LEARNING ANALYSIS SPARSITY PRIORS

Gabriel Peyré[1], *Jalal Fadili*[2]

[1] Ceremade , CNRS-Université Paris-Dauphine
[2] GREYC, CNRS-ENSICAEN-Université de Caen
Emails: gabriel.peyre@ceremade.dauphine.fr, jalal.fadili@greyc.ensicaen.fr

ABSTRACT

This paper introduces a novel approach to learn a dictionary in a sparsity-promoting analysis-type prior. The dictionary is optimized in order to optimally restore a set of exemplars from their degraded noisy versions. Towards this goal, we cast our problem as a bilevel programming problem for which we propose a gradient descent algorithm to reach a stationary point that might be a local minimizer. When the dictionary analysis operator specializes to a convolution, our method turns out to be a way of learning generalized total variation-type prior. Applications to 1-D signal denoising are reported and potential applicability and extensions are discussed.

***Keywords*—** Dictionary learning, analysis prior, total variation, denoising.

1. INTRODUCTION

1.1. Analysis vs. Synthesis Priors

A popular approach for solving inverse problems in signal and image processing is through a variational formulation that consists in minimizing an objective functional that balances a data fidelity term against a regularization term. Such a formulation has also a nice Bayesian interpretation with the MAP estimator. Among the overwhelming number of existing priors in the literature, sparsity-promoting priors have received considerable attention over the last decade, and have been shown very effective in recovering complex structures of natural signals and images.

In general, the sparsity of the sought after signal or image manifests itself in a wisely chosen dictionary $D = (d_m)_{m=0}^{P-1}$ of P atoms in $\mathbb{R}^N$. The literature on regularization priors can be divided between analysis-based and synthesis-based priors. Let $y \in \mathbb{R}^N$ a noisy observed signal/image where the noise is assumed zero-mean additive white Gaussian. A synthesis-based prior for denoising seeks a sparse set of coefficients u that are solutions of

$$u^\star \in \underset{u \in \mathbb{R}^P}{\operatorname{argmin}} \; \frac{1}{2}\|y - Du\|^2 + \Gamma(u), \tag{1}$$

and the signal/image is then synthesized from these representation coefficients as $Du^\star$, where $\Gamma(u)$ is a sparsity-promoting proper lower-semicontinuous penalty function. A popular choice of Γ is the now celebrated convex ℓ^1 norm $\Gamma(u) = \sum_m |u_m|$, in which case (1) is coined basis pursuit denoising [3]. Sparsity synthesis priors in redundant dictionaries such as translation invariant wavelets have been popular to perform denoising, and also to solve more complicated inverse problems; see for instance [12].

An analysis-based prior seeks a signal or image x whose forward (analysis) transform coefficients are sparse. This corresponds to the minimization problem

$$\min_{x \in \mathbb{R}^N} \frac{1}{2}\|y - x\|^2 + \Gamma(D^* x) \; . \tag{2}$$

One can think of choosing D^* as the analysis operator of a redundant dictionary such as a translation invariant wavelet frame. A more frequent use of analysis priors is though via a differential operator D^* to enforce some regularity on the signal, while still allowing to restore discontinuities. The action of D^* can also be interpreted as a convolution with some finite impulse response filter. This leads for instance to a discrete version of the popular total variation (TV) prior introduced by Rudin Osher and Fatemi [14].

The work of [5] was the first to give some insights into the connections between analysis and synthesis-based priors. For instance, if the dictionary D is square and invertible (bijective), then the class of problems (1) and (2) are equivalent. More precisely, a solution to the analysis prior problem with the forward operator D^{-1} is uniquely recovered from that of the equivalent synthesis prior problem with the dictionary D. In our phrasing of (1)-(2), this amounts to saying that $D^{-1} = D^*$, i.e. D is orthonormal. In the general case this equivalence does not hold as is the case for redundant dictionaries or the TV prior.

1.2. Dictionary Learning

Starting from the seminal work of Olshausen and Field [13], several methods have attempted to learn a dictionary D from a set of exemplars. All previous approaches have focused on the synthesis-type prior (1) where the objective functional is now jointly minimized in the coefficients and the dictionary D –with a norm constraint on its columns to avoid the classical scale indeterminacy– so that the latter allows a sparse representation of all exemplars; see for instance [10, 9, 1, 6] to name a few. The objective functional (1) is however non-convex even if Γ is

This work is supported by ANR grant NatImages, ANR-08-EMER-009.

convex, and different methods have been proposed to compute a stationary point, which may happen to be a local minimizer under appropriate circumstances. Synthesis prior-based learned dictionaries have been successfully applied to denoising [1] and more generally to inverse problems such as inpainting [11].

1.3. Task-driven Dictionary Learning

A recent line of research has promoted a new framework for dictionary learning driven by the task one is ultimately interested in achieving on the signal [7, 8]. In a nutshell, the sparsity-regularized optimization problem (1) is used to solve a task problem such as denoising, super-resolution or discrimination, and the dictionary is learned in order to achieve the best performance for that task on a set of training samples. This leads to a bilevel programming problem [4], that can be solved for D using gradient descent-type algorithms.

1.4. Contributions

This paper introduces a novel approach to learn a dictionary in a sparsity-promoting analysis-type prior. To the best of our knowledge, this is the first work in this direction. It performs a task-driven learning by solving a bilevel programming problem. The method is specialized to learn a convolution dictionary. We report numerical examples to support the usefulness of our approach.

2. ANALYSIS PRIOR SOLUTION SENSITIVITY

2.1. Smoothed Sparsity Regularization

At first glance, it is tempting to use the ℓ^1 sparsity penalty for Γ. However, such a choice will entail important difficulties in the bilevel programming problem, mainly because of the non-smoothness of the ℓ^1-norm in the dictionary (and even less for its sub-gradients). Much more complicated tools should be used in this case which can be borrowed from variational and set-valued analysis theory.

To alleviate these difficulties, we propose in the present work a smooth continuously differentiable version of the ℓ_1 penalty, that we define as

$$\forall\, u \in \mathbb{R}^P, \quad \Gamma(u) = \sum_{m=0}^{P-1} \sqrt{|u_m|^2 + \varepsilon^2}, \tag{3}$$

for a small smoothing parameter $\varepsilon > 0$. This turns also to be a strictly convex functional. Such a smoothing is quite familiar in the variational image processing community, and was also used for synthesis-prior sparse coding in [2].

It is important to note that our setting is quite different in several aspects with regard to the synthesis prior learning with ℓ^1 penalty. In the synthesis formulation (1), the functional is smooth with respect to the dictionary, but is not strictly convex with respect to the coefficients, so that the minimizer in the coefficients is not unique. As for smoothness, uniqueness is also important in sensitivity (perturbation) analysis of the regularized minimizer, to avoid treating the minimizer as a set-valued map. To mitigate this shortcoming, the authors of [8] replaced the ℓ^1 norm by the elastic net regularization that combines the ℓ^1 norm and the squared ℓ^2 norm, hence strictly (in fact strongly) convexifying the problem in the coefficients.

Let's rewrite the analysis regularization problem (2) with Γ as given by (3)

$$x(D, y) = \underset{x \in \mathbb{R}^N}{\operatorname{argmin}}\ \frac{1}{2}\|y - x\|^2 + \Gamma(D^* x)\ . \tag{4}$$

This is a strongly (hence strictly) convex problem and $x(D, y)$ is uniquely determined. In the sequel, we denote by $\nabla_\Gamma[u] \in \mathbb{R}^P$ the gradient mapping of Γ at $u \in \mathbb{R}^P$, and $\mathrm{H}_\Gamma[u] : \mathbb{R}^P \to \mathbb{R}^P$ its Hessian operator at u. In the case of (3), the gradient and the Hessian read

$$\nabla_\Gamma[u]_m = \frac{u_m}{\sqrt{\varepsilon^2 + |u_m|^2}},\ \mathrm{H}_\Gamma[u] = \operatorname{diag}\left(\frac{\varepsilon^2}{(\varepsilon^2 + |u_m|^2)^{3/2}}\right)_m .$$

2.2. Sensitivity of the Regularized Solution

Optimizing the dictionary D in a task-driven framework, as we will explain in Section 3, requires to characterize the mapping $D \mapsto x(D, y)$, and in particular its variations (also known as sensitivity or perturbation analysis in optimization theory) with respect to the dictionary D. The following proposition shows that this mapping is continuously differentiable and gives the formula of its transpose derivative operator. Since y is fixed in this section, we lighten the notation by writing $x(D) = x(D, y)$.

Theorem 1. *The mapping $D \mapsto x(D)$ is of class C^1 at $D \in \mathbb{R}^{N \times P}$, and its derivative at D satisfies for all $z \in \mathbb{R}^N$*

$$dx[D]^*(z) = -\bar{z} \times \nabla_\Gamma[u]^{\mathrm{T}} - x(D) \times (\mathrm{H}_\Gamma[u](D^* \bar{z}))^{\mathrm{T}}\ , \tag{5}$$

for $u = D^* x(D)$, $\bar{z} = \Delta^{-1} z$ and where

$$\Delta = \mathrm{Id} + D\mathrm{H}_\Gamma[u]D^* : \mathbb{R}^N \to \mathbb{R}^N\ ,$$

and Id *is the identity operator on $\mathbb{R}^N$.*

Proof. The first order sufficient and necessary optimality condition of (4) reads

$$x(D) - y + D\nabla_\Gamma[D^* x(D)] = 0. \tag{6}$$

Equation (6) defines implicitly the mapping $D \mapsto x(D)$. Its derivative and smoothness are obtained from the implicit function theorem applied to the mapping $S(x, D) = x - y + D\nabla_\Gamma[D^* x]$. Indeed, differentiating (6) with respect to D in the direction $\delta \in \mathbb{R}^{N \times P}$ gives

$$\begin{aligned} &dx[D](\delta) + \delta\nabla_\Gamma[D^* x(D)] + \\ &\qquad D\mathrm{H}_\Gamma[D^* x(D)](\delta^* x(D) + D^* dx[D](\delta)) = 0, \end{aligned} \tag{7}$$

Using convexity of Γ, Δ as defined by (5) is positive definite, hence invertible. Thus, the implicit function theorem allows to conclude that $x(D)$ is C^1, and the derivative is obtained by inverting (7)

$$dx[D](\delta) = -\Delta^{-1}\left(\delta\nabla_\Gamma[D^* x(D)] + D\mathrm{H}_\Gamma[D^* x(D)]\delta^* x(D)\right).$$

Transposing this equation leads to (5). □

3. UNSTRUCTURED DICTIONARY LEARNING WITH ANALYSIS PRIOR

3.1. Bilevel Programming for Analysis Dictionary Learning

In a dictionary learning framework, the object of interest is the dictionary $D = (d_m)_{m=0}^{P-1}$. In most previous works, a set of noisy exemplars $\{x_k\}_k$ is given and the dictionary is solved for so as to maximize the sparsity of the representation coefficients of the exemplars. In a task-driven framework for denoising, as explained in Section 1.3, the learned dictionary is optimized to achieve the best possible performance in the task one is targeting, for instance denoising. This is achieved from a set of pairs (y_k, x_k) where $x_k \in \mathbb{R}^N$ is a clean exemplar, $y_k = x_k + w_k$ its corresponding noisy version, and w_k is an additive noise, implicitly assumed to be white Gaussian in view of (4). The dictionary is then obtained by minimizing the empirical denoising risk

$$\min_{D\in\mathbb{R}^{N\times P}} \mathcal{E}(D) = \frac{1}{2}\sum_k \|x_k - x(D, y_k)\|^2. \tag{8}$$

This optimization problem together with (4) form a hierarchical mathematical programming problem known as bilevel programming [4].

3.2. Analysis Prior Dictionary Learning Algorithm

According to Theorem 1, $\mathcal{E}$ is a smooth functional. A local minimizer (or a stationary point) of (8) can then be found using a gradient descent

$$D^{(t+1)} = D^{(t)} - \eta_t \nabla\mathcal{E}(D^{(t)}), \tag{9}$$

where $0 < \eta_t < \eta$ is a small enough sequence of step-sizes. The gradient of the energy reads

$$\nabla\mathcal{E}(D) = \sum_k dx[D, y_k]^*(x(D, y_k) - x_k)\ , \tag{10}$$

where $dx(D, y_k) : \mathbb{R}^{P\times N} \to \mathbb{R}^N$ is the differential of the mapping $D \mapsto x(D, y_k)$ at D as exhibited in the proof of Theorem 1, and $dx[D, y_k]^*$ its adjoint operator, computed from (5).

4. CONVOLUTION DICTIONARY LEARNING WITH ANALYSIS PRIOR

A popular family of priors is obtained when specializing D to be translation invariant and defined from a single atom $\gamma \in \mathbb{R}^N$, so that $d_i = \gamma(\cdot - i)$ where we use periodic boundary conditions for simplicity. This means that the dictionary prior is defined as a circular convolution with γ, since $D^* x = \gamma \star x$. Such a prior is reminiscent of TV regularization.

We denote $\varphi(\gamma) \in \mathbb{R}^{N\times N}$ the circular convolution operator defined by γ, and $\varphi(\gamma)^*$ its adjoint associated with $\widetilde{\gamma}$, where for any vector x, $\widetilde{x}$ is its reversed version, i.e. $\widetilde{x}_i = x_{-i}$. The convolution kernel γ is learned by solving again a bilevel program with (8) and (4) specialized to a dictionary solely parametrized by γ,

$$\min_{\gamma\in\mathbb{R}^N} \bar{\mathcal{E}}(\gamma) = \frac{1}{2}\sum_k \|x_k - \bar{x}(\gamma, y_k)\|^2,$$

where $\bar{x}(\gamma, y) = x(\varphi(\gamma), y)$. This energy is minimized using a gradient descent as in (9). The following proposition gives the expression of the adjoint derivative of $\bar{x}(\gamma)$ involving only convolutions (dependence on y was dropped to lighten notation).

Proposition 1. *For any $z \in \mathbb{R}^N$, the adjoint derivative of $\bar{x}(\gamma)$ is*

$$d\bar{x}[\gamma]^*(z) = -\bar{z}\star\widetilde{\nabla_\Gamma[u]} - \bar{x}(\gamma)\star(\mathrm{H}_\Gamma\widetilde{[u](\gamma\star\bar{z})}), \tag{11}$$

where $u = \gamma\star\bar{x}(\gamma)$, $\bar{z} = \Delta^{-1}z$ and where we have written

$$\Delta = \mathrm{Id} + \varphi(\gamma)^*\mathrm{H}_\Gamma[u]\varphi(\gamma) : \mathbb{R}^N \to \mathbb{R}^N\ .$$

Proof. Using the chain rule, we have

$$d\bar{x}[\gamma]^*(z) = d\varphi[\gamma]^*\left(dx[\varphi(\gamma)]^*(z)\right). \tag{12}$$

For 1-D convolutions, it can be shown that the adjoint derivative $d\varphi[\gamma]^*$ corresponds to summing along the diagonals of a matrix $A \in \mathbb{R}^{N\times N}$,

$$d\varphi[\gamma]^*(A) = \alpha \quad \text{where} \quad \alpha_i = \sum_{s-t=i} A_{s,t}\ , \tag{13}$$

and equality of indices should be understood modulo N. This expression extends to higher dimensional convolution. It is not difficult to verify that if $A = uv^{\mathrm{T}}$, meaning in particular that $A_{i,j} = u_i v_j$, then $d\varphi[\gamma]^*(A) = u\star\widetilde{v}$. Thus, piecing together (5) (with $\varphi(\gamma)$ in lieu of D^*), (12) and (13) yields the expression of the adjoint derivative (11), where (13) allows to simplify (11) by involving only convolutions. □

The computation of $d\bar{x}[\gamma]^*(z)$ from (11) requires only four convolutions, and the resolution of the linear system $\Delta\bar{z} = z$. This system is solved efficiently with a few conjugate gradient iterations capitalizing on the special structure of the operator $\varphi(\gamma)$ and its adjoint $\varphi(\gamma)^*$.

5. NUMERICAL EXPERIMENTS

In the first experiment, we consider 1-D binary piecewise constant signals $x_k = 1_{[a_k, b_k]}$ of size $n = 128$ where $0 \leqslant a_k < b_k < n$ are distinct discretized uniform random locations. In this setting the discrete TV regularization, computed with the FIR filter $\gamma_{\mathrm{TV},\lambda}(0) = \lambda$, $\gamma_{\mathrm{TV},\lambda}(1) = -\lambda$ and 0 otherwise, is used as a gold-standard since it is known to recover almost perfectly piecewise constant signals. In the following, λ is chosen in order to minimize $\bar{\mathcal{E}}(\gamma_{\text{TV},\lambda}) = \mathcal{E}(D_{\text{TV},\lambda})$. The unstructured dictionary learning scheme is performed with an initial dictionary $D^{(0)}$ which is a random white noise, and $\varepsilon = 10^{-3}$. Figure 1, left, shows that the denoising energy $\mathcal{E}(D^{(t)})$ converges to that of the TV dictionary $\mathcal{E}(D_{\text{TV},\lambda})$. Though the convergence is rather slow given the poor initialization and the small value of ε. This shows the ability of the method to recover the TV filter from an arbitrary initialization.

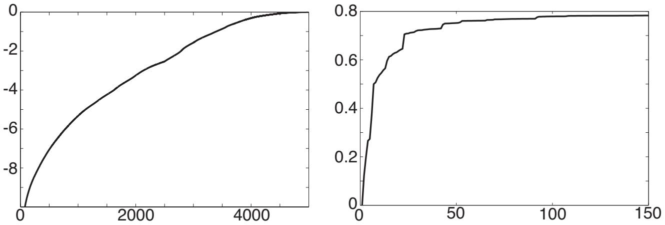

Fig. 1. *Evolution of the denoising energy $-10\log_{10}(\mathcal{E}(D^{(t)})/\mathcal{E}(D_{\text{TV},\lambda}))$ with the number of iterations t. Left: unstructured dictionary, step exemplar signals. Right: convolution dictionary, natural signals.*

In the second experiment, a convolution analysis prior is learned on a set of 1D natural signals of size $n = 256$ with $\varepsilon = 10^{-2}$. Each signal $x_k(i) = f(a_k, i)$ is extracted uniformly randomly as a row a_k of the "lena" image f of size 256^2, see Figure 2, left. We initialize $\gamma^{(0)} = \gamma_{\text{TV},\lambda}$ as the TV filter with the optimal value of λ. Figure 1, right, shows how the learning improves the denoising performance up to 0.8dB on the training data. Note now how dictionary structuring together with good initialization speed up convergence. Figure 2, right, displays the optimal filter computed with our method. We then applied this learned filter to denoise another set of natural signals, obtained from the rows of the "Boat" image. This yielded an average denoising improvement of 0.3dB with respect to the TV filter. This is less that for Lena which can be explained by the fact that the "Boat" has stronger edges and is less oscillatory, which makes the learned filter somewhat less adapted.

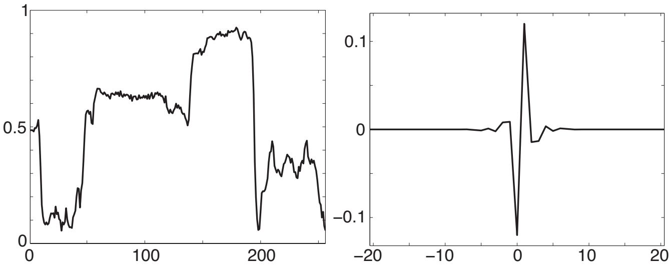

Fig. 2. *Left: example of 1-D natural signal x_k used for the learning. Right: optimal filter $\gamma^{(\infty)}$ learned by our method.*

6. CONCLUSION AND PERSPECTIVES

This paper has introduced a novel approach for dictionary learning of a sparsity-based analysis prior. Although we only reported some preliminary experiments to exemplify potential applications to 1-D signal denoising, we believe our method has very promising potentials. Some future investigations which will certainly be our next milestones include:

- Extend the framework to higher dimensions starting with 2-D image processing while incorporating invariances, such as translation invariance through 2-D convolutions, and exemplars extracted from image patches.
- Extend to multiple kernels learning to better capture complicated regularity patterns while preserving translation invariance.
- Extend to other tasks such as linear inverse problems regularization and classification.

7. REFERENCES

[1] M. Aharon, M. Elad, and A.M. Bruckstein. The k-svd: An algorithm for designing of overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Proc.*, 54(11):4311–4322, 2006.

[2] J. A. Bagnell and D. M. Bradley. Differentiable sparse coding. In Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Léon Bottou, editors, *NIPS*, pages 113–120. MIT Press, 2008.

[3] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20:33–61, December 1998.

[4] B. Colson, P. Marcotte, and G. Savard. An overview of bilevel optimization. *Annals of Operations Research*, 153(1):235–256, 2007.

[5] M. Elad, P. Milanfar, and R. Rubinstein. Analysis versus synthesis in signal priors. *Inverse Problems*, 23(3):947–968, 2007.

[6] K. Engan, S. O. Aase, and J. Hakon Husoy. Method of optimal directions for frame design. In *Proc. ICASSP '99*, pages 2443–2446, Washington, DC, USA, 1999. IEEE Computer Society.

[7] L. Horesh and E. Haber. Sensitivity computation of the ℓ^1 minimization problem and its application to dictionary design. *Inverse Problems*, 25:025002, 2009.

[8] F. Bach J. Mairal and J. Ponce. Task-driven dictionary learning. *Preprint arXiv:1009.5358v1*, 2010.

[9] 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 Comput.*, 15(2):349–396, 2003.

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

[11] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Trans. Image Proc.*, 17(1):53–69, January 2008.

[12] S. Mallat. *A Wavelet Tour of Signal Processing, 3^{rd} edition*. Elsevier, 2009.

[13] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive-field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, June 1996.

[14] L.I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60(1–4):259–268, 1992.