Apprentissage de dictionnaire pour les représentations parcimonieuses

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 Dictionary learning and matrix factorization

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 [18, 6, 7]. That is, $\mathbf{x}$ is the linear combination of columns of $\mathbf{D}$, referred to as *atoms*. Formally, this reads as $\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}$, 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. 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 $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$ be the matrix containing the n training samples arranged as its columns, and let $\mathbf{A} = [\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_n] \in \mathbb{R}^{d \times n}$ contain the corresponding n coefficient vectors, a common approach to the dictionary learning process is the optimization problem

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

Therein, $g\colon \mathbb{R}^d \to \mathbb{R}$ is a penalty function promoting certain constraints for the coefficients vectors, e.g. sparsity or positivity, and $\mathfrak{D}$ 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 $\mathbf{x}_i$ 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 (1). 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, 26] 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 by our results include: the ℓ^1 penalty and its powers; any mixed norm [12] or quasi-norm [3]; the characteristic function of the ℓ^1-ball, of the set of k-sparse vectors [26], or of non-negative vectors [23].
- **Various classes of dictionaries $\mathfrak{D}$ 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 $\mathfrak{P}$ of probability distributions $\mathbb{P}$.** 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 $\mathbf{D} \in \mathbb{R}^{m \times d}$ 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 minimizing

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

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

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

Given n training samples $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$, the average fit of dictionary to $\mathbf{X}$ given the penalty is

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

We show[1] the uniform convergence of the empirical cost function to its expectation, that is to say

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

[1] All technical details can be found in the preprint [10].

holds with high probability, where $\eta_n(g, \mathfrak{D}, \mathfrak{P}) \propto \sqrt{\frac{\log n}{n}}$. As a consequence, we control the generalization bound with the empirical optimum $\hat{\mathbf{D}}$: with controlled high probability,

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

4 Outline of the approach

For this, we primarily show that for a given draw $\mathbf{X}$ of the training set, the function $\mathbf{D} \mapsto F_{\mathbf{X}}(\mathbf{D})$ is Lipschitz with a constant expressed in terms of $\mathbf{X}$ 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 $F_{\mathbf{X}}(\cdot)$ is *uniformly Lipschitz* as a function of $\mathbf{D} \in \mathbb{R}^{m\times p}$ (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 $\mathfrak{D}$ 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 $\mathcal{K}$ 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 [26].

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

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

[3] Rick Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Processing Letters*, 14(10):707–710, 2007.

[4] Alexandre d'Aspremont, Laurent El Ghaoui, Michael I Jordan, and Gert R G Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *arXiv*, June 2004.

[5] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[6] M. Elad, M. A. T. Figueiredo, and Yi M. On the role of sparse and redundant representations in image processing. *Proceedings of the IEEE*, 98(6):972–982, 2010.

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

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

[9] Allen Gersho and Robert M Gray. *Vector Quantization and Signal Compression*. Springer, 1992.

[10] Remi Gribonval, Rodolphe Jenatton, Francis Bach, Martin Kleinsteuber, and Matthias Seibert. Sample Complexity of Dictionary Learning and other Matrix Factorizations. December 2013.

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

[12] Rodolphe Jenatton, Jean-Yves Audibert, and Francis Bach. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12:2777–2824, 2011.

[13] Rodolphe Jenatton, Remi Gribonval, and Francis Bach. Local stability and robustness of sparse dictionary learning in the presence of noise. Technical report, CMAP, 2012.

[14] Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, and Francis Bach. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12:2297–2334, 2011.

[15] 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*, 15(2):349–396, 2003.

[16] Boris Mailhé, Sylvain Lesage, Rémi Gribonval, Frédéric Bimbot, Pierre Vandergheynst, et al. Shift-invariant dictionary learning for sparse representations: extending K-SVD. In *16th EUropean SIgnal Processing COnference (EUSIPCO'08)*, 2008.

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

[18] Stéphane Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 3rd edition, 2008.

[19] Andreas Maurer and Massimiliano Pontil. K-dimensional coding schemes in hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, 2010.

[20] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by VI? *Vision Research*, 37(23):3311–3326, 1997.

[21] R. Rubinstein, M. Zibulevsky, and M. Elad. Double sparsity: Learning sparse dictionaries for sparse signal approximation. *IEEE Transactions on Signal Processing*, 58(3):1553–1564, 2010.

[22] Ron Rubinstein, A M Bruckstein, and Michael Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.

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

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

[25] Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[26] Daniel Vainsencher, Shie Mannor, and Alfred M. Bruckstein. The sample complexity of dictionary learning. *Journal of Machine Learning Research*, 12:3259–3281, 2011.

[27] Daniela M. Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

[28] Youwei Zhang and Laurent El Ghaoui. Large-scale sparse principal component analysis with application to text data. *arXiv*, October 2012.