Whiteness-based bilevel learning of regularization parameters in imaging

Carlo Santambrogio, Monica Pragliola, Alessandro Lanza, Marco Donatelli, Luca Calatroni

Whiteness-based bilevel learning of regularization parameters in imaging

Carlo Santambrogio[1,2], Monica Pragliola[3], Alessandro Lanza[4], Marco Donatelli[1], Luca Calatroni[2]
[1] *Science and High Technology Department, University of Insubria, Italy*
[2] *Laboratoire I3S, CNRS, UniCA, Inria, Sophia-Antipolis, France*
[3] *Dept. of Mathematics and Applications, University of Naples Federico II, Italy*
[4] *Dept. of Mathematics, University of Bologna, Italy*

***Abstract*—We consider an unsupervised bilevel optimization strategy for learning regularization parameters in the context of imaging inverse problems in the presence of additive white Gaussian noise. Compared to supervised and semi-supervised metrics relying either on the prior knowledge of reference data and/or on some (partial) knowledge on the noise statistics, the proposed approach optimizes the whiteness of the residual between the observed data and the observation model with no need of ground-truth data. We validate the approach on standard Total Variation-regularized image deconvolution problems which show that the proposed quality metric provides estimates close to the mean-square error oracle and to discrepancy-based principles.**

***Index Terms*—Imaging inverse problems, parameter estimation, bilevel learning, residual whiteness principle.**

I. Introduction

Variational methods are a reference paradigm in the field of ill-posed imaging inverse problems. They aim at stabilizing the unstable inversion process by minimizing a suitable energy functional encoding prior information available both on the desired image (such as sparsity, smoothness) and the noise statistics. Given a blurred, noisy and possibly incomplete (vectorized) image $\mathbf{y} \in \mathbb{R}^m$ and a linear observation model $\mathbf{A} \in \mathbb{R}^{m\times n}$, the task of retrieving a degradation-free image $\mathbf{x}^* \in \mathbb{R}^n, m \leq n$ from $\mathbf{y}$ in the presence of Additive White Gaussian Noise (AWGN) can be reformulated in variational terms as the optimization problem:

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\arg\min} \left(F(\mathbf{x}; \mathbf{A}, \mathbf{y}, \boldsymbol{\lambda}) := \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 + R(\mathbf{x}; \boldsymbol{\lambda}) \right), \quad (1)$$

where $R : \mathbb{R}^n \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$ enforces prior knowledge through regularization and $\boldsymbol{\lambda} \in \Lambda \subseteq \mathbb{R}^\ell$ is a vector of hyper-parameters tailoring the amount of regularization in terms of specific local/global features and against the quadratic data term. A popular choice for R consists in promoting sparsity w.r.t. to some specific representation of the image by choosing for $\lambda > 0$ $R(\mathbf{x}; \boldsymbol{\lambda}) = \lambda\|\boldsymbol{\Phi}\mathbf{x}\|_1$, where $\boldsymbol{\Phi} \in \mathbb{R}^{d\times n}$ or $R(\mathbf{x}; \boldsymbol{\lambda}) = \lambda\|\mathbf{D}\mathbf{x}\|_{2,1}$ where $\mathbf{D} \in \mathbb{R}^{2n\times n}$ denotes the discrete image gradient. The former choice has been the object of study in several works in the field of compressed sensing [1], while the latter has been extensively employed starting from the pioneering work [2] under the name of Total Variation (TV) regularization which is still a benchmark for model-based approaches for imaging.

Note that while the choice of a scalar $\lambda > 0$ essentially serves to balance the (global) action of the regularizer against the data term, local choices in the form $R(\mathbf{x}; \boldsymbol{\lambda}) = \sum_{j=1}^n \lambda_j r_j(\mathbf{x})$ enforces regularization at a local level weighted by a vector of space-variant parameters $\boldsymbol{\lambda}$, see, e.g., [3] for a straightforward space-variant generalization of the TV regularizer and [4] for some more sophisticated variants. Even in the scalar case, however, choosing an optimal $\lambda = \widehat{\lambda}$ is a challenging problem. Classical approaches rely on the use of cross-validation or on the (heuristic) study of the Pareto frontier as in the case of L-curve [5]. Provided that some information on the AWGN component is available (i.e., its standard deviation value $\sigma > 0$), Morozov-type approaches aim at estimating $\widehat{\lambda}$ by imposing that the solution $\mathbf{x}^*(\widehat{\lambda})$ of (1) satisfies $\|\mathbf{A}\mathbf{x}^*(\widehat{\lambda}) - \mathbf{y}\|_2^2 \approx m\sigma^2$ [6] or via more refined unbiased estimators [7] depending on σ. We remark that in practice brute-force methods are still often employed whenever ℓ is reasonably small. However, over-parametrized variational and deep-learning models often require estimations of millions of parameters which make their use prohibitive.

Bilevel learning [8]–[13] is a powerful paradigm for the estimation of optimal hyper-parameters $\widehat{\boldsymbol{\lambda}}$. There, the idea consists in optimizing a certain quality measure $\mathcal{Q}(\boldsymbol{\lambda}, \mathbf{x}^*(\boldsymbol{\lambda}))$ assessing the goodness of the solution $\mathbf{x}^*(\boldsymbol{\lambda})$ to (1) with respect to the specific task considered. In the context of imaging, standard bilevel approaches make use of classical mean-squared error (MSE) metrics mimicking SNR-type optimization [11] and/or other ones related to more perception-inspired metrics such as the Structural Similarity Index (SSIM), see, e.g. [12]. While this choice is natural, it suffers from the major problem of requiring reference ground-truth data for its assessment, which may be restrictive in applications. Other quality measures can be used within a semi-supervised framework providing optimal estimations depending only on information coming from the noise, in a discrepancy-based fashion [14].

In this work, we propose an unsupervised bilevel learning approach where optimality of $\widehat{\boldsymbol{\lambda}}$ is assessed by maximizing the whiteness of the residual between the observations $\mathbf{y}$ and the observation model $\mathbf{A}\mathbf{x}^*(\widehat{\boldsymbol{\lambda}})$ [15], [16]. As a proof of concept, we validate the approach for the case of simple TV-regularized

The authors acknowledge the support by the ANR JCJC TASKABILE grant ANR-22-CE48-0010.

image deconvolution problem in the scalar case $\boldsymbol{\lambda} = \lambda > 0$. Our preliminary results show that the proposed whiteness measure allows to achieve results almost as good as in the fully supervised and semi-supervised case, but depending only on the model $\mathbf{A}$ and the measurements $\mathbf{y}$, making it interesting for applications where the use of a large number of examples is prohibitive.

II. Methods

We review in this section the main ingredients of our approach, that is the bilevel learning paradigm for the estimation of $\widehat{\boldsymbol{\lambda}}$ in both the supervised and semi-supervised case. We then propose an unsupervised metric based on residual whiteness.

A. Hyper-parameter bilevel learning

Bilevel optimization approaches compute optimal parameters $\widehat{\boldsymbol{\lambda}}$ by solving a nested optimization problem in the form

$$\begin{aligned} \widehat{\boldsymbol{\lambda}} \in \underset{\boldsymbol{\lambda} \in \Lambda}{\arg\min} \ & \sum_{k=1}^{K} \mathcal{Q}(\mathbf{x}_k^*(\boldsymbol{\lambda})) \\ \text{s.t.} \quad & \mathbf{x}_k^*(\boldsymbol{\lambda}) \in \underset{\mathbf{x} \in \mathbb{R}^n}{\arg\min} \ F(\mathbf{x}; \mathbf{A}, \mathbf{y}_k, \boldsymbol{\lambda}), \quad k = 1, \dots, K, \end{aligned} \tag{2}$$

where, for $k = 1, \dots, K$, $\mathcal{Q}$ is a metric assessing the quality of the reconstructed images $\mathbf{x}_k^*(\boldsymbol{\lambda})$ given the measurements $\mathbf{y}_k$ (typically, some blurred, noisy and possibly undersampled image patches) w.r.t. some reference quantities. Solving (2) usually requires some smoothness assumption on the lower-level variational constraint described by F so as to allow implicit differentiation w.r.t. $\boldsymbol{\lambda}$ by means of the implicit function theorem, see, e.g. [8], [11] and Section III.

Regarding the choice of $\mathcal{Q}$, supervised approaches (see, e.g., [11], [12]) make use of pairs of exemplar ground-truth images corresponding to the measurements $\{\bar{\mathbf{x}}_k, \mathbf{y}_k\}$, $k = 1, \dots, K$ so that $\mathcal{Q}$ enforces proximity between $\mathbf{x}_k^*(\boldsymbol{\lambda})$ and $\bar{\mathbf{x}}_k$, i.e.:

$$\mathcal{Q}_{\text{MSE}}(\mathbf{x}^*(\boldsymbol{\lambda}); \bar{\mathbf{x}}) := \frac{1}{2} \|\mathbf{x}^*(\boldsymbol{\lambda}) - \bar{\mathbf{x}}\|_2^2, \tag{3}$$

which practically corresponds to choose $\boldsymbol{\lambda}$ so as to optimize the Signal to Noise Ratio (SNR) of the reconstructions.

In [14] a semi-supervised quality metric was employed to select the optimal $\widehat{\boldsymbol{\lambda}}$. Differently from (3), the quality loss employed therein relates more to a discrepancy-type measure assessing whether the residual quantity $\boldsymbol{r}(\boldsymbol{\lambda}) := \mathbf{A}\mathbf{x}^*(\boldsymbol{\lambda}) - \mathbf{y}$ has magnitude close to the noise intensity, i.e. $\|\boldsymbol{r}(\boldsymbol{\lambda})\|_2^2 \approx m\sigma^2$, so that a natural semi-supervised choice for dealing with AWGN which does not require the ground truth data $\{\bar{\mathbf{x}}_k\}$ but an estimate of σ is the Gaussianity loss:

$$\mathcal{Q}_{\text{Gauss}}(\mathbf{x}^*(\boldsymbol{\lambda}); \sigma) := \frac{1}{2} \left(\|\boldsymbol{r}(\boldsymbol{\lambda})\|_2^2 - m\sigma^2 \right)^2. \tag{4}$$

B. Residual whiteness principle

To avoid the prior knowledge of both the reference data $\bar{\mathbf{x}}_k$ and the Gaussian noise standard deviation σ, a different quality measure optimizing the whiteness of the residual $\boldsymbol{r}(\boldsymbol{\lambda})$ can be used. Employed in an heuristic fashion under the name of *residual whiteness principle* in several papers, see, e.g., [15], [16], we propose here to use such loss as an unsupervised quality metric for (2). We thus consider:

$$\mathcal{Q}_{\text{White}}(\mathbf{x}^*(\boldsymbol{\lambda})) := \frac{1}{2} \left\| \frac{\boldsymbol{r}(\boldsymbol{\lambda}) \circledast \boldsymbol{r}(\boldsymbol{\lambda})}{\|\boldsymbol{r}(\boldsymbol{\lambda})\|_2^2} \right\|_2^2, \tag{5}$$

where, with a little abuse of notation, we denote by $\mathbf{x}_1 \circledast \mathbf{x}_2$ the discrete circular cross-correlation between the matrices $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathbb{R}^{n_1 \times n_2}$ such that $\text{vec}(\boldsymbol{X}_d) = \mathbf{x}_d \in \mathbb{R}^n$, $d = 1, 2$ and $n_1 n_2 = n$ which is defined for $(j_1, j_2) \in \{0, \dots, n_1 - 1\} \times \{0, \dots, n_2 - 1\}$ by

$$(\boldsymbol{X}_1 \circledast \boldsymbol{X}_2)_{j_1, j_2} = \sum_{k_1=0}^{n_1-1} \sum_{k_2=0}^{n_2-1} (\boldsymbol{X}_1)_{k_1, k_2} (\boldsymbol{X}_2)_{(j_1+k_1) \bmod n_1, (j_2+k_2) \bmod n_2}.$$

Note that as observed in [15] the normalization term in (5) eliminates the dependence on σ.

We remark that to deal with noise scenarios different than AWGN, in [17] a whiteness measure tailored for Poisson noise was considered. We thus expect that, under suitable modifications, our proposed approach could suit to more general noise scenarios as well for tailored choice of $\mathcal{Q}_{\text{White}}$.

III. Optimization algorithms

In this section we discuss the algorithms employed to solve both the lower- and the upper-level optimization problems in (2). For simplicity we consider in the following discussion $\Lambda = \mathbb{R}_{>0}$ and $K = 1$, that is we look for an optimal positive parameter $\widehat{\lambda}$ based only on the observed image $\mathbf{y}$. In order to exploit implicit differentiation for the computation of the gradient of the bilevel problem (2), we consider a smoothed version F_ε of the ℓ_2-TV functional, defined in terms of a C^2 Huber smoothing of the TV regularization term given by:

$$H_\varepsilon(\mathbf{D}\mathbf{x}) := \|\mathbf{D}\mathbf{x}\|_{2,1,\varepsilon} = \sum_{j=1}^{n} h_\varepsilon((\mathbf{D}\mathbf{x})_j), \tag{6}$$

where $(\mathbf{D}\mathbf{x})_j \in \mathbb{R}^2$ is the j-th component of the discrete image gradient $\mathbf{D}\mathbf{x} \in \mathbb{R}^{2n}$ and $h_\varepsilon : \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$ is a C^2 Huber smoothing function defined by:

$$h_\varepsilon(\boldsymbol{v}) = \begin{cases} \frac{3}{4\varepsilon} \|\boldsymbol{v}\|_2^2 - \frac{1}{8\varepsilon^3} \|\boldsymbol{v}\|_2^4 & \text{if } \|\boldsymbol{v}\|_2 < \varepsilon \\ \|\boldsymbol{v}\|_2 - \frac{3\varepsilon}{8} & \text{if } \|\boldsymbol{v}\|_2 \geq \varepsilon \end{cases}. \tag{7}$$

The smoothed ℓ_2-TV functional

$$F_\varepsilon(\mathbf{x}; \mathbf{A}, \mathbf{y}, \lambda) := \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda H_\varepsilon(\mathbf{D}\mathbf{x}) \tag{8}$$

is thus C^2. Its optimization properties are reported in the following proposition.

Proposition III.1. *The functional F_ϵ defined in (6)-(8) is twice continuously differentiable and convex on $\mathbb{R}^n$. Moreover, if $\ker(\mathbf{A}) \cap \ker(\mathbf{D}) = \{\mathbf{0}_n\}$, F_ε is also coercive and, hence, admits a compact convex set of global minimisers. Its gradient $\nabla_{\mathbf{x}} F_\varepsilon \in \mathbb{R}^n$ and Hessian $\nabla_{\mathbf{x}}^2 F_\varepsilon \in \mathbb{R}^{n \times n}$ are given by*

$$\begin{aligned} \nabla_{\mathbf{x}} F_\varepsilon(\mathbf{x}; \mathbf{A}, \mathbf{y}, \lambda) &= \mathbf{A}^{\mathrm{T}} (\mathbf{A}\mathbf{x} - \mathbf{y}) + \lambda \mathbf{D}^{\mathrm{T}} \nabla H_\varepsilon(\mathbf{D}\mathbf{x}), \\ \nabla_{\mathbf{x}}^2 F_\varepsilon(\mathbf{x}; \mathbf{A}, \mathbf{y}, \lambda) &= \mathbf{A}^{\mathrm{T}}\mathbf{A} + \lambda \mathbf{D}^{\mathrm{T}} \nabla^2 H_\varepsilon(\mathbf{D}\mathbf{x}) \mathbf{D}, \end{aligned} \tag{9}$$

where ∇H_ε and $\nabla^2 H_\varepsilon$ denote the vector (respectively, matrix) of first-(respectively, second-)order derivatives of $H_\varepsilon(\boldsymbol{z})$

w.r.t. $\boldsymbol{z} = \mathbf{D}\mathbf{x}$. *For any* $\lambda, \varepsilon \in \mathbb{R}_{++}$, *the gradient* $\nabla_{\mathbf{x}} F_\varepsilon$ *is thus* $L_{\lambda,\varepsilon}$*-Lipschitz continuous, and* $L_{\lambda,\varepsilon}$ *can be bounded as:*

$$L_{\lambda,\varepsilon} = \max_{\mathbf{x}\in\mathbb{R}^n} \left\|\nabla^2_{\mathbf{x}} F_\varepsilon(\mathbf{x};\mathbf{A},\mathbf{y},\lambda)\right\|_2 \leq \|\mathbf{A}\|_2^2 + \frac{12\,\lambda}{\varepsilon} =: \overline{L}_{\lambda,\varepsilon}. \quad (10)$$

Proof. It follows easily from definitions (6)-(8) that F_ε is $C^2(\mathbb{R}^n)$ and convex on $\mathbb{R}^n$. Then, if $\ker(\mathbf{A}) \cap \ker(\mathbf{D}) = \{\mathbf{0}_n\}$, F_ε is coercive as both the fidelity and regularization terms are compositions of a linear map - with coefficient matrix $\mathbf{A}$ and $\mathbf{D}$, respectively - and a coercive function. It follows that F_ε admits a compact convex set of global minimizers. The gradient and Hessian expressions in (9) are derived easily by applying the chain rule of differentiation (Jacobian of composite functions). Then, based on (9) and on the sub-additivity and sub-multiplicativity properties of the spectral matrix norm, the smallest gradient Lipschitz constant $L_{\lambda,\varepsilon}$ in (10) satisfies

$$\begin{aligned} L_{\lambda,\varepsilon} &= \max_{\mathbf{x}\in\mathbb{R}^n} \left\| \mathbf{A}^{\mathrm{T}}\mathbf{A} + \lambda\, \mathbf{D}^{\mathrm{T}} \nabla^2 H_\varepsilon(\mathbf{D}\mathbf{x}) \mathbf{D} \right\|_2 \\ &\leq \|\mathbf{A}\|_2^2 + \lambda\, \|\mathbf{D}\|_2^2 \max_{\mathbf{x}\in\mathbb{R}^n} \left\|\nabla^2 H_\varepsilon(\mathbf{D}\mathbf{x})\right\|_2 \\ &\leq \|\mathbf{A}\|_2^2 + 8\,\lambda \max_{\mathbf{x}\in\mathbb{R}^n} \left\|\nabla^2 H_\varepsilon(\mathbf{D}\mathbf{x})\right\|_2 , \end{aligned} \quad (11)$$

where (11) comes from recalling that $\|\mathbf{D}\|_2^2 \leq 8$ (with $\|\mathbf{D}\|_2^2 \approx 8$) when the discrete gradient operator $\mathbf{D}$ approximates horizontal and vertical partial derivatives by means of (unscaled) standard finite differences [18]. It can be proved (we omit the proof due to the page limit) that $\nabla^2 H_\varepsilon(\mathbf{D}\mathbf{x}) \in \mathbb{R}^{2n\times 2n}$ is a real symmetric 2×2 block matrix with diagonal blocks which admits $\mathbf{x}$-dependent eigenvalue decomposition

$$\nabla^2 H_\varepsilon(\mathbf{D}\mathbf{x}) = \mathbf{V}_\varepsilon^{\mathrm{T}}(\mathbf{D}\mathbf{x})\, \mathbf{E}_\varepsilon(\mathbf{D}\mathbf{x})\, \mathbf{V}_\varepsilon(\mathbf{D}\mathbf{x}), \quad (12)$$

with orthogonal modal matrix $\mathbf{V}_\varepsilon(\mathbf{D}\mathbf{x})$ and eigenvalue matrix $\mathbf{E}_\varepsilon(\mathbf{D}\mathbf{x}) = \mathrm{diag}(e_\varepsilon^{(1)}(\mathbf{D}\mathbf{x}), \ldots, e_\varepsilon^{(2n)}(\mathbf{D}\mathbf{x}))$ satisfying

$$0 \leq e_\varepsilon^{(i)}(\mathbf{D}\mathbf{x}) \leq \frac{3}{2\,\varepsilon}, \quad \forall\, i = 1,\ldots,2n, \quad \forall\, \mathbf{x}\in\mathbb{R}^n. \quad (13)$$

It thus follows that

$$\max_{\mathbf{x}\in\mathbb{R}^n} \left\|\nabla^2 H_\varepsilon(\mathbf{D}\mathbf{x})\right\|_2 = \max_{\mathbf{x}\in\mathbb{R}^n} \left\|\mathbf{E}_\varepsilon(\mathbf{D}\mathbf{x})\right\|_2 = \frac{3}{2\,\varepsilon}, \quad (14)$$

where the last equality comes from the fact that there always exists $\mathbf{x}\in\mathbb{R}^n$ such that at least one among the eigenvalues is equal to the upper bound $3/(2\,\varepsilon)$. By replacing (14) into (11), we obtain (10). □

Remark. *We remark that for many inverse imaging problems the value* $\|\mathbf{A}\|_2^2$ *in* (10) *is known a priori or easily computable. For instance,* $\|\mathbf{A}\|_2^2 = 1$ *in image deblurring (with normalized blur functions), inpainting problems.*

We now want to make precise the optimization algorithms required to solve both the lower and the upper-level problem. To do that, and similarly as in [10], [14], we consider at first a change of variables $\lambda = \exp(\beta)$ in order to deal with an unconstrained problem depending on a parameter $\beta\in\mathbb{R}$. Upon this choice note that there holds:

$$\mathcal{Q}(\mathbf{x}^*(\lambda)) = \mathcal{Q}(\mathbf{x}^*(w(\beta))),$$

where $w : \beta \mapsto \lambda$ is the exponential function. Neglecting for simplicity the dependence of F_ε on the problem ingredients $\mathbf{A}$ and $\mathbf{y}$ and denoting for ease of notation by $\mathbf{x}^* = \mathbf{x}^*(\lambda) = \mathbf{x}^*(w(\beta))$ the solution of the lower-level problem in (2) for a fixed λ, by optimality we get:

$$\nabla_{\mathbf{x}}\, F_\varepsilon(\mathbf{x}^*; w(\beta)) = \mathbf{0}, \quad (15)$$

which, by differentiating the left-hand-side w.r.t. β and applying the chain rule entails:

$$\begin{aligned} \frac{\partial \nabla_{\mathbf{x}}\, F_\varepsilon(\mathbf{x}^*; w(\beta))}{\partial \beta} &= \frac{\partial \nabla_{\mathbf{x}} F_\varepsilon(\mathbf{x}^*; w(\beta))}{\partial w} \frac{\partial w}{\partial \beta} \\ &+ \nabla^2_{\mathbf{x}} F_\varepsilon(\mathbf{x}^*; w(\beta)) \frac{\partial \mathbf{x}^*}{\partial \beta}, \end{aligned} \quad (16)$$

whence, by (15) and the implicit function theorem, entails:

$$\frac{\partial \mathbf{x}^*}{\partial \beta} = -\left(\nabla^2_{\mathbf{x}} F_\varepsilon(\mathbf{x}^*; w(\beta))\right)^{-1} \frac{\partial \nabla_{\mathbf{x}} F_\varepsilon(\mathbf{x}^*; w(\beta)}{\partial w} \frac{\partial w}{\partial \beta},$$

which can be used for the computation of the derivative of the nested problem so as to get:

$$\begin{aligned} \frac{\partial \mathcal{Q}(\mathbf{x}^*)}{\partial \beta} &= \nabla_{\mathbf{x}} \mathcal{Q}(\mathbf{x}^*)^T \frac{\partial \mathbf{x}^*}{\partial \beta} \\ &= -\nabla_{\mathbf{x}} \mathcal{Q}(\mathbf{x}^*)^T \left(\nabla^2_{\mathbf{x}} F_\varepsilon(\mathbf{x}^*; w(\beta))\right)^{-1} \frac{\partial \nabla_{\mathbf{x}} F_\varepsilon(\mathbf{x}^*; w(\beta))}{\partial w} \frac{\partial w}{\partial \beta}, \end{aligned} \quad (17)$$

where the expression of $\nabla_{\mathbf{x}} \mathcal{Q}(\mathbf{x}^*(\lambda))$ depends on the specific expression of the assessment loss considered. Formula (17) is classically used for standard gradient-type algorithms (such as gradient-descent, quasi-Newton…) addressing the bilevel problem (2).

We now describe in Section III-A the accelerated first-order solver of the lower-level problem computing (an approximation of) $\mathbf{x}^*$ and describe in Section III-B a Gauss-Newton strategy using an expression similar to (17) to address the solution of the nested bilevel problem.

A. Accelerated gradient-descent lower-level solver

To compute (approximate) minimizers $\mathbf{x}^*$ of the smooth and convex functional F_ε in (8) we consider the Nesterov's accelerated gradient-descent algorithm, see Algorithm 1. We preferred such approach to Newton-type techniques in order to reduce the computational costs required for the inversion of (approximations of) the Hessian along the iterations. Still, to get a good approximation of $\mathbf{x}^*$ a fairly small relative tolerance parameter ϵ should be employed to assess optimality. Note that at each iteration $i \geq 1$ of the outer optimization solver, a fixed step-size $L_i = \overline{L}_{\lambda_i,\varepsilon}$ in (10) depending on the current estimate λ_i is used.

B. Gauss-Newton upper-level solver

We now describe the optimization algorithm solving the outer problem in (2) for optimizing over λ the different quality losses $\mathcal{Q}$ described above. Note that independently on the convexity of the loss function $\mathcal{Q}$ (which holds for instance both for (3) and (4)), the nested problem is generally non-convex hence only convergence to local minima $\widehat{\lambda} = \exp(\widehat{\beta})$

Algorithm 1 Nesterov AGD, $\texttt{Nesterov}_{\text{AGD}}(\mathbf{x}_0, L_i, \lambda_i, \epsilon)$

Initialize: $\theta_0 = 1, \tau = 1/L_i,\ \mathbf{x}_{-1} = \mathbf{x}_0, t = 0$
while $\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2 > \epsilon$ **do**
$\theta_{t+1} = \frac{1+\sqrt{1+4\theta_t^2}}{2}$
$\boldsymbol{z}_{t+1} = \mathbf{x}_t + \frac{\theta_t - 1}{\theta_{t+1}}(\mathbf{x}_t - \mathbf{x}_{t-1})$
$\mathbf{x}_{t+1} = \boldsymbol{z}_{t+1} - \tau \nabla_{\mathbf{x}} F_\varepsilon(\boldsymbol{z}_{t+1}; \mathbf{A}, \mathbf{y}, \lambda_i)$
$t = t + 1$
end while
return $\mathbf{x}^* = \mathbf{x}^*(\lambda_i)$

is expected. In the literature, several algorithms have addressed this task. In [11], [12], for instance, a semi-smooth Newton algorithm was employed. Here, similarly as in [14] we employ a Gauss-Newton algorithm which suits well to the squared ℓ^2-type structure of the three losses employed which can indeed be all expressed as

$$\mathcal{Q}(\mathbf{x}^*(\lambda)) = \frac{1}{2}\|\rho(\mathbf{x}^*(\lambda))\|_2^2 \tag{18}$$

for suitable choices and dimensionality of $\rho(\mathbf{x}^*(\lambda))$ depending on the choice of $\mathcal{Q} \in \{\mathcal{Q}_{\text{MSE}}, \mathcal{Q}_{\text{Gauss}}, \mathcal{Q}_{\text{White}}\}$. By (18), we observe that

$$\frac{\partial \mathcal{Q}(\mathbf{x}^*(\exp(\beta_i)))}{\partial \beta} = \rho(\mathbf{x}^*(\exp(\beta_i)))^T \boldsymbol{J}_\rho(\beta_i),$$

where $\boldsymbol{J}_\rho(\beta_i) = \frac{\partial \rho(\mathbf{x}^*(\exp(\beta_i)))}{\partial \beta}$ can be computed similarly as in (17) depending on the particular choice of ρ. We can now define in Algorithm 2 a Gauss-Newton solver for the bilevel problem (2) making explicit use of the residual function $\rho(\cdot)$. The algorithm depends on a tolerance parameter ϵ_1 which assesses stationarity and a line-search parameter $\alpha \in (0, 1)$ which could potentially be estimated on the fly by imposing, e.g., Wolfe-type conditions, but which we preferred to fix beforehand.

Algorithm 2 Gauss-Newton bilevel solver, $\text{GN}_{\text{bil}}(\beta_0, \epsilon_1, \alpha, \text{max\_it})$

Initialize: $i = 0$, $\mathbf{x}^*(w(\beta_{-1})) = \mathbf{y}$
while $\|d_i\|_2 > \epsilon_1$ and $i <$ max_it **do**
compute $\mathbf{x}^*(w(\beta_i))$=$\texttt{Nesterov}_{\text{AGD}}(\mathbf{x}^*(w(\beta_{i-1})), L_i, w(\beta_i), \epsilon)$ using (17), compute descent direction by solving:

$$d_i = -\left(\boldsymbol{J}_\rho(\beta_i)^T \boldsymbol{J}_\rho(\beta_i)\right)^{-1} \left(\boldsymbol{J}_\rho(\beta_i)^T \rho(\mathbf{x}^*(w(\beta_i)))\right)$$

update using $\beta_{i+1} = \beta_i + \alpha d_i$
$i = i + 1$
end while
return $\widehat{\lambda} = w(\widehat{\beta})$

IV. Experimental results

We now compare the proposed bilevel approaches on two exemplar image deconvolution problems with different types of blur (Gaussian, motion) and AWGN of different magnitude. The algorithmic parameters for both Algorithm 1 and 2 are chosen as $\epsilon = 10^{-6}$, $\epsilon_1 = 10^{-5}$, $\alpha = 0.1$, max_it = 60. The Huber smoothing parameter in (7) is chosen as $\varepsilon = 10^{-3}$, while the initial β-value in Algoritgm 2 has been set as $\beta_0 = 2$.

For our tests, we considered a dataset of 30 test images with size 180×180 pixels from the BSD400 repository [19]. The generic test image $\bar{\mathbf{x}}$ has been corrupted by space-invariant blur defined by a convolution kernel; the Gaussian blur kernel used has square support of side 9 pixels and standard deviation 2, while for the motion blur kernel the support size is 10 pixels and the direction angle is 60°. The generic blurred image $\mathbf{A}\bar{\mathbf{x}}$ has then been corrupted by realizations of AWGN with different magnitudes. In our set-up, the noise level is quantified by the Blurred Signal-to-Noise Ratio (BSNR) defined by:

$$\text{BSNR}(\mathbf{y}, \bar{\mathbf{x}}) = 10 \log_{10} \frac{\|\mathbf{A}\bar{\mathbf{x}} - \mathbb{E}(\mathbf{A}\bar{\mathbf{x}})\|_2^2}{\|\mathbf{A}\bar{\mathbf{x}} - \mathbf{y}\|_2^2}, \tag{19}$$

where $\mathbb{E}(\mathbf{A}\bar{\mathbf{x}})$ denotes the average intensity of the blurred image $\mathbf{A}\bar{\mathbf{x}}$. Notice that there exists a one-to-one *inverse* relationship between the BSNR and the noise standard deviation and the larger the noise the smaller the BSNR. We thus selected increasing values of $\text{BSNR} \in \{10, 17.5, 25, 32.5, 40\}$.

For each test image and BSNR value we evaluated the quality of the image reconstructed by bilevel optimization (2) of the regularization parameter λ when using the three quality metrics (3) (supervised, S), (4) (semi-supervised, SS) and (5) (unsupervised, U) in terms of average values of PSNR and SSIM computed over the set of images.

In Figure 1 we report the results obtained in the case of Gaussian blur for the estimation of a scalar parameter $\widehat{\lambda}$. We observe that compared to the MSE oracle and the semi-supervised Gaussianity loss, the proposed whiteness-based procedure provides results which are as good but do not require the use of prior information.

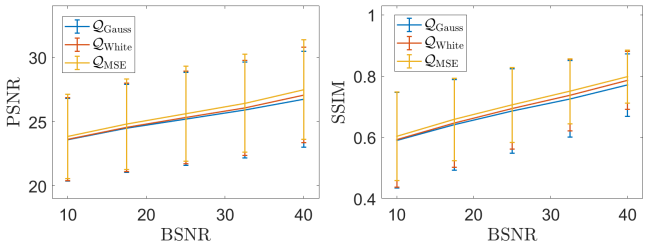

Fig. 1: Average PSNR and SSIM and dispersion bands for MSE (S), Gaussianity (SS) and Whiteness (U) loss computed over 30 test images corrupted by Gaussian blur and AWGN of different levels.

A similar behavior is observed for a TV deconvolution problem in the presence of motion blur, see Figure 2. For this second test, we report in Table I and in Figure 3 the numerical values and the visual results obtained for three different images in the dataset, respectively.

	MSE (S)		Gaussianity (SS)		Whiteness (U)	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
#1	**26.50**	**0.66**	26.03 (1.8%)	0.63 (4.1%)	26.05 (1.7%)	0.64 (3.8%)
#13	**22.67**	**0.54**	22.43 (1.1%)	0.51 (6.1%)	22.50 (0.7%)	0.52 (4.6%)
#22	**19.50**	0.65 (8.8%)	18.91 (3.0%)	0.70 (0.3%)	19.02 (2.4%)	**0.71**

TABLE I: PSNR and SSIM achieved by optimizing the MSE (S), Gaussianity (SS) and Whiteness (U) loss for the three different images in Figure 3 of the BSD400 dataset corrupted by motion blur and AWGN with BSNR=10. Percentages w.r.t. the maximum values (in bold) are reported.

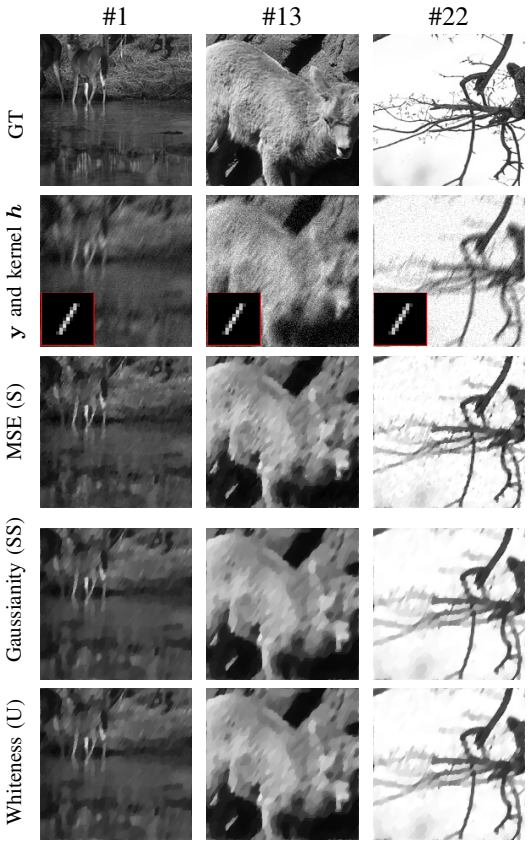

Fig. 3: From top to bottom, row-wise: ground-truth images, data $\mathbf{y}$ corrupted by motion blur and AWGN with BSNR=10, optimal reconstructions achieved by bilevel optimization of MSE (S), Gaussianity (SS) and Whiteness (U) loss.

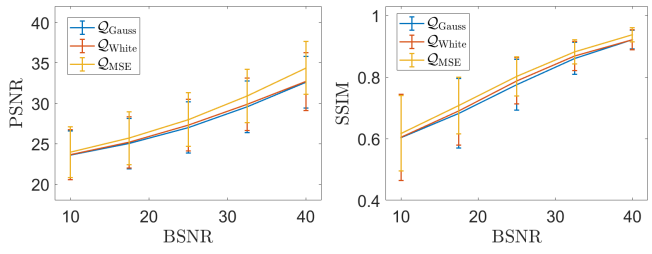

Fig. 2: Average PSNR and SSIM and dispersion bands for MSE (S), Gaussianity (SS) and Whiteness (U) loss computed over 30 test images corrupted by motion blur and AWGN of different levels.

V. Conclusions

We proposed an unsupervised bilevel learning strategy based on residual whiteness for estimating the regularization parameters in exemplar TV-regularized image deconvolution problems in the presence of AWGN. Our results suggest that such quality measure performs as well as standard MSE-based and discrepancy-type alternatives, but does not rely on any ground truth data nor noise magnitude estimation. Further work should address its use in more challenging inverse problems and comparisons with recent approaches [20] proposed in the context of deep learning for, e.g., image denoising.

References

[1] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, 2006.

[2] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phen.*, vol. 60, no. 1, 1992.

[3] L. Calatroni, A. Lanza, M. Pragliola, and F. Sgallari, "Adaptive parameter selection for weighted-TV image reconstruction problems," *Journal of Physics: Conference Series*, vol. 1476, no. 1, p. 012003, 2020.

[4] M. Pragliola, L. Calatroni, A. Lanza, and F. Sgallari, "On and beyond total variation regularization in imaging: The role of space variance," *SIAM Rev.*, vol. 65, no. 3, pp. 601–685, 2023.

[5] P. C. Hansen, "Analysis of discrete ill-posed problems by means of the L-curve," *SIAM Rev.*, vol. 34, no. 4, pp. 561–580, 1992.

[6] V. A. Morozov, "On the solution of functional equations by the method of regularization," *Doklady Mathematics*, vol. 7, pp. 414–417, 1966.

[7] J. Pesquet, A. Benazza-Benyahia, and C. Chaux, "A SURE approach for digital signal/image deconvolution problems," *IEEE Trans. Signal Process.*, vol. 57, no. 12, 2009.

[8] E. Haber and L. Tenorio, "Learning regularization functionals: a supervised training approach," *Inverse Probl.*, vol. 19, no. 3, pp. 611–626, 2003.

[9] K. G. G. Samuel and M. F. Tappen, "Learning optimized MAP estimates in continuously-valued MRF models," in *2009 CVPR*, 2009.

[10] F. Pedregosa, "Hyperparameter optimization with approximate gradient," in *ICML*, vol. 48. New York, USA: PMLR, 2016, pp. 737–746.

[11] K. Kunisch and T. Pock, "A bilevel optimization approach for parameter learning in variational models," *SIAM J. on Imaging Sci.*, vol. 6, no. 2, 2013.

[12] J. C. De los Reyes, C.-B. Schönlieb, and T. Valkonen, "Bilevel parameter learning for higher-order total variation regularisation models," *J. of Math. Imaging Vis.*, vol. 57, no. 1, 2017.

[13] C. Crockett and J. A. Fessler, "Bilevel methods for image reconstruction," *Found. Trends Signal Process.*, vol. 15, no. 2-3, 2022.

[14] J. Fehrenbach, M. Nikolova, G. Steidl, and P. Weiss, "Bilevel image denoising using gaussianity tests," in *SSVM*. Cham: Springer International Publishing, 2015, pp. 117–128.

[15] A. Lanza, M. Pragliola, and F. Sgallari, "Residual whiteness principle for parameter-free image restoration," *Electron. Trans. Numer. Anal.*, vol. 53, pp. 329–351, 2020.

[16] M. Pragliola, L. Calatroni, A. Lanza, and F. Sgallari, "ADMM-based residual whiteness principle for automatic parameter selection in single image super-resolution problems," *J. Math. Imaging Vis.*, vol. 65, no. 1, pp. 99–123, 2023.

[17] F. Bevilacqua, A. Lanza, M. Pragliola, and F. Sgallari, "Whiteness-based parameter selection for Poisson data in variational image processing," *Appl. Math. Model.*, vol. 117, pp. 197–218, 2023.

[18] A. Chambolle, "An algorithm for total variation minimization and applications," *J. Math. Imaging Vis.*, vol. 20, no. 1, pp. 89–97, 2004.

[19] D. R. Martin, C. C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," *ICCV 2001*, vol. 2, pp. 416–423 vol.2, 2001.

[20] A. Floquet, S. Dutta, E. Soubies, D.-H. Pham, and D. Kouame, "Automatic tuning of denoising algorithms parameters without ground truth," *IEEE Signal Process. Lett.*, vol. 31, pp. 381–385, 2024.