dictionary $\mathbf{D}$ are subject to some constraints. Such a setting covers prominent examples like Principal Component Analysis (PCA), where $\mathbf{D}$ has orthogonal columns, thus representing the subspace where the signal in the given class is contained. Another example is the sparse synthesis model, also known as sparse coding, where typically $\mathbf{D}$ consists of normalized columns that form an overcomplete basis of the signal space, and $\boldsymbol{\alpha} \in \mathbb{R}^d$ is assumed to be sparse.

The task of learning such dictionaries from a given set of training data is related to matrix factorization problems. Important examples include Higher-Order SVD (also known as multilinear SVD) [4], sparse coding also called dictionary learning [5]–[11], its variants with separable [12] or sparse [13] dictionaries, Non-negative Matrix Factorization (NMF) [14], K-means clustering [15], sparse PCA [16]–[18], and more. The learning task is expressed formally as follows.

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$ be the matrix containing the n training samples arranged as its columns, and let $\mathbf{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n] \in \mathbb{R}^{d \times n}$ contain the corresponding n coefficient vectors, a common approach to the dictionary learning process is the optimization problem

$$\underset{\mathbf{A},\mathbf{D}}{\text{minimize}} \ \tfrac{1}{2}\|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 + \sum_{i=1}^{n} g(\boldsymbol{\alpha}_i) \ \text{ s.t. } \ \mathbf{D} \in \mathfrak{D}. \tag{2}$$

Therein, $g\colon \mathbb{R}^d \to \mathbb{R}$ is a function that promotes the constraints for the coefficient vectors, e.g., sparsity or positivity, and $\mathfrak{D}$ is some predefined admissible set of solutions for the dictionary. Note that for PCA there are no constraints on the coefficient vectors, which implies the penalty function $g \equiv 0$ in this case.

A fundamental question in such a learning process is the sample complexity issue. Assuming that the training samples are drawn according to some distribution $\mathbb{P}$ representing the class of signals of interest, one would ideally like to select the dictionary $\mathbf{D}^\star$ yielding the minimum expected value of (2). However, having only access to n training samples, one can at best select an empirical minimizer $\hat{\mathbf{D}}$. Is this empirical minimizer useful beyond the training set from which it has been selected? This depends on how much the empirical cost function deviates from its expectation.

State of the art sample complexity estimates [19], [20] primarily consider the case where g is the indicator function of a set, such as an ℓ^1 or an ℓ^0 ball, $\mathfrak{D}$ is the set of all unit norm dictionaries or a subset with a restricted isometry property, and the distribution $\mathbb{P}$ is in the class $\mathfrak{P}$ of distributions on the unit sphere of $\mathbb{R}^m$. We generalize these results to:

- **General classes of penalty functions.** Examples covered

R. Gribonval heads the PANAMA project-team (Inria & CNRS), Rennes, France. This work was partially supported by the EU FP7, SMALL project, FET-Open grant number 225913, and by the European Research Council, PLEASE project (ERC-StG-2011-277906). remi.gribonval@inria.fr

R. Jenatton contributed during his Ph.D. with the Inria - SIERRA Project-Team, Laboratoire d'Informatique de l'Ecole Normale Supérieure, Paris, France and his postdoc at CMAP, Ecole Polytechnique, UMR CNRS 7641, Palaiseau, France. He is now with Amazon, Berlin, Germany. jenatton@amazon.com

F. Bach heads the Inria - SIERRA Project-Team, Laboratoire d'Informatique de l'Ecole Normale Supérieure, Paris, France. francis.bach@ens.fr

M. Kleinsteuber & M. Seibert are with the Department of Electrical and Computer Engineering, Technische Universität München, Munich, Germany. This work has been supported by the German DFG funded Cluster of Excellence CoTeSys - Cognition for Technical Systems, and by the DFG grant KL 2189/8-1. {kleinsteuber, m.seibert}@tum.de

This paper was presented in part at SSP 2014 - the IEEE Workshop on Statistical Signal Processing, Jupiters, Gold Coast, Australia.

by our results include: the ℓ^1 penalty and its powers; any mixed norm [21] or quasi-norm [22]; the characteristic function of the ℓ^1-ball, of the set of k-sparse vectors [20], or of non-negative vectors [14].
- **Various classes of dictionaries $\mathfrak{D}$ that can incorporate structures.** Examples covered include: dictionaries with unit norm atoms which are used in many dictionary learning schemes, e.g., K-SVD [8]; sparse dictionaries [13]; shift-invariant dictionaries [23]; tensor product dictionaries [12]; orthogonal dictionaries; non-negative dictionaries [14]; topic models [24]; and tensor products of Stiefel matrices used for Higher-order SVDs [4], [25].
- **Various classes $\mathfrak{P}$ of probability distributions $\mathbb{P}$.** Examples include distributions on the unit sphere which are tackled in [20], but also certain distributions built using sub-Gaussian random variables [26]. For more information on sub-Gaussian random variables, see [27].

After formalizing the problem and setting our notations in Section II, we state our main results in Section III and illustrate them with families of examples in Section IV. The obtained sample complexity estimates depend on the Lipschitz constant of the optimized cost function (2), seen as a function of the unknown dictionary $\mathbf{D}$. This Lipschitz constant is essentially driven by the penalty function g, as discussed in Section V. Our results rely on the assumption that the distributions $\mathbb{P}$ in the class $\mathfrak{P}$ satisfy a concentration of measure property, and the structured constraint set $\mathfrak{D}$ plays a role in the sample complexity estimates through its covering number, as discussed in Section VI where the main theoretical results are established. Covering numbers for a number of structure sets $\mathfrak{D}$ are considered in Section VII. Section VIII gathers worked examples in relation to previous work. Finally, Section IX discusses sharper sample complexity estimates in the high-dimensional setting $m \gg d$.

II. Notations

Sets are denoted with gothic face as $\mathfrak{D}, \mathfrak{P}, \mathfrak{A}$. Matrices are written as boldface capital letters like $\mathbf{X}, \mathbf{D}$, column vectors are denoted by boldfaced small letters, e.g., $\boldsymbol{\alpha}, \mathbf{d}$ whereas scalars are either capital or small letters like n, N. By v_i we denote the i^{th} element of the vector $\mathbf{v}$, v_{ij} denotes the i^{th} element in the j^{th} column of a matrix $\mathbf{V}$, while the i^{th} column is referred to by $\mathbf{v}_i$. The Frobenius norm for matrices is denoted $\|\cdot\|_F$ and the corresponding inner product between matrices is denoted $\langle \cdot, \cdot \rangle_F$. The operator norm $\|\cdot\|_{1\to 2}$ of an $m \times d$ matrix $\boldsymbol{\Delta} = [\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_d]$ is

$$\|\boldsymbol{\Delta}\|_{1\to 2} \triangleq \sup_{\|\boldsymbol{\alpha}\|_1 \leq 1} \|\boldsymbol{\Delta}\boldsymbol{\alpha}\|_2 = \max_{1\leq i\leq d} \|\boldsymbol{\delta}_i\|_2.$$

Finally, $\log$ is the natural logarithm so that $e^{\log t} = t$ for $t > 0$.

Given a dictionary $\mathbf{D} \in \mathbb{R}^{m\times d}$ that fulfills certain structural properties and a signal $\mathbf{x} \in \mathbb{R}^m$, a representation vector $\boldsymbol{\alpha} \in \mathbb{R}^d$ is typically obtained by solving the minimization problem

$$\hat{\boldsymbol{\alpha}} \in \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^d} \tfrac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + g(\boldsymbol{\alpha}),$$

where $g : \mathbb{R}^d \to \mathbb{R}^+ \cup \{+\infty\}$ is a penalty function promoting constraints for the coefficient vector. For our purposes, the question of whether a minimizer $\hat{\boldsymbol{\alpha}}$ actually exists is irrelevant, and we define the quality of how well a signal $\mathbf{x}$ can be coded by a dictionary $\mathbf{D}$ by

$$f_{\mathbf{x}}(\mathbf{D}) \triangleq \inf_{\boldsymbol{\alpha}\in\mathbb{R}^d} \mathcal{L}_{\mathbf{x}}(\mathbf{D}, \boldsymbol{\alpha}),$$

with

$$\mathcal{L}_{\mathbf{x}}(\mathbf{D}, \boldsymbol{\alpha}) \triangleq \tfrac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + g(\boldsymbol{\alpha}).$$

Given n training samples $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{m\times n}$, the average quality of how well a dictionary allows a representation for $\mathbf{X}$ while considering the constraints imposed on the coefficients is

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

The cost function $F_{\mathbf{X}}$ can be written in short form as $F_{\mathbf{X}}(\mathbf{D}) = \inf_{\mathbf{A}\in\mathbb{R}^{d\times n}} \mathcal{L}_{\mathbf{X}}(\mathbf{D}, \mathbf{A})$ with

$$\mathcal{L}_{\mathbf{X}}(\mathbf{D}, \mathbf{A}) \triangleq \tfrac{1}{2n}\|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 + G(\mathbf{A})$$

where $\mathbf{A} \triangleq [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n]$ and $G(\mathbf{A}) \triangleq \frac{1}{n}\sum_{i=1}^n g(\boldsymbol{\alpha}_i)$.

III. Main results & Outline of our Approach

The main contribution of this paper is a general framework to establish the uniform convergence of the function $F_{\mathbf{X}}(\cdot)$ to its expectation when the samples $\mathbf{x}_i$ are all drawn independently according to an unknown probability distribution $\mathbb{P}$. We show in this work that

$$\sup_{\mathbf{D}\in\mathfrak{D}} |F_{\mathbf{X}}(\mathbf{D}) - \mathbb{E}_{\mathbf{x}\sim\mathbb{P}} f_{\mathbf{x}}(\mathbf{D})| \leq \eta_n(g, \mathfrak{D}, \mathfrak{P}).$$

holds with "overwhelming" probability (that will be controlled explicitly in due time as well as the precision bound $\eta_n(g, \mathfrak{D}, \mathfrak{P}) \propto \sqrt{\log n / n}$).

A particular consequence is in terms of generalization bound. An ideal objective in dictionary learning and related matrix factorization problems would be to select the optimal dictionary for the underlying distribution $\mathbb{P}$ of training samples,

$$\mathbf{D}^\star \in \arg\min_{\mathbf{D}\in\mathfrak{D}} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}} f_{\mathbf{x}}(\mathbf{D}).$$

In practice, one can at best access an empirically optimal dictionary $\hat{\mathbf{D}}_n \in \mathfrak{D}$, which minimizes $F_{\mathbf{X}}(\mathbf{D})$ given n training samples gathered in $\mathbf{X}$. How close is its performance to that of $\mathbf{D}^\star$? A consequence of our main result is a generalization bound for the empirical optimum $\hat{\mathbf{D}}_n$: with controlled high probability,

$$\mathbb{E}_{\mathbf{x}\sim\mathbb{P}} f_{\mathbf{x}}(\hat{\mathbf{D}}_n) \leq \mathbb{E}_{\mathbf{x}\sim\mathbb{P}} f_{\mathbf{x}}(\mathbf{D}^\star) + 2\eta_n.$$

Note that the uniform convergence result holds for all dictionaries $\mathbf{D} \in \mathfrak{D}$, and not only at the global optimum of the learning problem. Moreover, even though finding the empirical minimizer $\hat{\mathbf{D}}_n$ often means facing a difficult non-convex optimization problem, recent work on sparse dictionary learning establishes that certain polynomial time algorithms can provably find it with high probability in certain scenarii [28]–[30]. Uniform convergence results as established here

should enable further investigation of the finite sample analysis of such algorithms.

The main result, formally stated in Theorem 1 in Section III-E, combines a standard argument based on covering numbers with an estimation of the expected Lipschitz constant of the function $\mathbf{D} \mapsto F_{\mathbf{X}}(\mathbf{D})$. Our main technical contribution is the estimation of this Lipschitz constant, which is summarized in Theorems 2 and 3 in Section V-B. To prove the main result, we first show that $\mathbf{D} \mapsto F_{\mathbf{X}}(\mathbf{D})$ is Lipschitz with a constant expressed in terms of $\mathbf{X}$ and g. To be more precise, we prove the property $|F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}')| \leq L \cdot \|\mathbf{D} - \mathbf{D}'\|_{1\to 2}$ for all admissible dictionaries $\mathbf{D}, \mathbf{D}' \in \mathfrak{D}$, where L is dependent on $\mathbf{X}$ and g. In the course of our work we provide more tangible bounds for the Lipschitz constant. However, this requires certain assumptions on the penalty function g. These assumptions will be motivated and introduced in the remainder of this section. Therefore, we refrain from presenting the concrete bounds at this point. They can be found in Corollary 3 and Lemma 6. The discussion is concluded using an argument based on covering numbers and concentration of measure. While the latter technique is fairly standard, a major contribution of this paper lies in the identification of two large classes of penalty functions for which the desired Lipschitz property holds and is nicely controlled. In the following, we will give a detailed insight into the required components to provide the generalization bound. Furthermore, we motivate the assumptions that we enforce on the penalty function g (these are labeled with **A**), or jointly on the penalty function g and the set of admissible dictionaries $\mathfrak{D}$ (labeled with **B**), and on the probability distribution (labeled with **C**).

A. Role of the constraint set $\mathfrak{D}$

The structure of the constraint set $\mathfrak{D}$ is incorporated in the analysis by employing a standard ϵ-net argument. A compact constraint set $\mathfrak{D} \subset \mathbb{R}^{m\times d}$ can be covered with a finite number of balls of small radius, i.e., it has a finite covering number

$$\mathcal{N}(\mathfrak{D}, \epsilon) \triangleq \min\{\sharp\mathfrak{Q} : \mathfrak{Q} \subset \mathbb{R}^{m\times d}, \mathfrak{D} \subset \bigcup_{\mathbf{q}\in\mathfrak{Q}} B_\epsilon(\mathbf{q})\}$$

where $\sharp$ denotes the cardinality of a set. For further background about covering numbers, we refer the reader to [31] and references therein. In our setting the ϵ-balls $B_\epsilon(\cdot)$ are defined with respect to the metric $\|\cdot\|_{1\to 2}$ since the unit ball in this metric is closely connected with the common constraint that a dictionary has unit norm columns.

We postpone explicit estimates of covering numbers to Section VII, but notice that covering numbers for all the considered classes satisfy the generic upper bound

$$\mathcal{N}(\mathfrak{D}, \epsilon) \leq (C/\epsilon)^h, \ \forall\ 0 < \epsilon \leq 1 \tag{3}$$

where $h \geq 1$ is a measure of the "dimension" of $\mathfrak{D}$, and $C \geq 1$. In particular, for the set of all dictionaries in $\mathbb{R}^{m\times d}$ with unit-norm columns, which is one of the most common constraint sets, these constants take the values $h = md$ and $C = 3$. The sample complexity will be essentially driven by the "dimension" h, while the constant C will only play a logarithmic role.

B. Lipschitz property for "norm-like" penalty functions g

In short, we will prove in Section V that under assumptions

- **A1:** g is non-negative;
- **A2:** g is lower semi-continuous;
- **A3:** g is coercive: $g(\boldsymbol{\alpha}) \to \infty$ as $\|\boldsymbol{\alpha}\| \to \infty$;

the function $\mathbf{D} \mapsto F_{\mathbf{X}}(\mathbf{D})$ is Lipschitz over $\mathbb{R}^{m\times d}$ with controlled constant. With the additional assumption

- **A4:** $g(\mathbf{0}) = 0$;

the Lipschitz constant is bounded by

$$L_{\mathbf{X}}(\bar{g}) \triangleq \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i\|_2 \cdot \bar{g}(\|\mathbf{x}_i\|_2^2/2) \tag{4}$$

where we define the auxiliary function $\bar{g}$ as follows.

Definition 1. For penalty functions g satisfying **A1-A3**,

$$\bar{g}(t) \triangleq \sup_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^d \\ g(\boldsymbol{\alpha})\leq t}} \|\boldsymbol{\alpha}\|_1, \ t \geq 0.$$

The presence of the ℓ^1 norm in this definition results from the choice of the metric $\|\cdot\|_{1\to 2}$ to measure covering numbers. It will become clearer in the proof of Lemma 5 where we need to control $\|\boldsymbol{\alpha}\|_1$ where $\boldsymbol{\alpha}$ is a (near) minimizer of $\mathcal{L}_{\mathbf{x}}(\mathbf{D}, \cdot)$.

The following properties of this function will be useful.

Lemma 1. *The function $\bar{g}$ as defined in Definition 1 is non-increasing and takes its values in $\mathbb{R}^+$.*

Proof of Lemma 1: The sublevel sets of g are nested, hence $\bar{g}$ is non-decreasing. Since g is coercive, its sublevel sets are bounded, hence $\bar{g}$ takes its values in $\mathbb{R}^+$. ∎

Note that Assumption **A4** is a convenience that should not be taken too literally, as penalty functions that fulfill **A1**-**A3** and have a global minimum at $\mathbf{0}$ can be manipulated to fulfill **A4** by subtracting the value of g at $\mathbf{0}$.

C. Lipschitz property under joint assumptions on g and $\mathfrak{D}$

While Assumptions **A1-A4** cover a wide range of penalty functions, they do not cover popular penalties related to the ℓ^0 quasi-norm, such as the indicator function (also called characteristic function) of k-sparse vectors:

$$\chi_{k\text{-sparse}}(\boldsymbol{\alpha}) \triangleq \begin{cases} 0, & \text{if } \|\boldsymbol{\alpha}\|_0 \leq k, \\ +\infty, & \text{otherwise;} \end{cases}$$

Vainsencher *et al.* [20] deal with the latter case under an incoherence assumption on $\mathbf{D}$, i.e., by restricting $\mathfrak{D}$ to be the class of dictionaries with small (cumulative) coherence. A careful study of their technique shows that the results are actually valid under an assumption related to the well-known restricted isometry property (RIP) [3], [32]. In fact, while the RIP is usually expressed in its symmetric form, only its lower bound actually plays a role in the considered context. This justifies the following definition.

Definition 2 (Lower RIP)**.** For $k \leq \min(m, d)$, we denote $\delta_k(\mathbf{D})$ the smallest $0 < \delta \leq 1$ such that for all k-sparse vectors $\boldsymbol{\alpha}$

$$(1 - \delta)\|\boldsymbol{\alpha}\|_2^2 \leq \|\mathbf{D}\boldsymbol{\alpha}\|_2^2.$$

For $k \leq \min(m,d)$ and $0 < \delta < 1$ we define the compact set

$$\mathfrak{D}_{\text{LRIP}}(k,\delta) \triangleq \left\{ \mathbf{D} \in \mathbb{R}^{m\times d} \,:\, \delta_k(\mathbf{D}) \leq \delta \right\}. \tag{5}$$

Similarly, Assumptions **A1-A4** do not cover the indicator function of non-negative coefficients,

$$\chi_{\text{NMF}}(\boldsymbol{\alpha}) \triangleq \begin{cases} 0, & \text{if } \alpha_i \geq 0, \forall 1 \leq i \leq d, \\ +\infty, & \text{otherwise;} \end{cases}$$

which is typically used in conjunction with a non-negativity constraint on $\mathbf{D}$ as well, namely $\mathbf{D} \in \mathfrak{D}_{\text{NMF}}$, where

$$\mathfrak{D}_{\text{NMF}} \triangleq \left\{ \mathbf{D} \in \mathbb{R}^{m\times d} \,:\, d_{ij} \geq 0,\ \forall i,j \right\}. \tag{6}$$

To unify the treatment of such penalty functions not covered by Assumptions **A1-A4**, we develop complementary results based on joint assumptions on g and the constraint set $\mathfrak{D}$:

- **B1:** $g = \chi_{\mathcal{K}}$ is the indicator function of a set $\mathcal{K}$;
- **B2:** there is $\kappa > 0$ such that for any $\boldsymbol{\alpha} \in \mathcal{K}$ and $\mathbf{D} \in \mathfrak{D}$,

$$\kappa \|\boldsymbol{\alpha}\|_1^2 \leq \|\mathbf{D}\boldsymbol{\alpha}\|_2^2$$

- **B3:** $\mathcal{K}$ contains the origin: $\mathbf{0} \in \mathcal{K}$.

Note that **B2** is a type of restricted eigenvalue condition, see e.g., [33, Equation (30)].

These assumptions hold for $g = \chi_{k\text{-sparse}}$ and $\mathfrak{D} = \mathfrak{D}_{\text{LRIP}}(k,\delta)$ with $\kappa = (1-\delta)/d$, and they will be shown to also hold for $g = \chi_{\text{NMF}}$ and $\mathfrak{D} = \mathfrak{D}_{\text{NMF}}(\kappa) \subset \mathfrak{D}_{\text{NMF}}$ where

$$\mathfrak{D}_{\text{NMF}}(\kappa) \triangleq \left\{ \mathbf{D} \in \mathbb{R}^{m\times d} \,:\, d_{ij} \geq 0,\ \|\mathbf{d}_j\|_2^2 \geq \kappa d,\ \forall i,j \right\}. \tag{7}$$

Under Assumptions **B1-B3** we show that the function $\mathbf{D} \mapsto F_{\mathbf{X}}(\mathbf{D})$ is Lipschitz over $\mathfrak{D}$. Its Lipschitz constant can be expressed as $L_{\mathbf{X}}(\bar{g})$ (see Eq. (4)) where $\bar{g}$ is defined in the context of assumptions **B1-B3** as:

Definition 3. For penalty functions g and $\mathfrak{D}$ satisfying **B1-B3**, we define

$$\bar{g}(t) \triangleq 2\sqrt{2t/\kappa}$$

Occasionally, an additional assumption will be helpful:

- **B4:** $\mathfrak{D}$ is convex.

D. Role of the class $\mathfrak{P}$ of probability distributions

Finally, the results rely on two assumptions on the probability distribution $\mathbb{P}$ from which the training samples $\mathbf{x}_i$ are assumed drawn i.i.d.:

- first, we need to control the Lipschitz constant $L_{\mathbf{X}}(\bar{g})$ when the sample size n is large.
- second, given $\mathbf{D}$, we need to control the concentration of the empirical average $F_{\mathbf{X}}(\mathbf{D})$ around its expectation;

By the law of large numbers, the first condition holds under assumption

- **C1:** bounded moment

$$L_{\mathbb{P}}(\bar{g}) \triangleq \mathbb{E}_{\mathbf{x}\sim\mathbb{P}} \|\mathbf{x}_i\|_2 \bar{g}\left(\tfrac{\|\mathbf{x}_i\|_2^2}{2}\right) < +\infty.$$

We will see on many examples (Table I) that $t\bar{g}(t^2/2) \propto t^2$ or $t\bar{g}(t^2/2) \propto t^3$, hence this is a relatively mild condition. From a quantitative perspective our results will be expressed using

$$\Lambda_n(L,\bar{g}) \triangleq \mathbb{P}\left(\tfrac{1}{n}\sum_{i=1}^n \|\mathbf{x}_i\|_2 \bar{g}\left(\tfrac{\|\mathbf{x}_i\|_2^2}{2}\right) > L \right). \tag{8}$$

By abuse of notation we will simply write $\Lambda_n(L)$ and will exploit the fact that $\lim_{n\to\infty} \Lambda_n(L) = 0$ for $L > L_{\mathbb{P}}(\bar{g})$.

The second condition is measured through

$$\Gamma_n(\gamma) \triangleq \sup_{\mathbf{D}\in\mathfrak{D}} \mathbb{P}\left(\left| \tfrac{1}{n}\sum_{i=1}^n f_{\mathbf{x}_i}(\mathbf{D}) - \mathbb{E} f_{\mathbf{x}}(\mathbf{D}) \right| > \gamma \right). \tag{9}$$

As discussed in Section VI, our main results exploit **C1** and

- **C2:** there are $c > 0$ and $T \in (0, +\infty]$ such that

$$\Gamma_n(c\tau) \leq 2\exp(-n\tau^2),\ \forall 0 \leq \tau \leq T,\ \forall n.$$

As shown in Section VI, this covers the case of probability distributions on the unit sphere in $\mathbb{R}^m$ (see, e.g., [20]),

$$\mathfrak{S}^{m-1} \triangleq \{\mathbf{x} \in \mathbb{R}^m \,:\, \|\mathbf{x}\|_2 = 1\},$$

and more generally on Euclidean balls B_R of given radius R,

Definition 4 (Probability distributions on a ball)**.** The set of probability distribution within a ball of radius R is given by

$$\mathfrak{P}_{B_R} \triangleq \{\mathbb{P} \,:\, \mathbb{P}(\|\mathbf{x}\|_2 \leq R) = 1\}.$$

For $\mathbb{P} \in \mathfrak{P}_{B_R}$, **C2** holds with $c = R^2/\sqrt{8}$, $T = +\infty$, and **C1** holds, with $\Lambda_n(L) = 0$ for $L = R\bar{g}(R^2/2)$ (Lemma 11).

Assumption **C2** also covers the following classes which contain the sub-Gaussian sparse signal model proposed in [26]. More details will be given in Section IV.

Definition 5. A distribution is in $\mathfrak{P}_A$, $A > 0$ if

$$\mathbb{P}(\|\mathbf{x}\|_2^2 \geq At) \leq \exp(-t),\ \forall t \geq 1. \tag{10}$$

For $\mathbb{P} \in \mathfrak{P}_A$, **C2** holds with $c = 12A$, $T = 1$ (Lemma 12), and **C1** holds as soon as $\bar{g}$ has at most polynomial growth.

E. Main result

Our main result is obtained using a standard union bound argument. The details are in Section VI. In short, under assumptions **A1-A4** or **B1-B3**, together with **C1-C2**, we show:

Theorem 1. *Consider $L > L_{\mathbb{P}}(\bar{g})$ and define*

$$\beta \triangleq h \cdot \max\left(\log \tfrac{2LC}{c}, 1\right), \tag{11}$$

$$\eta_n(g,\mathfrak{D},\mathfrak{P}) \triangleq 3c\sqrt{\tfrac{\beta\log n}{n}} + c\sqrt{\tfrac{\beta+x}{n}}. \tag{12}$$

Then, given $0 \leq x \leq nT^2 - \beta\log n$ we have: except with probability at most $\Lambda_n(L) + 2e^{-x}$,

$$\sup_{\mathbf{D}\in\mathfrak{D}} |F_{\mathbf{X}}(\mathbf{D}) - \mathbb{E}_{\mathbf{x}} f_{\mathbf{x}}(\mathbf{D})| \leq \eta_n(g,\mathfrak{D},\mathfrak{P}). \tag{13}$$

Note that $\Lambda_n(L)$ is primarily characterized by the penalty function g and the class of probability distributions $\mathfrak{P}$ (see (8)*), while the constants $C, h \geq 1$ depend on the class of dictionaries $\mathfrak{D}$, see* (3)*, and $c > 0$, $0 < T \leq \infty$ depend on the class of probability distributions $\mathfrak{P}$, see* **C2***.*

The constant $3c$ in the first term of the right hand side of (12) is improved to $2c$ for penalties that satisfy **A1-A4** or **B1-B4**, see Lemma 10 in Section VI where Theorem 1 is proved.

IV. Examples

We now discuss concrete settings covered by our results. More detailed worked examples will be given in Section VIII.

A. Penalty functions satisfying A1-A4

Many classical penalty functions are covered by Assumptions **A1-A4**: norms, quasi-norms, their powers, indicator functions of compact sets containing the origin, and more.

As a side note, if the penalty function g is invariant by sign flips (this is the case, e.g., for standard norms, mixed norms, etc.), one may notice that $\bar{g}(t)$ is related to the Fenchel conjugate [34] of the function $h(\boldsymbol{\alpha}) = \chi_{g(\boldsymbol{\alpha})\leq t}$, which reads as

$$h^\star(\mathbf{q}) \triangleq \sup_{\boldsymbol{\alpha}} \left(\langle \mathbf{q}, \boldsymbol{\alpha}\rangle - h(\boldsymbol{\alpha})\right)$$

since $\bar{g}(t) = h^\star(\mathbf{1})$ where $\mathbf{1}$ is the vectors with all ones.

1) Norms and quasi-norms: All standard ℓ^p norms ($1 \leq p \leq \infty$), quasi-norms ($0 < p < 1$), and mixed $\ell^{p,q}$ norms used in structured sparse regularization satisfy **A1-A4**, and we have $\bar{g}(t) = C_g t$ with

$$C_g = \sup_{\boldsymbol{\alpha}\neq\mathbf{0}} \frac{\|\boldsymbol{\alpha}\|_1}{g(\boldsymbol{\alpha})} < \infty$$

For example, for $g(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_p$ with $\boldsymbol{\alpha} \in \mathbb{R}^d$, we have

$$C_g = d^{(1-1/p)_+}$$

where $(\cdot)_+ \triangleq \max(\cdot, 0)$. We let the reader check that for mixed $\ell^{1,2}$ norms, $\|\boldsymbol{\alpha}\|_{1,2} = \sum_i \|\boldsymbol{\alpha}_{J_i}\|_2$, where the sets J_i partition the index set $[\![1; d]\!]$, we have $C_g = j^{1/2}$ where $j = \max_i \sharp J_i$.

2) Indicator functions of compact sets: Assumptions **A1-A3** also hold for the indicator function of a compact set $\mathcal{K}$,

$$g(\boldsymbol{\alpha}) = \chi_{\boldsymbol{\alpha}\in\mathcal{K}} \triangleq \begin{cases} 0, & \text{if } \boldsymbol{\alpha} \in \mathcal{K}, \\ +\infty, & \text{otherwise.} \end{cases}$$

For such penalties we have $\bar{g}(t) = \sup_{\boldsymbol{\alpha}\in\mathcal{K}} \|\boldsymbol{\alpha}\|_1$ for all $t \geq 0$. Assumption **A4** further requires that the compact set contains the origin, $\mathbf{0} \in \mathcal{K}$.

In particular, assumptions **A1-A4** hold for the indicator function of a ball defined by a (quasi-)norm, i.e., with $\mathcal{K} = \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\| \leq \lambda\}$. For an ℓ^p (quasi-)norm, we have $\bar{g}(t) = d^{(1-1/p)_+}\lambda$.

3) More examples: Assumptions **A1-A4** hold when g is a power of a (quasi-)norm, $g(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}/\lambda\|_p^r$, with $r > 0$, leading to $\bar{g}(t) = d^{(1-1/p)_+}\lambda t^{1/r}$. Note that the indicator function of $\mathcal{K} = \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\| \leq \lambda\}$ reads as the limit when $r \to \infty$.

There are of course measures that do not fit in our framework, such as the arctangent sparsity measure, which mimics the ℓ^0 one [35] and is defined as

$$g_{\mathrm{atan}}(\boldsymbol{\alpha}) \triangleq \sum_{j=1}^{d} \mathrm{arc\,tan}^2\left(\alpha_j\right).$$

It is not coercive and does thereby not meet assumption **A3**.

B. Penalty functions and constraint sets that satisfy B1-B3

Consider $g(\boldsymbol{\alpha}) = \chi_{k\text{-sparse}}(\boldsymbol{\alpha}) = \chi_{\mathcal{K}}$ with the set $\mathcal{K} = \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\|_0 \leq k\}$ and $\mathfrak{D} = \mathfrak{D}_{\mathrm{LRIP}}(k, \delta)$. Assumptions **B1** and **B3** obviously hold since $\mathcal{K}$ is a finite union of subspaces. Moreover, by definition of $\mathfrak{D}_{\mathrm{LRIP}}(k, \delta)$, for any $\boldsymbol{\alpha} \in \mathcal{K}$ and $\mathbf{D} \in \mathfrak{D}$ we have

$$\|\mathbf{D}\boldsymbol{\alpha}\|_2^2 \geq (1-\delta)\|\boldsymbol{\alpha}\|_2^2 \geq \tfrac{1-\delta}{k}\|\boldsymbol{\alpha}\|_1^2$$

where the rightmost inequality follows from the fact that $\boldsymbol{\alpha}$ is k-sparse. Hence assumption **B2** holds with $\kappa = (1-\delta)/k$.

Consider now $g(\boldsymbol{\alpha}) = \chi_{\mathrm{NMF}}(\boldsymbol{\alpha}) = \chi_{\mathcal{K}}$ with $\mathcal{K} = \{\boldsymbol{\alpha} : \alpha_i \geq 0,\ 1 \leq i \leq d\}$, and $\mathfrak{D} = \mathfrak{D}_{\mathrm{NMF}}(\kappa)$. Assumptions **B1** and **B3** obviously hold since $\mathcal{K}$ is an orthant. Moreover, since $\mathbf{D} \in \mathfrak{D}$ has non-negative entries[1], for any $\boldsymbol{\alpha} \in \mathcal{K}$ we have

$$\begin{aligned}\|\mathbf{D}\boldsymbol{\alpha}\|_2^2 &= \sum_{i,j=1}^{d} \alpha_i\alpha_j\langle \mathbf{d}_i, \mathbf{d}_j\rangle \geq \sum_{i=1}^{d} \alpha_i^2\|\mathbf{d}_i\|_2^2 \\ &\geq \|\boldsymbol{\alpha}\|_2^2 \cdot \min_{1\leq j\leq d} \|\mathbf{d}_j\|_2^2 \geq \|\boldsymbol{\alpha}\|_1^2 \cdot \min_{1\leq j\leq d} \|\mathbf{d}_j\|_2^2/d.\end{aligned}$$

Hence assumption **B2** holds.

Table I summarizes, for standard penalty functions satisfying either **A1-A4** or **B1-B3**, the expression of $\bar{g}$ and that of $t\bar{g}(t^2/2)$ which appears in the expression of $L_{\mathbf{X}}(\bar{g})$ (see (4)).

C. Covering numbers

Table II summarizes the covering numbers of the structured classes of dictionaries considered in Section VII. The provided covering numbers all depend on the signal dimension m, which may lead to sub-optimal sample complexity estimates for high-dimensional problems, i.e., when $m \gg d$. This issue is discussed in section Section IX.

D. Probability distributions

While previous work [19], [20] only covers distributions on the unit sphere or in the volume of the unit ball, our results cover more complex models, such as the model of sparse signals with sub-Gaussian non-zero coefficients introduced in [26].

Definition 6 (Sub-Gaussian model of sparse signals)**.** Given a fixed reference dictionary $\mathbf{D}_0 \in \mathfrak{D}$, each noisy sparse signal $\mathbf{x} \in \mathbb{R}^m$ is built from the following steps:

1) *Support generation:* Draw uniformly without replacement k atoms out of the d available in $\mathbf{D}_0$. This procedure thus defines a support $J \triangleq \{j \in [\![1; d]\!] : \delta(j) = 1\}$ whose size is $|J| = k$, and where $\delta(j)$ denotes the indicator function equal to one if the j-th atom is selected, zero otherwise. Note that $\mathbb{E}[\delta(j)] = \frac{k}{d}$ and for $i \neq j$ we further have $\mathbb{E}[\delta(j)\delta(i)] = \frac{k(k-1)}{d(d-1)}$.
2) *Coefficient generation:* Define a sparse vector $\boldsymbol{\alpha}_0 \in \mathbb{R}^d$ 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; otherwise, we assume there exists some $\sigma > 0$ such that for $j \in J$ we have, for all $t \in$

[1] In fact, as in [19] we observe it is sufficient to have $\langle \mathbf{d}_i, \mathbf{d}_j\rangle \geq 0$, $i \neq j$

TABLE I
PENALTY FUNCTIONS $g : \mathbb{R}^d \to \mathbb{R}$ AND ASSOCIATED $\bar{g}$ (FROM EITHER DEFINITION 1 OR DEFINITION 3)

$g(\boldsymbol{\alpha})$	remark	$\bar{g}(t)$	$t\bar{g}(t^2/2)$
(quasi)norm $g(\boldsymbol{\alpha})$	$C_g \triangleq \sup_{\boldsymbol{\alpha}\neq\mathbf{0}} \|\boldsymbol{\alpha}\|_1/g(\boldsymbol{\alpha})$	$C_g t$	$\frac{C_g}{2}\cdot t^3$
$\|\boldsymbol{\alpha}/\lambda\|_p^r$,	$0<p\leq\infty$, $0<r<\infty$	$d^{(1-1/p)_+}\cdot\lambda\cdot t^{1/r}$	$d^{(1-1/p)_+}\cdot\lambda\cdot t\cdot(t^2/2)^{1/r}$
$\chi_{\|\boldsymbol{\alpha}\|_p\leq\lambda}$,	$0<p\leq\infty$	$d^{(1-1/p)_+}\cdot\lambda$	$d^{(1-1/p)_+}\cdot\lambda\cdot t$
$\chi_{k\text{-sparse}}(\boldsymbol{\alpha})$	on $\mathfrak{D}_{\mathrm{LRIP}}(k,\delta)$, see (5)	$2\sqrt{k/(1-\delta)}\sqrt{2t}$	$2t^2\sqrt{k/(1-\delta)}$
$\chi_{\mathrm{NMF}}(\boldsymbol{\alpha})$	on $\mathfrak{D}_{\mathrm{NMF}}(\kappa)$, see (7)	$2\sqrt{2t/\kappa}$	$2t^2/\sqrt{\kappa}$

TABLE II
CLASSES OF STRUCTURED DICTIONARIES $\mathfrak{D}$ WITH COVERING NUMBERS $\mathcal{N}(\mathfrak{D},\epsilon)\leq(C/\epsilon)^h$, $0<\epsilon\leq 1$. SEE SECTION VII

$\mathfrak{D}$	exponent h	constant C
Unit norm $\mathfrak{D}(m,d)$	md	3
Separable $\mathfrak{D}_{\mathrm{sep}}$	$\sum_i m_i d_i$	3
Sparse $\mathfrak{D}_{\mathrm{sparse}}(m,d)$	sd	$3\binom{m}{s}^{1/s}$
Orthogonal $\mathfrak{D}_{\mathrm{orth}}(m)$	$m(m-1)/2$	$3\pi e^{\pi}$
Stiefel $\mathfrak{D}_{\mathrm{St}}(m,d)$	$md-d(d+1)/2$	$3\pi e^{\pi}$
Stiefel Tensor $\mathfrak{D}_{\otimes\mathrm{St}}$	$\sum_i m_i d_i - d_i(d_i+1)/2$	$3\pi e^{\pi}$

$\mathbb{R}$, $\mathbb{E}\{\exp(t[\boldsymbol{\alpha}_0]_j)\}\leq\exp(\sigma^2t^2/2)$. We denote σ_α the smallest value of σ such that this property holds. For background on sub-Gaussianity, see [27], [36].

3) *Noise:* Eventually generate the signal $\mathbf{x}=\mathbf{D}_0\boldsymbol{\alpha}_0+\boldsymbol{\epsilon}$, where the entries of the additive noise $\boldsymbol{\epsilon}\in\mathbb{R}^m$ are assumed i.i.d. sub-Gaussian with parameter σ_ε.

Remark 1. The model in [26] is more restricted: it assumes that, for $j\in J$, $|[\boldsymbol{\alpha}_0]_j|\geq\underline{\alpha}$ almost surely, where $\underline{\alpha}>0$.

The distribution $\mathbb{P}$ of this sub-Gaussian sparse signal model belongs to $\mathfrak{P}_A$ with $A=k\sigma_\alpha^2+m\sigma_\varepsilon^2$, as shown below (the argument can be originally found in [26]).

Lemma 2 (From [37])**.** *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 any $t\geq 1$, we have*

$$\mathbb{P}\Big(\|\mathbf{Az}\|_2^2>5\sigma^2\|\mathbf{A}\|_F^2t\Big)\leq\exp(-t).$$

Corollary 1. *Let $\mathbf{x}$ be a signal following the model of Definition 6. For any $t\geq 1$ we have*

$$\mathbb{P}\left(\|\mathbf{x}\|_2^2>5(k\sigma_\alpha^2+m\sigma_\varepsilon^2)t\right)\leq\exp(-t).$$

Proof: The considered norm can be expressed as follows

$$\|\mathbf{x}\|_2^2=\left\|\begin{bmatrix}\sigma_\alpha[\mathbf{D}_0]_J & \sigma_\varepsilon\mathbf{Id}\end{bmatrix}\begin{pmatrix}\frac{1}{\sigma_\alpha}[\boldsymbol{\alpha}_0]_J\\ \frac{1}{\sigma_\varepsilon}\boldsymbol{\epsilon}\end{pmatrix}\right\|_2^2.$$

The result is a direct application of Lemma 2 conditioned to the draw of J, and the observation that

$$\left\|\begin{bmatrix}\sigma_\alpha[\mathbf{D}_0]_J & \sigma_\varepsilon\mathbf{Id}\end{bmatrix}\right\|_F^2=\|[\mathbf{D}_0]_J\|_F^2\cdot\sigma_\alpha^2+m\sigma_\varepsilon^2=k\sigma_\alpha^2+m\sigma_\varepsilon^2.$$

The bound being independent of J, the result is also true without conditioning. ∎

Section VIII will detail worked examples with results for specific combinations of dictionary class $\mathfrak{D}$, penalty g, and probability distribution $\mathbb{P}$, relating our results to prior work.

V. LIPSCHITZ CONTINUITY OF $F_{\mathbf{X}}$

Under appropriate assumptions on the penalty function g, we prove below that the function $F_{\mathbf{X}}(\mathbf{D})$ is Lipschitz continuous with a controlled constant $L_{\mathbf{X}}(\bar{g})$. We begin by a one-sided Lipschitz property with an additional quadratic term that we will soon get rid of.

Lemma 3. *Let $\|\cdot\|$ be some norm for $m\times d$ matrices and $\|\cdot\|_\star$ its dual norm*[2]*. For any $\epsilon>0$, the set*

$$\mathfrak{A}_\epsilon(\mathbf{X},\mathbf{D})=\{[\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_n]:\boldsymbol{\alpha}_i\in\mathbb{R}^d,\ \mathcal{L}_{\mathbf{x}_i}(\mathbf{D},\boldsymbol{\alpha}_i)\leq f_{\mathbf{x}_i}(\mathbf{D})+\epsilon\}.$$

is not empty, and for any $\mathbf{D}'$ we have

$$F_{\mathbf{X}}(\mathbf{D}')\leq F_{\mathbf{X}}(\mathbf{D})+L_{\mathbf{X}}(\mathbf{D})\cdot\|\mathbf{D}'-\mathbf{D}\|+C_{\mathbf{X}}(\mathbf{D})\cdot\|\mathbf{D}'-\mathbf{D}\|^2.$$

with

$$L_{\mathbf{X}}(\mathbf{D})\triangleq\inf_{\epsilon>0}\sup_{\mathbf{A}\in\mathfrak{A}_\epsilon}\tfrac{1}{n}\cdot\|(\mathbf{X}-\mathbf{DA})\mathbf{A}^\top\|_\star,\qquad(14)$$

$$C_{\mathbf{X}}(\mathbf{D})\triangleq\inf_{\epsilon>0}\sup_{\mathbf{A}\in\mathfrak{A}_\epsilon}\tfrac{C}{2n}\sum_{i=1}^n\|\boldsymbol{\alpha}_i\|_1^2.\qquad(15)$$

The constant C in (15) *depends only on the norm $\|\cdot\|$ and the dimensions m,d. When $\|\cdot\|=\|\cdot\|_{1\to 2}$ we have $C=1$. When $\mathfrak{A}_0\neq\emptyset$, $L_{\mathbf{X}}(\mathbf{D})$ and $C_{\mathbf{X}}(\mathbf{D})$ are suprema over $\mathfrak{A}_0$.*

The choice of the ℓ^1 norm in the bound (15) is an arbitrary convenience with no impact on the nature of the results to follow. More important will be the choice of the metric $\|\cdot\|$ which will be related to how we measure covering numbers for $\mathfrak{D}$. This will be discussed later.

Proof: Fix $\epsilon>0$ and $\mathbf{A}\in\mathfrak{A}_\epsilon$. From the identity

$$\mathcal{L}_{\mathbf{X}}(\mathbf{D}',\mathbf{A})=\mathcal{L}_{\mathbf{X}}(\mathbf{D},\mathbf{A})+\tfrac{1}{n}\langle\mathbf{X}-\mathbf{DA},(\mathbf{D}'-\mathbf{D})\mathbf{A}\rangle_F+\tfrac{1}{2n}\|(\mathbf{D}'-\mathbf{D})\mathbf{A}\|_F^2\qquad(16)$$

[2] $\|\mathbf{U}\|_\star\triangleq\sup_{\mathbf{D},\|\mathbf{D}\|\leq 1}\langle\mathbf{U},\mathbf{D}\rangle_F$ with $\langle\cdot,\cdot\rangle_F$ the Frobenius inner product.

and the definition of the dual norm we have

$$\frac{|\langle \mathbf{X}-\mathbf{D}\mathbf{A}, (\mathbf{D}'-\mathbf{D})\mathbf{A}\rangle_F|}{\|\mathbf{D}'-\mathbf{D}\|} \leq \left\|(\mathbf{X}-\mathbf{D}\mathbf{A})\mathbf{A}^\top\right\|_\star .$$

Moreover, by the equivalence of all norms on the finite dimensional space of $m \times d$ matrices, there is a constant C (equal to one when $\|\cdot\| = \|\cdot\|_{1\to 2}$) such that

$$\begin{aligned}\tfrac{1}{2n}\|(\mathbf{D}'-\mathbf{D})\mathbf{A}\|_F^2 &\leq \|\mathbf{D}'-\mathbf{D}\|_{1\to 2}^2 \cdot \tfrac{1}{2n}\sum_{i=1}^n \|\boldsymbol{\alpha}_i\|_1^2 \\ &\leq \|\mathbf{D}'-\mathbf{D}\|^2 \cdot \tfrac{C}{2n}\sum_{i=1}^n \|\boldsymbol{\alpha}_i\|_1^2.\end{aligned}$$

Therefore, we deduce from (16) that

$$\begin{aligned}F_{\mathbf{X}}(\mathbf{D}') = \inf_{\mathbf{A}} \mathcal{L}_{\mathbf{X}}(\mathbf{D}', \mathbf{A}) &\leq \sup_{\mathbf{A}\in\mathfrak{A}_\epsilon} \mathcal{L}_{\mathbf{X}}(\mathbf{D}', \mathbf{A}) \\ &\leq \sup_{\mathbf{A}\in\mathfrak{A}_\epsilon} \mathcal{L}_{\mathbf{X}}(\mathbf{D}, \mathbf{A}) \\ &\quad + \sup_{\mathbf{A}\in\mathfrak{A}_\epsilon} \tfrac{1}{n}\|(\mathbf{X}-\mathbf{D}\mathbf{A})\mathbf{A}^\top\|_\star \cdot \|\mathbf{D}'-\mathbf{D}\| \\ &\quad + \sup_{\mathbf{A}\in\mathfrak{A}_\epsilon} \tfrac{C}{2n}\sum_{i=1}^n \|\boldsymbol{\alpha}_i\|_1^2 \cdot \|\mathbf{D}'-\mathbf{D}\|^2 \qquad (17)\end{aligned}$$

By definition, $\sup_{\mathbf{A}\in\mathfrak{A}_\epsilon} \mathcal{L}_{\mathbf{X}}(\mathbf{D}, \mathbf{A}) \leq F_{\mathbf{X}}(\mathbf{D}) + \epsilon$. We conclude by taking the infimum of (17) over all $\epsilon > 0$. ∎

Corollary 2. *Let $\|\cdot\|$ be some norm for $m\times d$ matrices and $\|\cdot\|_\star$ its dual norm, and $\mathfrak{D}$ a class of dictionaries. If*

$$\sup_{\mathbf{D}\in\mathfrak{D}} L_{\mathbf{X}}(\mathbf{D}) \leq L_{\mathbf{X}}(\mathfrak{D}); \qquad (18)$$

$$\sup_{\mathbf{D}\in\mathfrak{D}} C_{\mathbf{X}}(\mathbf{D}) \leq C_{\mathbf{X}}(\mathfrak{D}), \qquad (19)$$

then for any $\mathbf{D} \neq \mathbf{D}' \in \mathfrak{D}$ we have

$$\frac{|F_{\mathbf{X}}(\mathbf{D}') - F_{\mathbf{X}}(\mathbf{D})|}{\|\mathbf{D}'-\mathbf{D}\|} \leq L_{\mathbf{X}}(\mathfrak{D}) \cdot \left(1 + \tfrac{C_{\mathbf{X}}(\mathfrak{D})}{L_{\mathbf{X}}(\mathfrak{D})}\|\mathbf{D}'-\mathbf{D}\|\right).$$

In particular, if we can establish the bounds (18) and (19) then $F_{\mathbf{X}}$ is uniformly locally Lipschitz (with respect to $\|\cdot\|$) over the class $\mathfrak{D}$ for any constant $L > L_{\mathbf{X}}(\mathfrak{D})$.

A. Uniform local Lipschitz constants

Given Corollary 2 we now bound $L_{\mathbf{X}}(\mathbf{D})$ and $C_{\mathbf{X}}(\mathbf{D})$ when the norm $\|\cdot\|$ is the operator norm $\|\cdot\|_{1\to 2}$. This is motivated by the fact that standard sparse coding is often performed with dictionaries constrained to have unit ℓ^2 norm, which are closely connected with the unit sphere of this norm. With this choice, the shape of $\bar{g}$ in Definition 1 should appear natural in the context of the proof of Lemma 5 below.

Extensions to other definitions of $\bar{g}$ can be envisioned if we change the metric. In particular, when g itself is a norm, one could consider the operator norm $\|\cdot\|_{g\to 2}$, at the price of possibly more tricky estimates of covering numbers.

1) Penalty functions satisfying **A1-A3***, with arbitrary* $\mathbf{D}$*:*

Lemma 4. *If g satisfies Assumptions* **A1-A3***, then the set $\mathfrak{A}_0$ is not empty, and it is bounded.*

Proof: By the non-negativity (**A1**) and coercivity (**A3**) of $g(\cdot)$, $\mathcal{L}_{\mathbf{X}}(\mathbf{D}, \mathbf{A})$ is non-negative and $\lim_{k\to\infty} \mathcal{L}_{\mathbf{X}}(\mathbf{D}, \mathbf{A}_k) = \infty$ whenever $\lim_{k\to\infty} \|\mathbf{A}_k\| = \infty$. Therefore, the function $\mathbf{A} \mapsto \mathcal{L}_{\mathbf{X}}(\mathbf{D}, \mathbf{A})$ has bounded sublevel sets. Moreover, since g is lower semi-continuous (**A2**), then so is $\mathbf{A} \mapsto \mathcal{L}_{\mathbf{X}}(\mathbf{D}, \mathbf{A})$, therefore it attains its infimum value. ∎

Lemma 5 (Penalty functions that satisfy **A1-A3**)**.** *Under Assumptions* **A1-A3***, for any training set* $\mathbf{X}$ *and dictionary* $\mathbf{D}$*, the constants defined in* (14) *and* (15) *with the norm $\|\cdot\| = \|\cdot\|_{1\to 2}$ satisfy the bounds*

$$L_{\mathbf{X}}(\mathbf{D}) \leq \tfrac{1}{n}\sum_{i=1}^n \sqrt{2 f_{\mathbf{x}_i}(\mathbf{D})} \cdot \bar{g}(f_{\mathbf{x}_i}(\mathbf{D})) \qquad (20)$$

$$C_{\mathbf{X}}(\mathbf{D}) \leq \tfrac{1}{2n}\sum_{i=1}^n [\bar{g}(f_{\mathbf{x}_i}(\mathbf{D}))]^2. \qquad (21)$$

Proof: By Lemma 4 the set $\mathfrak{A}_0$ is not empty, and by definition, for any $\mathbf{A} \triangleq [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n] \in \mathfrak{A}_0$,

$$\tfrac{1}{2}\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + g(\boldsymbol{\alpha}_i) \leq f_{\mathbf{x}_i}(\mathbf{D}). \qquad (22)$$

Therefore, **A1** (non-negativity of g) implies

$$0 \leq g(\boldsymbol{\alpha}_i) \leq f_{\mathbf{x}_i}(\mathbf{D}) \qquad (23)$$

$$\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2 \leq \sqrt{2 f_{\mathbf{x}_i}(\mathbf{D})}, \qquad (24)$$

for $i = 1, \ldots, n$. Combined with Definition 1, (23) implies

$$\|\boldsymbol{\alpha}_i\|_1 \leq \bar{g}(f_{\mathbf{x}_i}(\mathbf{D})), \qquad (25)$$

$$C_{\mathbf{X}}(\mathbf{D}) \leq \tfrac{1}{2n}\sum_{i=1}^n [\bar{g}(f_{\mathbf{x}_i}(\mathbf{D}))]^2, \qquad (26)$$

which proves the bound in (21). We now prove the inequality (20). For any $m\times d$ matrix $\boldsymbol{\Delta}$ we have

$$\begin{aligned}\langle \mathbf{X}-\mathbf{D}\mathbf{A}, \boldsymbol{\Delta}\mathbf{A}\rangle_F &= \sum_{i=1}^n \langle \mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i, \boldsymbol{\Delta}\boldsymbol{\alpha}_i\rangle \\ &\leq \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2 \cdot \|\boldsymbol{\Delta}\boldsymbol{\alpha}_i\|_2 \\ &\leq \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2 \cdot \|\boldsymbol{\Delta}\|_{1\to 2} \cdot \|\boldsymbol{\alpha}_i\|_1.\end{aligned} \qquad (27)$$

This allows us to provide an upper bound for the dual norm by exploiting Equations (24) and (25)

$$\tfrac{1}{n}\|(\mathbf{X}-\mathbf{D}\mathbf{A})\mathbf{A}^\top\|_\star \leq \tfrac{1}{n}\sum_{i=1}^n \sqrt{2 f_{\mathbf{x}_i}(\mathbf{D})} \cdot \bar{g}(f_{\mathbf{x}_i}(\mathbf{D})) \qquad (28)$$

which shows (20). ∎

Corollary 3 (Penalty functions that satisfy **A1-A4**)**.** *Under Assumptions* **A1-A4***, for any training set* $\mathbf{X}$ *and dictionary* $\mathbf{D}$*,*

the constants defined in (14) *and* (15) *with the norm* $\|\cdot\| = \|\cdot\|_{1\to 2}$ *satisfy the bounds*

$$L_{\mathbf{X}}(\mathbf{D}) \leq \tfrac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i\|_2 \cdot \bar{g}(\tfrac{1}{2}\|\mathbf{x}_i\|_2^2) \qquad \triangleq L_{\mathbf{X}}(\bar{g}) \qquad (29)$$

$$C_{\mathbf{X}}(\mathbf{D}) \leq \tfrac{1}{2n}\sum_{i=1}^{n} [\bar{g}(\|\mathbf{x}_i\|_2^2/2)]^2 \qquad \triangleq C_{\mathbf{X}}(\bar{g}). \qquad (30)$$

Proof: If the condition **A4** ($g(\mathbf{0}) = 0$) is fulfilled, we can extend Equations (23,24) to

$$0 \leq g(\boldsymbol{\alpha}_i) \leq f_{\mathbf{x}_i}(\mathbf{D}) \leq \mathcal{L}_{\mathbf{x}_i}(\mathbf{D}, \mathbf{0}) = \tfrac{1}{2}\|\mathbf{x}_i\|_2^2 \qquad (31)$$

$$\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2 \leq \sqrt{2 f_{\mathbf{x}_i}(\mathbf{D})} \leq \|\mathbf{x}_i\|_2, \qquad (32)$$

for $i = 1, \ldots, n$. Due to the fact that $\bar{g}$ is non-decreasing, the inequality in (30) follows directly from Equation (26).

It remains to show the inequality in (29), which is verified by plugging Equations (31,32) in (28). This yields the upper bound

$$\tfrac{1}{n}\sum_{i=1}^{n} \sqrt{2 f_{\mathbf{x}_i}(\mathbf{D})} \cdot \bar{g}(f_{\mathbf{x}_i}(\mathbf{D})) \leq \tfrac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i\|_2 \cdot \bar{g}(\tfrac{1}{2}\|\mathbf{x}_i\|_2^2)$$

which is independent of the constraint set $\mathfrak{D}$. ∎

2) Penalty functions and constraint sets that satisfy **B1-B3***:*

Lemma 6 (Penalty functions g and constraint sets $\mathfrak{D}$ that satisfy **B1-B3**)**.** *Under assumptions* **B1-B3***, for any training set* $\mathbf{X}$ *and dictionary* $\mathbf{D} \in \mathfrak{D}$*, the constants defined in* (14) *and* (15) *with the norm* $\|\cdot\| = \|\cdot\|_{1\to 2}$ *satisfy the bounds*

$$L_{\mathbf{X}}(\mathbf{D}) \leq \tfrac{2}{n\sqrt{\kappa}}\|\mathbf{X}\|_F^2 \qquad \triangleq L_{\mathbf{X}}(\bar{g}) \qquad (33)$$

$$C_{\mathbf{X}}(\mathbf{D}) \leq \tfrac{2}{n\kappa}\|\mathbf{X}\|_F^2 \qquad \triangleq C_{\mathbf{X}}(\bar{g}). \qquad (34)$$

Proof: Fix $\epsilon > 0$ and consider $\boldsymbol{\alpha}_i$, $i = 1, \ldots, n$ such that $\mathbf{A} \triangleq [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n] \in \mathfrak{A}_\epsilon$. One can check that by **B1**, Equations (22)-(24) hold up to an additive $O(\epsilon)$ term[3] on the right hand side, as well as (27). By **B3** we also have (31)-(32) with a similar additive term, yielding

$$\|\mathbf{D}\boldsymbol{\alpha}_i\|_2 \leq \|\mathbf{x}_i\|_2 + \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\| \leq 2\|\mathbf{x}_i\| + O(\epsilon). \qquad (35)$$

Moreover, by **B2**, we have

$$\|\boldsymbol{\alpha}_i\|_1^2 \leq \tfrac{\|\mathbf{D}\boldsymbol{\alpha}_i\|_2^2}{\kappa} \leq \frac{4\|\mathbf{x}_i\|_2^2}{\kappa} + O(\epsilon). \qquad (36)$$

Taking the infimum over $\epsilon > 0$ shows (34).

To prove (33) we combine (27) with (32) and (36), yielding an upper bound to $\frac{1}{n}\|(\mathbf{X} - \mathbf{D}\mathbf{A})\mathbf{A}^\top\|_\star$:

$$\tfrac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2 \cdot \|\boldsymbol{\alpha}_i\|_1 \leq \frac{2}{n\sqrt{\kappa}}\sum_{i=1}^{n} \|\mathbf{x}_i\|_2^2 + O(\epsilon). \quad (37)$$

∎

Observe the similarity between the expressions (33) and (4). It justifies Definition 3 which yields the identity $L_{\mathbf{X}}(\mathfrak{D}) = L_{\mathbf{X}}(\bar{g})$. With this notation we will be able to jointly cover the different penalty functions dealt with so far.

[3] $f(\epsilon) = O(\epsilon)$ if there exists $C < \infty$ such that $|f(\epsilon)| \leq C\epsilon$.

B. Main results on Lipschitz properties

We are ready to state the main results on the Lipschitz property of $F_{\mathbf{X}}(\mathbf{D})$. First we go for a *global* Lipschitz property. Note that Theorem 2 below is expressed over a *convex* class of dictionaries: under assumptions **A1-A4** the result is valid uniformly over $\mathbf{D} \in \mathbb{R}^{m\times d}$ (in particular, *it is not restricted to dictionaries with unit norm columns*)[4]; under assumptions **B1-B3**, we add the explicit assumption **B4** that $\mathfrak{D}$ is convex.

Theorem 2. *Assume either* **A1-A4** *or* **B1-B4***. Then, for any* $\mathbf{X}$*, and any* $\mathbf{D}, \mathbf{D}' \in \mathfrak{D}$*,*

$$|F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}')| \leq L_{\mathbf{X}}(\bar{g}) \cdot \|\mathbf{D} - \mathbf{D}'\|_{1\to 2}. \qquad (38)$$

Proof: Fix $\epsilon > 0$. By Corollaries 2 and 3 (resp. Lemma 6):

$$|F_{\mathbf{X}}(\mathbf{D}') - F_{\mathbf{X}}(\mathbf{D})| \leq (1+\epsilon) L_{\mathbf{X}}(\bar{g}) \cdot \|\mathbf{D}' - \mathbf{D}\|_{1\to 2} \quad (39)$$

whenever $\delta = \|\mathbf{D}' - \mathbf{D}\|_{1\to 2} \leq \frac{\epsilon L_{\mathbf{X}}(\bar{g})}{C_{\mathbf{X}}(\bar{g})}$.

When δ exceeds this bound, we choose an integer $k \geq 1$ such that $\delta/k \leq \frac{\epsilon L_{\mathbf{X}}(\bar{g})}{C_{\mathbf{X}}(\bar{g})}$ and define $\mathbf{D}_i = \mathbf{D} + i(\mathbf{D}' - \mathbf{D})/k$, $0 \leq i \leq k$. Note that this sequence is a priori constructed in the surrounding space of $\mathfrak{D}$. Under assumptions **B1-B4**, the convexity of $\mathfrak{D}$ (**B4**) ensures $\mathbf{D}_i \in \mathfrak{D}$. Under assumptions **A1-A4**, the local Lipschitz constant $L_{\mathbf{X}}(\bar{g})$ defined in (29) actually holds independently of a particular dictionary structure $\mathfrak{D}$. Hence, in both cases, since $\|\mathbf{D}_{i+1} - \mathbf{D}_i\|_{1\to 2} \leq \frac{\epsilon L_{\mathbf{X}}(\bar{g})}{C_{\mathbf{X}}(\bar{g})}$ for $i = 0, \ldots, k-1$, the bound we just obtained yields:

$$\begin{aligned} |F_{\mathbf{X}}(\mathbf{D}_{i+1}) - F_{\mathbf{X}}(\mathbf{D}_i)| &\leq (1+\epsilon)\ L_{\mathbf{X}}(\bar{g})\ \|\mathbf{D}_{i+1} - \mathbf{D}_i\|_{1\to 2} \\ &\leq (1+\epsilon)\ L_{\mathbf{X}}(\bar{g})\ \|\mathbf{D}' - \mathbf{D}\|_{1\to 2}/k \\ |F_{\mathbf{X}}(\mathbf{D}') - F_{\mathbf{X}}(\mathbf{D})| &\leq \textstyle\sum_{i=0}^{k-1} |F_{\mathbf{X}}(\mathbf{D}_{i+1}) - F_{\mathbf{X}}(\mathbf{D}_i)| \\ &\leq (1+\epsilon)\ L_{\mathbf{X}}(\bar{g})\ \|\mathbf{D}' - \mathbf{D}\|_{1\to 2}. \end{aligned}$$

Thus, the bound (39) can be extended to any pair $\mathbf{D}, \mathbf{D}' \in \mathfrak{D}$. Since the choice of $\epsilon > 0$ is arbitrary, Equation (38) follows. ∎

As an example, consider NMF expressed with $g = \chi_{\text{NMF}}$ and $\mathfrak{D}_{\text{NMF}\&\ell^1}$ the class of non-negative dictionaries with ℓ_1 normalized columns, used in topic models [24],

$$\mathfrak{D}_{\text{NMF}\&\ell^1} \triangleq \{\mathbf{D} : \mathbf{d}_i \in \mathbb{R}^m; \|\mathbf{d}_i\|_1 = 1;\ d_{ij} \geq 0\},$$

which is convex (**B4**) as it is the Cartesian product of d copies of the simplex. Since $\mathfrak{D} \subset \mathfrak{D}_{\text{NMF}}(\kappa)$ with $\kappa = 1/md$, assumptions **B1-B3** hold.

A slightly weaker result expressed in terms of *uniform local* Lipschitz property holds under **B1-B3** for *non-convex* $\mathfrak{D}$.

Theorem 3. *Under assumptions* **B1-B3***, for any training set* $\mathbf{X}$ *and any dictionaries* $\mathbf{D}' \neq \mathbf{D} \in \mathfrak{D}$ *we have*

$$\frac{|F_{\mathbf{X}}(\mathbf{D}') - F_{\mathbf{X}}(\mathbf{D})|}{\|\mathbf{D}' - \mathbf{D}\|_{1\to 2}} \leq L_{\mathbf{X}}(\bar{g}) \cdot \left(1 + \sqrt{\tfrac{1}{\kappa}}\|\mathbf{D}' - \mathbf{D}\|_{1\to 2}\right).$$

Proof: This is the direct result of Lemma 3. The upper bounds for $C_{\mathbf{X}}(\mathfrak{D}), L_{\mathbf{X}}(\mathfrak{D})$ provided in Lemma 6 yield the factor $\frac{C_{\mathbf{X}}(\bar{g})}{L_{\mathbf{X}}(\bar{g})} = \sqrt{\frac{1}{\kappa}}$. ∎

[4] This requirement will only arise from the need to have a finite covering number for the sample complexity estimate

Remark 2. The reader may wonder why we state a global Lipschitz property in Theorem 2 but only a uniform local one in Theorem 3. First, the triangle inequality argument does not seem to extend: the line joining $\mathbf{D}$ to $\mathbf{D}'$ cannot be cut in small segments so that each intermediate pair remains in $\mathfrak{D}$. In fact, there may even be several connected components in the class. Moreover, even in a given connected component, it is not clear what the length of the geodesics that would mimic such a triangle inequality argument is. Answering such questions would require a deeper investigation of the topology and geometry of $\mathfrak{D}$. Since Theorem 3 is sufficient for our needs, we leave them to future work.

C. Lipschitz-continuity for high-dimensional settings

When the signal dimension m is larger than the number d of dictionary elements, then the Lipschitz-continuity studied above is not appropriate as it leads to sample complexities that grow with m. Instead, we may consider Lipschitz-continuity with respect to $\mathbf{D}^\top\mathbf{D}$ and $\mathbf{D}^\top\mathbf{X}$, which we show below and use in Section IX to sketch improved results in such high-dimensional settings.

Lemma 7. *Under the same assumptions as in Lemma 3, for any $\mathbf{D}'$, we have*

$$F_{\mathbf{X}}(\mathbf{D}') \leq F_{\mathbf{X}}(\mathbf{D}) + C_{\mathbf{X}}^{(1)}(\mathbf{D})\cdot\|(\mathbf{D}')^\top\mathbf{D}' - \mathbf{D}^\top\mathbf{D}\|_F + C_{\mathbf{X}}^{(2)}(\mathbf{D})\cdot\|\mathbf{X}^\top\mathbf{D}' - \mathbf{X}^\top\mathbf{D}\|_F.$$

with

$$C_{\mathbf{X}}^{(1)}(\mathbf{D}) \triangleq \inf_{\epsilon>0}\sup_{\mathbf{A}\in\mathfrak{A}_\epsilon} \tfrac{1}{2n}\cdot\|\mathbf{A}\mathbf{A}^\top\|_F,$$

$$C_{\mathbf{X}}^{(2)}(\mathbf{D}) \triangleq \inf_{\epsilon>0}\sup_{\mathbf{A}\in\mathfrak{A}_\epsilon} \tfrac{1}{n}\cdot\|\mathbf{A}\|_F.$$

Proof: We follow the same principle as in the proof of Lemma 3, but we use the following equality

$$\mathcal{L}_{\mathbf{X}}(\mathbf{D}',\mathbf{A}) = \mathcal{L}_{\mathbf{X}}(\mathbf{D},\mathbf{A}) + \tfrac{1}{2n}\langle(\mathbf{D}')^\top\mathbf{D}' - \mathbf{D}^\top\mathbf{D}, \mathbf{A}\mathbf{A}^\top\rangle_F - \tfrac{1}{n}\langle(\mathbf{D}')^\top\mathbf{X} - \mathbf{D}^\top\mathbf{X}, \mathbf{A}\rangle_F.$$

Note that we use the Frobenius norm for this lemma, but that we could use any norm on matrices. ∎

D. Other losses

One can envision extensions of Lemmata 3 and 7 and their consequences to matrix factorization problems where the ℓ_2 data-fidelity term, $\frac{1}{2}\|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}\|_2^2$, is replaced by a more general loss $\ell(\mathbf{x},\mathbf{D}\boldsymbol{\alpha})$. In particular, Lemma 3 can be extended to losses $\ell(\cdot,\cdot)$ satisfying for any $\mathbf{x},\mathbf{y}$

$$\ell(\mathbf{x},\mathbf{y}+\mathbf{h}) \leq \ell(\mathbf{x},\mathbf{y}) + \langle\nabla_{\mathbf{y}}\ell(\mathbf{x},\mathbf{y}),\mathbf{h}\rangle + C\|\mathbf{h}\|_2^2,$$

i.e., with a local quadratic behavior. This would be reminiscent of the work of Negahban *et al* [33] which covers M-estimation problems where the loss, which is convex in the unknown parameter is regularized with a convex decomposable penalty. Here the considered problem is intrinsically non convex in $\mathbf{D}$.

A generalization of Corollary 3 would further require assuming that $\ell(\mathbf{x},\mathbf{y}) \geq 0$ for any $\mathbf{x},\mathbf{y}$, to replace (31) with $0 \leq g(\boldsymbol{\alpha}_i) \leq \ell(\mathbf{x},\mathbf{0})$, and identifying conditions on the loss ensuring that for some function $B(.)$, an analogue of the bound (32) holds: $\|\nabla_{\mathbf{y}}\ell(\mathbf{x}_i,\mathbf{D}\boldsymbol{\alpha}_i)\|_2 \leq B(\mathbf{x}_i)$. The resulting Lipschitz constant would read $L_{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^n B(\mathbf{x}_i)\bar{g}(\ell(\mathbf{x}_i,\mathbf{0}))$. The full characterization of such extensions and of the families of losses that can be considered is, however, beyond the scope of this paper.

VI. Sample Complexity (Proof of Theorem 1)

Given the global (resp. uniform local) Lipschitz property of $F_{\mathbf{X}}$, a standard route (see, e.g., [19], [20]) to control the sample complexity via a uniform convergence result is to rely on concentration of measure and covering numbers for the considered class of dictionaries (see Section VII for details on covering numbers).

A. Lipschitz property of the expected cost function

Mild assumptions on the quantities $\Lambda_n(L)$ and $\Gamma_n(\gamma)$ as defined in (8) and (9) are sufficient to control the Lipschitz constants of $F_{\mathbf{X}}$ and of its expectation:

Lemma 8. *Under assumptions* **A1-A4** *or* **B1-B4**, *and* **C1***:

1) *the function $\mathbf{D} \to F_{\mathbf{X}}(\mathbf{D})$ is Lipschitz with constant L , with probability at least $1-\Lambda_n(L)$.*
2) *the expected cost function $\mathbf{D} \mapsto \mathbb{E}f_{\mathbf{x}}(\mathbf{D})$ is Lipschitz with constant L as soon as $L > L_{\mathbb{P}}(\bar{g})$ and:*
   C2' *there exists a sequence γ_n such that*

$$\lim_{n\to\infty}\gamma_n = 0; \qquad \lim_{n\to\infty}\Gamma_n(\gamma_n) = 0$$

Proof: The first result trivially follows from Section V and the definition of $\Lambda_n(L)$. For the second one, given $\mathbf{D},\mathbf{D}'$, consider an i.i.d. draw $\mathbf{X}$ of n samples from $\mathbb{P}$. We have

$$\begin{aligned}|\mathbb{E}f_{\mathbf{x}}(\mathbf{D}') - \mathbb{E}f_{\mathbf{x}}(\mathbf{D})| &\leq |\mathbb{E}f_{\mathbf{x}}(\mathbf{D}') - F_{\mathbf{X}}(\mathbf{D}')| \\ &\quad + |F_{\mathbf{X}}(\mathbf{D}') - F_{\mathbf{X}}(\mathbf{D})| \\ &\quad + |F_{\mathbf{X}}(\mathbf{D}) - \mathbb{E}f_{\mathbf{x}}(\mathbf{D})| \\ &\leq L\|\mathbf{D}'-\mathbf{D}\|_{1\to2} + 2\gamma_n\end{aligned}$$

except with probability at most $\Lambda_n(L) + 2\Gamma_n(\gamma_n)$. The limit for large n yields the desired bound with probability one. ∎

Remark 3. Under **B1-B3** only (without the convexity assumption on $\mathfrak{D}$), an analogon of Lemma 8 holds where the conclusion that $\mathbf{D} \mapsto F_{\mathbf{X}}(\mathbf{D})$ (resp. its expectation) is "Lipschitz with constant L" is replaced with:

$$\frac{|F_{\mathbf{X}}(\mathbf{D}') - F_{\mathbf{X}}(\mathbf{D})|}{\|\mathbf{D}-\mathbf{D}'\|_{1\to2}} \leq L\cdot\left(1+\sqrt{\tfrac{1}{\kappa}}\|\mathbf{D}-\mathbf{D}'\|_{1\to2}\right). \quad (40)$$

(resp. (40) holds with $F_{\mathbf{X}}(\cdot)$ replaced by $\mathbb{E}f_{\mathbf{x}}(\cdot)$).

We are now ready to state a first uniform convergence result.

B. Abstract uniform convergence result

Lemma 9. *Assume* **C1-C2**. *For any $\epsilon,\gamma > 0$ we have, except with probability at most*

$$\Lambda_n(L) + \mathcal{N}(\mathfrak{D},\epsilon)\cdot\Gamma_n(\gamma) \quad (41)$$

- *under* **A1-A4** *or* **B1-B4**:

$$\sup_{\mathbf{D}\in\mathfrak{D}}|F_{\mathbf{X}}(\mathbf{D}) - \mathbb{E}f_{\mathbf{x}}(\mathbf{D})| \leq 2L\epsilon + \gamma \quad (42)$$

- *under* **B1-B3** *only*:

$$\sup_{\mathbf{D}\in\mathfrak{D}} |F_{\mathbf{X}}(\mathbf{D}) - \mathbb{E}f_{\mathbf{x}}(\mathbf{D})| \le 2L\epsilon\cdot\left(1+\sqrt{\tfrac{1}{\kappa}\epsilon}\right)+\gamma \quad (43)$$

Proof: We only give the proof under **A1-A4** or **B1-B4**. A straightforward adaptation yields the proof under **B1-B3** only.

Fix $\epsilon > 0$ and consider an ϵ-cover of $\mathfrak{D}$ with respect to the $\|\cdot\|_{1\to 2}$ metric with at most $\mathcal{N} = \mathcal{N}(\mathfrak{D},\epsilon)$ elements $\{\mathbf{D}_j\}$. Fixing $\mathbf{D}\in\mathfrak{D}$ there is an index j such that $\|\mathbf{D}_j - \mathbf{D}\|_{1\to 2} \le \epsilon$, hence

$$\begin{aligned}|F_{\mathbf{X}}(\mathbf{D}) - \mathbb{E}f_{\mathbf{x}}(\mathbf{D})| \le\ & |F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_j)| \\ & + |F_{\mathbf{X}}(\mathbf{D}_j) - \mathbb{E}f_{\mathbf{x}}(\mathbf{D}_j)| \\ & + |\mathbb{E}f_{\mathbf{x}}(\mathbf{D}_j) - \mathbb{E}f_{\mathbf{x}}(\mathbf{D})| \\ \le\ & |F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_j)| \\ & + \sup_{1\le j\le\mathcal{N}} |F_{\mathbf{X}}(\mathbf{D}_j) - \mathbb{E}f_{\mathbf{x}}(\mathbf{D}_j)| \\ & + L\epsilon\end{aligned}$$

where we used the Lipschitz property of the expected cost function (Lemma 8). By a union bound, for any $\gamma > 0$ we obtain the bound (42) except with probability at most (41). ■

C. Main uniform convergence result

The above sample complexity estimate is not quite explicit since it combines covering numbers and concentration bounds. In the following, we give more explicit sample complexity estimates under more specific assumptions on $\mathcal{N}(\mathfrak{D},\epsilon)$ and $\Gamma_n(\gamma)$. The assumption on $\mathcal{N}(\mathfrak{D},\epsilon)$ is justified by the fact that all examples of classes $\mathfrak{D}$ developed in Section VII have covering numbers with bounds of the type expressed in (3). The assumption **C2** on $\Gamma_n(\gamma)$ is further discussed at the end of this section.

Lemma 10. *Assume* **C1-C2** *holds and that $\mathcal{N}(\mathfrak{D},\epsilon)$ satisfy the bound* (45)*, i.e., for $c > 0$, $T\in(0,\infty]$, and $C, h \ge 1$:*

$$\Gamma_n(c\tau) \le 2\exp(-n\tau^2),\ \forall 0\le\tau\le T; \quad (44)$$

$$\mathcal{N}(\mathfrak{D},\epsilon) \le \left(\frac{C}{\epsilon}\right)^h,\ \forall 0<\epsilon\le 1. \quad (45)$$

Define

$$\beta \triangleq h\cdot\max(\log\tfrac{2LC}{c}, 1)$$

*and $D = 1$ (under **A1-A4** or **B1-B4**) or $D = \max(1/\kappa, 1)$ (under **B1-B3**). Assume that the sample size n satisfies*

$$\frac{n}{\log n} \ge \max\left(8, \tfrac{\beta}{T^2}, D\cdot\left(\tfrac{c}{2L}\right)^2\beta\right). \quad (46)$$

Then, for any

$$0 \le x \le nT^2 - \beta\log n \quad (47)$$

we have, except with probability at most $\Lambda_n(L) + 2e^{-x}$

$$\sup_{\mathbf{D}\in\mathfrak{D}} |F_{\mathbf{X}}(\mathbf{D}) - \mathbb{E}f_{\mathbf{x}}(\mathbf{D})| \le \eta_n(L,C,h,c).$$

In the case of **A1-A4** *or* **B1-B4***, we have*

$$\eta_n \triangleq 2c\sqrt{\tfrac{\beta\log n}{n}} + c\sqrt{\tfrac{\beta+x}{n}},$$

whereas for **B1-B3***, we obtain*

$$\eta_n \triangleq 3c\sqrt{\tfrac{\beta\log n}{n}} + c\sqrt{\tfrac{\beta+x}{n}}.$$

Note that Theorem 1 follows from Lemma 10.

Proof: First we observe that the condition **C2** on Γ_n implies **C2'** (defined in Lemma 8), and since we assume **C1** we can apply Lemma 9. Notice that (46) implies that one can indeed find x that satisfies assumption (47). We set

$$\epsilon = \tfrac{c\sqrt{\beta}}{2L}\sqrt{\tfrac{\log n}{n}},$$

$$\tau = \sqrt{\tfrac{h[\log C/\epsilon]+x}{n}} = \sqrt{\tfrac{h[\log(2LC/c\sqrt{\beta})]+\frac{h}{2}\log\frac{n}{\log n}+x}{n}}.$$

Since (46) implies $n \ge 8$, hence $\frac{\log n}{2} \ge 1$, we have

$$0 \le x \le nT^2 - \tfrac{\beta}{2}\log n - \tfrac{\beta}{2}\log n \le nT^2 - \tfrac{h}{2}\log n - \beta.$$

By definition of β and h we have $\beta \ge h \ge 1$ hence

$$\begin{aligned}n\tau^2 - \tfrac{h}{2}\log\tfrac{n}{\log n} &= h\log\tfrac{2LC}{c\sqrt{\beta}} + x \le h\log\tfrac{2LC}{c} + x \\ &\le \beta + x \le nT^2 - \tfrac{h}{2}\log\tfrac{n}{\log n}.\end{aligned}$$

This shows that $0 \le \tau \le T$. Moreover, by (46) we further have $0 < \epsilon \le 1$, hence we can apply (44) and (45) to obtain

$$\mathcal{N}(\mathfrak{D},\epsilon)\cdot\Gamma_n(c\tau) \le 2\cdot(C/\epsilon)^h\cdot\exp(-n\tau^2) = 2\cdot e^{-x}.$$

Using **A1-A4** or **B1-B4** we conclude as follows: since $\beta \ge 1$ and $\log n \ge 1$ we have

$$\begin{aligned}2L\epsilon + c\tau &= c\sqrt{\tfrac{\beta\log n}{n}} \\ &\quad + c\cdot\sqrt{h\log\tfrac{2LC}{c\sqrt{\beta}} + \tfrac{h}{2}\log\tfrac{n}{\log n} + x}\cdot\tfrac{1}{\sqrt{n}} \\ &\le c\sqrt{\tfrac{\beta\log n}{n}} + c\cdot\sqrt{\tfrac{\beta}{2}\log n + \beta + x}\cdot\tfrac{1}{\sqrt{n}} \\ &\le c\sqrt{\tfrac{\beta\log n}{n}}\cdot(1+\tfrac{1}{\sqrt{2}}) + c\cdot\sqrt{\beta+x}\cdot\sqrt{\tfrac{1}{n}} \\ &\le 2c\sqrt{\tfrac{\beta\log n}{n}} + c\sqrt{\tfrac{\beta+x}{n}}.\end{aligned}$$

Under **B1-B3**, the definition of D and assumption (46) imply $0 < \sqrt{\frac{1}{\kappa}}\epsilon \le 1$ hence we get similarly:

$$2L\epsilon\cdot\left(1+\sqrt{\tfrac{1}{\kappa}\epsilon}\right) + c\tau \le 3c\sqrt{\tfrac{\beta\log n}{n}} + c\sqrt{\tfrac{\beta+x}{n}}.$$

■

D. On assumptions *C1* and *C2*

Assumptions **C1-C2** are actually satisfied under rather standard hypotheses:

Lemma 11. *Assume that $\mathbb{P}\in\mathfrak{P}_{R\mathfrak{S}^{m-1}}$. Then*

$$\Gamma_n(R^2\tau/\sqrt{8}) \le 2\exp(-n\tau^2), \qquad \forall n, \forall\tau\ge 0 \quad (48)$$

$$\Lambda_n(R\ \bar{g}(R^2/2)) = 0, \qquad \forall n.$$

In other words, **C2** *holds with $c = R^2/\sqrt{8}$ and $T = +\infty$, and* **C1** *holds with $L_{\mathbb{P}}(\bar{g}) \le R\bar{g}(R^2/2)$).*

Proof: For any $\mathbf{D}$, the random variables $y_i = f_{\mathbf{x}_i}(\mathbf{D})$ satisfy $0 \le y_i \le \frac{1}{2}\|\mathbf{x}\|_2^2 \le \frac{R^2}{2}$ almost surely. Applying Hoeffding's inequality yields (48). We conclude by observing that

since $\|\mathbf{x}_i\|_2 \leq R$ almost surely, we have $L_{\mathbf{X}}(\bar{g}) \leq R\,\bar{g}(R^2/2)$ almost surely. ∎

Lemma 12. *Assume there is a constant $A > 0$ such that $\mathbb{P} \in \mathfrak{P}_A$ and that the penalty function g satisfies* **A1** *(non-negativity) and* **A4** *($g(\mathbf{0}) = 0$). Then*

$$\Gamma_n(12A\tau) \leq 2\exp(-n\tau^2),\ \forall 0 \leq \tau \leq 1.$$

In other words, **C2** *holds with $c = 12A$ and $T = 1$.*

Moreover, in all examples of penalties g considered in Table I, the growth of $t\bar{g}(t^2/2)$ for large t is at most polynomial, so that $L_{\mathbb{P}}(\bar{g}) < +\infty$, and **C1** *holds.*

Proof: We follow the argument in [26]. This primarily results from Bernstein's inequality, see, e.g., [38].

Lemma 13 (Bernstein's Inequality)**.** *Let $\{z_j\}_{j\in[\![1;n]\!]}$ be independent, zero-mean random variables. If there exist $v, M > 0$ such that for any integer $q \geq 2$ and any $j \in [\![1;n]\!]$, it holds*

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

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

$$\mathbb{P}\Big(\tfrac{1}{n}\sum_{j=1}^{n} z_j > \gamma\Big) \leq \exp\Big(-\frac{n\gamma^2}{2(v^2 + M\gamma)}\Big).$$

In particular, for any $\tau \leq \frac{v}{2M}$, we have

$$\mathbb{P}\Big(\tfrac{1}{n}\sum_{j=1}^{n} z_j > 2v\tau\Big) \leq \exp\big(-n\tau^2\big).$$

We can now state a simplification of [26, Lemma 24].

Lemma 14. *Let y be a random variable and assume there is a constant $B > 0$ such that, for any $t \geq 1$,*

$$\mathbb{P}\left(|y| > Bt\right) \leq \exp(-t). \tag{49}$$

Then, for any $u \geq 1$, any integer $q \geq 1$, and $0 < p \leq 1$, we have

$$\mathbb{E}\Big[|y|^{pq}\Big] \leq q!\Big[B^p u\Big]^q [1 + \exp(3-u)] \tag{50}$$

$$\mathbb{E}\Big[\Big||y|^p - \mathbb{E}\big[|y|^p\big]\Big|^q\Big] \leq q!\Big[2B^p u\Big]^q [1 + \exp(3-u)]. \tag{51}$$

To keep the flow of the paper we postpone the proof to the appendix. We now have the tools to state a simplified version of [26, Corollary 6] which suits our needs.

Corollary 4. *Consider n independent draws $\{y_i\}_{i\in[\![1;n]\!]}$ satisfying the hypothesis* (49) *for some $B > 0$. Then, for any $0 < p \leq 1$ and $0 \leq \tau \leq 1$, we have*

$$\mathbb{E}\{|y_i|^p\} \leq 6B^p \tag{52}$$

$$\mathbb{P}\left(\left|\tfrac{1}{n}\sum_{i=1}^{n}\left(|y_i|^p - \mathbb{E}\,|y_i|^p\right)\right| \geq 24B^p \cdot \tau\right) \leq 2\exp(-n\tau^2) \tag{53}$$

Proof: We apply Lemma 14 with $u = 3$. For $q = 1$, Equation (50) yields (52). For $q \geq 2$, Equation (14) shows that we can apply Bernstein's inequality (Lemma 13) with $z_i = |y_i|^p - \mathbb{E}\,|y_i|^p$, $M = 2B^p u = 6B^p$ and $v = \sqrt{2}M\sqrt{1+e^{3-u}} = 2M = 12B^p$. This shows that for $0 \leq \tau \leq v/2M = 1$ we have (53). ∎

We are now ready to prove Lemma 12. First, we notice that for any $\mathbf{D}$, $0 \leq y_i = f_{\mathbf{x}}(\mathbf{D}) \leq \frac{1}{2}\|\mathbf{x}\|_2^2$, hence we have

$$\sup_{\mathbf{D}\in\mathfrak{D}} \mathbb{P}\left(f_{\mathbf{x}}(\mathbf{D}) \geq \tfrac{A}{2}t\right) \leq \exp(-t),\ \forall t \geq 1.$$

Applying Corollary 4 with $B = A/2$ and $p = 1$ yields $\Gamma_n(12A\tau) \leq 2\exp(-n\gamma^2)$ for $0 \leq \tau \leq 1$. ∎

VII. Constraint set structures and covering numbers

Dictionaries learned with standard methods are unstructured matrices $\mathbf{D} \in \mathbb{R}^{m\times d}$ that allow factored representations of the signals of interest. However, in practice the dimension of the signals which are being represented and consequently the possible dictionaries' dimensions are inherently restricted by limited memory and limited computational resources. Furthermore, when used within signal reconstruction algorithms where many matrix vector multiplications have to be performed, those dictionaries are computationally expensive to apply. In the following we first describe various classes of dictionaries and their motivation in signal processing. Then we turn to the estimation of their covering numbers.

A. Unstructured Dictionaries with Unit Norm Atoms

Imposing unit norm on the atoms is a common technique to avoid trivial solutions for the learning problem. An exemplary learning algorithm that employs this kind of restriction is the famous K-SVD algorithm proposed in [8]. All following dictionary structures base upon this restriction. The class of dictionaries of this form is a product of spheres defined as

$$\mathfrak{D}(m,d) \triangleq \{\mathbf{D} \in \mathbb{R}^{m\times d} \,:\, \mathrm{ddiag}(\mathbf{D}^\top\mathbf{D}) = \mathbf{I}_d\}.$$

The operator ddiag maps matrices $\mathbb{R}^{d\times d}$ to $\mathbb{R}^{d\times d}$ by leaving the diagonal entries unaltered while setting all others to zero.

In NMF, this dictionary structure is further limited by requiring non-negative entries, yielding the class $\mathfrak{D}_{\text{NMF}} \cap \mathfrak{D}(m,d)$, or $\mathfrak{D}_{\text{NMF}\&\ell^1}$ if the ℓ^1 normalization is preferred.

Lemma 15 (Covering number bound for $\mathfrak{D}(m,d)$)**.** *For the Euclidean metric, $\mathfrak{S}^{m-1}$ the unit sphere in $\mathbb{R}^m$ and for any $\epsilon > 0$, we have*

$$\mathcal{N}(\mathfrak{S}^{m-1}, \epsilon) \leq \big(1 + 2/\epsilon\big)^m.$$

Moreover, for the metric induced on $\mathbb{R}^{m\times d}$ by $\|\cdot\|_{1\to 2}$, and for any $\epsilon > 0$, we have

$$\mathcal{N}(\mathfrak{D}(m,d), \epsilon) \leq \big(1 + 2/\epsilon\big)^{md}.$$

Proof: Lemma 2 in [36] gives the first conclusion for the sphere in $\mathbb{R}^m$. As for the second result, remember that the set $\mathfrak{D}(m,d)$ is a Cartesian product of d $(m-1)$-spheres. ∎

B. Sparse Dictionaries

The authors of [13] propose a technique to learn so-called sparse dictionaries. A sparse dictionary is the representation of a learned dictionary $\mathbf{D}$ and a given base dictionary $\mathbf{\Theta}$ (e.g., DCT, ODCT, Wavelet) with the sparse matrix $\bar{\mathbf{D}}$ via $\mathbf{D} = \mathbf{\Theta}\bar{\mathbf{D}}$. This learning algorithm offers a reduction in training complexity while aiming at maintaining the data modeling abilities of unconstrained dictionary learning. We show in the appendix that:

$$\mathcal{N}(\mathfrak{D}_{\text{sparse}}(m,d,s,\mathbf{\Theta}),\epsilon) \leq \left(\tbinom{m}{s}(1+2/\epsilon)^s\right)^d.$$

C. Orthogonal Dictionaries

Orthogonal dictionaries are square matrices belonging to

$$\mathfrak{D}_{\text{orth}}(d) \triangleq \{\mathbf{D} \in \mathbb{R}^{d\times d} : \mathbf{D}^\top\mathbf{D} = \mathbf{I}_d, \det(\mathbf{D}) = 1\}. \quad (54)$$

They represent the simplest form of dictionaries that allow a unique representation of the signal. Computing the inner product of the signal and the atoms of the dictionary provides the representation coefficients. We show in the appendix that:

$$\mathcal{N}(\mathfrak{D}_{\text{orth}}(d),\epsilon) \leq (3\pi e^\pi/\epsilon)^{d(d-1)/2}.$$

D. Stiefel manifold

In PCA (see Section VIII), the dictionary is a matrix consisting of d orthonormal columns where $d \leq m$, i.e., $\mathbf{D}$ is a tall matrix. This constraint set is the so called Stiefel manifold

$$\mathfrak{D}_{\text{St}}(m,d) \triangleq \{\mathbf{D} \in \mathbb{R}^{m\times d} : \mathbf{D}^\top\mathbf{D} = \mathbf{I}_d\}. \quad (55)$$

We show in the appendix that:

$$\mathcal{N}(\mathfrak{D}_{\text{St}}(d,m),\epsilon) \leq (3\pi e^\pi/\epsilon)^{md-d(d+1)/2}.$$

E. Separable Dictionaries

In [12] a dictionary learning scheme is developed that learns a separable dictionary for sparse image representations: instead of sparsely representing $\mathbf{x} = \text{vec}(\mathbf{Y})$ as $\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}$, where $\mathbf{Y}$ is a matrix containing an image, one directly represents $\mathbf{Y}$ sparsely as $\mathbf{B}$ via two dictionaries $\mathbf{D}_1$, $\mathbf{D}_2$ by the equation

$$\mathbf{Y} = \mathbf{D}_1\mathbf{B}\mathbf{D}_2^\top. \quad (56)$$

Note that in order to relate this method to standard dictionary learning techniques Equation (56) can be rewritten as

$$\text{vec}(\mathbf{Y}) = (\mathbf{D}_2 \otimes \mathbf{D}_1)\cdot\text{vec}(\mathbf{B}).$$

We show in the appendix that:

$$\mathcal{N}(\mathfrak{D}_{\text{sep}},\epsilon) \leq (3/\epsilon)^{\sum_i m_i d_i}.$$

F. Other Tensor Dictionaries

Similar to the structure of Separable dictionaries we can define other constraint sets that are combinations of dictionary structures. Examples for this are tensor products of Stiefel matrices defined as

$$\mathfrak{D}_{\otimes\text{St}} \triangleq \mathfrak{D}_{\text{St}}(m_1,d_1)\otimes\ldots\otimes\mathfrak{D}_{\text{St}}(m_z,d_z) \quad (57)$$

which are used for Higher-order SVD. For the sake of readability the dimensions are omitted in this definition.

We show in the appendix that:

$$\mathcal{N}(\mathfrak{D}_{\otimes\text{St}},\epsilon) \leq (3\pi e^\pi/\epsilon)^{\sum_i m_i d_i - d_i(d_i+1)/2}.$$

VIII. Worked Examples

In this section we propose some settings that fit to our framework and provide the corresponding sample complexity estimates, i.e., the value of β.

Example 1 (PCA). PCA can be expressed as the optimization problem

$$\min_{\mathbf{A};\mathbf{D}\in\mathfrak{D}_{\text{St}}(m,d)} \tfrac{1}{2}\|\mathbf{X}-\mathbf{D}\mathbf{A}\|_F^2 + \tfrac{1}{n}\sum g(\boldsymbol{\alpha}_i)$$

with the penalty function $g = \chi_{d\text{-sparse}}$. Since $\boldsymbol{\alpha}_i \in \mathbb{R}^d$, the penalty function is equivalent to 0. Its sole purpose is to shape the problem in a way that fits into the general framework (2). The constraint set is $\mathfrak{D}_{\text{St}}(m,d)$, the set of $m\times d$-matrices with orthonormal columns.

a) Training data in the unit ball: Initially, we consider a distribution $\mathbb{P}$ in the unit ball ($R = 1$, therefore, $c = 1/\sqrt{8}$, cf. Lemma 11). Since $\mathfrak{D}_{\text{St}}(m,d) \subset \mathfrak{D}_{\text{LRIP}}(k,\delta)$ with $k = d$ and $\delta = 0$, the Lipschitz constant is $L = 2\sqrt{d}$, cf. Table I. Table II yields $h = md - d(d+1)/2$ and $C = 3\pi e^\pi$. Hence, the constant driving the sample complexity is

$$\beta_{\text{PCA}} = \left(md - \tfrac{d(d+1)}{2}\right)\cdot\log(12\pi e^\pi\sqrt{8d}).$$

Compared to existing sample complexity results for PCA [39], [40], this seems rather pessimistic: in [40] the sample complexity is driven by $\beta \propto d$, and we lose a large dimension-dependent m factor[5] as well as an additional mild $\log n$ factor. In fact, the assumption that the data is distributed in the unit ball is crucial in the results of [39], [40] as it allows the use of the McDiarmid concentration inequality, while, in our general context, we exploit a standard but somewhat crude argument using union bounds and covering numbers. This however means we can handle data distributions not restricted to live on the unit ball, as we will see next.

b) Sub-Gaussian training data: The sub-optimality of our results for PCA with respect to state of the art is probably the price to pay for the generality of the approach. For example, in contrast to previous work, our approach provides results for other distributions. For example, given the distribution $\mathbb{P} \in \mathfrak{P}_A$, we obtain the constant $c = 12A$ (Lemma 12), while all other constants remain the same. Thus, we get

$$\beta_{\text{PCA}} = \left(md - \tfrac{d(d+1)}{2}\right)\cdot\max\left(\log\left(\tfrac{\pi e^\pi\sqrt{d}}{A}\right),1\right), \quad (58)$$

Example 2 (Sparse coding, training data in the unit ball). Consider $\mathbb{P}$ a distribution in the unit ball ($R = 1$). We have $c = 1/\sqrt{8}$ (Lemma 11), and Table I immediately provides values of the Lipschitz constant $L = \bar{g}(1/2)$ such that $\Lambda_n(L) = 0$. For example, for $0 < p < \infty$

- for $g_{\chi_p}(\boldsymbol{\alpha}) = \chi_{\|\boldsymbol{\alpha}\|_p\leq\lambda}$, we have $L_p = d^{(1-1/p)_+}\lambda$;
- for $g_{\chi_0}(\boldsymbol{\alpha}) = \chi_{\|\boldsymbol{\alpha}\|_0\leq k}$, we have $L_0 = 2\sqrt{\frac{k}{1-\delta}}$

For $\mathfrak{D} = \mathfrak{D}(m,d)$ (resp. $\mathfrak{D} = \mathfrak{D}_{\text{LRIP}}(k,\delta)$), Table II provides $h = md$ and $C = 3$, and Theorem 1 yields

$$\beta_{\chi_p} = md\cdot\max\left(\log 6\sqrt{8}\lambda d^{(1-1/p)_+},1\right),$$
$$\beta_{\chi_0} = md\cdot\log 12\sqrt{8k/(1-\delta)}.$$

[5] See Section IX for techniques to handle this high-dimensional scaling.

The cases g_{χ_1} and g_{χ_0} are dealt with by Vainsencher *et al.* [20, Theorem 7] who provide a comparable result, however restricted to distributions on the unit sphere whereas our framework allows for more complex distributions. Moreover while they make assumptions on the cumulative coherence of $\mathbf{D}$ for their results with the ℓ^0 constraint, we only rely on the restricted isometry property.

The bound of Maurer and Pontil [19] for g_{χ_p}, $p \geq 1$, $\lambda = 1$, is of the order $\beta'_p \propto d^{4-2/p}$. We extend it to $p < 1$ as well, and improve the case $p = 1$ as soon as $m \leq d$.

In addition to the above penalty functions associated to the indicator functions of the ℓ^1 (resp. ℓ^0) ball, our results allow dealing with the ℓ^1 norm itself –a very popular proxy for the ℓ^0 "norm"– or its powers as penalty functions. For $\mathfrak{D} = \mathfrak{D}(m,d)$ (resp. $\mathfrak{D} = \mathfrak{D}_{\mathrm{LRIP}}(k,\delta)$), the sparsity measures

- $g_{\ell^1}(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_1/\lambda$, yielding $L = \frac{\lambda}{2}$;
- $g_{(\ell^1)^2}(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_1^2/\lambda^2$, yielding $L = \frac{\lambda}{\sqrt{2}}$;

are associated to the sample complexities

$$\beta_{\ell^1} = md \cdot \max\left(\log 3\sqrt{8}\lambda, 1\right)$$
$$\beta_{(\ell^1)^2} = md \cdot \max\left(\log 12\lambda, 1\right)$$

for probability distributions in the unit ball.

Example 3 (Sparse coding, sub-Gaussian training data). For a distribution $\mathbb{P}$ satisfying assumption (10) we have $c = 12A$. For $\mathfrak{D} = \mathfrak{D}(m,d)$ we again have the constants $C = 3$, $h = md$ and Theorem 1 then yields sample complexity for $g_{(\ell^1)^2}$

$$\beta = md \cdot \max\left(\log \tfrac{\lambda}{2\sqrt{2}A}, 1\right).$$

Example 4 ("Doubly sparse" coding, training data in the unit ball). Our results also allow to deal with other classes of structured dictionaries, such as the sparse dictionary $\mathfrak{D}_{\mathrm{sparse}}(m,d)$ defined in Equation (65), yielding $h = sd$ and $C = 3\binom{m}{s}^{1/d}$. With g_{χ_1} (respectively g_{χ_0} with a restriction to $\mathfrak{D}_{\mathrm{sparse}}(m,d) \cap \mathfrak{D}_{\mathrm{LRIP}}(k,\delta)$), using Stirling's approximation, Theorem 1 yields

$$\beta^s_{\chi_1} \leq sd \cdot \max\left(\log 6\sqrt{8}\lambda + \log \tfrac{me}{s}, 1\right)$$
$$\beta^s_{\chi_0} \leq sd \cdot \left(\log 12\sqrt{8k/(1-\delta)} + \log \tfrac{me}{s}\right).$$

Example 5 (Non-negative Matrix Factorization, training data in the unit ball). In Non-negative Matrix Factorization *(NMF)* (cf. [14]) a matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ is decomposed into a product of two matrices with non-negative entries. It can be expressed as the optimization problem

$$\min_{\mathbf{A};\mathbf{D} \in \mathfrak{D}_{\mathrm{NMF}\&\ell^1}} \tfrac{1}{2}\|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 + \sum_{i=1}^{n} \chi_{\mathrm{NMF}}(\boldsymbol{\alpha}_i). \tag{59}$$

Since the constraint set $\mathfrak{D}_{\mathrm{NMF}\&\ell^1}$ is a subset of $\mathfrak{D}_{\mathrm{NMF}}(1/md)$, the penalty $g = \chi_{\mathrm{NMF}}$ and the set $\mathfrak{D}_{\mathrm{NMF}\&\ell^1}$ satisfy **B1-B3** with $\kappa = 1/md$, and Table I gives the Lipschitz constant $L = 2\sqrt{md}$. Moreover, since $\mathfrak{D}_{\mathrm{NMF}\&\ell^1} \subset \mathfrak{D}_{\mathrm{ball}}(m,d)$, where

$$\mathfrak{D}_{\mathrm{ball}}(m,d) \triangleq \{\mathbf{D} \in \mathbb{R}^{m \times d} \,:\, \|\mathbf{d}_i\|_2 \leq 1\}$$

has the same covering number as $\mathfrak{D}(m,d)$, we get $\mathcal{N}(\mathfrak{D},\epsilon) \leq \left(\frac{3}{\epsilon}\right)^{md}$ as a direct consequence of the covering number bound for the product of spheres. This yields the constants $C = 3$ and $h = md$, and together with the constant $c = 1/\sqrt{8}$, which we get for distributions on the unit sphere, we obtain

$$\beta_{\mathrm{NMF}} = md \cdot \log 12\sqrt{8md}. \tag{60}$$

In a similar setting Maurer and Pontil [19] obtain a sample complexity estimate $\beta \propto d^3$ which has the advantage of being independent of the signal dimension m. Yet, our result seems sharper unless the signal dimension m exceeds d^2, in which case the approach sketched in Section IX could be followed to recover improved estimates in this high-dimensional scaling.

Example 6 (K-means clustering, training data in the unit ball). Given a set of n data points K-means clustering algorithms learn a dictionary with K columns and assign each data point to one of these columns while minimizing the sum of squared distances. These atoms represent the K distinct classes. In our notation so far $d = K$. Then the problem of K-means clustering can be expressed as the minimization problem

$$\min_{\mathbf{A};\mathbf{D} \in \mathbb{R}^{m \times K}} \tfrac{1}{2}\|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 + \sum_{i=1}^{n} g(\boldsymbol{\alpha}_i) \tag{61}$$

with the penalty function

$$g(\boldsymbol{\alpha}) = \chi_{\|\boldsymbol{\alpha}\|_0 = 1} + \chi_{\sum_k \alpha_k = 1}. \tag{62}$$

This penalty function ensures that all columns of $\mathbf{A}$ have exactly one entry with the value 1. The centers of the clusters are represented in the columns of $\mathbf{D}$.

This penalty function fulfills **A1-A3**, and its auxiliary function is $\bar{g}(t) \equiv 1$ for all $t \in \mathbb{R}$. However, g does not fulfill **A4**. This has the consequence that we cannot merely apply Corollary 3. Instead, we have to rely on the preceding Lemma 5 which leaves us with the upper bound $\frac{1}{n}\sum_{i=1}^n \sqrt{2f_{\mathbf{x}_i}(\mathbf{D})} \cdot \bar{g}(f_{\mathbf{x}_i}(\mathbf{D})) = \frac{1}{n}\sum_{i=1}^n \sqrt{2f_{\mathbf{x}_i}(\mathbf{D})}$ for the Lipschitz constant. To proceed, we recall the definition $f_{\mathbf{x}}(\mathbf{D}) = \inf_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + g(\boldsymbol{\alpha})$. For the function $\frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + g(\boldsymbol{\alpha})$ to be finite, the coefficient vector $\boldsymbol{\alpha}$ has to fulfill $g(\boldsymbol{\alpha}) = 0$. Due to the construction of g, there is only a finite number of possible choices for $\boldsymbol{\alpha}$ (namely $\boldsymbol{\alpha}$ is a standard basis vector in $\mathbb{R}^d$), and the Lipschitz constant can be upper bounded by

$$L_{\mathbf{X}}(\mathbf{D}) \leq \tfrac{1}{n}\sum_{i=1}^{n} \max_{j=1,\ldots,K} \|\mathbf{x}_i - \mathbf{d}_j\|_2 \leq \max_{i,j} \|\mathbf{x}_i - \mathbf{d}_j\|_2.$$

Restricting our analysis to the case of data samples $\mathbf{x}_i$ lying within the unit ball ($\|\mathbf{x}_i\|_2 \leq 1$), the optimum dictionary coincides with the optimum where the matrix $\mathbf{D}$ is constrained to belong to $\mathfrak{D}_{\mathrm{ball}}(m,K)$. It follows that the Lipschitz constant is simply bounded by $L = 2$. Moreover, $\mathfrak{D}_{\mathrm{ball}}(m,K)$ has the same covering number as $\mathfrak{D}(m,K)$. From the structural constraints to the dictionary and the signal distribution we obtain the constants $L = 2$, $C = 3$, $h = mK$, $c = 1/\sqrt{8}$. The above implies the constant

$$\beta_{K-\mathrm{means}} = mK \cdot \log(12\sqrt{8}).$$

The literature on the sample complexity of K-means is abundant. Following the early work of Pollard [41], a series of authors have established worst-case lower bounds on the

excess risk decaying at best as $1/\sqrt{n}$ [42]–[44], as well as upper bounds of the order of $O(1/\sqrt{n})$ (for a given dimension m) [45]. More concrete estimates, where the role of K and m is explicated, are needed to compare with our sample complexity estimate. Bartlett *et al.* [43] bound the excess risk by a constant times $\min\left(\sqrt{Km/n}, \sqrt{K^{1-2/m}m\log n/n}\right)$ which, for large m and arbitrary K, is essentially matched by our estimate. Estimates independent of the dimension m have been obtained by Biau *et al.* [46] (see also [19]) of the order $K/\sqrt{n}$. These correspond to $\beta \propto K^2$, hence our result seems sharper when $m \lesssim K$, i.e., when the number of clusters exceeds the dimension.

Example 7 (Higher-order SVD). A natural extension of PCA is the so called Higher-order SVD *(HOSVD)* or multilinear SVD introduced in [25]. Let the training signals $\mathbf{x}_i$ be real z-tensors, i.e., $\mathbf{X}$ is an element of $\mathbb{R}^{m_1\times m_2\times\ldots\times m_z\times n}$ with $\mathbf{X} = [\mathbf{X}_1\ \ldots\ \mathbf{X}_n]$ where $\mathbf{X}_i \in \mathbb{R}^{m_1\times m_2\times\ldots\times m_z}$. The HOSVD $\mathbf{X}$ is obtained by solving the minimization problem

$$\min_{\mathbf{A},\mathbf{D}\in\mathfrak{D}_{\otimes\mathrm{St}}} \tfrac{1}{2}\sum_{i=1}^{n}\|\mathbf{X}_i - \mathbf{D}\circ\mathbf{A}_i\|_F^2.$$

In this equation $\mathbf{D}$ is an element of the constraint set $\mathfrak{D}_{\otimes\mathrm{St}}$ as defined in (57) and therefore a concatenation of Stiefel matrices $\mathbf{D}_1,\ldots,\mathbf{D}_z$ with $\mathbf{D}_i \in \mathfrak{D}_{\mathrm{St}}(m_i,d_i)$, and the coefficient tensor $\mathbf{A}\in\mathbb{R}^{d_1\times\ldots d_z\times n}$ has the form $\mathbf{A} = [\mathbf{A}_1\ \ldots\ \mathbf{A}_n]$ with $\mathbf{A}_i \in \mathbb{R}^{d_1\times\ldots d_z}$. The operator $\circ$ denotes the operation

$$\mathbf{D}\circ\mathbf{A}_i \triangleq \mathbf{A}_i \times_1 \mathbf{D}_1 \times_2 \mathbf{D}_2 \ldots \times_z \mathbf{D}_z,$$

where $\times_N$ denotes the so-called N-mode product introduced in [25]. Just as for PCA the trivial penalty $G(\mathbf{A})\equiv 0$ can be written using $g(\boldsymbol{\alpha}) = \chi_{\|\boldsymbol{\alpha}\|_0\leq\prod_{i=1}^{z} d_i}$ where $\boldsymbol{\alpha}_i$ is the vector resulting from the unfolding of the i-th component of $\mathbf{A}$, $1\leq i\leq n$, along its rightmost dimension.

Table II yields the constants $C = 3\pi e^{\pi}$ and $h = \sum_{i=1}^{z} m_i d_i - d_i(d_i+1)/2$. To obtain the Lipschitz constant L we unfold the tensor $\mathbf{A}$ to the matrix $\mathbb{R}^{(\prod_{i=1}^{z} d_i)\times n}$ and then use a penalty definition similar to the one used for PCA. This yields the constant $L = 2\cdot\sqrt{\prod_{i=1}^{z} d_i}$. For distributions $\mathbb{P}$ in the ball with $R=1$, Theorem 1 provides

$$\beta_{\mathrm{HOSVD}} = \left(\sum_{i=1}^{z} m_i d_i - \tfrac{d_i(d_i+1)}{2}\right)\cdot\log\left(12\pi e^{\pi}\sqrt{8\prod_{i=1}^{z} d_i}\right).$$

In light of state of the art results [40] on PCA, this apparently new sample complexity estimate is probably quite pessimistic.

IX. High-dimensional settings

When m gets larger, the reasoning in Section VI based on covering numbers of the set of admissible dictionaries $\mathfrak{D}$ leads to unfavorable scaling with respect to m. This factor is introduced through the ϵ-net argument following Lemma 9. In order to avoid the issues that arise in the high-dimensional setting, we could envision using an alternative to Lemma 3 which served as a precursor to the discussion of the Lipschitz property. Such an alternative is proposed in Lemma 7, where we derive a Lipschitz property of $F_{\mathbf{X}}$ w.r.t. the Frobenius norms of $\mathbf{D}^\top\mathbf{D}$ and $\mathbf{D}^\top\mathbf{X}$ which are in $\mathbb{R}^{d\times d}$ and $\mathbb{R}^{d\times n}$, respectively. Covering net arguments revolving around these quantities clearly become independent of the signal dimension m, but would yield difficulties due to their dependence on the draw of $\mathbf{X}$. Providing concrete results for this cases is beyond the scope of this paper, instead, we sketch below an alternative approach.

Another option to handle the high-dimensional setting is to extend the approach proposed in [19]. We provide a rough outline below. First, recall that the initial problem is to provide an upper bound for the expression

$$\sup_{\mathbf{D}\in\mathfrak{D}}\left(\tfrac{1}{n}\sum_{i=1}^{n} f_{\mathbf{x}_i}(\mathbf{D}) - \mathbb{E}_{\mathbf{x}} f_{\mathbf{x}}(\mathbf{D})\right).$$

We can use McDiarmid's inequality in combination with a symmetrization argument and Rademacher averages to provide the upper bound

$$\sup_{\mathbf{D}\in\mathfrak{D}}\left(\tfrac{1}{n}\sum_{i=1}^{n} f_{\mathbf{x}_i}(\mathbf{D}) - \mathbb{E}_{\mathbf{x}} f_{\mathbf{x}}(\mathbf{D})\right) = \mathrm{C}_1\cdot\left[\mathbb{E}_\eta \sup_{\mathbf{D}\in\mathfrak{D}} A_{\mathbf{D}}(\eta)\right] + \mathrm{C}_2\cdot\sqrt{2t/n}$$

with probability at least $1-e^{-t}$ and the Gaussian process $A_{\mathbf{D}}(\eta) = \frac{1}{n}\sum_{i=1}^{n}\eta_i f_{\mathbf{x}_i}(\mathbf{D})$ where the variables η_i are i.i.d. standard Gaussian. Therefore, it remains to find an upper bound for $\mathbb{E}_\eta \sup_{\mathbf{D}\in\mathfrak{D}} A_{\mathbf{D}}(\eta)$. Slepian's Lemma states that for any Gaussian process $B_{\mathbf{D}}$ that fulfills the condition $\mathbb{E}_\eta|A_{\mathbf{D}}(\eta) - A_{\mathbf{D}'}(\eta)|^2 \leq \mathbb{E}_\xi|B_{\mathbf{D}}(\xi) - B_{\mathbf{D}'}(\xi)|^2$ the bound $\mathbb{E}_\eta[\sup_{\mathbf{D}\in\mathfrak{D}} A_{\mathbf{D}}(\eta)] \leq \mathbb{E}_\xi[\sup_{\mathbf{D}\in\mathfrak{D}} B_{\mathbf{D}}(\xi)]$ holds. In the case at hand we can observe that $\mathbb{E}_\eta|A_{\mathbf{D}}(\eta) - A_{\mathbf{D}'}(\eta)|^2 = \frac{1}{n^2}\sum_{i=1}^{n}|f_{\mathbf{x}_i}(\mathbf{D}) - f_{\mathbf{x}_i}(\mathbf{D}')|^2$. Lemma 7 yields the upper bound

$$|f_{\mathbf{x}}(\mathbf{D}) - f_{\mathbf{x}}(\mathbf{D}')| \leq C_{\mathbf{X}}^{(1)}\|\mathbf{D}^\top\mathbf{D} - (\mathbf{D}')^\top\mathbf{D}'\|_F + C_{\mathbf{X}}^{(2)}\|(\mathbf{D}-\mathbf{D}')^\top\mathbf{x}\|_2, \qquad (63)$$

which implies that the Gaussian process

$$B_{\mathbf{D}}(\xi) = C'\langle\mathbf{D}^\top\mathbf{D}, \xi^{(1)}\rangle_F + C''\langle\mathbf{D}^\top\mathbf{x}, \xi^{(2)}\rangle_2 \qquad (64)$$

fulfills the condition of Slepian's Lemma for appropriate choices of the scalars C', C''. Here, the first scalar product operates in $\mathbb{R}^{d\times d}$, whereas the second one is defined on $\mathbb{R}^d$. Thus, computing the supremum over all $\mathbf{D}$ would result in an expression which is independent of the signal dimension m. Especially in the case of PCA, the final sample complexity result would benefit since the first term in Equation (63), and thereby the first term in (64), vanishes, yielding a tighter bound. However, it is not within the scope of this work to provide concrete results for this alternative approach.

X. Conclusion and discussion

We proposed a general framework to determine the sample complexity of dictionary learning and related matrix factorization problems. The generality of the framework makes it applicable for a variety of structure constraints, penalty functions, and signal distributions beyond previous work. In particular, it covers formulations such as principal component analysis, sparse dictionary learning, non-negative matrix

factorization, or K-means clustering, for which we provide sample complexity bounds in the worked examples section.

To keep the exposition of our results simple, we focused on matrix factorization where the data fidelity term is expressed with squared ℓ^2 norm. A straightforward adaptation of the computation of the Lipschitz constant of Section V can easily be made to achieve similar results with the plain ℓ^2 norm [20], and easy adaptations can be envisioned when the data fidelity term is $\ell(\mathbf{x}-\mathbf{D}\boldsymbol{\alpha})$ where the gradient $\nabla\ell(\cdot)$ is Lipschitz. More general loss functions are briefly discussed in Section V-D.

The obtained sample complexity results applied to sparse coding extend those of Maurer and Pontil [19] and Vainsencher *et al.* [20] in primarily two ways.

First, we relax the assumption that the training data lives in the unit ball [19] or even the unit Euclidean sphere [20] by showing that it is sufficient to have sufficient decay of the probability of drawing training samples with "large" norm. This is essentially achieved by replacing Hoeffding's inequality with a more refined Bernstein inequality argument.

Second, and more importantly, we handle penalty functions g beyond indicator functions of compact sets [19, Theorem 1], or of sets $\mathcal{K}$ such that $\sup_{\mathbf{D}\in\mathfrak{D},\boldsymbol{\alpha}\in\mathcal{K}}\|\mathbf{D}\boldsymbol{\alpha}\|_2 < \infty$ [19, Theorem 2], or of ℓ^1 or ℓ^0 balls [20]. Indeed, the first generic case dealt with in this paper involves penalty functions that only need to be non-negative, lower semi-continuous, coercive (and for convenience be minimum at zero). The second generic case also covers as particular cases the indicator function of k-sparse vectors and that of non-negative vectors, with restrictions on $\mathfrak{D}$.

Beyond sparse coding, our results provide apparently new sample complexity estimates for K-means. Compared to state of the art results [43] and [46] we pay an additional $\log n$ factor in our excess risk analysis, but improve the sample complexity estimate when the number of clusters K exceeds the ambient dimension m. A brief discussion on possible strategies of how to obtain results independent of m is provided in Section IX.

Similarly, our results provide new sample complexity estimates for NMF improving over state of the art results [19] unless the ambient dimension m exceeds the order d^2. Again, we refer to Section IX for a discussion of the high-dimensional setting.

Despite its successes, the main limitation of the proposed approach seems to be in the rather crude use of union bounds and covering numbers. One can envision sharper sample complexity estimates using more refined concentration tools. For example, sample complexity results for PCA for signal distributions in the unit ball were studied in [40]. Although our approach achieves results for more general distributions such as the class $\mathfrak{P}_A$, the sample complexity results we obtain for this setting seem disappointingly pessimistic, both paying an additional $\log n$ factor and overestimating the sample complexity as $md-d(d+1)/2$ rather than d. We expect this may be due to the use by the authors in [40] of more refined tools such as McDiarmid's bounded difference concentration inequality for the supremum of the deviation, and more importantly of a clever dimension-independent bound (i.e., independent of m) on the expectation of the supremum of the deviation. We dedicated Section IX to a brief discussion of how we would envision to apply these tools to the more general context considered here, to obtain dimension-independent results.

The reader may have noticed that the rate of convergence of our estimates is in $1/\sqrt{n}$, which is typical when using techniques based on empirical processes. In certain settings it is possible with substantially more work to achieve fast rates in $1/n$, see, e.g., [47] in a much simpler setting, or the discussions of fast rates in [19], [20], [40], [46]. It is not clear at this stage whether it is a realistic objective to achieve fast rates in the investigated general setting. In any case this is expected to require well chosen "margin conditions" to hold, see, e.g., [48], [49] for the case of K-means.

Improving the estimate of Lipschitz constants is another avenue for improvement, although its role in our sample complexity estimates is already only logarithmic. In fact, the generality of our approach comes at a price. The bound on $L_{\mathbf{X}}(\mathbf{D})$ in Lemma 5 (resp. Lemma 6) is a worst case estimate given the bounds (24) and (25) (resp. (36)), and therefore rather crude. In a probabilistic setting, where the training samples $\mathbf{x}_i$ are drawn i.i.d. according to some distribution $\mathbb{P}$, one can envision much sharper bounds for $\frac{1}{n}\|(\mathbf{X}-\mathbf{D}\mathbf{A})\mathbf{A}^\top\|_\star$ using matrix concentration inequalities. A key difficulty will come from the control of $\mathfrak{A}$ which is dependent on $\mathbf{X}$ and $\mathbf{D}$. Moreover, while we concentrated on the exploitation of Lemma 3 and its corollary for the special metric $\|\cdot\|_{1\to 2}$ on dictionaries, better choices may be possible, including metrics defined in terms of the penalty g when applicable (potentially leading to alternate definitions of $\bar{g}$), or metrics dependent on the data distribution $\mathbb{P}$.

The reader may object that the role of the Lipschitz constant L in our final results (see, e.g., Theorem 1) is only logarithmic so the added value of such technicalities might be limited. Here, we notice that for the sake of simplicity we expressed Corollary 3 with a uniform Lipschitz constant independent of $\mathbf{D}$. Clearly, some intermediate steps yield finer estimates $L_{\mathbf{X}}(\mathbf{D})$ that depend on the considered dictionary $\mathbf{D}$, i.e., *local* Lipschitz constants. In certain scenarii these may provide more precise estimates that may turn out to be useful especially in the analysis of local properties of $F_{\mathbf{X}}(\mathbf{D})$. This is best illustrated with an example.

Example 8. When $g(\boldsymbol{\alpha}) = \lambda\|\boldsymbol{\alpha}\|_1^2$, we have $\bar{g}(t) = \sqrt{t/\lambda}$ hence $L_{\mathbf{X}}(\mathbf{D}) = \frac{1}{n}\sum_{i=1}^n \sqrt{2f_{\mathbf{x}_i}(\mathbf{D})}\cdot\sqrt{f_{\mathbf{x}_i}(\mathbf{D})/\lambda} = F_{\mathbf{X}}(\mathbf{D})\sqrt{2/\lambda}$. As a result, Corollary 2 implies that we locally have $F_{\mathbf{X}}(\mathbf{D}')/F_{\mathbf{X}}(\mathbf{D}) \le 1 + \sqrt{2/\lambda}\|\mathbf{D}'-\mathbf{D}\|_{1\to 2}$. This implies that $\log F_{\mathbf{X}}(\mathbf{D})$ is uniformly Lipschitz with constant $L = \sqrt{2/\lambda}$. In other words, $F_{\mathbf{X}}(\mathbf{D})$ is more regular where it takes small values.

This is likely to have an impact when performing a local stability analysis of sparse coding with the penalty $\lambda\|\boldsymbol{\alpha}\|_1^2$ rather than $\lambda\|\boldsymbol{\alpha}\|_1$, in the spirit of [26], see also [50]–[52].

Finally, observing that this paper provides sample complexity estimates of matrix factorization with penalties $g(\boldsymbol{\alpha})$ much beyond indicator functions of (often compact) sets, it is natural to wonder if we could also replace the constraint $\mathbf{D}\in\mathfrak{D}$, i.e., the penalty $\chi_{\mathfrak{D}}(\mathbf{D})$, by a more general penalty on $\mathbf{D}$ promoting certain dictionaries. This would somehow lead to a model selection process where one could envision, e.g., adapting the

effective number of nonzero columns of $\mathbf{D}$, but would raise hard questions regarding the usability of covering numbers.

ACKNOWLEDGMENTS

R. Gribonval would like to thank Daniel Vainsencher and Karin Schnass for memorable discussions motivating this work and Gilles Blanchard for invaluable comments.

APPENDIX A
PROOF OF LEMMA 14

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

$$\left(\mathbb{E}\left\{||y|^p - \mathbb{E}\left\{|y|^p\right\}|^q\right\}\right)^{1/q} \leq \left(\mathbb{E}\left\{|y|^{pq}\right\}\right)^{1/q} + \left(\mathbb{E}\left\{\left(\mathbb{E}\left\{|y|^p\right\}\right)^q\right\}\right)^{1/q},$$

so that by using Jensen's inequality, we obtain

$$\mathbb{E}\left\{||y|^p - \mathbb{E}\left[|y|^p\right]|^q\right\} \leq 2^q \mathbb{E}\left[|y|^{pq}\right],$$

thus proving (51) provided that (50) holds. We now focus on these raw moments. Let us fix some $u \geq 1$. We introduce the event

$$\mathcal{K} \triangleq \{\omega \,:\, |y(\omega)|/B \leq u\},$$

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

$$\mathcal{K}^c \subseteq \cup_{l=l_u}^{\infty} \mathcal{K}_l^c, \quad \text{with } \mathcal{K}_l^c = \{\omega \,:\, |y(\omega)|/B \in [l, l+1)\}.$$

We have

$$\begin{aligned}
\mathbb{E}\{|y|^{pq}\} &= \mathbb{E}\{\mathbb{1}_{\mathcal{K}}|y|^{pq}\} + \mathbb{E}\{\mathbb{1}_{\mathcal{K}^c}|y|^{pq}\} \\
&\leq \left(Bu\right)^{pq} + \textstyle\sum_{l=l_u}^{\infty} \mathbb{E}\{\mathbb{1}_{\mathcal{K}_l^c}|y|^{pq}\} \\
&\leq B^{pq} \cdot \left[u^{pq} + \textstyle\sum_{l=l_u}^{\infty} (l+1)^{pq} \cdot \mathbb{E}\{\mathbb{1}_{\mathcal{K}_l^c}\}\right] \\
&\leq B^{pq} \cdot \left[u^{q} + \textstyle\sum_{l=l_u}^{\infty} (l+1)^{pq} \cdot \mathbb{E}[\mathbb{1}_{\{\omega;\ |y(\omega)| \geq Bl\}}]\right]
\end{aligned}$$

where in the last line we used $u^p \leq u$ since $u \geq 1$ and $p \leq 1$. Using the hypothesis (49), we continue

$$\mathbb{E}\{|y|^{pq}\} \leq B^{pq} \cdot \left[u^q + \textstyle\sum_{l=l_u}^{\infty} (l+1)^{pq} \exp(-l)\right].$$

Upper bounding the discrete sum by a continuous integral, we recognize here the incomplete Gamma function [53],

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

where again we used $t^{pq} \leq t^q$ for $t \geq 1$. A standard formula [53, equation (1.3)] leads to, for $u \geq 1$,

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

Putting all the pieces together we reach the conclusion.

APPENDIX B
COVERING NUMBERS

To estimate covering numbers in Lemma 15 we used implicitly the following Lemma, which follows directly from the definition of covering numbers and will serve to extend this result to other constraint sets.

Lemma 16. *Consider compact constraint sets $\mathfrak{D}_1, \ldots, \mathfrak{D}_k$ with respective covering number bounds $N_1, \ldots, N_k$ with respect to metrics $\rho_1(\cdot,\cdot), \ldots, \rho_k(\cdot,\cdot)$. Then the covering number of the Cartesian product $\mathfrak{D} \triangleq \mathfrak{D}_1 \times \ldots \times \mathfrak{D}_k$ (with respect to the $\max_i \rho_i(\mathbf{D}_i, \mathbf{D}_i')$ metric) has the upper bound*

$$\mathcal{N}(\mathfrak{D}, \epsilon) \leq \prod_{i=1}^{k} N_i.$$

In order to further extend these results to more elaborate constraint sets we resort to a result from [54].

Lemma 17. *Given a constraint set $\mathfrak{D}$ with distance measure ρ_1, a normed space $\mathfrak{M}$ with distance measure ρ, and a mapping $\Phi\colon \mathfrak{M} \to \mathfrak{D}$ that fulfill the conditions*

1) *There exists an $r \in \mathbb{R}, r > 0$ such that $\Phi(B_r) \supset \mathfrak{D}$ where B_r is a ball around $0 \in \mathfrak{M}$ with radius r with respect to the distance metric ρ (Surjectivity).*
2) *There exists an $L \in \mathbb{R}$ such that $\rho_1(\Phi(\mathbf{D}_1), \Phi(\mathbf{D}_2)) \leq L \cdot \rho(\mathbf{D}_1, \mathbf{D}_2)$ for $\mathbf{D}_1, \mathbf{D}_2 \in B_r$ (Lipschitz property).*

Then the covering number of $\mathfrak{D}$ has the upper bound

$$\mathcal{N}(\mathfrak{D}, \epsilon) \leq \left(3rL/\epsilon\right)^h$$

where the exponent h is the dimension of the constraint set $\mathfrak{D}$ in the sense of its manifold structure.

These lemmata imply covering number bounds for all the constraint sets introduced in the previous section.

1) Sparse dictionaries: The class of s-sparse dictionaries $\mathfrak{D}_{\text{sparse}}(m, d, s, \mathbf{\Theta})$ with $\mathbf{\Theta} \in \mathfrak{D}_{\text{orth}}(m)$ is the Cartesian product of d copies of

$$\mathfrak{D}_{\text{sparse}}(m, 1, s, \mathbf{\Theta}) \triangleq \{\mathbf{\Theta}\mathbf{d} \,:\, \mathbf{d} \in \mathbb{R}^d, \|\mathbf{d}\|_0 \leq s\}. \tag{65}$$

By Lemma 16 its covering number with the $\|\cdot\|_{1\to 2}$ metric is

$$\mathcal{N}(\mathfrak{D}_{\text{sparse}}(m, d, s, \mathbf{\Theta}), \epsilon) \leq \left(\mathcal{N}(\mathfrak{D}_{\text{sparse}}(m, 1, s, \mathbf{\Theta}), \epsilon)\right)^d.$$

Since the set $\mathfrak{D}_{\text{sparse}}(m, 1, s, \mathbf{\Theta})$ is isometric to $\mathfrak{D}_{\text{sparse}}(m, 1, s, \mathbf{I}_m)$, we have $\mathcal{N}(\mathfrak{D}_{\text{sparse}}(m, 1, s, \mathbf{\Theta}), \epsilon) = \mathcal{N}(\mathfrak{D}_{\text{sparse}}(m, 1, s, \mathbf{I}_m), \epsilon)$ these covering numbers are with respect to the Euclidean metric). Now, observe that $\mathfrak{D}_{\text{sparse}}(m, 1, s, \mathbf{I}_m)$ is simply the set of normalized s-sparse vectors in $\mathbb{R}^m$, which is a union of $\binom{m}{s}$ $(s-1)$-spheres from as many subspaces. Hence, we obtain $\mathcal{N}(\mathfrak{D}_{\text{sparse}}(m, 1, s, \mathbf{I}_m), \epsilon) \leq \binom{m}{s}\left(1 + 2/\epsilon\right)^s$, yielding

$$\mathcal{N}(\mathfrak{D}_{\text{sparse}}(m, d, s, \mathbf{\Theta}), \epsilon) \leq \left(\binom{m}{s}\left(1 + 2/\epsilon\right)^s\right)^d.$$

2) Orthogonal Dictionaries: The orthogonal dictionaries as defined in Equation (54) are the group of special orthogonal matrices. The special orthogonal group is a Lie group, and the corresponding Lie algebra is the vector space of skew symmetric matrices $\mathfrak{so}(d)$, i.e., $\exp\colon \mathfrak{so}(d) \to \mathfrak{D}_{\mathrm{orth}}(d)$ is a surjective map. Furthermore, it can be shown that it is sufficient to restrict $\mathfrak{so}(d)$ to a ball around $\mathbf{0}$ with radius π in order to obtain a surjective map. The exponential mapping is Lipschitz since it holds that

$$\rho(\exp(\mathbf{S}_1), \exp(\mathbf{S}_2)) \leq e^a \|\mathbf{S}_1 - \mathbf{S}_2\|, \tag{66}$$

with $\mathbf{S}_1, \mathbf{S}_2 \in B_\pi$ and $a = \max(\|\mathbf{S}_1\|, \|\mathbf{S}_2\|) \leq \pi$ and we obtain the Lipschitz constant $L = e^\pi$. Finally, the dimension of $\mathfrak{D}_{\mathrm{orth}}(d)$ is $d(d-1)/2$ which yields the covering number bound

$$\mathcal{N}(\mathfrak{D}_{\mathrm{orth}}(d), \epsilon) \leq (3\pi e^\pi/\epsilon)^{d(d-1)/2}.$$

3) Stiefel matrices: The Stiefel manifold, defined in Equation (55), can also be defined as the quotient space $\mathfrak{D}_{\mathrm{orth}}(m)/\mathfrak{D}_{\mathrm{orth}}(m-d)$, cf. [55]. In the previous section, we already mentioned that the Lie algebra of $\mathfrak{D}_{\mathrm{orth}}(m)$ is given by $\mathfrak{so}(m)$. Now, we define the function

$$\begin{aligned}&\Phi\colon \mathfrak{so}(m) \to \mathfrak{D}_{\mathrm{St}}(d, m),\\ &\Phi \triangleq q \circ \exp,\end{aligned}$$

where $q\colon \mathfrak{D}_{\mathrm{orth}}(m) \to \mathfrak{D}_{\mathrm{orth}}(m)/\mathfrak{D}_{\mathrm{orth}}(m-d)$ denotes the quotient mapping. The Lipschitz constant for the exponential mapping has been previously established in Equation (66). The function q is a quotient map and therefore a contraction, and as the mapping Φ is a combination of two Lipschitz mappings it is Lipschitz itself. Thus, conditions *(1)* and *(2)* in Lemma 17 are fulfilled and taking into account that $\dim(\mathfrak{D}_{\mathrm{St}}(d, m)) = md - d(d+1)/2$ we obtain the covering number bound

$$\mathcal{N}(\mathfrak{D}_{\mathrm{St}}(d, m), \epsilon) \leq (3\pi e^\pi/\epsilon)^{md - d(d+1)/2}.$$

4) Separable Dictionaries: The mapping

$$\begin{aligned}\Phi\colon \mathfrak{D}(m_1, d_1) \times \ldots \times \mathfrak{D}(m_z, d_z) &\to \mathfrak{D}_{\mathrm{sep}},\\ (\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_2) &\mapsto \mathbf{D}_1 \otimes \mathbf{D}_2 \otimes \ldots \otimes \mathbf{D}_z\end{aligned}$$

is surjective and Lipschitz with a Lipschitz constant smaller than one. Thus, Lemmata 17 & 16 yield

$$\mathcal{N}(\mathfrak{D}_{\mathrm{sep}}, \epsilon) \leq (3/\epsilon)^{\sum_i m_i d_i}.$$

5) Tensor Product of Stiefel matrices: Analogous to above we introduce the mapping

$$\begin{aligned}\Phi\colon \mathfrak{D}_{\mathrm{St}}(m_1, d_1) \times \ldots \times \mathfrak{D}_{\mathrm{St}}(m_z, d_z) &\to \mathfrak{D}_{\otimes \mathrm{St}},\\ (\mathbf{D}_1, \ldots, \mathbf{D}_z) &\mapsto \mathbf{D}_1 \otimes \ldots \otimes \mathbf{D}_z.\end{aligned}$$

As this is just a special case of the mapping in the previous constraint set, this mapping is surjective as well as a contraction. Hence, according to Lemmata 17 and16 the covering number bound is given by

$$\mathcal{N}(\mathfrak{D}_{\otimes \mathrm{St}}, \epsilon) \leq (3\pi e^\pi/\epsilon)^{\sum_i m_i d_i - d_i(d_i+1)/2}.$$

REFERENCES

[1] S. Mallat, *A Wavelet Tour of Signal Processing*, 3rd ed. Academic Press, 2008.

[2] M. Elad, M. A. T. Figueiredo, and Y. M., "On the role of sparse and redundant representations in image processing," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 972–982, 2010.

[3] M. Elad, *Sparse and Redundant Representations*, ser. From Theory to Applications in Signal and Image Processing. Springer, 2010.

[4] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.

[5] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by VI?" *Vision Research*, vol. 37, no. 23, pp. 3311–3326, 1997.

[6] K. Engan, S. Aase, and J. Hakon Husoy, "Method of optimal directions for frame design," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1999, pp. 2443–2446.

[7] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T. W. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Computation*, vol. 15, no. 2, pp. 349–396, 2003.

[8] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[9] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research*, vol. 11, no. 1, pp. 19–60, 2010.

[10] I. Tošić and P. Frossard, "Dictionary learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 27–38, 2011.

[11] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.

[12] S. Hawe, M. Seibert, and M. Kleinsteuber, "Separable dictionary learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[13] R. Rubinstein, M. Zibulevsky, and M. Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1553–1564, 2010.

[14] D. Seung and L. Lee, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[15] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Springer, 1992.

[16] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet, "A direct formulation for sparse PCA using semidefinite programming," *arXiv*, Jun. 2004.

[17] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, 2009.

[18] Y. Zhang and L. El Ghaoui, "Large-scale sparse principal component analysis with application to text data," *arXiv*, Oct. 2012.

[19] A. Maurer and M. Pontil, "K-dimensional coding schemes in hilbert spaces," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5839–5846, 2010.

[20] D. Vainsencher, S. Mannor, and A. M. Bruckstein, "The sample complexity of dictionary learning," *Journal of Machine Learning Research*, vol. 12, pp. 3259–3281, 2011.

[21] R. Jenatton, J.-Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *Journal of Machine Learning Research*, vol. 12, pp. 2777–2824, 2011.

[22] R. Chartrand, "Exact reconstruction of sparse signals via nonconvex minimization," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 707–710, 2007.

[23] B. Mailhé, S. Lesage, R. Gribonval, F. Bimbot, P. Vandergheynst *et al.*, "Shift-invariant dictionary learning for sparse representations: extending K-SVD," in *16th EUropean SIgnal Processing COnference (EUSIPCO'08)*, 2008.

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

[25] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM Journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, 2000.

[26] R. Jenatton, R. Gribonval, and F. Bach, "Local stability and robustness of sparse dictionary learning in the presence of noise," CMAP, Tech. Rep., 2012.

[27] V. V. Buldygin and I. U. V. Kozachenko, *Metric characterization of random variables and random processes*. American Mathematical Society, 2000, vol. 188.
[28] D. A. Spielman, H. Wang, and J. Wright, "Exact recovery of sparsely-used dictionaries," *Journal of Machine Learning Research: Workshop and Conference Proceeedings; 25th Annual Conference on Learning Theory*, vol. 23, pp. 1–18, 2012.
[29] S. Arora, R. Ge, and A. Moitra, "New Algorithms for Learning Incoherent and Overcomplete Dictionaries," *arXiv*, Aug. 2013.
[30] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon, "Learning Sparsely Used Overcomplete Dictionaries via Alternating Minimization," *arXiv*, Oct. 2013.
[31] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bulletin of the American Mathematical Society*, vol. 39, pp. 1–49, 2002.
[32] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.
[33] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers," *Statistical Science*, vol. 27, no. 4, pp. 538–557, Nov. 2012.
[34] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
[35] C. Hage and M. Kleinsteuber, "Robust PCA and subspace tracking from incomplete observations using ℓ_0-surrogates," *Computational Statistics*, vol. 29, no. 2, 2014.
[36] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," Preprint arXiv:1011.3027, Tech. Rep., 2010.
[37] D. Hsu, S. M. Kakade, and T. Zhang, "A tail inequality for quadratic forms of subgaussian random vectors," *Electronic Communications in Probability*, vol. 17, no. 52, pp. 1–6, 2012.
[38] G. Bennett, "Probability inequalities for the sum of independent random variables," *Journal of the American Statistical Association*, 1962.
[39] J. Shawe-Taylor, C. K. I. Williams, N. Cristianini, and J. Kandola, "On the eigenspectrum of the gram matrix and the generalization error of kernel-PCA," *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2510–2522, Jul. 2005.
[40] G. Blanchard, O. Bousquet, and L. Zwald, "Statistical properties of kernel principal component analysis," *Machine Learning*, vol. 66, no. 2-3, pp. 259–294, 2007.
[41] D. Pollard, "A central limit theorem for k-means clustering," *The Annals of Probability*, vol. 10, no. 4, pp. 919–926, Nov. 1982.
[42] T. Linder, "On the training distortion of vector quantizers," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1617–1623, 2000.
[43] P. L. Bartlett, T. Linder, and G. Lugosi, "The minimax distortion redundancy in empirical quantizer design," *IEEE Transactions on Information Theory*, vol. 44, no. 5, pp. 1802–1813, 1998.
[44] A. Antos, "Improved minimax bounds on the test and training distortion of empirically designed vector quantizers," *IEEE Transactions on Information Theory*, vol. 51, no. 11, pp. 4022–4032, 2005.
[45] T. Linder, G. Lugosi, and K. Zeger, "Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding," *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 1728–1740, 1994.
[46] G. Biau, L. Devroye, and G. Lugosi, "On the performance of clustering in hilbert spaces," *IEEE Transactions on Information Theory*, vol. 54, no. 2, pp. 781–790, 2008.
[47] K. Sridharan, S. Shalev-Shwartz, and N. Srebro, "Fast Rates for Regularized Objectives," *Advances in Neural Information Processing Systems*, vol. 21, pp. 1545–1552, 2008.
[48] C. Levrard, "Fast rates for empirical vector quantization," *Electronic Journal of Statistics*, vol. 7, pp. 1716–1746, 2013.
[49] ——, "Non Asymptotic Bounds for Vector Quantization in Hilbert Spaces," *Annals of Statistics*, vol. 43, pp. 592–619, Apr. 2015.
[50] R. Gribonval and K. Schnass, "Dictionary identification - Sparse matrix-factorisation via ℓ_1 minimisation," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3523–3539, 2010.
[51] Q. Geng, H. Wang, and J. Wright, "On the local correctness of ℓ^1 minimization for dictionary learning," *arXiv*, Jan. 2011.
[52] K. Schnass, "On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD," *CoRR*, vol. abs/1301.3375, 2013.
[53] 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.
[54] S. J. Szarek, "Metric entropy of homogeneous spaces and finsler geometry of classical lie groups," in *Quantum Probability*, ser. Banach Center Publications, vol. 43. Polish Academy of Sciences, 1998, pp. 395–410.
[55] A. Edelman, T. A. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM Journal on Matrix Analysis and Applications*, vol. 20, no. 2, pp. 303–353, 1998.

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

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

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

Martin Kleinsteuber received his PhD in Mathematics from the University of Würzburg, Germany, in 2006. After post-doc positions at National ICT Australia Ltd, the Australian National University, Canberra, Australia, and the University of Würzburg, he has been appointed assistant professor for geometric optimization and machine learning at the Department of Electrical Engineering and Information Technology, TU München, Germany, in 2009. He won the SIAM student paper prize in 2004 and the Robert-Sauer-Award of the Bavarian Academy of Science in 2008 for his works on Jacobi-type methods on Lie algebras. His research interests are in the areas of statistical signal processing, machine learning and computer vision, with an emphasis on optimization and learning methods for sparse and robust signal representations and blind signal separation.

Matthias Seibert received the Dipl.-math. degree from the Julius-Maximilians-Universität Würzburg, Würzburg, Germany, in 2012. He is currently pursuing the doctoral degree at the Department of Electrical and Computer Engineering, Technische Universitt München, Munich, Germany. His current research interests include sparse signal models, geometric optimization, and learning complexity.