Smooth Optimization of Orthogonal Wavelet Basis

Jordan Frecon[1], Riccardo Grazzi[2], Saverio Salzo[2], and Massimiliano Pontil[2,3]

[1]Normandie Univ, INSA Rouen, UNIROUEN, UNIHAVRE, LITIS.
Saint-Etienne-du-Rouvray, France
[2]Computational Statistics and Machine Learning, Istituto Italiano di Tecnologia, Italy
[3]Department of Computer Science, University College London, United Kingdom

Abstract

Wavelets are a powerful tool for signal and image processing tasks. They allow to analyze the noise level separately at multiple scales and to adapt the denoising algorithm accordingly. However, the performance strongly rely on the choice of the wavelet basis. The aim of this work is to learn the wavelet basis that is adapted to both the denoising task and the class of images at hand.We tackle this problem by a smooth bilevel approach where the wavelet coefficients are optimized at the lower-level and the wavelet filters are learned at the upper-level. Numerical experiments support the added benefits over classical wavelets.
Keywords: Bilevel optimization, wavelet denoising.

1 Introduction

Over the past decades, the discrete wavelet transform has been widely used in numerous applications in signal and image processing ranging from denoising to deblurring [DJ94, PBBZP16, Mal08]. Wavelets bases provide a mixed spatial-scale representation in which non stationary processes can be represented by few coefficients. This has made wavelets very suitable for compressing and denoising natural images. Henceforth, an active area of research has emerged to design and to find the optimal wavelet bases for compacting purposes.

Related works: Typical methods to design wavelets mainly focus upon the wavelet function and its properties such as regularity, smoothness and symmetry. In order to choose the wavelet adapted to some class of images parametric models have been proposed such as in [Thi04] where a lifting method is devised to create a discrete wavelet similar to a given pattern. In the same line of work, [HRCD07] provides a way to determine the lifting parameters that yield the sparsest wavelet coefficients according to the Gini Index. We also point out the works of [LDHD02, NKA$^+$06] which optimize the wavelet for classification purposes. More recently, non parametric models have been devised. In [Søg17, RM18b, RM18a], the discrete wavelet transform is framed as a modified convolutional neural network and the wavelet filters are optimized by backpropagation to promote sparsity under some admissibility constraints. The optimization of the wavelet filters has been recently addressed for image compression and denoising in [GP18]. The authors considered the lasso model in the wavelet domain, also known as wavelet soft-thresholding, to either reduce the number of wavelet coefficients or to denoise the observation, depending on the application. A minimization scheme is then proposed to alternatively estimate the thresholded wavelet coefficients and the wavelet basis which minimize the lasso objective.

Contributions and outline: The main contribution of the paper is to frame the learning of orthogonal wavelet basis as a bilevel optimization problem. Given some denoising task, the wavelet coefficients are learned at the lower-level while the wavelet basis is optimized at the upper-level by judging upon the estimated wavelet coefficients. In addition, we make two advances. The first is in terms of application since that, contrary to [GP18], here the observations are corrupted by a linear degradation. The second is algorithmic as the bilevel framework of [FSP18, ORBP16] is extended by replacing the lower-level problem by a smooth primal forward-backward algorithm with Bregman distances. After some preliminaries reported in Section 2, we present in Section 3 the proposed bilevel optimization problem. The smooth algorithmic solution is then devised in Section 4. Deblurring experiments are conducted in Section 5 on a benchmark image.

2 Wavelet denoising

In this section, we recall the basics of the discrete wavelet transform and present the most popular variational problem for denoising in the wavelet domain.

2.1 Discrete wavelet transform

The principle of wavelet transform lies in multiresolution analysis where any given observation $y \in \mathbb{R}^N$ is viewed through the scope at different resolutions $\{2^{-j}\}_{j=1\dots J}$. Given a filter $h \in \mathbb{R}^K$, the wavelet transform of x, denoted by $W(h)x$, is the collection of details coefficients d_j and the coarse approximation a_J, i.e.,

$$W(h)x = [d_1(h), \dots, d_J(h), a_J(h)]. \tag{1}$$

These coefficients can be computed hierarchically across the scales by successively applying low-pass and high-pass filters [Mal08, Theorem 7.10]

$$(\forall p \in [N], \forall j \in [J]),\ \begin{cases} a_{j+1,p}(h) = a_j(h) * \bar{h}_{2p}, \\ d_{j+1,p}(h) = a_j(h) * \bar{g}_{2p}(h), \end{cases} \tag{2}$$

where we have introduced the notation $\bar{x}_p = x_{K+1-p}$. This can be seen as a filter bank decomposition where the approximation coefficient $a_{j+1}(h)$ (resp. details coefficients $d_{j+1}(h)$) can be computed by applying a low-pass filter $\bar{h}$ (resp. high-pass filter $\bar{g}(h)$) to $a_j(h)$ followed by a down-sampling step. We now recall the conditions on h and $g(h)$ in order to yield an admissible wavelet representation with perfect reconstruction. In particular, we restrict our study to the case where $g(h)$ is the quadrature mirror filter of h, i.e., $(\forall k \in \{1, \dots, K\})$, $g_k(h) = (-1)^k h_{K-k-1}$ [CEG76, Mal08].

Definition 2.1 (Admissible set of wavelet filters)**.** Let $K \in \mathbb{N}_+$ be the width of the wavelet support. Then, the set of admissible wavelet filters is $\mathcal{C} = \mathcal{C}_1 \cap \mathcal{C}_2$ where $\mathcal{C}_1 = \left\{h \in \mathbb{R}^K \;\middle|\; \mathbf{1}_K^\top h = \sqrt{2},\ \mathbf{1}_K^\top g(h) = 0\right\}$ is a linear constraint which includes normalization and smoothness conditions while

$$\mathcal{C}_2 = \left\{h \in \mathbb{R}^K \;\middle|\; (\forall p \in \{0, \dots, \frac{K}{2}\}), \frac{1}{2} h^\top O_{2p} h = \delta_{0,2p}\right\}$$

encodes the orthogonality conditions through the matrices $O_{2p} \in \mathbb{R}^{K \times K}$ where $(O_{2p})_{i,j} = 2$ if $j = 2p + i$ and 0 otherwise.

2.2 Denoising in the wavelet domain

In this section, we consider the inverse problem of reconstructing an unknown original image $\bar{y} \in \mathbb{R}^N$ from an observation corrupted by linear and additive degradations. We model the noisy image by $y = A\bar{y} + \varepsilon$ where $A \in \mathbb{R}^{N \times N}$ can be blurring operator and $\varepsilon \in \mathbb{R}^N$ is a realization of white Gaussian noise.

A simple but efficient non-linear denoising estimator can be obtained by regularizing the coefficients of the noisy image in a well chosen basis. Here, we focus on an orthogonal wavelet basis, which is efficient to denoise piecewise regular images. We consider the problem of finding $\hat{x}(h)$ such that [BT09, PBBZP16]

$$\hat{x}(h) = \operatorname*{argmin}_{x \in \mathbb{R}^N} \frac{1}{2}\|y - AW(h)^{-1}x\|^2 + \lambda\|x\|_1, \tag{3}$$

where $\lambda > 0$ is a regularization parameter acting as a tradeoff between the amount of reconstruction and the sparsity of the wavelet coefficients.

3 Bilevel framework

The main goal of this paper is to investigate the benefits of a bilevel scheme to additionally optimize $\hat{x}(h)$ over the wavelet filter h. In order to learn an admissible wavelet filter $h \in \mathcal{C}$ adapted to a class of images, we propose to consider the following bilevel problem.

Problem 3.1. *Given T noisy images $y_t \in \mathbb{R}^N$, $t = 1 \dots, T$, a function $E\colon \mathbb{R}^N \to \mathbb{R}$ and a regularization parameter $\lambda > 0$, solve*

$$\operatorname*{minimize}_{h \in \mathcal{C}} \left\{\mathcal{U}(h) := \frac{1}{T}\sum_{t=1}^{T} E(\hat{x}_t(h))\right\}, \tag{4}$$

where $\mathcal{C}$ is given in Definition 2.1 and, for every $t \in \{1, \dots, T\}$, the denoised wavelet coefficients $\hat{x}_t(h)$ are are given by

$$\hat{x}_t(h) \in \operatorname*{argmin}_{x \in \mathbb{R}^N} \frac{1}{2}\|y_t - AW(h)^{-1}x\|^2 + \lambda\|x\|_1. \tag{5}$$

For instance, for $E = \|\cdot\|_1$, the upper objective $\mathcal{U}$ in (4) promotes wavelet filter h for which the denoised coefficients, computed at the lower-level (5), are the sparsest. Since we do not have a closed form expression for $\hat{x}_t(h)$ but we rather have an iterative scheme converging to $\hat{x}_t(h)$, we embrace the idea proposed in [FSP18] and approximate $\hat{\boldsymbol{x}}(h) = (\hat{x}_1(h), \dots, \hat{x}_T(h))$ by the Q-th iterate $\boldsymbol{x}^{(Q)}$. Here we propose to resort to a smooth primal algorithm. The associated bilevel problem that we are actually solving reads:

Problem 3.2. *Given a mapping $\mathcal{M}$, the set of constraints $\mathcal{C}$ of Definition 2.1 as well as a maximum number of inner iterations $Q \in \mathbb{N}$, solve*

$$\underset{h\in\mathcal{C}}{\text{minimize}} \left\{ \mathcal{U}_Q(h) := \frac{1}{T}\sum_{t=1}^{T} E(x_t^{(Q)}(h)) \right\}, \quad \text{where} \begin{cases} \boldsymbol{x}^{(0)}(\theta) \text{ is chosen arbitrarily} \\ \text{for } q = 0, 1, \ldots, Q-1 \\ \quad \boldsymbol{x}^{(q+1)}(\theta) = \mathcal{M}(\boldsymbol{x}^{(q)}(\theta), \theta) \end{cases} \tag{6}$$

4 Algorithmic solution

In this section, we tackle the solving of Problem 3.2. The properties of the objective function $\mathcal{U}_Q$ depend on the choice of algorithm (6) minimizing the lower-level objective in Problem 3.1. However, since the latter is nonsmooth, the mapping $\mathcal{M}$ is usually nonsmooth as well. In that case, computing a subgradient of $\mathcal{U}_Q$ is a challenge since it is the composition of multiple nonsmooth functions. Designing a smooth algorithm is the cornerstone of the proposed method and is presented in Section 4.2. The corresponding proposed bilevel optimization scheme is the subject of the next section.

4.1 Upper-level solver

In order to solve Problem 3.2, we propose to resort to the following subgradient descent algorithm

$$(\forall n \in \{1, \ldots, n_{\max}\}),\ h^{(n+1)} = \mathcal{P}_{\mathcal{C}}\left(h^{(n)} - \gamma_n s_n\right), \tag{7}$$

where $\mathcal{P}_{\mathcal{C}}$ denotes the projection onto the set $\mathcal{C}$ in Definition 2.1, $(\gamma_n)_n$ is a decreasing sequence of step-sizes and $s_n \in \partial\mathcal{U}_Q(h^{(n)})$. On one hand, when $\boldsymbol{x}^{(Q)}(h)$ is smooth, s_n can be computed by using the basic chain rule. On the other hand, since $\mathcal{C}$ is nonconvex, any algorithm for projecting onto $\mathcal{C}$ might lead to a sub-optimal solution. While a computationally intense projection algorithm has been devised in [GP18], here we suggest the following alternating projection algorithm $h \leftarrow \mathcal{P}_{\mathcal{C}_1}(\tilde{\mathcal{P}}_{\mathcal{C}_2}(h))$ [DL18], where $\tilde{\mathcal{P}}_{\mathcal{C}_2}$ is an inexact projection onto $\mathcal{C}_2$ obtained by linearizing the quadratic constraints, i.e.,

$$\tilde{\mathcal{P}}_{\mathcal{C}_2}(h) = \underset{v\in\mathbb{R}^K}{\operatorname{argmin}} \frac{1}{2}\|h - v\|^2$$
$$\text{s.t. } \forall p \in \{0, \ldots, K/2\},\ \frac{1}{2}h^\top O_{2p} h + O_{2p}(v - h) = \delta_{0,2p}.$$

4.2 Smooth lower-level solver

We now turn to solving (5) by a smooth algorithm. Without loss of generality, we can deal with a single observation omitting the index t. By splitting the variable x into positive and negative parts, $x = x^+ - x^-$ with $x^+ \geq 0$ and $x^- \geq 0$, the ℓ_1 regularization term can be expressed as $\mathbb{1}^\top\lambda(x^+ + \mathbb{1}^\top x^-)$ which is linear in x^+ and x^-. Therefore, (5) can be rewritten as the following bound-constrained quadratic program [FNW07]

$$\underset{\mathrm{x}\in\mathbb{R}^{2N}_{\geq 0}}{\text{minimize}} \underbrace{c(h)^\top \mathrm{x} + \frac{1}{2}\mathrm{x}^\top B(h)\mathrm{x}}_{f(\mathrm{x},h)}, \tag{8}$$

where $\mathrm{x} = [x^+; x^-]$, $c(h) = \lambda\mathbb{1} + [-H(h)^\top y; H(h)^\top y]$, $H(h) = AW(h)^{-1}$ and

$$B(h) = \begin{bmatrix} H(h)^\top H(h) & -H(h)^\top H(h) \\ -H(h)^\top H(h) & H(h)^\top H(h) \end{bmatrix}. \tag{9}$$

While (8) can be solved by gradient projection methods, they involve a projection onto the non-negative orthant at each interation, thus making the algorithm nonsmooth. We point out that these methods can be seen as particular instances of the forward-backward algorithm with Bregman distances [BBT16, VN17] where the Legendre function Φ defining the Bregman distance is the squared norm $\frac{1}{2}\|\cdot\|^2$. In order to find a smooth algorithm, we consider an interior point variant where Φ is ϵ-strongly convex and $\overline{\operatorname{dom}\Phi} = \mathbb{R}^{2N}_{\geq 0}$. By elaborating on [BBT16, VN17], we consider the following algorithm

$$\begin{aligned} &\mathrm{x}^{(0)}(h) \in \operatorname{int}\operatorname{dom}\Phi \\ &\text{for } q = 0, 1, \ldots, Q-1 \\ &\quad \mathrm{x}^{(q+1)}(h) = \nabla\Phi^*\left(\nabla\Phi(\mathrm{x}^{(q)}(h)) - \gamma\nabla f(\mathrm{x}^{(q)}(h), h)\right) \\ &\boldsymbol{x}^{(Q)}(h) = \mathrm{x}^{(Q)}_{1:N}(h) - \mathrm{x}^{(Q)}_{N+1:2N}(h) \end{aligned}$$

which converges for any step-size $0 < \gamma < \epsilon/\|B\|_2$. Note that since $W(h)$ is orthonormal, we have that $\|B\|_2 = 2\|A^\top A\|_2$. We propose to resort to the Burg's entropy $\Phi\colon \mathrm{x} \mapsto \sum_{n=1}^{2N}\phi(\mathrm{x}_n)$ where $\phi\colon t \mapsto -\log t + (\epsilon/2)t^2$, for which $(\forall t \in \mathbb{R})$, $\phi'(t) = (\epsilon t^2 - 1)/t$ and $\phi^{*\prime}(t) = (t + \sqrt{t^2 + 4\epsilon})/2\epsilon$. Note that for such choice $\operatorname{dom}\Phi = \mathbb{R}^{2N}_{>0}$. Thus the resulting algorithm only yields an approximate solution of (5). We suggest to initialize $\mathrm{x}^{(0)} = 1/\sqrt{\epsilon}\mathbb{1}_{2N}$ so that to enforce $\nabla\Phi(\mathrm{x}^{(0)}) = 0_{2N}$.

Figure 1: From left to right: original, noisy and reconstructed images by the proposed method, db16, sym16 and coif6.

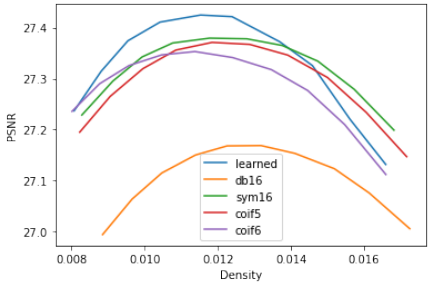

Figure 2: Comparison with standard wavelets filters.

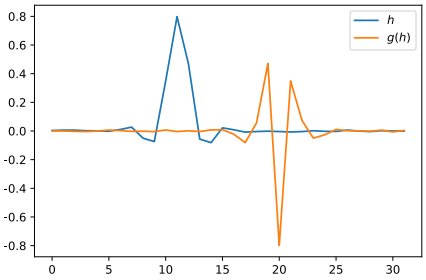

Figure 3: Learned low-pass and high-pass filters.

5 Experiments

Experimental setting We consider a bitmap gray version of the Lena image[1], denoted by $\bar{y}$ and made of $N = 512 \times 512$ pixels as well as its noisy version $y = A\bar{y} + \epsilon$ where ϵ is a realization of Gaussian noise of standard deviation 0.05 and A denotes a Gaussian blur of width 15 and standard deviation 3. Both the original and noisy image are displayed in Fig.1. The wavelet length is set to $K = 32$ and the wavelet filter is initialized from some Gaussian noise $\varepsilon \sim \mathcal{N}(0_K, 1)$ by projecting it into $\mathcal{C}$, i.e., $h^{(0)} = \mathcal{P}_{\mathcal{C}}(\varepsilon)$. We considered 5 different realizations of random noise and picked the one which minimizes the upper-level objective in Problem 3.2. The proposed method is compared against the following classical wavelets: Daubechies (db16), symlet (sym16) and coiflets (coif5 and coif6). We choose $J = 4$ decomposition layers and we ran the experiments for 10 values of λ equally spaced, in a $\log_{10}$-scale, between 10^{-2} and 10^{-1}.

Quantification of performance. The benefits of using a particular wavelet basis is quantified by two measures: the peak signal-to-noise ratio of the reconstructed image (PSNR) and the density $\|\cdot\|_0/N$ of the wavelet coefficients. As λ varies, we report the corresponding performance in Figure 2. Overall, it shows that the learned wavelet permits to achieve a higher PSNR with fewer wavelet coefficients than the classical wavelets. The closest is coif6 a lower density is attained at the price of a small decrease in PSNR. Since the PSNR does not capture how well edges are recovered, we also report in Fig. 1 the reconstructed images for deblurring. A visual inspection shows that the proposed method better reconstructs the edges, notably around the mouth, nose and the hat.
In addition, we report in Fig. 3 the learned wavelet filters. We remark that they strongly look like those of the coiflet while we learned them from random noise.

[1] `http://eeweb.poly.edu/~yao/EL5123/image/lena_gray.bmp`

6 Conclusion

In this work, we have proposed a bilevel framework for learning orthogonal wavelets adapted to some given task. While we have considered image deblurring and reconstruction, the proposed method could also be extended to other purposes provided that a smooth algorithm can be devised. In addition, although the corresponding bilevel optimization problem is nonconvex, we are able learn wavelets from random noise which perform equally well to classical wavelets resulting from years of mathematical modeling. Perspectives include the extension to biorthogonal and anisotropic wavelets. More interestingly, the proposed framework paves the way to smooth deep unfolding techniques where the dictionary $W(h)$ could vary at each iteration.

References

[BBT16] H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2016.

[BT09] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[CEG76] A. Croisier, D. Esteban, and C. Galand. Perfect channel splitting by use of interpolation/decimation/tree decomposition techniques. *Proc. Int. Symp. on Info., Circuits and Systems, (Patras, Greece)*, pages 443–446, August 1976.

[DJ94] D. L. Donoho and J. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81(3):425–455, 1994.

[DL18] D. Drusvyatskiy and A. S. Lewis. Inexact alternating projections on nonconvex sets. *arXiv preprint arXiv:1811.01298*, 2018.

[FNW07] M. Figueiredo, R. D. Nowak, and S. J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of selected topics in signal processing*, 1(4):586–597, 2007.

[FSP18] J. Frecon, S. Salzo, and M. Pontil. Bilevel learning of the group lasso structure. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8301–8311. Curran Associates, Inc., 2018.

[GP18] T. Grandits and T. Pock. Optimizing wavelet bases for sparser representations. In M. Pelillo and E. Hancock, editors, *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 249–262, Cham, 2018. Springer International Publishing.

[HRCD07] N. Hurley, S. Rickard, P. Curran, and K. Drakakis. Maximizing sparsity of wavelet representations via parameterized lifting. In *15th International Conference on Digital Signal Processing*, pages 631–634. IEEE, 2007.

[LDHD02] M. Lucas, C. Doncarli, E. Hitti, and N. Dechamps. Wavelet optimization for classification. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages II–1249–II–1252, May 2002.

[Mal08] S. Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, Inc., Orlando, FL, USA, 3rd edition, 2008.

[NKA+06] M. Nielsen, E. N. Kamavuako, M. M. Andersen, M.-F. Lucas, and D. Farina. Optimal wavelets for biomedical signal compression. *Medical and Biological Engineering and Computing*, 44(7):561–568, Jul 2006.

[ORBP16] P. Ochs, R. Ranftl, T. Brox, and T. Pock. Techniques for gradient-based bilevel optimization with non-smooth lower level problems. *Journal of Mathematical Imaging and Vision*, 56(2):175–194, 2016.

[PBBZP16] N. Pustelnik, A. Benazza-Benhayia, Y. Zheng, and J.-C. Pesquet. Wavelet-based image deconvolution and reconstruction. *Wiley Encyclopedia of Electrical and Electronics Engineering*, pages 1–34, 2016.

[RM18a] D. Recoskie and R. Mann. Gradient-based filter design for the dual-tree wavelet transform. *CoRR*, abs/1806.01793, 2018.

[RM18b] D. Recoskie and R. Mann. Learning filters for the 2d wavelet transform. In *2018 15th Conference on Computer and Robot Vision (CRV)*, pages 198–205, May 2018.

[Søg17] A. Søgaard. Learning optimal wavelet bases using a neural network approach. *arXiv preprint arXiv:1706.03041*, 2017.

[Thi04] H. Thielemann. Optimally matched wavelets. In *PAMM: Proceedings in Applied Mathematics and Mechanics*, volume 4, pages 586–587. Wiley Online Library, 2004.

[VN17] Q. Van Nguyen. Forward-backward splitting with Bregman distances. *Vietnam Journal of Mathematics*, 45(3):519–539, 2017.