Learning a fast transform with a dictionary

Olivier Chabiron[1], François Malgouyres[2], Jean-Yves Tourneret[1] and Nicolas Dobigeon[1].

[1]Institut de Recherche en Informatique de Toulouse, IRIT-CNRS UMR 5505, ENSEEIHT, Toulouse, France.

[2]Institut de Mathématiques de Toulouse, IMT-CNRS UMR 5219, Université de Toulouse, Toulouse, France.

***Abstract*— A powerful approach to sparse representation, dictionary learning consists in finding a redundant frame in which the representation of a particular class of images is sparse. In practice, all algorithms performing dictionary learning iteratively estimate the dictionary and a sparse representation of the images using this dictionary. However, the numerical complexity of dictionary learning restricts its use to atoms with a small support.**

A way to alleviate these issues is introduced in this paper, consisting in dictionary atoms obtained by translating the composition of K convolutions with S-sparse kernels of known support. The dictionary update step associated with this strategy is a non-convex optimization problem, which we study here.

A block-coordinate descent or Gauss-Seidel algorithm is proposed to solve this problem, whose search space is of dimension KS, which is much smaller than the size of the image. Moreover, the complexity of the algorithm is linear with respect to the size of the image, allowing larger atoms to be learned (as opposed to small patches). An experiment is presented that shows the approximation of a large cosine atom with $K = 7$ sparse kernels, demonstrating a very good accuracy.

1 Introduction

The problem we introduce in this paper is motivated by the dictionary learning (DL) field. DL was pioneered by [8, 9] and has received a growing attention since then. The principle behind DL is to find a representation for data that makes it simpler, sparser. We invite the reader to consult [4] for more details about sparse representations and DL. The archetype of the DL strategy is to look for a dictionary as the solution of the following optimization problem

$$\text{argmin}_{\mathbf{D},(\mathbf{x}_i)_{1\le i\le I}} \quad \sum_{i=1}^{I} \|\mathbf{D}\mathbf{x}_i - \mathbf{y}_i\|_2^2 + f(\mathbf{x}_i),$$

where $\mathbf{y}_i$ are the learning database, $\mathbf{D}$ is the dictionary matrix, whose columns are the atoms, and f is a sparsity-inducing function. The resulting problem can be solved (or approximatively solved) by many methods including MOD [5] and K-SVD [1]. All these approaches rely on alternatively updating the codes $\mathbf{x}_i$ and the dictionary $\mathbf{D}$.

Our primary motivation for considering the observation model (1) comes from computational issues. Usually, DL is applied to small patches, because of the computational cost of repeatedly computing of the matrix-vector products $\mathbf{D}\mathbf{x}_i$, which is worse than $\mathcal{O}(N^2)$. Moreover, the cost of the dictionary update is usually worse than $\mathcal{O}(N^3)$.

We propose a model where the learned atoms are a composition of K convolutions with S-sparse kernels. The interest for such a constraint is to provide numerically effective dictionaries and allow to consider larger atoms. Indeed, the search space is only of dimension KS, which is typically smaller than the size of the target atom.

The present work focuses on the dictionary update step of one atom. In this context, the code $\mathbf{x}$ is known. Our goals are both to approximate a large target atom $\boldsymbol{\kappa}$ with our model and to obtain target atoms whose manipulation is numerically efficient. This translates into a non-convex optimization problem.

2 Problem formulation

Let consider an observed d-dimensional signal $\mathbf{y}$ of $(\mathbb{R}^N)^d$, assumed to result from the convolution of a known input signal $\mathbf{x} \in (\mathbb{R}^N)^d$ with an unknown target kernel $\boldsymbol{\kappa} \in (\mathbb{R}^N)^d$, contaminated by an additive noise $\mathbf{b}$ following the linear model

$$\mathbf{y} = \boldsymbol{\kappa} * \mathbf{x} + \mathbf{b}, \tag{1}$$

where $*$ stands for the circular discrete convolution[1] in dimension d. For instance, the unknown target kernel $\boldsymbol{\kappa} \in (\mathbb{R}^N)^d$ may refer to the unknown impulse response of a 1D-filter or, conversely, to the point spread function of a 2D-filtering operator.

The problem addressed in this paper consists of approximating the unknown kernel $\boldsymbol{\kappa}$ by a composition of convolutions with $K \ge 2$ sparse kernels $(\mathbf{h}^k)_{1\le k\le K} \in ((\mathbb{R}^N)^d)^K$

$$\boldsymbol{\kappa} \approx \hat{\boldsymbol{\kappa}} \triangleq \mathbf{h}^1 * \cdots * \mathbf{h}^K. \tag{2}$$

The kernels $\mathbf{h}^1, \ldots, \mathbf{h}^K$ are constrained to have less than a fixed number S of non-zero elements, i.e., they are assumed to be at most S-sparse. As stated before, this assumption aims at providing a cost-effective dictionary by reducing the computations for $\mathbf{x} * \boldsymbol{\kappa}$. Furthermore, the locations of their non-zero elements in $\{0, \ldots, N\}^d$ are assumed to be known or pre-set. More precisely, the support of the kth kernel (i.e., the locations of the non zero elements of $\mathbf{h}^k$), denoted $\text{supp}\left(\mathbf{h}^k\right)$, is constrained to a fixed set of discrete indexes $\mathcal{S}_k$

$$\text{supp}\left(\mathbf{h}^k\right) \subset \mathcal{S}_k\ ,\ \forall k \in \{1, \ldots, K\} \tag{3}$$

An example of indexes for 1D convolution kernel would be

$$\mathcal{S}_k = \{k-1, 2k-1, \ldots, Sk-1\}. \tag{4}$$

Assuming that the noise vector $\mathbf{b}$ is an independent and identically distributed Gaussian sequence, approximating the unknown convolution kernel $\boldsymbol{\kappa}$ from the observed measurements $\mathbf{y}$ can be formulated as the following optimization problem

$$(P_0): \left\{ \begin{array}{l} \text{argmin}_{\mathbf{h}\in((\mathbb{R}^N)^d)^K} \|\mathbf{y} - \mathbf{h}^1 * \cdots * \mathbf{h}^K * \mathbf{x}\|_2^2, \\ \text{subject to supp}\left(\mathbf{h}^k\right) \subset \mathcal{S}_k\ , \forall k \in \{1, \ldots, K\} \end{array} \right.$$

where $\|\cdot\|_2$ stands for the usual Euclidean norm in $(\mathbb{R}^N)^d$. The problem (P_0) is non convex. Thus, depending on the values of K, $(\mathcal{S}_k)_{1\le k\le K}$, $\mathbf{x}$ and $\mathbf{y}$, it might be difficult or impossible

[1] All the elements of $(\mathbb{R}^N)^d$ are extended over $\mathbb{Z}^d$ by periodization.

to find a good approximation of a global minimizer of (P_0). Our objectives are to study if such a problem bends itself to global optimization, and to assess the approximation power of the computed compositions of convolutions.

3 Block-coordinate descent

The problem (P_0), as formulated in the previous section, is unhandy for global optimization. As detailed in [2], it has irrelevant stationary points and is non-convex (though infinitely differentiable). To adress these issues, a scalar weight λ is introduced and kernels are constrained to have a unit norm. Moreover, we elect a block-coordinate formulation in order to solve the problem with a Gauss-Seidel type algorithm (called Alternate Least Squares, sharing many similarities with the one used in [7]).

$$(P_k): \begin{cases} \text{argmin}_{\lambda\in\mathbb{R},\mathbf{h}\in\mathbb{R}^N} \|\mathbf{y}-\lambda\mathbf{h}*\mathbf{x}^k\|_2^2, \\ \text{subject to supp}(\mathbf{h})\subset\mathcal{S}_k \text{ and } \|\mathbf{h}\|_2=1 \end{cases}$$

with

$$\mathbf{x}^k = \mathbf{h}^1 * \cdots * \mathbf{h}^{k-1} * \mathbf{h}^{k+1} * \cdots * \mathbf{h}^K * \mathbf{x}, \tag{5}$$

where the kernels $\mathbf{h}^{k'}$ are fixed $\forall k' \neq k$. The problem (P_k) is linear and can be expressed as a matrix-vector product considering only the elements of $\mathbf{h}$ that belong to its support: The idea is to alternatively solve (P_k) by iterating on k. The support constraint significantly reduces the search space of the problem, and thus the amount of calculations needed to solve it for a stationary point. Algorithm 1 shows an overview of the resolution of (P_k). The computational complexity associated with a passage in the while loop is $O((K+S)KSN^d)$, i.e., it is linear with respect to the size N^d of the signal. The detailed steps to solving (P_k) are given in [2].

Algorithm 1: ALS algorithm

```
Input:
$\mathbf{y}$: target measurements;
$\mathbf{x}$: known coefficients;
$(\mathcal{S}_k)_{1\le k\le K}$: supports of the kernels $(\mathbf{h}^k)_{1\le k\le K}$.
Output:
$(\mathbf{h}^k)_{1\le k\le K}$: convolution kernels such that
$\mathbf{h}^1 * \ldots * \mathbf{h}^K \approx \boldsymbol{\kappa}$.
begin
    Initialize the kernels $((\mathbf{h}_p^k)_{p\in N})_{1\le k\le K}$;
    while not converged do
        for $k$ = 1 ,..., $K$ do
            Update $\mathbf{h}^k$ and $\lambda$ with a minimizer of $(P_k)$ ;
```

4 Synthetic example

In this section, we show an experiment consisting of approximating a 2D cosine atom $\boldsymbol{\kappa}$ in an image $\mathbf{y}$ of size 64×64 (i.e., $d = 2$ and $N = 64$). Such an atom can be seen as a large local cosine or a Fourier atom, both widely used in image processing. The interest of this atom is that it covers the whole image and is of a rather large support, making it difficult to handle with existing dictionary learning strategies.

$$\boldsymbol{\kappa}_p = \cos\left(2\pi\frac{\langle p,(2,5)\rangle}{N}\right) \quad ,\forall p\in\{0,\ldots,63\}^2.$$

The code $\mathbf{x}$ is a sparse image whose elements are chosen independent and identically distributed according to a Bernoulli-Gaussian distribution, widely used in sparse signal and image deconvolution [3, 6, 10]. Therefore, $\mathbf{y}$ contains a few weighted translations of the cosine atom $\boldsymbol{\kappa}$[2]. The target $\mathbf{y}$ is built with additive Gaussian noise of variance $\sigma^2 = 0.5$. Kernel supports have been set to a simple 5×5 square, linearly dilated with k, similar to the 1-D example given in (4).

Figures 1 and 2 show the cosine image $\mathbf{y}$, its approximation $\lambda\mathbf{x}*\mathbf{h}^1*\cdots*\mathbf{h}^K$, the actual atom $\boldsymbol{\kappa}$ and $\lambda\mathbf{h}^1*\cdots*\mathbf{h}^K$, for $K = 7$ and $S = 25$. The results obtained here are quite accurate even though the cosine image was corrupted by additive noise.

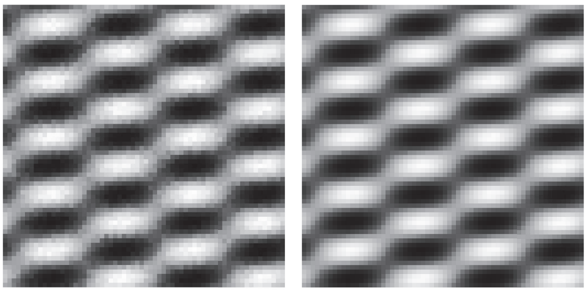

Figure 1: Cosine approximation with $K = 7$, $S = 25$, and Gaussian noise of variance $\sigma^2 = 0.5$. Cosine image $\mathbf{y}$ (left) and approximation $\lambda\mathbf{x}*\mathbf{h}^1*\cdots*\mathbf{h}^K$ (right).

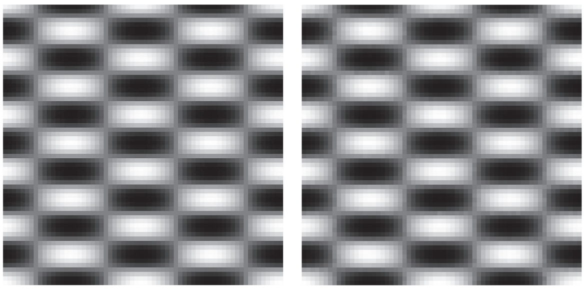

Figure 2: Cosine approximation with $K = 7$, $S = 25$, and Gaussian noise of variance $\sigma^2 = 0.5$. True atom $\boldsymbol{\kappa}$ (left) and approximation $\lambda\mathbf{h}^1 * \cdots * \mathbf{h}^K$ (right).

5 Conclusion

This work shows that simple atoms can be accurately approximated with a composition of convolutions. The kernels used in the approximation are constrained to be sparse (i.e., with sparse supports), leading to a computationally efficient algorithm, despite the non-convexity of the function to optimize. This efficiency was illustrated on a 2D-cosine function, but similar experiments conducted with archetypal kernels (e.g., wavelets or curvelets) show similar performances [2].

The proposed modeling and algorithmic schemes open new perspectives on the general problem of dictionary learning. More specifically, it seems reasonable to derive a DL technique which recovers large structured dictionary whose atoms consist of compositions of convolutions.

Finally, how to choose, set or draw the kernel supports remains a large and yet unexplored issue, that may have significant impact on the method performances.

[2] A sum of cosines of same frequency and different phases will yield a cosine of unchanged frequency.

References

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

[2] Olivier Chabiron, Francois Malgouyres, Jean-Yves Tourneret, and Nicolas Dobigeon. Toward fast transform learning. 2013. submitted to IJCV.

[3] F. Champagnat, Y. Goussard, and J. Idier. Unsupervised deconvolution of sparse spike trains using stochastic approximation. *IEEE Trans. Signal Process.*, 44(12):29882998, 1996.

[4] M. Elad. *Sparse and redundant representations: From theory to applications in signal and image processing*. Springer, 2010.

[5] K. Engan, S. O. Aase, and J. Hakon Husoy. Method of optimal directions for frame design. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, pages 2443–2446, Washington, DC, USA, 1999.

[6] G. Kail, J.-Y. Tourneret, N. Dobigeon, and F. Hlawatsch. Blind deconvolution of sparse pulse sequences under a minimum distance constraint: A partially collapsed Gibbs sampler method. *IEEE Trans. Signal Process.*, 60(6):2727–2743, june 2012.

[7] L. De Lathauwer, B. De Moor, and J. Vandewalle. On the best rank-1 and rank-(r1,r2,. . .,rn) approximation of higher-order tensors. *SIAM J. Matrix Anal. Appl.*, 21(4):1324–1342, 2000.

[8] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2):337–365, 2000.

[9] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311 – 3325, 1997.

[10] C. Quinsac, N. Dobigeon, A. Basarab, J.-Y. Tourneret, and D. Kouamé. Bayesian compressed sensing in ultrasound imaging. In *Proc. of Third International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP11)*, San Juan, Puerto Rico, 2011.