Regularized Bregman Approach for Total-Variation Parameter Learning

Kristian Bredies, Enis Chenchene
Department of Mathematics and Scientific Computing, University of Graz
Heinrichstraße 36, 8010 Graz, Austria
kristian.bredies@uni-graz.at, enis.chenchene@uni-graz.at

Abstract

We present a new approach for data-driven tuning of regularization parameters for total-variation denoising. The proposed approach hinges on a specific proxy for the underlying bilevel problem, which admits a tractable monolevel reformulation that can be efficiently solved with a new conditional-gradient-type method. We show numerical experiments, and open avenues for promising extensions.

1. Introduction

Tuning the regularization parameter in the classical total-variation denoising model by Rudin, Osher and Fatemi, *i.e.*,

$$\min_{u\in\mathbb{R}^n} \tfrac{1}{2}\|u-\xi\|^2 + \alpha\,\mathrm{TV}(u), \tag{1}$$

where $\xi \in \mathbb{R}^n$ is noisy image and TV is the total-variation seminorm [1, Eq. (4.1)], is a delicate challenge. A standard approach to do so [1,3] aims at solving the *bilevel* problem:

$$\min_{\alpha\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} \|u_i^\dagger - u_i^{\alpha(\xi_i)}\|^2, \tag{2}$$

where $u_i^{\alpha(\xi_i)}$ is the solution to (1) with parameter $\alpha(\xi_i)$ and data ξ_i, and $\mathcal{F} = \{\alpha\colon \mathbb{R}^n \to \mathbb{R}_+\}$ is a given *model*. For a noisy image ξ, an optimal solution to (2) should provide a regularization parameter $\alpha(\xi)$ such that the denoised image is close to the corresponding ground-truth $u^\dagger$.

2. Regularized Bregman Learning

We consider the following proxy for (2):

$$\min_{\alpha\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{2}\|u^\dagger - u_i^{\alpha(\xi_i)}\|^2 + D_{\alpha(\xi_i)}(u_i^\dagger, u_i^{\alpha(\xi_i)})\right), \tag{3}$$

where D_{α_i} is the *Bregman*-divergence associated to α_i TV, with $\alpha_i := \alpha(\xi_i)$ for all $i \in [N] := \{1,\dots,N\}$, *i.e.*,

$$\begin{aligned} D_{\alpha_i}(u_i^\dagger, u_i^{\alpha_i}) := {}& \alpha_i\,\mathrm{TV}(u_i^\dagger) - \alpha_i\,\mathrm{TV}(u_i^{\alpha_i}) \\ & - \langle \xi_i - u_i^{\alpha_i}, u_i^\dagger - u_i^{\alpha_i}\rangle. \end{aligned} \tag{4}$$

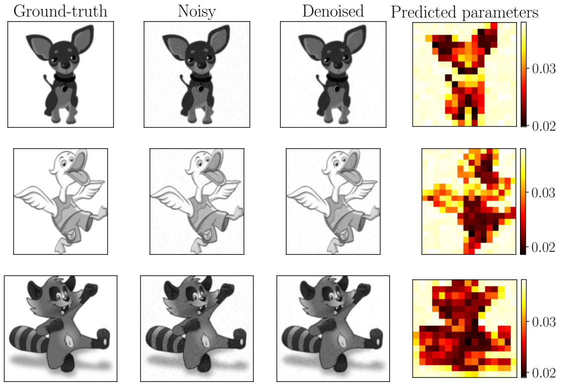

Figure 1. Denoising with predicted TV-parameters for each patch. *cf*. Experiment 1 in Sec. 4.

Note that D_{α_i} is always non-negative since TV is convex and $\xi_i - u_i^{\alpha_i} \in \alpha_i \partial\,\mathrm{TV}(u_i^{\alpha_i})$, hence (3) is indeed an upper bound for (2).

Monolevel reformulation. At first sight, problem (3) has again a complex bilevel nature. However, applying the polarization identity to the inner product in the right hand-side of (4), and using standard duality theory (see, *e.g.*, [2, Section III]), we get for all $i \in [N]$ that

$$\begin{aligned} &\tfrac{1}{2}\|u_i^\dagger - u_i^{\alpha_i}\|^2 + D_{\alpha_i}(u_i^\dagger, u_i^{\alpha_i}) \\ &= \alpha_i \mathrm{TV}(u_i^\dagger) - \left(\tfrac{1}{2}\|u_i^{\alpha_i} - \xi_i\|^2 + \alpha_i\,\mathrm{TV}(u_i^{\alpha_i})\right) + C \\ &= \alpha_i \mathrm{TV}(u_i^\dagger) + \inf_{\boldsymbol{v}_i\in\boldsymbol{B}_{\alpha_i}^\infty} \tfrac{1}{2}\|\operatorname{div}\boldsymbol{v}_i + \xi_i\|^2 + C', \end{aligned} \tag{5}$$

where C and C' are constants that do not depend on α, $\|\boldsymbol{v}\|_{\infty,2} := \max_{j\in[n]} \|\boldsymbol{v}_j\|_2$ for all $\boldsymbol{v} \in \mathbb{R}^{n\times 2}$, and $\boldsymbol{B}_{\alpha_i}^\infty$ is the α_i-ball with respect to $\|\cdot\|_{\infty,2}$. Plugging (5) into (3) we get the following equivalent formulation of (3):

$$\min_{\substack{\alpha\in\mathcal{F},\\ \boldsymbol{v}_i\in\boldsymbol{B}_{\alpha(\xi_i)}^\infty}} \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{2}\|\operatorname{div}\boldsymbol{v}_i + \xi_i\|^2 + \alpha(\xi_i)\,\mathrm{TV}(u_i^\dagger)\right). \tag{6}$$

It only remains to fix the model $\mathcal{F}$.

Model selection. We investigate the performance of quadratic models, *i.e.*,

$$\mathcal{F} = \{\alpha\colon \mathbb{R}^n \to \mathbb{R}_+ \mid \alpha(\xi) = \bar{\xi}^* A\bar{\xi},\ A \succcurlyeq 0\}, \qquad (7)$$

where for every $\xi \in \mathbb{R}^n$, $\bar{\xi} = (\xi, 1) \in \mathbb{R}^{n+1}$ and $A \succcurlyeq 0$ stands for symmetric, positive semidefinite matrices. With this choice, the monolevel problem (6) turns into

$$\min_{(A,\boldsymbol{v})\in\mathcal{C}} \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{2}\|\operatorname{div}\boldsymbol{v}_i+\xi_i\|^2+\bar{\xi}_i^{\,*}A\bar{\xi}_i\,\mathrm{TV}(u_i^\dagger)\right), \qquad (8)$$

where $\mathcal{C}$ is the closed convex set of tuples $(A, \boldsymbol{v}) \in \mathbb{R}^{(n+1)^2} \times \mathbb{R}^{n\times 2} \times \cdots \times \mathbb{R}^{n\times 2}$ such that $A \succcurlyeq 0$ and

$$\|\boldsymbol{v}_i\|_{\infty,2} \le \bar{\xi}_i^{\,*}A\bar{\xi}_i \quad \text{for all } i \in [N]. \qquad (9)$$

3. Training procedure

Solving (8) is challenging due to the complex and unbounded constraint set, making traditional methods such as proximal- or conditional-gradient impractical. Instead, we employ a more efficient *hybrid* approach introduced in [2]. For $(A^k, \boldsymbol{v}^k)$, and $k \in \mathbb{N}$, we first compute

$$\begin{cases} \widetilde{A}^{k+1} = \mathrm{Proj}_{\succcurlyeq}\left(A^k - \dfrac{1}{\lambda N}\displaystyle\sum_{i=1}^{N} c_i^k\bar{\xi}_i\bar{\xi}_i^{\,*}\right), \\ (\widetilde{\boldsymbol{v}}_i^{k+1})_j = \dfrac{(\nabla\widetilde{u}_i^k)_j}{\|(\nabla\widetilde{u}_i^k)_j\|_2}\bar{\xi}_i^{\,*}\widetilde{A}^{k+1}\bar{\xi}_i, \quad i \in [N],\ j \in [n], \end{cases} \qquad (10)$$

where $\mathrm{Proj}_{\succcurlyeq}$ is the projection onto the set of symmetric, positive semidefinite matrices, $c_i^k := \mathrm{TV}(u_i^\dagger) - \mathrm{TV}(\widetilde{u}_i^k)$, $\lambda > 0$, and $\widetilde{u}_i^k := \operatorname{div}\boldsymbol{v}_i^k + \xi_i$. Then, we update the current iterate via

$$\begin{cases} A^{k+1} = A^k + \theta_k(\widetilde{A}^{k+1} - A^k), \\ \boldsymbol{v}^{k+1} = \boldsymbol{v}^k + \theta_k(\widetilde{\boldsymbol{v}}^{k+1} - \boldsymbol{v}^k), \end{cases} \qquad (11)$$

where θ_k is a step-size given by

$$\min\left\{1, \frac{\frac{1}{N}\sum_{i=1}^{N}\left(G_i(\widetilde{u}_i^k, \boldsymbol{v}_i^k) - c_i^k\bar{\xi}_i^{\,*}(\widetilde{A}^{k+1} - A^k)\bar{\xi}_i\right)}{\frac{4}{N}\|\widetilde{A}^{k+1} - A^k\|^2 + \|\widetilde{\boldsymbol{v}}^{k+1} - \boldsymbol{v}^k\|^2}\right\},$$

where G_i is the primal-dual gap associated to (1) with data ξ_i and parameter $\alpha(\xi_i)$, *i.e.*, for all $u \in \mathbb{R}^n$ and $\boldsymbol{v} \in \mathbb{R}^{n\times 2}$

$$\begin{aligned} G_i(u, \boldsymbol{v}) :=& \tfrac{1}{2}\|u - \xi_i\|^2 + \alpha(\xi_i)\,\mathrm{TV}(u) \\ &+ \tfrac{1}{2}\|\operatorname{div}\boldsymbol{v}_i + \xi_i\|^2 - \tfrac{1}{2}\|\xi_i\|^2. \end{aligned} \qquad (12)$$

According to [2], employing (11) to solve (8), we can expect that i) the objective function of (8) evaluated on $(A^k, \boldsymbol{v}^k)$ converges monotonically to the infimum value with a $o(k^{-1/3})$ worst-case rate, ii) the iterates remain bounded with cluster points lying in the set of optimal solutions, iii) $A^k \to A^*$ with $(A^*, \boldsymbol{v}^*)$ optimal for some $\boldsymbol{v}^*$. The latter is particularly beneficial in our application.

Models	Quadratic	Constant $\alpha = \eta\ 10^{-4}$		
		$\eta = 13.9$	$\eta = 27.13$	$\eta = 37.3$
MSE	**0.1529**	0.2917	0.1833	0.1777

Table 1. Results of Experiment 2 in Sec. 4.

4. Numerical experiments.

We use a training set of $N = 101440$ images with $p = 16$ and Gaussian noise of variance 0.05. We set $\lambda = 50$ and run (11) until a residual (*cf*. D in [2]) reaches 10^{-4}.

Experiment 1. We use the trained model to denoise a new test image split into 16×16 patches. Each patch's TV-parameter is computed using the trained model, showing higher values for flatter regions (like backgrounds) and lower values for complex image parts, as seen in Fig. 1.

Experiment 2. We assess our model performance against 8 constant parameters choices spaced evenly from 10^{-4} to 10^{-1} on a test set of $N_t = 200$ images. As performance metrics, we consider

$$\mathrm{MSE} := \tfrac{1}{N_t}\textstyle\sum_{i=1}^{N_t}\|u_i^\dagger - u_i^{\alpha(\xi_i)}\|^2, \qquad (13)$$

or (13) replacing $\alpha(\xi_i)$ with the above constant values. The three best results are contained in Tab. 1.

Results and conclusions. As expected, flexible (non-constant) models can improve performances (see Tab. 1). These can be efficiently trained minimizing the loss function (3) via the hybrid method (11). Future work includes investigating stochastic variants of (11), exploring different models such as Neural Networks, and extending the learning framework to diverse variational denoising models.

Acknowledgments. This work has received funding from the European Union's Framework Programme for Research and Innovation Horizon 2020 (2014–2020) under the Marie Skłodowska-Curie Grant Agreement No. 861137. The Department of Mathematics and Scientific Computing is a member of NAWI Graz (https://www.nawigraz.at/).

References

[1] Luca Calatroni, Chung Cao, Juan Carlos De los Reyes, Carola-Bibiane Schönlieb, and Tuomo Valkonen. *Bilevel approaches for learning of variational imaging models*, pages 252–290. De Gruyter, Berlin, Boston, 2017.

[2] Enis Chenchene, Alireza Hosseini, and Kristian Bredies. A hybrid proximal generalized conditional gradient method and application to total variation parameter learning. In *2023 European Control Conference (ECC)*, pages 1–6, 2023.

[3] Karl Kunisch and Thomas Pock. A bilevel optimization approach for parameter learning in variational models. *SIAM Journal on Imaging Sciences*, 6(2):938–983, 2013.