Restart strategies enabling automatic differentiation for hyperparameter tuning in inverse problems

Leo Davy, Nelly Pustelnik, Patrice Abry

Restart strategies enabling automatic differentiation for hyperparameter tuning in inverse problems

Leo Davy, Nelly Pustelnik, Patrice Abry
CNRS, ENS de Lyon
Laboratoire de physique, F-69007 Lyon, France
leo.davy@ens-lyon.fr, nelly.pustelnik@ens-lyon.fr, patrice.abry@ens-lyon.fr

***Abstract*—Numerous signal/image processing tasks can be formulated as variational problems, whose solutions depend, often crucially, on the values of hyperparameters. Their automated selection usually involves the computation of gradients of a well chosen loss function, which often turns unfeasible analytically. The deep-learning inspired use of automatic differentiation to compute such gradients, though appealing, is significantly impaired by the usually large number of iterations inherently attached to functional minimization in variational problems. The present work proposes and assesses the use of a restart strategy for automated hyperparameter tuning, combining the benefits of automatic differentiation with properties of proximal iterative algorithms. It studies theoretically its conditions of applicability in a generic algorithmic framework and its specification to accelerated Chambolle-Pock iterations when dealing with strongly convex objective function. The effectiveness is illustrated for image denoising and texture segmentation problems.**

***Index Terms*—Hyperparameter selection, proximal algorithms, restart strategy, hypergradient computation**

I. Introduction

Context. Numerous image/signal processing problems, e.g., restoration [1], [2], segmentation [2], pandemic intensity monitoring [3], can be modeled by the following functional minimization formulation:

$$\widehat{x}(z,\theta) := \operatorname{argmin}_x f(x,z) + \lambda g(Lx) \tag{1}$$

with f a data-fidelity term, g a prior/expert information, and where $\theta = \{\lambda\}$ (or possibly $\theta = \{\lambda, L\}$ in the context of dictionary learning) are hyperparameters, whose selection drastically impacts the achieved solution.

Solving minimization (1) generally entails an iterative scheme involving implicit or explicit (sub)-gradient descent steps on f and g, leading to algorithmic steps denoted $\phi_{z,\theta,k}$ such that the procedure

$$\Phi_{z,\theta,K} := \phi_{z,\theta,K} \circ \phi_{z,\theta,K-1} \cdots \circ \phi_{z,\theta,1}, \tag{2}$$

in the limit of infinitely many iterations leads to $\widehat{x}(z,\theta)$. A large panel of (proximal) algorithms have been developed over the last twenty years to efficiently solve (1) depending on the properties of the involved functions (cf. [4] and references therein).

This work is funded by the Fondation Simone et Cino Del Duca - Institut de France and also supported by ANR-19-CE48-0009 Multisc'In, DGA/AID (01D22020572), IADoc@UdL (ANR-20-THIA-0007-01).

Regarding hyperparameter selection, the strategies are often based on minimizing a well-chosen (supervised or unsupervised) loss $\mathcal{L}$:

$$\widehat{\theta} \in \operatorname{Argmin}_{\theta\in\Theta}\ \mathcal{L}(\widehat{x}(z,\theta)). \tag{3}$$

To solve (3), the most naive strategy is a brute force grid search but it is obviously infeasible in most realistic problems. Therefore, optimization tools can be considered, such as BFGS, ADAM, or SGD, entailing the computation of the so-called *hypergradient* $\partial_\theta(\mathcal{L} \circ \widehat{x}(z,\theta))$ via automatic differentiation (AD) [5], [6], [7].

However, it is well known [5] that AD's memory footprint increases linearly with the number of iterations which prohibits its direct use when $\widehat{x}(z,\theta)$ is obtained from standard optimization algorithms $\Phi_{z,\theta,K}$ which typically have more than $K = 10^3$ steps to solve (1). Effectively minimizing (3) for large K is the difficult yet critical challenge addressed here.

State-of-the-art for hypergradient computation. Several attempts addressed the issue of coping with large numbers of iterations in using AD.

In the context of unfolded strategies [8], [9], [10], $\widehat{x}(z,\theta)$ is replaced with the solution obtained after a limited number of iterations thus AD can be applied. This however induces a modification of the optimization landscape for $\mathcal{L}(\widehat{x}(z,\cdot))$ and does not exactly solve the bilevel formulation (1)-(3). Another strategy involving unfolded schemes (i.e. truncated iterations) referred to as Deep Equilibrium allows to better solve this bilevel problem by using the fixed-point equation $\widehat{x} = \Phi_{z,\theta,K}(\widehat{x})$ and the implicit function theorem to compute the hypergradient but at the price of a complex implementation [11], [12].

Alternatively, *iterative differentiation* in reverse mode has been proposed in [13], [14], [15], [16]. This strategy offers very good performance but at the price of a complex implementation as requiring to derive the closed form of the derivative of each algorithmic step $\phi_{z,\theta,k}$.

Acceleration and restart. Along a different line, as AD is adapted for a limited number of iterations, a large panel of optimization strategies were devised aiming to decrease the number of iterations: e.g. inertia [17], [18], preconditioning [19], multilevel [20], [21], or restart [22], [23], [24]. Though promising, these accelerated optimization schemes were massively explored to solve (1) but much less to solve the more complex bilevel problem (3), a gap the present work aims to contribute to fill.

Goals, contributions and outline. In this work, we devise and assess theoretically and practically the use of a restart strategy for the automated selection of hyperparameters, in the generic class of problems defined by (1)-(3). The proposed restart strategy for solving the bilevel problem is defined and analyzed in Section II. This generic theoretical analysis is specified for the specific case of the so-called Chambolle-Pock algorithm, classically used to solve (1) (Section III). Finally, the proposed restart strategy is illustrated at work on two classical problems (image denoising and texture segmentation) in Section IV, which quantifies its effective ability to perform efficiently and relevantly hyperparameter selection for these standard inverse problems.

II. Proposed restart strategy

A. Principle and definitions

The key ingredient of the proposed restart strategy is to replace K iterations in the algorithm solving (3) by T (the restart time) repeated applications of $K_0 \ll K$ iterations such that $K \sim TK_0$.

Formally, for solving (1), starting from any initialization x_0, we approximate K iterations of the form (2), i.e.:

$$\widehat{x}(z,\theta) \simeq \widehat{x}_K = \Phi_{z,\theta,K}(x_0) \tag{4}$$

with K_0 iterations of the form (2) repeated $T \geq 1$ times:

$$\widehat{x}^T_{K_0} = \Phi^T_{z,\theta,K_0}(x_0) := \underbrace{(\Phi_{z,\theta,K_0} \circ \cdots \circ \Phi_{z,\theta,K_0})}_{T \text{ times}}(x_0). \tag{5}$$

Based on this reformulation, we propose to perform hyperparameter selection through AD only on the last t-th iteration of Φ_{z,θ,K_0} leading to the proposed Algorithm 1.

Algorithm 1

Require: Set $K_0, T \geq 0$.
Initialize $\theta^{[0]}$ and $x_T^{[0]}$.
For $\ell = 0, 1, \ldots$
- $x_0 = x_T^{[\ell]}$
- For $t = 1, \ldots, T-1$
  - $x_t = \Phi_{z,K_0,\theta^{[\ell]}}(x_{t-1})$
- $\theta^{[\ell+1]} = \theta^{[\ell]} - \eta \partial_\theta (\mathcal{L} \circ \Phi_{z,\cdot,K_0}(x_{T-1}))(\theta^{[\ell]})$
- $x_T^{[\ell+1]} = \Phi_{z,K_0,\theta^{[\ell]}}(x_{T-1})$

The key contribution of this work is to first prove that the proposed approximation, $\widehat{x}_K \simeq \widehat{x}^T_{K_0}$, is consistent for K_0 much smaller than the number of iterations K usually needed to solve (1). The second contribution is to establish that, for T sufficiently large, x_{T-1} will be close to the fixed point, and considering recent advances in Deep Equilibrium formalism [11], [25], we can establish that the hypergradient computation thus only involves Φ_{z,θ,K_0} such that:

$$\partial_\theta \mathcal{L} \circ \Phi_{z,\theta,K}(x_0) \approx \partial_\theta \mathcal{L} \circ \Phi_{z,\cdot,K_0}(x_{T-1}) \tag{6}$$

which allows to use AD only on the last t-th iteration.

B. Fixed point analysis of restart strategy

We denote by $\mathcal{H}$ a (finite-dimensional) real Hilbert space and recall that an operator $\Phi \colon \mathcal{H} \to \mathcal{H}$ is ω-Lipschitz continuous for some $\omega \in [0,1)$ if

$$(\forall x \in \mathcal{H})(\forall y \in \mathcal{H}) \quad \|\Phi x - \Phi y\| \leq \omega \|x - y\|. \tag{7}$$

For such a class of operators, Banach-Picard theorem [26, Theorem 1.48] asserts that for $x_0 \in \mathcal{H}$ and $\mathrm{Fix}\Phi = \{\widehat{x}\}$ for some $\widehat{x} \in \mathcal{H}$, the sequence $(x_t)_{t\in\mathbb{N})}$ such that $x_{t+1} = \Phi x_t$, converges strongly to $\widehat{x}$ with linear convergence rate ω,

$$(\forall t \in \mathbb{N}) \quad \|x_t - \widehat{x}\| \leq \omega^t \|x_0 - \widehat{x}\|. \tag{8}$$

As a direct consequence of Banach-Picard theorem and Lipschitz continuity definition we obtain the following result.

Lemma 1. *Let $\mathcal{X}_0 \subset \mathcal{H}$ and $\Phi_{z,\theta,K_0} \colon \mathcal{X}_0 \to \mathcal{X}_0$ ω_{K_0}-Lipschitz with $\omega_{K_0} \in [0,1)$, then the sequence $(x_t)_{t\in\mathbb{N}}$ generated as $x_{t+1} = \Phi_{z,\theta,K_0} x_t$ converges strongly for any choice of $x_0 \in \mathcal{X}_0$ to $\mathrm{Fix}\Phi_{z,\theta,K_0}$ with linear rate ω_{K_0} and Φ^T_{z,θ,K_0} is $\omega^T_{K_0}$-Lipschitz.*

The next result gives a sufficient condition to ensure that $\widehat{x}_K \simeq \widehat{x}^T_{K_0}$ when $\lim_{K\to\infty} \omega_K \to 0$.

Proposition 1. *Let $\Phi_{z,\theta,k} \colon \mathcal{X}_0 \to \mathcal{X}_0$ ω_k-Lipschitz with $\omega_k \in [0,1)$. If there exists K_0 such that $\widehat{x} \in \mathrm{Fix}\Phi_{K_0} = \mathrm{Fix}\Phi_K$ for all $K \geq K_0$, then $\|\widehat{x}_K - \widehat{x}^T_{K_0}\| \leq (\omega^T_{K_0} + \omega_K)\|\widehat{x} - x_0\|$.*

This is a direct consequence of the following inequalities:

$$\begin{aligned}
&\|\Phi^T_{K_0}(x_0) - \Phi_K(x_0)\| \\
&\leq \|\Phi^T_{K_0}(x_0) - \widehat{x}\| + \|\widehat{x} - \Phi_K(x_0)\| \\
&\leq \omega^T_{K_0}\|x_0 - \widehat{x}\| + \omega_K\|\widehat{x} - x_0\| = (\omega^T_{K_0} + \omega_K)\|\widehat{x} - x_0\|.
\end{aligned}$$

C. Efficient computation of the hypergradient

Following [11], [12], Deep-Equilibrium framework allows us to compute hypergradients at a fixed-point $\widehat{x}$ as:

$$\frac{\partial \mathcal{L} \circ \widehat{x}(z,\cdot)}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial x}(\widehat{x})(\mathcal{J}_\theta(\widehat{x}))^{-1} \frac{\partial \Phi_{z,\cdot,K}(\widehat{x})}{\partial \theta} \tag{9}$$

$$\text{where} \quad \mathcal{J}_\theta(\widehat{x}) = \mathrm{I} - \frac{\partial \Phi_{z,\theta,K}}{\partial x}(\widehat{x}). \tag{10}$$

In practice, $\mathcal{J}_\theta$ can be challenging to invert. Recent works [25] have shown that if $\partial_x \Phi_{z,\theta,K}$ has a small operator norm, then $\mathcal{J}_\theta$ could be replaced by I leading to a much simpler implementation with similar performances as standard Deep Equilibrium. The context of Lipschitz operator with $\omega_{K_0} < 1$ fits this framework so that we use the Jacobian-free approximation [25]

$$\frac{\partial \mathcal{L}(\widehat{x}(z,\theta))}{\partial \theta} \simeq \frac{\partial \mathcal{L}}{\partial x}(\widehat{x}) \frac{\partial \Phi_{z,\theta,K_0}(x_T)}{\partial \theta} \tag{11}$$

where x_T is an approximation of the fixed point $\widehat{x}$ according to Prop. 1. An interpretation of that assumption is that the hypergradient can be computed by backpropagating AD only over the last fixed-point iteration.

Next section allows us to specify this hypergradient restart procedure to Chambolle-Pock algorithm considered in the experimental part.

III. Strongly convex Chambolle-Pock algorithm

The Chambolle-Pock algorithm proposes an iterative scheme for solving (2), whose Φ_{z,θ,K_0} is obtained by the following procedure

$$\text{For } k = 0, 1, \ldots, K_0 \quad \left\lfloor \begin{array}{ll} u_{k+1} & = \text{prox}_{\sigma_k(\lambda g)^*}(u_k + \sigma_k L \tilde{x}_k) \\ x_{k+1} & = \text{prox}_{\tau_k f(\cdot,z)}(x_k + \sigma_k L^* u_{k+1}) \\ \tilde{x}_k & = x_{k+1} + \beta_k(x_{k+1} - x_k) \end{array} \right. \tag{12}$$

where g^* denotes the Fenchel conjugate of g. The convergence relies on assumptions recalled below.

Assumption 1. *$f \in \Gamma^0(\mathcal{H}), g \in \Gamma^0(\mathcal{U})$ with $\mathcal{H}$,$\mathcal{U}$ Hilbert spaces and $L : \mathcal{H} \to \mathcal{U}$ a bounded linear operator with norm $M = ||L||_{\text{op}}$. f is γ-strongly convex (with $\gamma > 0$) and the parameters of (12) are defined as $\beta_k = 1/\sqrt{1 + 2\gamma\tau_k}; \tau_{k+1} = \beta_k\tau_k; \sigma_{k+1} = \sigma_k/\beta_k$.*

As shown in [18], under Assumption 1 and $\tau_0\sigma_0 M^2 < 1$, the sequence of primal-dual iterates $(x_k, u_k) \in \mathcal{H} \times \mathcal{U}$ converges to a solution $(\widehat{x}, \widehat{u})$ of

$$\min_x \max_u \langle Lx, u\rangle + f(x, z) - (\lambda g)^*(u). \tag{13}$$

and the sequence $(x_k)_{k\in\mathbb{N}}$ converges to a solution of (1).

This setting has the practical advantage that few regularity assumptions are required over f and g in order to minimize (1). If both f and g^* are strongly convex it has been shown [18] that $\Phi_{z,\theta,K}$ is ω-Lipschitz, with $\omega \in [0, 1)$ which allows for a straightforward application of results from the previous section since $\text{Fix}\Phi_{K_0} = \text{Fix}\Phi_K$.

However, in order to fit our application context where only f is strongly convex, the next result investigates convergence of the restarted scheme (5) when strong convexity is only assumed for f. The proof is provided in Appendix VI.

Proposition 2. *Under Assumption 1, and considering $\sigma_0 = 1/(\tau_0 M^2)$ for any $\tau_0 > 0$, then there exists some $K'(\varepsilon, \gamma\tau_0) > \frac{\gamma\tau_0}{1+\varepsilon}$ s.t. for all $K_0 \geq K'$ the following holds for any $T \geq 1$*

$$||\widehat{x} - x^T_{K_0}||^2 \leq \left(\frac{1+\varepsilon}{\gamma^2\tau_0^2 {K_0}^2}\right)^T ||\widehat{x} - x_0||^2 + \kappa||\widehat{u} - u_0||^2 \tag{14}$$

where $\kappa = \frac{\tau_0^2 M^2}{1 - \frac{1+\varepsilon}{\gamma^2\tau_0^2 {K_0}^2}}$.

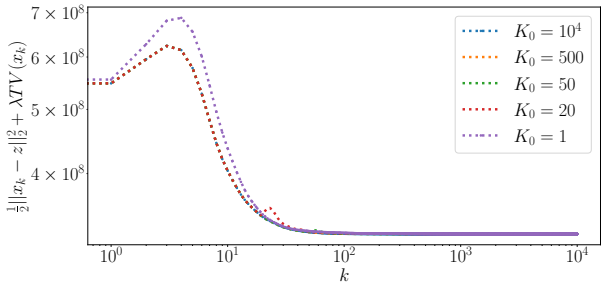
For τ_0 small enough and K_0 large enough, the restart algorithmic procedure Φ^T_{z,θ,K_0} is close to be Lipschitz. In this context, the assumptions are close to those of Prop. 1 and its application in (11) leading to Alg. 1.

The above estimate is pessimistic as the proof relies on the dual estimate initialization with the same u_0 for all T where in practice we will use u_K^{T-1} for $T > 1$ since the sequence $(u_k)_k$ is known to converge (at least weakly) to $\widehat{u}$.

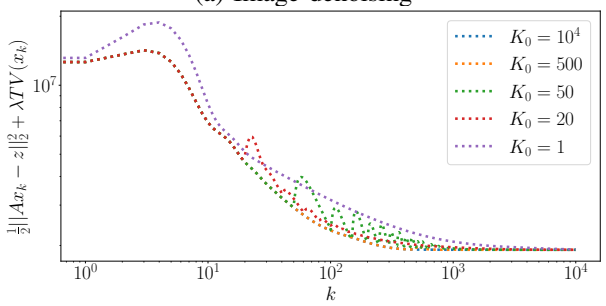
IV. Performance assessment

A. Denoising and texture segmentation framework

We assess the performance of the proposed hyperparameter restart selection procedure applied in the context of image denoising and texture segmentation, problems that can both be cast in the form of (1).



(a) Image denoising



(b) Texture segmentation

Fig. 1: Evolution of the objective function considering Φ^T_{z,θ,K_0} such that $K_0 T = 10^4$ in the context of (a) image denoising and (b) texture segmentation.

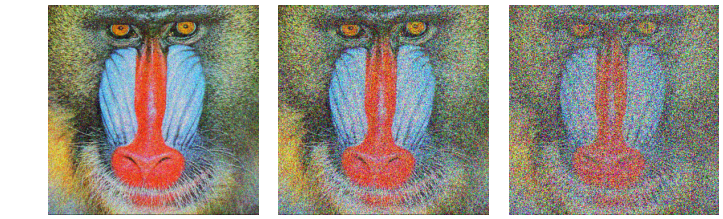
We define an input x as a map $\Omega \times \mathcal{B} \to \mathbb{R}$ involving a coordinate space denoted by a graph Ω and a set of bands $\mathcal{B} = \{1, \ldots, B\}$. For images, Ω corresponds to the lattice defined by $n = (n_x, n_y) \in \Omega = \{1, \ldots, N_x\} \times \{1, \ldots, N_y\}$. For gray-scale images $B = 1$ and for RGB images $B = 3$. The multiband isotropic total variation of an input x is defined below for $\widetilde{\Lambda} = (\widetilde{\lambda}^{(1)}, \ldots, \widetilde{\lambda}^{(B)})$ with $\widetilde{\lambda}^{(b)} \geq 0, \forall b \in \mathcal{B}$

$$\text{TV}_{\widetilde{\Lambda}}(x) = \sum_{n\in\Omega} \sqrt{\sum_{b\in\mathcal{B}} \widetilde{\lambda}^{(b)} \sum_{n'\sim n} |x_n^{(b)} - x_{n'}^{(b)}|^2}. \tag{15}$$

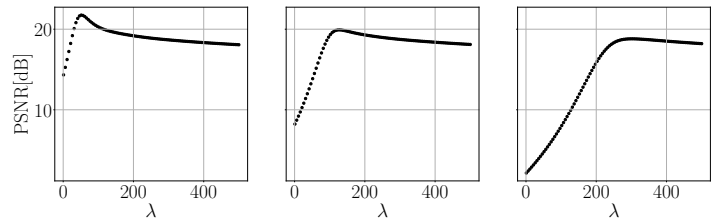
For gray-scale images and $\widetilde{\Lambda} = \mathbf{1}$ we recover the standard isotropic total variation [18]. The choice of setting $\sum_b$ inside the square-root favors simultaneous changes through all bands [27]. We may also observe that with the following linear operator $(L_\Lambda x)_n^{(b)} = \sqrt{\widetilde{\lambda}^{(b)}}(x_{n_x}^{(b)} - x_{n_x-1}^{(b)}, x_{n_y}^{(b)} - x_{n_y-1}^{(b)})$ then $\text{TV}_\Lambda(x) = ||L_\Lambda x||_{2,1}$.

For image denoising we set $f(x, z) = \frac{1}{2}||x - z||_2^2$ where z denotes the noisy version of an original image $\overline{x}$ (i.e. $z = \overline{x} + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$) and $g(x) = \|x\|_{2,1}$ where the hyperparameter to estimate is $\theta = \{\lambda\}$.

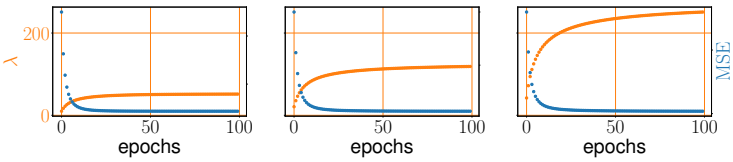
For the texture segmentation purpose, we first recall that non-smooth textures can be analyzed from multiscale local quantities $|c_{j,n}^{(b)}|$ (e.g. wavelet coefficients where j denotes the scale, n the location in space and b the band to capture anisotropy) [27], [29], [30] and more specically by considering their scale-free behaviour allows us to write that $z_{j,n}^{(b)} := \log|c_{j,n}^{(b)}| \sim v_n^{(b)} + jh_n^{(b)}$. This leads to the data-fidelity term $\frac{1}{2}||Ax - z||_2^2$ where A is the linear map defined as $(Ax)_n^{(b)} = (v_n^{(b)} + jh_n^{(b)})_{j=0}^{J-1} \in \mathbb{R}^J$ with $x_n^{(b)} = (h_n^{(b)}, v_n^{(b)})$.



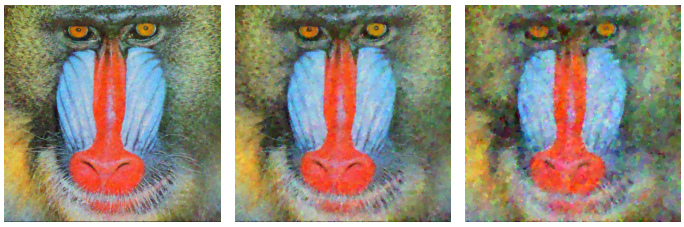
(a) $\sigma = 50$ (b) $\sigma = 100$ (c) $\sigma = 200$



(d) PSNR(λ)



(e) $\lambda^{[\ell]}$ and $\mathrm{MSE}(\widehat{x}_{\lambda^{[\ell]}} - \bar{x})$



(f) $\sigma = 50$ (g) $\sigma = 100$ (h) $\sigma = 200$

Fig. 2: Denoising. From top to bottom: noisy observations with corresponding noise; PSNR optimization landscape; loss and hyperparameter of Alg. 1 through learning; denoised estimates from Alg. 1 after 100 epochs with $T = 10, K_0 = 50$.

In that setting we consider a TV penalization as expressed in (15) and proposed in [14], [31].

The considered multiscale quantity is the undecimated dual-tree complex wavelet transform [32] which gives 6-wavelet coefficients per pixels so that there are $B = 12$ coefficients per pixel.

B. Convergence of forward mode (solving (1))

We numerically assess the convergence behaviour of the proposed primal-dual restart scheme in the context of image denoising and texture segmentation. The results are displayed in Fig 1. More precisely, we evaluate the impact of performing either $\Phi_{z,\theta,K}$ or Φ^T_{z,θ,K_0} for several choices of K_0.

For denoising, we consider the image in Fig. 2 (a) with $\sigma = 50$ and we set $\lambda = 50$. For texture segmentation we consider the clean image of Fig. 2 with $\lambda = 50$.

In Figures 1 (a)-(b), we compare the evolution of the objective function (1) when $\Phi^T_{K_0}$ described in (12) is applied for five choices of K_0 where the number of restart is s.t. $K_0 T_0 = 10^4$. In particular the CP algorithm with (resp. without) strongly convex acceleration correspond respectively to the choices $K_0 = 10^4$ (resp. $K_0 = 1$). All methods converge to the same minimal value of the criterion (1) and we also notice that although at restart points, the criterion might increase, it is overall decreasing at about the same rate as the strongly convex CP algorithm.

C. Hyperparameter estimation in the denoising framework

We apply Alg. 1 in the context of denoising and display results (Figs. 2 (f)-(h)) for 3 noise levels (Figs. 2 (a)-(c)) with same variance across channels so that we only learn $\theta = \{\lambda\}$. We consider a MSE loss $\mathcal{L}(x) = ||x - \bar{x}||_2^2$ and observe (Figs. 2 (e)) that the sequence $(\theta^{[\ell]})_\ell$ converges quickly and monotonically to an optimal value of λ in terms of PSNR (Figs. 2 (d)). Similar results are obtained when considering the unsupervised loss SURE [33].

D. Multiband hyperparameter estimation for segmentation

In order to assess the versatility of the method with respect to the choices of hyperparameters, we propose to also learn a multiband penalization $\widetilde{\Lambda}$ in (15) in the context of texture segmentation. We display in Fig. 3 (e) the convergence for $\theta = \{\lambda, (1, \dots, 1)\}$ and $\theta = \{\lambda, (\widetilde{\lambda}^{(b)})_{b \in \mathcal{B}}\}$. In both settings the values of λ converge approximately to the same value while the $\widetilde{\lambda}^{(b)}$ converge to different values corresponding to different regularization levels which are to be applied for each band $b \in \mathcal{B}$. In Fig. 3 (c) and (d) we display the results of segmentation when $\theta = \{\lambda, (1, \dots, 1)\}$ or $\theta = \{\lambda, \widetilde{\Lambda}\}$ illustrating the benefits of considering different weights through the bands. We observe that considering $\theta = \{\lambda, \widetilde{\Lambda}\}$ instead of $\theta = \{\lambda\}$ remains stable, improves performance in terms of misclassified pixels and sharpness of contours at a minimal difficulty of implementation thanks to the AD framework.

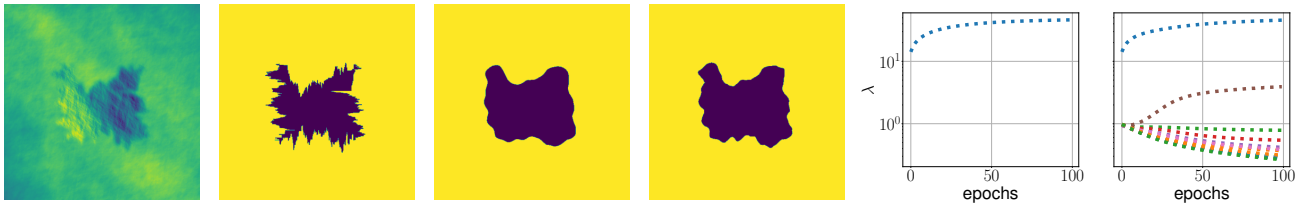
V. Conclusions and Perspectives

Our contributions leverages AD to tune hyperparameters for variational problems solved through iterative optimization algorithms. We showed that under Lipschitz continuity assumption over successive steps of iterative algorithm, their very large number of steps can be replaced by a small number of iterations restarted a large number of times. This assumption also justifies selecting hyperparameters by backpropagating AD only through the last restart. Extending this work to other iterative algorithms, possibly dictionary learning $\theta = \{\lambda, L\}$ or unrolled, might be a subject of future work.

VI. Appendix

Under the assumptions of the proposition for any $\varepsilon > 0$ there exists some K' s.t. for all $K_0 \geq K'$ [18]

$$||\widehat{x} - x_{K_0}||^2 \leq \frac{1+\varepsilon}{\gamma^2 \tau_0^2 {K_0}^2} \left(||\widehat{x} - x_0||^2 + \tau_0^2 M^2 ||\widehat{u} - u_0||^2 \right).$$



(a) Observed texture (b) Mask (c) $x_{\lambda^*,\Lambda^*\equiv I}$ (2.918%) (d) $x_{\lambda^*,\Lambda}$ (2.779%) (e) Learning $\theta = \{\lambda\}$ and $\theta = \{\lambda, \Lambda\}$

Fig. 3: Segmentation. (a) heterogeneous texture obtained from two homogeneous Anisotropic Fractional Brownian Fields [28] according to the mask (b). (c)-(d) K-class segmentations from the $(\mathbb{R}^B)^\Omega$ estimates on which apply a K-means segmentation over the set of B-dimensional features to group pixels by similarity, percentage of misclassified pixels is also reported. In Fig. (e) we display the evolution of $\{\lambda\}$ and $\{\lambda, \widetilde{\Lambda}\}$ through epochs.

Computing iteratively the bound for a number of restarts T we obtain

$$
\begin{aligned}
||\widehat{x} - x_{K_0}^T||^2 &\leq \frac{1+\varepsilon}{\gamma^2\tau_0^2{K_0}^2}\left(||\widehat{x} - x_{K_0}^{T-1}||^2 + \tau_0^2M^2||\widehat{u} - u_{K_0}^{T-1}||^2\right) \\
&\leq \frac{1+\varepsilon}{\gamma^2\tau_0^2{K_0}^2}\left(\frac{1+\varepsilon}{\gamma^2\tau_0^2{K_0}^2}(||\widehat{x} - x_K^{T-2}||^2\right. \\
&\quad \left. + \tau_0^2M^2||\widehat{u} - u_{K_0}^{T-2}||^2) + \tau_0^2M^2||\widehat{u} - u_{K_0}^{T-1}||^2\right) \\
&\leq \left(\frac{1+\varepsilon}{\gamma^2\tau_0^2{K_0}^2}\right)^T ||\widehat{x} - x_0||^2 + \tau_0^2M^2||\widehat{u} - u_0||^2 \sum_{t=1}^{T}\left(\frac{1+\varepsilon}{\gamma^2\tau_0^2{K_0}^2}\right)^t
\end{aligned}
$$

which is upper bounded by the result in (14).

REFERENCES

[1] N. Pustelnik, A. Benazza-Benhayia, Y. Zheng, and J.-C. Pesquet, "Wavelet-based image deconvolution and reconstruction," Wiley Encyclopedia of EEE, 2016.

[2] A. Chambolle and T. Pock, "An introduction to continuous optimization for imaging," Acta Numerica, vol. 25, pp. 161–319, 2016.

[3] P. Abry, N. Pustelnik, S. Roux, P. Jensen, P. Flandrin, R. Gribonval, C.-G. Lucas, É. Guichard, P. Borgnat, and N. Garnier, "Spatial and temporal regularization to estimate COVID-19 reproduction number R(t): Promoting piecewise smoothness via convex optimization," Plos one, vol. 15, no. 8, p. e0237901, 2020.

[4] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in Fixed-Point Algorithms for Inverse Problems in Science and Engineering. New York: Springer-Verlag, 2011, pp. 185–212.

[5] A. Griewank, "On automatic differentiation," Math. Program., vol. 6, no. 6, pp. 83–107, 1989.

[6] Y. Bengio, "Gradient-based optimization of hyperparameters," Neural Computation, vol. 12, no. 8, pp. 1889–1900, 2000.

[7] L. Franceschi, M. Donini, P. Frasconi, and M. Pontil, "Forward and reverse gradient-based hyperparameter optimization," in Proceedings of the 34th International Conference on Machine Learning, vol. 70. PMLR, 06–11 Aug 2017, pp. 1165–1173.

[8] M. Jiu and N. Pustelnik, "A deep primal-dual proximal network for image restoration," IEEE J. Sel. Top. Signal Process., vol. 15, no. 2, pp. 190–203, 2021.

[9] P. Nguyen, E. Soubies, and C. Chaux, "MAP-informed unrolled algorithms for hyper-parameter estimation," in IEEE ICIP, 2023, pp. 2160–2164.

[10] M. Savanier, E. Chouzenoux, J.-C. Pesquet, and C. Riddell, "Deep unfolding of the DBFB algorithm with application to ROI CT imaging with limited angular density," IEEE Trans. Comput. Imaging, 2023.

[11] S. Bai, J. Z. Kolter, and V. Koltun, "Deep equilibrium models," in NeurIPS, vol. 32. Curran Associates, Inc., 2019.

[12] D. Gilton, G. Ongie, and R. Willett, "Deep equilibrium architectures for inverse problems in imaging," IEEE Trans. Comput. Imaging, vol. 7, pp. 1123–1133, 2021.

[13] C.-A. Deledalle, S. Vaiter, J. Fadili, and G. Peyré, "Stein unbiased gradient estimator of the risk (sugar) for multiple parameter selection," SIAM J. Imaging Sci., vol. 7, no. 4, pp. 2448–2487, 2014.

[14] B. Pascal, S. Vaiter, N. Pustelnik, and P. Abry, "Automated data-driven selection of the hyperparameters for total-variation-based texture segmentation," J. Math. Imag. Vis., vol. 63, no. 7, pp. 923–952, 2021.

[15] Q. Bertrand, Q. Klopfenstein, M. Massias, M. Blondel, S. Vaiter, A. Gramfort, and J. Salmon, "Implicit differentiation for fast hyperparameter selection in non-smooth convex learning," J. Mach. Learn. Res., vol. 23, no. 149, pp. 1–43, 2022.

[16] C. Pouliquen, P. Gonçalves, M. Massias, and T. Vayer, "Implicit differentiation for hyperparameter tuning the weighted graphical lasso," 2023.

[17] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," SIAM J. Imaging Sci., vol. 2, no. 1, pp. 183–202, 2009.

[18] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," J. Math. Imag. Vis., vol. 40, no. 1, pp. 120–145, 2011.

[19] E. Chouzenoux, J.-C. Pesquet, and A. Repetti, "A block coordinate variable metric forward–backward algorithm," Journal of Global Optimization, vol. 66, no. 3, pp. 457–485, 2016.

[20] P. Parpas, "A Multilevel Proximal Gradient Algorithm for a Class of Composite Optimization Problems," SIAM J. Sci. Comput., vol. 39, no. 5, 2017.

[21] G. Lauga, E. Riccietti, N. Pustelnik, and P. Gonçalves, "Multilevel Fista For Image Restoration," IEEE ICASSP, 4-10 June 2023.

[22] I. Necoara, Y. Nesterov, and F. Glineur, "Linear convergence of first order methods for non-strongly convex optimization," Math. Program., vol. 175, pp. 69–107, 2019.

[23] J. Aujol, L. Calatroni, C. Dossal, H. Labarrière, and A. Rondepierre, "Parameter-free fista by adaptive restart and backtracking," arXiv preprint arXiv:2307.14323, 2023.

[24] V. Roulet and A. d'Aspremont, "Sharpness, restart and acceleration," in Proc. Ann. Conf. Neur. Inform. Proc. Syst., vol. 30. Curran Associates, Inc., 2017.

[25] S. W. Fung, H. Heaton, Q. Li, D. McKenzie, S. Osher, and W. Yin, "Jfb: Jacobian-free backpropagation for implicit networks," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 6, 2022, pp. 6648–6656.

[26] H. Bauschke and P. Combettes, Convex Analysis and Monotone Operator Theory in Hilbert Spaces, ser. CMS Books in Mathematics. Springer New York, 2011.

[27] B. Pascal, N. Pustelnik, and P. Abry, "Strongly convex optimization for joint fractal feature estimation and texture segmentation," Appl. Comp. Harm. Analysis, vol. 54, pp. 303–322, 2021.

[28] H. Biermé, M. Moisan, and F. Richard, "A turning-band method for the simulation of anisotropic fractional Brownian field," J. Comput. Graph. Statist., vol. 24, no. 3, pp. 885–904, 2015.

[29] J. D. B. Nelson and N. C. Kingsbury, "Dual-Tree wavelets for estimation of locally varying and anisotropic fractal dimension," in IEEE ICIP, Hong Kong, Sept. 26-29, 2010.

[30] L. Davy, N. Pustelnik, and P. Abry, "Combining dual-tree wavelet analysis and proximal optimization for anisotropic scale-free texture segmentation," in IEEE ICASSP 2023, Rhodes Island, Greece, 2023.

[31] ——, "Sélection d'hyperparamètres non supervisée par différentiation automatique, application à la segmentation de textures," no. 2023-1256. Grenoble: Proc. GRETSI, Aout 6 - Sept 9 2023, pp. p. 737–740.

[32] I. Selesnick, R. Baraniuk, and N. Kingsbury, "The Dual-Tree Complex wavelet transform," IEEE Signal Process. Mag., vol. 22, no. 6, pp. 123–151, 2005.

[33] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," Ann. Stat., vol. 9, no. 6, pp. 1135–1151, 1981.