Rémi Gribonval 1 & Rodolphe Jenatton 2 & Francis Bach 3
& Martin Kleinsteuber 4 & Matthias Seibert 5

1 Inria - PANAMA Project-Team, Rennes             remi.gribonval@inria.fr
2 Inria - SIERRA Project-Team, Paris              r.jenatton@criteo.com
3 Inria - SIERRA Project-Team, Paris              francis.bach@ens.fr
4 Dept of Elec. Eng. and Information Techn., Techn. Univ. München, kleinsteuber@tum.de
5 Dept of Elec. Eng. and Information Techn., Techn. Univ. München, m.seibert@tum.de

Résumé. La modélisation de données de grande dimension comme combinaisons
linéaires parcimonieuses d’atomes d’un dictionnaire est devenu un outil très populaire en
traitement du signal et de l’image. Etant donné l’importance du choix du dictionnaire
pour le déploiement opérationnel de ces outils, des approches basées sur l’apprentissage à
partir d’une collection ont connu un bel essor. Les techniques les plus populaires abordent
le problème sous l’angle de la factorisation de grandes matrices via la minimisation d’une
fonction de coût non-convexe. Si des progrès importants en terme d’efficacité algorithmique
ont favorisé leur diffusion, ces approches restaient jusqu’à récemment essentiellement
empiriques. Nous présenterons des travaux récents abordant les aspects statistiques
de ces techniques et contribuant à caractériser l’excès de risque en fonction du nombre
d’exemples disponibles. Les résultats couvrent non seulement l’apprentissage de dictionnaire
pour les représentations parcimonieuses, mais également une classe sensiblement
plus large de factorisations de matrices sous contraintes.

Mots-clés. Apprentissage de dictionnaire, parcimonie.

Abstract. A popular approach within the signal processing and machine learning
communities consists in modelling high-dimensional data as sparse linear combinations of
atoms selected from a dictionary. Given the importance of the choice of the dictionary
for the operational deployment of these tools, a growing interest for learned dictionaries
has emerged. The most popular dictionary learning techniques, which are expressed as
large-scale matrix factorization through the optimization of a non convex cost function,
have been widely disseminated thanks to extensive empirical evidence of their success and
steady algorithmic progress. Yet, until recently they remained essentially heuristic. We
will present recent work on statistical aspects of sparse dictionary learning, contributing
to the characterization of the excess risk as a function of the number of training samples.
The results cover non only sparse dictionary learning but also a much larger class of
constrained matrix factorization problems.

Keywords. Dictionary learning, sparse representation.

1 This work was supported in part by the European Research Council, PLEASE project (ERC-StG-
2011-277906).
1 Dictionary learning and matrix factorization

The fact that a signal $x \in \mathbb{R}^m$ which belongs to a certain class has a representation over some class dependent dictionary $D \in \mathbb{R}^{m \times d}$ is the backbone of many successful signal reconstruction and data analysis algorithms [18, 6, 7]. That is, $x$ is the linear combination of columns of $D$, referred to as atoms. Formally, this reads as $x = D\alpha$, where the coefficient vector $\alpha \in \mathbb{R}^d$ as well as the dictionary $D$ are subject to some constraints. Such a setting covers prominent examples like Principal Component Analysis (PCA), where $D$ 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 $D$ consists of normalized columns that form an overcomplete basis of the signal space, and $\alpha \in \mathbb{R}^d$ is assumed to be sparse.

The task of learning such dictionaries from a given set of training data is related to matrix factorization. Important examples include Higher-Order SVD (also known as multilinear SVD) [25], sparse coding also called dictionary learning [20, 8, 15, 1, 17, 24, 22], its variants with separable [11] or sparse [21] dictionaries, Non-negative Matrix Factorization (NMF) [23], $K$-means clustering [9], sparse PCA [4, 27, 28], and more.

The learning task is expressed formally as follows. Let $X = [x_1, \ldots, x_n] \in \mathbb{R}^{m \times n}$ be the matrix containing the $n$ training samples arranged as its columns, and let $A = [\alpha_1, \ldots, \alpha_n] \in \mathbb{R}^{d \times n}$ contain the corresponding $n$ coefficient vectors, a common approach to the dictionary learning process is the optimization problem

$$\minimize_{A,D} \frac{1}{2}\|X - DA\|_F^2 + \sum_{i=1}^n g(\alpha_i) \quad \text{subject to} \quad D \in \mathcal{D}. \quad (1)$$

Therein, $g: \mathbb{R}^d \to \mathbb{R}$ is a penalty function promoting certain constraints for the coefficients vectors, e.g. sparsity or positivity, and $\mathcal{D}$ is some predefined admissible set of solutions for the dictionary.

2 Sample complexity

A fundamental question in such a learning process is the sample complexity issue. Assuming that the training samples $x_i$ are drawn according to some distribution $P$ representing the class of signals of interest, one would ideally like to select the dictionary $D^*$ yielding the minimum expected value of (1). However, having only access to $n$ training samples, one can at best select an empirical minimizer $\hat{D}$. 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, 26] primarily consider the case where $g$ is the indicator function of a set, such as an $\ell^1$ or an $\ell^0$ ball, $\mathcal{D}$ is the set of all unit
norm dictionaries or a subset with a restricted isometry property, and the distribution \( P \) is in the class \( \mathcal{P} \) of distributions on the unit sphere of \( \mathbb{R}^m \). We generalize these results to:

- **General classes of penalty functions.** Examples covered by our results include: the \( \ell^1 \) penalty and its powers; any mixed norm [12] or quasi-norm [3]; the characteristic function of the \( \ell^1 \)-ball, of the set of \( k \)-sparse vectors [26], or of non-negative vectors [23].

- **Various classes of dictionaries \( \mathcal{D} \) that can incorporate structures.** Examples covered by our results include: dictionaries with unit norm atoms which are used in many dictionary learning schemes, e.g. K-SVD [1]; sparse dictionaries [21]; shift-invariant dictionaries [16]; tensor product dictionaries [11]; orthogonal dictionaries; non-negative dictionaries [23]; topic models [14]; and tensor products of Stiefel matrices used for Higher-order SVDs [25, 5].

- **Various classes \( \mathcal{P} \) of probability distributions \( P \).** Examples include distributions on the unit sphere which are tackled in [26], but also certain distributions built using sub-Gaussian random variables [13]. For more information on sub-Gaussian random variables, see [2].

3 Notations and main results

Given a dictionary \( D \in \mathbb{R}^{m \times d} \) that fulfills certain structural properties and a signal \( x \in \mathbb{R}^m \), a representation vector \( \alpha \in \mathbb{R}^d \) is typically obtained by minimizing

\[
L_x(D, \alpha) \triangleq \frac{1}{2} \|x - D\alpha\|_2^2 + g(\alpha).
\]

where \( g : \mathbb{R}^d \to \mathbb{R}^+ \cup \{+\infty\} \) is a penalty function promoting constraints for the coefficient vector. For our purposes, we define the quality of how well a signal \( x \) can be coded by a dictionary \( D \) by

\[
f_x(D) \triangleq \inf_{\alpha \in \mathbb{R}^d} L_x(D, \alpha),
\]

Given \( n \) training samples \( X = [x_1, \ldots, x_n] \in \mathbb{R}^{m \times n} \), the average fit of dictionary to \( X \) given the penalty is

\[
F_X(D) \triangleq \frac{1}{n} \sum_{i=1}^n f_{x_i}(D).
\]

We show\(^2\) the uniform convergence of the empirical cost function to its expectation, that is to say

\[
\sup_{D \in \mathcal{D}} |F_X(D) - \mathbb{E}_{x \sim P} f_{x}(D)| \leq \eta_n(g, \mathcal{D}, \mathcal{P})
\]

\(^2\)All technical details can be found in the preprint [10].
holds with high probability, where \( \eta_n(g, \mathcal{D}, \mathcal{P}) \propto \sqrt{\log \frac{n}{n}} \). As a consequence, we control the generalization bound with the empirical optimum \( \hat{D} \): with controlled high probability,

\[
E_{x \sim \mathcal{P}} f_x(\hat{D}_n) \leq E_{x \sim \mathcal{P}} f_x(D^*) + 2\eta_n.
\]

### 4 Outline of the approach

For this, we primarily show that for a given draw \( X \) of the training set, the function \( D \mapsto F_X(D) \) is Lipschitz with a constant expressed in terms of \( X \) and \( g \), and conclude using a classical argument based on covering numbers and concentration of measure. While the latter technique is fairly standard, our main contribution lies in the identification of two large classes of penalty functions \( g \) for which the desired Lipschitz property holds and is nicely controlled. The first class we handle consists of non-negative, lower semi-continuous and coercive penalty function \( g \) such that \( g(0) = 0 \). This notably covers all norms, quasi-norms, and their powers. For such penalty functions we show that \( F_X(\cdot) \) is uniformly Lipschitz as a function of \( D \in \mathbb{R}^{m \times p} \) (unconstrained dictionary). The second class covers as particular cases the indicator function of \( k \)-sparse vectors and that of non-negative vectors, with restrictions on the class \( \mathcal{D} \) of admissible dictionaries, which must satisfy certain properties related to the popular restricted isometry property extensively used in the analysis of sparse recovery for inverse problems with a fixed dictionary.

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.

The obtained sample complexity results applied to sparse coding extend those of Maurer and Pontil [19] and Vainsencher et al. [26] 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 [26] 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 \( K \) such that \( \sup_{D \in \mathcal{D}, \alpha \in K} \|D\alpha\|_2 < \infty \) [19, Theorem 2], or of \( \ell^1 \) or \( \ell^0 \) balls [26].

### References


