Chapter 4
Subdifferential and proximity operator

Nelly Pustelnik

Chapter objectives.

- Replace the gradient by the *subdifferential* for non-differentiable convex functions (ℓ_1, total variation, constraints, hinge loss).
- Know how to compute a subdifferential (calculus rules and qualification conditions) and write Fermat's rule $0 \in \partial f(\widehat{x})$.
- Define the *proximity operator* and the Moreau envelope, characterize them, and know their properties (non-expansiveness, fixed points).
- Compute the usual proximity operators (projections, soft thresholding, powers, quadratics) and the calculus rules (separability, orthonormal basis, composition with a linear operator).

Contents

1 Motivation: when the gradient is no longer enough **2**

2 Subdifferential **3**

2.1 Definition . . . 3

2.2 Fundamental properties . . . 4

3 Subdifferential calculus **5**

3.1 Examples . . . 6

3.2 Application: optimality conditions for the Lasso . . . 8

4 Proximity operator **8**

4.1 Definition . . . 8

4.2 Characterization and interpretation . . . 9

4.3 Examples of proximity operators . . . 11

4.4 Calculus rules . . . 12

4.5 Examples: denoising as a prox computation . . . 13

4.6 Resources . . . 15

5 Towards proximal algorithms **16**

5.1 Proximal point algorithm . . . 16

5.2 Proximal gradient (forward–backward) . . . 17

5.3 Why not simply use a subgradient? . . . 17

6 Exercises **19**

The needs of mechanics, more precisely the dynamics of material systems with unilateral constraints, and also the statics of such systems (elastic medium with limited compressibility of W. Prager), have led us to the following extension of the idea of projection onto a closed convex set.
– Jean-Jacques Moreau –
Proximité et dualité dans un espace hilbertien, 1965

Notation and reminders (Chapter 2). $\mathcal{H}$ and $\mathcal{G}$ denote real Hilbert spaces, $\langle \cdot \,|\, \cdot \rangle$ their inner product, $\mathcal{B}(\mathcal{H}, \mathcal{G})$ the set of bounded linear operators and L^* the adjoint of L. We denote by $\Gamma_0(\mathcal{H})$ the set of proper, lower semicontinuous and convex functions $f : \mathcal{H} \to]-\infty, +\infty]$, by ι_C the indicator function of a set C, and by $2^{\mathcal{H}}$ the power set of $\mathcal{H}$. For $C \subset \mathcal{H}$, $\operatorname{int} C$ denotes the interior of C; for a set-valued operator $A : \mathcal{H} \to 2^{\mathcal{G}}$, $\operatorname{ran} A = \bigcup_{x \in \mathcal{H}} A(x)$.

1 Motivation: when the gradient is no longer enough

The previous chapters assumed f to be differentiable, and even β-smooth, in order to use gradient descent

$$x^{[k+1]} = x^{[k]} - \gamma \nabla f(x^{[k]}), \qquad \gamma \in \left]0, \tfrac{2}{\beta}\right[.$$

However, many criteria encountered in Chapter 2 (§ 5) are not differentiable:

- the sparsity-promoting penalty $\lambda \|x\|_1$ (Lasso, denoising in a wavelet basis), which is non-differentiable as soon as a component is zero, that is, precisely at the points we are looking for;
- the total variation $\|Dx\|_1$ or $\|Dx\|_{1,2}$;
- constraints, through ι_C, which equals $+\infty$ outside C;
- the Poisson data-fidelity term (non-Lipschitz gradient);
- the hinge loss of SVMs.

An implicit scheme. For $f \in \mathcal{C}^{1,1}_\beta$, the explicit scheme can be replaced by the *implicit* scheme (backward Euler)

$$x^{[k+1]} = x^{[k]} - \gamma \nabla f(x^{[k+1]}). \tag{1}$$

Three questions arise:

1. How can we compute $x^{[k+1]}$, which appears on both sides?
2. Which values of γ guarantee convergence?
3. What if f is not differentiable?

The subdifferential (Section 2) answers the third question by replacing $\nabla f(x^{[k+1]})$ with an element of $\partial f(x^{[k+1]})$; the proximity operator (Section 4) answers the first one: $x^{[k+1]} = \operatorname{prox}_{\gamma f}(x^{[k]})$. Finally, we will see that this scheme converges for *any* $\gamma > 0$ (next chapter).

2 Subdifferential

2.1 Definition

Definition 2.1 – (Moreau subdifferential)

Let $f : \mathcal{H} \to]-\infty, +\infty]$ be a proper function. The *subdifferential* of f is the set-valued operator

$$\partial f : \mathcal{H} \to 2^{\mathcal{H}} : \; x \mapsto \{u \in \mathcal{H} \mid (\forall y \in \mathcal{H}) \quad \langle y - x \,|\, u \rangle + f(x) \leq f(y)\}.$$

An element $u \in \partial f(x)$ is called a *subgradient* of f at x.

Geometric interpretation.

- $u \in \partial f(x)$ if and only if the affine function $y \mapsto f(x) + \langle u \,|\, y - x \rangle$, which coincides with f at x, is a *minorant* of f on the whole of $\mathcal{H}$.
- In dimension 1, the subgradients are the slopes of the lines passing through $(x, f(x))$ and lying below the graph (i.e. $\mathrm{gra} f = \{(x, f(x)) \; : \; x \in \mathcal{H}\}$).
- For a differentiable convex function, we recover the first-order inequality of Chapter 2, $f(y) \geq f(x) + \langle \nabla f(x) \,|\, y - x \rangle$: the gradient is a subgradient.

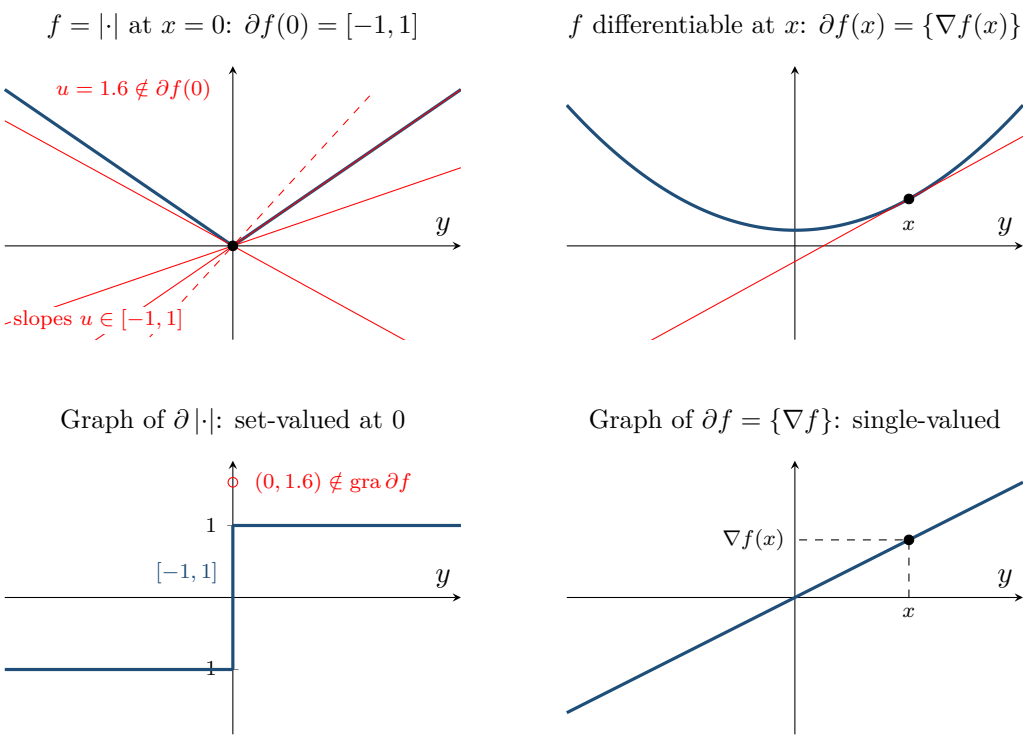


Figure 1: Top: subgradients, i.e., slopes of the affine minorants of f that are exact at x. At a kink there are infinitely many of them; at a point of differentiability (convex function) there is only one, the gradient. Bottom: graphs of the corresponding subdifferentials. For $|\cdot|$, the jump of the derivative is "filled" by a vertical segment at 0: ∂f is set-valued, whereas for a differentiable f its graph is that of ∇f.

Remark (non-convex functions) The definition makes sense for any proper function, but it is tailored to convex functions: a subgradient provides a *global* affine minorant. For $f = -|\cdot|$, no line passing through $(0,0)$ stays below the graph: $\partial f(0) = \varnothing$ (and even $\partial f(x) = \varnothing$ for all x). For non-convex functions, other subdifferentials are used (Fréchet, limiting, Clarke), which are beyond the scope of this chapter.

2.2 Fundamental properties

Theorem 2.2 – (Fermat's rule)

Let $f : \mathcal{H} \to]-\infty, +\infty]$ be proper. Then

$$\widehat{x} \in \operatorname*{Argmin}_{x \in \mathcal{H}} f(x) \quad \Longleftrightarrow \quad 0 \in \partial f(\widehat{x}).$$

Proof.

$$\begin{aligned} 0 \in \partial f(\widehat{x}) &\Longleftrightarrow (\forall y \in \mathcal{H})\ \langle y - \widehat{x} \,|\, 0 \rangle + f(\widehat{x}) \leq f(y) \\ &\Longleftrightarrow (\forall y \in \mathcal{H})\ f(\widehat{x}) \leq f(y). \end{aligned}$$

■

This equivalence is immediate but fundamental: it holds without any convexity or differentiability assumption, and generalizes the condition $\nabla f(\widehat{x}) = 0$ of Chapter 2. The whole difficulty is shifted to the *computation* of ∂f.

Proposition 2.3 – (Differentiable case)

If $f : \mathcal{H} \to]-\infty, +\infty]$ is convex and Gâteaux differentiable at $x \in \operatorname{dom} f$, then $\partial f(x) = \{\nabla f(x)\}$.

Proof. $\nabla f(x) \in \partial f(x)$. For $y \in \mathcal{H}$ and $t \in]0,1]$, convexity gives $f(x+t(y-x)) \leq (1-t)f(x)+tf(y)$, that is, $\frac{f(x+t(y-x))-f(x)}{t} \leq f(y) - f(x)$. Letting $t \to 0^+$: $\langle \nabla f(x) \,|\, y - x \rangle \leq f(y) - f(x)$.

Uniqueness. Let $u \in \partial f(x)$. For $h \in \mathcal{H}$ and $t > 0$, $f(x + th) - f(x) \geq t \langle u \,|\, h \rangle$; dividing by t and letting $t \to 0^+$, $\langle \nabla f(x) - u \,|\, h \rangle \geq 0$ for all h. With $h = u - \nabla f(x)$, we obtain $\|u - \nabla f(x)\|^2 \leq 0$. ■

Proposition 2.4 – (Monotonicity)

Let $f : \mathcal{H} \to]-\infty, +\infty]$ be proper. ∂f is *monotone*:

$$(\forall (x,u), (y,v) \text{ with } u \in \partial f(x),\ v \in \partial f(y)) \qquad \langle u - v \,|\, x - y \rangle \geq 0.$$

Proof. Summing $f(y) \geq f(x)+\langle u \,|\, y - x \rangle$ and $f(x) \geq f(y)+\langle v \,|\, x - y \rangle$, we obtain $0 \geq \langle u - v \,|\, y - x \rangle$. ■

Monotonicity generalizes the fact that the derivative of a convex function of one variable is nondecreasing; for a differentiable f, it reads $\langle \nabla f(x) - \nabla f(y) \,|\, x - y \rangle \geq 0$.

3 Subdifferential calculus

Proposition 3.1 – (Elementary rules)

Let $f : \mathcal{H} \to]-\infty, +\infty]$ be proper, $g : \mathcal{G} \to]-\infty, +\infty]$ proper and $L \in \mathcal{B}(\mathcal{H}, \mathcal{G})$.

(i) $(\forall \lambda \in]0, +\infty[)\ \partial(\lambda f) = \lambda\, \partial f$.

(ii) If $\operatorname{dom} g \cap L(\operatorname{dom} f) \neq \varnothing$, then

$$(\forall x \in \mathcal{H}) \qquad \partial f(x) + L^* \partial g(Lx) \subset \partial(f + g \circ L)(x).$$

Proof. (i) follows directly from the definition. (ii) Let $u \in \partial f(x)$ and $v \in \partial g(Lx)$. For all $y \in \mathcal{H}$, $f(y) \geq f(x) + \langle u \,|\, y - x \rangle$ and $g(Ly) \geq g(Lx) + \langle v \,|\, Ly - Lx \rangle = g(Lx) + \langle L^* v \,|\, y - x \rangle$. Summing, $(f + g \circ L)(y) \geq (f + g \circ L)(x) + \langle u + L^* v \,|\, y - x \rangle$. ■

The reverse inclusion is false in general: a *qualification condition* is needed.

Theorem 3.2 – (Sum and composition)

Let $f \in \Gamma_0(\mathcal{H})$, $g \in \Gamma_0(\mathcal{G})$ and $L \in \mathcal{B}(\mathcal{H}, \mathcal{G})$. If

$$\operatorname{int}(\operatorname{dom} g) \cap L(\operatorname{dom} f) \neq \varnothing \qquad \text{or} \qquad \operatorname{dom} g \cap \operatorname{int}\big(L(\operatorname{dom} f)\big) \neq \varnothing,$$

then

$$\partial(f + g \circ L) = \partial f + L^* \circ \partial g \circ L.$$

Special cases:

- if $f, g \in \Gamma_0(\mathcal{H})$ and $\operatorname{dom} g = \mathcal{H}$ (or $\operatorname{dom} f = \mathcal{H}$), then $\partial(f + g) = \partial f + \partial g$;
- if $g \in \Gamma_0(\mathcal{G})$, $L \in \mathcal{B}(\mathcal{H}, \mathcal{G})$ and $\operatorname{int}(\operatorname{dom} g) \cap \operatorname{ran} L \neq \varnothing$, then $\partial(g \circ L) = L^* \circ \partial g \circ L$.

The proof of the inclusion $\subset$ relies on a separation theorem (Hahn–Banach); see [1, Th. 16.47].

Warning Qualification is not a technical subtlety. On $\mathbb{R}$, let $f(x) = -\sqrt{x}$ if $x \geq 0$ ($+\infty$ otherwise) and $g = \iota_{]-\infty,0]}$, both in $\Gamma_0(\mathbb{R})$. Then $f + g = \iota_{\{0\}}$, hence $\partial(f + g)(0) = \mathbb{R}$. But $\partial f(0) = \varnothing$ (vertical tangent at 0), hence $\partial f(0) + \partial g(0) = \varnothing$. Here $\operatorname{dom} f \cap \operatorname{dom} g = \{0\}$, $\operatorname{int}(\operatorname{dom} g) \cap \operatorname{dom} f = \varnothing$ and $\operatorname{dom} g \cap \operatorname{int}(\operatorname{dom} f) = \varnothing$: none of the conditions is satisfied.

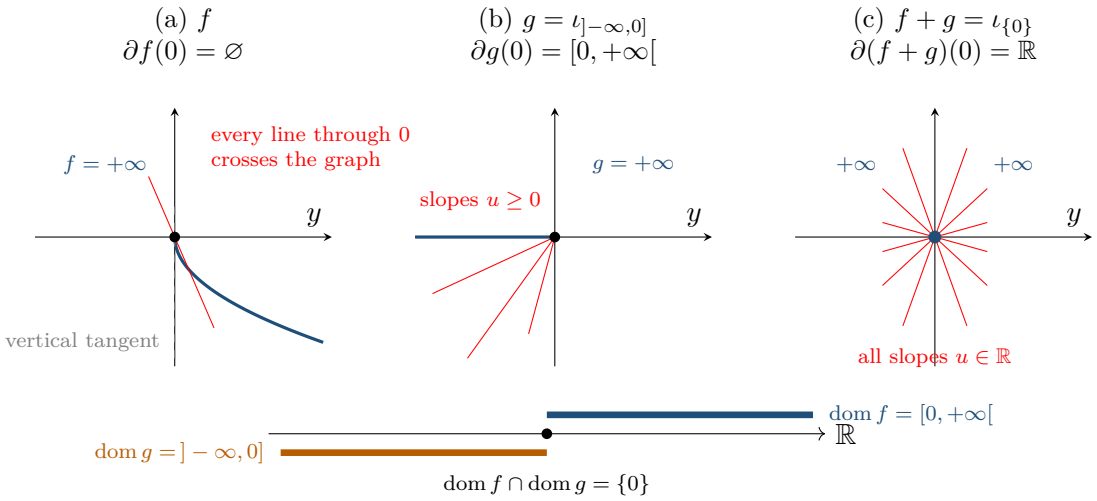


Figure 2: Failure of the sum rule without qualification. (a) f has a vertical tangent at 0: no line passing through $(0,0)$ minorizes f, hence $\partial f(0) = \varnothing$. (b) g admits all nonnegative slopes at 0. (c) $f + g$ is finite only at 0: every line passing through the origin minorizes it, hence $\partial(f+g)(0) = \mathbb{R} \neq \partial f(0) + \partial g(0) = \varnothing$. Bottom: the domains meet at a single point, on their boundaries; the interior of neither domain meets the other domain.

Proposition 3.3 – (Separable sum)

Let $(\mathcal{H}_i)_{i\in I}$ be Hilbert spaces, with $I \subset \mathbb{N}$ finite, and $\mathcal{H} = \bigoplus_{i\in I} \mathcal{H}_i$. For every i, let $f_i : \mathcal{H}_i \to]-\infty, +\infty]$ be proper, and

$$f : \mathcal{H} \to]-\infty, +\infty] : \; x = (x_i)_{i\in I} \mapsto \sum_{i\in I} f_i(x_i).$$

Then

$$(\forall x = (x_i)_{i\in I} \in \mathcal{H}) \qquad \partial f(x) = \bigtimes_{i\in I} \partial f_i(x_i).$$

Proof.
Inclusion $\supset$. If $t = (t_i)_{i\in I} \in \bigtimes_i \partial f_i(x_i)$, then for every i and every $y_i \in \mathcal{H}_i$, $f_i(y_i) \geq f_i(x_i) + \langle t_i \,|\, y_i - x_i\rangle$; summing over i, $f(y) \geq f(x) + \langle t \,|\, y - x\rangle$ for every $y = (y_i)_i$.

Inclusion $\subset$. Let $t \in \partial f(x)$, that is, $\sum_i f_i(y_i) \geq \sum_i f_i(x_i) + \sum_i \langle t_i \,|\, y_i - x_i\rangle$ for every y. Fix $j \in I$ and take $y_i = x_i$ for $i \neq j$ (we may assume $x \in \operatorname{dom} f$; otherwise $\partial f(x) = \varnothing$ and, since one of the x_i lies outside $\operatorname{dom} f_i$, the product is empty as well). The terms $i \neq j$, which are finite, cancel out: $f_j(y_j) \geq f_j(x_j) + \langle t_j \,|\, y_j - x_j\rangle$ for every y_j, that is, $t_j \in \partial f_j(x_j)$. ∎

3.1 Examples

Example 3.1 (Absolute value)**.** For $f = |\cdot|$ on $\mathbb{R}$:

$$\partial\,|\cdot|\,(\zeta) = \begin{cases} \{-1\} & \text{if } \zeta < 0, \\ [-1,1] & \text{if } \zeta = 0, \\ \{1\} & \text{if } \zeta > 0. \end{cases}$$

For $\zeta \neq 0$, f is differentiable (Proposition 2.3). At 0, $u \in \partial f(0) \iff (\forall y)\; uy \leq |y| \iff |u| \leq 1$.

Example 3.2 (ℓ_1 norm)**.** For $f = \|\cdot\|_1 : \mathbb{R}^N \to \mathbb{R} : x \mapsto \sum_{i=1}^N |x_i|$, the separable sum rule

(Proposition 3.3 above) gives

$$\partial \|\cdot\|_1 (x) = \partial |\cdot| (x_1) \times \cdots \times \partial |\cdot| (x_N) = \{u \in \mathbb{R}^N \mid u_i = \operatorname{sign}(x_i) \text{ if } x_i \neq 0,\ u_i \in [-1,1] \text{ if } x_i = 0\}.$$

Example 3.3 (Euclidean norm)**.** For $f = \|\cdot\|$ on $\mathcal{H}$:

$$\partial \|\cdot\| (x) = \begin{cases} \left\{\dfrac{x}{\|x\|}\right\} & \text{if } x \neq 0, \\ \{u \in \mathcal{H} \mid \|u\| \leq 1\} & \text{if } x = 0. \end{cases}$$

At $x \neq 0$, $\|\cdot\|$ is differentiable. At 0, $u \in \partial f(0) \iff (\forall y)\ \langle u \,|\, y \rangle \leq \|y\|$, which is equivalent to $\|u\| \leq 1$ (Cauchy–Schwarz in one direction, $y = u$ in the other).

Example 3.4 (Indicator function and normal cone)**.** Let $C \subset \mathcal{H}$ be nonempty. For every $x \in \mathcal{H}$, $\partial \iota_C(x)$ is the *normal cone* to C at x:

$$\partial \iota_C(x) = N_C(x) = \begin{cases} \{u \in \mathcal{H} \mid (\forall y \in C)\ \langle u \,|\, y - x \rangle \leq 0\} & \text{if } x \in C, \\ \varnothing & \text{otherwise.} \end{cases}$$

Indeed, if $x \in C$, the inequality $\langle y - x \,|\, u \rangle + \iota_C(x) \leq \iota_C(y)$ is trivial for $y \notin C$ and reads $\langle y - x \,|\, u \rangle \leq 0$ for $y \in C$. If $x \notin C$, $\iota_C(x) = +\infty$ and no u is suitable. At an interior point, $N_C(x) = \{0\}$; at a boundary point, $N_C(x)$ contains the outward normals.

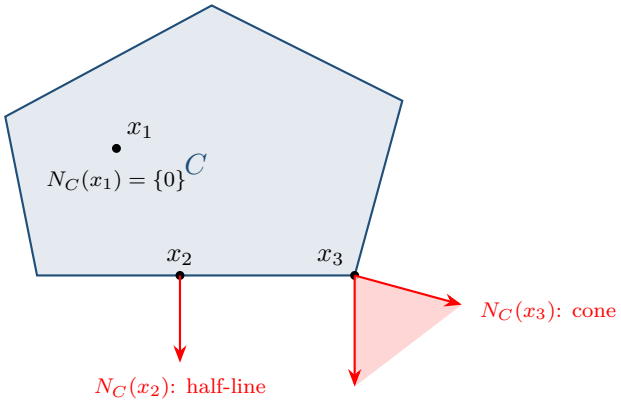


Figure 3: Normal cone to a convex polygon: reduced to $\{0\}$ at an interior point, a half-line (outward normal) on an edge, the cone generated by the two outward normals at a vertex.

Constrained minimization. For f convex and differentiable and C closed and convex, Fermat's rule applied to $f + \iota_C$ (together with the sum rule, Theorem 3.2) gives

$$\widehat{x} \in \operatorname*{Argmin}_{x \in C} f(x) \iff -\nabla f(\widehat{x}) \in N_C(\widehat{x}) \iff \widehat{x} \in C \text{ and } (\forall y \in C)\ \langle \nabla f(\widehat{x}) \,|\, y - \widehat{x} \rangle \geq 0.$$

This is a *variational inequality*: at $\widehat{x}$, no feasible direction is a descent direction. If $\widehat{x}$ lies in the interior of C, we recover $\nabla f(\widehat{x}) = 0$.

Example 3.5 (Huber function)**.** The Huber function with parameter $\mu > 0$ is a smooth approximation of $|\cdot|$:

$$h_\mu(\zeta) = \begin{cases} |\zeta| - \dfrac{\mu}{2} & \text{if } |\zeta| > \mu, \\ \dfrac{\zeta^2}{2\mu} & \text{if } |\zeta| \leq \mu, \end{cases}$$

$$\partial h_\mu(\zeta) = \{h'_\mu(\zeta)\} = \begin{cases} \left\{\frac{\zeta}{|\zeta|}\right\} & \text{if } |\zeta| > \mu, \\ \left\{\frac{\zeta}{\mu}\right\} & \text{if } |\zeta| \leq \mu. \end{cases}$$

h_μ is convex and differentiable, with a $\frac{1}{\mu}$-Lipschitz derivative: it is $\frac{1}{\mu}$-smooth. We will see that h_μ is exactly the Moreau envelope of $|\cdot|$.

3.2 Application: optimality conditions for the Lasso

Consider, for $\Phi \in \mathbb{R}^{M\times N}$, $z \in \mathbb{R}^M$ and $\lambda > 0$,

$$\min_{x\in\mathbb{R}^N} \; F(x) = \tfrac{1}{2}\left\|\Phi x - z\right\|^2 + \lambda \left\|x\right\|_1 . \tag{2}$$

The domain of the first term is $\mathbb{R}^N$: the sum rule applies without restriction. With Example 3.2, Fermat's rule reads

$$0 \in \Phi^\top(\Phi\widehat{x} - z) + \lambda\,\partial\left\|\cdot\right\|_1(\widehat{x}) \iff \begin{cases} \left(\Phi^\top(z - \Phi\widehat{x})\right)_i = \lambda\ \mathrm{sign}(\widehat{x}_i) & \text{if } \widehat{x}_i \neq 0, \\ \left|\left(\Phi^\top(z - \Phi\widehat{x})\right)_i\right| \leq \lambda & \text{if } \widehat{x}_i = 0. \end{cases}$$

The correlated residual $\Phi^\top(z - \Phi\widehat{x})$ is "saturated" at $\pm\lambda$ on the support of $\widehat{x}$ and stays within $[-\lambda, \lambda]$ elsewhere. In particular, $\widehat{x} = 0$ is a solution if and only if $\left\|\Phi^\top z\right\|_\infty \leq \lambda$ (Exercise **??**): beyond this threshold, the solution is identically zero. This is the sparsity induced by the non-differentiability of $\left\|\cdot\right\|_1$ at 0.

Example 3.6 (Scalar case: soft thresholding)**.** For $N = M = 1$, $\Phi = 1$: $\min_{x\in\mathbb{R}} \frac{1}{2}(x - z)^2 + \lambda\left|x\right|$. The condition $0 \in \widehat{x} - z + \lambda\,\partial\left|\cdot\right|(\widehat{x})$ gives

- if $\widehat{x} > 0$: $\widehat{x} = z - \lambda$, possible only if $z > \lambda$;
- if $\widehat{x} < 0$: $\widehat{x} = z + \lambda$, possible only if $z < -\lambda$;
- if $\widehat{x} = 0$: $z \in [-\lambda, \lambda]$.

Hence $\widehat{x} = \mathrm{sign}(z)\max\{|z| - \lambda, 0\}$: this is *soft thresholding*, which we will meet again as the proximity operator of $\lambda\left|\cdot\right|$.

4 Proximity operator

4.1 Definition

Definition 4.1 – (Moreau envelope and proximity operator)

Let $f \in \Gamma_0(\mathcal{H})$ and $\gamma > 0$.

- The *Moreau envelope* of f with parameter γ is

$$^\gamma f : \mathcal{H} \to \mathbb{R} : \; x \mapsto \inf_{y\in\mathcal{H}} \; f(y) + \frac{1}{2\gamma}\left\|y - x\right\|^2 .$$

- The *proximity operator* of f is

$$\mathrm{prox}_f : \mathcal{H} \to \mathcal{H} : \; x \mapsto \arg\min_{y\in\mathcal{H}} \; f(y) + \frac{1}{2}\left\|y - x\right\|^2 .$$

Thus $\mathrm{prox}_{\gamma f}(x)$ is the unique point where the infimum defining $^\gamma f(x)$ is attained: $^\gamma f(x) = f\left(\mathrm{prox}_{\gamma f}\, x\right) + \frac{1}{2\gamma}\left\|x - \mathrm{prox}_{\gamma f}\, x\right\|^2$.

Proposition 4.2 – (Well-posedness)

If $f \in \Gamma_0(\mathcal{H})$, then for every $x \in \mathcal{H}$, the function $\psi(\cdot\,; x) = f + \frac{1}{2}\|\cdot - x\|^2$ has a unique minimizer; prox_f is therefore a (single-valued) mapping from $\mathcal{H}$ to $\mathcal{H}$, and ${}^{\gamma}f$ is real-valued, with ${}^{\gamma}f \leq f$.

Proof. $\psi(\cdot\,; x)$ is proper (since f is), l.s.c. (sum of l.s.c. functions) and 1-strongly convex (a convex function plus $\frac{1}{2}\|\cdot - x\|^2$). A function in Γ_0 has an affine minorant [1, Th. 9.20]; added to a quadratic term, it makes $\psi(\cdot\,; x)$ coercive. The existence and uniqueness theorem of Chapter 2 applies. Finally, taking $y = x$ in the infimum, ${}^{\gamma}f(x) \leq f(x)$. ■

The proximity operator looks for a point that achieves a trade-off between *minimizing* f and *staying close to* x. The parameter γ controls this trade-off: $\operatorname{prox}_{\gamma f}\, x \to x$ when $\gamma \to 0$, and $\operatorname{prox}_{\gamma f}\, x$ tends to a minimizer of f when $\gamma \to +\infty$.

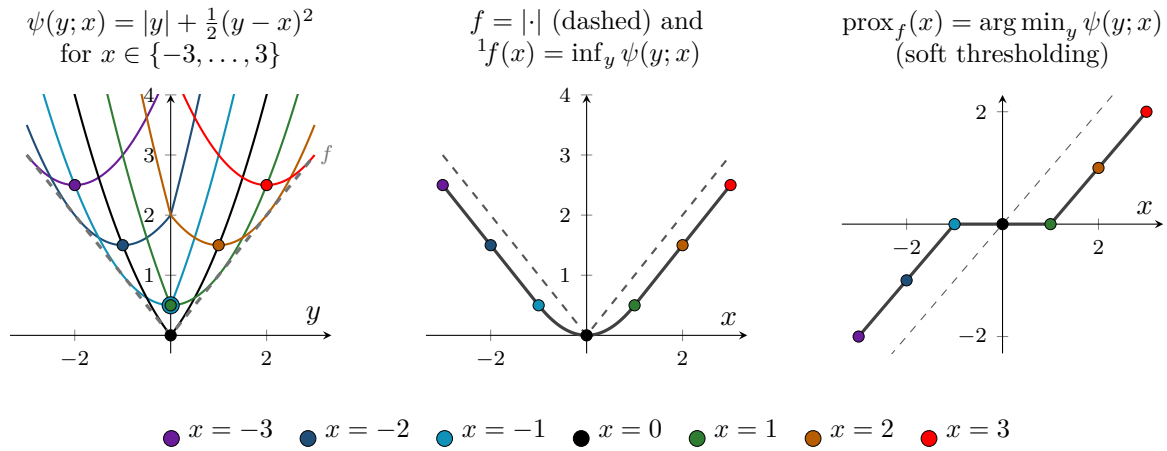


Figure 4: For $f = |\cdot|$ and $\gamma = 1$. Left: the function $\psi(\cdot\,; x)$ minimized for seven values of x, each with its own color (legend below the figure), and $f = |\cdot|$ as a dashed gray line; the point of the same color is its minimum, located at $(\operatorname{prox}_f x,\ {}^{1}f(x))$ (for $x = \pm 1$, the two minima coincide at $(0, 0.5)$). Center: the *height* of this minimum, plotted against x, describes the Moreau envelope, drawn on the same vertical scale as the left panel; it is the Huber function h_1 (Example 3.5), smooth and below f. Right: its *position*, plotted against x, describes prox_f, the soft thresholding operator, which vanishes on $[-1, 1]$.

4.2 Characterization and interpretation

Theorem 4.3 – (Characterization of the prox)

Let $f \in \Gamma_0(\mathcal{H})$ and $\gamma > 0$. For all $x, p \in \mathcal{H}$,

$$p = \operatorname{prox}_{\gamma f}(x) \quad \Longleftrightarrow \quad x - p \in \gamma\, \partial f(p).$$

In other words, $\operatorname{prox}_{\gamma f} = (\mathrm{Id} + \gamma \partial f)^{-1}$.
The proximity operator is the *resolvent* of $\gamma \partial f$, denoted by $\mathrm{J}_{\gamma \partial f}$.

Proof. By Fermat's rule, then the sum rule (the domain of the quadratic term is $\mathcal{H}$):

$$\begin{aligned} p = \arg\min_y \ \gamma f(y) + \tfrac{1}{2}\|y - x\|^2 &\iff 0 \in \partial(\gamma f + \tfrac{1}{2}\|\cdot - x\|^2)(p) \\ &\iff 0 \in \gamma \partial f(p) + p - x \\ &\iff x \in (\mathrm{Id} + \gamma \partial f)(p). \end{aligned}$$

■

Answer to the question of the implicit scheme. The theorem can also be written

$$x^{[k+1]} = \mathrm{prox}_{\gamma f}(x^{[k]}) \iff x^{[k+1]} = x^{[k]} - \gamma u^{[k]} \quad \text{with} \quad u^{[k]} \in \partial f(x^{[k+1]}).$$

A proximal step is thus an *implicit* subgradient step: a subgradient at the arrival point is used. For a differentiable f, this is exactly the scheme (1).

Proposition 4.4 – (Fixed points)

Let $f \in \Gamma_0(\mathcal{H})$ and $\gamma > 0$. Then

$$\mathrm{Fix}\,\mathrm{prox}_{\gamma f} = \{x \in \mathcal{H} \mid \mathrm{prox}_{\gamma f}\, x = x\} = \mathrm{Argmin}\, f.$$

Proof.

$$\begin{aligned} x = \mathrm{prox}_{\gamma f}\, x &\iff x - x \in \gamma \partial f(x) \\ &\iff 0 \in \partial f(x) \\ &\iff x \in \mathrm{Argmin}\, f \end{aligned}$$

(Theorem 4.3, then Fermat's rule). ■

Proposition 4.5 – (Non-expansiveness / 1-Lipschitz continuity)

Let $f \in \Gamma_0(\mathcal{H})$. For all $x, y \in \mathcal{H}$, with $p = \mathrm{prox}_f\, x$ and $q = \mathrm{prox}_f\, y$,

$$\|p - q\|^2 \leq \langle p - q \,|\, x - y \rangle, \qquad \text{and hence} \qquad \|p - q\| \leq \|x - y\|.$$

Proof. By Theorem 4.3, $x - p \in \partial f(p)$ and $y - q \in \partial f(q)$. By monotonicity of ∂f (Proposition 2.4), $\langle (x - p) - (y - q) \,|\, p - q \rangle \geq 0$, that is, $\langle x - y \,|\, p - q \rangle \geq \|p - q\|^2$. Cauchy–Schwarz then gives $\|p - q\|^2 \leq \|p - q\|\,\|x - y\|$. ■

Non-expansiveness (1-Lipschitz continuity) is what makes proximal iterations stable: an error on x is never amplified.

Proposition 4.6 – (Regularity of the Moreau envelope)

Let $f \in \Gamma_0(\mathcal{H})$ and $\gamma > 0$. Then ${}^{\gamma}f$ is convex and differentiable, with a $\frac{1}{\gamma}$-Lipschitz gradient:

$$(\forall x \in \mathcal{H}) \qquad \nabla\, {}^{\gamma}f(x) = \frac{1}{\gamma}(x - \mathrm{prox}_{\gamma f}\, x).$$

Moreover, $\inf {}^{\gamma}f = \inf f$ and $\mathrm{Argmin}\, {}^{\gamma}f = \mathrm{Argmin}\, f$.

See [1, Prop. 12.30]. The Moreau envelope is thus a *smooth regularization* of f that changes neither the minimum value nor the minimizers. For $f = |\cdot|$, it is the Huber function with parameter γ.

4.3 Examples of proximity operators

Projection. Let $C \subset \mathcal{H}$ be nonempty, closed and convex. Then $\iota_C \in \Gamma_0(\mathcal{H})$ and

$$(\forall x \in \mathcal{H}) \qquad \operatorname{prox}_{\iota_C}(x) = \operatorname*{arg\,min}_{y \in C} \tfrac{1}{2}\|y - x\|^2 = P_C(x),$$

the orthogonal projection onto C (and $\operatorname{prox}_{\gamma\iota_C} = P_C$ for every $\gamma > 0$). Theorem 4.3 gives back the classical characterization: $p = P_C x \iff p \in C$ and $(\forall y \in C)\ \langle x - p \,|\, y - p\rangle \leq 0$. Some explicit projections on $\mathbb{R}^N$:

Set C	**Projection** $P_C(x)$
nonnegative orthant $[0, +\infty[^N$	$\big(\max\{x_i, 0\}\big)_{1\leq i\leq N}$
hypercube $[a, b]^N$ (dynamic range constraint)	$\big(\min\{\max\{x_i, a\}, b\}\big)_{1\leq i\leq N}$
ball $\{y \mid \|y\| \leq r\}$	$\min\{1, \frac{r}{\|x\|}\}\, x$
hyperplane $\{y \mid \langle a \,|\, y\rangle = b\}$, $a \neq 0$	$x + \frac{b - \langle a \,|\, x\rangle}{\|a\|^2}\, a$

ℓ_1 norm: soft thresholding. For $\lambda > 0$, by Example 3.6,

$$\operatorname{prox}_{\lambda|\cdot|}(\xi) = \operatorname{sign}(\xi)\max\{|\xi| - \lambda, 0\} = \begin{cases} \xi - \lambda & \text{if } \xi > \lambda, \\ 0 & \text{if } |\xi| \leq \lambda, \\ \xi + \lambda & \text{if } \xi < -\lambda, \end{cases}$$

and, by separability (Proposition 4.7), $\operatorname{prox}_{\lambda\|\cdot\|_1}(x) = \big(\operatorname{prox}_{\lambda|\cdot|}(x_i)\big)_{1\leq i\leq N}$.

Euclidean norm: block thresholding. For $\lambda > 0$ (Exercise **??**),

$$\operatorname{prox}_{\lambda\|\cdot\|}(x) = \max\Big\{1 - \frac{\lambda}{\|x\|}, 0\Big\}\, x \qquad \text{(with the convention } \operatorname{prox}_{\lambda\|\cdot\|}(0) = 0\text{)}.$$

Applied block by block, it gives the prox of the mixed norm $\|x\|_{1,2} = \sum_g \|x_g\|$ (group-structured sparsity, isotropic TV on the gradients).

Powers. Let $\chi > 0$, $q \in [1, +\infty[$ and $\varphi : \mathbb{R} \to \mathbb{R} : \xi \mapsto \chi\,|\xi|^q$. For some values of q, $\operatorname{prox}_\varphi$ has a closed form [4]:

$$\operatorname{prox}_\varphi(\xi) = \begin{cases} \operatorname{sign}(\xi)\max\{|\xi| - \chi, 0\} & q = 1, \\ \xi + \dfrac{4\chi}{3\cdot 2^{1/3}}\Big((\epsilon - \xi)^{1/3} - (\epsilon + \xi)^{1/3}\Big), \quad \epsilon = \sqrt{\xi^2 + 256\chi^3/729} & q = \frac{4}{3}, \\ \xi + \dfrac{9\chi^2 \operatorname{sign}(\xi)}{8}\Big(1 - \sqrt{1 + \dfrac{16\,|\xi|}{9\chi^2}}\Big) & q = \frac{3}{2}, \\ \dfrac{\xi}{1 + 2\chi} & q = 2, \\ \operatorname{sign}(\xi)\dfrac{\sqrt{1 + 12\chi\,|\xi|} - 1}{6\chi} & q = 3, \\ \Big(\dfrac{\epsilon + \xi}{8\chi}\Big)^{1/3} - \Big(\dfrac{\epsilon - \xi}{8\chi}\Big)^{1/3}, \quad \epsilon = \sqrt{\xi^2 + 1/(27\chi)} & q = 4. \end{cases}$$

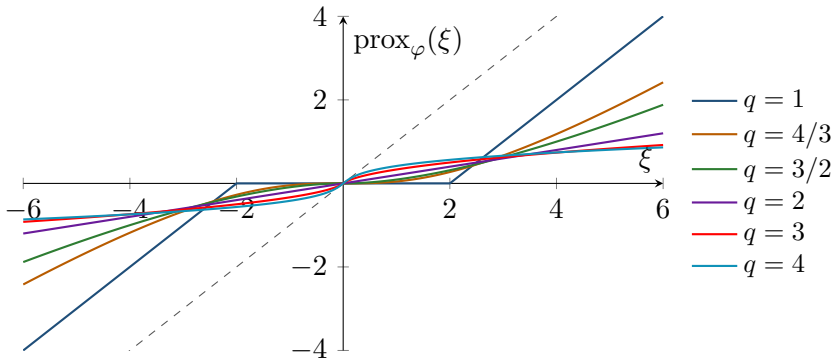


Figure 5: $\mathrm{prox}_{\chi|\cdot|^q}$ for $\chi = 2$. Only $q = 1$ sets a whole interval to zero (and thus produces exact zeros); for $q > 1$, the prox is a smooth shrinkage towards 0, all the stronger near the origin as q gets closer to 1.

Quadratic function. Let $L \in \mathcal{B}(\mathcal{H}, \mathcal{G})$, $\gamma > 0$ and $z \in \mathcal{G}$. Then

$$f = \frac{\gamma}{2}\|L \cdot - z\|^2 \quad \Longrightarrow \quad \mathrm{prox}_f = (\mathrm{Id} + \gamma L^* L)^{-1}(\cdot + \gamma L^* z).$$

Indeed, f is differentiable with gradient $\gamma L^*(L\cdot - z)$, and $x - p = \gamma L^*(Lp - z) \iff (\mathrm{Id} + \gamma L^* L)p = x + \gamma L^* z$; the operator $\mathrm{Id} + \gamma L^* L \succeq \mathrm{Id}$ is invertible. For $L = \mathrm{Id}$, $z = 0$: $\mathrm{prox}_{\frac{\gamma}{2}\|\cdot\|^2} = \frac{1}{1+\gamma}\mathrm{Id}$. If L is a circular convolution, the inversion is performed by FFT (Chapter 2, exercise on Tikhonov).

4.4 Calculus rules

Property	$g(x)$	$\mathrm{prox}_g(x)$
Translation	$f(x - z)$, $z \in \mathcal{H}$	$z + \mathrm{prox}_f(x - z)$
Quadratic perturbation	$f(x) + \frac{\alpha}{2}\|x\|^2 + \langle z \,\vert\, x\rangle + c$	$\mathrm{prox}_{\frac{f}{\alpha+1}}\left(\frac{x-z}{\alpha+1}\right)$
	$z \in \mathcal{H}$, $\alpha \geq 0$, $c \in \mathbb{R}$	
Scaling	$f(\rho x)$, $\rho \in \mathbb{R} \setminus \{0\}$	$\frac{1}{\rho}\mathrm{prox}_{\rho^2 f}(\rho x)$
Reflection	$f(-x)$	$-\mathrm{prox}_f(-x)$
Moreau envelope	${}^\gamma f(x)$, $\gamma > 0$	$\frac{1}{1+\gamma}\left(\gamma x + \mathrm{prox}_{(1+\gamma)f}(x)\right)$

Table 1: Calculus rules for $f \in \Gamma_0(\mathcal{H})$ [1, Prop. 24.8]. Each of them is proved with Theorem 4.3 (Exercise 1).

Proposition 4.7 – (Separable sum)

For $i \in \{1, \ldots, n\}$, let $f_i \in \Gamma_0(\mathcal{H}_i)$, and $f(x_1, \ldots, x_n) = \sum_{i=1}^n f_i(x_i)$ on $\mathcal{H}_1 \times \cdots \times \mathcal{H}_n$. Then

$$(\forall x = (x_1, \ldots, x_n)) \qquad \mathrm{prox}_f(x) = \left(\mathrm{prox}_{f_i}(x_i)\right)_{1 \leq i \leq n}.$$

Proof. The criterion $\sum_i (f_i(y_i) + \frac{1}{2}\|y_i - x_i\|^2)$ is a sum of terms that each depend on a single block y_i: it is minimized block by block. ∎

Proposition 4.8 – (Decomposition in an orthonormal basis)

Let $\mathcal{H}$ be separable and $(b_i)_{i\in I}$ an orthonormal basis of $\mathcal{H}$. For every i, let $\varphi_i \in \Gamma_0(\mathbb{R})$ with $\varphi_i \geq 0$ (an assumption that is unnecessary in finite dimension). If $f(x) = \sum_{i\in I} \varphi_i(\langle x \,|\, b_i\rangle)$, then

$$\operatorname{prox}_f(x) = \sum_{i\in I} \operatorname{prox}_{\varphi_i}(\langle x \,|\, b_i\rangle)\, b_i.$$

Proof in finite dimension. Let W be the orthogonal matrix whose rows are the $b_i^\top$, and $c = Wy$. Since $\|y - x\| = \|Wy - Wx\|$, minimizing $\sum_i \varphi_i((Wy)_i) + \frac{1}{2}\|Wy - Wx\|^2$ amounts to minimizing with respect to c the separable function $\sum_i(\varphi_i(c_i) + \frac{1}{2}(c_i - (Wx)_i)^2)$, hence $c_i = \operatorname{prox}_{\varphi_i}((Wx)_i)$ and $y = W^\top c$. ■

Example. In the canonical basis of $\mathbb{R}^N$, with $\varphi_i = \lambda|\cdot|$, we recover the componentwise soft thresholding of $\operatorname{prox}_{\lambda\|\cdot\|_1}$.

Proposition 4.9 – (Composition with a linear operator)

Let $f \in \Gamma_0(\mathcal{H})$ and $L \in \mathcal{B}(\mathcal{G}, \mathcal{H})$ such that $LL^* = \mu\mathrm{Id}$ with $\mu \in]0, +\infty[$. Then

$$\operatorname{prox}_{f\circ L} = \mathrm{Id} - \mu^{-1}L^* \circ (\mathrm{Id} - \operatorname{prox}_{\mu f}) \circ L.$$

In particular, if L is unitary ($L^*L = LL^* = \mathrm{Id}$), $\operatorname{prox}_{f\circ L} = L^* \circ \operatorname{prox}_f \circ L$.

Proof. Since $LL^* = \mu\mathrm{Id}$, L is surjective, $\operatorname{ran} L = \mathcal{H}$, and $V = \operatorname{ran} L^* = (\ker L)^\perp$ is closed; the orthogonal projection onto V is $P_V = L^*(LL^*)^{-1}L = \mu^{-1}L^*L$. By Theorem 3.2 ($\operatorname{ran} L = \mathcal{H}$), $\partial(f \circ L) = L^* \circ \partial f \circ L$. Let $x \in \mathcal{G}$ and $p = \operatorname{prox}_{f\circ L} x$:

$$x - p \in L^*\partial f(Lp), \quad \text{that is,} \quad x - p = L^*v,\ v \in \partial f(Lp).$$

Component on $V^\perp$. $x - p \in V$, hence $P_{V^\perp}p = P_{V^\perp}x = x - \mu^{-1}L^*Lx$. *Component on V.* Applying L: $Lx - Lp = LL^*v = \mu v \in \mu\,\partial f(Lp)$, that is, $Lp = \operatorname{prox}_{\mu f}(Lx)$ (Theorem 4.3). Hence $P_V p = \mu^{-1}L^*Lp = \mu^{-1}L^* \operatorname{prox}_{\mu f}(Lx)$. Summing: $p = x - \mu^{-1}L^*(Lx - \operatorname{prox}_{\mu f}(Lx))$. ■

4.5 Examples: denoising as a prox computation

Throughout this section, we observe a noisy signal

$$\mathrm{z} = \overline{\mathrm{x}} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2\mathrm{Id}),$$

and we seek to estimate $\overline{\mathrm{x}} \in]-\infty, +\infty]^N$. Recall that, for a proper, lower semicontinuous and convex $g :]-\infty, +\infty]^N \to]-\infty, +\infty]$, and for $\theta > 0$,

$$\operatorname{prox}_{\theta g}(\mathrm{z}) = \arg\min_{\mathrm{x}} \ \frac{1}{2}\|\mathrm{x} - \mathrm{z}\|_2^2 + \theta\, g(\mathrm{x}).$$

The term $\frac{1}{2}\|\mathrm{x} - \mathrm{z}\|_2^2$ is, up to a constant, the negative Gaussian log-likelihood. The function g encodes the prior on $\overline{\mathrm{x}}$. The variational denoising estimators

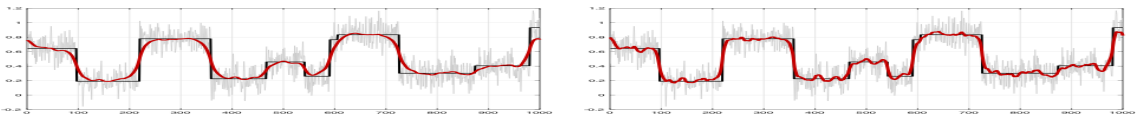
$$\widehat{\mathrm{x}}(\mathrm{z};\theta) = \arg\min_{\mathrm{x}\in]-\infty,+\infty]^N} \ \frac{1}{2}\|\mathrm{x} - \mathrm{z}\|_2^2 + \theta\,\|\mathrm{Lx}\|_\bullet \tag{3}$$

are therefore exactly proximity operators: $\widehat{\mathrm{x}}(\mathrm{z};\theta) = \operatorname{prox}_{\theta\|\mathrm{L}\cdot\|_\bullet}(\mathrm{z})$. The examples below show when this prox has an explicit form, and when an iterative algorithm is required.

Quadratic regularization: linear denoising Let $\mathrm{Lx} = \psi * \mathrm{x}$ with $\psi = \begin{bmatrix} 1 & -1 \end{bmatrix}$ (finite differences) and $g = \|\mathrm{L}\cdot\|_2^2$. The function to be minimized in (3) is differentiable and strongly convex. Its unique minimizer makes the gradient vanish:

$$\widehat{\mathrm{x}} - \mathrm{z} + 2\theta\,\mathrm{L}^*\mathrm{L}\,\widehat{\mathrm{x}} = 0 \quad \Longleftrightarrow \quad \mathrm{prox}_{\theta\|\mathrm{L}\cdot\|_2^2}(\mathrm{z}) = (\mathrm{Id} + 2\theta\,\mathrm{L}^*\mathrm{L})^{-1}\mathrm{z}.$$

The matrix $\mathrm{Id} + 2\theta\,\mathrm{L}^*\mathrm{L}$ is symmetric positive definite, hence invertible. The estimator is *linear* in z: it is a low-pass filter whose cutoff frequency decreases as θ increases (Figure 6). The discontinuities of $\overline{\mathrm{x}}$ are smoothed out, whatever the value of θ.



large θ small θ

Figure 6: Linear denoising: $\psi = [1 \;\; -1]$, $\|\cdot\|_\bullet = \|\cdot\|_2^2$.

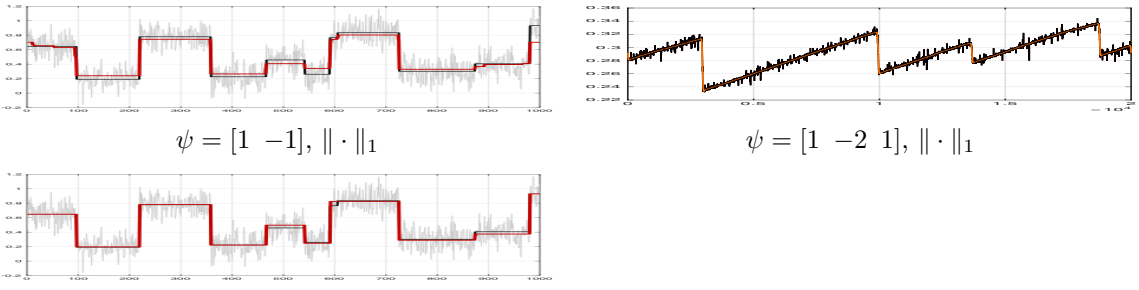
Total variation: piecewise constant denoising Let us now take $\|\cdot\|_\bullet = \|\cdot\|_1$ in (3):

$$\widehat{\mathrm{x}}(\mathrm{z};\theta) = \mathrm{prox}_{\theta\|\mathrm{L}\cdot\|_1}(\mathrm{z}).$$

The choice of the filter ψ sets the prior (Figure 7).

- With $\psi = [1 \;\; -1]$, first-order differences are penalized: this is 1D total variation. The estimator is *piecewise constant* and preserves jumps.
- With $\psi = [1 \;\; -2 \;\; 1]$, second-order differences are penalized: the estimator is *piecewise affine*.
- The non-convex choice $\|\cdot\|_\bullet = \|\cdot\|_0$, with $\psi = [1 \;\; -1]$, also yields a piecewise constant solution, with jumps that are not shrunk. The problem, however, is combinatorial.

Although $\mathrm{prox}_{\theta\|\cdot\|_1}$ is explicit (Example 3.6), $\mathrm{prox}_{\theta\|\mathrm{L}\cdot\|_1}$ is not: L is not an isometry ($\mathrm{LL}^* \neq \mathrm{Id}$). It therefore cannot be reduced to soft thresholding by a change of variable. Computing this prox requires an iterative algorithm, for instance forward–backward on the dual problem.



$\psi = [1 \;\; -1]$, $\|\cdot\|_1$ $\psi = [1 \;\; -2 \;\; 1]$, $\|\cdot\|_1$

$\psi = [1 \;\; -1]$, $\|\cdot\|_0$

Figure 7: Non-linear denoising: piecewise constant and piecewise affine estimators.

Wavelet denoising Let L be an orthonormal wavelet transform. Figure 8 shows one- and two-level decompositions. By Proposition 4.8 and Example 3.6,

$$\widehat{\mathrm{x}} = \arg\min_{\mathrm{x}} \; \frac{1}{2}\|\mathrm{x} - \mathrm{z}\|_2^2 + \theta\|\mathrm{Lx}\|_1 = \mathrm{prox}_{\theta\|\mathrm{L}\cdot\|_1}(\mathrm{z}) = \mathrm{L}^*\,\mathrm{soft}_\theta(\mathrm{Lz}).$$

The estimator is computed in three explicit steps, illustrated in Figure 9:

1. analysis: $\boldsymbol{\zeta} = \mathrm{Lz}$;
2. soft thresholding of the coefficients: $\mathrm{soft}_\theta(\boldsymbol{\zeta})$;

3. synthesis: $\widehat{x} = L^* \mathrm{soft}_\theta(\zeta)$.

The noise, spread over all the coefficients, is removed by thresholding. The image, which is sparse in the wavelet basis, is concentrated on a few large coefficients that survive the threshold.

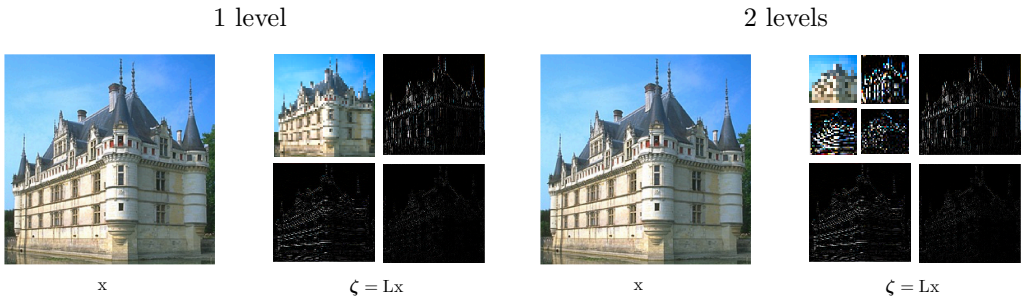


Figure 8: One- and two-level wavelet decompositions.

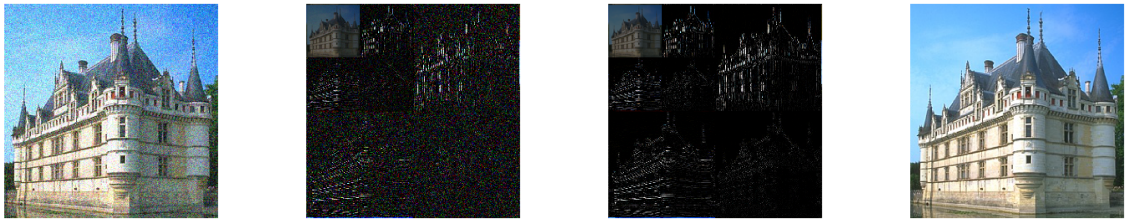


Figure 9: Denoising by soft thresholding of the wavelet coefficients

g	$\mathrm{prox}_{\theta g}(z)$	explicit form?
$\|L \cdot\|_2^2$	$(\mathrm{Id} + 2\theta L^* L)^{-1} z$	yes (linear system)
$\|\cdot\|_1$	$\mathrm{soft}_\theta(z)$	yes
$\|L \cdot\|_1$, L orthonormal	$L^* \mathrm{soft}_\theta(Lz)$	yes
$\|L \cdot\|_1$, L = finite differences	–	no: iterative algorithm

4.6 Resources

Many proximity operators have a closed form: ℓ_1 norm (soft thresholding), $\ell_{1,2}$ norm, powers $|\cdot|^q$, Kullback–Leibler divergence (Poisson noise), projections onto boxes, onto ℓ_1 or $\ell_{1,p}$ balls, overlapping group norms, composition with an operator L such that LL^* is diagonal. Lists and code are available:

- `http://proximity-operator.net`: list of proximity operators with Matlab and Python code (Chierchia, Chouzenoux, Combettes, Pesquet);
- the table by A. Beck in [2];
- SPAMS (Mairal, Bach, Ponce, Sapiro, Jenatton, Obozinski);
- the fast implementations by L. Condat (`https://lcondat.github.io/software.html`);
- `https://github.com/cvxgrp/proximal` (Parikh, Chu, Boyd) and the surveys [6, 11].

```
1 import numpy as np
2
```

```python
def prox_l1(x, lam):
    """Soft thresholding: prox of lam*||.||_1."""
    return np.sign(x) * np.maximum(np.abs(x) - lam, 0)

def prox_l2(x, lam):
    """Block thresholding: prox of lam*||.||_2."""
    nx = np.linalg.norm(x)
    return np.zeros_like(x) if nx <= lam else (1 - lam / nx) * x

def proj_box(x, a, b):
    """Projection onto [a,b]^N: prox of the indicator function."""
    return np.clip(x, a, b)

def prox_quad(x, L, z, gamma):
    """prox of (gamma/2)||L . - z||^2 = (Id + gamma L^T L)^{-1}(x + gamma L^T z).
    """
    return np.linalg.solve(np.eye(L.shape[1]) + gamma * L.T @ L, x + gamma * L.T @
    z)

# Numerical check: compare with a direct minimization
from scipy.optimize import minimize
rng = np.random.default_rng(0)
x, lam = rng.standard_normal(5), 0.7
res = minimize(lambda y: lam * np.linalg.norm(y) + 0.5 * np.sum((y - x)**2), x)
print(np.allclose(res.x, prox_l2(x, lam), atol=1e-5))   # True
```

5 Towards proximal algorithms

This section closes the loop opened in Section 1; the algorithms themselves will be studied in the next chapter.

5.1 Proximal point algorithm

Algorithm 5.1 – (Proximal point)

Given $f \in \Gamma_0(\mathcal{H})$, $x^{[0]} \in \mathcal{H}$ and $\gamma > 0$. For $k = 0, 1, \ldots$: $\quad x^{[k+1]} = \operatorname{prox}_{\gamma f}(x^{[k]})$.

Theorem 5.2 – (Convergence of the proximal point algorithm)

If $\operatorname{Argmin} f \neq \varnothing$, the sequence $(x^{[k]})_{k\in\mathbb{N}}$ converges (weakly; strongly in finite dimension) to a minimizer of f, for *any* $\gamma > 0$, and

$$f(x^{[k]}) - \inf f \leq \frac{\left\|x^{[0]} - \widehat{x}\right\|^2}{2\gamma k} \qquad \text{for every } \widehat{x} \in \operatorname{Argmin} f.$$

See [9, 14] for convergence and [8] for the rate. One way to understand it: by Proposition 4.6,

$$x^{[k+1]} = \operatorname{prox}_{\gamma f}(x^{[k]}) = x^{[k]} - \gamma\, \nabla\, {}^{\gamma}f(x^{[k]}) :$$

the proximal point algorithm is a *gradient descent* on the Moreau envelope ${}^{\gamma}f$, which is $\frac{1}{\gamma}$-smooth and has the same minimizers as f, with step size $\gamma = 1/\beta$. The results of Chapter 3 therefore apply to ${}^{\gamma}f$. No regularity assumption is made on f, and no bound on γ is needed, which answers

the three questions of Section 1. The price to pay: the prox of the whole function f must be computable, which is rarely the case for a realistic criterion (data fidelity plus penalty).

5.2 Proximal gradient (forward–backward)

Consider the composite problem

$$\min_{x\in\mathcal{H}} f(x)+g(x), \qquad f\in\Gamma_0(\mathcal{H})\ \beta\text{-smooth}, \quad g\in\Gamma_0(\mathcal{H})\text{ with a computable prox}, \tag{4}$$

which covers the Lasso, wavelet denoising, constraints ($g=\iota_C$) and sparse logistic regression. We combine an explicit gradient step (*forward*) on f and an implicit proximal step (*backward*) on g:

$$x^{[k+1]} = \operatorname{prox}_{\gamma g}(x^{[k]} - \gamma\nabla f(x^{[k]})). \tag{5}$$

Proposition 5.3 – (Fixed points of forward–backward)

For every $\gamma>0$: $x=\operatorname{prox}_{\gamma g}(x-\gamma\nabla f(x)) \iff x\in\operatorname{Argmin}(f+g)$.

Proof. By Theorem 4.3, $x=\operatorname{prox}_{\gamma g}(x-\gamma\nabla f(x)) \iff -\gamma\nabla f(x)\in\gamma\partial g(x) \iff 0\in\nabla f(x)+\partial g(x)$. Since $\operatorname{dom} f=\mathcal{H}$, Theorem 3.2 gives $\nabla f(x)+\partial g(x)=\partial(f+g)(x)$, and we conclude with Fermat's rule. ■

Remarks:

- For $\gamma\in\left]0,\frac{2}{\beta}\right[$, the sequence (5) converges to a minimizer of $f+g$ [5], with $(f+g)(x^{[k]})-\inf(f+g)=O(1/k)$ for $\gamma\le 1/\beta$.
- For $g=0$, we recover gradient descent.
- For $f=0$, the proximal point algorithm.
- For $g=\iota_C$, the *projected gradient*.
- For $g=\lambda\|\cdot\|_1$ (i.e., the Lasso), it is the ISTA algorithm [7]:

$$x^{[k+1]} = \mathcal{S}_{\gamma\lambda}(x^{[k]} - \gamma\Phi^\top(\Phi x^{[k]}-z)), \qquad \mathcal{S}_{\gamma\lambda}=\operatorname{prox}_{\gamma\lambda\|\cdot\|_1}.$$

5.3 Why not simply use a subgradient?

The *subgradient method*

$$x^{[k+1]} = x^{[k]} - \gamma_k u^{[k]} \qquad \text{with} \qquad u^{[k]}\in\partial f(x^{[k]}), \tag{6}$$

applies to any $f\in\Gamma_0(\mathbb{R}^N)$ with full domain: it is enough to be able to compute *one* subgradient at each point. It suffers, however, from three shortcomings.

1. $-u^{[k]}$ is not necessarily a descent direction. For a differentiable f, $-\nabla f(x)$ is a descent direction as soon as $\nabla f(x)\neq 0$. This no longer holds for an arbitrary subgradient, even far from a minimizer. Let $f(x)=|x_1|+2\,|x_2|$ on $\mathbb{R}^2$ and $x=(1,0)$: then $\partial f(x)=\{1\}\times[-2,2]$. For a subgradient $u=(1,s)$, $s\in[-2,2]$, and a step size $\gamma\in\,]0,1]$,

$$f(x-\gamma u) = |1-\gamma|+2\,|-\gamma s| = 1+\gamma(2\,|s|-1). \tag{7}$$

Hence $-u$ is a descent direction if and only if $|s|<\frac{1}{2}$. For $|s|=\frac{1}{2}$, f remains constant along the direction, and for $|s|>\frac{1}{2}$ it *increases* whatever the step size: with $u=(1,2)$, $f(x-\gamma u)=$

$1+3\gamma > f(x)$. Only a quarter of the segment $\partial f(x)$ thus yields a descent direction, and nothing in the method guarantees that $u^{[k]}$ is chosen there. The minimal-norm subgradient, here $u = (1, 0)$, is always a descent direction when $0 \notin \partial f(x)$, but computing it generally requires solving an auxiliary problem. One can therefore neither guarantee $f(x^{[k+1]}) \leq f(x^{[k]})$ nor use an Armijo-type line search.

What does decrease, however, is the *distance to the set of minimizers*. For $x^\star \in \operatorname{Argmin} f$ and $u^{[k]} \in \partial f(x^{[k]})$, the subgradient inequality $f(x^\star) \geq f(x^{[k]}) + \langle u^{[k]}, x^\star - x^{[k]} \rangle$ gives

$$\begin{aligned} \|x^{[k+1]} - x^\star\|^2 &= \|x^{[k]} - x^\star\|^2 - 2\gamma_k \langle u^{[k]}, x^{[k]} - x^\star \rangle + \gamma_k^2 \|u^{[k]}\|^2 \\ &\leq \|x^{[k]} - x^\star\|^2 - 2\gamma_k (f(x^{[k]}) - f^\star) + \gamma_k^2 \|u^{[k]}\|^2, \end{aligned} \tag{8}$$

with $f^\star = \min f$. The term $-2\gamma_k(f(x^{[k]}) - f^\star) \leq 0$ brings the iterate closer to $x^\star$; the term $\gamma_k^2 \|u^{[k]}\|^2 \geq 0$ pushes it away. For a small γ_k, the first one prevails: the iterate gets closer to $x^\star$ *even if f increases*. Geometrically, $-u^{[k]}$ makes an acute angle with $x^\star - x^{[k]}$.

2. The step size must tend to 0. For a smooth function, the gradient vanishes at the minimizer: $\nabla f(x^{[k]}) \to 0$ when $x^{[k]} \to x^\star$, so the displacement $\gamma \nabla f(x^{[k]})$ tends to 0 even with a constant step size. For a non-smooth function, nothing of the sort: subgradients do not tend to 0 near the minimizer. For $f = |\cdot|$, $\partial f(x) = \{\operatorname{sign}(x)\}$ for $x \neq 0$, so $\left|u^{[k]}\right| = 1$ as long as $x^{[k]} \neq 0$. The mapping $x \mapsto \partial f(x)$ is not continuous at $x^\star$: the only zero value, $0 \in \partial f(0)$, is attained only at the exact minimizer.

With a constant step size γ, the iteration $x^{[k+1]} = x^{[k]} - \gamma \operatorname{sign}(x^{[k]})$ decreases $\left|x^{[k]}\right|$ by γ at each step as long as $\left|x^{[k]}\right| > \gamma$. Then, unless $x^{[k]}$ lands exactly on 0, it oscillates forever between two values a and $a - \gamma$ of opposite signs, with $a \in]0, \gamma[$: it does not converge. For example, for $x^{[0]} = 1.3$ and $\gamma = 1$, we obtain $1.3,\ 0.3,\ -0.7,\ 0.3,\ -0.7, \ldots$

Inequality (8) quantifies this phenomenon. Assume $\|u^{[k]}\| \leq G$, for instance if f is G-Lipschitz. Summing from 0 to k, we obtain

$$\min_{0 \leq i \leq k} f(x^{[i]}) - f^\star \leq \frac{\|x^{[0]} - x^\star\|^2 + G^2 \sum_{i=0}^k \gamma_i^2}{2 \sum_{i=0}^k \gamma_i}. \tag{9}$$

- **Constant step size $\gamma_i = \gamma$.** The right-hand side tends to $G^2\gamma/2$: we only converge to a *neighborhood* of $f^\star$, whose size is proportional to the step size. This is the oscillation observed above.
- **Decreasing step size.** For the bound to tend to 0, we need $\sum_i \gamma_i = +\infty$, to be able to reach $x^\star$ whatever the initialization, and $\gamma_k \to 0$, for the error term $\sum \gamma_i^2 / \sum \gamma_i$ to vanish. For example $\gamma_k \propto 1/\sqrt{k+1}$, or $\gamma_k \propto 1/(k+1)$, which even satisfies $\sum \gamma_k^2 < +\infty$, which additionally ensures the convergence of the iterates.

3. The iterates almost never have exactly zero components. The update (6) sets $x_n^{[k+1]}$ exactly to 0 only if $x_n^{[k]} = \gamma_k u_n^{[k]}$, which happens only by coincidence. Even when the solution is sparse, the iterates oscillate around 0 with an amplitude of the order of γ_k, without settling there: sparsity is only asymptotic. In contrast, the proximity operator of $\lambda|\cdot|$, the soft thresholding of Example 3.6, maps a whole interval $[-\lambda, \lambda]$ exactly to 0. Proximal algorithms thus produce exactly sparse iterates from the very first iterations.

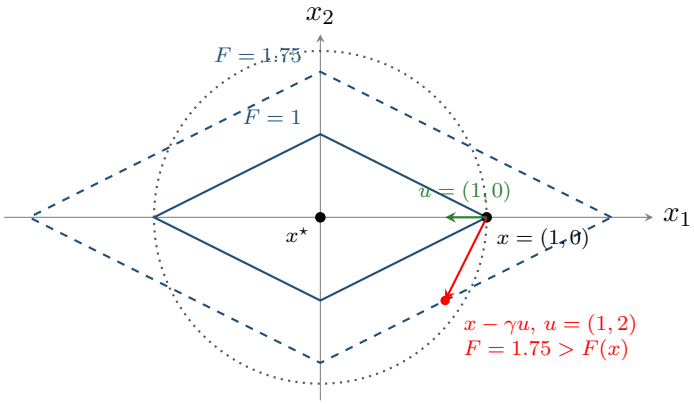


Figure 10: Shortcoming 1 for $F(x) = |x_1| + 2\,|x_2|$ at $x = (1, 0)$, with $\gamma = 0.25$. With the subgradient $u = (1, 2)$ (red), the resulting point leaves the level line $F = 1$ and reaches $F = 1.75$: F increases. It is nevertheless closer to $x^\star$, since it enters the dotted disk of radius $\|x - x^\star\|$. Among the subgradients $u = (1, s)$, $s \in [-2, 2]$, only those with $|s| < \frac{1}{2}$ decrease F, such as $u = (1, 0)$ (green), which has minimal norm.

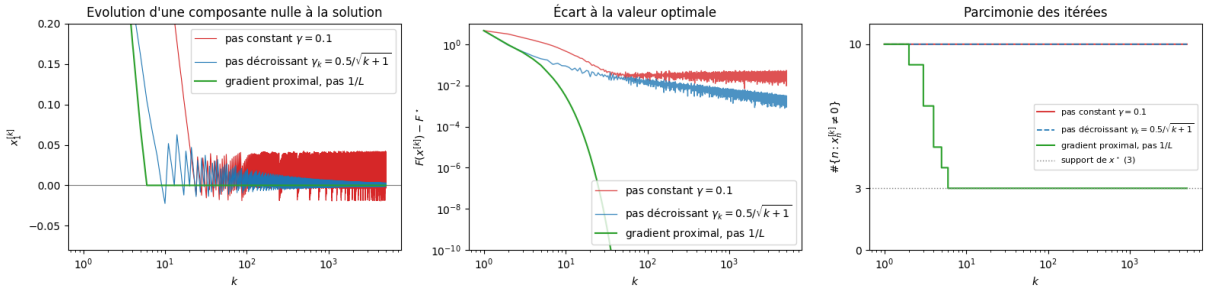


Figure 11: Shortcoming 3 on a problem $\min_x \frac{1}{2}\|Ax - z\|^2 + \lambda\|x\|_1$ with a random $A \in \mathbb{R}^{30\times 10}$, whose solution has 3 nonzero components. Left: the component x_1, which is zero at the solution, oscillates around 0 with the subgradient method ($\gamma_k = 0.5/\sqrt{k+1}$), whereas the soft thresholding of the proximal gradient (step size $1/L$) sets it exactly to zero within a few iterations. Right: the proximal gradient finds the correct support from $k = 5$ on, whereas all the components of the subgradient iterates remain nonzero.

6 Exercises

Exercise 1 (Calculus rules)**.** Prove the translation, scaling and quadratic perturbation rules. Deduce the prox of the *elastic net* $\xi \mapsto \lambda\,|\xi| + \frac{\alpha}{2}\xi^2$.

Solution hints

References

[1] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2nd ed., Springer, 2017.

[2] A. Beck, *First-Order Methods in Optimization*, SIAM, 2017.

[3] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems", *SIAM J. Imaging Sciences*, 2(1):183–202, 2009.

[4] C. Chaux, P. L. Combettes, J.-C. Pesquet and V. R. Wajs, "A variational formulation for frame-based inverse problems", *Inverse Problems*, 23(4):1495–1518, 2007.

[5] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting", *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

[6] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing", in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, Springer, 2011, pp. 185–212.

[7] I. Daubechies, M. Defrise and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint", *Comm. Pure Appl. Math.*, 57(11):1413–1457, 2004.

[8] O. Güler, "On the convergence of the proximal point algorithm for convex minimization", *SIAM J. Control Optim.*, 29(2):403–419, 1991.

[9] B. Martinet, "Régularisation d'inéquations variationnelles par approximations successives", *Revue française d'informatique et de recherche opérationnelle*, 4(R3):154–158, 1970.

[10] J.-J. Moreau, "Proximité et dualité dans un espace hilbertien", *Bulletin de la Société Mathématique de France*, 93:273–299, 1965.

[11] N. Parikh and S. Boyd, "Proximal algorithms", *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

[12] N. Pustelnik and L. Condat, "Proximity operator of a sum of functions; application to depth map estimation", *IEEE Signal Processing Letters*, 24(12):1827–1831, 2017.

[13] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.

[14] R. T. Rockafellar, "Monotone operators and the proximal point algorithm", *SIAM J. Control Optim.*, 14(5):877–898, 1976.

[15] Y.-L. Yu, "On decomposing the proximal map", *Advances in Neural Information Processing Systems*, 2013.