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 to 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 [1] 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 [2]. 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. [3, 4].

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 [5], and extended to the general setting in [6]. When no closed-form is available, $\mathrm{df}(y)$ can be estimated using data perturbation and Monte-Carlo integration, see e.g. [7]. 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 [4] and refined in [8], this allows to compute the GSURE of sparse synthesis regularization.

3. Differentiating Proximal Splitting Schemes

Due to obvious space limitations, we only describe here the derivative of a primal-dual proximal splitting algorithm. But the same idea remains valid for other splitting schemes as it is described in a longer version of this paper [9].

3.1. Proximal Operator

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

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

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: $x = \mathrm{Prox}_{\tau G^*}(x, y) + \tau\,\mathrm{Prox}_{G/\tau}(x/\tau, y),\ \ \tau > 0$.

3.2. Primal-Dual Splitting

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

$$x(y) \in \underset{x \in \mathcal{X}}{\mathrm{argmin}}\ E(x, y) = H(x, y) + G(Kx, y)\ , \tag{1}$$

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 [10] to solve (1) reads

$$\begin{aligned}
u^{(\ell+1)} &= \mathrm{Prox}_{\sigma H^*}(U^{(\ell)}) \quad \text{where} \quad U^{(\ell)} = u^{(\ell)} + \sigma K \tilde{x}^{(\ell)},\\
x^{(\ell+1)} &= \mathrm{Prox}_{\tau G}(X^{(\ell)}) \quad \text{where} \quad X^{(\ell)} = x^{(\ell)} - \tau K^* u^{(\ell)},\\
\tilde{x}^{(\ell+1)} &= x^{(\ell+1)} + \theta(x^{(\ell+1)} - x^{(\ell)})
\end{aligned} \tag{2}$$

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 (1). $\theta = 0$ corresponds to the Arrow-Hurwitz algorithm, and for $\theta = 1$ a convergence rate of $O(1/\ell)$ was established on the restricted duality gap [10].

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

$$\begin{aligned}
\upsilon^{(\ell+1)} &= \mathcal{H}_1^{(\ell)}(\Upsilon^{(\ell)}) + \mathcal{H}_2^{(\ell)}(\delta) \quad \text{where} \quad \Upsilon^{(\ell)} = \upsilon^{(\ell)} + \sigma K \tilde{\xi}^{(\ell)}, \\
\xi^{(\ell+1)} &= \mathcal{G}_1^{(\ell)}(\Xi^{(\ell)}) + \mathcal{G}_2^{(\ell)}(\delta) \quad \text{where} \quad \Xi^{(\ell)} = \xi^{(\ell)} - \tau K^* \upsilon^{(\ell)}, \\
\tilde{\xi}^{(\ell+1)} &= \xi^{(\ell+1)} + \theta(\xi^{(\ell+1)} - \xi^{(\ell)})
\end{aligned} \tag{3}$$

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

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

3.3. 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, $\operatorname{Prox}_G(\cdot, y)$ is a 1-Lipschitz mapping of its first argument. Furthermore, in all the considered application, $\operatorname{Prox}_G(x, \cdot)$ 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. 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.

4. Risk Estimator

Projected GSURE. Recall that $\Pi = \Phi^*(\Phi\Phi^*)^+\Phi$ is the orthogonal projector on $\ker(\Phi)^\perp = \operatorname{Im}(\Phi^*)$, 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 (1) (see e.g. example (5)).

Let $\mu_0(y) = \Phi^*(\Phi\Phi^*)^+ y$ the maximum likelihood estimator. By generalizing the projected GSURE of [3] to any linear operator Φ, we have

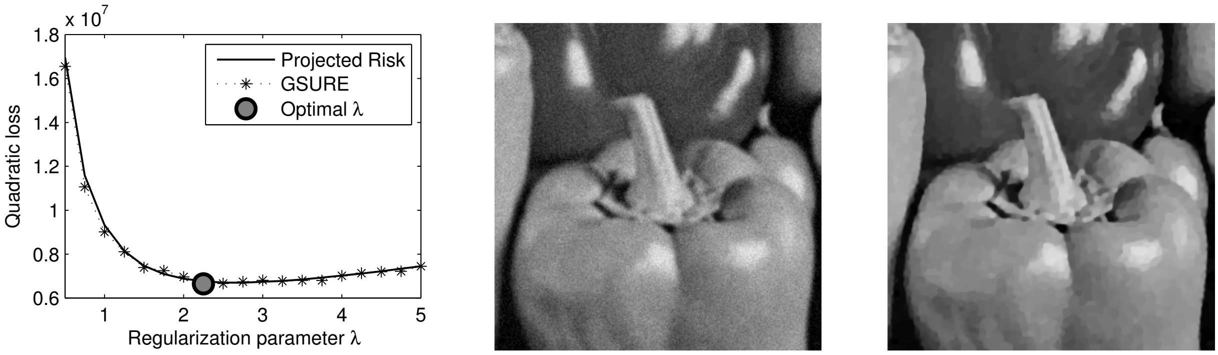
$$\operatorname{GSURE}(y) = \|\mu_0(y) - \mu(y)\|^2 - \sigma^2 \operatorname{tr}((\Phi\Phi^*)^+) + 2\sigma^2 \operatorname{div}((\Phi\Phi^*)^+ \Phi f(y))$$

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 $\operatorname{Im}(\Phi^*)$, i.e. $\mathbb{E}_w(\operatorname{GSURE}(y)) = \mathbb{E}_w(\|\Pi f_0 - \mu(y)\|^2)$.

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

$$\operatorname{div}((\Phi\Phi^*)^+ \Phi f(y)) = \mathbb{E}_z(\langle \partial f(y)[z], \, \mu_0(z) \rangle) \approx \frac{1}{k} \sum_{i=1}^{k} \langle \partial f(y)[z_i], \, \mu_0(z_i) \rangle \tag{4}$$

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 5 for a detailed example), the GSURE(y) can in turn be iteratively estimated by plugging $\partial x^{(\ell)}(y)[z_i]$ provided by (3) into (4).



(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.

5. Numerical Results

Total variation regularization of linear inverse problems amounts to solving

$$f(y) \in \underset{f}{\operatorname{argmin}}\ \frac{1}{2}\|\Phi f - y\|^2 + \lambda \|\nabla f\|_1 \ , \tag{5}$$

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 $\|t\|_1 = \sum_i \|t_i\|$. Problem (5) is a special instance of (1) 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

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

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

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

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

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

Fig. 1 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 (4) with $k = 4$ realizations z_i.

References

[1] Stein C 1981 *The Annals of Statistics* **9** 1135–1151

[2] Eldar Y C 2009 *IEEE Transactions on Signal Processing* **57** 471–481

[3] Pesquet J C, Benazza-Benyahia A and Chaux C 2009 *IEEE Transactions on Signal Processing* **57** 4616–4632

[4] Vonesch C, Ramani S and Unser M 2008 *ICIP* (IEEE) pp 665–668

[5] Zou H, Hastie T and Tibshirani R 2007 *The Annals of Statistics* **35** 2173–2192

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

[7] Ye J 1998 *Journal of the American Statistical Association* **93** 120–131

[8] Giryes R, Elad M and Eldar Y 2011 *Applied and Computational Harmonic Analysis* **30** 407–422

[9] Deledalle C, Vaiter S, Peyré G, Fadili J and Dossal C 2012 Proximal splitting derivatives for unbiased risk estimation Tech. rep. Preprint Hal-00662719 URL http://hal.archives-ouvertes.fr/hal-00662719

[10] Chambolle A and Pock T 2011 *Journal of Mathematical Imaging and Vision* **40** 120–145

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