Apprentissage d'a priori analyse

Gabriel PEYRÉ[1], Jalal FADILI[2]

[1]CEREMADE CNRS-Université Paris Dauphine

[2]GREYC CNRS-ENSICAEN-Université de Caen

gabriel.peyre@ceremade.dauphine.fr, jalal.fadili@greyc.ensicaen.fr

Résumé – Ce papier étudie l'optimisation d'un a priori analyse parcimonieux pour le débruitage d'images. Un dictionnaire redondant ou bien un filtre de convolution est optimisé afin de minimiser l'erreur de débruitage sur des signaux donnés en exemples. L'apprentissage de dictionnaires pour les représentations parcimonieuses est traditionnellement cantonné à une formulation synthèse. Ce travail généralise les travaux précédents à l'apprentissage pour une formulation analyse. Pour ce faire, nous formulons le problème comme un programme d'optimisation bi-niveau pour lequel une analyse et un algorithme de descente sont fournis. L'algorithme est appliqué à des signaux synthétiques et naturels 1D afin d'apprendre un dictionnaire non-structuré et un dictionnaire de convolution. L'a priori ainsi obtenu améliore les résultats de débruitage par rapport à un a priori de variation totale. Ces résultats sont encourageants en vue d'une extension en 2D. *

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 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 1D signal denoising are reported and potential applicability and extensions are discussed.

1 Introduction

Un méthode populaire pour débruiter les signaux et les images consiste à résoudre un problème d'optimisation variationnel comportant un terme d'attache aux données et un terme de régularisation. Cette approche se généralise aux problèmes inverses (super-résolution, inpainting, etc.). Parmi le grand nombre d'approches existantes, les a priori parcimonieux ont été beaucoup étudiés ces dernières années. De tels a priori sont efficaces pour retrouver des structures complexes présentes dans les signaux et les images naturels.

1.1 A priori analyse et synthèse

On considère un dictionnaire redondant $D = (d_m)_{m=0}^{P-1}$ de P atomes dans $\mathbb{R}^N$ pour représenter les données. Les travaux antérieurs sur la régularisation parcimonieuse se divisent en deux grandes familles : celle avec formulation dite analyse et la seconde dite synthèse. Si $y \in \mathbb{R}^N$ est un signal (ou une image) bruité, un a priori synthèse calcule un ensemble de coefficients u qui sont solutions de

$$u(D,y) \in \operatorname*{argmin}_{u \in \mathbb{R}^P} \frac{1}{2}\|y - Du\|^2 + \Gamma(u), \qquad (1)$$

et le signal débruité est ensuite synthétisé comme $Du(D,y)$. Ici $\Gamma(u)$ est un une fonctionnelle de pénalisation convexe favorisant la parcimonie. Un choix populaire de fonction Γ est la norme ℓ^1, $\Gamma(u) = \sum_m |u_m|$, [3]. L'a priori synthèse dans des dictionnaires redondants comme une trame d'ondelettes invariantes par translation est largement utilisé pour des problèmes de débruitage ainsi que pour résoudre des problèmes inverses, voir par exemple [7].

Un a priori analyse cherche un signal ou une image $x = x(D,y)$ dont les produits scalaires avec les atomes du dictionnaire D sont parcimonieux. Ceci correspond à la minimisation de

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

On peut choisir D^* comme étant l'opérateur d'analyse d'un dictionnaire redondant comme une trame d'ondelettes invariantes par translation. Une utilisation plus fréquente de l'a priori analyse définit D^* comme un opérateur différentiel pour imposer une certaine régularité aux signaux, tout en permettant de restaurer des discontinuités. Ceci mène par exemple à une discrétisation de l'a priori de variation totale (TV) introduit par Rudin, Osher et Fatemi [9].

Le travail de [5] a été le premier à étudier les liens entre les a priori analyse et synthèse. Par exemple, si le dictionnaire D est carré est inversible, alors les classes de problème (1) et (2) son équivalentes. Dans le cas où D est unitaire, alors les estimateurs $x(D,y)$ et $Du(D,y)$ calculés par (1) et (2) sont égaux. Lorsque le dictionnaire D est redondant, les problèmes (1) et (2) ne sont pas équivalents. C'est par exemple le cas pour l'a priori de variation totale.

*Ce travail a été en partie financé par le projet ANR NatImages, ANR-08-EMER-009.

1.2 Apprentissage de dictionnaires

Les travaux pionniers de Olshausen et Field [8] ont initié de nombreuses approches pour apprendre un dictionnaire D à partir d'un ensemble d'exemples. Toutes les approches précédentes se sont concentrées sur les a priori de type synthèse (1). Elles effectuent une minimisation de la fonctionnelle à la fois sur les coefficients et sur le dictionnaire, voir par exemple [8, 1]. Il faut se garder toutefois de mimer cette approche en l'appliquant hâtivement à la fonctionnelle (2), car le minimiseur global en (D, x) est trivial dans ce cas et sans intérêt. Récemment, différents travaux ont proposé d'apprendre un dictionnaire optimisé pour une tâche précise pour laquelle le dictionnaire est utilisé (par exemple débruitage, classification ou super-résolution), voir [6]. Ces approches nécessitent la résolution d'un programme d'optimisation bi-niveau, car une partie des variables dans la fonctionnelle à optimiser (par exemple les coefficients parcimonieux) est elle-même solution d'un autre problème d'optimisation.

1.3 Contributions

Nous proposons une méthode pour apprendre un dictionnaire en formulation analyse. Le dictionnaire adapté peut être non-structuré ou convolutif, et est la solution d'un problème d'optimisation bi-niveau. Des résultats numériques sur des signaux 1D montre l'efficacité de notre approche.

2 Apprentissage de dictionnaire avec a priori analyse

2.1 A prior de parcimonie lissé

Il est difficile d'utiliser directement la norme ℓ^1 comme pénalité Γ de parcimonie, car elle n'est pas différentiable. Afin de pouvoir utiliser des méthodes de descente gradient pour l'optimisation bi-niveau du dictionnaire, nous proposons d'utiliser une version lissée de la norme ℓ^1, définie comme

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

pour un paramètre de lissage $\varepsilon > 0$ suffisamment petit. Ce type de lissage est fréquent dans la littérature du traitement variationel des images, et a été également utilisé dans le cadre de l'apprentissage de dictionnaire avec a priori synthèse [2].

Dans la suite, nous notons $\nabla_\Gamma[u] \in \mathbb{R}^P$ le gradient de Γ au point $u \in \mathbb{R}^P$, et $\mathrm{H}_\Gamma[u] : \mathbb{R}^P \to \mathbb{R}^P$ sa Hessienne. Pour le cas particulier défini en (3), nous avons

$$\nabla_\Gamma[u]_m = \frac{u_m}{\sqrt{\varepsilon^2 + |u_m|^2}},$$

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

2.2 Sensibilité de la solution pour l'a priori analyse

Le théorème suivant donne la dérivée de la solution $x(D, y)$ par rapport au dictionnaire D. Afin d'alléger les notations, on note $x(D) = x(D, y)$ car y est fixé dans ce théorème.

Theorem 1. *L'application $D \mapsto x(D)$ est de classe C^1, et sa dérivée au point $D \in \mathbb{R}^{N\times P}$ satisfait pour tout $z \in \mathbb{R}^N$*

$$dx[D]^*(z) = -\bar{z} \times \nabla_\Gamma[u]^{\mathrm{T}} - x(D) \times \left(\mathrm{H}_\Gamma[u](D^*\bar{z})\right)^{\mathrm{T}}, \tag{4}$$
$$\text{où} \quad \Delta = \mathrm{Id} + D\mathrm{H}_\Gamma[u]D^* : \mathbb{R}^N \to \mathbb{R}^N,$$

et $u = D^*x(D)$, $\bar{z} = \Delta^{-1}z$ et où Id *est l'opérateur identité sur $\mathbb{R}^N$.*

Démonstration. Par convexité, la condition nécessaire et suffisante d'optimalité du premier ordre de (2) donne

$$x(D) - y + D\nabla_\Gamma[D^*x(D)] = 0. \tag{5}$$

L'équation (5) définit de façon implicite la fonctionnelle $D \mapsto x(D)$. Sa dérivabilité et sa dérivée sont obtenues en appliquant le théorème des fonctions implicites à la fonctionnelle $S(x, D) = x - y + D\nabla_\Gamma[D^*x]$. En effet, en différentiant (5) par rapport à D dans la direction $\delta \in \mathbb{R}^{N\times P}$, on obtient

$$dx[D](\delta) + \delta\nabla_\Gamma[D^*x(D)] + D\mathrm{H}_\Gamma[D^*x(D)](\delta^*x(D) + D^*dx[D](\delta)) = 0, \tag{6}$$

La convexité de Γ, implique que Δ défini par (4) est définie positive, et donc inversible. Le théorème des fonctions implicite permet de conclure que $D \mapsto x(D)$ est de classe C^1, et que sa dérivée est obtenue en inversant (6)

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

Transposer cette équation donne (4). □

Il est important de remarquer que notre étude de la sensibilité de la régularisation analyse au dictionnaire diffère nettement de celle de l'a priori synthèse [6]. Afin d'obtenir la différentiabilité de $D \mapsto x(D, y)$, nous avons besoin de lisser la norme ℓ^1 ce qui justifie notre pénalité $\Gamma(D^*x)$. Dans [6], les auteurs strict-convexifient $\Gamma(u)$ en ajoutant une perturbation quadratique à la norme ℓ^1, et ce afin de garantir l'unicité du minimiseur $u(D, y)$ et d'obtenir la différentiabilité de $D \mapsto u(D, y)$.

3 Apprentissage de dictionnaire avec a priori analyse

3.1 Apprentissage de dictionnaire non-structuré

Le dictionnaire est optimisé afin d'obtenir les meilleures performances sur un ensemble de paires (y_k, x_k) d'exemples, où $x_k \in \mathbb{R}^N$ est un signal sans bruit, et $y_k = x_k + w_k$ est une version bruitée, avec w_k un bruit additif, qui est ici supposé centré

blanc Gaussien. Le dictionnaire est alors obtenu en minimisant le risque empirique moyen sur l'ensemble des exemples

$$\min_{D\in\mathbb{R}^{N\times P}} \left\{ \mathcal{E}(D) = \frac{1}{2}\sum_k \|x_k - x(D,y_k)\|^2 \right\}. \quad (7)$$

Ceci correspond à un problème d'optimisation non-convexe. C'est un problème bi-niveau (voir [4]) car $x(D,y_k)$ est lui-même la solution du problème d'optimisation (2).

D'après le Théorème 1, $\mathcal{E}$ est une fonctionnelle C^1, et l'on peut calculer un point stationnaire de (7) à l'aide d'une descente de gradient

$$D^{(t+1)} = D^{(t)} - \eta_t \nabla\mathcal{E}(D^{(t)}), \quad (8)$$
$$\text{avec} \quad \nabla\mathcal{E}(D) = \sum_k dx[D,y_k]^*(x(D,y_k) - x_k)$$

où $0 < \eta_t < \eta$ est une suite de pas de descente suffisamment petits et où $dx[D,y_k)] : \mathbb{R}^{N\times P} \to \mathbb{R}^P$ est la dérivée de $D \mapsto x(D,y_k)$ prise en D, dont l'expression est donné au Théorème 1, et $dx[D,y_k]^*$ est son opérateur adjoint.

3.2 Apprentissage d'un dictionnaire convolutif

Une famille populaire d'a priori analyse est obtenue en forçant D à être invariant par translation. Il est ainsi défini à l'aide d'un seul atome $\gamma \in \mathbb{R}^N$, de sorte que $d_i = \gamma(\cdot - i)$, où l'on a utiliser des conditions au bord périodique par simplicité. Ceci signifie que l'a priori est défini comme une convolution circulaire avec γ, puisque $D^*x = \gamma \star x$. Un tel a priori généralise l'a priori de variation totale, où en 1D, γ est un filtre calculant une dérivée première par différences finies (dérivateur discret).

On note $\varphi(\gamma) \in \mathbb{R}^{N\times N}$ l'opérateur de convolution circulaire définit par γ, et $\varphi(\gamma)^*$ son adjoint, qui est associé au filtre $\widetilde{\gamma}$, où, pour un vecteur x, $\widetilde{x}_i = x_{-i}$. Le noyau de convolution γ est appris en résolvant un problème bi-niveau avec (7) et (2) particularisé à un dictionnaire paramétré par γ,

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

où $\bar{x}(\gamma,y) = x(\varphi(\gamma),y)$. Cette énergie est minimisée par une descente de gradient similaire à (8). La proposition suivante donne l'expression de l'adjoint de la dérivée de $\bar{x}(\gamma)$ en utilisant uniquement des convolutions (la dépendance par rapport à y a été enlevée pour alléger les expressions).

Proposition 1. *Pour un vecteur $z \in \mathbb{R}^N$, l'opérateur adjoint de la dérivée de $\gamma \mapsto \bar{x}(\gamma)$ et*

$$d\bar{x}[\gamma]^*(z) = -\bar{z} \star \widetilde{\nabla_\Gamma[u]} - \bar{x}(\gamma) \star (\mathrm{H}_\Gamma\widetilde{[u](\gamma \star \bar{z})}), \quad (9)$$

où $u = \gamma \star \bar{x}(\gamma)$, $\bar{z} = \Delta^{-1}z$ et

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

Démonstration. A l'aide de la règle de dérivation des fonctions composées, on obtient

$$d\bar{x}[\gamma]^*(z) = d\varphi[\gamma]^* \left(dx[\varphi(\gamma)]^*(z)\right). \quad (10)$$

Pour une convolution 1D, on peut montrer que l'adjoint de la dérivée $d\varphi[\gamma]^*$ correspond à la sommation le des diagonales de la matrice $A \in \mathbb{R}^{N\times N}$,

$$d\varphi[\gamma]^*(A) = \alpha \quad \text{où} \quad \alpha_i = \sum_{s-t=i} A_{s,t}\,, \quad (11)$$

et où l'égalité des indices est entendue modulo N. Cette expression s'étend à des convolutions en dimensions supérieures. On peut alors vérifier que si $A = uv^{\mathrm{T}}$, ce qui signifie que $A_{i,j} = u_i v_j$, alors $d\varphi[\gamma]^*(A) = u \star \widetilde{v}$. Ainsi, en regroupant (4) (avec $\varphi(\gamma)$ à la place de D^*), (10) et (11), on obtient l'expression de l'adjoint de la dérivée (9), où (11) permet de simplifier (9) à l'aide de seulement deux convolutions. □

Le calcul de $d\bar{x}[\gamma]^*(z)$ à l'aide (9) ne nécessite que quatre convolutions, et la résolution du système linéaire $\Delta\bar{z} = z$. Ce système peut se résoudre efficacement à l'aide de quelque itérations de gradients conjugués en exploitant la structure particulière de l'opérateur $\varphi(\gamma)$ et de son adjoint $\varphi(\gamma)^*$ (i.e. diagonaux dans Fourier).

4 Expériences numériques

Dans les expériences numériques qui suivent, les signaux observés $y_k = x_k + w_k$ sont contaminés par un bruit additif centré blanc Gaussien w_k d'écart-type 0.03, avec la normalisation $\|x_k\|_\infty = 1\ \forall k$.

4.1 Dictionnaire non-structuré

Dans la première expérience, nous considérons des signaux 1D constants par morceaux $x_k = 1_{[a_k,b_k]}$ de taille $N = 128$ où $0 \leqslant a_k < b_k < N$ sont des positions aléatoires distinctes uniformes sur $[0, N-1]$. Pour ce type de signaux, la régularisation TV est très efficace, car elle permet de retrouver de façon quasi-parfaite les signaux constants par morceaux. Elle est obtenue à l'aide du filtre $\gamma_{\mathrm{TV},\lambda}$ tel que $\gamma_{\mathrm{TV},\lambda}(0) = \lambda$, $\gamma_{\mathrm{TV},\lambda}(1) = -\lambda$ et 0 sinon. La valeur de λ est calculée afin de minimiser $\bar{\mathcal{E}}(\gamma_{\mathrm{TV},\lambda}) = \mathcal{E}(D_{\mathrm{TV},\lambda})$. La méthode d'apprentissage du dictionnaire non-structuré est appliquée avec un dictionnaire initial aléatoire $D^{(0)}$ dont les entrées sont iid Gaussiennes, et avec $\varepsilon = 10^{-3}$. La figure 1, gauche, montre que l'énergie de débruitage $\mathcal{E}(D^{(t)})$ converge vers l'énergie du dictionnaire TV $\mathcal{E}(D_{\mathrm{TV},\lambda})$. La convergence est assez lente à cause de la mauvaise initialisation du dictionnaire et la faible valeur de ε. Ceci démontre la capacité de notre méthode à retrouver le filtre TV optimal à partir d'une initialisation arbitraire.

4.2 Dictionnaire convolutif

Dans la seconde expérience, un a priori analyse convolutif est appris à partir d'un ensemble de signaux naturels de taille $N = 256$ avec $\varepsilon = 10^{-2}$. Chaque signal $x_k(i) = f(a_k,i)$ est extrait comme une ligne d'indice a_k uniformément aléatoirement de l'image "lena" f de taille 256^2, voir figure 2, gauche.

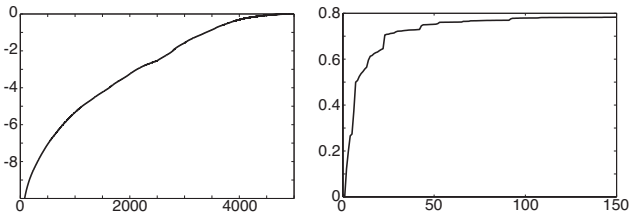

FIGURE 1 – *Évolution de l'énergie de débruitage* $-10\log_{10}(\mathcal{E}(D^{(t)})/\mathcal{E}(D_{TV,\lambda}))$ *au cours des itérations* t*. Gauche : dictionnaire non structuré, signaux constants par morceaux en exemples. Droite : dictionnaire convolutif, signaux naturels.*

Nous initialisons $\gamma^{(0)} = \gamma_{\text{TV},\lambda}$ comme le filtre TV avec une valeur optimale pour λ. La figure 1, droite, montre comment l'apprentissage améliore la performance de débruitage d'environ 0.8 dB par rapport à une régularisation TV sur les données d'apprentissage. On peut noter que la structuration sous forme de convolution ainsi que la bonne initialisation améliore la convergence de la méthode. Nous avons ensuite appliqué le filtre appris sur un autre ensemble de signaux naturels, obtenus comme des lignes de l'image "Boat". Ceci a donné lieu à une amélioration des performance de débruitage d'environ 0.3 dB par rapport au filtre TV. Ceci est moins que pour l'image "Lena", ce qui peut s'expliquer par le faire que l'image "Boat" comporte plus de contours et est moins oscillante (moins de texture), ce qui rend le filtre appris moins adapté.

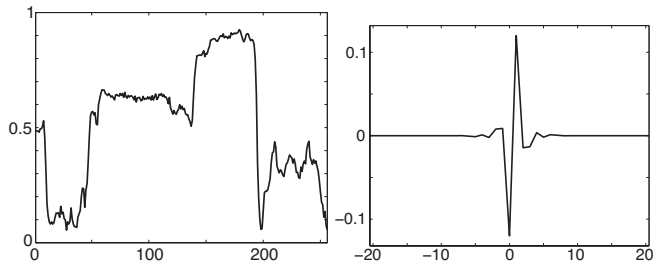

FIGURE 2 – *Gauche : exemple de signal naturel 1D* x_k *utilisé pour l'apprentissage. Droite : filtre optimal* $\gamma^{(\infty)}$ *appris par notre méthode.*

Conclusion et perspectives

Nous avons introduit dans ce papier une nouvelle méthode pour l'apprentissage d'un a priori parcimonieux avec une formulation analyse. Nous avons montré quelques résultats préliminaire pour illustrer les applications potentielles pour le débruitage de signaux 1D. Plusieurs extensions de cette méthodes sont envisageables :

- Extension en dimension plus élevée, comme par exemple pour des images en 2D.
- Extension à l'apprentissage de plusieurs noyaux de convolution simultanément.
- Extension à d'autres problèmes inverses comme la déconvolution, ou encore à la classification.

Références

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

[7] S. Mallat. *A Wavelet Tour of Signal Processing, 3rd edition*. Elsevier, 2009.

[8] 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.

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