HAL
archives-ouvertes.fr

Proximal Splitting Derivatives for Risk Estimation

Charles Deledalle, Samuel Vaiter, Gabriel Peyré, Jalal M. Fadili, Charles Dossal

▶ **To cite this version:**

Charles Deledalle, Samuel Vaiter, Gabriel Peyré, Jalal M. Fadili, Charles Dossal. Proximal Splitting Derivatives for Risk Estimation. NCMIP'12, Apr 2012, France. 386, pp.012003, 2012, .

HAL Id: hal-00670213

https://hal.archives-ouvertes.fr/hal-00670213

Submitted on 14 Feb 2012

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.

Proximal Splitting Derivatives for Risk Estimation

C. Deledalle[1], S. Vaiter[1], G. Peyré[1], J. Fadili[2], C. Dossal[3]

[1] CEREMADE, CNRS-Université Paris-Dauphine
[2] GREYC, CNRS-ENSICAEN-Université de Caen
[3] IMB, CNRS-Université Bordeaux 1

Abstract. This paper develops a novel framework to compute a projected Generalized Stein Unbiased Risk Estimator (GSURE) for a wide class of sparsely regularized solutions of inverse problems. This class includes arbitrary convex data fidelities with both analysis and synthesis mixed $\ell^1 - \ell^2$ norms. The GSURE necessitates to compute the (weak) derivative of a solution w.r.t. the observations. However, as the solution is not available in analytical form but rather through iterative schemes such as proximal splitting, we propose to iteratively compute the GSURE by differentiating the sequence of iterates. This provides us with a sequence of differential mappings, which, hopefully, converge to the desired derivative and allows to compute the GSURE. We illustrate this approach on total variation regularization with Gaussian noise and to sparse regularization with poisson noise,o automatically select the regularization parameter.

1. Introduction

This paper focuses on unbiased estimation of the ℓ^2-risk of recovering an image $f_0 \in \mathbb{R}^N$ from low-dimensional noisy observations $y = \Phi f_0 + w$, where $w \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_P)$. The linear bounded imaging operator $\Phi : \mathbb{R}^N \to \mathcal{Y} = \mathbb{R}^P$ entails loss of information so that $P < N$ or is rank-deficient for $P = N$, and the recovery problem is typically ill-posed.

In the following we denote $f(y) \in \mathbb{R}^N$ the estimator of f_0 from the observations $y \in \mathcal{Y}$. More specifically, we consider an estimator $f(y)$ defined as a function of coefficients $x(y) \in \mathcal{X}$ (where $\mathcal{X}$ is a suitable finite-dimensional Hilbert space) that solves $x(y) \in \operatorname{argmin}_{x \in \mathcal{X}} E(x, y)$ where the set of minimizers is nonempty. Here $E(x, y)$ is an energy functional parameterized by the observations $y \in \mathcal{Y}$. In some cases (e.g. total variation regularization), one directly has $f(y) = x(y)$, but for sparse regularization in a redundant synthesis dictionary, the latter maps coefficients $x(y)$ to images $f(y)$. We then make a distinction between $f(y)$ and $x(y)$ in the following.

This work proposes a versatile approach for unbiased risk estimation in the case where $x(y)$ is computed by proximal splitting algorithms. These methods have become extremely popular to solve inverse problems with convex non-smooth regularizations, e.g. those encountered in the sparsity field.

2. Previous Works

Unbiased Risk Estimation. The SURE [14] is an unbiased ℓ^2-risk estimator. For denoising $\Phi = \mathrm{Id}$, it provides an unbiased estimate $\mathrm{SURE}(y)$ of the risk $\mathbb{E}(\|f(y) - f_0\|^2)$ that depends solely on y, without prior knowledge of f_0. This can prove very useful for objective choice of parameters that minimize the recovery risk of f_0. A generalized SURE (GSURE) has been developed for noise models within the multivariate canonical exponential family [7]. In the context of inverse problems, this allows to compute the projected risk $\mathbb{E}(\|\Pi(f(y) - f_0)\|^2)$ where

Π is the orthogonal projector on $\ker(\Phi)^\perp$. Similar GSURE versions have been proposed for Gaussian noise and special regularizations or/and inverse problems, e.g. [12, 17].

SURE and GSURE have been applied to various inverse problem, to estimate the reconstruction risk $\mathbb{E}(\|\Phi(f(y)-f_0)\|^2)$, for wavelet denoising with linear expansion of thresholds [10], wavelet-vaguelet non-iterative thresholding [12], synthesis sparsity [6, 17, 8] and analysis sparsity [5].

Unbiased Estimation of the Degrees of Freedom. A prerequisite to compute the SURE or GSURE is an unbiased estimate of the degrees of freedom $\mathrm{df}(y)$. Roughly speaking, for overdetermined linear models, $\mathrm{df}(y)$ is the number of free parameters in modeling $f(y)$ from y. There are situations where $\mathrm{df}(y)$ can be estimated in closed-form from $f(y)$. This occurs e.g. in synthesis ℓ^1 regularization, as established in the overdetermined case in [19], and extended to the general setting in [9]. These results have been extended to analysis sparsity (for instance total variation) [16, 15]. When no closed-form is available, $\mathrm{df}(y)$ can be estimated using data perturbation and Monte-Carlo integration, see e.g. [18]. Alternatively, an estimate can be obtained by formally differentiating the sequence of iterates provided by an algorithm that converges to $f(y)$. As proposed initially by [17] and refined in [8], this allows to compute the GSURE of sparse synthesis regularization.

Proximal Splitting Algorithm. Convex optimization problems that appear in imaging applications usually enjoy a lot of structure, that is efficiently captured using proximal splitting methods. These methods are tailored to tackle large non-smooth convex optimization problems and are now mainstream in image processing. See for instance [3] for an overview of these methods. The precise algorithm to be used depends on the specific structure of the problem. The forward-backward (FB) algorithm handles the sum of a smooth and a simple function (for which the proximal mapping can be computed in closed form), see for instance [4]. The Douglas-Rachford (DR) algorithm [2] does not make any smoothness assumption, but requires that the functional is split into a sum of simple functions. The generalized foward-backward (GFB) algorithm [13] is an hybridization between FB and DR, thus enabling to take into account a smooth function and an arbitrary number of simple functions. Lastly, let us mention primal-dual schemes, such as the one recently proposed by Chambolle and Pock [1], that enables to minimize compositions of linear operators and simple functions. In this paper, we illustrate the versatility of our method by applying it to both primal (FB, DR and GFB) and primal-dual (CP) algorithms. The proposed methodology can however be adapted to any other proximal splitting method.

Contributions The main contribution of this paper is a a new risk estimation for arbitrary sparse regularizations when the solution is computed using a proximal splitting algorithm. This extends the previous iterative computation methods [17, 8] to a broader class of regularizations (e.g. total variation) without the need to resort to Monte-Carlo integration (such as [11]) which are numerically costly.

3. Differentiating Iterative Schemes

Iterative minimization scheme. The computation of $x(y)$ as a minimizer of some energy $E(x,y)$ is usually solved using an iterative scheme, initializing $x^{(0)}(y)$ (for instance to 0) and then updating

$$x^{(\ell+1)}(y) = \psi(x^{(\ell)}(y), y), \tag{1}$$

where $\psi : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ is a mapping, and so that $x^{(\ell)}(y)$ converge to a solution $x(y)$ that satisfies the (non-expansive) fixed-point equation

$$x(y) = \psi(x(y), y). \tag{2}$$

Note that we have made here explicit the dependency of the iterates $x^{(\ell)} = x^{(\ell)}(y)$ with the observation y.

A simple, but instructive example, is the case where $x \mapsto E(x, y)$ is a C^1 function with $1/L$ Lipschitz gradient, in which case one can use $\psi(x, y) = x - \tau \nabla_1 E(x, y)$ where $0 < \tau < 2/L$ and ∇_1 is the gradient of E with respect to the first variable. However, the functional E we are interested in are typically not C^1. The remaining sections of this paper develops more advanced proximal splitting schemes to define the iterations.

Iterative differentiation. The goal is then derive an algorithm to compute $\partial x(y)$, the derivative of $y \in \mathcal{Y} \mapsto x(y) \in \mathcal{X}$. A natural way to achieve this goal is to apply an implicit function theorem to (2). This is equivalent to applying this implicit function theorem to the first order condition $0 \in \partial_1 E(x(y), y)$ where $\partial_1 E$ is the sub-derivative of E with respect to the first variable. This is essentially the methods used in [9, 16] for the special case of ℓ^1 norm regularization.

While this leads to important theoretical results, the resulting closed form formula are in practice quite unstable and require the computation of $x(y)$ with high precision. Furthermore, the explicit computation of this differential is out of reach for many variational regularization functionals E.

A more practical way to achieve this goal, introduced in [17], and that we pursue here, is to compute iteratively this derivative. This is achieved by deriving formula (1), which allows, for any vector $\delta \in \mathcal{X}$ to compute $\xi^{(\ell)} = \partial x^{(\ell)}(y)[\delta]$ (the derivative of $y \mapsto x^{(\ell)}(y)$ applied at δ) as

$$\xi^{(\ell+1)} = \Psi_1^{(\ell)}(\xi^{(\ell)}) + \Psi_2^{(\ell)}(\delta),$$

where we have defined, for $k = 1, 2$, the following linear operators

$$\Psi_k^{(\ell)} = \partial_k \psi(x^{(\ell)}(y), y).$$

The following sections are devoted to the extension of this approach to more complicated iterative algorithm that are able to handle non-smooth functionals E.

4. Differentiating Proximal Splitting Schemes

4.1. Proximal Operator

The proximal operator associated to a proper lower semi-continuous (lsc) and convex function $x \mapsto G(x, y)$ is

$$\mathrm{Prox}_G(x, y) = \operatorname*{argmin}_z \frac{1}{2}\|x - z\|^2 + G(z, y).$$

A function for which $\mathrm{Prox}_G(x, y)$ can be computed in closed-form is dubbed simple. A distinctive property of $\mathrm{Prox}_G(\cdot, y)$ that plays a central role in the sequel is that its is a 1-Lipschitz mapping. When y is fixed, we will denote $\mathrm{Prox}_G(x)$ instead of $\mathrm{Prox}_G(x, y)$ to lighten the notation.

The Legendre-Fenchel conjugate of G is $G^*(z, y) = \max_x \langle x, z \rangle - G(x, y)$. A useful proximal calculus rule is Moreau's identity: $x = \mathrm{Prox}_{\tau G^*}(x, y) + \tau \,\mathrm{Prox}_{G/\tau}(x/\tau, y),\ \tau > 0$.

4.2. Generalized Forward Backward Splitting

The Generalized Forward Backward (GFB) splitting [13] is allows one to solve a variation problem of the form

$$x(y) \in \operatorname*{argmin}_{x \in \mathcal{X}} E(x,y) = F(x,y) + \sum_{i=1}^{Q} G_i(x,y) \tag{3}$$

under the hypothesis that F is C^1 with $1/L$ gradient and the G_i functions are simple. It reads, for all $i = 1, \ldots, Q$,

$$z_i^{(\ell+1)} = z_i^{(\ell)} - x^{(\ell)} + \operatorname{Prox}_{n\gamma G_i}(X^{(\ell)}) \quad \text{where} \quad X^{(\ell)} = 2x^{(\ell)} - z_i^{(\ell)} - \gamma \nabla_1 F(x^{(\ell)})$$

and $x^{(\ell+1)} = \frac{1}{Q}\sum_{i=1}^{Q} z_i^{(\ell+1)}$. One recovers as a special case the Foward-Backward method [4] when $Q = 1$ and the Douglas-Rachford [2] when $F = 0$.

Please note that in this algorithm, the iterates $z_i^{(\ell)}, x^{(\ell)}, \tilde{x}^{(\ell)}$ are actually functions of the observations y.

For any vector $\delta \in \mathcal{X}$, we wish to compute $\xi^{(\ell)} = \partial x^{(\ell)}(y)[\delta]$, $\zeta_i^{(\ell)} = \partial z_i^{(\ell)}(y)[\delta]$ and $\tilde{\xi}^{(\ell)} = \partial \tilde{x}^{(\ell)}(y)[\delta]$. The iterations on the derivatives reads

$$\zeta_i^{(\ell+1)} = \zeta_i^{(\ell)} - \xi^{(\ell)} + \mathcal{G}_{i,1}^{(\ell)}(\Xi^{(\ell)}) + \mathcal{G}_{i,2}^{(\ell)}(\delta) \quad \text{where} \quad \Xi^{(\ell)} = 2\xi^{(\ell)} - \zeta_i^{(\ell)} - \gamma(\mathcal{F}_1^{(\ell)}(\xi^{(\ell)}) + \mathcal{F}_2^{(\ell)}(\delta))$$

and $\xi^{(\ell+1)} = \frac{1}{Q}\sum_{i=1}^{Q} \zeta_i^{(\ell+1)}$, where we have defined the following linear operators for $k = 1, 2$:

$$\mathcal{G}_{i,k}^{(\ell)} = \partial_k \operatorname{Prox}_{n\gamma G_i}(X^{(\ell)}, y) \quad \text{and} \quad \mathcal{F}_k^{(\ell)} = \partial_k \nabla_1 F(x^{(\ell)}, y)\ .$$

4.3. Primal-Dual Splitting

Proximal splitting schemes can be used to solve the large class of variational problems

$$x(y) \in \operatorname*{argmin}_{x \in \mathcal{X}} E(x,y) = H(x,y) + G(Kx,y)\ , \tag{4}$$

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

The primal-dual relaxed Arrow-Hurwicz algorithm as proposed in [1] to solve (4) reads

$$\begin{aligned} u^{(\ell+1)} &= \operatorname{Prox}_{\sigma H^*}(U^{(\ell)}) \quad \text{where} \quad U^{(\ell)} = u^{(\ell)} + \sigma K \tilde{x}^{(\ell)}, \\ x^{(\ell+1)} &= \operatorname{Prox}_{\tau G}(X^{(\ell)}) \quad \text{where} \quad X^{(\ell)} = x^{(\ell)} - \tau K^* u^{(\ell)}, \\ \tilde{x}^{(\ell+1)} &= x^{(\ell+1)} + \theta(x^{(\ell+1)} - x^{(\ell)}) \end{aligned} \tag{5}$$

where $u^{(\ell)} \in \mathcal{U}$, $x^{(\ell)} \in \mathcal{X}$ and $\tilde{x}^{(\ell)} \in \mathcal{X}$. The parameters $\sigma > 0, \gamma > 0$ are chosen such that $\sigma\gamma\|K\|^2 < 1$, and $\theta \in [0,1]$ to ensure convergence of $x^{(\ell)}$ toward a global minimizer of (4). $\theta = 0$ corresponds to the Arrow-Hurwitz algorithm, and for $\theta = 1$ a convergence rate of $O(1/\ell)$ was established on the restricted duality gap [1].

For any vector $\delta \in \mathcal{Y}$, our goal is to compute the derivatives $\xi^{(\ell)} = \partial x^{(\ell)}(y)[\delta]$, $\upsilon^{(\ell)} = \partial u^{(\ell)}(y)[\delta]$ and $\tilde{\xi}^{(\ell)} = \partial \tilde{x}^{(\ell)}(y)[\delta]$. Using the chain rule, the sequence of derivatives then reads

$$\begin{aligned} \upsilon^{(\ell+1)} &= \mathcal{H}_1^{(\ell)}(\Upsilon^{(\ell)}) + \mathcal{H}_2^{(\ell)}(\delta) \quad \text{where} \quad \Upsilon^{(\ell)} = \upsilon^{(\ell)} + \sigma K \tilde{\xi}^{(\ell)}, \\ \xi^{(\ell+1)} &= \mathcal{G}_1^{(\ell)}(\Xi^{(\ell)}) + \mathcal{G}_2^{(\ell)}(\delta) \quad \text{where} \quad \Xi^{(\ell)} = \xi^{(\ell)} - \tau K^* \upsilon^{(\ell)}, \\ \tilde{\xi}^{(\ell+1)} &= \xi^{(\ell+1)} + \theta(\xi^{(\ell+1)} - \xi^{(\ell)}) \end{aligned} \tag{6}$$

where we have defined the following linear mappings for $k = 1, 2$ with ∂_k the derivative w.r.t. the k-th argument

$$\mathcal{H}_k^{(\ell)}(\cdot) = \partial_k \operatorname{Prox}_{\sigma H^*}(U^{(\ell)}, y)[\cdot] \quad \text{and} \quad \mathcal{G}_k^{(\ell)}(\cdot) = \partial_k \operatorname{Prox}_{\tau G}(X^{(\ell)}, y)[\cdot].$$

4.4. Discussion on convergence issues

One has to be aware that given that the proximal mappings are not necessarily differentiable everywhere, its differential is actually set-valued. Therefore, one should appeal to involved tools from non-smooth analysis to make the above statements rigorous. We prefer not to delve into these technicalities for the lack of space.

Another major issue is to theoretically ensure the existence of a proper sequence $\xi^{(\ell)}$ that converges toward $\partial x(y)[\delta]$. Regarding existence, as $\mathrm{Prox}_G(\cdot, y)$ is a 1-Lipschitz mapping of its first argument. Furthermore, in all the considered application, $\mathrm{Prox}_G(x, \cdot)$ is also Lipschitz with respect to its second argument. If one starts at an appropriate initialization, by induction, $y \mapsto x^{(\ell)}(y)$ is also Lipschitz, hence differentiable almost everywhere. Note that for sparse synthesis regularization, convergence can be ensured as for ℓ large enough, $\partial x^{(\ell)}(y)$ is constant equal to $\partial x(y)$. As far as convergence 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.

5. Risk Estimator

Projected GSURE. Recall that $\Pi = \Phi^*(\Phi\Phi^*)^+\Phi$ is the orthogonal projector on $\ker(\Phi)^\perp = \mathrm{Im}(\Phi^*)$, where $^+$ stands for the Moore-Penrose pseudo-inverse. Let $\mu(y) = \Pi f(y)$ the projected estimator of Πf_0. While $f(y)$ is not necessarily uniquely defined, we assume that $\mu(y)$ is unambiguously defined as a single-valued mapping of the observation y. This can be ensured under a strict convexity condition on H or G in (4) (see e.g. example (9)).

Let $\mu_0(y) = \Phi^*(\Phi\Phi^*)^+y$ the maximum likelihood estimator. By generalizing the projected GSURE of [12] to any linear operator Φ, we have

$$\mathrm{GSURE}(y) = \|\mu_0(y) - \mu(y)\|^2 - \sigma^2 \operatorname{tr}((\Phi\Phi^*)^+) + 2\sigma^2 \operatorname{div}((\Phi\Phi^*)^+\Phi f(y)) \tag{7}$$

where $\operatorname{div}(g)(y) = \operatorname{tr}(\partial g(y))$ is the divergence of the mapping $g : \mathcal{Y} \to \mathcal{Y}$. Under weak differentiability of $y \mapsto \mu(y)$, one can prove that the GSURE is an unbiased estimate of the risk on $\mathrm{Im}(\Phi^*)$, i.e. $\mathbb{E}_w(\mathrm{GSURE}(y)) = \mathbb{E}_w(\|\Pi f_0 - \mu(y)\|^2)$ (see Appendix A).

Iterative Numerical Computation. One of the bottlenecks in calculating the GSURE(y) is to efficiently compute the divergence term. Using the Jacobian trace formula of the divergence, it can be easily seen that

$$\operatorname{div}((\Phi\Phi^*)^+\Phi f(y)) = \mathbb{E}_z(\langle \partial f(y)[z],\, \mu_0(z)\rangle) \approx \frac{1}{k}\sum_{i=1}^{k} \langle \partial f(y)[z_i],\, \mu_0(z_i)\rangle \tag{8}$$

where $z \sim \mathcal{N}(0, \mathrm{Id}_P)$ and z_i are k realizations of z. Since $f(y)$ and its iterates $f^{(\ell)}(y)$ are defined as explicit functions of $x(y)$ and $x^{(\ell)}(y)$ (see Section 6 for a detailed example), the GSURE(y) can in turn be iteratively estimated by plugging $\partial x^{(\ell)}(y)[z_i]$ provided by (6) into (8).

6. Numerical Results

6.1. Total Variation Regularization

Total variation regularization of linear inverse problems amounts to solving

$$f(y) \in \operatorname*{argmin}_f \frac{1}{2}\|\Phi f - y\|^2 + \lambda\|\nabla f\|_1 \ , \tag{9}$$

where $\nabla f \in \mathbb{R}^{N\times 2}$ is a discrete gradient. 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 $\|t\|_1 = \sum_i \|t_i\|$.

CP formulation. Problem (9) is a special instance of (4) letting $x = f$, $H(x, y) = 0, \forall (x, y)$ and

$$K(x) = (\Phi x, \nabla x) \quad \text{and} \quad \forall\, u = (s, t) \in \mathbb{R}^P \times \mathbb{R}^{N\times 2}, \quad G(u, y) = \frac{1}{2}\|s - y\|^2 + \lambda\|t\|_1 \ .$$

Separability of G in s and t entails that

$$\mathrm{Prox}_{\tau G}(u, y) = ((1 - \tau)s + \tau y, T_{\lambda\tau}(t)) \ ,$$

where $T_\rho, \rho > 0$, is the component-wise $\ell^1 - \ell^2$ soft-thresholding, defined for $i = 1, \ldots, N$ as

$$T_\rho(t)_i = \max(0, 1 - \rho/\|t_i\|)t_i \quad \text{and} \quad \partial T_\rho(t)[\delta_t]_i = \begin{cases} 0 & \text{if} \quad \|t_i\| \leqslant \rho \\ \delta_{t,i} - \frac{\rho}{\|t_i\|} P_{t_i}(\delta_{t,i}) & \text{otherwise} \end{cases} \ , \tag{10}$$

where P_α is the orthogonal projector on $\alpha^\perp$ for $\alpha \in \mathbb{R}^2$, and $T_\rho(t)_i$ although not differentiable on the sphere $\{t_i : \|t_i\| = \rho\}$, is directionally differentiable there. Therefore

$$\partial_1 \mathrm{Prox}_{\tau G}(u, y)[\delta_s, \delta_t] = ((1 - \tau)\delta_s, \partial T_{\lambda\tau}(\delta_t)) \quad \text{and} \quad \partial_2 \mathrm{Prox}_{\tau G}(u, y)[\delta_y] = (\tau\delta_y, 0) \ .$$

GFB formulation. It can also be recast as (3) using $x = (f, u) \in \mathcal{X} = \mathbb{R}^N \times \mathbb{R}^{N\times 2}$, $Q = 2$ simple functional, and for $x = (f, u)$

$$F(x, y) = \frac{1}{2}\|\Phi f - y\|^2, \quad G_1(x, y) = \lambda\|u\|_1, \quad \text{and}$$

$$G_2(x, y) = \iota_{\mathcal{C}}(x) \quad \text{where} \quad \mathcal{C} = \{x = (f, u) \setminus u = \nabla f\} \ .$$

One has

$$\nabla_1 F(x, y) = (\Phi^*(\Phi f - y), 0),$$

and thus

$$\partial_1 \nabla_1 F(x, y)[\delta_f, \delta_u] = (\Phi^*\Phi\delta_f, 0) \quad \text{and} \quad \partial_2 \nabla_1 F(x, y)[\delta_y] = (-\Phi^*\delta_y, 0)$$

and

$$\mathrm{Prox}_{\tau G_1}(x, y) = (f, T_{\lambda\tau}(u))$$

where T_ρ is the vectorial soft thresholding defined in eq. (10), and thus

$$\partial_1 \mathrm{Prox}_{\tau G_1}(x, y)[\delta_f, \delta_u] = (\delta_f, \partial T_{\lambda\tau}(\delta_u)) \quad \text{and} \quad \partial_2 \mathrm{Prox}_{\tau G_1}(x, y)[\delta_y] = (0, 0)$$

and

$$\mathrm{Prox}_{\tau G_2}(x, y) = ((\mathrm{Id} + \Delta)^{-1}(u + \mathrm{div}\, f), \nabla(\mathrm{Id} + \delta)^{-1}(u + \mathrm{div}\, f))$$

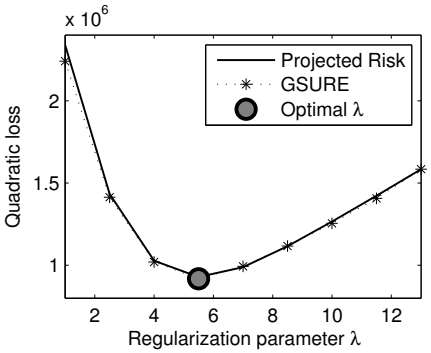
where Δ is the Laplace operator and div corresponds to the adjoint operator of ∇, and thus

$$\begin{aligned} \partial_1 \mathrm{Prox}_{\tau G_2}(x, y)[\delta_f, \delta_u] &= ((\mathrm{Id} + \Delta)^{-1}(\delta_u + \mathrm{div}\, \delta_f), \nabla(\mathrm{Id} + \Delta)^{-1}(\delta_u + \mathrm{div}\, \delta_f)), \quad \text{and} \\ \partial_2 \mathrm{Prox}_{\tau G_2}(x, y)[\delta_y] &= (0, 0) \ . \end{aligned}$$

Figure 1 shows an application of our GSURE to estimate the optimal λ parameter. It is applied to a super-resolution problem, where $\Phi \in \mathbb{R}^{P\times N}$ is a vertical sub-sampling matrix, where $(P/N = 0.5)$, $\sigma = 10$ (for an image f_0 with a range $[0, 255]$). For each value of λ in the tested range, GSURE(y) is computed for a single realization of y using (8) with $k = 1$ realizations z_i.

Fig. 2 depicts an application of our GSURE to adjust the value of λ optimizing the recovery for a deblurring problem, where $\Phi \in \mathbb{R}^{N\times N}$ is a convolution matrix (Gaussian kernel of width 2 px), $\sigma = 10$ (for an image f_0 with a range $[0, 255]$). For each value of λ in the tested range, GSURE(y) is computed for a single realization of y using (8) with $k = 4$ realizations z_i.

[1] Without impacting the optimal choice of λ, the two curves have been vertically shifted for visualization purposes.

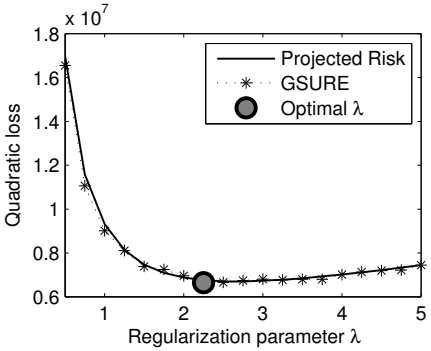


(a)

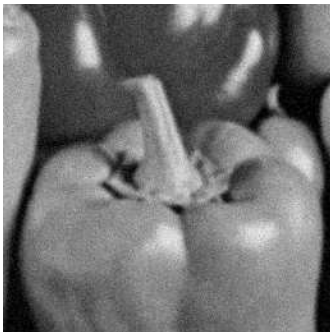
(b) y

(c) $f(y)$ at the optimal λ

Figure 1. (a) Projected risk and its GSURE estimate[1] (b) y. (c) $f(y)$ at the optimal λ.



(a)



(b) y

(c) $f(y)$ at the optimal λ

Figure 2. (a) Projected risk and its GSURE estimate[1] (b) y. (c) $f(y)$ at the optimal λ.

6.2. Sparse synthesis with block sparsity

Mixed $\ell^1 - \ell^2$ norm promoting block sparsity of linear inverse problems amounts to solving

$$f(y) = \Psi^* x(y) \quad \text{and} \quad x(y) \in \operatorname*{argmin}_x \frac{1}{2}\|\Phi\Psi x - y\|^2 + \lambda \sum_{i=1}^{Q} \|\mathcal{B}_i x\|_1 \ , \tag{11}$$

where $\Psi \in \mathbb{R}^{N\times N}$ is an orthonormal synthesis dictionary and $\mathcal{B}_i : \mathbb{R}^N \to \mathbb{R}^{N'\times B}$ are Q linear operators extracting blocks of size B such as $N = N' \times B$. The operators $\mathcal{B}_i$ are designed such that each of them corresponds to a different partition of $\mathbb{R}^N$ into N' non-overlapping blocks of size B. The $\ell^1 - \ell^2$ norm of $t = (t_i)_{i=1}^N \in \mathbb{R}^{N'\times B}$, with $t_i \in \mathbb{R}^B$, is defined as $\|t\|_1 = \sum_i \|t_i\|$.

CP formulation. When $Q = 1$, problem (9) is a special instance of (4) letting $x = \Phi^* f$, $H(x, y) = 0, \forall (x, y)$ and

$$K(x) = (\Phi x, \mathcal{B}_1 x) \quad \text{and} \quad \forall\, u = (s, t) \in \mathbb{R}^P \times \mathbb{R}^{N\times 2}, \quad G(u, y) = \frac{1}{2}\|s - y\|^2 + \lambda\|t\|_1 \ .$$

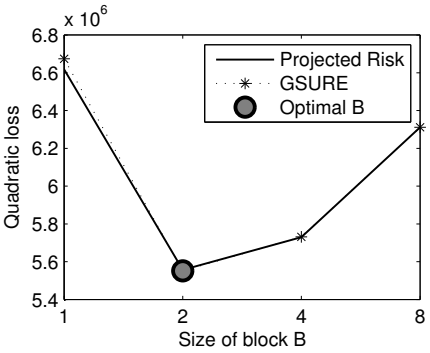
The proximal operator associated to G has already been given in the total variation case. When $Q > 1$, CP formulation cannot be used and one could instead use the GFB formulation.

GFB formulation. It can also be recast as (3) using $x = \Phi^* f$, Q simple functionals, and for all $i = 1, \ldots, Q$

$$F(x, y) = \frac{1}{2}\|\Phi\Psi x - y\|^2, \quad G_i(x, y) = \lambda\|\mathcal{B}_i x\|_1 \ .$$

One has

$$\nabla_1 F(x, y) = \Psi^*\Phi^*(\Phi\Psi x - y),$$



(a)

(b) $\mu_0(y)$

(c) $f(y)$ at the optimal λ

Figure 3. (a) Projected risk and its GSURE estimate[1] (b) $\mu_0(y)$. (c) $f(y)$ at the optimal λ.

and thus

$$\partial_1 \nabla_1 F(x,y)[\delta_x] = \Psi^* \Phi^* \Phi \Psi \delta_x \quad \text{and} \quad \partial_2 \nabla_1 F(x,y)[\delta_y] = -\Psi^* \Phi^* \delta_y$$

and

$$\mathrm{Prox}_{\tau G_i}(x,y) = \mathcal{B}_i^* T_{\lambda\tau}(\mathcal{B}_i x)$$

where T_ρ is the vectorial soft thresholding defined in eq. (10), and thus

$$\partial_1 \,\mathrm{Prox}_{\tau G_i}(x,y)[\delta_x] = \mathcal{B}_i^* \partial T_{\lambda\tau}(\mathcal{B}_i \delta_x) \quad \text{and} \quad \partial_2 \,\mathrm{Prox}_{\tau G_i}(x,y)[\delta_y] = 0 \ .$$

Fig. 3 depicts an application of our GSURE to adjust the size of blocks B optimizing the recovery for a compressed sensing problem, where $\Phi \in \mathbb{R}^{P\times N}$ is a realization of a random matrix distribution (randomized sub-sampling of a random convolution), where $(P/N = 0.5)$, $\sigma = 10$ (for an image f_0 with a range $[0, 255]$). For each size B, $Q = B^2$, i.e., all possible partitions have been used, and λ has been set to the value $0.7\sigma/B$. For each value of λ in the tested range, GSURE(y) is computed for a single realization of y using (8) with $k = 1$ realizations z_i.

7. Conclusion

We obtained proximal splitting derivatives for unbiasedly estimate the projected risk in regularized inverse problems handling both synthesis and analysis sparsity priors as well as mixed norms for block structured sparsity. Its usefulness has been illustrated on automatic choice of the regularization parameter for total variation regularization and automatic choice of the size of blocks for sparse synthesis with block sparsity.

Appendix A. Proof of equation (7)

Assume $w \mapsto g(w)$ is weakly differentiable in the sense of [14], Stein's lemma reads

$$\mathbb{E}_w \left\langle w, g(w) \right\rangle = \sigma^2 \,\mathrm{div}\, g(w) \ .$$

Under weak differentiability of $y \mapsto \mu(y)$ and using the fact that $\mu_0(y) = \Pi f_0 + \Phi^*(\Phi\Phi^*)^+ w$ and $\mu(y) = \Pi f(y)$, one has

$$\begin{aligned}
&\mathbb{E}\|\mu_0(y) - \mu(y)\|^2 \\
&= \mathbb{E}\|\mu_0(y)\|^2 - 2\mathbb{E}\left\langle \mu_0(y), \mu(y) \right\rangle + \mathbb{E}\|\mu(y)\|^2 \\
&= \mathbb{E}\|\Pi f_0 + \Phi^*(\Phi\Phi^*)^+ w\|^2 - 2\mathbb{E}\left\langle \Pi f_0 + \Phi^*(\Phi\Phi^*)^+ w, \mu(y) \right\rangle + \mathbb{E}\|\mu(y)\|^2 \\
&= \mathbb{E}\|\Pi f_0\|^2 + \sigma^2 \,\mathrm{tr}((\Phi\Phi^*)^+) - 2\mathbb{E}\left\langle \Pi f_0, \mu(y) \right\rangle - 2\mathbb{E}\left\langle w, (\Phi\Phi^*)^+ \Phi f(y) \right\rangle + \mathbb{E}\|\mu(y)\|^2 \\
&= \mathbb{E}\|\Pi f_0 - \mu(y)\|^2 + \sigma^2 \,\mathrm{tr}((\Phi\Phi^*)^+) - 2\sigma^2 \,\mathrm{div}((\Phi\Phi^*)^+ \Phi f(y)) \ .
\end{aligned}$$

□

References

[1] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[2] P. L. Combettes and J.-. Pesquet. A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE J. Selected Topics in Signal Processing*, 1(4):564–574, 2007.

[3] P. L. Combettes and J.-C. Pesquet. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, chapter Proximal Splitting Methods in Signal Processing, pages 185–212. Springer-Verlag, 2011.

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

[5] C. Deledalle, S. Vaiter, G. Peyré, J. Fadili, and C. Dossal. Unbiased risk estimation for sparse analysis regularization. Technical report, Preprint Hal-??, 2012.

[6] F.-X. Dupé, M.J. Fadili, and J-L. Starck. A proximal iteration for deconvolving Poisson noisy images using sparse representations. *IEEE Transactions on Image Processing*, 18(2), 2009. 310–321.

[7] Y. C. Eldar. Generalized SURE for exponential families: Applications to regularization. *IEEE Transactions on Signal Processing*, 57(2):471–481, 2009.

[8] R. Giryes, M. Elad, and Y.C. Eldar. The projected GSURE for automatic parameter tuning in iterative shrinkage methods. *Applied and Computational Harmonic Analysis*, 30(3):407–422, 2011.

[9] M. Kachour, C. Dossal, J. Fadili, G. Peyré, and C. Chesneau. The degrees of freedom of penalized l1 minimization. Technical report, Preprint Hal-00638417, 2011.

[10] F. Luisier. *The SURE-LET approach to image denoising*. PhD thesis, École polytechnique fédérale de lausanne, 2010.

[11] Lusier. ? *IEEE Transactions on Image Processing*, 2011.

[12] J-C. Pesquet, A. Benazza-Benyahia, and C. Chaux. A SURE approach for digital signal/image deconvolution problems. *IEEE Transactions on Signal Processing*, 57(12):4616–4632, 2009.

[13] H. Reguet, J. Fadili, and G. Peyré. Generalized forward-backward splitting. Technical report, Preprint Hal-??, 2011.

[14] C.M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

[15] R.J. Tibshirani and J. Taylor. The solution path of the generalized Lasso. *The Annals of Statistics*, 39(3):1335–1371, 2011.

[16] S. Vaiter, G. Peyré, C. Dossal, and M.J. Fadili. Robust sparse analysis regularization. Technical report, Preprint Hal-00627452, 2011.

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

[18] J. Ye. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93, 1998. 120–131.

[19] H. Zou, T. Hastie, and R. Tibshirani. On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.