Stein Unbiased GrAdient estimator of the Risk (SUGAR) for multiple parameter selection
Charles-Alban Deledalle, Samuel Vaiter, Jalal M. Fadili, Gabriel Peyré

To cite this version:

HAL Id: hal-00987295
https://hal.archives-ouvertes.fr/hal-00987295v2
Submitted on 6 Aug 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Stein Unbiased GrAdient estimator of the Risk (SUGAR) for multiple parameter selection

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

August 6, 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 $\ell_1$-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 [51]. 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, it is proved to be also a consistent estimator. This gradient estimate 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 methods including proximal splitting ones 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, risk estimation, SURE, parameter selection, proximal splitting, sparsity

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

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

(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.
§GREYCY, CNRS-ENSICAEN-Université de Caen, France.
¶Part of this work was completed while the first author was at CEREMADE‡.
where $W \sim \mathcal{N}(0, \sigma^2 \text{Id}_P)$, and the linear imaging operator $\Phi : \mathcal{X} \to \mathcal{Y}$ entails some loss of information. Typically, $P = \dim(\mathcal{Y})$ is smaller than $N = \dim(\mathcal{X})$, or $\Phi$ 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, $\Theta$ 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 $(y, \theta)$ 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 $\theta$ (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 $\theta$.

Recall that for a locally integrable function $f : a \in \Omega \mapsto \mathbb{R}$, $\Omega$ is an open subset of $\mathbb{R}^N$, its weak partial derivative with respect to $a_i$ in $\Omega$ is the locally integrable function $g_i$ on $\Omega$ such that
\[
\int_{\Omega} g_i(a) \varphi(a) da = -\int_{\Omega} f(a) \frac{\partial \varphi(a)}{\partial a_i} da
\]
holds for all functions $\varphi \in C^1_c(\Omega)$, i.e. the space of continuously differentiable functions of compact support. The weak partial derivative, if it exists, is uniquely defined Lebesgue-almost everywhere (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. Similarly, a vector-valued function $h : a \in \mathbb{R}^N \mapsto h(a) = (h_1(a), \ldots, h_P(a)) \in \mathbb{R}^P$ is weakly differentiable if $h_k(a)$ is weakly differentiable $\forall k \in \{1, \ldots, P\}$, 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 $C^1$. A comprehensive account on weak differentiability can be found in e.g. [28,30].

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

- **Given $(y, \theta)$**, consider a minimizer of a convex variational problem of the form
\[
x(y, \theta) = \text{Argmin} \{ E(x, y, \theta) = H(y, \Phi x) + R(x, \theta) \} \tag{2}
\]
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 [46], or gradient [59]. Such regularizers are usually parametrized with a collection of parameters $\theta$. 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 \( \Phi \), see e.g. [66]. 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) [67], hence weakly differentiable [28, Theorem 5, Section 4.2.3].

- Consider now the \( \ell \)-th iterate, denoted by \( x^{(\ell)}(y, \theta) \), of an iterative algorithm converging to a fixed point of an operator acting on \( X \). In this case, \( \theta \) 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, 12] for an overview. Some of these algorithms are considered in detail in Section 4.

The choice of \( \theta \) is generally a challenging task, especially as the dimension of \( \Theta \) gets large. Ideally, one would like to choose the parameters \( \theta^* \) that makes \( \mu(y, \theta^*) \) (or some appropriate image of it) as faithful as possible to \( \mu_0 \) (or some appropriate image of it). Formally, this can be cast as selecting \( \theta^* \) that minimizes the expected reconstruction error (a.k.a., mean-squared error or quadratic risk), i.e.

\[
\theta^* \in \operatorname{Argmin}_{\theta \in \Theta} \left\{ R^A(\mu)(\mu_0, \theta) = \mathbb{E}_W \| A(\mu(Y, \theta) - \mu_0) \|^2 \right\}
\]

where the matrix \( A \in \mathbb{R}^{M \times P} \) is typically chosen to counterbalance the effect of \( \Phi \), 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 \( \theta \). However, this would only apply if \( \mu_0 \) were available. In the context of our observation model (1), \( \mu_0 \) 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 \( \mu_0 \).

Toward this goal, we adopt the framework of the (generalized) Stein Unbiased Risk Estimator (SURE) [27, 48, 55, 62, 66]. For a fixed \( \theta \), the celebrated Stein’s lemma [62] 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 \( y \). This idea have been exploited for years in several statistical and signal processing applications, typically for selecting thresholds in wavelet based reconstruction algorithms, see e.g. [3, 5, 10, 16, 22, 48, 52, 54].

Given such an estimator \( \hat{R}^A(\mu)(y, \theta) \) (see Section 2.1), the idea is so to replace the optimization problem (3) with

\[
\theta^* \in \operatorname{Argmin}_{\theta \in \Theta} \hat{R}^A(\mu)(y, \theta).
\]
Providing that the variance of $\frac{1}{n} \hat{R}^A\{\mu\}(Y, \theta)$ can be made arbitrarily small or even asymptotically vanishing as $P$ increases, so that it becomes a consistent estimator of $\frac{1}{n} R^A\{\mu\}(\mu_0, \theta)$, one can expect that the minimizers of (4) becomes close to those of (3).

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 $\theta \mapsto \hat{R}^A\{\mu\}(y, \theta)$ were sufficiently smooth. To our knowledge, only [17] have performed such an optimization with Newton’s method where $(y, \theta) \mapsto \hat{R}^A\{\mu\}(y, \theta)$ was $C^\infty$. Unfortunately, being a function of $\partial_1 \mu(y, \theta)$, $\theta \mapsto \hat{R}^A\{\mu\}(y, \theta)$ 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 $\hat{R}^A\{\mu\}(y, \theta)$ at different values of $\theta$. Even if in some particular cases this can be done efficiently (see for instance [22]), the computational expense can become prohibitive in general especially as $\dim(\Theta)$ increases.

Derivative-free optimization algorithms have also been investigated (see for instance [51] for the case of 2 parameters). But such approaches typically do not scale up to problems where $\Theta$ has a linear vector space structure with dimension larger than 2. Their performance are known to degrade exponentially with problem size, and they require to compute a lower and an upper bound on the optimal value over a given region.

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 $\theta$ 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 [51] 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 $\partial_2 \mu(y, \theta)$. Such optimization technique can be provably faster thanks to first-order information compared to derivative-free approaches. 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 it 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 [13], Douglas-Rachford [11] and generalized forward-backward [50]) and primal-dual [9] algorithms (see [41] for a recent review). 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 [33]). From \cite{26,66}, the degrees of freedom of a function \( y \mapsto \mu(y,\theta) \) relatively to a matrix \( A \in \mathbb{R}^{M \times P} \) is given by

\[
df^A\{\mu\}(\mu_0,\theta) = \sum_{i=1}^{P} \frac{\text{cov}(AY_i,(A\mu(Y,\theta)))_i}{\sigma^2},
\]

such that \( df^A\{\mu\}(\mu_0,\theta) \) is maximal when \( A\mu(Y,\theta) \) is highly correlated with the random vector \( AY \). Taking \( A = \text{Id} \), leads to the standard definition of the DOF defined in the seminal work of Efron \cite{26}. But other choices of \( A \) allow to counterbalance the undesirable effect of the linear operator \( \Phi \) (recall that \( \mu_0 = \Phi x_0 \)). For instance, setting \( A = (\Phi^*\Phi)^{-1}\Phi^* \) when \( \Phi \) has full-rank, or \( A = \Phi^*(\Phi\Phi^*)^+ \) when \( \Phi \) is rank deficient*, provides a measure of the DOF relatively to the least-squares estimate of \( x_0 \), i.e. \( x_{\text{LS}}(y) = Ay \) \cite{27,48,66}.

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 \cite{53}) given as

\[
\text{SURE}^A\{\mu\}(y,\theta) = \|A(\mu(y,\theta) - y)\|^2 - \sigma^2 \text{tr}(A^*A) + 2\sigma^2 df^A\{\mu\}(y,\theta) \quad (6)
\]

with

\[
\tilde{df}^A\{\mu\}(y,\theta) = \text{tr}(A\partial_1(\mu(y,\theta))A^*).
\]

where we recall that \( \partial_1(\mu(y,\theta)) \) is the weak Jacobian of \( \mu(y,\theta) \) with respect to the first argument \( y \). It can be shown that (see e.g. \cite{27,62,66})

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

Expression (6) is general enough to encompass unbiased estimates of the prediction risk \( \mathbb{E}_W \|\mu(Y,\theta) - \mu_0\|^2 \) (i.e. \( A = \text{Id} \)), the projection risk \( \mathbb{E}_W \|\Pi(x(Y,\theta) - x_0)\|^2 \), where \( \Pi \) 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 \( \Phi \) has full rank (i.e. \( A = (\Phi^*\Phi)^{-1}\Phi^* \)). This can prove useful when \( \Phi \) is rank deficient, since in this case, the minimizers of the prediction risk can be far away from the minimizers of the estimation risk \cite{56}. 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 \cite{27}.

Note that generalization of the SURE have been developed for other noise models, typically within the multivariate canonical exponential family (see e.g. \cite{27,37,38,55}).

Applications of SURE emerged for choosing the smoothing parameters in families of linear estimates \cite{44} such as for model selection, ridge regression, smoothing splines, etc. After its introduction in the wavelet community with the SURE-Shrink algorithm \cite{22}, it has been widely used for various image restoration problems, e.g. with sparse regularizations \cite{3,5,6,10,16,45,48,51-54,70} or with non-local filters \cite{15,24,68,69}.

*\((\cdot)^+\) stands for the Moore-Penrose pseudo-inverse.
However, a major practical difficulty when using the SURE lies in the numerical computation of the DOF estimate, i.e. the quantity $\hat{df}^A\{\mu\}(y, \theta)$ 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 $\hat{df}^A\{\mu\}(Y, \theta)$ 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.

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 $\ell^1$-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 [23, 65, 73]. These results have in turn been extended to more general sparsity promoting regularizations [20, 40, 61, 64–66, 72], and spectral regularizations (e.g. nuclear norm) [8, 19].

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 $A = \text{Id}$ and $\Phi = \text{Id}$, this trace can generally be obtained by closed-form computations of the $P$ diagonal elements of $\partial_1 \mu(y, \theta)$ (see e.g. [22, 68]). This can also be done for some particular inverse problems. For instance, the authors of [48] provide an expression of this trace for the wavelet-vaguelette estimator when $\Phi$ 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 $\Phi$ a mixing operator or $\mu$ an iterative estimator). To avoid such a costly procedure, the authors of [31, 51] suggest making use of the following trace equality

$$\hat{df}^A\{\mu\}(y, \theta) = \text{tr} (A \partial_1 \mu(y, \theta) A^*) = \mathbb{E}_\Delta \langle \partial_1 \mu(y, \theta)[\Delta], A^*A\Delta \rangle$$

where $\Delta \sim \mathcal{N}(0, \text{Id}_P)$ 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 $\delta$. Remark that $\Delta$ does not necessary have to be Gaussian and higher precisions can be reached in some specific cases, see for instance [1, 21, 39, 58]. As shown in [58], 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,
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 $\mu^{(i)}(y, \theta)$ that eventually converge to some $\mu(y, \theta)$ as $\ell \to +\infty$. It appears then natural to estimate the DOF of $\mu^{(i)}(Y, \theta)$ 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 $\mu^{(i)}(Y, \theta)$ by evaluating its directional derivative $\partial_1 \mu^{(i)}(y, \theta)[\delta]$. A practical way, initiated by [70], to compute this quantity, consists in recursively differentiating the sequence of iterates. The authors of [70] 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 $D_\mu^{(\ell+1)} = \partial_1 \mu^{(\ell+1)}(y, \theta)[\delta]$, is obtained iteratively as a function of $\mu^{(\ell)}(y, \theta)$ and $D_\mu^{(\ell)} = \partial_1 \mu^{(\ell)}(y, \theta)[\delta]$. 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

$$\text{SURE}_\text{MC}^A\{\mu^{(\ell)}\}(y, \theta, \delta) = \|A(\mu^{(\ell)}(y, \theta) - y)\|^2 - \sigma^2 \text{tr}(A^*A) + 2\sigma^2 \hat{d}_\text{MC}^A\{\mu^{(\ell)}\}(y, \theta, \delta)$$

with $\hat{d}_\text{MC}^A\{\mu^{(\ell)}\}(y, \theta, \delta) = \langle \partial_1 \mu^{(\ell)}(y, \theta)[\delta], A^*\delta \rangle$.

A similar approach is described in [32]. Pursuing this idea, the authors of [52,53] have recently provided such closed-form expressions in the case of the split Bregman method. Concurrently, in an early short version of this paper [18], 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 [60, 71] and rediscovered in [51] consists in estimating $\text{tr}(A \partial_1 \mu(y, \theta)A^*)$ via finite differences given, for $\varepsilon > 0$, by

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

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

$$\text{SURE}_\text{FD}^A\{\mu\}(y, \theta, \varepsilon) = \|A(\mu(y, \theta) - y)\|^2 - \sigma^2 \text{tr}(A^*A) + 2\sigma^2 \hat{d}_\text{FD}^A\{\mu\}(y, \theta, \varepsilon)$$

with $\hat{d}_\text{FD}^A\{\mu\}(y, \theta, \varepsilon) = \frac{1}{\varepsilon} \sum_{i=1}^P (A^*A(\mu(y + \varepsilon e_i, \theta) - \mu(y, \theta)))_i$.  \(\text{(11)}\)
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 \(\varepsilon\) 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. [28, Theorem 1-2, Section 6.2], which in turn implies
\[
\lim_{\varepsilon \to 0} \operatorname{SURE}^A_{\text{FD}}(\mu)(y, \theta, \varepsilon) = \operatorname{SURE}^A(\mu)(y, \theta) \quad \text{Lebesgue a.e.} \quad (12)
\]
The value of \(\varepsilon\) can so be chosen as small as possible as soon as it does not rise to numerical instabilities due to limited machine precision. To avoid numerical instabilities, \(\varepsilon\) should be chosen in a reasonable range of values with respect to the amplitudes of the data. In practice, we observe that precise results can be reached compared to the closed-form derivation, hence 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) \(\operatorname{SURE}^A\) given by
\[
\operatorname{SURE}^A_{\text{FD}}(\mu)(y, \theta, \varepsilon) = \|A(\mu(y, \theta) - y)\|^2 - \sigma^2 \text{tr}(A^*A) + 2\sigma^2 \hat{d}^A_{\text{FD}}(\mu)(y, \theta, \varepsilon)
\]
with \(\hat{d}^A_{\text{FD}}(\mu)(y, \theta, \varepsilon) = \frac{1}{\varepsilon} (\mu(y + \varepsilon \delta, \theta) - \mu(y, \theta), A^*A\delta).
\]

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 \(\operatorname{SURE}^A(\mu)\), \(\operatorname{SURE}^A_{\text{FD}}(\mu)\) is a smooth function of \(\theta\) 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 controlled 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 \(\theta \mapsto \operatorname{SURE}^A(\mu)(y, \theta)\) might contain discontinuities. Typically, \(\hat{d}^A_{\text{FD}}(\mu)(y, \theta)\) has discontinuities where \((y, \theta) \mapsto \mu(y, \theta)\) is not differentiable.

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

**Proposition 1.** Assume \(\mu(y, \theta)\) is weakly differentiable with respect to \(y\) and \(\theta\). Given \(\varepsilon > 0\), \(\hat{d}^A_{\text{FD}}(\mu)(y, \theta, \varepsilon)\) and \(\operatorname{SURE}^A_{\text{FD}}(\mu)(y, \theta, \varepsilon)\) are also weakly differentiable with respect to \(y\) and \(\theta\), and their (weak) gradients with respect to \(\theta\) are given, for almost all \(\theta \in \Theta\), as
\[
\operatorname{SUGAR}^A_{\text{FD}}(\mu)(y, \theta, \varepsilon) = \nabla_{\theta} \{ \operatorname{SURE}^A_{\text{FD}}(\mu) \}(y, \theta, \varepsilon)
\]
\[
= 2\partial_2 \mu(y, \theta)^* A^*A(\mu(y, \theta) - y) + 2\sigma^2 \nabla \{ \hat{d}^A_{\text{FD}}(\mu) \}(y, \theta, \varepsilon)
\]
where \(\nabla \{ \hat{d}^A_{\text{FD}}(\mu) \}(y, \theta, \varepsilon) = \frac{1}{\varepsilon} \sum_{i=1}^{P} \left( \partial_2 \mu(y + \varepsilon e_i, \theta) - \partial_2 \mu(y, \theta) \right)^* A^*A e_i\).
Thanks to Proposition 1, a quasi-Newton-like method can now be used to optimize $\text{SURE}_{\text{FD}}^A\{\mu\}(y, \theta, \varepsilon)$ for the vector of continuous parameters $\theta$ by implementing the iteration

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

where $B_n \in \mathbb{R}^{\dim(\Theta) \times \dim(\Theta)}$ is a sequence of definite-positive matrices. Typically, if $\theta \mapsto \text{SURE}_{\text{FD}}^A\{\mu\}(y, \theta, \delta, \varepsilon)$ behaves locally as a $C^2$ function, $B_n$ should approach the inverse of the corresponding Hessian at $\theta_n$. Remark that in general there is no guarantee that the risk has a unique global minimizer, though in the 1-D case it is generally the case. When several parameters are involved, such objective can have several local minima and specific quasi-Newton-like methods might be developed to avoid being stuck in one of them.

In practice, the calculation of SUGAR$_{\text{FD}}^A$ depends on the computation of the Jacobian matrices with respect to the parameters $\theta$. We will see in Section 4 how this quantity can be efficiently computed when $\mu$ results from an iterative algorithm.

We now turn to the asymptotic unbiasedness of the proposed gradient estimator of the risk as $\varepsilon$ 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 $\theta$.

(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, as considered in Section 4 (see also Section 3.2 for soft-thresholding). Standard convex analysis arguments [36] show that the proximity operator is indeed a uniformly Lipschitz of its argument $y$ with constant $L_1 = 1$, independently of $\theta$.

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 $\mu(y, \theta) = \text{Prox}_{\theta G}(y)$, $\theta > 0$, 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 $\mu(y, \theta) = \text{Prox}_{\theta_1 G_1} \circ \cdots \circ \text{Prox}_{\theta_m G_m}(y)$, $\theta \in [0, +\infty]^m$, 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. $\ell_1$, $\ell_1 - \ell_2$ 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.

---

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

\[
\lim_{\varepsilon \to 0} E_W \left[ \text{SUGAR}^A_{\text{FD}}\{\mu\}\{Y, \theta, \varepsilon\} \right] = \nabla^2 \{ R^A\{\mu\}\}(\mu_0, \theta) \]

and

\[
\lim_{\varepsilon \to 0} E_W \left[ \nabla^2 \{ df^A\{\mu\}\}(Y, \theta, \varepsilon) \right] = \nabla^2 \{ df^A\{\mu\}\}(\mu_0, \theta) .
\]

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

However, even if \( \varepsilon \) should decrease towards 0, it should not decrease too fast. In particular, for a fixed dimension \( P \), the step \( \varepsilon \) 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 SUGAR\(_{\text{FD}}^A\{\mu\}\{y, \theta, \varepsilon\} \) 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 \( \frac{1}{P}\text{SUGAR}_{\text{FD}}^A\{\mu\}\{Y, \theta, \varepsilon\} \) can become very unstable when \( \varepsilon \) decreases too fast with the dimension \( P \). The underlying statistical question is whether one can control the variance of \( \frac{1}{P}\text{SUGAR}_{\text{FD}}^A\{\mu\}\{Y, \theta, \varepsilon\} \) as \( P \) increases, and make arbitrarily small or even asymptotically vanishing, so that \( \frac{1}{P}\text{SUGAR}_{\text{FD}}^A\{\mu\}\{Y, \theta, \varepsilon\} \) becomes a consistent estimator. Unfortunately, consistency of our gradient estimator of the risk is very intricate to get in the general case, as it is the case for the consistency of the SURE. However, when \( \mu \) specializes to soft-thresholding, such a result can be achieved.

### 3.2 SUGAR for Soft-thresholding

In this section, we show that the proposed gradient estimator of the risk can be consistent in the case where \( \mu \) is the soft-thresholding (ST) function and \( A = \text{Id}_P \). The ST is the proximal operator of the \( \ell_1 \)-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 \( 1 \leq i \leq P \), as

\[
\text{ST}(y, \lambda)_i = \begin{cases} y_i + \lambda & \text{if } y_i \leq -\lambda \\ 0 & \text{if } -\lambda < y_i < \lambda \\ y_i - \lambda & \text{otherwise} \end{cases} .
\]  

(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 our gradient estimator of the soft-thresholding risk is asymptotically unbiased.

We start with following lemma which collects the statistics of the gradient of the finite difference DOF estimator.

\footnote{as we will see, the higher the dimension \( P \), the smaller \( \varepsilon \) could be.}
Lemma 1 (Statistics of the gradient of the finite difference DOF estimator). Let $0 < \varepsilon < 2\lambda$. The weak gradient of $\lambda \mapsto \hat{d}_F \{ST\}(Y, \lambda, \varepsilon)$ is such that

\[
\mathbb{E}_W \left[ \nabla_2 \{\hat{d}_F \{ST\}\}(Y, \lambda, \varepsilon) \right] = -\frac{1}{2} \sum_{i=1}^{P} \frac{\varphi[(\mu_0)_i, \lambda, \varepsilon]}{\varepsilon},
\]

and

\[
\forall_W \left[ \nabla_2 \{\hat{d}_F \{ST\}\}(Y, \lambda, \varepsilon) \right] = \frac{1}{2\varepsilon} \sum_{i=1}^{P} \frac{\varphi[(\mu_0)_i, \lambda, \varepsilon]}{\varepsilon} - \frac{1}{4} \sum_{i=1}^{P} \left( \frac{\varphi[(\mu_0)_i, \lambda, \varepsilon]}{\varepsilon} \right)^2.
\]

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

We now turn to the asymptotic behavior of the proposed gradient estimator of the risk 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 $\{\Psi_p\}_{p \geq 1}$ where $\Psi_p \in \mathbb{R}^{P \times N}$ and such that, for all $P > 1$, $\Psi_p$ is the sub-matrix obtained by cutting down one line of $\Psi_{p+1}$. We can then define a sequence of observation models as a the sequence of random vector $\{Y_p\}_{P \geq 1}$ defined as

\[
Y_p = \Psi_p z_0 + W_p \quad \text{where} \quad W_p \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_p).
\]

We also define the sequence $\{(\mu_p)_p\}_{P \geq 1}$ where $(\mu_0)_p = \Psi_p z_0$. In the following, for the sake of clarity, we omit the dependency of $Y_p$, $W_p$ and $(\mu_0)_p$ on $P$.

We can now state our consistency result 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 $\lambda \neq |Y_i + \hat{\varepsilon}(P) e_i|$)

\[
\lim_{P \to \infty} \frac{1}{P} \left( \text{SUGAR}_F \{ST\}(Y, \lambda, \hat{\varepsilon}(P)) - \nabla_2 \{R \{ST\}\}(\mu_0, \lambda) \right) = 0
\]

and

\[
\lim_{P \to \infty} \frac{1}{P} \left( \nabla_2 \{\hat{d}_F \{ST\}\}(Y, \lambda, \hat{\varepsilon}(P)) - \nabla_2 \{d_f \{ST\}\}(\mu_0, \lambda) \right) = 0.
\]

In plain words, Theorem 2 asserts that for our gradient estimator of the soft-thresholding risk 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, $\frac{1}{P} \text{SUGAR}_F \{ST\}(y, \lambda, \hat{\varepsilon}(P))$ is guaranteed to come close to $\frac{1}{P} \nabla_2 \{R \{ST\}\}(\mu_0, \lambda)$ 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 $\varepsilon$ through a bias-variance trade-off.

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

\[
\mathbb{E}_W \left[ \frac{1}{P} \left( \nabla_2 \{\hat{d}_f \{ST\}\}(Y, \lambda, \varepsilon) - \nabla_2 \{d_f \{ST\}\}(\mu_0, \lambda) \right)^2 \right] = \frac{1}{P^2} \mathbb{E}_W \left[ \nabla_2 \{\hat{d}_f \{ST\}\}(Y, \lambda, \varepsilon) - \nabla_2 \{d_f \{ST\}\}(\mu_0, \lambda) \right]^2 + \frac{1}{P^2} \forall_W \left[ \nabla_2 \{\hat{d}_f \{ST\}\}(Y, \lambda, \varepsilon) \right].
\]
where the statistics of $\nabla_2\{\hat{d}\{ST\}\}(Y,\lambda,\varepsilon)$ are given in Lemma 1 and

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

Thus, if $\mu_0$ 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 $\varepsilon$ for a fixed dimension $P$ and a given threshold $\lambda$. See the following numerical experiments which illustrate this relationship. When $\mu_0$ is unknown, an a priori model can be imposed, such as for instance belonging to some ball promoting sparsity, e.g. a weak $\ell_\gamma$-ball for $\gamma > 0$. For $\gamma$ sufficiently small, this ball corresponds to compressible or nearly sparse vectors $\mu_0$ whose entries $|\mu_i|$ sorted in descending order of magnitude behave as $O(t^{-1/\gamma})$. With such a model at hand, the MSE in Proposition 2 can be optimized for $\varepsilon$ given $P$, $\sigma$, $\lambda$ and $\gamma$. This however entails a highly non-linear equation...
that cannot be solved in closed form. We defer such a development to a future work.

Figure 1.(a) shows the evolution of the bias and the variance as a function of the ratio \( \varepsilon/\sigma \) for fixed values of \( \sigma, \lambda \) and a compressible vector \( \mu_0 \), i.e. \(|\langle \mu_0 \rangle_i| = O(i^{-1/\gamma})\), 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 \( 0.1 \sigma \). Figure 1.(c) shows the evolution of the MSE as a function of the dimension \( P \) and the ratio \( \varepsilon/\sigma \) 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 \( \varepsilon^*(P) = C\sigma/P^\alpha \) with \( C > 0 \) and \( 0 < \alpha < 1 \). Of course the optimal constants \( C \) and \( \alpha \) depend on the choice of \( \mu_0, \sigma \) and \( \lambda \). However, whatever \( C > 0 \) and \( 0 < \alpha < 1 \), or more generally for any admissible choice of \( \varepsilon \) such that \( \lim_{P \to \infty} \varepsilon(P) = 0 \) and \( \lim_{P \to \infty} P^{-1}\varepsilon(P)^{-1} = 0 \), 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 \( \lim_{P \to \infty} P^{-1}\varepsilon(P)^{-1} = 0 \).

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 \( \varepsilon^*(P) = C\sigma/P^\alpha \), as suggested by our study on the soft thresholding, resulted in a reliable way to parametrize 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 \( \ell \) 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 \( \delta \)) or \( \theta \). This proves useful to respectively estimate the risk via SURE\(_A\) (see Section 2) and estimate its gradient via SUGAR\(_{\text{FDMC}}\) (see Section 3).

The iterative schemes we consider can be cast in the same framework, which subsumes 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

\[
\begin{align*}
    x^{(\ell)} &= \gamma(a^{(\ell)}) \\
    a^{(\ell+1)} &= \psi(a^{(\ell)}, y, \theta),
\end{align*}
\]

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

To make our ideas clear, consider the instructive example where \( x \mapsto E(x, y, \theta) \) is convex and \( C^1(X) \) with \( L \)-Lipschitz gradient, in which case \( A = X \), \( a = x \) 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.
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 \( x(y, \theta) \in X \) and its risk estimate \( \hat{R}^A_x \{ y \} (y, \theta) \)

Sample a vector \( \delta \) from \( \mathcal{N}(0, \text{Id}_P) \)

Initialize \( a^{(0)} \leftarrow 0 \)

Initialize \( D_a^{(0)} \leftarrow 0 \)

for \( \ell \) from 0 to \( L - 1 \) do

\[
\begin{align*}
 a^{(\ell+1)} & \leftarrow \psi(a^{(\ell)}, y, \theta) \\
 D_a^{(\ell+1)} & \leftarrow \Psi_a^{(\ell)}(D_a^{(\ell)}) + \Psi_y^{(\ell)}(\delta)
\end{align*}
\]

end for

\[
\begin{align*}
 x^{(L)} & \leftarrow \gamma(a^{(L)}) \\
 D_x^{(L)} & \leftarrow \Gamma_a^{(L)}(D_a^{(L)}) \\
 \hat{d}^A_{\text{MC}} & \leftarrow \langle \Phi D_x^{(L)}, A^* \delta \rangle \\
 \text{SURE}^A_{\text{MC}} & \leftarrow \| A(y - \Phi x^{(L)}) \|^2 - \sigma^2 \text{tr}(A^* A) + 2\sigma^2 \hat{d}^A_{\text{MC}} \\
 \text{return} \quad x(y, \theta) \leftarrow x^{(L)} \quad \text{and} \quad \hat{R}^A_x \{ y \} (y, \theta) \leftarrow \text{SURE}^A_{\text{MC}}
\end{align*}
\]

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 \( \hat{R}^A \) using Monte Carlo simulation. Even if computing the risk requires more operations, the global complexity of the algorithm is unchanged.

However, two major issues have to be taken care of. First, one has to ensure weak differentiability 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 \( x^* \).

Regarding weak differentiability of the iterates, it relies essentially on regularity conditions to apply the chain rule, e.g. [28, Section 4.2.2], i.e. regularity properties of the iteration mappings \( \gamma \) and \( \psi \) and of the initialization. For instance, for proximal splitting algorithms, it turns out that \( \gamma \) is the composition of one or several non-expansive operators, hence 1-Lipschitz, operators. In turn, \( \gamma \) is 1-Lipschitz. Furthermore, in all examples we consider, \( \psi \) 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 also Lipschitz. Using the chain rule for Lipschitz mappings [28, Theorem 4 and Remark, Section 4.2.2], weak differentiability 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 \( \gamma, \psi \) 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 SURE\(^A\)\(_{\text{MC}}\) and SUGAR\(^A\)\(_{\text{FD}}\(_{\text{MC}}\).
**Algorithm** Risk and gradient risk estimation of an iterative scheme

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

Sample a vector $\delta$ from $\mathcal{N}(0, \text{Id}_P)$
Choose $\varepsilon = C\sigma / P^\alpha$

**for** $y' = y$ and $y' = y + \varepsilon \delta$ **do**

  Initialize $a^{(0)} \leftarrow 0$
  Initialize $J_a^{(0)} \leftarrow 0$

  **for** $\ell$ from 0 to $\mathcal{L} - 1$ **do**
  
    $a^{(\ell+1)} \leftarrow \psi(a^{(\ell)}, y', \theta)$
    $J_a^{(\ell+1)} \leftarrow \Psi_a^{(\ell)}(J_a^{(\ell)}) + \Psi_\theta^{(\ell)}$

  **end for**

  $x^{(\ell)}(y') \leftarrow \gamma(a^{(\ell)})$
  $J_x^{(\ell)}(y') \leftarrow \Gamma_a^{(\ell)}(J_a^{(\ell)})$

**end for**

$d_{\text{FDMC}}^A \leftarrow \frac{1}{\varepsilon} \left< \Phi(x^{(L)}(y + \varepsilon \delta) - x^{(L)}(y)), A^*A\delta \right>$

$\text{SURE}_{\text{FDMC}}^A \leftarrow ||A(y - \Phi x^{(L)})||^2 - \sigma^2 \text{tr}(A^*A) + 2\sigma^2 d_{\text{FDMC}}^A$

$\text{SUGAR}_{\text{FDMC}}^A \leftarrow 2 J_x^{(L)}(y)^* \Phi^* A^* A (\Phi x^{(L)} - y) + \frac{2\sigma^2}{\varepsilon} \left( J_x^{(L)}(y + \varepsilon \delta) - J_x^{(L)}(y) \right)^* \Phi^* A^* A \delta$

**return** $x(y, \theta) \leftarrow x^{(L)}(y)$, $\hat{R}^A\{x\}(y, \theta) \leftarrow \text{SURE}_{\text{FDMC}}^A$ and $\hat{\nabla}_2 R^A\{x\}(y, \theta) \leftarrow \text{SUGAR}_{\text{FDMC}}^A$

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 $\hat{R}^A$ using approximated Monte Carlo simulation, i.e., as described in [51]. 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.

**4.2 Computation of SURE$_{\text{MC}}^A$ 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 [70] (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 $\delta$ (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 $D_x^{(\ell)} = \partial_1 x^{(\ell)}(y, \theta)[\delta]$. 
is given by

\[ D_x^{(ℓ)} = Γ^{(ℓ)}(D_a^{(ℓ)}) \]

with

\[ D_a^{(ℓ+1)} = Ψ^{(ℓ)}(D_a^{(ℓ)}) + Ψ_y^{(ℓ)}(δ), \]

where \( D_a^{(ℓ)} = \partial_1 a^{(ℓ)}(y, \theta)[δ] \) and we have defined the following linear mappings

\[ Γ^{(ℓ)}(·) = \partial_1 γ(a^{(ℓ)})[·], \]
\[ Ψ^{(ℓ)}(·) = \partial_1 ψ(a^{(ℓ)}, y, \theta)[·], \]
\[ Ψ_y^{(ℓ)}(·) = \partial_2 ψ(a^{(ℓ)}, y, \theta)[·]. \]

Plugging \( \partial_2 x^{(ℓ)}(y, \theta)[δ] \) in (8), and in turn in (6), gives iteratively an unbiased \(^5\) estimate of the risk at the current iterate \( x^{(ℓ)}(y, \theta) \). The whole procedure is summarized in Fig. 2. It is worth point out that although estimating the risk entails additional operations, the global complexity is the same as for the original iterative algorithm without risk estimation.

4.3 Computation of SUGAR\(^A\)\(_{FDMC}\) for Risk Optimization

We now focus on the computation of the weak Jacobian \( \partial_2 x^{(ℓ)}(y, \theta) \). Unlike for risk estimation that required only weak directional derivatives, for risk optimization we need the full weak Jacobian matrix \( \partial_2 x(y, \theta) \in \mathbb{R}^{\text{dim}(Θ) \times N} \). 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 [34, 47] for a comprehensive account of these approaches.

In our case, we consider that, unlike for \( \partial_1 x^{(ℓ)}(y, \theta) \), the matrix \( \partial_2 x(y, \theta) \) is in practice quite small since \( \text{dim}(Θ) \ll P \), 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 \( \partial_2 x(y, \theta) \).

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

**Proposition 4.** The weak Jacobian \( J_x^{(ℓ)} = \partial_2 x^{(ℓ)}(y, \theta) \) is given by

\[ J_x^{(ℓ)} = Γ_a^{(ℓ)}(J_a^{(ℓ)}) \]

with

\[ J_a^{(ℓ+1)} = Ψ^{(ℓ)}(J_a^{(ℓ)}) + Ψ_θ^{(ℓ)}, \]

where \( J_a^{(ℓ)} = \partial_2 a^{(ℓ)}(y, \theta) \) and we have defined

\[ Γ_a^{(ℓ)}(·) = \partial_1 γ(a^{(ℓ)})[·], \]
\[ Ψ_a^{(ℓ)}(·) = \partial_1 ψ(a^{(ℓ)}, y, \theta)[·], \]
\[ Ψ_θ^{(ℓ)} = \partial_2 ψ(a^{(ℓ)}, y, \theta). \]

\(^5\)Expectation is to be taken here with respect to both the Gaussian measure of the noise \( W \) and the direction \( Δ \).
Plugging $\partial_2 x^{(\ell)}(y, \theta)$ in the expression of SUGAR$^A_{\text{FDMC}}$ 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 [51]. 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 $\theta$, 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 [50] allows one to find one element belonging to the set $x(y, \theta)$ solution of the structured convex optimization problem

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

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 $\ell_1$ norm is simple since its proximal operator is explicitly the soft-thresholding). Recall that the proximal mapping of a proper closed convex function $G$ is defined as

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

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

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 $a^{(\ell+1)} = \psi(a^{(\ell)}, y, \theta)$ chosen such that for all $k = 1, \ldots, Q$,

$$(x^{(\ell+1)} = \frac{1}{Q} \sum_{k=1}^{Q} z_k^{(\ell+1)} $$

and

$$ z_k^{(\ell+1)} = z_k^{(\ell)} - x^{(\ell)} + \operatorname{Prox}_{\nu Q G_k}(z_k^{(\ell)}, y, \theta) $$

with

$$ z_k^{(\ell)} = 2x^{(\ell)} - z_k^{(\ell)} - \nu \nabla_1 F(x^{(\ell)}, y, \theta). $$

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 [13] when $Q = 1$ and the Douglas-Rachford splitting [11] when $F = 0$.

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

$$ D_x^{(\ell+1)} = \frac{1}{Q} \sum_{k=1}^{Q} D_z_k^{(\ell+1)} $$

and

$$ D_z_k^{(\ell+1)} = D_z_k^{(\ell)} - D_x^{(\ell)} + G_{k,x}(D_z_k^{(\ell)}) + G_{k,y}(\delta) $$

with

$$ D_z_k^{(\ell)} = 2D_z^{(\ell)} - D_z_k^{(\ell)} - \nu(F_2^{(\ell)}(D_z^{(\ell)}) + F_y^{(\ell)}(\delta)). $$
where we have defined the following linear mappings
\[
G_{k,x}^{(\ell)}(\cdot) = \partial_1\{\text{Prox}_{\omega G_{k}}\}(Z_{k}^{(\ell)}, y, \theta)[], \\
G_{k,y}^{(\ell)}(\cdot) = \partial_2\{\text{Prox}_{\omega G_{k}}\}(Z_{k}^{(\ell)}, y, \theta)[], \\
F_{x}^{(\ell)}(\cdot) = \partial_1\{\nabla F\}(x^{(\ell)}, y, \theta)[], \\
\text{and } F_{y}^{(\ell)}(\cdot) = \partial_2\{\nabla F\}(x^{(\ell)}, y, \theta)[].
\]

\[\textbf{Corollary 2.}\] The GFB weak Jacobian \(J_{a}^{(\ell)} = \Gamma_{a}^{(\ell)}(J_{a}^{(\ell)})\), where \(J_{a}^{(\ell+1)} = \Psi_{a}^{(\ell)}(J_{a}^{(\ell)}) + \Psi_{\theta}^{(\ell)}\), is computed by evaluating iteratively
\[
J_{a}^{(\ell+1)} = \frac{1}{\alpha} \sum_{k=1}^{Q} J_{z_{k}}^{(\ell+1)}
\]
and \(J_{a}^{(\ell+1)} = J_{a}^{(\ell)} - J_{a}^{(\ell)} + \psi_{k,\theta}(J_{a}^{(\ell)} - G_{k,\theta}) + G_{k,x}^{(\ell)}
\]
with \(J_{z_{k}}^{(\ell)} = 2J_{z_{k}}^{(\ell)} - \nu F_{x}^{(\ell)}(J_{x}^{(\ell)}) + F_{y}^{(\ell)}\),
\]
where we have defined
\[
G_{k,x}^{(\ell)}(\cdot) = \partial_1\{\text{Prox}_{\omega G_{k}}\}(Z_{k}^{(\ell)}, y, \theta)[], \\
G_{k,y}^{(\ell)}(\cdot) = \partial_2\{\text{Prox}_{\omega G_{k}}\}(Z_{k}^{(\ell)}, y, \theta)[], \\
F_{x}^{(\ell)}(\cdot) = \partial_1\{\nabla F\}(x^{(\ell)}, y, \theta)[], \\
\text{and } F_{y}^{(\ell)}(\cdot) = \partial_2\{\nabla F\}(x^{(\ell)}, y, \theta).[\]

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
\[
x(y, \theta) = \text{Argmin}_{x \in X} \{E(x, y, \theta) = H(x, y, \theta) + G(K(x), y, \theta)\}, \tag{18}
\]
where both \(x \mapsto H(x, y, \theta)\) and \(u \mapsto G(u, y, \theta)\) are proper closed convex and simple functions, and \(K : X \rightarrow U\) is a bounded linear operator.

The primal-dual* relaxed Arrow-Hurwicz algorithm as revitalized recently in [9] (that we coin CP) to solve (18) implements (16) with \(a^{(\ell)} = (\xi^{(\ell)}, \tilde{x}^{(\ell)}, u^{(\ell)}) \in A = X^{2} \times U\), \(x^{(\ell)} = \gamma(a^{(\ell)}) = \xi^{(\ell)}\) and \(a^{(\ell+1)} = \psi(a^{(\ell)}, y, \theta)\) such that
\[
u^{(\ell+1)} = \text{Prox}_{\tau G^{*}}(U^{(\ell)}, y, \theta) \text{ where } U^{(\ell)} = u^{(\ell)} + \tau K(\tilde{x}^{(\ell)}),
\]
x^{(\ell+1)} = \text{Prox}_{\xi H}(X^{(\ell)}, y, \theta) \text{ where } X^{(\ell)} = x^{(\ell)} - \xi K^{*}(u^{(\ell+1)}), \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
\[
\text{Prox}_{\tau G^{*}}(u, y) = u - \tau \text{ Prox}_{G^{*}/\tau}(u/y, \tau).
\]
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 [9].

*We invite the interested reader to consult [41] for a detailed review on primal-dual algorithms.
Corollary 3. For any vector $\delta \in \mathcal{X}$, the CP weak directional derivatives $D_x^{(\ell)} = \Gamma_x^{(\ell)}(D_u^{(\ell)})$ and $D_{x^{(\ell+1)}} = \Psi^{(\ell)}_x(D_u^{(\ell)}) + \Psi^{(\ell)}_x(\delta)$ are computed by evaluating iteratively

$$D_u^{(\ell+1)} = G_u^{(\ell)}(D_U^{(\ell)}) + G_y^{(\ell)}(\delta) \quad \text{where} \quad D_U^{(\ell)} = D_u^{(\ell)} + \tau K(D_{\tilde{x}}^{(\ell)}),$$

$$D_{x^{(\ell+1)}} = \mathcal{H}_x^{(\ell)}(D_X^{(\ell)}) + \mathcal{H}_y^{(\ell)}(\delta) \quad \text{where} \quad D_X^{(\ell)} = D_x^{(\ell)} - \xi K^*(D_{x^{(\ell+1)}}),$$

and

$$D_{\tilde{x}}^{(\ell+1)} = D_x^{(\ell+1)} + \zeta(D_x^{(\ell+1)} - D_{\tilde{x}}^{(\ell)}),$$

where we have defined the following linear mappings

$$\mathcal{H}_x^{(\ell)}(\cdot) = \partial_1\{\text{Prox}_{\xi H}\}(X^{(\ell)}, y, \theta)[\cdot],$$
$$\mathcal{H}_y^{(\ell)}(\cdot) = \partial_2\{\text{Prox}_{\xi H}\}(X^{(\ell)}, y, \theta)[\cdot],$$
$$G_u^{(\ell)}(\cdot) = \partial_1\{\text{Prox}_{\tau G^*}\}(U^{(\ell)}, y, \theta)[\cdot],$$
$$G_y^{(\ell)}(\cdot) = \partial_2\{\text{Prox}_{\tau G^*}\}(U^{(\ell)}, y, \theta)[\cdot].$$

Corollary 4. Similarly to Corollary 3, the CP weak Jacobians $J_x^{(\ell)} = \Gamma_x^{(\ell)}(J_u^{(\ell)})$ and $J_{x^{(\ell+1)}} = \Psi^{(\ell)}_x(J_u^{(\ell)}) + \Psi^{(\ell)}_x(\xi)$ are computed by evaluating iteratively

$$J_u^{(\ell+1)} = G_u^{(\ell)}(J_U^{(\ell)}) + G_y^{(\ell)}(\xi) \quad \text{where} \quad J_U^{(\ell)} = J_u^{(\ell)} + \tau K(J_{\tilde{x}}^{(\ell)}),$$

$$J_x^{(\ell+1)} = \mathcal{H}_x^{(\ell)}(J_X^{(\ell)}) + \mathcal{H}_y^{(\ell)}(\xi) \quad \text{where} \quad J_X^{(\ell)} = J_x^{(\ell)} - \xi K^*(J_{x^{(\ell+1)}}),$$

and

$$J_{\tilde{x}}^{(\ell+1)} = J_x^{(\ell+1)} + \zeta(J_x^{(\ell+1)} - J_{\tilde{x}}^{(\ell)}),$$

where we have defined

$$\mathcal{H}_x^{(\ell)}(\cdot) = \partial_1\{\text{Prox}_{\xi H}\}(X^{(\ell)}, y, \theta)[\cdot],$$
$$\mathcal{H}_y^{(\ell)}(\cdot) = \partial_2\{\text{Prox}_{\xi H}\}(X^{(\ell)}, y, \theta)[\cdot],$$
$$G_u^{(\ell)}(\cdot) = \partial_1\{\text{Prox}_{\tau G^*}\}(U^{(\ell)}, y, \theta)[\cdot],$$
$$G_y^{(\ell)}(\cdot) = \partial_2\{\text{Prox}_{\tau G^*}\}(U^{(\ell)}, y, \theta).$$

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 $\ell_1$-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 $L = 100$ iterations. For quasi-Newton optimization, we used the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method with the implementation of [43]. The use of BFGS is just
given as an example to make the most of the proposed gradient estimator of the risk and seems to be good enough for the cases considered in all the following experiments examples. Of course, other first-order optimization methods can be considered for just as well, essentially if the risk presents several local minima.

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

\[
\min_{x} \frac{1}{2} \| \Phi x - y \|^2 + \sum_{k=1}^{K} \lambda^k R^k(x),
\]

where \( \lambda^k > 0, \forall k \in \mathbb{N} \), the initialization \( \lambda_0^k \) is chosen empirically as

\[
\lambda_0^k = \frac{P \sigma^2}{4 \sum_{k=1}^{K} R^k(x_{LS}(y))},
\]

where \( x_{LS}(y) \) is the least-square estimator. At the first iteration, the approximate inverse Hessian \( B_1 \) should be chosen such that, for all \( k > 0 \), \( \lambda_1^k \) is of the same order as \( \lambda_0^k \). To this end, we suggest initializing \( B_1 \) as a diagonal matrix with diagonal entries

\[
B_1^k = \frac{\alpha \lambda_0^k}{SUGAR_{FDMC}\{x\}(y, \lambda_0, \delta, \varepsilon)_k},
\]

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

\[
\frac{\| SUGAR_{FDMC}\{x\}(y, \lambda, \delta, \varepsilon) \|_\infty}{\| SUGAR_{FDMC}\{x\}(y, \lambda_0, \delta, \varepsilon) \|_\infty} \leq \tau
\]

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 \( Y = \Phi x_0 + W, W \sim \mathcal{N}(0, \sigma^2 \text{Id}_P) \), 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

\[
x^*(y, \lambda) \in \min_{x} \frac{1}{2} \| \Phi x - y \|^2 + \lambda \| x \|_s,
\]

where \( \lambda > 0 \) and \( \| \cdot \|_s \) is the nuclear norm (a.k.a., trace for the symmetric semi-definite positive case or Schatten 1-norm). This is a spectral function defined as the \( \ell_1 \) norm of the singular values \( \Lambda_x \in \mathbb{R}^{n = \min(n_1, n_2)} \), i.e.

\[
\| x \|_s = \| \Lambda_x \|_1.
\]

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., [29]). It is the convex hull of the
rank function restricted to the unit spectral ball \([7]\). The parameter \(\lambda\) 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 \(\lambda\) and the rank of \(x(y, \lambda)\). The optimal value of \(\lambda\) depends indeed on \(x_0\), \(\Phi\) and \(\sigma\) 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

\[
F(x, y, \lambda) = \frac{1}{2} \| \Phi x - y \|^2
\]

and \(G_1(x, y, \lambda) = \lambda \|x\|_\ast\).

Hence the GFB algorithm\(\footnote{which corresponds in this case where \(Q = 1\) to the forward-backward algorithm.}\) can be used to solve (24) by setting

\[
\nabla_1 F(x, y, \lambda) = \Phi^* (\Phi x - y)
\]

and \(\text{Prox}_{\tau G_1}(x, y, \lambda) = V_x \text{diag}(\text{ST}(\Lambda_x, \tau \lambda))U_x^*\)

where \(\text{diag} : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n_2}\) 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 \(x = V_x \text{diag}(\Lambda_x)U_x^*\) and \(\text{ST}\) is the 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

\[
\partial_1 \{\nabla_1 F\}(x, y, \lambda)[\delta_x] = \Phi^* \Phi \delta_x,
\]

\[
\partial_2 \{\nabla_1 F\}(x, y, \lambda)[\delta_y] = -\Phi^* \delta_y,
\]

\[
\partial_3 \{\nabla_1 F\}(x, y, \lambda) = 0
\]

and

\[
\partial_1 \{\text{Prox}_{\tau G_1}\}(x, y, \lambda)[\delta_x] = V_x (\mathcal{H}(\Lambda_x)[\delta_x] + \Gamma_S(\Lambda_x)[\delta_x] + \Gamma_A(\Lambda_x)[\delta_x])U_x^*,
\]

\[
\partial_2 \{\text{Prox}_{\tau G_1}\}(x, y, \lambda)[\delta_y] = 0,
\]

\[
\partial_3 \{\text{Prox}_{\tau G_1}\}(x, y, \lambda) = V_x \text{diag}(\partial_2 \text{ST}(\Lambda_x, \tau \lambda))U_x^*
\]

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

\[
\mathcal{H}(\Lambda_x)[\delta_x] = \text{diag}(\partial_1 \text{ST}(\Lambda_x, \rho \lambda)[\text{diag}(\tilde{\delta}_x)])
\]

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

\[
\Gamma_S(\Lambda_x)[\tilde{\delta}_x]_{i,j} = \frac{(\delta_x)_{i,j} + (\delta_x)_{j,i}}{2} \times \left\{ \begin{array}{ll} 0 & \text{if } i = j \\ \frac{\text{ST}(\Lambda_x, \rho \lambda)_{i,j} - \text{ST}(\Lambda_x, \rho \lambda)_{j,i}}{\text{ST}(\Lambda_x, \rho \lambda)_{i,j}} & \text{if } (\Lambda_x)_{i,j} \neq 0 \end{array} \right.
\]

otherwise,

\[
\Gamma_A(\Lambda_x)[\tilde{\delta}_x]_{i,j} = \frac{(\delta_x)_{i,j} - (\delta_x)_{j,i}}{2} \times \left\{ \begin{array}{ll} 0 & \text{if } i = j \\ \frac{\text{ST}(\Lambda_x, \rho \lambda) + \text{ST}(\Lambda_x, \rho \lambda)}{\text{ST}(\Lambda_x, \rho \lambda)_{i,j} + \text{ST}(\Lambda_x, \rho \lambda)_{j,i}} & \text{if } (\Lambda_x)_{i,j} > 0 \text{ or } (\Lambda_x)_{j,i} > 0 \\ \partial_1 \text{ST}(\Lambda_x, \rho \lambda)_{i,j} & \text{otherwise}, \end{array} \right.
\]

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

\[
\partial_1 \text{ST}(t, \rho)_{i,i} = \begin{cases} 0 & \text{if } |t_i| \leq \rho \\ 1 & \text{otherwise}, \end{cases}
\]

\[
\partial_1 \text{ST}(t, \rho)[\delta_t]_{i,i} = \partial_1 \text{ST}(t, \rho)_{i,i} \times (\delta_t)_{i,i}
\]

and

\[
\partial_2 \text{ST}(t, \rho)_{i,i} = \begin{cases} 0 & \text{if } |t_i| \leq \rho \\ -\text{sign}(t_i) & \text{otherwise}. \end{cases}
\]

\(\text{which corresponds in this case where } Q = 1\) to the forward-backward algorithm.
The closed-form expression we derived for \( \partial_t \{ \text{Prox}_{\mathcal{G}_1} \}(x, y, \lambda)[\delta_x] \) is far from trivial. It is essentially due to [25, 42, 63], see [8] for an expression similar to ours. The generalization of this result to other matrix-valued spectral function has been studied in [19].

**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 [4]. We therefore consider \( y \in \mathbb{R}^P \) with the forward model \( Y = \Phi x_0 + W \), \( W \sim \mathcal{N}(0, \sigma^2 \text{Id}_P) \), where \( x_0 \) is a dense but low-rank (or approximately so) matrix and \( \Phi \) 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 \( x_0 = V_{x_0} \text{diag} \Lambda_{x_0} U_{x_0}^* \) has been chosen with \( V_{x_0} \) and \( U_{x_0} \) two realizations of the uniform distribution of orthogonal matrices and \( \Lambda_{x_0} = (k^{-1})_{1 \leq k \leq n} \) such that \( x_0 \) is approximately low-rank with a rapidly decaying spectrum. The binary masking operator is such that for \( i = 1, \ldots, P \), \((\Phi x)_i = x_{\Sigma(i),\Sigma(i)} \) 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 deviation \( \sigma \) has been set such that the resulting minimum least-square estimate \( x_{\text{LS}}(y) = \Phi^* y \) has a relative error \( \|x_{\text{LS}}(y) - x_0\|_F / \|x_0\|_F = 0.9 \).

Figure 4.(a) and (b) depict the risk, the prediction risk and the SURE = SURE\(^A\) (with \( A = \text{Id} \)) estimates** as a function of \( \lambda \) obtained from a single realization of \( y \) and \( \delta \). In (a), SURE\(_{\text{MC}}\)\{\(x\}y, \lambda, \delta\) has been evaluated for 12 values of \( \lambda \) chosen in a suitable tested range using the algorithm given in Figure 2. Figure (b), shows the benefits of computing SURE\(_{\text{FDMC}}\)\{\(x\}y, \lambda, \delta, \varepsilon\) and SUGAR\(_{\text{FDMC}}\)\{\(x\}y, \lambda, \delta, \varepsilon\), as described in Figure 3, to realize a quasi-Newton optimization. The sequence of iterates \( \lambda_n \) is represented as well as the sequence of the slopes of SURE\(_{\text{FDMC}}\)\{\(x\}y, \lambda_n, \delta, \varepsilon\) given by SUGAR\(_{\text{FDMC}}\)\{\(x\}y, \lambda_n, \delta, \varepsilon\). The BFGS algorithm reaches to the optimal value in 5 iterations only. One can also notice that SURE\(_{\text{FDMC}}\)\{\(x\}y, \lambda, \delta, \varepsilon\) and SURE\(_{\text{MC}}\)\{\(x\}y, \lambda, \delta\) are both good, and visually equivalent, estimators of the prediction risk. At the optimum value \( \lambda^* \) 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 \( x(y, \lambda^*) \) is compared to \( x_0 \) and \( x_{\text{LS}}(y) \), and their respective spectrum in Figure 4.(d). The solution \( x(y, \lambda^*) \) 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 \( Y = \Phi x_0 + W \in \mathbb{R}^P \), \( W \sim \mathcal{N}(0, \sigma^2 \text{Id}_P) \), 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

\[
x^*(y, \lambda) \in \text{Argmin}_x \frac{1}{2} \|\Phi x - y\|^2 + \lambda \|\nabla x\|_{1,2},
\]

where \( \lambda > 0 \) and \( \nabla : \mathbb{R}^N \to \mathbb{R}^{N \times 2} \) is the two-dimensional discrete gradient operator. The \( \ell^1-\ell^2 \) 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 \( \lambda \) controls the regularity of the

**Without impacting the optimal choice of \( \lambda \), the curves have been rescaled for visualization purposes.**
Figure 4: (a-b) Risk, prediction risk and its SURE estimates as a function of the regularization parameter $\lambda$. (a) The 12 points where SURE$_{MC}\{x\}(y, \lambda, \delta)$ has been evaluated by exhaustive search. (b) The 5 evaluation points of SURE$_{FDMC}\{x\}(y, \lambda, \delta, \varepsilon)$ and SUGAR$_{FDMC}\{x\}(y, \lambda, \delta, \varepsilon)$ required by BFGS to reach the optimal one. (c-d) Respectively, a close in and the spectrum of the underlying matrix $x_0$, the least-square estimate $x_{LS}(y)$ and the solution $x(y, \lambda)$ at the optimal $\lambda$.

A large value of $\lambda$ 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 $\lambda$ 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 $x = (f, u)$

$$F(x, y, \lambda) = \frac{1}{2} \| \Phi f - y \|^2,$$
$$G_1(x, y, \lambda) = \lambda \| u \|_{1,2},$$

and $$G_2(x, y, \lambda) = \iota_C(x) \quad \text{where} \quad C = \left\{ x = (f, u) \setminus u = \nabla f \right\}.$$

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

$$\nabla_1 F(x, y, \lambda) = (\Phi^* (\Phi f - y), 0),$$
$$\text{Prox}_{\tau G_1}(x, y, \lambda) = (f, S_{T_{1,2}}(u, \tau \lambda))$$

and $$\text{Prox}_{\tau G_2}(x, y, \lambda) = ((\text{Id} + \Delta)^{-1}(f + \text{div} u), \nabla (\text{Id} + \Delta)^{-1}(f + \text{div} u)).$$
where \(\Delta\) is the Laplacian operator and div is the discrete divergence operator such that 
\[
\text{div} = -\nabla^*.
\]
The operator \(ST_{1,2}\) is the component-wise \(\ell^1-\ell^2\) soft-thresholding defined, for any 
dimensions \(N\) and \(D\), \(t \in \mathbb{R}^{N \times D}\) and \(\rho > 0\), by
\[
ST_{1,2}(t,\rho)_i = \begin{cases} 0 & \text{if } \|t_i\| \leq \rho \\ t_i - \rho \frac{t_i}{\|t_i\|} & \text{otherwise} \end{cases} \quad \text{for all } 1 \leq i \leq N. \tag{27}
\]

For \(D = 1\), the component-wise \(\ell^1-\ell^2\) 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

\[
\partial_1 \{\nabla F\}(x,y,\lambda)[\delta_x] = (\Phi^* \Phi \delta_f, 0),
\]
\[
\partial_2 \{\nabla F\}(x,y,\lambda)[\delta_y] = (-\Phi^* \delta_y, 0),
\]
\[
\partial_1 \{\text{Prox}_{\tau_G1}\}(x,y,\lambda)[\delta_x] = (\delta_f, \partial_1 ST_{1,2}(u,\tau \lambda)[\delta_u]),
\]
\[
\partial_2 \{\text{Prox}_{\tau_G1}\}(x,y,\lambda)[\delta_y] = (0,0),
\]
\[
\partial_3 \{\text{Prox}_{\tau_G1}\}(x,y,\lambda) = (0,0),
\]
and
\[
\partial_1 \{\text{Prox}_{\tau_G2}\}(x,y,\lambda)[\delta_x] = ((\text{Id} + \Delta)^{-1}(\delta_f + \text{div} \delta_u), \nabla (\text{Id} + \Delta)^{-1}(\delta_f + \text{div} \delta_u)),
\]
\[
\partial_2 \{\text{Prox}_{\tau_G2}\}(x,y,\lambda)[\delta_y] = (0,0),
\]
\[
\partial_3 \{\text{Prox}_{\tau_G2}\}(x,y,\lambda) = (0,0)
\]

where the weak derivatives of the component-wise \(\ell^1-\ell^2\) 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
\[
\partial_1 ST_{1,2}(t,\rho)[\delta_t]_i = \begin{cases} 0 & \text{if } \|t_i\| \leq \rho \\ \delta_{t,i} - \rho \frac{\delta_{t,i}}{\|t_i\|} P_t(\delta_{t,i}) & \text{otherwise} \end{cases} \tag{28}
\]
and

\[
\partial_2 ST_{1,2}(t,\rho)_i = \begin{cases} 0 & \text{if } \|t_i\| \leq \rho \\ -t_i/\|t_i\| & \text{otherwise} \end{cases}
\]

where \(P_\alpha\) 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 \(Y = \Phi x_0 + W \in \mathbb{R}^P\),
\(W \sim \mathcal{N}(0,\sigma^2 \text{Id}_P)\), where \(x_0\) is a piece-wise constant (or approximately so) image and \(\Phi\) 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

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

Figure 5.(a) and (b) display the risk, the projection risk and the GSURE = SURE\(^A\) (with \(A = \Pi\)) estimates as a function of \(\lambda\) obtained from a single realization of \(y\) and \(\delta\). In
(a), GSURE\(\text{MC}\{x\}(y,\lambda,\delta)\) has been evaluated for 12 values of \(\lambda\) chosen in a suitable tested
range using the algorithm given in Figure 2. Figure (b), shows the benefit of computing
GSURE\(\text{FDMC}\{x\}(y,\lambda,\delta,\varepsilon)\) and GSUGAR\(\text{FDMC}\{x\}(y,\lambda,\delta,\varepsilon)\), as described in Figure 3, to re-
alize a quasi-Newton optimization. The sequence of iterates \(\lambda_n\) is represented as well as the
sequence of the slopes of $\text{SURE}_{\text{FDMC}}(x)(y, \lambda_n, \delta, \varepsilon)$ given by $\text{SURE}_{\text{GDUARFDMC}}(x)(y, \lambda_n, \delta, \varepsilon)$. The BFGS algorithm reaches to the optimal value in 4 iterations. The deviation of $\text{SURE}_{\text{FDMC}}(x)(y, \lambda_n, \delta, \varepsilon)$ from the projection risk is of the same order as the deviation of $\text{SURE}_{\text{MC}}(x)(y, \lambda, \delta)$. At the optimum value $\lambda^*$ minimizing the GSURE, the true risk is not too far from its minimum showing that relatively to the range of variation of the risk, in this case, the projection risk is indeed a good objective in order to minimize the risk. In Figure 5.(c-e) the solution $x(y, \lambda^*)$ is compared to $x_0$ and $y$. The solution $x(y, \lambda^*)$ has a PSNR of 24.98 dB (i.e., a gain of about +3.94 dB). Remark that given such noise level and convolution operator, masking 80% of (high) frequencies, the solution selected by minimizing the GSURE criterion is still a bit blurred and exhibit a staircasing effect. Using a larger regularization parameter $\lambda$ would result in a more “cartoon” result with less blur. This would however entail a larger bias, which corresponds to a loss of contrast inherent to the convexity of the TV prior. This larger bias subsequently degrades the mean square error, which explains why it is not selected by the GSURE criterion.
Table 1: Illustration of the minimization of SURE_{FDNC} 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 $\lambda^*$ are given. The associated value of SURE and the PSNR are compared to neighbors of $\lambda^*$ located at $0.75\lambda^*$ and $1.25\lambda^*$.

<table>
<thead>
<tr>
<th>Input</th>
<th>Optimal parameters</th>
<th>SURE/PSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$J$</td>
<td>$\dim \Lambda$</td>
</tr>
<tr>
<td>Mandrill</td>
<td>1</td>
<td>1 (7.58)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1 (5.63)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1 (4.54)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2 (5.94, 4.24)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1 (7.51, 1.07, 0.99)</td>
</tr>
<tr>
<td>House</td>
<td>1</td>
<td>1 (18.38)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1 (11.11)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1 (8.73)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2 (14.47, 5.20)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3 (15.00, 2.50, 2.83)</td>
</tr>
<tr>
<td>Cameraman</td>
<td>1</td>
<td>1 (13.50)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1 (8.78)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1 (7.14)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2 (10.98, 3.74)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3 (11.56, 3.31, 0.97)</td>
</tr>
</tbody>
</table>

5.4 Weighted $\ell_1$-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 $Y = \Phi x_0 + W \in \mathbb{R}^P$, $W \sim \mathcal{N}(0, \sigma^2 \text{Id}_P)$, using a $J$-scale undecimated wavelet analysis regularization of the form

$$x^*(y, \lambda) \in \text{Argmin}_x \frac{1}{2} \| \Phi x - y \|^2 + \| \Psi x \|_{1, \lambda}$$

where

$$\Psi = \begin{pmatrix} \Psi^h_1 \\ \Psi^v_1 \\ \vdots \\ \Psi^h_J \\ \Psi^v_J \end{pmatrix}$$

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

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

Multi-scale wavelet analysis promotes piece-wise regular images by enforcing smoothness while preserving sharp discontinuities at different scales and orientations. Each parameter $\lambda_j$ controls the regularity at scale $j$. A large value of $\lambda_j$ tends to over-smooth structures at scale $j$, while a small value leads to under-smoothing. As noted in several papers, see e.g. [10, 49], the optimal values $\lambda_j$ 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 \( \lambda = \theta \in \Theta = \mathbb{R}^+ \), and

\[
H(x, y, \lambda) = \frac{1}{2} \| \Phi x - y \|^2,
\]

\[
G(u, y, \lambda) = \| u \|_{1, \lambda}^2
\]

and \( K(x) = \Psi x \).

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

\[
\text{Prox}_{\xi H}(x, y, \lambda) = x + \xi \Phi^* y - \xi \Phi^*(\text{Id} + \xi \Phi \Phi^*)^{-1}\Phi(x + \xi \Phi^* y),
\]

\[
\text{Prox}_{\tau G^*}(u, y, \lambda) = u - \tau ST(u/\tau, \lambda/\tau)
\]

and \( K^*(u, \lambda) = \sum_{j=1}^{J} (\Psi h^*_j u_j^h + \Psi v^*_j u_j^v) \)

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

\[
ST(t, \rho)^o_j = ST(t^o_j, \rho_j)
\]

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

\[
\partial_1 \{ \text{Prox}_{\xi H} \}(x, y, \lambda)[\delta_x] = \delta_x + \xi \Phi^*(\text{Id} + \xi \Phi \Phi^*)^{-1}\Phi \delta_x,
\]

\[
\partial_2 \{ \text{Prox}_{\xi H} \}(x, y, \lambda)[\delta_y] = \xi \Phi^* \delta_y - \xi^2 \Phi^*(\text{Id} + \xi \Phi \Phi^*)^{-1}\Phi \Phi^* \delta_y,
\]

\[
\partial_3 \{ \text{Prox}_{\xi H} \}(x, y, \lambda) = 0
\]

and \( \partial_1 \{ \text{Prox}_{\tau G^*} \}(u, y, \lambda)[\delta_u] = \delta_u - \partial_1 ST(u/\tau, \lambda/\tau)[\delta_u], \)

\[
\partial_2 \{ \text{Prox}_{\tau G^*} \}(u, y, \lambda)[\delta_y] = 0,
\]

\[
\partial_3 \{ \text{Prox}_{\tau G^*} \}(u, y, \lambda) = -\partial_2 ST(u/\tau, \lambda/\tau).
\]

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

\[
\partial_1 \text{ST}(t, \rho)[\delta_t]^o_j = \partial_1 \text{ST}(t^o_j, \rho_j)[\delta_t]^o_j \quad \text{and} \quad \partial_2 \text{ST}(t, \rho)^o_j = \partial_2 \text{ST}(t^o_j, \rho_j)
\]

for all scales \( 1 \leq j \leq J \) and orientations \( o = v, h \).

**Application to compressed sensing** We illustrate the multi-scale wavelet \( \ell_1 \)-analysis regularization on a compressed sensing problem. We therefore consider the forward model \( Y = \Phi x_0 + W \in \mathbb{R}^P \), \( W \sim \mathcal{N}(0, \sigma^2 \text{Id}_P) \), where \( x_0 \) is a piece-wise multi-scale regular (or approximately so) image and \( \Phi \) is a random matrix. Here the multi-scale transform \( W \) is constructed from undecimated Daubechies 4 wavelets [14].

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 \( x_{LS}(y) = \Phi^+ y \) has a PSNR given by \( 10 \log_{10}(255^2 / \| x_{LS}(y) - x_0 \|_F^2) \approx 16 \text{ dB} \).

Table 1 and Figure 6 illustrates the multi-scale regularization obtained by minimizing the SURE \( A^4 \) (with \( A = \text{Id} \)) for three different images \( x_0 \), known as Mandrill, House and Cameraman, and a single realization of \( y \) and \( \delta. \) 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 \( \lambda^* \) 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
(a) Underlying image $x_0$. (b) Least-square estimate $x_{LS}(y)$. (c) Result with $\lambda^*$ 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.

Figure 6: From top to bottom, a close up on Mandrill, House and Cameraman. (a) Underlying image $x_0$. (b) Least-square estimate $x_{LS}(y)$. (c) Result with $\lambda^*$ 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.

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 $\text{SURE}_{\text{FDMC}}$, we have compared the SURE and the PSNR values at $0.75\lambda^*$ and $1.25\lambda^*$. At the optimal $\lambda^*$, the SURE is as expected minimal. Furthermore, at $\lambda^*$, 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 asymptotically 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 [34]. 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, \( \hat{d}_F^A\{\mu\}(y, \theta, \varepsilon) \) is just the sum of \( P \) weakly differentiable functions, and hence is weakly differentiable with the weak derivative with respect to \( \theta \) as given. Moreover, \( \|A(\mu(y, \theta) - y)\|^2 = \sum_{i=1}^{P} ((A(\mu(y, \theta) - y))_i)^2 \). 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 [28, 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 \( \theta \) 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 $\Theta$, and choose $\varphi \in C^1_c(\Theta)$ with support in $V$. We have

$$
\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 
\quad \text{[Stein Lemma] (S.1)} = \int_{V} \mathbb{E}_W \left[ \text{SURE}^A\{\mu\}(Y,\theta) \right] \frac{\partial \varphi(\theta)}{\partial \theta_i} d\theta 
\quad \text{[12]} = \int_{V} \mathbb{E}_W \left[ \lim_{\varepsilon \to 0} \text{SURE}_{FD}^A\{\mu\}(Y,\theta,\varepsilon) \right] \frac{\partial \varphi(\theta)}{\partial \theta_i} d\theta 
\quad \text{[Dominated convergence] (S.2)} = \lim_{\varepsilon \to 0} \int_{V} \mathbb{E}_W \left[ \text{SURE}_{FD}^A\{\mu\}(Y,\theta,\varepsilon) \right] \frac{\partial \varphi(\theta)}{\partial \theta_i} d\theta 
\quad \text{[Fubini] (S.3)} = \lim_{\varepsilon \to 0} \mathbb{E}_W \left[ \int_{V} \text{SURE}_{FD}^A\{\mu\}(Y,\theta,\varepsilon) \frac{\partial \varphi(\theta)}{\partial \theta_i} d\theta \right] 
\quad \text{[Weak differentiability, Proposition 1] (S.5)} = -\lim_{\varepsilon \to 0} \mathbb{E}_W \left[ \int_{V} \frac{\partial}{\partial \theta_i} \text{SURE}_{FD}^A\{\mu\}(Y,\theta,\varepsilon) \varphi(\theta) d\theta \right] 
\quad \text{[Proposition 1] (S.6)} = -\lim_{\varepsilon \to 0} \mathbb{E}_W \left[ \int_{V} \left( \text{SUGAR}_{FD}^A\{\mu\}(Y,\theta,\varepsilon) \right)_i \varphi(\theta) d\theta \right] 
\quad \text{[Fubini] (S.7)} = -\lim_{\varepsilon \to 0} \int_{V} \mathbb{E}_W \left[ \left( \text{SUGAR}_{FD}^A\{\mu\}(Y,\theta,\varepsilon) \right)_i \varphi(\theta) d\theta \right] 
\quad \text{[Dominated convergence] (S.8)} = -\int_{V} \left( \lim_{\varepsilon \to 0} \mathbb{E}_W \left[ \left( \text{SUGAR}_{FD}^A\{\mu\}(Y,\theta,\varepsilon) \right)_i \right] \right) \varphi(\theta) d\theta 
= -\int_{\Theta} \left( \lim_{\varepsilon \to 0} \mathbb{E}_W \left[ \left( \text{SUGAR}_{FD}^A\{\mu\}(Y,\theta,\varepsilon) \right)_i \right] \right) \varphi(\theta) d\theta .
$$

From the definition of weak derivative, we get the claimed result on the asymptotic unbiasedness of $\text{SUGAR}_{FD}^A\{\mu\}$. 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 $\text{SURE}_{FD}^A\{\mu\}(\mu_0,\theta,\varepsilon) = \|A(\mu(y,\theta) - y)\|^2 - \sigma^2 \text{tr}(A^2A) + 2\sigma^2 \text{SURE}_{FD}^A\{\mu\}(\mu_0,\theta,\varepsilon)$.

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 $\sigma^2$, and $g_{\sigma}$ 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. [28, Theorem 1-2, Section 6.2]. Moreover, we have for any $\theta$

$$
\|\mu(y,\theta) - \mu(y',\theta)\| \leq L_1 \|y - y'\| \Rightarrow |\mu_i(y,\theta) - \mu_i(y',\theta)| \leq L_1 \|y - y'\| , \quad (31)
$$

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

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

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

\[
\| \mu(y, \theta) - y \| \leq \| y \| + \| \mu(y, \theta) - \mu(0, \theta) \| \leq (1 + L_1) \| y \| .
\]

Combining this with (31) leads to

\[
|f_\varepsilon(y, \theta)| = \left| A(\mu(y, \theta) - y) \right|^2 - \sigma^2 \text{tr}(A^*A) + 2\sigma^2 \frac{1}{\varepsilon} \sum_{i=1}^{P} (A^*A(\mu(y + \varepsilon e_i, \theta) - \mu(y, \theta)))_i \leq \| A \|^2 P^2 \left( 1 + \frac{L_1^2 \| y \|^2}{P^2} + 1 + 2L_1 \right) .
\]

Note that the bound is independent of \( \theta \). Thus

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

This bound together with the fact that \( \varphi \) is continuously differentiable with compact support in \( V \) means that \( f_\varepsilon \frac{\partial \mu}{\partial \theta} \) is dominated by an integrable function on \( Y \times V \). 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 SUGAR^A_{\text{FD}} \{ \mu \} in Proposition 1.

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

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

Thus

\[
|g_\sigma * h(\cdot, \theta)(\mu_0)| \leq 2 \int_{Y} g_\sigma(y - \mu_0) |(\partial_2 \mu(y, \theta)^*A^*A(\mu(y, \theta) - y))_i| \, dy
\]

[Assumption (A.3)] \leq 2L_2 \frac{\| A \|^2}{\sigma^2} \int_{Y} g_\sigma(y - \mu_0) \| \mu(y, \theta) - y \| \, dy
\]

[Assumptions (A.1)-(A.2)] \leq 2(1 + L_1) L_2 \frac{\| A \|^2}{\sigma^2} \int_{Y} g_\sigma(y - \mu_0) \| y \| \, dy
\]

\leq 2(1 + L_1) L_2 \frac{\| A \|^2}{\sigma^2} \mathbb{E}_{W} \| y \| \, dy
\]

[Jensen inequality] \leq 2(1 + L_1) L_2 \frac{\| A \|^2}{\sigma^2} \mathbb{E}_{W} \| y \|^2 \frac{1}{2} \, dy
\]

\leq 2(1 + L_1) L_2 \frac{\| A \|^2}{\sigma^2} \| \mu_0 \|^2 + P^2 < \infty .
\]
Let \( f \) defined as in (S.7). We have just shown that the integrand in \( \theta \), i.e. \( \mathbb{E}_W[f_\varepsilon(Y, \cdot)] \), 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 \( f_\varepsilon \) 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.

\[ \text{Proposition 1.} \] For a fixed \( \lambda \), it can be shown similarly to [28, Theorem 4(iii), Section 4.2.2], that \( \text{ST}(\cdot, \lambda) \) is weakly differentiable and that its weak Jacobian \( h(y) = \partial_2 \text{ST}(y, \lambda) \) is diagonal, with diagonal elements, for \( 1 \leq i \leq P \),

\[
h(y)_i = \begin{cases} +1 & \text{if } y_i \leq -\lambda \\ 0 & \text{if } -\lambda < y_i < \lambda \\ -1 & \text{otherwise} \end{cases}
\] (32)

We next define, for a fixed \( \lambda \), the quantity \( h'(y, \varepsilon) = \nabla_2 \{ \tilde{d}_F \{ \text{ST} \} \} (y, \lambda, \varepsilon) \). Using Proposition 1 and the fact that \( \varepsilon < 2\lambda \) give

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

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

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

The Cauchy-Schwartz inequality implies that

$$\mathbb{V}_W\left[\frac{1}{P}SUGAR_{FD}\{ST\}(Y, \lambda, \varepsilon)\right]^{1/2} \leq 2\mathbb{V}_W\left[\frac{1}{P}h(y)^*(ST(Y, \lambda) - Y)\right]^{1/2} + 2\sigma^2 \mathbb{V}_W\left[\frac{1}{P}h'(Y, \hat{\varepsilon}(P))\right]^{1/2}.$$

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

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

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

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

where $\#\{|y| > \lambda\}$ denotes the number of entries of $|y|$ greater than $\lambda$. We have $\#\{|Y_i| > \lambda\} \sim \text{iid Bernoulli}(p_i)$ whose variance is $p_i(1 - p_i)$, where $p_i = \frac{1}{2} \left( \text{erf} \left( \frac{(\mu_0) + \lambda}{\sqrt{2} \sigma} \right) - \text{erf} \left( \frac{(\mu_0) - \lambda}{\sqrt{2} \sigma} \right) \right)$.

It follows that $\mathbb{V}_W[\#\{|Y| > \lambda\}] = \sum_{i=1}^{P} p_i(1 - p_i) \leq P$ and hence

$$\lim_{P \to \infty} \mathbb{V}_W\left[\frac{1}{P}h'(Y, \hat{\varepsilon}(P))\right] = \lim_{P \to \infty} \mathbb{V}_W\left[\frac{1}{P}h'(Y, \hat{\varepsilon}(P))\right] = 0.$$

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 $\lim_{\varepsilon \to 0} \mathbb{E}_W \nabla_2 \{d f\{ST\}\}(Y, \lambda, \varepsilon) = \nabla_2 \{d f\{ST\}\}(\mu_0, \lambda)$ concludes the second part.

B Regularity of the proximal operator of a gauge

We first provide a glimpse of gauges.

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

$$\gamma_C(y) = \inf \{ \omega > 0 \mid y \in \omega C \}.$$

As usual, $\gamma_C(y) = +\infty$ if the infimum is not attained.

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

$$C^\circ = \{ z \in \mathbb{R}^N \mid \langle z, x \rangle \leq 1 \text{ for all } x \in C \}$$

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

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

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

(i) $\gamma_C$ is a non-negative, closed, convex and positively homogeneous function.

(ii) $C$ is the unique closed convex set containing the origin such that

$$C = \{ y \in \mathcal{Y} \mid \gamma_C(y) \leq 1 \}.$$

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

(iv) The gauge of $C$ and the support function $\sigma_{C^0}(y) = \max_{z \in C^0} \langle y, z \rangle$ coincide, i.e.

$$\gamma_C = \sigma_{C^0}.$$

**Proof.** (i)-(ii) follow from [36, Theorem V.1.2.5]. (iii) is a consequence of [36, Theorem V.1.2.5(ii) and Corollary V.1.2.6]. (iv) [36, Proposition V.3.2.4].

We are now equipped to prove our regularity result.

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

$$\| \text{Prox}_{\theta G}(y) - \text{Prox}_{\theta' G}(y) \| \leq L_2 |\theta - \theta'| \| y - \text{Prox}_{\theta G}(y) \| / \theta.$$  \hfill (33)

Now, we have

$$\theta G(y) = \gamma_{C/\theta}(y) = \sigma_{\theta C^0}(y),$$  \hfill (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

$$y - \text{Prox}_{\theta G}(y) = y - \text{Prox}_{\sigma_{\theta C^0}}(y) = \text{Proj}_{\theta C^0}(y).$$

By virtue of Lemma 2(iii), there exists a constant $L_2 > 0$, independent of $y$, such that\footnote{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.}

$$\| y - \text{Prox}_{\theta G}(y) \| = \| \text{Proj}_{\theta C^0}(y) \| \leq L_2 \gamma_{C^0}(\text{Proj}_{\theta C^0}(y)).$$

Applying (34) to $\gamma_{C^0}$, we get

$$\| y - \text{Prox}_{\theta G}(y) \| \leq L_2 \theta \gamma_{C^0}(\text{Proj}_{\theta C^0}(y)) \leq L_2 \theta,$$  \hfill (35)

where the last inequality follows from Lemma 2(ii) since obviously $\text{Proj}_{\theta C^0}(y) \in \theta C^0$. Combining (33) and (35), we get the desired result. \qed
Corollary 5. Let $C_i, i = 1, \ldots, m$, be compact convex sets containing the origin as an interior point, i.e. convex bodies, and $G_i = \gamma C_i$ the associated gauges. For any $\theta, \theta' \in [0, +\infty]^m$, and any $y \in Y$, the following holds

$$
\left\| \Prox_{\theta_1 G_1} \cdots \Prox_{\theta_m G_m}(y) - \Prox_{\theta'_1 G_1} \cdots \Prox_{\theta'_m G_m}(y) \right\| \leq \sqrt{m} \max_i L_{2,i} \| \theta - \theta' \|
$$

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

Proof. Using repeatedly the triangle inequality, Proposition 5 and the fact that the mapping $y \mapsto \Prox_{\theta_i G_i}(y)$ is 1-Lipschitz [57], we obtain

$$
\left\| \Prox_{\theta_1 G_1} \cdots \Prox_{\theta_m G_m}(y) - \Prox_{\theta'_1 G_1} \cdots \Prox_{\theta'_m G_m}(y) \right\| = \\
\left\| \left( \Prox_{\theta_1 G_1} \circ \Prox_{\theta_2 G_2} \circ \cdots \circ \Prox_{\theta_m G_m}(y) - \Prox_{\theta'_1 G_1} \circ \Prox_{\theta'_2 G_2} \circ \cdots \circ \Prox_{\theta'_m G_m}(y) \right) + \\
\left( \Prox_{\theta'_1 G_1} \circ \Prox_{\theta'_2 G_2} \circ \cdots \circ \Prox_{\theta'_m G_m}(y) - \Prox_{\theta'_1 G_1} \circ \Prox_{\theta'_2 G_2} \circ \cdots \circ \Prox_{\theta'_m G_m}(y) \right) \right\| \\
\leq L_{2,1} |\theta_1 - \theta'_1| + \sum_i L_{2,i} |\theta_i - \theta'_i| \leq \max_i L_{2,i} \| \theta - \theta' \|_1 \leq \sqrt{m} \max_i L_{2,i} \| \theta - \theta' \|
$$

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 [28, 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 $\theta$ where the required Lipschitz continuity assumptions w.r.t. $\theta$ hold true.

Corollary 1. We first notice that $D_a^{(\ell)} = (D_{\xi}^{(\ell)}, D_{z_1}^{(\ell)}, \ldots, D_{z_{2q}}^{(\ell)})$ where $D_{\xi}^{(\ell)} = \partial_1 \xi^{(\ell)}(y, \theta)[\delta]$ and $D_{z_k}^{(\ell)} = \partial_1 z_k^{(\ell)}(y, \theta)[\delta]$. Hence, applying again the chain rule [28, Theorem 4 and Remark, Section 4.2.2] to the sequence of iterates and using the fact that all involved mappings are Lipschitz, and $D_{\xi}^{(\ell)} = \Gamma_a^{(\ell)}(D_{\xi}^{(\ell)}) = D_{\xi}^{(\ell)}$ concludes the proof.

Corollary 2. Observe that $J_a^{(\ell)} = (J_{\xi}^{(\ell)}, J_{z_1}^{(\ell)}, \ldots, J_{z_{2q}}^{(\ell)})$ where $J_{\xi}^{(\ell)} = \partial_2 \xi^{(\ell)}(y, \theta)$ and $J_{z_k}^{(\ell)} = \partial_2 z_k^{(\ell)}(y, \theta)$. Arguing as in the proof of Corollary 1, using now that $J_a^{(\ell)} = \Gamma_a^{(\ell)}(J_a^{(\ell)}) = J_a^{(\ell)}$ yields the formula.

Corollary 3. As before, but now with $D_a^{(\ell)} = (D_{\xi}^{(\ell)}, D_{z}^{(\ell)}, \ldots, D_{u}^{(\ell)})$ where $D_{\xi}^{(\ell)} = \partial_1 \xi^{(\ell)}(y, \theta)[\delta]$, $D_{z}^{(\ell)} = \partial_1 z^{(\ell)}(y, \theta)[\delta]$, $D_{u}^{(\ell)} = \partial_1 u^{(\ell)}(y, \theta)[\delta]$, and $D_{x}^{(\ell)} = \Gamma_a^{(\ell)}(D_{x}^{(\ell)}) = D_{x}^{(\ell)}$. The chain rule completes the proof.

Corollary 4. As before, but now with $J_a^{(\ell)} = (J_{\xi}^{(\ell)}, J_{z}^{(\ell)}, \ldots, J_{u}^{(\ell)})$ where $J_{\xi}^{(\ell)} = \partial_1 \xi^{(\ell)}(y, \theta)$, $J_{z}^{(\ell)} = \partial_1 z^{(\ell)}(y, \theta)$, $J_{u}^{(\ell)} = \partial_1 u^{(\ell)}(y, \theta)$ and $J_{x}^{(\ell)} = \Gamma_a^{(\ell)}(J_x^{(\ell)}) = J_x^{(\ell)}$. The chain rule completes the proof.
References


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


