Structured sparsity: towards a "deep" understanding

Angélique Drémeau
ENSTA Bretagne, Lab-STICC (UMR 6285)
Brest, France
Email: angelique.dremeau@ensta-bretagne.fr

Florent Krzakala
UPMC Paris 6, LPS-ENS (UMR 8550)
Paris, France
Email: florent.krzakala@ens.fr

***Abstract*—Although of proven interest for decomposition algorithms, structures in sparse representations are rarely known in practice. In this work, we propose to model structures through so-called *restricted Boltzmann machines*, for which efficient learning algorithms exist. The model is then exploited into a variational Bayesian procedure. The approach is shown to present a good behavior with regard to its non-structured counterpart.**

***Index Terms*—Structured sparse representations, restricted Boltzmann machine, variational Bayesian approximations**

I. Introduction

Taking into account the structures naturally living in signal representations has proved to be relevant for the performance of the sparse decomposition algorithms (*e.g.* [1], [2]). Within this context, we proposed and developed in [3] a generic Bayesian algorithm, exploiting a Boltzmann machine (BM) (also considered in [4], [5], [6]) to model various types of structures. Formally, considering the observation model $\mathbf{y} = \sum_{i=1}^{M} s_i\, x_i\, \mathbf{d}_i + \mathbf{n}$, the support $\mathbf{s} \in \{0,1\}^M$ is assumed to obey

$$p(\mathbf{s}) \propto \exp(\mathbf{b}^T\mathbf{s} + \mathbf{s}^T\mathbf{W}\mathbf{s}), \tag{1}$$

while we classically moreover suppose $\mathbf{n} \sim \mathcal{N}(0, \sigma_n^2 \mathbf{I}_N)$ and $x_i \sim \mathcal{N}(0, \sigma_{x_i}^2)$, $\forall i \in \{1, \ldots, M\}$. BM encompasses many well-known probabilistic models as particular cases and offers then a nice option for a wide range of dictionaries and classes of signals. However the learning of its parameters is a difficult problem, which largely limits its practical use (structures, and thus parameters, are rarely known).

Inspired by recent works in neural networks, we propose here to replace model (1) by a so-called "restricted" BM (RBM) as

$$p(\mathbf{s}) = \sum\nolimits_{\mathbf{h}} p(\mathbf{s},\mathbf{h}) \propto \sum\nolimits_{\mathbf{h}} \exp(\mathbf{a}^T\mathbf{h} + \mathbf{b}^T\mathbf{s} + \mathbf{s}^T\mathbf{W}\mathbf{h}), \tag{2}$$

where $\mathbf{h}$ is a L-dimensional binary hidden variable. The RBM is the building block of "deep belief networks" [7] and has recently sparked a surge of interest partly because of the efficient algorithms developed to train it (as the Contrastive Divergence (CD) [8]).

II. Deep Structured SoBaP

Based on this model, we consider the following marginalized Maximum A Posteriori (MAP) estimation problem

$$\hat{\mathbf{s}} = \underset{\mathbf{s}\in\{0,1\}^M}{\operatorname{argmax}} \log p(\mathbf{s}|\mathbf{y}), \tag{3}$$

where $p(\mathbf{s}|\mathbf{y}) = \int_{\mathbf{x}} p(\mathbf{x},\mathbf{s}|\mathbf{y}) d\mathbf{x}$. To solve this problem, different sub-optimal techniques can be used. In the continuation of previous works [3], [9], [10], we are interested here in the solutions brought by variational approaches, which aim to approximate the posterior distribution $p(\mathbf{x},\mathbf{s}|\mathbf{y})$ by a distribution $q(\mathbf{x},\mathbf{s})$ leading to the minimum of the Kullback-Leibler divergence under specific sets of constraints. In particular, considering the factorization constraint $q(\mathbf{x},\mathbf{s}) = \prod_{i=1}^{M} q(x_i, s_i) = \prod_{i=1}^{M} q(x_i|s_i)\ q(s_i)$, we focus on a mean-field (MF) approximation, which can be in practice efficiently solved by an iterative algorithm, called "variational Bayes EM algorithm" [11]. Particularized to our model, the method gives rise to the following updates:

$$q(x_i|s_i) = \mathcal{N}(m(s_i), \Sigma(s_i)), \quad q(s_i) \propto \sqrt{\Sigma(s_i)} \exp\left(\frac{1}{2}\frac{m(s_i)^2}{\Sigma(s_i)}\right) \tilde{p}(s_i),$$

$$\text{where } \Sigma(s_i) = \frac{\sigma_{x_i}^2 \sigma_n^2}{\sigma_n^2 + s_i \sigma_{x_i}^2 \mathbf{d}_i^T \mathbf{d}_i}, \; m(s_i) = s_i \frac{\sigma_{x_i}^2}{\sigma_n^2 + s_i \sigma_{x_i}^2 \mathbf{d}_i^T \mathbf{d}_i} \langle \mathbf{r}_i \rangle^T \mathbf{d}_i,$$

$$\langle \mathbf{r}_i \rangle = \mathbf{y} - \sum\nolimits_{j\neq i} q(s_j = 1)\, m(s_j = 1)\, \mathbf{d}_j,$$

$$\tilde{p}(s_i) \propto \exp(b_i s_i) \prod\nolimits_{l=1}^{L} (1 + \exp(a_l + s_i w_{li} + \sum\nolimits_{j\neq i} q(s_j{=}1)\, w_{lj})).$$

The procedure presents a complexity $\mathcal{O}(M^2)$ per iteration, similar to the one of the approach proposed in [3]. Furthermore, it is of a more practical interest when the structures have to be learned. The use of RBMs being a natural bridge towards deep networks, we will refer to the proposed procedure as the "Deep Structured Soft Bayesian Pursuit" (*DSSoBaP*).

III. Proof of concept

To illustrate this advantage, we consider the MNIST database [12], widely used in the field of machine learning. The database is composed by 60000 training and 10000 testing handwritten digits, labelled from 0 to 9, in grayscale levels and of dimension $M = 28 \times 28$. The images are sparse, with $K = 150$ non-zero coefficients on average. The experimental procedure is as follows. We first train the RBM parameters on the sole supports of the training set, using Constrative Divergence [8] and setting $L = 10$. 100 images (10 per label) are then extracted from the testing set and reconstructed through a compressed sensing framework using a normalized zero-mean Gaussian sensing matrix $\mathbf{D}$. We evaluate the performance of *DSSoBaP*, and the one of its unstructured, Bernoulli-based counterpart *SoBaP* [3], in terms of the normalized mean-squared error (MSE) in function of the number of measurements N. Two different setups are then considered: $\sigma_n^2 = 0$ for Fig. 1 and $\sigma_n^2 = 0.01$ for Fig. 2. We can see here that *DSSoBaP* outperforms *SoBaP*, illustrating, to the extent of these experiments, the interest of exploiting structures through RBMs.

IV. Conclusion

In this paper, we have shown that RBMs can favorably be used to model (unknown) structures in sparse representations. As a proof of concept, its exploitation through a variational MF approximation leads to a promising approach. Future work should also investigate the strengths of such deep prior models with regard to the characterization of the structures.

Acknowledgment

This work has been supported in part by the ERC under the European Union's 7th Framework Programme Grant Agreement 307087-SPARCS.

REFERENCES

[1] L. Daudet, "Sparse and structured decompositions of signals with the molecular matching pursuit," *IEEE Trans. On Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1808 – 1816, September 2006.

[2] M. Kowalski, K. Siedenburg, and M. Dorfler, "Social sparsity! neighborhood systems enrich structured shrinkage operators," *IEEE Trans. On Signal Processing*, vol. 61, no. 10, pp. 2498 – 2511, May 2013.

[3] A. Drémeau, C. Herzet, and L. Daudet, "Boltzmann machine and mean-field approximation for structured sparse decompositions," *IEEE Trans. On Signal Processing*, vol. 60, no. 7, pp. 3425–3438, July 2012.

[4] P. J. Garrigues and B. A. Olshausen, "Learning horizontal connections in a sparse coding model of natural images," in *Advances in Neural Information Processing Systems (NIPS)*, December 2008, pp. 505–512.

[5] V. Cevher, M. F. Duarte, C. Hegde, and R. G. Baraniuk, "Sparse signal recovery using markov random fields," in *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2008.

[6] T. Peleg, Y. C. Eldar, and M. Elad, "Exploiting statistical dependencies in sparse representations for signal recovery," *IEEE Trans. On Signal Processing*, vol. 60, no. 5, pp. 2286 – 2303, May 2012.

[7] Y. Bengio, "Learning deep architectures for ai," *Foundations and trends in machine learning*, vol. 2, 2009.

[8] G. E. Hinton, "A practical guide to training restricted boltzmann machines," *Lecture Notes in Computer Science*, vol. 7700, pp. 599 – 619, 2012.

[9] F. Krzakala, M. Mezard, F. Sausset, Y. F. Sun, and L. Zdeborova, "Statistical-physics-based reconstruction in compressed sensing," *Physical Review X*, vol. 2, no. 021005, May 2012.

[10] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *IEEE Int'l Symposium on Information Theory (ISIT)*, July 2011.

[11] M. J. Beal and Z. Ghahramani, "The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures," *Bayesian Statistics*, vol. 7, pp. 453–463, 2003.

[12] Y. LeCun, C. Cortes, and C. Burges, "http://yann.lecun.com/exdb/mnist/," .

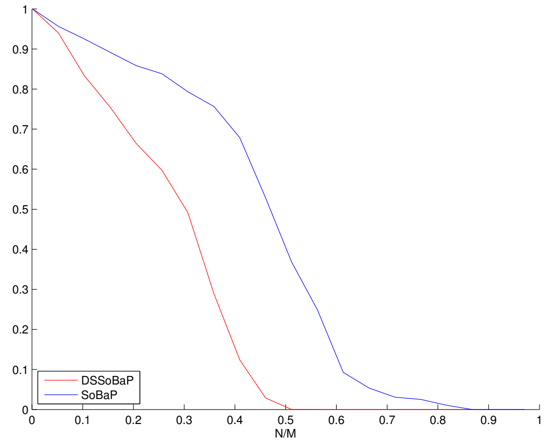

Fig. 1. Mean-squared error as a function of the number of measurements N (x-axis is N/M with $M = 784$) in the noiseless case.

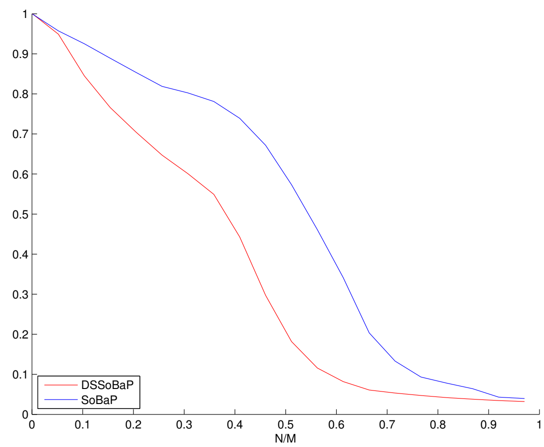

Fig. 2. Mean-squared error as a function of the number of measurements N (x-axis is N/M with $M = 784$) in the noisy case ($\sigma_n^2 = 0.01$).