Stein Unbiased GrAdient estimator of the Risk (SUGAR) for multiple parameter selection

Charles-Alban Deledalle, Samuel Vaiter, Gabriel Peyré, Jalal M. Fadili

Stein Unbiased GrAdient estimator of the Risk (SUGAR) for multiple parameter selection*

Charles-Alban Deledalle[†¶] Samuel Vaiter[‡] Gabriel Peyré[‡] Jalal Fadili[§]

May 5, 2014

Abstract

Algorithms to solve variational regularization of ill-posed inverse problems usually involve operators that depend on a collection of continuous parameters. When these operators enjoy some (local) regularity, these parameters can be selected using the so-called Stein Unbiased Risk Estimate (SURE). While this selection is usually performed by exhaustive search, we address in this work the problem of using the SURE to efficiently optimize for a collection of continuous parameters of the model. When considering non-smooth regularizers, such as the popular ℓ_1-norm corresponding to soft-thresholding mapping, the SURE is a discontinuous function of the parameters preventing the use of gradient descent optimization techniques. Instead, we focus on an approximation of the SURE based on finite differences as proposed in [44]. Under mild assumptions on the estimation mapping, we show that this approximation is a weakly differentiable function of the parameters and its weak gradient, coined the Stein Unbiased GrAdient estimator of the Risk (SUGAR), provides an asymptotically (with respect to the data dimension) unbiased estimate of the gradient of the risk. Moreover, in the particular case of soft-thresholding, the SUGAR is proved to be also a consistent estimator. The SUGAR can then be used as a basis to perform a quasi-Newton optimization. The computation of the SUGAR relies on the closed-form (weak) differentiation of the non-smooth function. We provide its expression for a large class of iterative proximal splitting methods and apply our strategy to regularizations involving non-smooth convex structured penalties. Illustrations on various image restoration and matrix completion problems are given.

Keywords: Inverse problem, SURE, risk estimation, parameter selection, proximal splitting, sparsity, low rank

1 Introduction

In this paper, we consider the recovery problem of a signal $x_0 \in \mathcal{X}$ (where $\mathcal{X} = \mathbb{R}^N$ or is a suitable finite-dimensional Hilbert space that can be identified to $\mathbb{R}^N$) from a realization $y \in \mathcal{Y} = \mathbb{R}^P$ of the normal random vector

$$Y = \mu_0 + W \quad \text{with} \quad \mu_0 = \Phi x_0 \tag{1}$$

*This work has been supported by the European Research Council (ERC project SIGMA-Vision)

[†]IMB, CNRS-Université Bordeaux, Bordeaux, France.

[‡]CEREMADE, CNRS-Paris-Dauphine, Paris, France.

[§]GREYC, CNRS-ENSICAEN-Université de Caen, France.

[¶]Part of this work was completed while the first author was at CEREMADE[‡].

where $W \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_P)$, and the linear imaging operator $\Phi : \mathcal{X} \to \mathcal{Y}$ entails some loss of information. Typically, $P = \dim(\mathcal{X})$ is smaller than $N = \dim(\mathcal{X})$, or Φ is rank-deficient, and the recovery problem is ill-posed.

Let $(y, \theta) \mapsto x(y, \theta)$ be some recovery mapping, possibly multivalued, which attempts to approach x_0 from a given realization $y \in \mathcal{Y}$ of Y and is parametrized by a collection of continuous parameters $\theta \in \Theta$. Throughout, Θ is considered as a subset of a linear subspace of dimension $\dim(\Theta)$. We also denote $\mu(y, \theta) = \Phi x(y, \theta) \in \mathcal{Y}$ and assume in the rest of the paper that it is always a single-valued mapping though $x(y, \theta)$ may not.

Depending on the smoothness of the mapping $y \mapsto \mu(y, \theta)$, the recovered estimate enjoys different regularity properties. For instance, $\mu(y, \theta)$ can be built by solving a variational problem with some regularizing penalty parametrized by θ (see the example in (2), as well as Section 4 and 5). This regularization is generally chosen so as to preserve/promote the interesting structure underlying x_0, e.g. singularities, textures, etc.. Also, depending on its choice and that of the data fidelity, the resulting mapping $y \mapsto \mu(y, \theta)$ may be smooth or not. To cover most of these situations, throughout the paper, we will assume that $(y, \theta) \mapsto \mu(y, \theta)$ is *weakly differentiable* with respect to both the observation y, and the collection of parameters θ.

Recall that for a locally integrable function $f : a \in \Omega \mapsto \mathbb{R}$, Ω is an open subset of $\mathbb{R}^N$, its weak partial derivative with respect to a_i in Ω is the locally integrable function g_i on Ω such that

$$\int_\Omega g_i(a)\varphi(a)\mathrm{d}a = -\int_\Omega f(a)\frac{\partial \varphi(a)}{\partial a_i}\mathrm{d}a$$

holds for all functions $\varphi \in C_c^1(\Omega)$, i.e. the space of continuously differentiable functions of compact support. The weak partial derivative, if it exists, is uniquely defined Lebesgue-a.e.. Thus we write

$$g_i = \frac{\partial f}{\partial a_i}$$

and all such pointwise relations involving weak derivatives will be accordingly understood to hold Lebesgue-a.e.. A function is said to be weakly differentiable if all its weak partial derivatives exist. Similarily, a vector-valued function $h : a \in \mathbb{R}^N \mapsto h(a) = (h_1(a), \dots, h_P(a)) \in \mathbb{R}^P$ is weakly differentiable if $h_k(a)$ is weakly differentiable $\forall k \in \{1, \dots, P\}$, and we will denote $\partial h(a)$ its weak Jacobian, and $\nabla g(a) = \partial h(a)^*$ its adjoint. Remark that weak differentiation concepts boil down to the classical ones when the considered function is $\mathcal{C}^1$. A comprehensive account on weak differentiability can be found in e.g. [24, 26].

Getting back to the estimator $x(y, \theta)$, we now discuss some typical examples covered in this paper.

- Given (y, θ), consider a minimizer of a convex variational problem of the form

$$x(y, \theta) = \operatorname*{Argmin}_{x \in \mathcal{X}} \{E(x, y, \theta) = H(y, \Phi x) + R(x, \theta)\} \tag{2}$$

  where $x(y, \theta)$ is the set of minimizers of $x \mapsto E(x, y, \theta)$ which is considered nonempty (the minimizer may not be unique but is assumed to exist). The data fidelity term $x \mapsto H(y, \Phi x)$ is defined using a strongly convex map $\mu \mapsto H(y, \mu)$. The regularization term $x \mapsto R(x, \theta)$ is assumed to be a closed proper and convex function, that accounts for the prior structure of x_0. Typical priors correspond to non-smooth regularizers such as sparsity in a suitable domain, e.g. Fourier, wavelet [40], or gradient [50]. Such regularizers are usually parametrized with a collection of parameters θ. A typical example is $R(x, \theta) = \theta R_0(x)$ where $\theta \in \mathbb{R}^+$ is a scaling which controls the strength of the regularization. Of course, more

complicated (multi-parameters) regularizations, are often considered in the applications, and our methodology aims at dealing with these higher dimensional sets of parameters.

An important observation is that even though $x(y,\theta)$ may not be a singleton (minimizer of $E(x,y,\theta)$ may not be unique), strict convexity of $H(y,\cdot)$ implies that all minimizers share the same image under Φ, see e.g. [57]. Hence $(y,\theta)\mapsto\mu(y,\theta)$ is defined without ambiguity as a single-valued mapping. Moreover, strong convexity of $H(y,\cdot)$ implies that $y\mapsto\mu(y,\theta)$ is non-expansive (i.e. uniformly 1-Lipschitz) [58], hence weakly differentiable [24, Theorem 5, Section 4.2.3].

- Consider now the ℓ-th iterate, denoted by $x^{(\ell)}(y,\theta)$, of an iterative algorithm converging to a fixed point of an operator acting on $\mathcal{X}$. In this case, θ can include the parameters of the fixed point operator, as well as other continuous parameters inherent to the fixed point iteration (such as, e.g., step sizes). Section 4 is completely dedicated to this setting, and appropriate sufficient conditions will be exhibited to ensure weak differentiability of $x^{(\ell)}(y,\theta)$ with respect to both its arguments.

  This general setting encompasses the case of proximal splitting methods that have become popular to solve large-scale optimization problems of the form (2), especially with convex non-smooth terms, e.g. those encountered in sparsity regularization. The precise splitting algorithm to be used depends on the structure of the optimization problem at hand. See for instance [2, 11] for an overview. Some of these algorithms are considered in detail in Section 4.

The choice of θ is generally a challenging task, especially as the dimension of Θ gets large. Ideally, one would like to choose the parameters $\theta^\star$ that makes $\mu(y,\theta^\star)$ (or some appropriate image of it) as faithful as possible to μ_0 (or some appropriate image of it). Formally, this can be cast as selecting $\theta^\star$ that minimizes the expected reconstruction error (a.k.a., mean-square error or quadratic risk), i.e.

$$\theta^\star \in \operatorname*{Argmin}_{\theta\in\Theta} \{R^A\{\mu\}(\mu_0,\theta) = \mathbb{E}_W \|A(\mu(Y,\theta)-\mu_0)\|^2\} \tag{3}$$

where the matrix $A\in R^{M\times P}$ is typically chosen to counterbalance the effect of Φ, see Section 2.1 for a precise discussion.

If $\theta\mapsto R^A\{\mu\}(\mu_0,\theta)$ were sufficiently smooth, at least locally (e.g. Lipschitz), one could expect to solve (3) using a (sub)gradient-descent scheme relying on the (weak) gradient of the risk $\nabla_2\{R^A\{\mu\}\}(\mu_0,\theta)$, where the subscript 2 specifies that the (weak) gradient is with respect to the second argument θ. However, this would only apply if μ_0 were available. In the context of our observation model (1), μ_0 is however considered to be unknown. Our motivation is then to build an estimator of $\nabla_2\{R^A\{\mu\}\}(\mu_0,\theta)$ that depends solely on y, without prior knowledge of μ_0.

Toward this goal, we adopt the framework of the (generalized) Stein Unbiased Risk Estimator (SURE) [23, 42, 53, 57]. For a fixed θ, the celebrated Stein's lemma [53] allows to unbiasedly estimate $R^A\{\mu\}(\mu_0,\theta)$ through the weak Jacobian $\partial_1\mu(y,\theta)$, where the subscript 1 specifies that the (weak) Jacobian is with respect to the first argument y. Given such an estimator $\widehat{R}^A\{\mu\}(y,\theta)$ (see Section 2.1), the idea is so to replace the optimization problem (3) with

$$\theta^\star \in \operatorname*{Argmin}_{\theta\in\Theta} \widehat{R}^A\{\mu\}(y,\theta)\ . \tag{4}$$

It remains to find an efficient way to solve the optimization (4). Again, a (sub)gradient-descent algorithm can qualify as a good candidate if $\theta\mapsto\widehat{R}^A\{\mu\}(y,\theta)$ were sufficiently smooth again.

To our knowledge, only [15] have performed such an optimization with Newton's method where $(y,\theta) \mapsto \widehat{R}^A\{\mu\}(y,\theta)$ was C^∞. Unfortunately, being a function of $\partial_1\mu(y,\theta)$, $\theta \mapsto \widehat{R}^A\{\mu\}(y,\theta)$ is in general not differentiable, not even continuous (think of a simple soft-thresholding). This then precludes the use of standard descent schemes.

The common practice has been to apply an exhaustive search by evaluating the risk estimate $\widehat{R}^A\{\mu\}(y,\theta)$ at different values of θ. Even if in some particular cases this can be done efficiently (see for instance [19]), the computational expense can become prohibitive in general especially as $\dim(\Theta)$ increases.

Derivative-free optimization algorithms have also been investigated (see for instance [44] for the case of 2 parameters). But such approaches typically do not scale up to problems where Θ has a linear vector space structure with dimension larger than 2.

Contributions

In this paper, we address the challenging problem of solving efficiently (4): a main subject of interest for applications that has been barely investigated. Our main contribution (Section 3) is an effective strategy to optimize automatically a collection of parameters θ independently of their dimension. While classical unbiased risk estimates entail optimizing a non-continuous function of the parameters, we show that the biased risk estimator introduced in [44] is differentiable in the weak sense. This allows us, whenever the derivatives exist, to perform a quasi-Newton optimization driven by a biased estimator of the gradient of the risk based on the evaluation of $\partial_2\mu(y,\theta)$. We prove that, under mild assumptions, this estimator is asymptotically (with respect to P) unbiased, hence the name: Stein Unbiased GrAdient estimator of the Risk (SUGAR). Moreover, in the particular case of soft-thresholding, we go a step further and show that SUGAR is actually a consistent estimator of the gradient of the risk.

As a second contribution (Section 4), we propose a versatile approach to compute the derivatives $\partial_1\mu(y,\theta)$ and $\partial_2\mu(y,\theta)$, involved respectively in the computation of the SURE and SUGAR, when $\mu(y,\theta)$ is computed through an iterative algorithm, typically proximal splitting methods. We illustrate the versatility of our method by applying it to both primal (forward-backward [12], Douglas-Rachford [10] and generalized foward-backward [43]) and primal-dual [8] algorithms. The proposed methodology can however be adapted to any other proximal splitting method and more generally to any algorithm whose iteration operator is weakly differentiable.

Numerical simulations involving multi-parameter selection for image restoration and matrix completion problems are reported in Section 5. The proofs of our results are collected in the appendix.

2 Overview on Risk Estimation

This section gives an overview of the literature to estimate the risk via the SURE and its variants for ill-posed inverse problems contaminated by additive white Gaussian noise.

2.1 Stein Unbiased Risk Estimator

Degrees of freedom (DOF) is often used to quantify the complexity of a statistical modeling procedure, see for instance, GCV (generalized cross-validation [29]). From [22,57], the degrees

of freedom of a function $y \mapsto \mu(y,\theta)$ relatively to a matrix $A \in R^{M\times P}$ is given by

$$df^A\{\mu\}(\mu_0,\theta) = \sum_{i=1}^{P} \frac{\operatorname{cov}(AY_i, (A\mu(Y,\theta))_i)}{\sigma^2} \ , \tag{5}$$

such that $df^A\{\mu\}(\mu_0,\theta)$ is maximal when $A\mu(Y,\theta)$ is highly correlated to the random vector AY. Taking $A = \mathrm{Id}$, leads to the standard definition of the DOF defined in the seminal work of Efron [22]. But other choices of A allow to counterbalance the undesirable effect of the linear operator Φ (recall that $\mu_0 = \Phi x_0$). For instance, setting $A = (\Phi^*\Phi)^{-1}\Phi^*$ when Φ has full-rank, or $A = \Phi^*(\Phi\Phi^*)^+$ when Φ is rank deficient*, provides a measure of the DOF relatively to the least-squares estimate of x_0, i.e. $x_{\mathrm{LS}}(y) = Ay$ [23, 42, 57].

With the proviso that $y \mapsto \mu(y,\theta)$ is weakly differentiable with essentially bounded weak partial derivatives, an unbiased estimate of the DOF can be used to unbiasedly estimate the risk in (3). This leads to the (generalized) SURE (also known as weighted SURE [46]) given as

$$\begin{aligned} &\mathrm{SURE}^A\{\mu\}(y,\theta) = \|A(\mu(y,\theta)-y)\|^2 - \sigma^2 \operatorname{tr}(A^*A) + 2\sigma^2 \widehat{df}^A\{\mu\}(y,\theta) \qquad (6)\\ &\text{with} \quad \widehat{df}^A\{\mu\}(y,\theta) = \operatorname{tr}\left(A\partial_1\mu(y,\theta)A^*\right) \end{aligned}$$

where we recall that $\partial_1\mu(y,\theta)$ is the weak Jacobian of $\mu(y,\theta)$. It can be shown that (see e.g. [23, 57])

$$\mathbb{E}_W[\widehat{df}^A\{\mu\}(Y,\theta)] = df^A\{\mu\}(\mu_0,\theta) \quad \text{and} \quad \mathbb{E}_W[\mathrm{SURE}^A\{\mu\}(Y,\theta)] = R^A\{\mu\}(\mu_0,\theta) \ .$$

Expression (6) is general enough to encompass unbiased estimates of the *prediction* risk $\mathbb{E}_W \|\mu(Y,\theta)-\mu_0\|^2$ (i.e. $A = \mathrm{Id}$), the *projection* risk $\mathbb{E}_W \|\Pi(x(Y,\theta)-x_0)\|^2$, where Π is the orthogonal projector on $\ker(\Phi)^\perp$ (i.e. $A = \Phi^*(\Phi\Phi^*)^+$), and the *estimation* risk $\mathbb{E}_W \|x(Y,\theta)-x_0\|^2$ when Φ has full rank (i.e. $A = (\Phi^*\Phi)^{-1}\Phi^*$). This can prove useful when Φ is rank deficient, since in this case, the minimizers of the prediction risk can be far away from the minimizers of the estimation risk [47]. The projection risk restricts the estimate to the subspace where there is a signal beside noise, and in this sense, is a good approximation of the estimation risk [23].

Applications of SURE emerged for choosing the smoothing parameters in families of linear estimates [38] such as for model selection, ridge regression, smoothing splines, etc. After its introduction in the wavelet community with the SURE-Shrink algorithm [19], it has been widely used to various image restoration problems, e.g. with sparse regularizations [4, 5, 9, 39, 42, 44–46, 61] or with non-local filters [14, 20, 59, 60].

However, a major practical difficulty when using the SURE lies in the numerical computation of the DOF estimate, i.e. the quantity $\widehat{df}^A\{\mu\}(y,\theta)$ for a given realization y. We now give a brief overview of some previous works to deal with this computation.

2.2 Closed-form SURE

The SURE is based on a DOF estimate $\widehat{df}^A\{\mu\}(Y,\theta)$ that can be sampled from the observation $y \in \mathbb{R}^P$ by evaluating the Jacobian $\partial_1\mu(y,\theta) \in \mathbb{R}^{N\times P}$. A natural way, to evaluate $\partial_1\mu(y,\theta)$ would be to derive its closed-form expression. This has been studied for some classes of variational problems.

*$(\cdot)^+$ stands for the Moore-Penrose pseudo-inverse.

In quadratic regularization (e.g. ridge regression), where solutions are of the form $x(y,\theta) = K(\theta)y$, where $K(\theta)$ is known as the hat or influence matrix, the Jacobian has a closed-form $\partial_1\mu(y,\theta) = \Phi K(\theta)$. In ℓ^1-synthesis regularization (a.k.a. the LASSO), the Jacobian matrix depends on the support (set of non-zero coefficients) of any LASSO solution $x(y,\theta)$. An estimator of the DOF can then be retrieved from the number of non-zero entries of this solution [34, 56, 64]. These results have in turn been extended to more general sparsity promoting regularizations [35, 52, 55–57, 63], and spectral regularizations (e.g. nuclear norm) [7, 17].

This approach however has three major bottlenecks. First, deriving the closed-form expression of the Jacobian is in general challenging and has to be addressed on a case by case basis. Second, in large-dimensional problems, evaluating numerically this Jacobian is barely possible. Even if it were possible, it might be subject to serious numerical instabilities. Indeed, solutions of variational problems are achieved via iterative schemes providing iterates $x^{(\ell)}(y,\theta)$ that eventually converge to the set of solutions as $\ell \to +\infty$. And yet, for instance, substituting the support of the true solution by the support of $x^{(\ell)}(y,\theta)$, obtained at a prescribed convergence accuracy, might be imprecise (all the more since the problem is ill-conditioned).

The three next sections review previous work to address one or some of these three points.

2.3 Monte-Carlo SURE

To deal with the large dimension of the Jacobian, the standard approach is to exploit the fact that the DOF only depends on the trace of $A\partial_1\mu(y,\theta)A^*$. In denoising applications where $A = \mathrm{Id}$ and $\Phi = \mathrm{Id}$, this trace can generally be obtained by closed-form computations of the P diagonal elements of $\partial_1\mu(y,\theta)$ (see e.g. [19, 59]). This can also be done for some particular inverse problems. For instance, the authors of [42] provide an expression of this trace for the wavelet-vaguelette estimator when Φ is a convolution matrix and $A = \Phi^+$. However, in more general settings, the complexity of the closed-form computation of the trace is non-linear, typically the number of operations is in $O(P \times P)$ (think of Φ a mixing operator or μ an iterative estimator). To avoid such a costly procedure, the authors of [27, 44] suggest making use of the following trace equality

$$\widehat{df}^{A}\{\mu\}(y,\theta) = \mathrm{tr}\left(A\partial_1\mu(y,\theta)A^*\right) = \mathbb{E}_\Delta\left\langle \partial_1\mu(y,\theta)[\Delta],\, A^*A\Delta\right\rangle \tag{7}$$

where $\Delta \sim \mathcal{N}(0, \mathrm{Id}_P)$ and $\partial_1\mu(y,\theta)[\delta] \in \mathbb{R}^P$ denotes the directional derivative of $y \mapsto \mu(y,\theta)$ at y in direction δ. Remark that Δ does not necessary have to be Gaussian and higher precisions can be reached in some specific cases, see for instance [1, 18, 33, 49]. As shown in [49], the performance of this trace estimator is governed by the distribution of the singular values of the operator $A\partial_1\mu(y,\theta)A^*$. More specifically, the slower the decay, the better the performance. While it is difficult to make a general claim, we observed numerically that for the recovery problems we consider, it provide a very accurate estimator of the trace. Hence, following [44,61], an estimate of $\mathrm{SURE}^A\{\mu\}(y,\theta)$ can be obtained by Monte-Carlo simulations using

$$\begin{aligned} &\mathrm{SURE}^A_{\mathrm{MC}}\{\mu\}(y,\theta,\delta) = \|A(\mu(y,\theta)-y)\|^2 - \sigma^2\,\mathrm{tr}(A^*A) + 2\sigma^2\widehat{df}^{A}_{\mathrm{MC}}\{\mu\}(y,\theta,\delta) \\ \text{with}\quad &\widehat{df}^{A}_{\mathrm{MC}}\{\mu\}(y,\theta,\delta) = \left\langle \partial_1\mu(y,\theta)[\delta],\, A^*A\delta\right\rangle \ . \end{aligned} \tag{8}$$

The evaluation of (8) necessitates only computing the P entries of $\partial_1\mu(y,\theta)[\delta]$.

It remains to find a stable and efficient way to evaluate for any vector $\delta \in \mathbb{R}^P$ the directional derivative $\partial_1\mu(y,\theta)[\delta] \in \mathbb{R}^P$.

2.4 Iterative Differentiation for Monte-Carlo SURE

When considering solutions $x(y,\theta)$ of a variational problem, the DOF cannot be robustly estimated if one knows only the iterates $\mu^{(\ell)}(y,\theta)$ that eventually converge to some $\mu(y,\theta)$ as $\ell \to +\infty$. It appears then natural to estimate the DOF of $\mu^{(\ell)}(Y,\theta)$ directly and make the assumption that it will converge to that of $\mu(y,\theta)$. For a realization $y \in \mathbb{R}^P$, one can sample an estimate of the DOF of the iterate $\mu^{(\ell)}(Y,\theta)$ by evaluating its directional derivative $\partial_1\mu^{(\ell)}(y,\theta)[\delta]$. A practical way, initiated by [61], to compute this quantity, consists in recursively differentiating the sequence of iterates. The authors of [61] have derived the closed-form expression of the directional derivative for the Forward-Backward (FB) algorithm. The directional derivative at iteration $\ell+1$, denoting by $\mathcal{D}_\mu^{(\ell+1)} = \partial_1\mu^{(\ell+1)}(y,\theta)[\delta]$, is obtained iteratively as a function of $\mu^{(\ell)}(y,\theta)$ and $\mathcal{D}_\mu^{(\ell)} = \partial_1\mu^{(\ell)}(y,\theta)[\delta]$. The Monte-Carlo DOF and the Monte-Carlo SURE can in turn be iteratively estimated by plugging $\partial_1\mu^{(\ell)}(y,\theta)[\delta]$ in (8) leading to

$$\begin{aligned} &\mathrm{SURE}^A_{\mathrm{MC}}\{\mu^{(\ell)}\}(y,\theta,\delta) = \left\|A(\mu^{(\ell)}(y,\theta)-y)\right\|^2 - \sigma^2\,\mathrm{tr}(A^*A) + 2\sigma^2\widehat{df}^A_{\mathrm{MC}}\{\mu^{(\ell)}\}(y,\theta,\delta) \\ \text{with}\quad &\widehat{df}^A_{\mathrm{MC}}\{\mu^{(\ell)}\}(y,\theta,\delta) = \left\langle \mathcal{D}_\mu^{(\ell)},\, A^*A\delta \right\rangle \ . \end{aligned} \tag{9}$$

A similar approach is described in [28]. Pursuing this idea, the authors of [45,46] have recently provided such closed-form expressions in the case of the split Bregman method. Concurrently, in an early short version of this paper [16], we have also considered this approach for general proximal splitting algorithms, an approach that we extend in Section 4.

2.5 Finite-Difference SURE

An alternative initiated in [51, 62] and rediscovered in [44] consists in estimating $\mathrm{tr}\,(A\partial_1\mu(y,\theta)A^*)$ via finite differences given, for $\varepsilon > 0$, by

$$\mathrm{tr}\,(A\partial_1\mu(y,\theta)A^*) \approx \sum_{i=1}^{P} \frac{[A^*A(\mu(y+\varepsilon e_i,\theta)-\mu(y,\theta))]_i}{\varepsilon}\ , \tag{10}$$

where $(e_i)_{1\leqslant i\leqslant P}$ is the canonical basis of $\mathbb{R}^P$. Plugging this expression in (8) yields the Finite-Difference (FD) SURE^A given by

$$\begin{aligned} &\mathrm{SURE}^A_{\mathrm{FD}}\{\mu\}(y,\theta,\varepsilon) = \|A(\mu(y,\theta)-y)\|^2 - \sigma^2\,\mathrm{tr}(A^*A) + 2\sigma^2\widehat{df}^A_{\mathrm{FD}}\{\mu\}(y,\theta,\varepsilon) \\ \text{with}\quad &\widehat{df}^A_{\mathrm{FD}}\{\mu\}(y,\theta,\varepsilon) = \frac{1}{\varepsilon}\sum_{i=1}^{P}(A^*A(\mu(y+\varepsilon e_i,\theta)-\mu(y,\theta)))_i. \end{aligned} \tag{11}$$

The main advantage of this method is that $(y,\theta) \mapsto \mu(y,\theta)$ can be used as a black-box, i.e., without knowledge on the underlying algorithm that provides $\mu(y,\theta)$, while, for ε small enough, it performs as well as the approach described in Section 2.4 that requires the knowledge of the derivatives in closed-form. In fact, if $y \mapsto \mu(y,\theta)$ is Lipschitz-continuous, then it is differentiable Lebesgue a.e. (Rademacher's theorem), and its derivative equals its weak derivative Lebesgue a.e. [24, Theorem 1-2, Section 6.2], which in turn implies

$$\lim_{\varepsilon\to 0} \mathrm{SURE}^A_{\mathrm{FD}}\{\mu\}(y,\theta,\varepsilon) = \mathrm{SURE}^A\{\mu\}(y,\theta) \quad \text{Lebesgue a.e.} \tag{12}$$

The value ε can so be chosen as small as possible (up to machine precision) yielding to a quasi unbiased risk estimator (i.e., with a negligible bias). It remains that when the data

dimension P is large, the evaluation of P finite differences along each axis might be numerically intractable. In that case, the Monte-Carlo approach (see Section 2.3) can also be used in conjunction with finite differences leading to the Finite-Difference Monte-Carlo (FDMC) SURE^A given by

$$\mathrm{SURE}^A_{\mathrm{FDMC}}\{\mu\}(y,\theta,\delta,\varepsilon) = \|A(\mu(y,\theta)-y)\|^2 - \sigma^2\,\mathrm{tr}(A^*A) + 2\sigma^2 \widehat{df}^A_{\mathrm{FDMC}}\{\mu\}(y,\theta,\delta,\varepsilon)$$

$$\text{with}\quad \widehat{df}^A_{\mathrm{FDMC}}\{\mu\}(y,\theta,\delta,\varepsilon) = \frac{1}{\varepsilon}\left\langle \mu(y+\varepsilon\delta,\theta)-\mu(y,\theta),\, A^*A\delta\right\rangle . \tag{13}$$

The originality of our approach described in the next section is to devise a grounded choice of $\varepsilon > 0$. This introduces a bias in the estimation of the risk. Nevertheless, as we will see, using $\varepsilon > 0$ plays an important role in risk optimization since, unlike $\mathrm{SURE}^A\{\mu\}$, $\mathrm{SURE}^A_{\mathrm{FD}}\{\mu\}$ is a smooth function of θ in the weak sense. This is the key point to optimize the risk. By choosing $\varepsilon > 0$ carefully, a smoother objective function can be used as a basis to perform a quasi-Newton-like optimization at the expense of a controled bias.

3 Risk Estimate Minimization

In this section, we investigate how risk estimates can be used for optimizing a collection of continuous parameters.

3.1 Stein's Unbiased GrAdient Risk (SUGAR) Estimator

The difficulty is that even if $\theta \mapsto R^A\{\mu\}(\mu_0,\theta)$ is differentiable in the weak sense, the function $\theta \mapsto \mathrm{SURE}^A\{\mu\}(y,\theta)$ might contain discontinuities. Typically, $\widehat{df}^A\{\mu\}(y,\theta)$ has discontinuities where $(y,\theta) \mapsto \mu(y,\theta)$ is not differentiable.

We start with a simple result showing that unlike $\mathrm{SURE}^A\{\mu\}(y,\theta)$, the finite-difference based mapping $\theta \mapsto \mathrm{SURE}^A_{\mathrm{FD}}\{\mu\}(y,\theta,\varepsilon)$, for $\varepsilon > 0$, is weakly differentiable.

Proposition 1. *Assume $\mu(y,\theta)$ is weakly differentiable with respect to y and θ. Given $\varepsilon > 0$, $\widehat{df}^A_{\mathrm{FD}}\{\mu\}(y,\theta,\varepsilon)$ and $\mathrm{SURE}^A_{\mathrm{FD}}\{\mu\}(y,\theta,\varepsilon)$ are also weakly differentiable with respect to y and θ, and their (weak) gradients with respect to θ are given, for almost all $\theta \in \Theta$, as*

$$\begin{aligned}
\mathrm{SUGAR}^A_{\mathrm{FD}}\{\mu\}(y,\theta,\varepsilon) &= \nabla_2\{\mathrm{SURE}^A_{\mathrm{FD}}\{\mu\}\}(y,\theta,\varepsilon)\\
&= 2\partial_2\mu(y,\theta)^*A^*A(\mu(y,\theta)-y) + 2\sigma^2\nabla_2\{\widehat{df}^A_{\mathrm{FD}}\{\mu\}\}(y,\theta,\varepsilon)\\
\textit{where}\quad \nabla_2\{\widehat{df}^A_{\mathrm{FD}}\{\mu\}\}(y,\theta,\varepsilon) &= \frac{1}{\varepsilon}\sum_{i=1}^{P}\left(\partial_2\mu(y+\varepsilon e_i,\theta)-\partial_2\mu(y,\theta)\right)^*A^*Ae_i\ .
\end{aligned}$$

Thanks to Proposition 1, a quasi-Newton-like method can now be used to optimize $\mathrm{SURE}^A_{\mathrm{FD}}\{\mu\}(y,\theta,\varepsilon)$ for the vector of continuous parameters θ by implementing the iteration

$$\theta_{n+1} = \theta_n - B_n\mathrm{SUGAR}^A_{\mathrm{FD}}\{\mu\}(y,\theta_n,\varepsilon)$$

where $B_n \in \mathbb{R}^{\dim(\Theta)\times\dim(\Theta)}$ is a sequence of definite-positive matrices. Typically, if $\theta \mapsto \mathrm{SURE}^A_{\mathrm{FD}}\{\mu\}(y,\theta,\delta,\varepsilon)$ behaves locally as a C^2 function, B_n should approach the inverse of the corresponding Hessian at θ_n. In practice, the calculation of $\mathrm{SUGAR}^A_{\mathrm{FD}}$ depends on the computation of the Jacobian matrices with respect to the parameters θ. We will see in Section 4 how this quantity can be efficiently computed when μ results from an iterative algorithm.

We now turn to the asymptotic unbiasedness of SUGAR as ε approaches 0. Toward this goal we need the following assumptions.

(A.1) The mapping $y \mapsto \mu(y, \theta)$ is uniformly Lipschitz continuous with Lipschitz constant L_1.

(A.2) The mapping $y \mapsto \mu(y, \theta)$ is such that $\mu(0, \theta) = 0$ for any θ.

(A.3) The mapping $\theta \mapsto \mu(y, \theta)$ is uniformly Lipschitz continuous with Lipschitz constant L_2 independently of y.

Remark 1 (Discussion of the assumptions)**.** *1. Assumption* **(A.1)** *is mild and is fulfilled in many situations of interest. In particular, this is the case when $y \mapsto \mu(y, \theta)$ is the proximal operator of proper closed and convex function, which is at the heart of Section 4 (see also Section 3.2 for soft-thresholding). Standard convex analysis arguments [32] show that the proximity operator is indeed a uniformly Lipschitz of its argument y with constant $L_1 = 1$, independently of θ.*

2. Assumption **(A.2)** *is very natural and does not entail any loss of generality. It basically states that, when the observations are zero, so is the estimator.*

3. As far as Assumption **(A.3)** *is concerned, it is verified under certain circumstances. This is for instance the case when $\mu(y, \theta) = \mathrm{Prox}_{\theta G}(y)$, $\theta > 0$, where G is the gauge (see Definition 2 in Appendix B) of any compact convex set containing the origin as an interior point[†]; see Proposition 5 in Appendix B. By induction, this also holds when $\mu(y, \theta) = \mathrm{Prox}_{\theta_1 G_1} \circ \cdots \circ \mathrm{Prox}_{\theta_m G_m}(y)$, $\theta \in]0, +\infty[^m$, and for any $i = 1, \ldots, m$, G_i is the gauge of any compact convex set containing the origin as an interior point, see Corollary 5. Typical instances of these gauges are norms, e.g. ℓ_1, $\ell_1 - \ell_2$ or nuclear norms very popular now in the signal and image processing community.*

4. Assumption **(A.3)** *can be relaxed to cover the case where L_2 depends on y. In such a situation, additional assumptions on the function $y \mapsto L_2(y)$ are needed for steps 3) and 4) in the proof of Theorem 1 to go through. We omit this case for the sake of clarity and to avoid further technicalities.*

We are now ready to state our theorem.

Theorem 1 (Asymptotic unbiasedness of SUGAR)**.** *Assume that* **(A.1)**-**(A.3)** *hold. Then, $R^A\{\mu\}(\mu_0, \cdot)$ and $df^A\{\mu\}(\mu_0, \cdot)$ are weakly differentiable, and for any Lebesgue point θ,*

$$\lim_{\varepsilon \to 0} \mathbb{E}_W \left[\mathrm{SUGAR}^A_{\mathrm{FD}}\{\mu\}(Y, \theta, \varepsilon) \right] = \nabla_2 \{R^A\{\mu\}\}(\mu_0, \theta)$$

$$\textit{and} \quad \lim_{\varepsilon \to 0} \mathbb{E}_W \left[\nabla_2 \{\widehat{df}^A_{\mathrm{FD}}\{\mu\}\}(Y, \theta, \varepsilon) \right] = \nabla_2 \{df^A\{\mu\}\}(\mu_0, \theta) \ .$$

Theorem 1 can be given the following interpretation. As ε gets close to 0, e.g. a decreasing function of the dimension P[‡], the gradient of $\mathrm{SURE}^A_{\mathrm{FD}}\{\mu\}(y, \cdot, \varepsilon)$ (normalized by P) can be used to estimate the gradient of the risk (also normalized by P) provided that P is large enough.

[†]Another case which is trivial corresponds to G being the indicator function of a non-empty closed convex set, in which case $L_2 = 0$.

[‡]as we will see, the higher the dimension P, the smaller ε could be.

However, even if ε should decrease towards 0, it should not decrease too fast. In particular, for a fixed dimension P, the step ε cannot be chosen arbitrarily small. This would not be an issue if $\mu(y,\cdot)$ were differentiable, but in general, there might be singularities. In fact, for a finite dimension P, the limit when $\varepsilon \to 0$ of the sample $\mathrm{SUGAR}^A_{\mathrm{FD}}\{\mu\}(y,\theta,\varepsilon)$ may not even exist, though that of its expectation does exist Lebesgue a.e. as shown in the proof of Theorem 1. As a consequence, the quantity $\frac{1}{P}\mathrm{SUGAR}^A_{\mathrm{FD}}\{\mu\}(Y,\theta,\varepsilon)$ can become very unstable when ε decreases too fast with the dimension P. The underlying statistical question is whether one can control the variance of $\frac{1}{P}\mathrm{SUGAR}^A_{\mathrm{FD}}\{\mu\}(Y,\theta,\varepsilon)$ as P increases, and make arbitrarily small or even asymptotically vanishing, to that $\frac{1}{P}\mathrm{SUGAR}^A_{\mathrm{FD}}\{\mu\}(Y,\theta,\varepsilon)$ becomes a consistent estimator. Unfortunatelly, consistency of SUGAR is very intricate to get in the general case, as it is the case for the consistency of the SURE. However, when μ specializes to soft-thresholding, such a result can be achieved.

3.2 SUGAR for Soft-tresholding

In this section, we show that SUGAR can consistently estimate the gradient of the risk in the case where μ is the soft-thresholding (ST) function and $A = \mathrm{Id}_P$. The ST is the proximal operator of the ℓ_1-norm. Understanding the ST is of chief interest since it is at the heart of any proximal splitting algorithm solving a regularized inverse problem involving terms of the form $\|D^*x\|_1$ where D is a linear operator.

Let first recall the definition of soft-thresholding.

Definition 1 (Soft-Thresholding)**.** *The soft-thresholding (ST) is defined, for $\lambda > 0$, and for all $1 \leqslant i \leqslant P$, as*

$$\mathrm{ST}(y,\lambda)_i = \begin{cases} y_i + \lambda & \text{if} \quad y_i \leqslant -\lambda \\ 0 & \text{if} \quad -\lambda < y_i < \lambda \\ y_i - \lambda & \text{otherwise} \end{cases} . \tag{14}$$

Observe that as a proximity operator of a norm, soft-thresholding satisfies Assumptions **(A.1)** through **(A.3)** of Theorem 1, see the corresponding discussion. Hence, we already anticipate from Theorem 1 that SUGAR provides an asymptotically unbiased estimate of soft-thresholding risk gradient.

We start with following lemma which collects the statistics of the gradient of the finite difference DOF estimator which will be at the heart of the next results.

Lemma 1 (Statistics of the gradient of the finite difference DOF estimator)**.** *Let $0 < \varepsilon < 2\lambda$. The weak gradient of $\lambda \mapsto \widehat{df}_{\mathrm{FD}}\{\mathrm{ST}\}(Y,\lambda,\varepsilon)$ is such that*

$$\mathbb{E}_W\left[\nabla_2\{\widehat{df}_{\mathrm{FD}}\{\mathrm{ST}\}\}(Y,\lambda,\varepsilon)\right] = \frac{-1}{2}\sum_{i=1}^{P}\frac{\varphi[(\mu_0)_i,\lambda,\varepsilon]}{\varepsilon},$$

$$\textit{and} \quad \mathbb{V}_W\left[\nabla_2\{\widehat{df}_{\mathrm{FD}}\{\mathrm{ST}\}\}(Y,\lambda,\varepsilon)\right] = \frac{1}{2\varepsilon}\sum_{i=1}^{P}\frac{\varphi[(\mu_0)_i,\lambda,\varepsilon]}{\varepsilon} - \frac{1}{4}\sum_{i=1}^{P}\left[\frac{\varphi[(\mu_0)_i,\lambda,\varepsilon]}{\varepsilon}\right]^2 .$$

where for $a \in \mathbb{R}$, $\varphi[a,\lambda,\varepsilon] = \mathrm{erf}\left(\frac{a+\lambda+\varepsilon}{\sqrt{2}\sigma}\right) - \mathrm{erf}\left(\frac{a+\lambda}{\sqrt{2}\sigma}\right) + \mathrm{erf}\left(\frac{a-\lambda+\varepsilon}{\sqrt{2}\sigma}\right) - \mathrm{erf}\left(\frac{a-\lambda}{\sqrt{2}\sigma}\right)$.

We now turn to the asymptotic behavior of SUGAR for large P, at a single realization of Y, i.e., our observation y. To this end, we first have to define how the observation model evolves with the dimension P. Given $z_0 \in \mathbb{R}^N$, we consider the sequence $\{\Psi_P\}_{P\geqslant 1}$ where $\Psi_P \in \mathbb{R}^{P\times N}$ and such that, for all $P > 1$, Ψ_P is the sub-matrix obtained by cutting down

one line of Ψ_{P+1}. We can then define a sequence of observation models as a the sequence of random vector $\{Y_P\}_{P\geqslant 1}$ defined as

$$Y_P = \Psi_P z_0 + W_P \quad \text{where} \quad W_P \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_P). \tag{15}$$

We also define the sequence $\{(\mu_0)_P\}_{P\geqslant 1}$ where $(\mu_0)_P = \Psi_P z_0$. In the following, for the sake of clarity, we ommit the dependency of Y_P, W_P and $(\mu_0)_P$ on P.

We can now state our consistency result of SUGAR for soft-thresholding.

Theorem 2 (Consistency of SUGAR)**.** *Take $\hat{\varepsilon}(P)$ such that $\lim_{P\to\infty} \hat{\varepsilon}(P) = 0$ and $\lim_{P\to\infty} P^{-1}\hat{\varepsilon}(P)^{-1} = 0$. Then for any Lebesgue point $\lambda > 0$ (i.e. such that $\forall (i,P), \lambda \neq |Y_i|$ and $\lambda \neq |Y_i + \hat{\varepsilon}(P)e_i|$)*

$$\begin{aligned} &\operatorname*{plim}_{P\to\infty} \left[\frac{1}{P}\left(\mathrm{SUGAR}_{\mathrm{FD}}\{\mathrm{ST}\}(Y,\lambda,\hat{\varepsilon}(P)) - \nabla_2\{R\{\mathrm{ST}\}\}(\mu_0,\lambda)\right)\right] = 0 \\ \text{and} \quad &\operatorname*{plim}_{P\to\infty} \left[\frac{1}{P}\left(\nabla_2\{\widehat{df}_{\mathrm{FD}}\{\mathrm{ST}\}\}(Y,\lambda,\hat{\varepsilon}(P)) - \nabla_2\{df\{\mathrm{ST}\}\}(\mu_0,\lambda)\right)\right] = 0\ . \end{aligned}$$

In plain words, Theorem 2 asserts that for the SUGAR of the soft-thresholding to be consistent, $\hat{\varepsilon}(P)$ should not decrease faster than the inverse of the dimension P. With the proviso that $\hat{\varepsilon}(P)$ fulfills the requirement, for P large enough, $\frac{1}{P}\mathrm{SUGAR}_{\mathrm{FD}}\{\mathrm{ST}\}(y,\lambda,\hat{\varepsilon}(P))$ is guaranteed to come close to $\frac{1}{P}\nabla_2\{R\{\mathrm{ST}\}\}(\mu_0,\lambda)$ with high probability.

Unfortunately, Theorem 2 does not dictate an explicit choice of $\hat{\varepsilon}(P)$, and the practitioner may wonder how to choose this value for a given P. It turns out that studying the mean squared error (MSE) of the gradient of the finite difference DOF estimator helps unveiling the link between P and ε through a bias-variance trade-off.

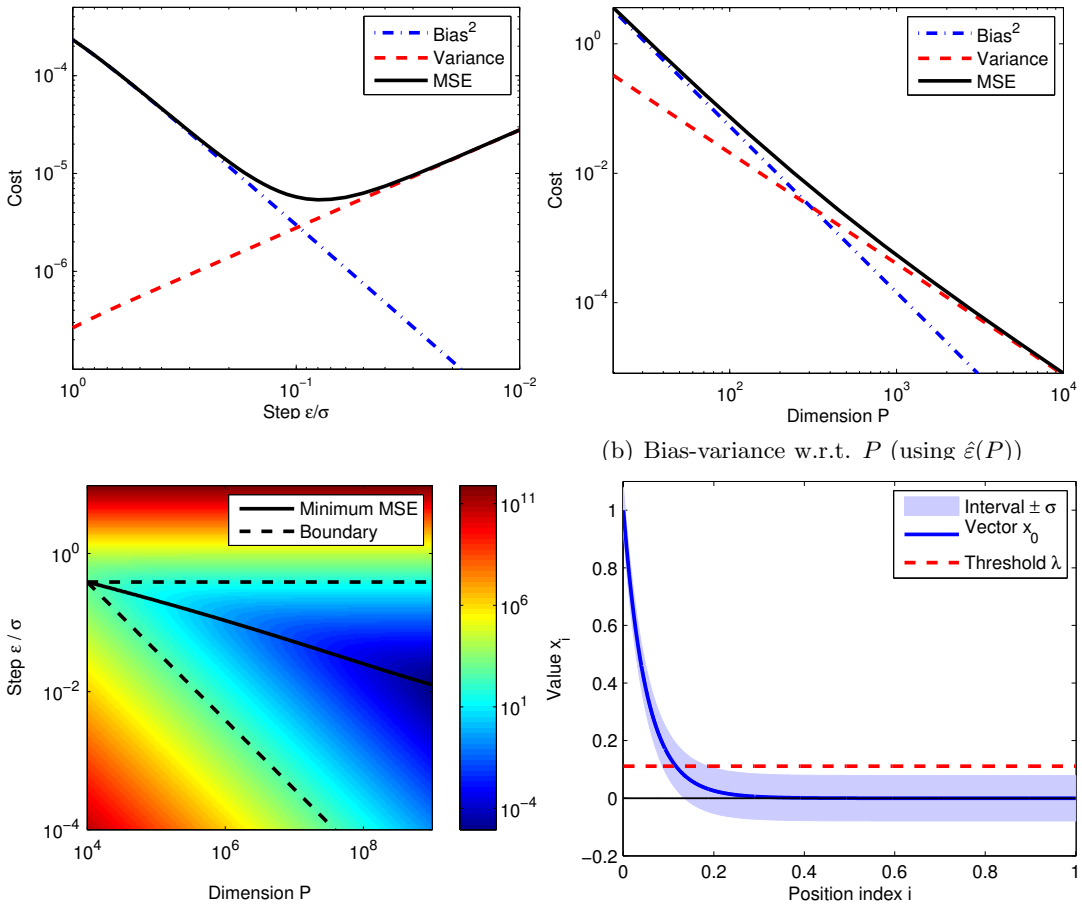
Proposition 2 (MSE of the gradient of the finite difference DOF estimator)**.** *The weak gradient of $\lambda \mapsto \widehat{df}\{\mathrm{ST}\}(Y,\lambda,\varepsilon)$ is such that*

$$\begin{aligned} &\mathbb{E}_W\left[\frac{1}{P}\left(\nabla_2\{\widehat{df}\{\mathrm{ST}\}\}(Y,\lambda,\varepsilon) - \nabla_2\{df\{\mathrm{ST}\}\}(\mu_0,\lambda)\right)\right]^2 = \\ &\underbrace{\frac{1}{P^2}\left(\mathbb{E}_W\left[\nabla_2\{\widehat{df}\{\mathrm{ST}\}\}(Y,\lambda,\varepsilon)\right] - \nabla_2\{df\{\mathrm{ST}\}\}(\mu_0,\lambda)\right)^2}_{Bias^2} + \underbrace{\frac{1}{P^2}\mathbb{V}_W\left[\nabla_2\{\widehat{df}\{\mathrm{ST}\}\}(Y,\lambda,\varepsilon)\right]}_{Variance}. \end{aligned}$$

where the statistics of $\nabla_2\{\widehat{df}\{\mathrm{ST}\}\}(Y,\lambda,\varepsilon)$ are given in Lemma 1 and

$$\nabla_2\{df\{\mathrm{ST}\}\}(\mu_0,\lambda) = \frac{-1}{\sqrt{2\pi}\sigma}\sum_{i=0}^{P}\left[\exp\left(-\frac{((\mu_0)_i+\lambda)^2}{2\sigma^2}\right) + \exp\left(-\frac{((\mu_0)_i-\lambda)^2}{2\sigma^2}\right)\right].$$

Thus, if μ_0 were given, the quantities in Proposition 2 could be computed in closed-form. The MSE can then be evaluated to select the optimal value of ε for a fixed dimension P and a given threshold λ. See the following numerical experiments which illustrate this relationship. When μ_0 is unknown, an a priori model can be imposed, such as for instance belonging to some ball promoting sparsity, e.g. a weak ℓ_γ-ball for $\gamma > 0$. For γ sufficiently small, this ball corresponds to compressible or nearly sparse vectors μ_0 whose entries $|\mu_i|$ sorted in descending order of magnitude behave as $O(i^{-1/\gamma})$. With such a model at hand, the MSE in Proposition 2 can be optimized for ε given P, σ, λ and γ. This however entails a highly non-linear equation that cannot be solved in closed form. We defer such a developement to a future work.



(b) Bias-variance w.r.t. P (using $\hat{\varepsilon}(P)$)

(c) MSE w.r.t. (ε, P)

(d) Choice of σ, μ_0 and λ

Figure 1: Bias-variance trade-off of the gradient estimator of the DOF of soft-thresholding, (a) with respect to the step ε and (b) with respect to the dimension P when using a power decay function $P \mapsto \hat{\varepsilon}(P)$. (c) Its mean squared error as a function of P and ε (in logarithmic scales). The solid line represents the pairs $(\hat{\varepsilon}^\star(P), P)$ where, for a fixed dimension P, $\hat{\varepsilon}^\star(P)$ minimizes the mean squared error. The function $\hat{\varepsilon}^\star(P)$ looks like a power function of the form of $C\sigma/P^\alpha$ with $C > 0$ and $0 < \alpha < 1$. The dashed lines represent respectively the power functions $\hat{\varepsilon}^{\inf}(P) = C\sigma$ and $\hat{\varepsilon}^{\sup}(P) = C\sigma/P$ outside which the mean squared error diverges when P increases. (d) Description of the settings of the experiments, i.e., the choice of σ, μ_0 and λ.

Figure 1.(a) shows the evolution of the bias and the variance as a function of the ratio ε/σ for fixed values of σ, λ and a compressible vector μ_0, i.e. $|(\mu_0)_i| = O(i^{-1/\gamma})$, chosen as illustrated on Figure 1.(d). When $\varepsilon \to 0$, for fixed P, the bias vanishes while the variance, and in turn the MSE, increases. However, for a step $\varepsilon > 0$, the MSE is finite and seems to be optimal around the value 0.1σ. Figure 1.(c) shows the evolution of the MSE as a function of the dimension P and the ratio ε/σ for the same fixed values as before. The optimal step, minimizing the MSE, seems to evolve as a power decay function (the scale is log-log) of the form $\varepsilon^\star(P) = C\sigma/P^\alpha$ with $C > 0$ and $0 < \alpha < 1$. Of course the optimal constants C and α depend on the choice of μ_0, σ and λ. However, whatever $C > 0$ and $0 < \alpha < 1$, or more generally for any admissible choice of $\hat{\varepsilon}$ such that $\lim_{P\to\infty} \hat{\varepsilon}(P) = 0$ and $\lim_{P\to\infty} P^{-1}\hat{\varepsilon}(P)^{-1} = 0$, the MSE vanishes with respect to P. Figure 1.(b) shows indeed the evolution of the bias, the variance and the MSE as a function of the dimension P when $\hat{\varepsilon}$ is

chosen as a power decay function. For $\alpha = 0$ or $\alpha = 1$, the MSE remains constant while, for $\alpha > 1$, the MSE diverges which suggests the necessity of $\lim_{P\to\infty} P^{-1}\hat{\varepsilon}(P)^{-1} = 0$.

4 Differentiation of an Iterative Scheme

We now turn to iterative algorithms that involve linear and soft-thresholding operators. We observed empirically that for all the inverse problems exposed in Section 5, setting $\varepsilon^\star(P) = C\sigma/P^\alpha$, as suggested our study the soft thresholding, resulted in a reliable way to parameterize our estimator. The effectiveness of this heuristic might be explained by the fact that the singularities encountered in most imaging problems are similar to absolute values, in order to encourage some sort of sparsity in the solution.

In this section, we focus on iterates, defined unambiguously as single-valued mappings $(y, \theta) \mapsto x^{(\ell)}(y, \theta)$, where ℓ is the iteration counter of the iterative algorithm. In this context, we propose to compute in closed-form the derivatives of $x^{(\ell)}(y, \theta)$ with respect to either y (in a direction δ) or θ. This proves useful to respectively estimate the risk via $\mathrm{SURE}^A_{\mathrm{MC}}$ (see Section 2) and estimate its gradient via $\mathrm{SUGAR}^A_{\mathrm{FDMC}}$ (see Section 3).

The iterative schemes we consider can be cast in the same framework, that of proximal splitting algorithms designed to minimize a proper, closed and convex objective function $x \mapsto E(x, y, \theta)$, whose set of minimizers is supposed non-empty. All these algorithms can be unified as an iterative scheme of the form

$$\begin{cases} x^{(\ell)} & = \gamma(a^{(\ell)}) \\ a^{(\ell+1)} & = \psi(a^{(\ell)}, y, \theta), \end{cases} \tag{16}$$

where $a^{(\ell)} \in \mathcal{A}$ is a sequence of auxiliary variables. $\psi : \mathcal{A} \times \mathcal{Y} \times \Theta \to \mathcal{A}$ is a fixed point operator in such a way that $a^{(\ell)}$ converge to a fixed point $a^\star$, and $\gamma : \mathcal{A} \to \mathcal{X}$ is non-expansive (i.e, $\|\gamma(a_1) - \gamma(a_2)\| \leqslant \|a_1 - a_2\|$ for any $a_1, a_2 \in \mathcal{A}$) entailing that $x^{(\ell)}$ will converge to $x^\star = \gamma(a^\star)$. Note that for the sake of clarity, we have dropped the dependencies of $a^\star$ and $x^\star$ to y and θ.

To make our ideas clear, consider the instructive example where $x \mapsto E(x, y, \theta)$ is convex and $C^1(\mathcal{X})$ with L-Lipschitz gradient, in which case $\mathcal{A} = \mathcal{X}$, $a = x$ and $\psi(x, y, \theta) = x - \tau\nabla_1 E(x, y, \theta)$ where $0 < \tau < 2/L$.

4.1 Iterative Weak Differentiability

A practical way to get the weak directional derivative $\partial_1 x(y, \theta)[\delta]$ and the weak Jacobian $\partial_2 x(y, \theta)$, is to compute them iteratively from the sequences (16) by relying on the chain rule. However, two major issues have to be taken care of. First, one has to ensure weak diffrentiability of the iterates (16) so that $\partial_1 x^{(\ell)}(y, \theta)[\delta]$ (or resp. to $\partial_2 x^{(\ell)}(y, \theta)$) exist Lebesgue a.e.. Second, one may legitimately ask whether the sequence of weak derivatives converges, and the properties of its cluster point, if any, with respect to the weak derivatives at a minimizer $x^\star$.

Regarding weak differentiability of the iterates, it relies essentially on regularity conditions to apply the chain rule, e.g. [24, Section 4.2.2], i.e. regularity properties of the iteration mappings γ and ψ and of the initialization. For instance, for proximal splitting algorithms, it turns out that γ is the composition of one or several non-expansive operators, hence 1-Lipschitz, operators. In turn, γ is 1-Lipschitz. Furthermore, in all examples we consider, ψ is also 1-Lipschitz with respect to its second and third arguments. Therefore, if one starts at a Lipschitz continuous initialization, by induction, $y \mapsto x^{(\ell)}(y, \theta)$ and $\theta \mapsto x^{(\ell)}(y, \theta)$ are

Algorithm Risk estimation of an iterative scheme

Inputs: observation $y \in \mathcal{Y} = \mathbb{R}^P$, collection of parameters $\theta \in \Theta$
Parameters: noise variance $\sigma^2 > 0$, linear operator $\Phi \in \mathbb{R}^{P \times N}$, matrix $A \in \mathbb{R}^{M \times P}$, number L of iterations
Output: solution $x(y, \theta) \in \mathcal{X}$ and its risk estimate $\widehat{R}^A\{x\}(y, \theta)$

Sample a vector δ from $\mathcal{N}(0, \mathrm{Id}_P)$
Initialize $a^{(0)} \leftarrow 0$ *
Initialize $\mathcal{D}_a^{(0)} \leftarrow 0$
for ℓ from 0 to $L-1$ **do** *
  $a^{(\ell+1)} \leftarrow \psi(a^{(\ell)}, y, \theta)$ *
  $\mathcal{D}_a^{(\ell+1)} \leftarrow \Psi_a^{(\ell)}(\mathcal{D}_a^{(\ell)}) + \Psi_y^{(\ell)}(\delta)$
end for *
$x^{(\ell)} \leftarrow \gamma(a^{(\ell)})$ *
$\mathcal{D}_x^{(\ell)} \leftarrow \Gamma_a^{(\ell)}(\mathcal{D}_a^{(\ell)})$
$\widehat{df}_{\mathrm{MC}}^A \leftarrow \left\langle \Phi \mathcal{D}_x^{(\ell)},\, A^*A\delta \right\rangle$
$\mathrm{SURE}_{\mathrm{MC}}^A \leftarrow \left\| A(y - \Phi x^{(\ell)}) \right\|^2 - \sigma^2 \operatorname{tr}(A^*A) + 2\sigma^2 \widehat{df}_{\mathrm{MC}}^A$
return $x(y, \theta) \leftarrow x^{(\ell)}$ and $\widehat{R}^A\{x\}(y, \theta) \leftarrow \mathrm{SURE}_{\mathrm{MC}}^A$

Figure 2: Pseudo-algorithm for risk estimation of an iterative scheme. The symbols * indicate the lines corresponding to the computation of x. The others are dedicated to the computation of the estimated risk $\widehat{R}^A$ using Monte Carlo simulation. Even if computing the risk requires more operations, the global complexity of the algorithm is unchanged.

also Lipschitz. Using the chain rule for Lipschitz mappings [24, Theorem 4 and Remark, Section 4.2.2], weak differentiabilty of $x^{(\ell)}$ follows with respect to both arguments.

As far as convergence of the sequence of weak Jacobians is concerned, this remains an open question in the general case, and we believe this would necessitate intricate arguments from non-smooth and variational analysis. This is left to future research.

From now on, we suppose that the Lipschitzian assumptions on γ, ψ and the initial points hold. The next two sections detail the computation of $\partial_1 x^{(\ell)}(y, \theta)[\delta]$ and $\partial_2 x^{(\ell)}(y, \theta)$ in order to get the estimates $\mathrm{SURE}_{\mathrm{MC}}^A$ and $\mathrm{SUGAR}_{\mathrm{FDMC}}^A$.

4.2 Computation of $\mathrm{SURE}_{\mathrm{MC}}^A$ for Risk Optimization

We describe here the iterative computation of the directional derivative $\partial_1 x^{(\ell)}(y, \theta)[\delta]$ following the idea introduced in [61] (see Section 2.4). Note that we focus on the directional derivative since, on the one hand, $\partial_1 x^{(\ell)}(y, \theta) \in \mathbb{R}^{N \times P}$ is never used explicitly but only its trace, not to mention its storage cost, and, on the other hand, the risk can be estimated by applying only the weak directional derivatives on random directions δ (see Section 2.3 for more details).

The next proposition summarizes a recursive scheme to compute the weak derivatives $\partial_1 x^{(\ell)}(y, \theta)[\delta]$.

Proposition 3. *For any vector $\delta \in \mathcal{X}$, the weak directional derivative $\mathcal{D}_x^{(\ell)} = \partial_1 x^{(\ell)}(y, \theta)[\delta]$*

Algorithm Risk and gradient risk estimation of an iterative scheme

Inputs: observation $y \in \mathcal{Y} = \mathbb{R}^P$, collection of parameters $\theta \in \Theta$
Parameters: noise variance $\sigma^2 > 0$, linear operator $\Phi \in \mathbb{R}^{P \times N}$, matrix $A \in \mathbb{R}^{M \times P}$, number L of iterations, decay parameters $C > 0$ and $0 < \alpha < 1$
Output: solution $x(y,\theta) \in \mathcal{X}$, its risk estimate $\widehat{R}^A\{x\}(y,\theta)$, and its gradient risk estimate $\hat{\nabla}_2 R^A\{x\}(y,\theta)$

Sample a vector δ from $\mathcal{N}(0, \mathrm{Id}_P)$ *
Choose $\varepsilon = C\sigma/P^\alpha$ *
for $y' = y$ and $y' = y + \varepsilon\delta$ **do** *
  Initialize $a^{(0)} \leftarrow 0$ *
  Initialize $\mathcal{J}_a^{(0)} \leftarrow 0$
  for ℓ from 0 to $L-1$ **do** *
    $a^{(\ell+1)} \leftarrow \psi(a^{(\ell)}, y', \theta)$ *
    $\mathcal{J}_a^{(\ell+1)} \leftarrow \Psi_a^{(\ell)}(\mathcal{J}_a^{(\ell)}) + \Psi_\theta^{(\ell)}$
  end for *
  $x^{(\ell)}(y') \leftarrow \gamma(a^{(\ell)})$ *
  $\mathcal{J}_x^{(\ell)}(y') \leftarrow \Gamma_a^{(\ell)}(\mathcal{J}_a^{(\ell)})$
end for *
$\widehat{df}_{\mathrm{FDMC}} \leftarrow \frac{1}{\varepsilon}\left\langle \Phi(x^{(\ell)}(y+\varepsilon\delta) - x^{(\ell)}(y)),\, A^*A\delta \right\rangle$ *
$\mathrm{SURE}^A_{\mathrm{FDMC}} \leftarrow \left\|A(y - \Phi x^{(\ell)})\right\|^2 - \sigma^2 \operatorname{tr}(A^*A) + 2\sigma^2 \widehat{df}_{\mathrm{FDMC}}$ *
$\mathrm{SUGAR}^A_{\mathrm{FDMC}} \leftarrow 2\mathcal{J}_x^{(\ell)}(y)^* \Phi^* A^* A(\Phi x^{(\ell)} - y) + \frac{2\sigma^2}{\varepsilon}\left(\mathcal{J}_x^{(\ell)}(y+\varepsilon\delta) - \mathcal{J}_x^{(\ell)}(y)\right)^* \Phi^* A^* A\delta$
return $x(y,\theta) \leftarrow x^{(\ell)}(y)$, $\widehat{R}^A\{x\}(y,\theta) \leftarrow \mathrm{SURE}^A_{\mathrm{FDMC}}$ and $\hat{\nabla}_2 R^A\{x\}(y,\theta) \leftarrow \mathrm{SUGAR}^A_{\mathrm{FDMC}}$

Figure 3: Pseudo-algorithm for risk and gradient risk estimation of an iterative scheme. The symbols * indicate the lines corresponding to the computation of x and its estimated risk R^A using approximated Monte Carlo simulation, i.e., as described in [44]. The others are dedicated to the computation of the estimated gradient of the risk $\hat{\nabla} R^A$. Even if computing the gradient of the risk requires more operations, the global complexity of the algorithm is unchanged.

is given by

$$\mathcal{D}_x^{(\ell)} = \Gamma_a^{(\ell)}(\mathcal{D}_a^{(\ell)})$$
$$\textit{with} \quad \mathcal{D}_a^{(\ell+1)} = \Psi_a^{(\ell)}(\mathcal{D}_a^{(\ell)}) + \Psi_y^{(\ell)}(\delta),$$

where $\mathcal{D}_a^{(\ell)} = \partial_1 a^{(\ell)}(y,\theta)[\delta]$ *and we have defined the following linear mappings*

$$\Gamma_a^{(\ell)}(\cdot) = \partial_1 \gamma(a^{(\ell)})[\cdot],$$
$$\Psi_a^{(\ell)}(\cdot) = \partial_1 \psi(a^{(\ell)}, y, \theta)[\cdot],$$
$$\Psi_y^{(\ell)}(\cdot) = \partial_2 \psi(a^{(\ell)}, y, \theta)[\cdot].$$

Plugging $\partial_2 x^{(\ell)}(y,\theta)[\delta]$ in (8), and in turn in (6), gives iteratively an unbiased[§] estimate of the risk at the current iterate $x^{(\ell)}(y,\theta)$. The whole procedure is summarized in Fig. 2. It

[§]Expectation is to be taken here with respect to both the Gaussian measure of the noise W and the direction Δ.

is worth point out that although estimating the risk entails additional operations, the global complexity is the same as for the iterative splitting algorithm without risk estimation.

4.3 Computation of $\mathrm{SUGAR}^{\boldsymbol{A}}_{\mathbf{FDMC}}$ for Risk Optimization

We now focus on the computation of the weak Jacobian $\partial_2 x^{(\ell)}(y,\theta)$. Unlike for risk estimation that required only weak directional derivatives, for risk optimization we need the full weak Jacobian matrix $\partial_2 x(y,\theta) \in \mathbb{R}^{\dim(\Theta)\times N}$. The proposed strategy, known as the forward accumulation, is one of the possible strategies to iteratively evaluate the derivatives by the use of the chain rule. The reverse accumulation is another strategy that does not require computing the full Jacobian matrix at the expense of a large memory load with respect to the number of iterations. Between these two extreme approaches, they are several hybrid strategies that can also be considered, knowing, that finding the optimal Jacobian accumulation strategy is an NP-complete problem. Such strategies have been studied in the field of "automatic differentiation" and the reader is invited to refer to [30, 41] for a comprehensive account of these approaches.

In our case, we consider that, unlike for $\partial_1 x^{(\ell)}(y,\theta)$, the matrix $\partial_2 x(y,\theta)$ is in practice quite small since $\dim(\Theta) \ll P$, hence implying only a memory load overhead of small fraction of P. Hence following the forward accumulation strategy, we propose a practical way to compute iteratively the full weak Jacobian matrix $\partial_2 x(y,\theta)$.

Note that due to the small dimension of Θ, evaluating this Jacobian matrix with Monte-Carlo simulations would not be a good strategy since it would require generating a large amount of random directions in Θ.

The next result describes an iterative scheme to compute $\partial_2 x^{(\ell)}(y,\theta)$.

Proposition 4. *The weak Jacobian $\mathcal{J}_x^{(\ell)} = \partial_2 x^{(\ell)}(y,\theta)$ is given by*

$$\begin{aligned} \mathcal{J}_x^{(\ell)} &= \Gamma_a^{(\ell)}(\mathcal{J}_a^{(\ell)}) \\ \textit{with} \quad \mathcal{J}_a^{(\ell+1)} &= \Psi_a^{(\ell)}(\mathcal{J}_a^{(\ell)}) + \Psi_\theta^{(\ell)}, \end{aligned}$$

where $\mathcal{J}_a^{(\ell)} = \partial_2 a^{(\ell)}(y,\theta)$ and we have defined

$$\begin{aligned} \Gamma_a^{(\ell)}(\cdot) &= \partial_1 \gamma(a^{(\ell)})[\cdot], \\ \Psi_a^{(\ell)}(\cdot) &= \partial_1 \psi(a^{(\ell)}, y, \theta)[\cdot], \\ \Psi_\theta^{(\ell)} &= \partial_3 \psi(a^{(\ell)}, y, \theta). \end{aligned}$$

Plugging $\partial_2 x^{(\ell)}(y,\theta)$ in the expression of $\mathrm{SUGAR}^{A}_{\mathrm{FDMC}}$ given by Proposition 1 provides iteratively an asymptotically (see Theorem 1) unbiased estimate of the gradient of the risk at the current iterate $x^{(\ell)}(y,\theta)$. The main steps of the procedure are summarized in Fig. 3. The estimation of the gradient of the risk entails only a small computational overhead compared to the risk estimation approach of [44]. Their respective complexity remains however the same.

Note finally that in both schemes, another initialization than $a^{(0)} = 0$ can be chosen, for instance depending on y and θ, in which case the respective derivatives require to be initialized accordingly.

The following sections are devoted to instantiate this approach to more specific iterative algorithms that are able to handle non-smooth convex objective functions E.

4.4 Application to Generalized Forward Backward Splitting

The Generalized Forward Backward (GFB) splitting [43] allows one to find one element belonging to the set $x(y,\theta)$ solution of the structured convex optimization problem

$$x(y,\theta) = \operatorname*{Argmin}_{x\in\mathcal{X}} \left\{ E(x,y,\theta) = F(x,y,\theta) + \sum_{k=1}^{Q} G_k(x,y,\theta) \right\} \tag{17}$$

under the assumptions that all functions are proper, closed and convex, F is $C^1(\mathcal{X})$ with L-Lipschitz continuous gradient, and the G_k functions are simple, in the sense their proximity operator can be computed in closed form (e.g. the ℓ_1 norm is simple since its proximal operator is explicitely the soft-thresholding). Recall that the proximal mapping of a proper closed convex function G is defined as

$$\mathrm{Prox}_G : x \in \mathcal{X} \mapsto \operatorname*{argmin}_{z\in\mathcal{X}} \frac{1}{2}\|z-x\|^2 + G(z) \ .$$

It is uniquely valued and non-expansive (in fact even firmly so, i.e., $\|\mathrm{Prox}_G(x_1) - \mathrm{Prox}_G(x_2)\| < \|x_1 - x_2\|$ for any $x_1, x_2 \in \mathcal{X}$).

The GFB implements iteration (16) with $a^{(\ell)} = (\xi^{(\ell)}, z_1^{(\ell)}, \ldots, z_Q^{(\ell)}) \in \mathcal{A} = \mathcal{X}^{1+Q}$, $x^{(\ell)} = \gamma(a^{(\ell)}) = \xi^{(\ell)}$ and $a^{(\ell+1)} = \psi(a^{(\ell)}, y)$ chosen such that for all $k = 1, \ldots, Q$,

$$\begin{array}{rrcl} & x^{(\ell+1)} & = & \frac{1}{Q}\sum_{k=1}^{Q} z_k^{(\ell+1)} \\ \text{and} & z_k^{(\ell+1)} & = & z_k^{(\ell)} - x^{(\ell)} + \mathrm{Prox}_{\nu Q G_k}(\mathcal{Z}_k^{(\ell)}, y, \theta) \\ \text{with} & \mathcal{Z}_k^{(\ell)} & = & 2x^{(\ell)} - z_k^{(\ell)} - \nu \nabla_1 F(x^{(\ell)}, y, \theta) \ . \end{array}$$

With the parameter $\nu \in]0, 2/L[$, the sequence of iterates $x^{(\ell)}$ is provably guaranteed to converge to a minimizer $x(y,\theta)$ of (17). One recovers as a special cases the Forward-Backward splitting [12] when $Q = 1$ and the Douglas-Rachford splitting [10] when $F = 0$.

Corollary 1. *For any vector $\delta \in \mathcal{X}$, the GFB weak directional derivatives $\mathcal{D}_x^{(\ell)} = \Gamma_a^{(\ell)}(\mathcal{D}_a^{(\ell)})$ and $\mathcal{D}_a^{(\ell+1)} = \Psi_a^{(\ell)}(\mathcal{D}_a^{(\ell)}) + \Psi_y^{(\ell)}(\delta)$ are computed by evaluating iteratively*

$$\begin{array}{rrcl} & \mathcal{D}_x^{(\ell+1)} & = & \frac{1}{Q}\sum_{k=1}^{Q} \mathcal{D}_{z_k}^{(\ell+1)} \\ \textit{and} & \mathcal{D}_{z_k}^{(\ell+1)} & = & \mathcal{D}_{z_k}^{(\ell)} - \mathcal{D}_x^{(\ell)} + \mathcal{G}_{k,x}^{(\ell)}(\mathcal{D}_{\mathcal{Z}_k}^{(\ell)}) + \mathcal{G}_{k,y}^{(\ell)}(\delta) \\ \textit{with} & \mathcal{D}_{\mathcal{Z}_k}^{(\ell)} & = & 2\mathcal{D}_x^{(\ell)} - \mathcal{D}_{z_k}^{(\ell)} - \nu(\mathcal{F}_x^{(\ell)}(\mathcal{D}_x^{(\ell)}) + \mathcal{F}_y^{(\ell)}(\delta)) \ , \end{array}$$

where we have defined the following linear mappings

$$\begin{array}{rrcl} & \mathcal{G}_{k,x}^{(\ell)}(\cdot) & = & \partial_1\{\mathrm{Prox}_{\nu Q G_k}\}(\mathcal{Z}_k^{(\ell)}, y, \theta)[\cdot], \\ & \mathcal{G}_{k,y}^{(\ell)}(\cdot) & = & \partial_2\{\mathrm{Prox}_{\nu Q G_k}\}(\mathcal{Z}_k^{(\ell)}, y, \theta)[\cdot], \\ & \mathcal{F}_x^{(\ell)}(\cdot) & = & \partial_1\{\nabla_1 F\}(x^{(\ell)}, y, \theta)[\cdot], \\ \textit{and} & \mathcal{F}_y^{(\ell)}(\cdot) & = & \partial_2\{\nabla_1 F\}(x^{(\ell)}, y, \theta)[\cdot] \ . \end{array}$$

Corollary 2. *In the same vein as Corollary 1, the GFB weak Jacobian $\mathcal{J}_x^{(\ell)} = \Gamma_a^{(\ell)}(\mathcal{J}_a^{(\ell)})$, where $\mathcal{J}_a^{(\ell+1)} = \Psi_a^{(\ell)}(\mathcal{J}_a^{(\ell)}) + \Psi_\theta^{(\ell)}$, is computed by evaluating iteratively*

$$\begin{array}{rrcl} & \mathcal{J}_x^{(\ell+1)} & = & \frac{1}{Q}\sum_{k=1}^{Q} \mathcal{J}_{z_k}^{(\ell+1)} \\ \textit{and} & \mathcal{J}_{z_k}^{(\ell+1)} & = & \mathcal{J}_{z_k}^{(\ell)} - \mathcal{J}_x^{(\ell)} + \mathcal{G}_{k,x}^{(\ell)}(\mathcal{J}_{\mathcal{Z}_k}^{(\ell)}) + \mathcal{G}_{k,\theta}^{(\ell)} \\ \textit{with} & \mathcal{J}_{\mathcal{Z}_k}^{(\ell)} & = & 2\mathcal{J}_x^{(\ell)} - \mathcal{J}_{z_k}^{(\ell)} - \nu(\mathcal{F}_x^{(\ell)}(\mathcal{J}_x^{(\ell)}) + \mathcal{F}_\theta^{(\ell)}) \ , \end{array}$$

where we have defined

$$\begin{aligned} \mathcal{G}_{k,x}^{(\ell)}(\cdot) &= \partial_1\{\mathrm{Prox}_{\nu Q G_k}\}(\mathcal{Z}_k^{(\ell)}, y, \theta)[\cdot], \\ \mathcal{G}_{k,\theta}^{(\ell)} &= \partial_3\{\mathrm{Prox}_{\nu Q G_k}\}(\mathcal{Z}_k^{(\ell)}, y, \theta), \\ \mathcal{F}_x^{(\ell)}(\cdot) &= \partial_1\{\nabla_1 F\}(x^{(\ell)}, y, \theta)[\cdot], \\ \textit{and}\quad \mathcal{F}_\theta^{(\ell)} &= \partial_3\{\nabla_1 F\}(x^{(\ell)}, y, \theta)\ . \end{aligned}$$

4.5 Application to Primal-dual Splitting

Proximal splitting schemes can be used to find an element of the set $x(y,\theta)$ defined as the solution of the large class of variational problems

$$x(y,\theta) = \underset{x\in\mathcal{X}}{\mathrm{Argmin}}\ \{E(x,y,\theta) = H(x,y,\theta) + G(K(x),y,\theta)\}\ , \tag{18}$$

where both $x \mapsto H(x,y,\theta)$ and $u \mapsto G(u,y,\theta)$ are proper closed convex and simple functions, and $K : \mathcal{X} \to \mathcal{U}$ is a bounded linear operator.

The primal-dual relaxed Arrow-Hurwicz algorithm as revitalized recently in [8] (that we coin CP) to solve (18) implements (16) with $a^{(\ell)} = (\xi^{(\ell)}, \tilde{x}^{(\ell)}, u^{(\ell)}) \in \mathcal{A} = \mathcal{X}^2 \times \mathcal{U}$, $x^{(\ell)} = \gamma(a^{(\ell)}) = \xi^{(\ell)}$ and $a^{(\ell+1)} = \psi(a^{(\ell)}, y)$ such that

$$\begin{aligned} &u^{(\ell+1)} = \mathrm{Prox}_{\tau G^*}(U^{(\ell)}, y, \theta) \quad \text{where} \quad U^{(\ell)} = u^{(\ell)} + \tau K(\tilde{x}^{(\ell)}), \\ &x^{(\ell+1)} = \mathrm{Prox}_{\xi H}(X^{(\ell)}, y, \theta) \quad \text{where} \quad X^{(\ell)} = x^{(\ell)} - \xi K^*(u^{(\ell+1)}), \\ &\tilde{x}^{(\ell+1)} = x^{(\ell+1)} + \zeta(x^{(\ell+1)} - x^{(\ell)}). \end{aligned} \tag{19}$$

where the Legendre-Fenchel conjugate of G is defined as $G^*(u,y,\tau) = \max_z \langle z,\, u\rangle - G(z,y,\tau)$, and its proximity operator is given by Moreau's identity as

$$\mathrm{Prox}_{\tau G^*}(u,y) = u - \tau\,\mathrm{Prox}_{G/\tau}(u/\tau, y)\ .$$

The parameters $\tau > 0, \xi > 0$ are chosen such that $\tau\xi\,\|K\|^2 < 1$, and $\zeta \in [0,1]$ to ensure provable convergence of $x^{(\ell)}$ toward an element in the set $x(y,\theta)$ of (18). $\zeta = 0$ corresponds to the Arrow-Hurwitz algorithm, and for $\zeta = 1$, a sublinear $O(1/\ell)$ convergence rate on the partial duality gap was established in [8].

Corollary 3. *For any vector $\delta \in \mathcal{X}$, the CP weak directional derivatives $\mathcal{D}_x^{(\ell)} = \Gamma_a^{(\ell)}(\mathcal{D}_a^{(\ell)})$ and $\mathcal{D}_a^{(\ell+1)} = \Psi_a^{(\ell)}(\mathcal{D}_a^{(\ell)}) + \Psi_y^{(\ell)}(\delta)$ are computed by evaluating iteratively*

$$\begin{aligned} \mathcal{D}_u^{(\ell+1)} &= \mathcal{G}_u^{(\ell)}(\mathcal{D}_U^{(\ell)}) + \mathcal{G}_y^{(\ell)}(\delta) \quad \textit{where} \quad \mathcal{D}_U^{(\ell)} = \mathcal{D}_u^{(\ell)} + \tau K(\mathcal{D}_{\tilde{x}}^{(\ell)}), \\ \mathcal{D}_x^{(\ell+1)} &= \mathcal{H}_x^{(\ell)}(\mathcal{D}_X^{(\ell)}) + \mathcal{H}_y^{(\ell)}(\delta) \quad \textit{where} \quad \mathcal{D}_X^{(\ell)} = \mathcal{D}_x^{(\ell)} - \xi K^*(\mathcal{D}_u^{(\ell+1)}), \\ \textit{and}\quad \mathcal{D}_{\tilde{x}}^{(\ell+1)} &= \mathcal{D}_x^{(\ell+1)} + \zeta(\mathcal{D}_x^{(\ell+1)} - \mathcal{D}_x^{(\ell)}) \end{aligned}$$

where we have defined the following linear mappings

$$\begin{aligned} \mathcal{H}_x^{(\ell)}(\cdot) &= \partial_1\{\mathrm{Prox}_{\xi H}\}(X^{(\ell)}, y, \theta)[\cdot], \\ \mathcal{H}_y^{(\ell)}(\cdot) &= \partial_2\{\mathrm{Prox}_{\xi H}\}(X^{(\ell)}, y, \theta)[\cdot], \\ \mathcal{G}_u^{(\ell)}(\cdot) &= \partial_1\{\mathrm{Prox}_{\tau G^*}\}(U^{(\ell)}, y, \theta)[\cdot], \\ \textit{and}\quad \mathcal{G}_y^{(\ell)}(\cdot) &= \partial_2\{\mathrm{Prox}_{\tau G^*}\}(U^{(\ell)}, y, \theta)[\cdot]\ . \end{aligned}$$

Corollary 4. *Similarly to Corollary 3, the CP weak Jacobians* $\mathcal{J}_x^{(\ell)} = \Gamma_a^{(\ell)}(\mathcal{J}_a^{(\ell)})$ *and* $\mathcal{J}_a^{(\ell+1)} = \Psi_a^{(\ell)}(\mathcal{J}_a^{(\ell)}) + \Psi_\theta^{(\ell)}$ *are computed by evaluating iteratively*

$$\begin{array}{rrcl} & \mathcal{J}_u^{(\ell+1)} & = & \mathcal{G}_u^{(\ell)}(\mathcal{J}_U^{(\ell)}) + \mathcal{G}_\theta^{(\ell)} \quad \textit{where} \quad \mathcal{J}_U^{(\ell)} = \mathcal{J}_u^{(\ell)} + \tau K(\mathcal{J}_{\tilde{x}}^{(\ell)}), \\ & \mathcal{J}_x^{(\ell+1)} & = & \mathcal{H}_x^{(\ell)}(\mathcal{J}_X^{(\ell)}) + \mathcal{H}_\theta^{(\ell)} \quad \textit{where} \quad \mathcal{J}_X^{(\ell)} = \mathcal{J}_x^{(\ell)} - \xi K^*(\mathcal{J}_u^{(\ell+1)}), \\ \textit{and} & \mathcal{J}_{\tilde{x}}^{(\ell+1)} & = & \mathcal{J}_x^{(\ell+1)} + \zeta(\mathcal{J}_x^{(\ell+1)} - \mathcal{J}_x^{(\ell)}) \end{array}$$

where we have defined

$$\begin{array}{rrcl} & \mathcal{H}_x^{(\ell)}(\cdot) & = & \partial_1\{\mathrm{Prox}_{\xi H}\}(X^{(\ell)}, y, \theta)[\cdot], \\ & \mathcal{H}_\theta^{(\ell)} & = & \partial_3\{\mathrm{Prox}_{\xi H}\}(X^{(\ell)}, y, \theta), \\ & \mathcal{G}_u^{(\ell)}(\cdot) & = & \partial_1\{\mathrm{Prox}_{\tau G^*}\}(U^{(\ell)}, y, \theta)[\cdot], \\ \textit{and} & \mathcal{G}_\theta^{(\ell)} & = & \partial_3\{\mathrm{Prox}_{\tau G^*}\}(U^{(\ell)}, y, \theta). \end{array}$$

Note that the two proximal splitting schemes described here were chosen for their flexibility and the richness of the class of problems they can handle. Obviously, the methodology and discussion extend easily to the reader's favorite proximal splitting algorithm.

5 Examples and Numerical Results

In this section, we exemplify the use of the formal differentiation of iterative proximal splitting algorithms for three popular variational problems: nuclear norm regularization, total-variation regularization and multi-scale wavelet ℓ_1-analysis sparsity prior. For each of them, the expressions of all quantities including the proximal operators and their derivatives are given in closed-form. On each problem, we illustrate the usefulness of our gradient risk estimators for (multi) continuous parameter optimization.

5.1 Implementation Details

All experiments reported below are based on the algorithms detailed in Figure 2 and 3 in conjunction with proximal splitting-algorithms presented in the previous section. The step of the finite difference is chosen as $\varepsilon = 2\sigma/P^{0.3}$. Iterative proximal splitting-algorithms will be used with 100 iterations. For quasi-Newton optimization, we used the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method with the implementation of [36]. An important issue in using quasi-Newton optimization is the choice of the initialization, the initial step and the stopping criteria. For a variation regularization problem expressed as

$$\underset{x}{\mathrm{Argmin}} \ \frac{1}{2}\left\|\Phi x - y\right\|^2 + \sum_{k=1}^{K} \lambda^k \mathcal{R}^k(x) \ , \tag{20}$$

where $\lambda^k > 0, \ \forall k \in \mathbb{N}$, the initialization λ_0^k is chosen empirically as

$$\lambda_0^k = \frac{P\sigma^2}{4\sum_{k=1}^{K} \mathcal{R}^k(x_{\mathrm{LS}}(y))} \ , \tag{21}$$

where $x_{\mathrm{LS}}(y)$ is the least-square estimator. At the first iteration, the approximate inverse Hessian B_1 should be chosen such that, for all $k > 0$, λ_1^k is of the same order as λ_0^k. To this end, we suggest initializing B_1 as a diagonal matrix with diagonal entries

$$B_1^k = \left| \frac{\alpha \lambda_0^k}{\mathrm{SUGAR}_{\mathrm{FDMC}}^{A}\{x\}(y, \lambda_0, \delta, \varepsilon)_k} \right| \tag{22}$$

such that, for all k, $\lambda_1^k = (1 \pm \alpha)\lambda_0^k$, where, in practice, we have chosen $\alpha = 0.9$. Finally, the BFGS method stops after the following criterion is reached

$$\frac{\left\|\mathrm{SUGAR}^A_{\mathrm{FDMC}}\{x\}(y, \lambda_n, \delta, \varepsilon)\right\|_\infty}{\left\|\mathrm{SUGAR}^A_{\mathrm{FDMC}}\{x\}(y, \lambda_0, \delta, \varepsilon)\right\|_\infty} \leqslant \tau \tag{23}$$

where we have chosen $\tau = 0.02$ meaning that the algorithm stops if all (weak) partial derivatives are at least 50 times lower than the maximal one at initialization.

For the sake of reproducibility, the Matlab scripts implementing the SURE and SUGAR for the different problems details hereafter are available online at `http://www.math.u-bordeaux1.fr/~cdeledal/sugar.php`.

5.2 Nuclear Norm Regularization

We consider the recovery of a low-rank matrix $x_0 \in \mathbb{R}^{n_1 \times n_2}$ from an observation $y \in \mathbb{R}^P$ of $Y = \Phi x_0 + W$, $W \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_P)$, where we have identified the matrix space $\mathbb{R}^{n_1 \times n_2}$ to the vector space $\mathbb{R}^N$ with $N = n_1 n_2$. To this end, we consider the following spectral regularization problem

$$x^\star(y, \lambda) \in \operatorname*{Argmin}_x \frac{1}{2} \|\Phi x - y\|^2 + \lambda \|x\|_* \ , \tag{24}$$

where $\lambda > 0$ and $\|\cdot\|_*$ is the nuclear norm (a.k.a., trace for the sdp case or Schatten 1-norm). This is a spectral function defined as the ℓ_1 norm of the singular values $\Lambda_x \in \mathbb{R}^{n=\min(n_1,n_2)}$, i.e.

$$\|x\|_* = \|\Lambda_x\|_1 \ .$$

The nuclear norm is a particular case of spectral regularization that accounts for prior knowledge on the spectrum of x, typically low-rank (see, e.g., [25]). It is the convex hull of the rank function restricted to the unit spectral ball [6]. The parameter λ balances the sparsity of the spectrum of the recovered matrix, and the tolerated amount of noise. However, except in the random measurements setting, there is no direct relation between λ and the rank of $x(y, \lambda)$. The optimal value of λ depends indeed on x_0, Φ and σ confirming the importance of automatic selection procedures.

Problem (24) is a special instance of (17) with the parameter $\lambda = \theta \in \Theta = \mathbb{R}^+$, $Q = 1$, and

$$\begin{aligned} F(x, y, \lambda) &= \frac{1}{2} \|\Phi x - y\|^2 , \\ \text{and} \quad G_1(x, y, \lambda) &= \lambda \|x\|_* \ . \end{aligned}$$

Hence the GFB algorithm[¶] can be used to solve (24) by setting

$$\begin{aligned} \nabla_1 F(x, y, \lambda) &= \Phi^*(\Phi x - y), \\ \text{and} \quad \mathrm{Prox}_{\tau G_1}(x, y, \lambda) &= V_x \, \mathrm{diag}(\mathrm{ST}(\Lambda_x, \tau\lambda)) U_x^* \end{aligned}$$

where $\mathrm{diag} : \mathbb{R}^n \to \mathbb{R}^{n_1 \times n_2}$ maps the entries of a vector in $\mathbb{R}^n$ to the main diagonal of a rectangular matrix in $\mathbb{R}^{n_1 \times n_2}$ filled with 0 elsewhere, $(V_x, U_x, \Lambda_x) \in \mathbb{R}^{n_1 \times n_1} \times \mathbb{R}^{n_2 \times n_2} \times \mathbb{R}^n$ is the singular value decomposition (SVD) of x such that $x = V_x \, \mathrm{diag}(\Lambda_x) U_x^*$ and ST is the

[¶] which corresponds in this case where $Q = 1$ to the forward-backward algorithm.

soft-thresholding operator (14). Corollary 1 and 2 can then be applied using, for any $\delta_x \in \mathcal{X}$ and $\delta_y \in \mathcal{Y}$, the relations

$$\begin{array}{rlcl} & \partial_1\{\nabla_1 F\}(x,y,\lambda)[\delta_x] & = & \Phi^*\Phi\delta_x, \\ & \partial_2\{\nabla_1 F\}(x,y,\lambda)[\delta_y] & = & -\Phi^*\delta_y, \\ & \partial_3\{\nabla_1 F\}(x,y,\lambda) & = & 0 \\ \text{and} & \partial_1\{\mathrm{Prox}_{\tau G_1}\}(x,y,\lambda)[\delta_x] & = & V_x(\mathcal{H}(\Lambda_x)[\bar{\delta}_x] + \Gamma_S(\Lambda_x)[\bar{\delta}_x] + \Gamma_A(\Lambda_x)[\bar{\delta}_x])U_x^*, \\ & \partial_2\{\mathrm{Prox}_{\tau G_1}\}(x,y,\lambda)[\delta_y] & = & 0, \\ & \partial_3\{\mathrm{Prox}_{\tau G_1}\}(x,y,\lambda) & = & V_x \operatorname{diag}(\partial_2 \mathrm{ST}(\Lambda_x,\tau\lambda))U_x^* \end{array}$$

where $\bar{\delta}_x = V_X^*\delta_x U_X \in \mathbb{R}^{n_1\times n_2}$, $\mathcal{H}(\Lambda_x)$ is defined as

$$\mathcal{H}(\Lambda_x)[\bar{\delta}_x] = \operatorname{diag}(\partial_1 \mathrm{ST}(\Lambda_x,\rho\lambda)[\operatorname{diag}(\bar{\delta}_x)])$$

and $\Gamma_S(\Lambda_x)$ and $\Gamma_A(\Lambda_x)$ are defined, for all $1 \leqslant i \leqslant n_1$ and $1 \leqslant j \leqslant n_2$, as

$$\Gamma_S(\Lambda_x)[\bar{\delta}_x]_{i,j} = \frac{(\bar{\delta}_x)_{i,j} + (\bar{\delta}_x)_{j,i}}{2} \times \begin{cases} 0 & \text{if} \quad i=j \\ \frac{\mathrm{ST}(\Lambda_x,\rho\lambda)_i - \mathrm{ST}(\Lambda_x,\rho\lambda)_j}{(\Lambda_x)_i - (\Lambda_x)_j} & \text{if} \quad (\Lambda_x)_i \neq (\Lambda_x)_j \\ \partial_1\mathrm{ST}(\Lambda_x,\rho\lambda)_{i,i} & \text{otherwise,} \end{cases}$$

$$\Gamma_A(\Lambda_x)[\bar{\delta}_x]_{i,j} = \frac{(\bar{\delta}_x)_{i,j} - (\bar{\delta}_x)_{j,i}}{2} \times \begin{cases} 0 & \text{if} \quad i=j \\ \frac{\mathrm{ST}(\Lambda_x,\rho\lambda)_i + \mathrm{ST}(\Lambda_x,\rho\lambda)_j}{(\Lambda_x)_i + (\Lambda_x)_j} & \text{if} \quad (\Lambda_x)_i > 0 \quad \text{or} \quad (\Lambda_x)_j > 0 \\ \partial_1\mathrm{ST}(\Lambda_x,\rho\lambda)_{i,i} & \text{otherwise,} \end{cases}$$

where for $i > n$ we have extended Λ_x and $\mathrm{ST}(\Lambda_x,\rho\lambda)$ as $(\Lambda_x)_i = 0$ and $\mathrm{ST}(\Lambda_x,\rho\lambda)_i = 0$, and for $j > n_1$ or $i > n_2$, $\bar{\delta}_x$ as $(\bar{\delta}_x)_{j,i} = 0$. Recall from (32) that the weak derivatives of the soft-thresholding are defined, for $t \in \mathbb{R}^N$, $\rho > 0$, $\delta_t \in \mathbb{R}^N$, $1 \leqslant i \leqslant N$, by

$$\begin{aligned} \partial_1\mathrm{ST}(t,\rho)_{i,i} &= \begin{cases} 0 & \text{if} \quad |t_i| \leqslant \rho \\ 1 & \text{otherwise} \end{cases} \\ \partial_1\mathrm{ST}(t,\rho)[\delta_t]_i &= \partial_1\mathrm{ST}(t,\rho)_{i,i} \times (\delta_t)_i \\ \text{and} \qquad \partial_2\mathrm{ST}(t,\rho)_i &= \begin{cases} 0 & \text{if} \quad |t_i| \leqslant \rho \\ -\operatorname{sign}(t_i) & \text{otherwise} \end{cases} \end{aligned} \tag{25}$$

The closed-form expression we derived for $\partial_1\{\mathrm{Prox}_{\tau G_1}\}(x,y,\lambda)[\delta_x]$ is far from trivial. It is essentially due to [21, 37, 54], see [7] for an expression similar to ours. The generalization of this result to other matrix-valued spectral function has been studied in [17].

Application to matrix completion We illustrate the nuclear norm regularization on a matrix completion problem encountered in recommendation systems such as the popular Netflix problem [3]. We therefore consider $y \in \mathbb{R}^P$ with the forward model $Y = \Phi x_0 + W$, $W \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_P)$, where x_0 is a dense but low-rank (or approximately so) matrix and Φ is a binary masking operator.

We have taken $(n_1, n_2) = (1000, 100)$ and $P = 25000$ observed entries (i.e., 25%). The underlying matrix $x_0 = V_{x_0} \operatorname{diag} \Lambda_{x_0} U_{x_0}^*$ has been chosen with V_{x_0} and U_{x_0} two realizations of the uniform distribution of orthogonal matrices and $\Lambda_{x_0} = (k^{-1})_{1\leqslant k\leqslant n}$ such that x_0 is approximately low-rank with a rapidly decaying spectrum. The binary masking operator is such that for $i = 1, \ldots, P$, $(\Phi x)_i = x_{\Sigma(i)_1,\Sigma(i)_2}$ where $\Sigma : [1, \ldots, n_1 \times n_2] \to [1, \ldots, n_1] \times [1, \ldots, n_2]$ is the realization of a random permutation of the $n_1 \times n_2$ entries of x. The standard

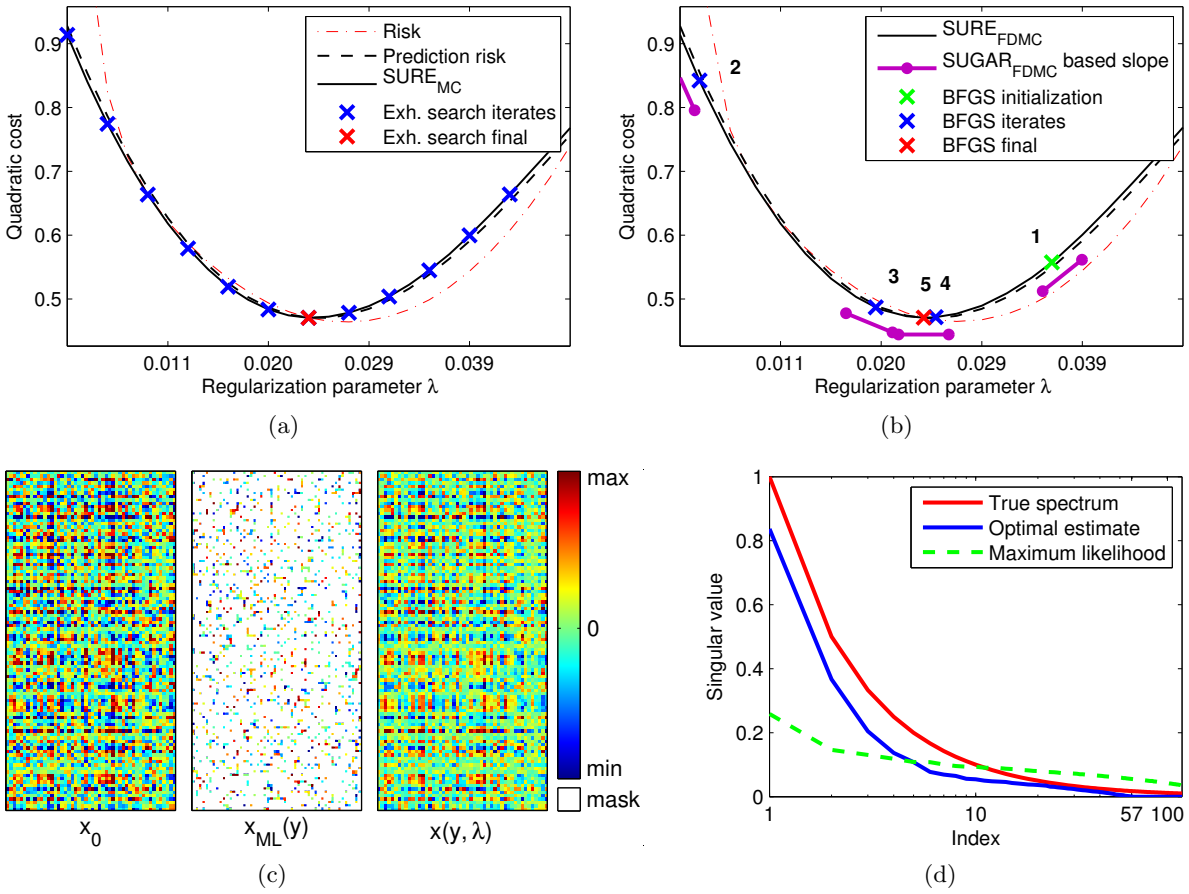


Figure 4: (a-b) Risk, prediction risk and its SURE estimates[‖] as a function of the regularization parameter λ. (a) The 12 points where $\mathrm{SURE}_{\mathrm{MC}}\{x\}(y,\lambda)$ has been evaluated by exhaustive search. (b) The 5 evaluation points of $\mathrm{SURE}_{\mathrm{FDMC}}\{x\}(y,\lambda)$ and $\mathrm{SUGAR}_{\mathrm{FDMC}}\{x\}(y,\lambda)$ required by BFGS to reach the optimal one. (c-d) Respectively, a close in and the spectrum of the underlying matrix x_0, the maximum likelihood estimate $x_{\mathrm{LS}}(y)$ and the solution $x(y,\lambda)$ at the optimal λ.

deviation σ has been set such that the resulting minimum least-square estimate $x_{\mathrm{LS}}(y) = \Phi^* y$ has a relative error $\|x_{\mathrm{LS}}(y) - x_0\|_F \,/\, \|x_0\|_F = 0.9$.

Figure 4.(a) and (b) depict the risk, the prediction risk and the SURE = SURE^A (with $A = \mathrm{Id}$) estimates[‖] as a function of λ obtained from a single realization of y and δ. In (a), $\mathrm{SURE}_{\mathrm{MC}}\{x\}(y,\lambda)$ has been evaluated for 12 values of λ chosen in a suitable tested range using the algorithm given in Figure 2. Figure (b), shows the benefit of computing $\mathrm{SURE}_{\mathrm{FDMC}}\{x\}(y,\lambda)$ and $\mathrm{SUGAR}_{\mathrm{FDMC}}\{x\}(y,\lambda)$, as described in Figure 3, to realize a quasi-Newton optimization. The sequence of iterates λ_n is represented as well as the sequence of the slopes of $\mathrm{SURE}_{\mathrm{FDMC}}\{x\}(y,\lambda_n)$ given by $\mathrm{SUGAR}_{\mathrm{FDMC}}\{x\}(y,\lambda_n)$. The BFGS algorithm reaches to the optimal value in 5 iterations only. One can also notice that $\mathrm{SURE}_{\mathrm{FDMC}}\{x\}(y,\lambda)$ and $\mathrm{SURE}_{\mathrm{MC}}\{x\}(y,\lambda)$ are both good, and visually equivalent, estimators of the prediction risk. At the optimum value $\lambda^\star$ minimizing the SURE, the true risk is not too far from its minimum showing that, in this case, the prediction risk is indeed a good objective in order to minimize the risk. In Figure 4.(c) a close in on the solution $x(y,\lambda^\star)$ is compared to x_0 and

[‖] Without impacting the optimal choice of λ, the curves have been rescaled for visualization purposes.

$x_{\mathrm{LS}}(y)$, and their respective spectrum in Figure 4.(d). The solution $x(y,\lambda^\star)$ has a rank of 57 with a relative error of 0.45 (i.e., a gain of about a factor 2 w.r.t. the least-square estimator).

5.3 Total-Variation Regularization

We consider the recovery of a piece-wise constant two dimensional image $x_0 \in \mathbb{R}^{n_1 \times n_2}$ from an observation y of $Y = \Phi x_0 + W \in \mathbb{R}^P$, $W \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_P)$, where we have identified the image space $\mathbb{R}^{n_1 \times n_2}$ to the vector space $\mathbb{R}^N$ with $N = n_1 n_2$. To this end, we suggest using (isotropic) total variation regularization of the form

$$x^\star(y,\lambda) \in \operatorname*{Argmin}_x \frac{1}{2} \|\Phi x - y\|^2 + \lambda \|\nabla x\|_{1,2} \ , \tag{26}$$

where $\lambda > 0$ and $\nabla : \mathbb{R}^N \to \mathbb{R}^{N \times 2}$ is the two-dimensional discrete gradient operator. The ℓ^1-ℓ^2 norm of a vector field $t = (t_i)_{i=1}^N \in \mathbb{R}^{N \times 2}$, with $t_i \in \mathbb{R}^2$, is defined as $\|t\|_{1,2} = \sum_i \|t_i\|$. Total-variation promotes the sparsity of the gradient field which turns out to be a prior that enforces smoothing while preserving edges. The parameter λ controls the regularity of the image. A large value of λ results to an image with large homogeneous areas while a small value results to an image with several small disconnected regions. The optimal value of λ is image and degradation dependent revealing the importance of automatic selection procedures.

Problem (26) is a special instance of (17) using $x = (f,u) \in \mathcal{X} = \mathbb{R}^N \times \mathbb{R}^{N \times 2}$, the parameter $\lambda = \theta \in \Theta = \mathbb{R}^+$, $Q = 2$ simple functionals, and for $x = (f,u)$

$$\begin{aligned} F(x,y,\lambda) &= \frac{1}{2} \|\Phi f - y\|^2 , \\ G_1(x,y,\lambda) &= \lambda \|u\|_{1,2} , \\ \text{and} \quad G_2(x,y,\lambda) &= \iota_{\mathcal{C}}(x) \quad \text{where} \quad \mathcal{C} = \{x = (f,u) \setminus u = \nabla f\} \ . \end{aligned}$$

Hence the GFB algorithm can be used to solve (26) using

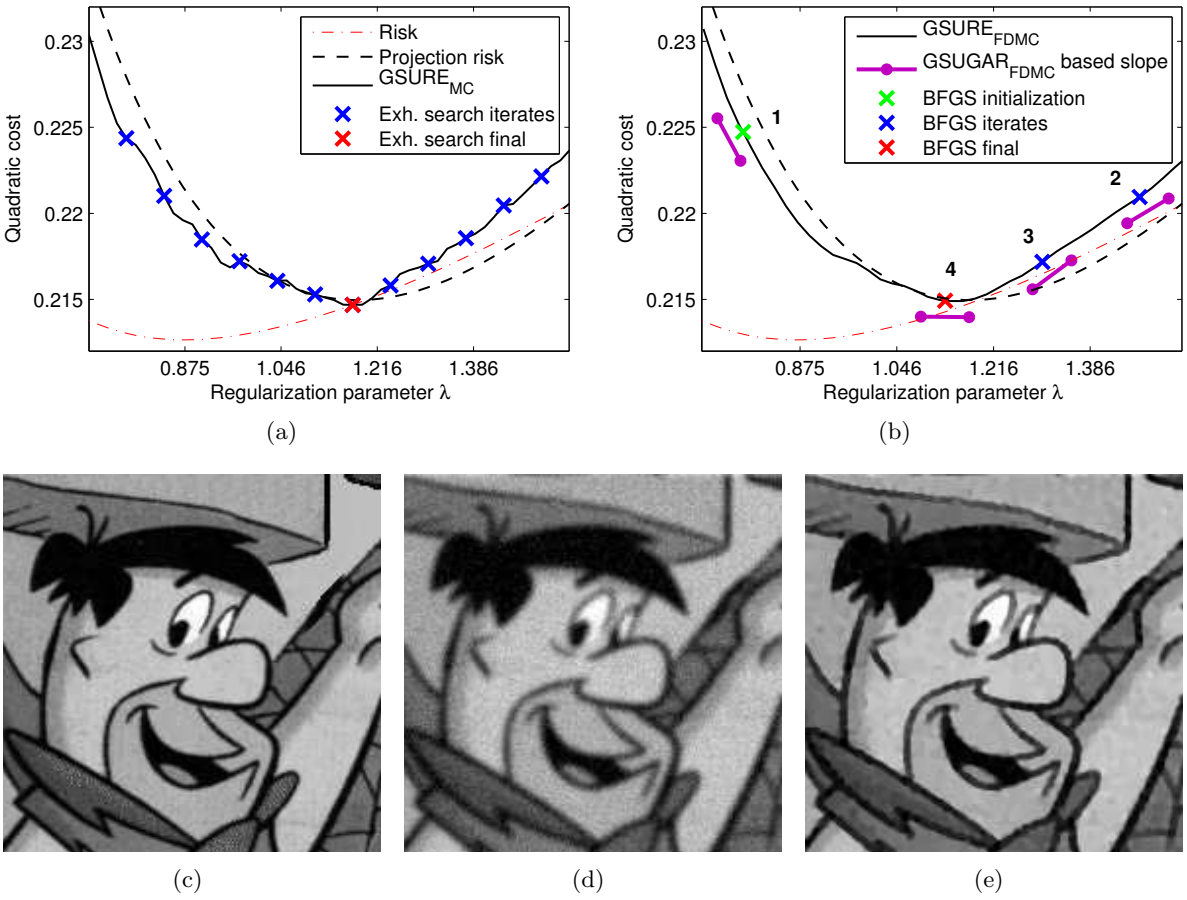
$$\begin{aligned} \nabla_1 F(x,y,\lambda) &= (\Phi^*(\Phi f - y), 0), \\ \mathrm{Prox}_{\tau G_1}(x,y,\lambda) &= (f, \mathrm{ST}_{1,2}(u, \tau\lambda)) \\ \text{and} \quad \mathrm{Prox}_{\tau G_2}(x,y,\lambda) &= ((\mathrm{Id} + \Delta)^{-1}(f + \operatorname{div} u), \nabla(\mathrm{Id} + \Delta)^{-1}(f + \operatorname{div} u)) \end{aligned}$$

where Δ is the Laplacian operator and div is the divergence operator such that $\operatorname{div} = -\nabla^*$. The operator $\mathrm{ST}_{1,2}$ is the component-wise ℓ^1-ℓ^2 soft-thresholding defined, for any dimensions N and D, $t \in \mathbb{R}^{N \times D}$ and $\rho > 0$, by

$$\mathrm{ST}_{1,2}(t,\rho)_i = \begin{cases} 0 & \text{if} \quad \|t_i\| \leqslant \rho \\ t_i - \rho \, t_i / \|t_i\| & \text{otherwise} \end{cases} , \quad \text{for all } 1 \leqslant i \leqslant N. \tag{27}$$

For $D = 1$, the component-wise ℓ^1-ℓ^2 soft-thresholding reduces to (14). Corollary 1 and 2 can then be applied, for any $\delta_x = (\delta_f, \delta_u) \in \mathcal{X}$ and $\delta_y \in \mathcal{Y}$, using the relations

$$\begin{aligned} \partial_1\{\nabla_1 F\}(x,y,\lambda)[\delta_f,\delta_u] &= (\Phi^*\Phi\delta_f, 0), \\ \partial_2\{\nabla_1 F\}(x,y,\lambda)[\delta_y] &= (-\Phi^*\delta_y, 0), \\ \partial_3\{\nabla_1 F\}(x,y,\lambda) &= (0,0), \\ \partial_1\{\mathrm{Prox}_{\tau G_1}\}(x,y,\lambda)[\delta_x] &= (\delta_f, \partial_1 \mathrm{ST}_{1,2}(u,\tau\lambda)[\delta_u]), \\ \partial_2\{\mathrm{Prox}_{\tau G_1}\}(x,y,\lambda)[\delta_y] &= (0,0), \\ \partial_3\{\mathrm{Prox}_{\tau G_1}\}(x,y,\lambda) &= (0, \partial_2 \mathrm{ST}_{1,2}(u,\tau\lambda)), \\ \text{and} \quad \partial_1\{\mathrm{Prox}_{\tau G_2}\}(x,y,\lambda)[\delta_x] &= ((\mathrm{Id} + \Delta)^{-1}(\delta_f + \operatorname{div} \delta_u), \nabla(\mathrm{Id} + \Delta)^{-1}(\delta_f + \operatorname{div} \delta_u)), \\ \partial_2\{\mathrm{Prox}_{\tau G_2}\}(x,y,\lambda)[\delta_y] &= (0,0), \\ \partial_3\{\mathrm{Prox}_{\tau G_2}\}(x,y,\lambda) &= (0,0) \end{aligned}$$



(a) (b)

(c) (d) (e)

Figure 5: (a-b) Risk, projection risk and its GSURE estimates[‖] as a function of the regularization parameter λ. (a) The 12 points where $\mathrm{GSURE}_{\mathrm{MC}}\{x\}(y,\lambda)$ has been evaluated by exhaustive search. (b) The 4 evaluation points of $\mathrm{GSURE}_{\mathrm{FDMC}}\{x\}(y,\lambda)$ and $\mathrm{GSUGAR}_{\mathrm{FDMC}}\{x\}(y,\lambda)$ required by BFGS to reach the optimal one. (c-d) Respectively, a close in of the underlying image x_0, the observation y and the solution $x(y,\lambda)$ at the optimal λ.

where the weak derivatives of the component-wise ℓ^1-ℓ^2 soft-thresholding are defined, for any dimensions N and D, $t \in \mathbb{R}^{N\times D}$, $\rho > 0$ and $\delta_t \in \mathbb{R}^{N\times D}$, by

$$\partial_1 \mathrm{ST}_{1,2}(t,\rho)[\delta_t]_i = \begin{cases} 0 & \text{if} \quad \|t_i\| \leqslant \rho \\ \delta_{t,i} - \frac{\rho}{\|t_i\|} P_{t_i}(\delta_{t,i}) & \text{otherwise} \end{cases} \tag{28}$$

$$\text{and} \qquad \partial_2 \mathrm{ST}_{1,2}(t,\rho)_i = \begin{cases} 0 & \text{if} \quad \|t_i\| \leqslant \rho \\ -t_i/\|t_i\| & \text{otherwise} \end{cases}$$

where P_α is the orthogonal projector on $\alpha^\perp$ for $\alpha \in \mathbb{R}^2$.

Application to image deblurring We illustrate the total-variation regularization on an image deblurring problem. We therefore consider the forward model $Y = \Phi x_0 + W \in \mathbb{R}^P$, $W \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_P)$, where x_0 is a piece-wise constant (or approximately so) image and Φ is a discrete convolution matrix.

We have taken a cartoon-like image of size $(n_1, n_2) = (512, 512)$ and $P = 512^2$ observations corresponding to noisy observations of a convolution product with a discrete Gaussian kernel of width 2 pixels. To ensure numerical stability of the pseudo-inverse (typically for the least-square estimate and the computation of the projection risk and its estimate), the kernel has been truncated in the Fourier domain such that too small contributions have been set to 0. The consequence is that around 80% of (high) frequencies are masked. The standard deviation has been set to $\sigma = 10$ (for an image x_0 with a range $[0, 255]$) such that the resulting minimum least-square estimate $x_{\mathrm{LS}}(y) = \Phi^+ y$ has a peak signal-to-noise ratio (PSNR) equals to $10 \log 10(255^2 / \|x_{\mathrm{LS}}(y) - x_0\|_F^2) = 21.02$ dB.

Figure 5.(a) and (b) display the risk, the projection risk and the GSURE = SURE^A (with $A = \Pi$) estimates as a function of λ obtained from a single realization of y and δ. In (a), $\mathrm{GSURE}_{\mathrm{MC}}\{x\}(y, \lambda)$ has been evaluated for 12 values of λ chosen in a suitable tested range using the algorithm given in Figure 2. Figure (b), shows the benefit of computing $\mathrm{GSURE}_{\mathrm{FDMC}}\{x\}(y, \lambda)$ and $\mathrm{GSUGAR}_{\mathrm{FDMC}}\{x\}(y, \lambda)$, as described in Figure 3, to realize a quasi-Newton optimization. The sequence of iterates λ_n is represented as well as the sequence of the slopes of $\mathrm{GSURE}_{\mathrm{FDMC}}\{x\}(y, \lambda_n)$ given by $\mathrm{GSUGAR}_{\mathrm{FDMC}}\{x\}(y, \lambda_n)$. The BFGS algorithm reaches to the optimal value in 4 iterations. Compared to the experiments in Figure 4, the variation of the risk with respect to the range of λ values, is smaller such that, at this scale, one can visually notice the distinctions between $\mathrm{GSURE}_{\mathrm{FDMC}}\{x\}(y, \lambda)$ and $\mathrm{GSURE}_{\mathrm{MC}}\{x\}(y, \lambda)$. As expected, $\mathrm{GSURE}_{\mathrm{FDMC}}\{x\}(y, \lambda)$ is smoother. Its deviation from the projection risk is of the same order as the deviation of $\mathrm{GSURE}_{\mathrm{MC}}\{x\}(y, \lambda)$. At the optimum value $\lambda^\star$ minimizing the GSURE, the true risk is not too far from its minimum showing that, in this case, the projection risk is indeed a good objective in order to minimize the risk. In Figure 5.(c-e) the solution $x(y, \lambda^\star)$ is compared to x_0 and y. The solution $x(y, \lambda^\star)$ has a PSNR of 24.98 dB (i.e., a gain of about +3.94 dB).

5.4 Weighted ℓ_1-analysis Wavelet Regularization

We focus on the recovery of a piece-wise regular image $x_0 \in \mathbb{R}^{n_1 \times n_2}$ from an observation y of $Y = \Phi x_0 + W \in \mathbb{R}^P$, $W \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_P)$, using a J-scale undecimated wavelet analysis regularization of the form

$$x^\star(y, \lambda) \in \operatorname*{Argmin}_x \frac{1}{2} \|\Phi x - y\|^2 + \|\Psi x\|_{1,\lambda} \quad \text{where} \quad \Psi = \begin{pmatrix} \Psi_1^h \\ \Psi_1^v \\ \vdots \\ \Psi_J^h \\ \Psi_J^v \end{pmatrix} \tag{29}$$

and $\lambda \in \mathbb{R}^{+J}$ and $\Psi \in \mathbb{R}^{2JN \times N}$ is the analysis operator of a two-orientation wavelet transform, where, for all scales $1 \leqslant j \leqslant J$, Ψ_j^h, Ψ_j^v are defined such that, for $x \in \mathbb{R}^N$, $u_j^h = \Psi_j^h x$ and $u_j^v = \Psi_j^v x$ are respectively the vectors of undecimated wavelet coefficients of x in the horizontal and vertical directions at the decomposition level j. The weighted ℓ^1-norm $\|.\|_{1,\lambda}$ is

$$\|\Psi x\|_{1,\lambda} = \sum_{j=1}^{J} \lambda_j \left(\left\|\Psi_j^h x\right\|_1 + \left\|\Psi_j^v x\right\|_1 \right) .$$

Multi-scale wavelet analysis promotes piece-wise regular images by enforcing smoothness while preserving sharp discontinuities at different scales and orientations. Each parameter λ_j controls the regularity at scale j. A large value of λ_j tends to over-smooth structures at scale

Table 1: Illustration of the minimization of $\mathrm{SURE}_{\mathrm{FDMC}}$ in multi-scale regularization obtained for the three images of Figure 6 with different numbers of scales from $J = 1$ to $J = 3$ using either one global parameter or one parameter per scales. For each case, the obtained optimal parameters $\lambda^\star$ are given. The associated value of SURE and the PSNR are compared to neighbors of $\lambda^\star$ located at $0.75\lambda^\star$ and $1.25\lambda^\star$.

Input		Optimal parameters			SURE/PSNR		
Image	PSNR	J	$\dim\Lambda$	$\lambda^\star$	$0.75\lambda^\star$	$\lambda^\star$	$1.25\lambda^\star$
`Mandrill`	17.37	1	1	(7.58)	7.53/24.84	**7.39**/24.90	7.43/**24.94**
		2	1	(5.63)	7.60/24.85	**7.45**/24.88	7.58/**24.89**
		3	1	(4.54)	7.87/24.04	**7.71**/**24.10**	7.83/**24.10**
		2	2	(5.94, 4.24)	7.49/25.02	**7.30**/25.06	7.38/**25.07**
		3	3	(7.51, 1.07, 0.99)	7.37/25.12	**7.22**/25.18	7.33/**25.20**
`House`	17.65	1	1	(18.38)	3.69/**31.16**	**3.51**/31.15	3.68/30.55
		2	1	(11.11)	3.72/31.31	**3.51**/**31.40**	3.81/31.05
		3	1	(8.73)	4.30/30.18	**4.08**/**30.31**	4.43/30.13
		2	2	(14.47, 5.20)	3.53/31.51	**3.34**/**31.57**	3.55/31.05
		3	3	(15.00, 2.50, 2.83)	3.52/31.55	**3.27**/**31.63**	3.44/31.14
`Cameraman`	15.13	1	1	(13.50)	5.29/28.61	**5.09**/**28.73**	5.35/28.64
		2	1	(8.78)	5.34/28.75	**5.09**/**28.83**	5.38/28.72
		3	1	(7.14)	5.84/28.03	**5.60**/**28.06**	5.88/27.99
		2	2	(10.98, 3.74)	5.16/28.91	**4.90**/**29.04**	5.09/28.96
		3	3	(11.56, 3.31, 0.97)	5.07/29.00	**4.86**/**29.11**	5.13/28.99

j, while a small value leads to under-smoothing. The optimal values λ_j are also image and degradation dependent revealing again the importance of automatic selection procedures.

Problem (29) is a special instance of (18) where the parameter $\lambda = \theta \in \Theta = \mathbb{R}^{+J}$, and

$$\begin{aligned} H(x,y,\lambda) &= \tfrac{1}{2}\left\|\Phi x - y\right\|^2, \\ G(u,y,\lambda) &= \left\|x\right\|_{1,\lambda}, \\ \text{and}\quad K(x) &= \Psi x\ . \end{aligned}$$

Hence the primal-dual CP splitting can be used to solve (29) using

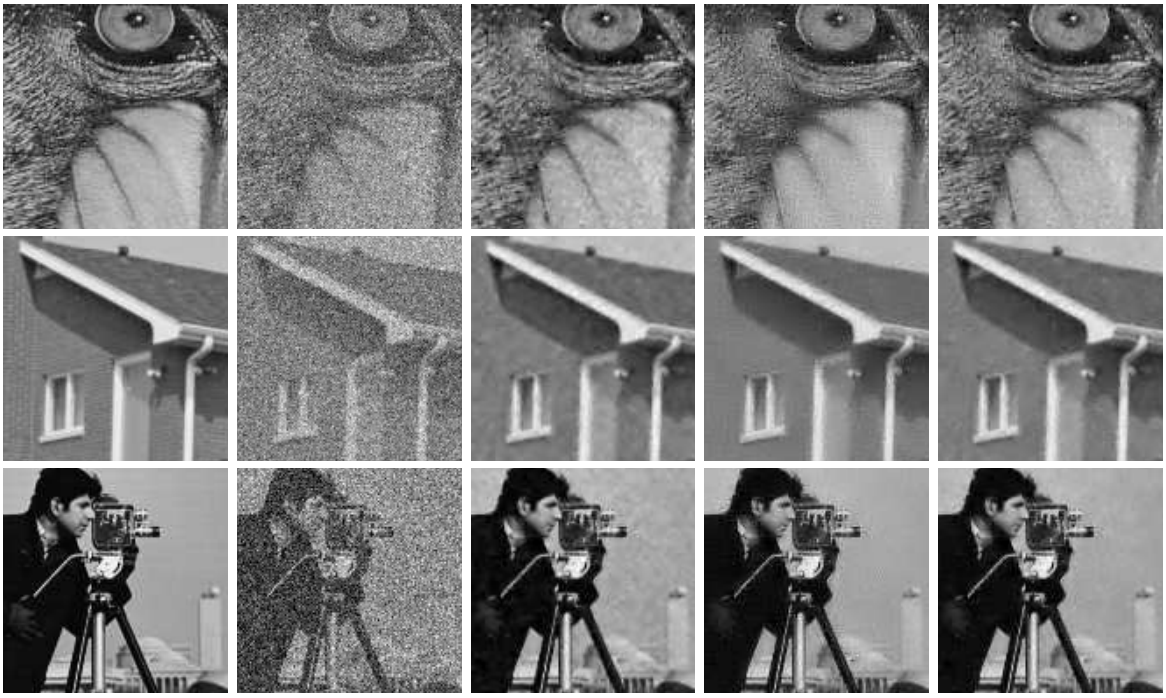
$$\begin{aligned} \mathrm{Prox}_{\xi H}(x,y,\lambda) &= x + \xi\Phi^* y - \xi\Phi^*(\mathrm{Id} + \xi\Phi\Phi^*)^{-1}\Phi(x + \xi\Phi^* y) \\ \mathrm{Prox}_{\tau G^*}(u,y,\lambda) &= u - \tau\mathrm{ST}(u/\tau, \lambda/\tau), \\ \text{and}\quad K^*(u,\lambda) &= \textstyle\sum_{j=1}^J (\Psi_j^{h^*} u_j^h + \Psi_j^{v*} u_j^v) \end{aligned}$$

where ST denotes here the multi-scale extension of the soft-thresholding operator (14), such that, for $t \in \mathbb{R}^{2JN}$ and $\rho \in \mathbb{R}^J$, we have

$$\mathrm{ST}(t,\rho)_j^o = \mathrm{ST}(t_j^o, \rho_j)$$

for all scales $1 \leqslant j \leqslant J$ and orientations $o = v, h$. Corollary 1 and 2 can then be applied using

$$\begin{aligned} \partial_1\{\mathrm{Prox}_{\xi H}\}(x,y,\lambda)[\delta_x] &= \delta_x + \xi\Phi^*(\mathrm{Id} + \xi\Phi\Phi^*)^{-1}\Phi\delta_x \\ \partial_2\{\mathrm{Prox}_{\xi H}\}(x,y,\lambda)[\delta_y] &= \xi\Phi^*\delta_y - \xi^2\Phi^*(\mathrm{Id} + \xi\Phi\Phi^*)^{-1}\Phi\Phi^*\delta_y \\ \partial_3\{\mathrm{Prox}_{\xi H}\}(x,y,\lambda) &= 0 \\ \text{and}\quad \partial_1\{\mathrm{Prox}_{\tau G^*}\}(u,y,\lambda)[\delta_u] &= \delta_u - \partial_1\mathrm{ST}(u/\tau, \lambda/\tau)[\delta_u] \\ \partial_2\{\mathrm{Prox}_{\tau G^*}\}(u,y,\lambda)[\delta_u] &= 0 \\ \partial_3\{\mathrm{Prox}_{\tau G^*}\}(u,y,\lambda) &= -\partial_2\mathrm{ST}(u/\tau, \lambda/\tau)[\delta_u]. \end{aligned}$$



(a) (b) (c) (d) (e)

Figure 6: From top to bottom, a close up on `Mandrill`, `House` and `Cameraman`. (a) Underlying image x_0. (b) Least-squared estimate $x_{\mathrm{LS}}(y)$. (c) Result with $\lambda^\star$ for one level of decomposition $J = 1$, (d) for three levels of decomposition $J = 3$ using one global parameter, and (e) for three levels of decomposition $J = 3$ using one parameter per scales.

where the derivatives of the multi-scale soft-thresholding are defined, for any $t \in \mathbb{R}^{2JN}$, $\rho \in \mathbb{R}^J$ and $\delta_t \in \mathbb{R}^{2JN}$, by

$$\partial_1 \mathrm{ST}(t,\rho)[\delta_t]_j^o = \partial_1 \mathrm{ST}(t_j^o, \rho_j)[\delta_{t_j}^o] \quad \text{and} \quad \partial_2 \mathrm{ST}(t,\rho)_j^o = \partial_2 \mathrm{ST}(t_j^o, \rho_j) \tag{30}$$

for all scales $1 \leqslant j \leqslant J$ and orientations $o = v, h$.

Application to compressed sensing We illustrate the multi-scale wavelet ℓ_1-analysis regularization on a compressed sensing problem. We therefore consider the forward model $Y = \Phi x_0 + W \in \mathbb{R}^P$, $W \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_P)$, where x_0 is a piece-wise multi-scale regular (or approximately so) image and Φ is a random matrix. Here the multi-scale transform W is constructed from undecimated Daubechies 4 wavelets [13].

We have taken a uniformly randomized sub-sampling of a uniform random convolution, where $(P/N = 0.5)$. The standard deviation has been set to $\sigma = 10$ (for an image x_0 with a range $[0, 255]$) such that the resulting minimum least-square estimate $x_{\mathrm{LS}}(y) = \Phi^+ y$ has a PSNR given by $10 \log 10(255^2 / \|x_{\mathrm{LS}}(y) - x_0\|_F) \approx 16$ dB.

Table 1 and Figure 6 illustrates the multi-scale regularization obtained by minimizing the $\mathrm{SURE} = \mathrm{SURE}^A$ (with $A = \mathrm{Id}$) for three different image x_0, known as `Mandrill`, `House` and `Cameraman`, and a single realization of y and δ. Three levels of decomposition from $J = 1$ to 3 are considered. We consider also to use either one global regularization parameter or one parameter per scales. Table 1 gives the selected optimal vector of parameters $\lambda^\star$ for each level of decomposition and their associated performance in terms of SURE and PSNR. We first

observe that compared to the global approach, optimizing one parameter per scale indeed adapts better to the regularity of the image. For instance, the image `Mandrill` contains fine scales with more energy than `House`, and then the obtained penalization of the first scale is smaller for `Mandrill` than for `House`. Visual inspection of these results on Figure 6 illustrates this automatic adaptation. In the same vein, with three levels of decomposition, the penalization is less severe for `Mandrill` than for `House` and `Cameraman`. We next observe that increasing the level of decomposition improves the PSNR when using one parameter per scale, while this is not the case when a global parameter is used. The gap is more important between $J = 1$ and $J = 2$. To assess the minimization of $\mathrm{SURE}_{\mathrm{FDMC}}$, we have compared the SURE and the PSNR values at $0.75\lambda^\star$ and $1.25\lambda^\star$. At the optimal $\lambda^\star$, the SURE is as expected minimal. Furthermore, at $\lambda^\star$, the PSNR is either maximal or not too far from its maximal value, showing that, in this case, the prediction risk is indeed a good objective in order to maximize the PSNR.

6 Conclusion

We have proposed a methodology for optimizing multiple continuous parameters of a weakly differentiable estimator that attempts solving a linear ill-posed inverse problem contaminated by additive white Gaussian noise. The proposed method selects the parameters minimizing an estimate of the risk and is driven by an estimate of its gradient. Classical unbiased estimators of the risk are generally non-continuous functions of the parameters, so that, their local variations cannot be used to estimate the gradient of the risk. These estimators require estimating the degree of freedom by evaluating the variations of the estimator with respect to the observations. We have shown that estimating the degree of freedom by finite differences leads to a weakly differentiable risk estimator. By carefully choosing the finite differences step and by computing explicitly the (weak) gradient of this estimate, an asymptotical unbiased estimator of the gradient of the risk is obtained. This estimator is numerically smooth enough to apply a quasi-Newton method. An explicit strategy to compute this (weak) gradient is given for a large class of (iterative) weakly differentiable algorithms. We exemplified our methodology on several popular proximal splitting methods. Numerical experiments have demonstrated the wide applicability and scope of the approach.

Our choice of the finite differences step size was essentially guided by a careful analysis of the soft-thresholding estimator. Choosing this step size with theoretical guarantees (such as consistency or optimality) in more general cases remains an open question.Beyond consistency and optimality, the question of quantifying the influence of the finite differences step on the smoothness of the risk gradient estimates and then on the performance of quasi-Newton methods is still open. To deal with parameter space of higher dimensions, other accumulation Jacobian strategies could be explored following [30]. Improvements could also be achieved on the settings of the quasi-Newton methods. In particular, a drawback of our approach is the sensitivity to local minima of the risk with respect to the collection of parameters. In some settings, more elaborated optimization strategies could be employed. Future work could also focus on the extensions to non-weakly differentiable estimators and/or inverse problems with non-Gaussian noises.

A Proofs of Section 3

Proposition 1. This is a consequence of the chain rule and linearity of the weak derivative. Indeed, $\widehat{df}^A_{\mathrm{FD}}\{\mu\}(y,\theta,\varepsilon)$ is just the sum of P weakly differentiable functions, and hence is weakly differentiable with the weak derivative with respect to θ as given. Moreover, $\|A(\mu(y,\theta)-y)\|^2 = \sum_{i=1}^P \left((A(\mu(y,\theta)-y))_i\right)^2$. Each term i is the composition of a weakly differentiable function $(A(\mu(y,\cdot)-y))_i$ and $(\cdot)^2$, where the latter is obviously continuously differentiable with bounded derivative, and takes 0 at the origin. It then follows from the chain rule [24, Theorem 4(ii), Section 4.2.2] that $(A(\mu(y,\cdot)-y))_i$ is weakly differentiable, and the weak derivative of $\|A(\mu(y,\cdot)-y)\|^2$ with respect to θ is indeed

$$2\partial_2\mu(y,\theta)^* A^* A(\mu(y,\theta)-y) \ .$$

□

Theorem 1. The proof strategy consists in commuting in an appropriate order the different signs (limit, integration and derivation) while checking that our assumptions provide sufficient conditions for this to hold.

Let V a compact subset of Θ, and choose $\varphi \in C^1_c(\Theta)$ with support in V. We have

$$\begin{aligned}
\int_\Theta R^A\{\mu\}(\mu_0,\theta)\frac{\partial\varphi(\theta)}{\partial\theta_i}d\theta &= \int_V R^A\{\mu\}(\mu_0,\theta)\frac{\partial\varphi(\theta)}{\partial\theta_i}d\theta \\
&= \int_V \mathbb{E}_W\left[\|A(\mu(Y,\theta)-y)\|^2\right]\frac{\partial\varphi(\theta)}{\partial\theta_i}d\theta \\
\text{[Stein Lemma]} \quad &\overset{(S.1)}{=} \int_V \mathbb{E}_W\left[\mathrm{SURE}^A\{\mu\}(Y,\theta)\right]\frac{\partial\varphi(\theta)}{\partial\theta_i}d\theta \\
\text{[(12)]} \quad &\overset{(S.2)}{=} \int_V \mathbb{E}_W\left[\lim_{\varepsilon\to0}\mathrm{SURE}^A_{\mathrm{FD}}\{\mu\}(Y,\theta,\varepsilon)\right]\frac{\partial\varphi(\theta)}{\partial\theta_i}d\theta \\
\text{[Dominated convergence]} \quad &\overset{(S.3)}{=} \lim_{\varepsilon\to0}\int_V \mathbb{E}_W\left[\mathrm{SURE}^A_{\mathrm{FD}}\{\mu\}(Y,\theta,\varepsilon)\right]\frac{\partial\varphi(\theta)}{\partial\theta_i}d\theta \\
\text{[Fubini]} \quad &\overset{(S.4)}{=} \lim_{\varepsilon\to0}\mathbb{E}_W\left[\int_V \mathrm{SURE}^A_{\mathrm{FD}}\{\mu\}(Y,\theta,\varepsilon)\frac{\partial\varphi(\theta)}{\partial\theta_i}d\theta\right] \\
\text{[Weak differentiability, Proposition 1]} \quad &\overset{(S.5)}{=} -\lim_{\varepsilon\to0}\mathbb{E}_W\left[\int_V \frac{\partial}{\partial\theta_i}\mathrm{SURE}^A_{\mathrm{FD}}\{\mu\}(Y,\theta,\varepsilon)\varphi(\theta)d\theta\right] \\
\text{[Proposition 1]} \quad &\overset{(S.6)}{=} -\lim_{\varepsilon\to0}\mathbb{E}_W\left[\int_V \left(\mathrm{SUGAR}^A_{\mathrm{FD}}\{\mu\}(Y,\theta,\varepsilon)\right)_i\varphi(\theta)d\theta\right] \\
\text{[Fubini]} \quad &\overset{(S.7)}{=} -\lim_{\varepsilon\to0}\int_V \mathbb{E}_W\left[\left(\mathrm{SUGAR}^A_{\mathrm{FD}}\{\mu\}(Y,\theta,\varepsilon)\right)_i\right]\varphi(\theta)d\theta \\
\text{[Dominated convergence]} \quad &\overset{(S.8)}{=} -\int_V \left(\lim_{\varepsilon\to0}\mathbb{E}_W\left[\left(\mathrm{SUGAR}^A_{\mathrm{FD}}\{\mu\}(Y,\theta,\varepsilon)\right)_i\right]\right)\varphi(\theta)d\theta \\
&= -\int_\Theta \left(\lim_{\varepsilon\to0}\mathbb{E}_W\left[\left(\mathrm{SUGAR}^A_{\mathrm{FD}}\{\mu\}(Y,\theta,\varepsilon)\right)_i\right]\right)\varphi(\theta)d\theta \ .
\end{aligned}$$

From the definition of weak derivative, we get the claimed result on the asymptotic unbiasedness of $\mathrm{SUGAR}^A_{\mathrm{FD}}$. The asymptotic unbiasedness of the gradient of the finite difference DOF naturally follows with the same proof strategy by ignoring the two first terms in the decomposition $\mathrm{SURE}^A_{\mathrm{FD}}\{\mu\}(\mu_0,\theta,\varepsilon) = \|A(\mu(y,\theta)-y)\|^2 - \sigma^2\,\mathrm{tr}(A^*A) + 2\sigma^2\widehat{df}^A_{\mathrm{FD}}\{\mu\}(\mu_0,\theta,\varepsilon)$.

We now justify each of the steps (S.1)-(S.8). We denote $g_{1,\sigma}$ the Gaussian probability density function of zero-mean and variance σ^2, and g_σ its P-dimensional version, i.e. $g_\sigma = (g_{1,\sigma})^P$.

(S.1) This is Stein lemma which applies owing to Assumption **(A.1)**. Indeed, $\mu(\cdot,\theta)$ is Lipschitz, hence weakly differentiable and its derivative equals its weak derivative Lebesgue a.e. [24, Theorem 1-2, Section 6.2]. Moreover, we have for any θ

$$\left\|\mu(y,\theta)-\mu(y',\theta)\right\| \leqslant L_1 \left\|y-y'\right\| \Rightarrow \left|\mu_i(y,\theta)-\mu_i(y',\theta)\right| \leqslant L_1 \left\|y-y'\right\| \ , \tag{31}$$

and thus, whenever the derivatives of $\mu_i(\cdot,\theta)$ exist, they are bounded by L_1. Consequently,

$$\mathbb{E}_W\left[\left|\frac{\partial \mu_i(Y)}{\partial y_i}\right|\right] \leqslant L_1 \ ,$$

i.e. the weak partial derivatives are essentially bounded.

(S.2) This is just (12) and the arguments justifying hold owing to Assumption **(A.1)**.

(S.3) Let $f_\varepsilon(y,\theta) = \mathrm{SURE}^A_{\mathrm{FD}}\{\mu\}(y,\theta,\varepsilon)$. From (12), $\lim_{\varepsilon\to 0} f_\varepsilon(y,\theta) = \mathrm{SURE}^A\{\mu\}(y,\theta)$ exists Lebesgue a.e. Assumptions **(A.1)**-**(A.2)** give

$$\|\mu(y,\theta)-y\| \leqslant \|y\| + \|\mu(y,\theta)-\mu(0,\theta)\| \leqslant (1+L_1)\|y\| \ .$$

Combining this with (31) leads to

$$\begin{aligned}|f_\varepsilon(y,\theta)| &= \left|\|A(\mu(y,\theta)-y)\|^2 - \sigma^2\operatorname{tr}(A^*A) + 2\sigma^2\frac{1}{\varepsilon}\sum_{i=1}^P (A^*A(\mu(y+\varepsilon e_i,\theta)-\mu(y,\theta)))_i\right| \\ &\leqslant \|A\|^2 P\sigma^2\left((1+L_1)^2\frac{\|y\|^2}{P\sigma^2}+1+2L_1\right) \ .\end{aligned}$$

Note that the bound is independent of θ. Thus

$$\begin{aligned}\mathbb{E}_W\left[\|A\|^2 P\sigma^2\left((1+L_1)^2\frac{\|y\|^2}{P\sigma^2}+1+2L_1\right)\right] = \\ \|A\|^2 P\sigma^2\left((1+L_1)^2\left(\frac{\|\mu_0\|^2}{P\sigma^2}+1\right)+1+2L_1\right) < \infty \ .\end{aligned}$$

This bound together with the fact that φ is continuously differentiable with compact support in V means that $f_\varepsilon \frac{\partial\varphi}{\partial\theta_i}$ is dominated by an integrable function on $\mathcal{Y}\times V$. The dominated convergence then applies which yields the claim.

(S.4) Fubini theorem surely applies in view of the integrability just shown at the end of (S.3).

(S.5) This is a consequence of Proposition 1 and definition of weak differentiability since $\mu(y,\cdot)$ is Lipschitz continuous independently of y.

(S.6) By definition of $\mathrm{SUGAR}^A_{\mathrm{FD}}\{\mu\}$ in Proposition 1.

(S.7) Let $f_\varepsilon(y,\theta) = (\mathrm{SUGAR}^A_{\mathrm{FD}}\{\mu\}(y,\theta,\varepsilon))_i$ and $h(y,\theta) = (2\partial_2\mu(y,\theta)^* A^* A(\mu(y,\theta)-y))_i$. By the translation invariance of the convolution product, we have

$$\mathbb{E}_W[f_\varepsilon(Y,\theta)] = 2(g_\sigma * h(\cdot,\theta))(\mu_0) + 2\sigma^2 \sum_{j=1}^P \frac{g_\sigma(\cdot+\varepsilon e_i) - g_\sigma}{\varepsilon} * \left(\partial_2\mu(\cdot,\theta)^* A^* A e_j\right)_i (\mu_0) \ .$$

Thus

$$\begin{aligned}
|(g_\sigma * h(\cdot,\theta))(\mu_0)| &\leqslant 2\int_{\mathcal{Y}} g_\sigma(y-\mu_0)\, |(\partial_2\mu(y,\theta)^* A^* A(\mu(y,\theta)-y))_i|\, dy \\
[\text{Assumption } \mathbf{(A.3)}] &\leqslant 2L_2 \|A\|^2 \int_{\mathcal{Y}} g_\sigma(y-\mu_0)\, \|\mu(y,\theta)-y\|\, dy \\
[\text{Assumptions } \mathbf{(A.1)}\text{-}\mathbf{(A.2)}] &\leqslant 2(1+L_1)L_2 \|A\|^2 \int_{\mathcal{Y}} g_\sigma(y-\mu_0)\, \|y\|\, dy \\
&\leqslant 2(1+L_1)L_2 \|A\|^2\, \mathbb{E}_W[\|y\|]dy \\
[\text{Jensen inequality}] &\leqslant 2(1+L_1)L_2 \|A\|^2\, \mathbb{E}_W[\|y\|^2]^{1/2}dy \\
&\leqslant 2(1+L_1)L_2 \|A\|^2 \left(\|\mu_0\|^2 + P\sigma^2\right)^{1/2} < \infty \ .
\end{aligned}$$

For the second term, we have

$$\begin{aligned}
&\left|\frac{g_\sigma(\cdot+\varepsilon e_i) - g_\sigma}{\varepsilon} * \left(\partial_2\mu(\cdot,\theta)^* A^* A e_j\right)_i (\mu_0)\right| \\
&\leqslant \int_{\mathcal{Y}} \left|\frac{g_\sigma(y-\mu_0+\varepsilon e_i) - g_\sigma(y-\mu_0)}{\varepsilon}\right| \left|(\partial_2\mu(y,\theta)^* A^* A e_j)_i\right| dy \\
[\text{Assumption } \mathbf{(A.3)}] &\leqslant L_2 \|A\|^2 \int_{\mathcal{Y}} \left|\frac{g_\sigma(y-\mu_0+\varepsilon e_i) - g_\sigma(y-\mu_0)}{\varepsilon}\right| dy \\
&\leqslant L_2 \|A\|^2 \int_{\mathbb{R}} \left|\frac{g_{1,\sigma}(t-(\mu_0)_i+\varepsilon) - g_{1,\sigma}(t-(\mu_0)_i)}{\varepsilon}\right| dt \\
[\text{Taylor}] &\leqslant L_2 \|A\|^2 \int_{\mathbb{R}} \int_0^1 \left|g'_{1,\sigma}(t-(\mu_0)_i+\tau)\right| dt d\tau \\
[\text{Fubini}] &\leqslant L_2 \|A\|^2 \int_{\mathbb{R}} \left|g'_{1,\sigma}(t)\right| dt < \infty \ .
\end{aligned}$$

In view of these bounds, and since φ is compactly supported in V, integrability of $f_\varepsilon\varphi$ on $\mathcal{Y}\times V$ is ensured, whence the claimed result follows.

(S.8) Let f_ε defined as in (S.7). We have just shown that the integrand in θ, i.e. $\mathbb{E}_W[f_\varepsilon(Y,\cdot))_i]\varphi$, is dominated by a function that is integrable on V. It remains to check that its limit exists Lebesgue a.e.. But this is yet again an application of the dominated convergence theorem to the sequence f_ε as an integrand with respect to the Gaussian measure $g_\sigma(y)dy$, which allows to deduce that $\lim_{\varepsilon\to 0}\mathbb{E}_W[f_\varepsilon(Y,\theta)\varphi(\theta)] = \mathbb{E}_W[\lim_{\varepsilon\to 0} f_\varepsilon(Y,\theta)\varphi(\theta)]$.

This completes the proof. □

Proposition 1. For a fixed λ, it can be shown similarly to [24, Theorem 4(iii), Section 4.2.2], that $\mathrm{ST}(\cdot,\lambda)$ is weakly differentiable and that its weak Jacobian $h(y)=\partial_2\mathrm{ST}(y,\lambda)$ is diagonal, with diagonal elements, for $1\leqslant i\leqslant P$,

$$h(y)_i=\begin{cases}+1 & \text{if} \quad y_i\leqslant-\lambda\\ 0 & \text{if} \quad -\lambda<y_i<\lambda\\ -1 & \text{otherwise}\end{cases}\quad . \tag{32}$$

We next define, for a fixed λ, the quantity $h'(y,\varepsilon)=\nabla_2\{\widehat{df}_{\mathrm{FD}}\{\mathrm{ST}\}\}(y,\lambda,\varepsilon)$. Using Proposition 1 and the fact that $\varepsilon<2\lambda$ give

$$h'(y,\varepsilon)=\sum_{i=1}^{P}\frac{h(y+\varepsilon e_i)_i-h(y)_i}{\varepsilon}=\sum_{i=1}^{P}\begin{cases}0 & \text{if} \quad y_i<-\lambda-\varepsilon\\ -1/\varepsilon & \text{if} \quad -\lambda-\varepsilon<y_i<-\lambda\\ 0 & \text{if} \quad -\lambda<y_i<\lambda-\varepsilon\\ -1/\varepsilon & \text{if} \quad \lambda-\varepsilon<y_i<\lambda\\ 0 & \text{if} \quad \lambda<y_i\end{cases}\quad .$$

Computing the expectation and the variance of $h'(Y,\varepsilon)$ in closed-form with truncated Gaussian statistics, and using the fact that h is separable in its arguments, give the proposed formula. □

Theorem 2. For P big enough, $\hat{\varepsilon}(P)<2\lambda$ since $\lim_{P\to\infty}\hat{\varepsilon}(P)=0$. Using the notations in the proof of Lemma 1 leads to

$$\mathrm{SUGAR}_{\mathrm{FD}}\{\mathrm{ST}\}(y,\lambda,\varepsilon)=2h(y)^*(\mathrm{ST}(y,\lambda)-y)+2\sigma^2h'(y,\hat{\varepsilon}(P))\ .$$

The Cauchy-Schwartz inequality implies that

$$\begin{aligned}\mathbb{V}_W\left[\frac{1}{P}\mathrm{SUGAR}_{\mathrm{FD}}\{\mathrm{ST}\}(Y,\lambda,\varepsilon)\right]^{1/2}\leqslant\ & 2\mathbb{V}_W\left[\frac{1}{P}h(y)^*(\mathrm{ST}(Y,\lambda)-Y)\right]^{1/2}\\ &+2\sigma^2\mathbb{V}_W\left[\frac{1}{P}h'(Y,\hat{\varepsilon}(P))\right]^{1/2}\ .\end{aligned}$$

Since $x\mapsto\sqrt{\pi}\,\mathrm{erf}\,(x/a)$ is Lipschitz continuous with a constant of $2/a$, Lemma 1 yields

$$\mathbb{V}_W\left[\frac{1}{P}h'(Y,\hat{\varepsilon}(P))\right]\leqslant\frac{\sqrt{2}}{\sqrt{\pi}\sigma P\hat{\varepsilon}(P)}\ .$$

By assumption, we have $\lim_{P\to\infty}P^{-1}\hat{\varepsilon}(P)^{-1}=0$ and then the variance of $\frac{1}{P}h'(Y,\hat{\varepsilon}(P))$ vanishes to zero. Next, remark that

$$h(y)^*(\mathrm{ST}(y,\lambda)-y)=\lambda\#\{|y|>\lambda\}$$

where $\#\{|y|>\lambda\}$ denotes the number of entries of $|y|$ greater than λ. We have $\#\{|Y_i|>\lambda\}\sim_{\mathrm{iid}}\mathrm{Bernoulli}(p_i)$ whose variance is $p_i(1-p_i)$, where $p_i=\frac{1}{2}\left(\mathrm{erf}\left(\frac{(\mu_0)_i+\lambda}{\sqrt{2}\sigma}\right)-\mathrm{erf}\left(\frac{(\mu_0)_i-\lambda}{\sqrt{2}\sigma}\right)\right)$. It follows that $\mathbb{V}_W[\#\{|Y|>\lambda\}]=\sum_{i=1}^{P}p_i(1-p_i)\leqslant P$ and hence

$$\lim_{P\to\infty}\mathbb{V}_W\left[\frac{1}{P}h(Y)^*(\mathrm{ST}(Y,\lambda)-Y)\right]=\lim_{P\to\infty}\mathbb{V}_W\left[\frac{1}{P}\lambda\#\{|Y|>\lambda\}\right]=0\ .$$

Consistency (i.e. convergence in probability) follows from traditional arguments by invoking Chebyshev inequality and using asymptotic unbiasedness (Theorem 1) and vanishing variance. □

Proposition 2. Developing the mean squared error in terms of bias and variance give the first part of the proposition. Lemma 1 and the fact that $\lim_{\varepsilon\to0}\mathbb{E}_W\nabla_2\{\widehat{df}\{\mathrm{ST}\}\}(Y,\lambda,\varepsilon)=\nabla_2\{df\{\mathrm{ST}\}\}(\mu_0,\lambda)$ concludes the second part. □

B Regularity of the proximal operator of a gauge

We first provide a glimpse of gauges.

Definition 2 (Gauge)**.** *Let $\mathcal{C}$ be a non-empty closed convex set containing the origin. The* gauge *of $\mathcal{C}$ is the function $\gamma_{\mathcal{C}}$ defined by*

$$\gamma_{\mathcal{C}}(y) = \inf \{\omega > 0 \setminus y \in \omega\mathcal{C}\} .$$

As usual, $\gamma_{\mathcal{C}}(y) = +\infty$ if the infimum is not attained.

Definition 3 (Polar set)**.** *Let $\mathcal{C}$ be a non-empty convex set. The set $\mathcal{C}^\circ$ given by*

$$\mathcal{C}^\circ = \left\{ z \in \mathbb{R}^N \setminus \langle z, x\rangle \leqslant 1 \text{ for all } y \in \mathcal{C} \right\}$$

is called the polar *of $\mathcal{C}$. $\mathcal{C}^\circ$ is a non-empty closed convex set containing the origin, and if $\mathcal{C}$ is closed and contains the origin as well, $\mathcal{C}^{\circ\circ} = \mathcal{C}$.*

We now summarize some key properties that will be needed in the main proof.

Lemma 2. *Let $\mathcal{C}$ be a non-empty closed convex set containing the origin. The following assertions hold.*

(i) $\gamma_{\mathcal{C}}$ is a non-negative, closed, convex and positively homogenenous function.

(ii) $\mathcal{C}$ is the unique closed convex set containing the origin such that

$$\mathcal{C} = \{y \in \mathcal{Y} \setminus \gamma_{\mathcal{C}}(y) \leqslant 1\} .$$

(iii) $\gamma_{\mathcal{C}}$ is bounded and coercive if, and only if, $\mathcal{C}$ is compact and contains the origin as an interior point.

(iv) The gauge of $\mathcal{C}$ and the support function $\sigma_{\mathcal{C}^\circ}(y) = \max_{z\in\mathcal{C}^\circ} \langle y, z\rangle$ coincide, i.e.

$$\gamma_{\mathcal{C}} = \sigma_{\mathcal{C}^\circ} .$$

Proof. (i)-(ii) follow from [32, Theorem V.1.2.5]. (iii) is a consequence of [32, Theorem V.1.2.5(ii) and Corollary V.1.2.6]. (iv) [32, Proposition V.3.2.4]. □

We are now equiped to prove our regularity result.

Proposition 5. *Let $\mathcal{C}$ be a compact convex set containing the origin as an interior point, i.e. a convex body, and $G = \gamma_{\mathcal{C}}$ is its gauge. For any $\theta > 0$, $\theta' > 0$ and any $y \in \mathcal{Y}$, the following holds*

$$\|\mathrm{Prox}_{\theta G}(y) - \mathrm{Prox}_{\theta' G}(y)\| \leqslant L_2|\theta - \theta'|$$

for some constant $L_2 > 0$ independent of y, i.e. for any y, $\theta \mapsto \mathrm{Prox}_{\theta G}(y)$ is Lipschitz continuous on $]0, +\infty[$.

Proof. From [31, Proposition 2.3(ii)], we have that for any y, the function $\theta \mapsto \mathrm{Prox}_{\theta G}(y)$ is such that

$$\|\mathrm{Prox}_{\theta G}(y) - \mathrm{Prox}_{\theta' G}(y)\| \leqslant |\theta - \theta'| \, \|y - \mathrm{Prox}_{\theta G}(y)\| / \theta . \tag{33}$$

Now, we have

$$\theta G(y) = \gamma_{\mathcal{C}/\theta}(y) = \sigma_{\theta\mathcal{C}^\circ}(y) , \tag{34}$$

where the first equality follows from positive homogeneity (Lemma 2(i)) and Definition 2, and the second equality is a consequence of Lemma 2(iv) and polarity.

Applying Moreau identity, we get that

$$y - \mathrm{Prox}_{\theta G}(y) = y - \mathrm{Prox}_{\sigma_{\theta\mathcal{C}^\circ}}(y) = \mathrm{Proj}_{\theta\mathcal{C}^\circ}(y) \ .$$

By virtue of Lemma 2(iii), there exists a constant $L_2 > 0$, independent of y, such that**

$$\|y - \mathrm{Prox}_{\theta G}(y)\| = \|\mathrm{Proj}_{\theta\mathcal{C}^\circ}(y)\| \leqslant L_2 \gamma_{\mathcal{C}^\circ}\big(\mathrm{Proj}_{\theta\mathcal{C}^\circ}(y)\big) \ .$$

Applying (34) to $\gamma_{\mathcal{C}^\circ}$, we get

$$\|y - \mathrm{Prox}_{\theta G}(y)\| \leqslant L_2 \theta \gamma_{\theta\mathcal{C}^\circ}\big(\mathrm{Proj}_{\theta\mathcal{C}^\circ}(y)\big) \leqslant L_2\theta \ , \tag{35}$$

where the last inequality follows from Lemma 2(ii) since obviously $\mathrm{Proj}_{\theta\mathcal{C}^\circ}(y) \in \theta\mathcal{C}^\circ$. Combining (33) and (35), we get the desired result. □

Corollary 5. *Let $\mathcal{C}_i$, $i = 1, \ldots, m$, be compact convex sets containing the origin as an interior point, i.e. a convex bodies, and $G_i = \gamma_{\mathcal{C}_i}$ the associated gauges. For any θ, $\theta' \in]0, +\infty[^m$, and any $y \in \mathcal{Y}$, the following holds*

$$\Big\|\mathrm{Prox}_{\theta_1 G_1} \circ \cdots \circ \mathrm{Prox}_{\theta_m G_m}(y) - \mathrm{Prox}_{\theta'_1 G_1} \circ \cdots \circ \mathrm{Prox}_{\theta'_m G_m}(y)\Big\| \leqslant \sqrt{m} \max_i L_{2,i} \left\|\theta - \theta'\right\|$$

where $L_{2,i} > 0$ is the same Lipschitz constant associated to $\mathcal{C}_i$ given in Proposition 5.

Proof. Using repeatedly the triangle inequality, Proposition 5 and the fact that the mapping $y \mapsto \mathrm{Prox}_{\theta_i G_i}(y)$ is 1-Lipschitz [48], we obtain

$$\begin{aligned}
&\Big\|\mathrm{Prox}_{\theta_1 G_1} \circ \cdots \circ \mathrm{Prox}_{\theta_m G_m}(y) - \mathrm{Prox}_{\theta'_1 G_1} \circ \cdots \circ \mathrm{Prox}_{\theta'_m G_m}(y)\Big\| = \\
&\Big\| \Big(\mathrm{Prox}_{\theta_1 G_1} \circ \mathrm{Prox}_{\theta_2 G_2} \circ \cdots \circ \mathrm{Prox}_{\theta_m G_m}(y) - \mathrm{Prox}_{\theta'_1 G_1} \circ \mathrm{Prox}_{\theta_2 G_2} \circ \cdots \circ \mathrm{Prox}_{\theta_m G_m}(y)\Big) + \\
&\Big(\mathrm{Prox}_{\theta'_1 G_1} \circ \mathrm{Prox}_{\theta_2 G_2} \circ \cdots \circ \mathrm{Prox}_{\theta_m G_m}(y) - \mathrm{Prox}_{\theta'_1 G_1} \circ \mathrm{Prox}_{\theta'_2 G_2} \circ \cdots \circ \mathrm{Prox}_{\theta_m G_m}(y)\Big) \Big\| \\
&\leqslant L_{2,1}|\theta_1 - \theta'_1| + \Big\|\mathrm{Prox}_{\theta_2 G_2} \circ \cdots \circ \mathrm{Prox}_{\theta_m G_m}(y) - \mathrm{Prox}_{\theta'_2 G_2} \circ \cdots \circ \mathrm{Prox}_{\theta'_m G_m}(y)\Big\| \\
&\leqslant \sum_i L_{2,i}|\theta_i - \theta'_i| \leqslant \max_i L_{2,i} \left\|\theta - \theta'\right\|_1 \leqslant \sqrt{m} \max_i L_{2,i} \left\|\theta - \theta'\right\|
\end{aligned}$$

as claimed. □

C Proofs of Section 4

Proposition 3. Since (16) is the composition of Lipschitz continuous mappings of y by assumption, applying the chain rule [24, Theorem 4 and Remark, Section 4.2.2] gives the formula. □

Proposition 4. The argument is exactly the same as that for Proposition 3 replacing y by θ where the required Lipschitz continuity assumptions w.r.t. θ hold true. □

**The constant L_2 can be given explicitly by bounding from below the support function of the inscribed ellipsoid of maximal volume, the so-called John ellipsoid. For symmetric convex bodies, L_2 can be made tightest possible. For simplicity, we avoid delving into these technicalities here.

Corollary 1. We first notice that $\mathcal{D}_a^{(\ell)} = (\mathcal{D}_\xi^{(\ell)}, \mathcal{D}_{z_1}^{(\ell)}, \ldots, \mathcal{D}_{z_Q}^{(\ell)})$ where $\mathcal{D}_\xi^{(\ell)} = \partial_1 \xi^{(\ell)}(y,\theta)[\delta]$ and $\mathcal{D}_{z_k}^{(\ell)} = \partial_1 z_k^{(\ell)}(y,\theta)[\delta]$. Hence, applying again the chain rule [24, Theorem 4 and Remark, Section 4.2.2] to the the sequence of iterates and using the fact that all involved mappings are Lipschitz, and $\mathcal{D}_x^{(\ell)} = \Gamma_a^{(\ell)}(\mathcal{D}_a^{(\ell)}) = \mathcal{D}_\xi^{(\ell)}$ concludes the proof. □

Corollary 2. Observe that $\mathcal{J}_a^{(\ell)} = (\mathcal{J}_\xi^{(\ell)}, \mathcal{J}_{z_1}^{(\ell)}, \ldots, \mathcal{J}_{z_Q}^{(\ell)})$ where $\mathcal{J}_\xi^{(\ell)} = \partial_2 \xi^{(\ell)}(y,\theta)$ and $\mathcal{J}_{z_k}^{(\ell)} = \partial_2 z_k^{(\ell)}(y,\theta)$. Arguing as in the proof of Corollary 1, using now that $\mathcal{J}_x^{(\ell)} = \Gamma_a^{(\ell)}(\mathcal{J}_a^{(\ell)}) = \mathcal{J}_\xi^{(\ell)}$ yields the formula. □

Corollary 3. As before, but now with $\mathcal{D}_a^{(\ell)} = (\mathcal{D}_\xi^{(\ell)}, \mathcal{D}_{\tilde{x}}^{(\ell)}, \ldots, \mathcal{D}_u^{(\ell)})$ where $\mathcal{D}_\xi^{(\ell)} = \partial_1 \xi^{(\ell)}(y,\theta)[\delta]$, $\mathcal{D}_{\tilde{x}}^{(\ell)} = \partial_1 \tilde{x}^{(\ell)}(y,\theta)[\delta]$ and $\mathcal{D}_u^{(\ell)} = \partial_1 u^{(\ell)}(y,\theta)[\delta]$. and $\mathcal{D}_x^{(\ell)} = \Gamma_a^{(\ell)}(\mathcal{D}_a^{(\ell)}) = \mathcal{D}_\xi^{(\ell)}$. The chain rule completes the proof. □

Corollary 4. As before, but now with $\mathcal{J}_a^{(\ell)} = (\mathcal{J}_\xi^{(\ell)}, \mathcal{J}_{\tilde{x}}^{(\ell)}, \ldots, \mathcal{J}_u^{(\ell)})$ where $\mathcal{J}_\xi^{(\ell)} = \partial_1 \xi^{(\ell)}(y,\theta)$, $\mathcal{J}_{\tilde{x}}^{(\ell)} = \partial_1 \tilde{x}^{(\ell)}(y,\theta)$ and $\mathcal{J}_u^{(\ell)} = \partial_1 u^{(\ell)}(y,\theta)$. and $\mathcal{J}_x^{(\ell)} = \Gamma_a^{(\ell)}(\mathcal{J}_a^{(\ell)}) = \mathcal{J}_\xi^{(\ell)}$. The chain rule completes the proof. □

References

[1] H. Avron and S. Toledo, *Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix*, Journal of the ACM (JACM), 58 (2011), p. 8.

[2] H. Bauschke and P. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books in Mathematics, Springer, 2011.

[3] J. Bennett and S. Lanning, *The Netflix prize*, in Proceedings of KDD Cup and Workshop, vol. 2007, 2007, p. 35.

[4] T. Blu and F. Luisier, *The SURE-LET approach to image denoising*, IEEE Trans. Image Process., 16 (2007), pp. 2778–2786.

[5] T. Cai and H. Zhou, *A data-driven block thresholding approach to wavelet estimation*, The Annals of Statistics, 37 (2009), pp. 569–595.

[6] E. J. Candès and B. Recht, *Exact matrix completion via convex optimization*, Found. Comput. Math., 9 (2009), pp. 717–772.

[7] E. J. Candes, C. A. Sing-Long, and J. D. Trzasko, *Unbiased risk estimates for singular value thresholding and spectral estimators*, IEEE Transactions on Signal Processing, 61 (2013), pp. 4643–4657.

[8] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, Journal of Mathematical Imaging and Vision, 40 (2011), pp. 120–145.

[9] C. Chaux, L. Duval, A. Benazza-Benyahia, and J.-C. Pesquet, *A nonlinear stein-based estimator for multichannel image denoising*, IEEE Transactions on Signal Processing, 56 (2008), pp. 3855–3870.

[10] P. L. Combettes and J.-C. Pesquet, *A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery*, IEEE J. Selected Topics in Signal Processing, 1 (2007), pp. 564–574.

[11] ——, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, Springer-Verlag, 2011, ch. Proximal Splitting Methods in Signal Processing, pp. 185–212.

[12] P. L. Combettes and V. R. Wajs, *Signal recovery by proximal forward-backward splitting*, SIAM Multiscale Modeling and Simulation, 4 (2005), p. 1168.

[13] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Communications on pure and applied mathematics, 41 (1988), pp. 909–996.

[14] C.-A. Deledalle, V. Duval, and J. Salmon, *Non-local Methods with Shape-Adaptive Patches (NLM-SAP)*, Journal of Mathematical Imaging and Vision, (2011), pp. 1–18.

[15] C.-A. Deledalle, F. Tupin, and L. Denis, *Poisson NL means: Unsupervised non local means for Poisson noise*, in IEEE Int. Conf. Image Process. (ICIP), IEEE, 2010, pp. 801–804.

[16] C.-A. Deledalle, S. Vaiter, G. Peyré, J. Fadili, and C. Dossal, *Proximal splitting derivatives for risk estimation*, in Journal of Physics: Conference Series 386 012003, IOP Publishing, 2012.

[17] ——, *Risk estimation for matrix recovery with spectral regularization*, in arXiv:1205.1482, 2012. Presented at ICML'2012 workshop on Sparsity, Dictionaries and Projections in Machine Learning and Signal Processing, Edinburgh, United Kingdom, 2012.

[18] S. Dong and K. Liu, *Stochastic estimation with¡ i¿ z¡/i¿¡ sub¿ 2¡/sub¿ noise*, Physics Letters B, 328 (1994), pp. 130–136.

[19] D. Donoho and I. Johnstone, *Adapting to Unknown Smoothness Via Wavelet Shrinkage.*, Journal of the American Statistical Association, 90 (1995), pp. 1200–1224.

[20] V. Duval, J.-F. Aujol, and Y. Gousseau, *A bias-variance approach for the non-local means*, SIAM Journal Imaging Sci., 4 (2011), pp. 760–788.

[21] A. Edelman, *Matrix jacobians with wedge products*, MIT Handout for 18.325, (2005).

[22] B. Efron, *How biased is the apparent error rate of a prediction rule?*, Journal of the American Statistical Association, 81 (1986), pp. 461–470.

[23] Y. C. Eldar, *Generalized SURE for exponential families: Applications to regularization*, IEEE Transactions on Signal Processing, 57 (2009), pp. 471–481.

[24] L. C. Evans and R. F. Gariepy, *Measure theory and fine properties of functions*, CRC Press, 1992.

[25] M. Fazel, *Matrix Rank Minimization with Applications*, PhD thesis, Stanford University, 2002.

[26] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Orde*, vol. 517 of Classics in Mathematics, Springer, 2nd ed., 1998.

[27] A. Girard, *A fast monte-carlo cross-validationprocedure for large least squares problems with noisy data*, Numerische Mathematik, 56 (1989), pp. 1–23.

[28] R. Giryes, M. Elad, and Y. Eldar, *The projected GSURE for automatic parameter tuning in iterative shrinkage methods*, Applied and Computational Harmonic Analysis, 30 (2011), pp. 407–422.

[29] G. Golub, M. Heath, and G. Wahba, *Generalized cross-validation as a method for choosing a good ridge parameter*, Technometrics, (1979), pp. 215–223.

[30] A. Griewank and A. Walther, *Evaluating derivatives: principles and techniques of algorithmic differentiation*, Society for Industrial and Applied Mathematics (SIAM), 2008.

[31] H.Attouch and B. F. Svaiter, *A continuous dynamical Newton-like approach to solving monotone inclusions*, SIAM J. Control Optim., 49 (2011), pp. 574–598.

[32] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis And Minimization Algorithms*, vol. I and II, Springer, 2001.

[33] M. Hutchinson, *A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines*, Communications in Statistics-Simulation and Computation, 18 (1989), pp. 1059–1076.

[34] M. Kachour, C. Dossal, J. Fadili, G. Peyré, and C. Chesneau, *The degrees of freedom of the Lasso in underdetermined linear regression models*, in Proc. SPARS 2011, 2011.

[35] K. Kato, *On the degrees of freedom in shrinkage estimation*, Journal of Multivariate Analysis, 100 (2009), pp. 1338–1352.

[36] A. Lewis and M. Overton, *Nonsmooth optimization via BFGS*, Submitted to SIAM J. Optimiz, (2009).

[37] A. Lewis and H. Sendov, *Twice differentiable spectral functions*, SIAM Journal on Matrix Analysis on Matrix Analysis and Applications, 23 (2001), pp. 368–386.

[38] K.-C. Li, *From Stein's unbiased risk estimates to the method of generalized cross validation*, Ann. Statist., 13 (1985), pp. 1352–1377.

[39] F. Luisier, T. Blu, and M. Unser, *Sure-let for orthonormal wavelet-domain video denoising*, Circuits and Systems for Video Technology, IEEE Transactions on, 20 (2010), pp. 913–919.

[40] S. Mallat, *A wavelet tour of signal processing, 3rd edition*, Elsevier, 2009.

[41] U. Naumann, *Optimal jacobian accumulation is np-complete*, Mathematical Programming, 112 (2008), pp. 427–441.

[42] J.-C. Pesquet, A. Benazza-Benyahia, and C. Chaux, *A SURE approach for digital signal/image deconvolution problems*, IEEE Transactions on Signal Processing, 57 (2009), pp. 4616–4632.

[43] H. Raguet, J. Fadili, and G. Peyré, *A generalized forward-backward splitting*, SIAM Journal on Imaging Sciences, 6 (2013), pp. 1199–1226.

[44] S. Ramani, T. Blu, and M. Unser, *Monte-Carlo SURE: a black-box optimization of regularization parameters for general denoising algorithms*, IEEE Trans. Image Process., 17 (2008), pp. 1540–1554.

[45] S. Ramani, Z. Liu, J. Rosen, J.-F. Nielsen, and J. A. Fessler, *Regularization parameter selection for nonlinear iterative image restoration and mri reconstruction using gcv and sure-based methods*, Image Processing, IEEE Transactions on, 21 (2012), pp. 3659–3672.

[46] S. Ramani, J. Rosen, Z. Liu, and J. Fessler, *Iterative weighted risk estimation for nonlinear image restoration with analysis priors*, Proc. SPIE Elec. Img, 8296 (2012), pp. 82960N1–12.

[47] J. Rice, *Choice of smoothing parameter in deconvolution problems*, Contemporary Mathematics, 59 (1986), pp. 137–151.

[48] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM Journal on Control and Optimization, 14 (1976), pp. 877–898.

[49] F. Roosta-Khorasani and U. Ascher, *Improved bounds on sample size for implicit matrix trace estimators*, arXiv preprint arXiv:1308.2475, (2013).

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

[51] X. Shen and J. Ye, *Adaptive model selection*, Journal of the American Statistical Association, 97 (2002), pp. 210–221.

[52] V. Solo and M. Ulfarsson, *Threshold selection for group sparsity*, in Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, IEEE, 2010, pp. 3754–3757.

[53] C. Stein, *Estimation of the mean of a multivariate normal distribution*, The Annals of Statistics, 9 (1981), pp. 1135–1151.

[54] D. Sun and J. Sun, *Nonsmooth matrix valued functions defined by singular values*, tech. rep., Department of Decision Sciences, National University of Singapore, 2003.

[55] R. Tibshirani and J. Taylor, *The solution path of the generalized Lasso*, The Annals of Statistics, 39 (2011), pp. 1335–1371.

[56] ———, *Degrees of freedom in lasso problems*, The Annals of Statistics, 40 (2012), pp. 1198–1232.

[57] S. Vaiter, C.-A. Deledalle, G. Peyré, C. Dossal, and J. Fadili, *Local behavior of sparse analysis regularization: Applications to risk estimation*, Applied and Computational Harmonic Analysis, 35 (2013), pp. 433–451.

[58] S. Vaiter, C.-A. Deledalle, G. Peyré, J. M. Fadili, and C. Dossal, *The degrees of freedom of partly smooth regularizers*, arXiv preprint arXiv:1404.5557, (2014).

[59] D. Van De Ville and M. Kocher, *SURE-based Non-Local Means*, IEEE Signal Process. Lett., 16 (2009), pp. 973–976.

[60] ——, *Non-local means with dimensionality reduction and SURE-based parameter selection*, IEEE Trans. Image Process., 9 (2011), pp. 2683–2690.

[61] C. Vonesch, S. Ramani, and M. Unser, *Recursive risk estimation for non-linear image deconvolution with a wavelet-domain sparsity constraint*, in ICIP, IEEE, 2008, pp. 665–668.

[62] J. Ye, *On measuring and correcting the effects of data mining and model selection*, Journal of the American Statistical Association, (1998), pp. 120–131.

[63] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, J. of The Roy. Stat. Soc. B, 68 (2006), pp. 49–67.

[64] H. Zou, T. Hastie, and R. Tibshirani, *On the "degrees of freedom" of the Lasso*, The Annals of Statistics, 35 (2007), pp. 2173–2192.