A characterization of proximity operators

Rémi Gribonval, Mila Nikolova

A CHARACTERIZATION OF PROXIMITY OPERATORS

RÉMI GRIBONVAL AND MILA NIKOLOVA

Abstract. We characterize proximity operators, that is to say functions that map a vector to a solution of a penalized least squares optimization problem. Proximity operators of convex penalties have been widely studied and fully characterized by Moreau. They are also widely used in practice with nonconvex penalties such as the ℓ^0 pseudo-norm, yet the extension of Moreau's characterization to this setting seemed to be a missing element of the literature. We characterize proximity operators of (convex or nonconvex) penalties as functions that are the subdifferential of some convex potential. This is proved as a consequence of a more general characterization of so-called Bregman proximity operators of possibly nonconvex penalties in terms of certain convex potentials. As a side effect of our analysis, we obtain a test to verify whether a given function is the proximity operator of some penalty, or not. Many well-known shrinkage operators are indeed confirmed to be proximity operators. However, we prove that windowed Group-LASSO and persistent empirical Wiener shrinkage – two forms of so-called social sparsity shrinkage– are generally *not* the proximity operator of any penalty; the exception is when they are simply weighted versions of group-sparse shrinkage with non-overlapping groups.

Keywords: Proximity operator; Convex regularization; Nonconvex regularization; Subdifferential; Shrinkage operator; Social sparsity; Group sparsity

1. Introduction and overview

Proximity operators have become an important ingredient of nonsmooth optimization, where a huge body of work has demonstrated the power of iterative proximal algorithms to address large-scale variational optimization problems. While these techniques have been thoroughly analyzed and understood for proximity operators involving convex penalties, there is a definite trend towards the use of proximity operators of nonconvex penalties such as the ℓ^0 penalty [7, 8].

This paper extends existing characterizations of proximity operators – which are specialized for convex penalties – to the nonconvex case. A particular motivation is to understand whether certain thresholding rules known as *social sparsity shrinkage*, which have been successfully exploited in the context of certain linear inverse problems, are proximity operators. Another motivation is to characterize when Bayesian estimation with the conditional mean estimator (also

This work and the companion paper [19] are dedicated to the memory of Mila Nikolova, who passed away prematurely in June 2018. Mila dedicated much of her energy to bring the technical content to completion during the spring of 2018. The first author did his best to finalize the papers as Mila would have wished. He should be held responsible for any possible imperfection in the final manuscript.

R. Gribonval (remi.gribonval@inria.fr) was with Univ Rennes, Inria, CNRS, IRISA when this work was conducted. He is now with Univ Lyon, Inria, CNRS, ENS de Lyon, UCB Lyon 1, LIP UMR 5668, F-69342, Lyon, France.
M. Nikolova, CMLA, CNRS and ENS de Cachan, Université Paris-Saclay, 94235 Cachan, France.

known as minimum mean square error estimation or MMSE) can be expressed as a proximity operator. This is the object of a companion paper [19] characterizing when certain variational approaches to address inverse problems can in fact be considered as Bayesian approaches.

1.1. **Characterization of proximity operators.** Let $\mathcal{H}$ be a Hilbert space equipped with an inner product $\langle \cdot, \cdot \rangle$ and a norm $\|\cdot\|$. This includes the case $\mathcal{H} = \mathbb{R}^n$, and most of the text can be read with this simpler setting in mind. The proximity operator of a function $\varphi : \mathcal{H} \to \mathbb{R}$ maps each $y \in \mathcal{H}$ to the solutions of a penalized least-squares problem

$$y \mapsto \mathrm{prox}_{\varphi}(y) := \arg\min_{x \in \mathcal{H}} \left\{ \tfrac{1}{2}\|y - x\|^2 + \varphi(x) \right\}$$

Formally, a proximity operator is set-valued as there may be several solutions to this problem, or the set of solutions may be empty. A primary example is the soft-thresholding function $f(y) := y \max(1 - 1/|y|, 0)$, $y \in \mathbb{R}$, which is the proximity operator of the absolute value function $\varphi(x) := |x|$.

Proximity operators can be defined for certain generalized functions $\varphi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$. A particular example is the projection onto a given closed convex set $\mathcal{C} \subset \mathcal{H}$, which can be written as $\mathrm{proj}_{\mathcal{C}} = \mathrm{prox}_{\varphi}$ with φ the indicator function of $\mathcal{C}$, i.e., $\varphi(x) = 0$ if $x \in \mathcal{C}$, $\varphi(x) = +\infty$ otherwise. For the sake of precision and brevity, we use the following definition:

Definition 1. *Let $\mathcal{Y} \subset \mathcal{H}$ be non-empty. A function $f : \mathcal{Y} \to \mathcal{H}$ is* a *proximity operator of a function $\varphi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ if, and only if, $f(y) \in \mathrm{prox}_{\varphi}(y)$ for each $y \in \mathcal{Y}$.*

In convex analysis, this corresponds to the notion of a *selection* of the set-valued mapping prox_{φ}.

A characterization of proximity operators *of convex lower semicontinuous (l.s.c.) functions* is due to Moreau. It involves the subdifferential $\partial\theta(x)$ of a convex l.s.c. function θ at x, i.e., the set of all its subgradients at x [14, Chapter III.2][1].

Proposition 1. [26, Corollary 10.c] *A function $f : \mathcal{H} \to \mathcal{H}$ defined everywhere is the proximity operator of a* proper convex l.s.c. *function $\varphi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ if, and only if the following conditions hold jointly:*

(a) *there exists a (convex l.s.c.) function ψ such that for each $y \in \mathcal{H}$, $f(y) \in \partial\psi(y)$;*
(b) *f is nonexpansive, i.e.*

$$\|f(y) - f(y')\| \leqslant \|y - y'\|, \quad \forall y, y' \in \mathcal{H}.$$

We extend Moreau's result to possibly nonconvex functions φ on subdomains of $\mathcal{H}$ by simply relaxing the non-expansivity condition:

Theorem 1. *Let $\mathcal{Y} \subset \mathcal{H}$ be non-empty. A function $f : \mathcal{Y} \to \mathcal{H}$ is a proximity operator of a function $\varphi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ if, and only if, there exists a convex l.s.c. function $\psi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ such that for each $y \in \mathcal{Y}$, $f(y) \in \partial\psi(y)$.*

[1]See Section 2.1 for detailed notations and reminders on convex analysis and differentiability in Hilbert spaces.

This is proved in Section 2 as a particular consequence of our main result, Theorem 3, which characterizes functions such that $f(y) \in \arg\min_{x\in\mathcal{H}}\{D(x,y)+\varphi(x)\}$ for certain types of data-fidelity terms $D(x,y)$. Among others, the data-fidelity terms covered by Theorem 3 include:

- the Euclidean distance $D(x,y) = \frac{1}{2}\|y-x\|^2$, which is the data-fidelity associated to proximity operators;
- its variant $D(x,y) = \frac{1}{2}\|y-Mx\|^2$ with M some linear operator; and
- Bregman divergences [9], leading to an analog of Theorem 1 to characterize so-called Bregman proximity operators [12] (see Corollary 5 in Section 2).

Theorem 3 further implies that the functions φ and ψ in Theorem 1 can be chosen such that

$$\psi(y) = \langle y, f(y)\rangle - \tfrac{1}{2}\|f(y)\|^2 - \varphi(f(y)), \quad \forall\, y \in \mathcal{Y}. \tag{1}$$

This is a particular instance of a more general result valid for all considered data-fidelity terms. Another consequence of Theorem 3 (see Corollary 4 in Section 2) is that for the considered data-fidelity terms $D(x,y)$, if $f: \mathcal{Y} \to \mathcal{H}$ can be written as $f(y) \in \arg\min_{x\in\mathcal{H}}\{D(x,y)+\varphi(x)\}$ for some (possibly nonconvex) function φ and if its image $\mathrm{Im}(f) := f(\mathcal{Y})$ is a convex set (e.g., if $\mathrm{Im}(f) = \mathcal{H}$) then

the function $x \mapsto D(x,y)+\varphi(x)$ is convex on $\mathrm{Im}(f)$.

This is reminiscent of observations on convex optimization with nonconvex penalties [28, 31] and on the hidden convexity of conditional mean estimation under additive Gaussian noise [17, 18, 25, 1]. The latter is extended to other noise models in the companion paper [19].

1.2. **The case of smooth proximity operators.** The smoothness of a proximity operator $f = \mathrm{prox}_\varphi$ and that of the corresponding functions φ and ψ, cf (1), are inter-related, leading to a characterization of *continuous* proximity operators[2].

Corollary 1. *Let $\mathcal{Y} \subset \mathcal{H}$ be non-empty and open and $f: \mathcal{Y} \to \mathcal{H}$ be C^0. The following are equivalent:*

(a) *f is a proximity operator of a function $\varphi: \mathcal{H} \to \mathbb{R}\cup\{+\infty\}$;*
(b) *there exists a convex $C^1(\mathcal{Y})$ function ψ such that $f(y) = \nabla\psi(y)$ for each $y \in \mathcal{Y}$.*

This is established in Section 2.6 as a particular consequence of our second main result, Corollary 6. There, we also prove that when f is a proximity operator of some φ, the Lipschitz property of f with Lipschitz constant L is equivalent to the convexity of $x \mapsto \varphi(x) + \left(1-\frac{1}{L}\right)\frac{\|x\|^2}{2}$. Moreau's characterization (Proposition 1) corresponds to the special case $L = 1$. Next, we characterize C^1 proximity operators on convex domains more explicitly using the differential of f.

Theorem 2. *Let $\mathcal{Y} \subset \mathcal{H}$ be non-empty, open and convex, and $f: \mathcal{Y} \to \mathcal{H}$ be C^1. The following properties are equivalent:*

(a) *f is a proximity operator of a function $\varphi: \mathcal{H} \to \mathbb{R}\cup\{+\infty\}$;*
(b) *there exists a convex $C^2(\mathcal{Y})$ function ψ such that $f(y) = \nabla\psi(y)$ for each $y \in \mathcal{Y}$;*

[2]See Section 2.1 for brief reminders on the notion of continuity / differentiability in Hilbert spaces.

(c) *the differential* $Df(y)$ *is a* symmetric positive semi-definite *operator*[3] *for each* $y \in \mathcal{Y}$.

Proof. Since f is C^1, the equivalence (a) $\Leftrightarrow$ (b) is a consequence of Corollary 1. We now establish (b) $\Leftrightarrow$ (c). Since $\mathcal{Y}$ is convex it is simply connected, and as $\mathcal{Y}$ is open by Poincar's lemma (see [16, Theorem 6.6.3] when $\mathcal{H} = \mathbb{R}^n$) the differential Df is symmetric if, and only if, f is the gradient of some C^2 function ψ. By [6, Proposition 17.7], the function ψ is convex iff $\nabla^2\psi \succeq 0$ on $\mathcal{Y}$, i.e. iff $Df \succeq 0$ on $\mathcal{Y}$. □

Corollary 2. *Let* $\mathcal{Y} \subset \mathcal{H}$ *be a set with non-empty interior* $\mathrm{int}(\mathcal{Y}) \neq \varnothing$, $y \in \mathrm{int}(\mathcal{Y})$, *and* $f : \mathcal{Y} \to \mathcal{H}$ *be a proximity operator. If* f *is* C^1 *in a neighborhood of* y, *then* $Df(y)$ *is symmetric positive semi-definite.*

Proof. Restrict f to any open convex neighborhood $\mathcal{Y}' \subset \mathcal{Y}$ of y and apply Theorem 2. □

Remark 1. Differentials are perhaps more familiar to some readers in the context of multivariate calculus: when $y = (y_j)_{j=1}^n \in \mathcal{H} = \mathbb{R}^n$ and $f(y) = (f_i(y))_{i=1}^n$, $Df(y)$ is identified to the Jacobian matrix

$$Jf(y) = (\tfrac{\partial f_i}{\partial y_j})_{1\leqslant i,j\leqslant n}.$$

The rows of $Jf(y)$ are the transposed gradients $\nabla f_i(y)$. The differential is symmetric if the mixed derivatives satisfy $\frac{\partial f_i}{\partial y_j} = \frac{\partial f_j}{\partial y_i}$ for all $i \neq j$. When $n = 3$, this corresponds to f being an *irrotational vector field*. More generally, this characterizes the fact that f is a so-called *conservative field*, i.e., a vector field that is the gradient of some potential function. As the Jacobian is the Hessian of this potential, it is positive definite if the potential is convex.

Finally we provide conditions ensuring that f is a proximity operator and that $f(y)$ is the only critical point of the corresponding optimization problem.

Corollary 3. *Let* $\mathcal{Y} \subset \mathcal{H}$ *be open and convex, and* $f : \mathcal{Y} \to \mathcal{H}$ *be* C^1 *with* $Df(y) \succ 0$ *on* $\mathcal{Y}$. *Then* f *is injective and there is* $\varphi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ *such that* $\mathrm{prox}_\varphi(y) = \{f(y)\}$, $\forall\, y \in \mathcal{Y}$ *and* $\mathrm{dom}(\varphi) = \mathrm{Im}(f)$. *Moreover, if* $Df(y)$ *is* boundedly invertible *on* $\mathcal{Y}$ *then* φ *is* C^1 *on* $\mathcal{Y}$ *and for each* $y \in \mathcal{Y}$, *the only critical point of* $x \mapsto \frac{1}{2}\|y - x\|^2 + \varphi(x)$ *is* $x = f(y)$.

This is established in Appendix A.6.

Remark 2. In finite dimension $\mathcal{H} = \mathbb{R}^n$, $Df(y)$ is boundedly invertible as soon as $Df(y) \succ 0$, hence we only need to assume that $Df(y) \succ 0$ to conclude that $f(y)$ is the unique critical point. This is no longer the case in infinite dimension. Indeed, consider $\mathcal{H} = \ell^2(\mathbb{N})$ and $f : y = (y_n)_{n\in\mathbb{N}} \mapsto f(y) := (y_n/(n+1))_{n\in\mathbb{N}}$. As f is linear, its differential is $Df(y) = f$ for every $y \in \mathcal{H}$. As $\langle f(y), y\rangle = \sum_{n\in\mathbb{N}} y_{n^2}/(n+1) > 0$ for each nonzero $y \in \mathcal{H}$ we have $Df(y) \succ 0$ but its inverse is unbounded. Given $n \in \mathbb{N}$ and $z \in \mathbb{R}$ we have $z/(n+1) = \arg\min_{x\in\mathbb{R}} \frac{1}{2}(z-x)^2 + nx^2/2$, hence $f = \mathrm{prox}_{\varphi_0}$ with $\varphi_0 : x = (x_n)_{n\in\mathbb{N}} \mapsto \varphi_0(x) := \sum_{n\in\mathbb{N}} nx_n^2/2$. Setting $\varphi(x) = \varphi_0(x)$ for $x \in \mathrm{Im}(f)$, $\varphi(x) = +\infty$ otherwise, we have $\mathrm{prox}_\varphi = f$ and $\mathrm{dom}(\varphi) = \mathrm{Im}(f) = \{x \in$

[3] A continuous linear operator $L : \mathcal{H} \to \mathcal{H}$ is symmetric if $\langle x, Ly\rangle = \langle Lx, y\rangle$ for each $x, y \in \mathcal{H}$. A symmetric continuous linear operator is positive semi-definite if $\langle x, Lx\rangle \geqslant 0$ for each $x \in \mathcal{H}$. This is denoted $L \succeq 0$. It is positive definite if $\langle x, Lx\rangle > 0$ for each nonzero $x \in \mathcal{H}$. This is denoted $L \succ 0$.

$\mathcal{H}, \sum_{n\in\mathbb{N}}(n+1)^2 x_n^2 < \infty\}$. Yet, as no point in $\mathrm{dom}(\varphi)$ admits any open neighborhood in $\mathcal{H}$, φ is nowhere differentiable and *every* $x \in \mathcal{H}$ is a critical point of $x \mapsto \frac{1}{2}\|y-x\|^2 + \varphi(x)$.

Terminology. Proximity operators often appear in the context of penalized least squares regression, where φ is called a *penalty*, and from now on we will adopt this terminology. In light of Corollary 1, a continuous proximity operator is exactly characterized as a gradient of a convex function ψ. In the terminology of physics, a proximity operator is thus a *conservative field* associated to a *convex potential*. In the language of convex analysis, subdifferentials of convex functions are characterized as maximal cyclically monotone operators [30, Theorem B].

1.3. **Organization of the paper.** The proof of our most general results, Theorem 3 and Corollary 6 (and the fact that they imply Theorem 1, (1), Corollary 4 and Corollary 1) are established in Section 2, where we also discuss their consequences in terms of Bregman proximity operators and illustrate them on concrete examples. As Theorem 1 and its corollaries characterize whether a function f is a proximity operator and study its smoothness in relation to that of the corresponding penalty and potential, they are particularly useful when f is not *explicitly built* as a proximity operator. This is the case of so-called social shrinkage operators (see e.g. [22]). We conclude the paper by showing in Section 3 that social shrinkage operators are generally *not* the proximity operator of any penalty.

1.4. **Discussion.** In light of the extension to nonconvex penalties of Moreau's characterization of proximity operators of convex (l.s.c.) penalties (Proposition 1), the nonexpansivity of the proximity operator f determines whether the underlying penalty φ is convex or not. While nonexpansivity certainly plays a role in the convergence analysis of iterative proximal algorithms based on convex penalties, the adaptation of such an analysis when the proximity operator is Lipschitz rather than nonexpansive, using Proposition 1, is an interesting perspective.

The characterization of smooth proximity operators as the gradients of convex potentials, which also appear in optimal transport (see e.g., [36]), suggests that further work is needed to better understand the connections between these concepts and tools. This could possibly lead to simplified arguments where the strong machinery of convex analysis may be used more explicitly despite the apparent lack of convexity of the optimization problems associated to nonconvex penalties.

2. Main results

We now state our main results, Theorem 3 and Corollary 6, and prove a number of their consequences including Theorem 1, (1), Corollary 4 and Corollary 1 which were advertized in Section 1. The most technical proofs are postponed to the Appendix.

2.1. **Detailed notations.** The indicator function of a set $\mathcal{S}$ is denoted

$$\chi_{\mathcal{S}}(x) := \begin{cases} 0 & \text{if} \quad x \in \mathcal{S}, \\ +\infty & \text{if} \quad x \notin \mathcal{S}. \end{cases}$$

The domain of a function $\theta : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is defined and denoted by $\mathrm{dom}(\theta) := \{x \in \mathcal{H} \mid \theta(x) < \infty\}$. Given $\mathcal{Y} \subset \mathcal{H}$ and a function $f : \mathcal{Y} \to \mathcal{H}$, the image of $\mathcal{Y}$ under f is denoted by $\mathrm{Im}(f)$. A function $\theta : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is proper iff there is $x \in \mathcal{H}$ such that $\theta(x) < +\infty$, i.e.,

$\text{dom}(\theta) \neq \varnothing$. It is lower semicontinuous (l.s.c.) if for each $x_0 \in \mathcal{H}$, $\liminf_{x \to x_0} \theta(x) \geqslant \theta(x_0)$, or equivalently if the set $\{x \in \mathcal{H} : \theta(x) > \alpha\}$ is open for every $\alpha \in \mathbb{R}$. A subgradient of a convex function $\theta : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ at x is any $u \in \mathcal{H}$ such that $\theta(x') - \theta(x) \geqslant \langle u, x' - x\rangle, \forall x' \in \mathcal{H}$. A function with k continuous derivatives[4] is called a C^k function. The notation $C^k(\mathcal{X})$ is used to specify a C^k function on an open domain $\mathcal{X}$. Thus C^0 is the space of continuous functions, whereas C^1 is the space of continuously differentiable functions [11, p. 327]. The gradient of a C^1 scalar function θ at x is denoted $\nabla\theta(x)$.

The segment between two elements $x, x' \in \mathcal{H}$ is the set $[x, x'] := \{tx + (1-t)x', t \in [0, 1]\}$. A finite union of segments $[x_{i-1}, x_i]$, $1 \leqslant i \leqslant n$, $n \in \mathbb{N}$, where $x_0 = x$ and $x_n = x'$ is called a polygonal path between x and x'. A non-empty subset $\mathcal{C} \subset \mathcal{H}$ is polygonally connected iff between each pair $x, x' \in \mathcal{C}$ there is a polygonal path with all its segments included in $\mathcal{C}$, $[x_{i-1}, x_i] \subset \mathcal{C}$.

Remark 3. The notion of polygonal-connectedness is a bit stronger than that of connectedness. Indeed, polygonal-connectedness implies the classical topological property of path-connectedness, which in turn implies connectedness. However there are path-connected sets that are not polygonally-connected – e.g., the unit circle in $\mathbb{R}^2$ is path-connected, but no two points are polygonally-connected, and there are connected sets that are not path-connected. Yet, every *open* connected set is polygonally-connected, see [16, Theorem 2.5.2] for a statement in $\mathbb{R}^n$.

2.2. **Main theorem.**

Theorem 3. *Consider $\mathcal{H}$ and $\mathcal{H}'$ two Hilbert spaces*[5]*, and $\mathcal{Y} \subset \mathcal{H}'$ a non-empty set. Let $a : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$, $b : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$, $A : \mathcal{Y} \to \mathcal{H}$ and $B : \mathcal{H} \to \mathcal{H}'$ be arbitrary functions. Consider $f : \mathcal{Y} \to \mathcal{H}$ and denote* $\text{Im}(f)$ *the image of $\mathcal{Y}$ under f.*

(a) *Let $D(x, y) := a(y) - \langle x, A(y)\rangle + b(x)$. The following properties are equivalent:*
   (i) *there is $\varphi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ such that $f(y) \in \arg\min_{x \in \mathcal{H}}\{D(x, y) + \varphi(x)\}$ for each $y \in \mathcal{Y}$;*
   (ii) *there is a convex l.s.c. $g : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ such that $A(f^{-1}(x)) \subset \partial g(x)$ for each $x \in \text{Im}(f)$;*

   When they hold, φ (resp. g) can be chosen given g (resp. φ) so that $g(x) + \chi_{\text{Im}(f)} = b(x) + \varphi(x)$.

(b) *Let φ and g satisfy* (ai) *and* (aii)*, respectively, and let $\mathcal{C} \subset \text{Im}(f)$ be polygonally connected. Then there is $K \in \mathbb{R}$ such that*

$$g(x) \quad = \quad b(x) + \varphi(x) + K, \quad \forall\, x \in \mathcal{C}. \tag{2}$$

(c) *Let $\widetilde{D}(x, y) := a(y) - \langle B(x), y\rangle + b(x)$. The following properties are equivalent:*
   (i) *there is $\varphi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ such that $f(y) \in \arg\min_{x \in \mathcal{H}}\{\widetilde{D}(x, y) + \varphi(x)\}$ for each $y \in \mathcal{Y}$;*

[4]see Appendix A.1 for some reminders on Fréchet derivatives in Hilbert spaces.

[5]For the sake of simplicity we use the same notation $\langle\cdot,\cdot\rangle$ for the inner products $\langle x, A(y)\rangle$ (between elements of $\mathcal{H}$) and $\langle B(x), y\rangle$ (between elements of $\mathcal{H}'$). The reader can inspect the proof of Theorem 3 to check that the result still holds if we consider *Banach spaces* $\mathcal{H}$ and $\mathcal{H}'$, $\mathcal{H}^\star$ and $(\mathcal{H}')^\star$ their duals, and $A : \mathcal{Y} \to \mathcal{H}^\star$, $B : \mathcal{H} \to (\mathcal{H}')^\star$.

(ii) *there is a convex l.s.c.* $\psi : \mathcal{H}' \to \mathbb{R} \cup \{+\infty\}$ *such that* $B(f(y)) \in \partial\psi(y)$ *for each* $y \in \mathcal{Y}$. φ *(resp.* ψ*) can be chosen given* ψ *(resp.* φ*) so that* $\psi(y) = \langle B(f(y), y\rangle - b(f(y)) - \varphi(f(y))$ *on* $\mathcal{Y}$.

(d) *Let* φ *and* ψ *satisfy* (ci) *and* (cii)*, respectively, and let* $\mathcal{C}' \subset \mathcal{Y}$ *be polygonally connected. Then there is* $K' \in \mathbb{R}$ *such that*

$$\psi(y) \quad = \quad \langle B(f(y)), y\rangle - b(f(y)) - \varphi(f(y)) + K', \quad \forall\, y \in \mathcal{C}'. \tag{3}$$

The proof of Theorem 3 is postponed to Appendix A.4. As stated in (a) (resp. (c)), the functions can be chosen such that the relation (2) (resp. (3)) holds on $\mathrm{Im}(f)$ (resp. on $\mathcal{Y}$) with $K = K' = 0$. As the functions φ, g, ψ are at best defined up to an additive constant, we provide in (b) (resp. (d)) conditions ensuring that adding a constant is indeed the unique degree of freedom. The role of polygonal-connectedness will be illustrated on examples in Section 2.7.

Example 1. In the context of linear inverse problems one often encounters optimization problems involving functions expressed as $\frac{1}{2}\|y-Mx\|^2+\varphi(x)$ with M some linear operator. Such functions fit into the framework of Theorem 3 using $a(y) := \frac{1}{2}\|y\|^2$, $b(x) := \frac{1}{2}\|Mx\|^2$, $A(y) := M^\star y$, and $B(x) := Mx$, where $M^\star$ is the adjoint of M. Among other consequences one gets that $f : \mathcal{Y} \to \mathcal{H}$ is a generalized proximity operator of this type for some penalty φ if, and only if, there is a convex l.s.c. ψ such that $Mf(y) \in \partial\psi(y)$ for each $y \in \mathcal{Y}$.

Examples where the data-fidelity term is a so-called Bregman divergence are detailed in Section 2.4 below. This covers the case of standard proximity operators where $D(x, y) = \frac{1}{2}\|y - x\|^2$.

2.3. **Convexity in proximity operators of nonconvex penalties.** An interesting consequence of Theorem 3 is that the optimization problem associated to (generalized) proximity operators is in a sense always convex, even when the considered penalty φ is not convex.

Corollary 4. *Consider* $\mathcal{H}, \mathcal{H}'$ *two Hilbert spaces. Let* $\mathcal{Y} \subset \mathcal{H}'$ *be non-empty and* $f : \mathcal{Y} \to \mathcal{H}$. *Assume that there is* $\varphi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ *such that* $f(y) \in \arg\min_{x\in\mathcal{H}} \{D(x, y) + \varphi(x)\}$ *for each* $y \in \mathcal{Y}$*, with* $D(x, y) = a(y) - \langle x, A(y)\rangle + b(x)$ *as in Theorem 3*(a)*. Then*

(a) *the function* $x \mapsto b(x) + \varphi(x)$ *is convex on each convex subset* $\mathcal{C} \subset \mathrm{Im}(f)$*;*

(b) *if* $\mathrm{Im}(f)$ *is convex, then the function* $x \in \mathrm{Im}(f) \mapsto D(x, y) + \varphi(x)$ *is convex,* $\forall\, y \in \mathcal{Y}$.

Similarly, if there is $\varphi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ *such that* $f(y) \in \arg\min_{x\in\mathcal{H}} \left\{\widetilde{D}(x, y) + \varphi(x)\right\}$ *for each* $y \in \mathcal{Y}$*, with* $\widetilde{D}(x, y) = a(y) - \langle B(x), y\rangle + b(x)$ *as in Theorem 3*(c) *then* $y \mapsto \langle B(f(y)), y\rangle - b(f(y)) - \varphi(f(y))$ *is convex on each convex subset* $\mathcal{C}' \subset \mathcal{Y}$.

Proof. (a) follows from Theorem 3(a)-(b). (b) follows from (a) and the definition of D. The proof of the result with $\widetilde{D}$ instead of D is similar. □

Corollary 4(b) might seem surprising as, given a nonconvex penalty φ, one may expect the optimization problem $\min_x D(x, y) + \varphi(x)$ to be nonconvex. However, as noticed e.g. by [27, 28, 31], there are nonconvex penalties such that this problem with $D(x, y) := \frac{1}{2}\|y - x\|^2$ is in fact convex. Corollary 4 establishes that this convexity property indeed holds whenever the image $\mathrm{Im}(f)$ of the resulting function f is a convex set. A particular case is that of functions f built as conditional expectations in the context of additive Gaussian denoising, which have been shown [17] to be proximity operators. Extensions of this phenomenon for conditional mean estimation with other noise models are discussed in the companion paper [19].

2.4. **Application to Bregman proximity operators.** The squared Euclidean norm is a particular *Bregman divergence*, and Theorem 3 characterizes generalized proximity operators defined with such divergences. The Bregman divergence, known also as D-function, was introduced in [9] for strictly convex differentiable functions on so-called linear topological spaces. For the goals of our study, it will be enough to consider that $h : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is proper, convex and differentiable on a Hilbert space.

Definition 2. *Let $h : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be proper convex and differentiable on its open domain* $\mathrm{dom}(h)$*. The Bregman divergence (associated with h) between x and y is defined by*

$$D_h : \mathcal{H} \times \mathcal{H} \to [0, +\infty] : (x, y) \to \begin{cases} h(x) - h(y) - \langle \nabla h(y), x - y\rangle, & \text{if } y \in \mathrm{dom}(h); \\ +\infty, & \text{otherwise} \end{cases} \tag{4}$$

In Theorem 3(a) one obtains $D(x, y) = D_h(x, y)$ by setting $a(y) = +\infty$ and $A(y)$ arbitrary if $y \notin \mathrm{dom}(h)$ and, for $y \in \mathrm{dom}(h)$ and each $x \in \mathcal{H}$,

$$a(y) := \langle \nabla h(y), y\rangle - h(y) \quad b(x) := h(x) \qquad \text{and} \quad A(y) = \nabla h(y) \tag{5}$$

The lack of symmetry of the Bregman divergence suggests to consider also $D_h(y, x)$. In Theorem 3(c) one obtains $\widetilde{D}(x, y) = D_h(y, x)$ using $b(x) = +\infty$ and $B(x)$ arbitrary for $x \notin \mathrm{dom}(h)$ and, for $x \in \mathrm{dom}(h)$ and each $y \in \mathcal{H}$,

$$a(y) := h(y) \qquad b(x) := \langle \nabla h(x), x\rangle - h(x) \quad \text{and} \quad B(x) = \nabla h(x) \tag{6}$$

The next claim is an application of Theorem 3 with $D(x, y) = D_h(x, y)$ and $\widetilde{D}(x, y) = D_h(y, x)$. We thus consider the so-called Bregman proximity operators which were introduced in [12]. We will focus on the characterization of these operators defined by $y \mapsto \arg\min_{x\in\mathcal{H}}\{D_h(x, y) + \varphi(x)\}$ and $y \mapsto \arg\min_{x\in\mathcal{H}}\{D_h(y, x) + \varphi(x)\}$. Such operators have been further studied in [5] with an emphasis on the notion of viability, which is essential for these operators to be useful in the context of iterative algorithms.

Corollary 5. *Consider $f : \mathcal{Y} \to \mathcal{H}$. Let $h : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function that is differentiable on its open domain* $\mathrm{dom}(h)$*. Let D_h read as in* (4)*.*

(a) *The following properties are equivalent:*
 (i) *there is $\varphi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ such that $f(y) \in \arg\min_{x\in\mathcal{H}}\{D_h(x, y) + \varphi(x)\}$, $\forall\, y \in \mathcal{Y}$;*
 (ii) *there is a convex l.s.c. $g : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ s.t. $\nabla h(f^{-1}(x)) \subset \partial g\,(x)$, $\forall\, x \in \mathrm{Im}(f)$;*
 When they hold, φ (resp. g) can be chosen given g (resp. φ) so that $g(x) + \chi_{\mathrm{Im}(f)} = h(x) + \varphi(x)$.

(b) *Let φ and g satisfy* (ai) *and* (aii)*, respectively, and let $\mathcal{C} \subset \mathrm{Im}(f)$ be polygonally connected. Then there is $K \in \mathbb{R}$ such that*

$$g(x) = h(x) + \varphi(x) + K, \quad \forall\, x \in \mathcal{C}.$$

(c) *The following properties are equivalent:*

(i) *there is* $\varphi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ *such that* $f(y) \in \arg\min_{x\in\mathcal{H}}\{D_h(y,x) + \varphi(x)\}$, $\forall\, y \in \mathcal{Y}$;
(ii) *there is a convex l.s.c.* $\psi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ *such that* $\nabla h(f(y)) \in \partial\psi(y)$, $\forall\, y \in \mathcal{Y}$.

φ can be chosen given ψ (resp. ψ given φ) s.t. $\psi(y) = \langle \nabla h(f(y)), y - f(y)\rangle + h(f(y)) - \varphi(f(y)), \forall y \in \mathcal{Y}$.

(d) *Let φ and ψ satisfy* (ci) (cii), *respectively, and let $\mathcal{C}' \subset \mathcal{Y}$ be polygonally connected. Then there is $K' \in \mathbb{R}$ such that*

$$\psi(y) = \big\langle \nabla h(f(y)), y - f(y)\big\rangle + h(f(y)) - \varphi(f(y)) + K', \quad \forall\, y \in \mathcal{C}'.$$

Proof. (a) and (b) use (5). Further, (c) and (d) use (6). □

2.5. **Specialization to (standard) proximity operators.** Standard (Hilbert space) proximity operators correspond to taking as the Bregman divergence $D_h(x,y) = \frac{1}{2}\|y-x\|^2$, which is associated to $h(x) := \frac{1}{2}\|x\|^2$. An immediate consequence of Corollary 2.4 is the following theorem, which implies Theorem 1 and (1).

Theorem 4. *Let $\mathcal{Y} \subset \mathcal{H}$ be non-empty, and $f : \mathcal{Y} \to \mathcal{H}$.*

(a) *The following properties are equivalent:*
(i) *there is* $\varphi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ *such that* $f(y) \in \operatorname{prox}_\varphi(y)$ *for each* $y \in \mathcal{Y}$;
(ii) *there is a convex l.s.c.* $g : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ *such that* $f^{-1}(x) \subset \partial g\,(x)$ *for each* $x \in \operatorname{Im}(f)$;
(iii) *there is a convex l.s.c.* $\psi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ *such that* $f(y) \in \partial\psi(y)$ *for each* $y \in \mathcal{Y}$.

When they hold, there exists a choice of φ, g, ψ satisfying (ai)-(aii)-(aiii) *such that*

$$\begin{aligned} g(x) + \chi_{\operatorname{Im}(f)} &= \tfrac{1}{2}\|x\|^2 + \varphi(x), \quad \forall\, x \in \mathcal{H};\\ \psi(y) &= \langle y, f(y)\rangle - \tfrac{1}{2}\|f(y)\|^2 - \varphi(f(y)), \quad \forall\, y \in \mathcal{Y}. \end{aligned}$$

(b) *Let φ, g and ψ satisfy* (ai), (aii) *and* (aiii), *respectively. Let $\mathcal{C} \subset \operatorname{Im}(f)$ and $\mathcal{C}' \subset \mathcal{Y}$ be polygonally connected. Then there exist $K, K' \in \mathbb{R}$ such that*

$$g(x) = \tfrac{1}{2}\|x\|^2 + \varphi(x) + K, \quad \forall\, x \in \mathcal{C}; \tag{7}$$

$$\psi(y) = \langle y, f(y)\rangle - \tfrac{1}{2}\|f(y)\|^2 - \varphi(f(y)) + K', \quad \forall\, y \in \mathcal{C}'. \tag{8}$$

2.6. **Local smoothness of proximity operators.** Theorem 4 characterizes proximity operators in terms of three functions: a (possibly nonconvex) penalty φ, a convex potential ψ, and another convex function g. As we now show, the properties of these functions are tightly inter-related. First we extend Moreau's characterization (Proposition 1) as follows:

Proposition 2. *Consider $f : \mathcal{H} \to \mathcal{H}$ defined everywhere, and $L > 0$. The following are equivalent:*

(1) there is $\varphi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ *s.t.* $f(y) \in \operatorname{prox}_\varphi(y)$ *on* $\mathcal{H}$, *and* $x \mapsto \varphi(x) + (1 - \frac{1}{L})\frac{\|x\|^2}{2}$ *is convex l.s.c;*
(2) the following conditions hold jointly:
(a) *there exists a (convex l.s.c.) function ψ such that for each $y \in \mathcal{H}$, $f(y) \in \partial\psi(y)$;*
(b) *f is L-Lipschitz, i.e.*

$$\|f(y) - f(y')\| \leqslant L\|y - y'\|, \quad \forall y, y' \in \mathcal{H}.$$

Proof. (1)⇒(2a). Simply observe that f is a proximity operator and use Theorem 4(ai)⇒(aiii).
(1)⇒(2b). The function $\tilde{\varphi}(z) := \frac{1}{L}(\varphi(Lz) + (1-\frac{1}{L})\frac{\|Lz\|^2}{2})$ is convex l.s.c. by assumption. We prove below that $\tilde{f} := f/L$ is a proximity operator of $\tilde{\varphi}$. By Proposition 1 $\tilde{f}$ is thus non-expansive, i.e., f is L-Lipschitz.
To show $\tilde{f}(y) \in \mathrm{prox}_{\tilde{\varphi}}(y)$ for each $y \in \mathcal{H}$, observe that $\varphi(x) = L\tilde{\varphi}(x/L) - (1-\frac{1}{L})\frac{\|x\|^2}{2}$. For each $x \in \mathcal{H}$

$$\begin{aligned}
\tfrac{1}{2}\|y-x\|^2 + \varphi(x) &= \tfrac{\|y\|^2}{2} - \langle y, x\rangle + \tfrac{\|x\|^2}{2} + L\tilde{\varphi}(x/L) - (1-\tfrac{1}{L})\tfrac{\|x\|^2}{2} = \tfrac{\|y\|^2}{2} - \langle y, x\rangle + \tfrac{\|x\|^2}{2L} + L\tilde{\varphi}(x/L)\\
&= \tfrac{\|y\|^2}{2} - L\langle y, z\rangle + L\tfrac{\|z\|^2}{2} + L\tilde{\varphi}(z)\\
&= (1-L)\tfrac{\|y\|^2}{2} + L\left(\tfrac{1}{2}\|y-z\|^2 + \tilde{\varphi}(z)\right), \quad \text{with } z = x/L.
\end{aligned}$$

Since $x = f(y)$ is a minimizer of the left-hand-side, $z = f(y)/L = \tilde{f}(y)$ is a minimizer of the right hand side, hence $\tilde{f}$ is a proximity operator of $\tilde{\varphi}$ as claimed.
(2a) and (2b)⇒(1). By (2a) the function $\tilde{\psi}(y) := \psi(y)/L$ is convex l.s.c and $f(y)/L \in \partial\tilde{\psi}(y)$. By Theorem 4(aiii)⇒(ai)$\tilde{f} := f/L$ is therefore a proximity operator. Since f is L-Lipschitz, $\tilde{f}$ is non-expansive hence by Proposition 1 $\tilde{f}$ is a proximity operator of some *convex l.s.c* penalty $\tilde{\varphi}$. The function $\varphi(x) := L\tilde{\varphi}(x/L) - (1-\frac{1}{L})\frac{\|x\|^2}{2}$ is such that $\varphi(x) + (1-\frac{1}{L})\frac{\|x\|^2}{2} = L\tilde{\varphi}(x/L)$ is convex l.s.c. as claimed. By the same argument as above, as $z = \tilde{f}(y)$ is a minimizer of $\frac{1}{2}\|y-z\|^2 + \tilde{\varphi}(x)$, $x = Lz = f(y)$ is a minimizer of $\frac{1}{2}\|y-x\|^2 + \varphi(x)$, showing that f is indeed a proximity operator of φ. □

Next we consider additional properties of these functions.

Corollary 6. *Let $\mathcal{Y} \subset \mathcal{H}$ and $f : \mathcal{Y} \to \mathcal{H}$. Consider three functions φ, g, ψ on $\mathcal{H}$ satisfying the equivalent properties* (ai), (aii) *and* (aiii) *of Theorem 4, respectively. Let $k \geqslant 0$ be an integer.*

(a) *Consider an open set $\mathcal{V} \subset \mathcal{Y}$. The following two properties are equivalent:*
  (i) *ψ is $C^{k+1}(\mathcal{V})$;*
  (ii) *f is $C^k(\mathcal{V})$;*
  When one of them holds, we have $f(y) = \nabla\psi(y), \forall y \in \mathcal{V}$.

(b) *Consider an open set $\mathcal{X} \subset \mathrm{Im}(f)$. The following three properties are equivalent:*
  (i) *φ is $C^{k+1}(\mathcal{X})$;*
  (ii) *g is $C^{k+1}(\mathcal{X})$;*
  (iii) *the restriction $\widetilde{f}$ of f to the set $f^{-1}(\mathcal{X})$ is injective and $(\widetilde{f})^{-1}$ is $C^k(\mathcal{X})$.*
  When one of them holds, $\widetilde{f}$ is a bijection between $f^{-1}(\mathcal{X})$ and $\mathcal{X}$, and we have

$$(\widetilde{f})^{-1}(x) = \nabla g(x) = x + \nabla\varphi(x), \forall x \in \mathcal{X}.$$

Before proving this corollary, let us first mention that the characterization of any *continuous* proximity operator f as the gradient of a C^1 convex potential ψ, i.e., $f = \nabla\psi$, is a direct consequence of Corollary 6(a) and Theorem 1. This establishes Corollary 1 from Section 1.

The proof of Corollary 6 relies on the following technical lemma which we prove in Appendix A.5 as a consequence of [6, Prop 17.41].

LEMMA 1. *Consider a function* $\varrho : \mathcal{H} \to \mathcal{H}$*, a function* $\theta : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ *and an open set* $\mathcal{X} \subset \mathrm{dom}(\varrho) \cap \mathrm{dom}(\theta) \subset \mathcal{H}$*. Assume that* θ *is subdifferentiable at each* $x \in \mathcal{X}$ *and that*

$$\forall\, x \in \mathcal{X} \qquad \varrho(x) \in \partial\theta(x) \tag{9}$$

Then the following statements are equivalent:

(a) ϱ *is continuous on* $\mathcal{X}$*;*

(b) θ *is continuously differentiable on* $\mathcal{X}$ *i.e., its gradient* $\nabla\theta(x)$ *is continuous on* $\mathcal{X}$*.*

When one of the statements holds, $\{\varrho(x)\} = \{\nabla\theta(x)\} = \partial\theta(x)$ *for each* $x \in \mathcal{X}$*.*

Proof. [Proof of Corollary 6]

(ai) $\Leftrightarrow$ (aii) By assumption ψ satisfies Theorem 4(cii), i.e., $f(y) \in \partial\psi(y)$, $\forall\, y \in \mathcal{V}$. By Lemma 1 with $\varrho := f$ and the convex function $\theta := \psi$, f is $C^0(\mathcal{V})$ if and only if ψ is $C^1(\mathcal{V})$ and when one of these holds, $f = \nabla\psi$ on $\mathcal{V}$. This proves the result for $k = 0$. The extension to $k \geqslant 1$ is trivial.

(bi) $\Leftrightarrow$ (bii) Consider $x \in \mathcal{V}$. As $\mathcal{V}$ is open there is an open ball $\mathcal{B}_x$ such that $x \in \mathcal{B}_x \subset \mathcal{V}$. Noticing that $\mathcal{B}_x$ is polygonally connected, by Theorem 4-(b), there is $K \in \mathbb{R}$ such that $g(x') = \frac{1}{2}\|x'\|^2 + \varphi(x') + K$ for each $x' \in \mathcal{B}_x$. Hence g is $C^{k+1}(\mathcal{B}_x)$ if and only if φ is $C^{k+1}(\mathcal{B}_x)$, and $\nabla g(x') = x' + \nabla\varphi(x')$ on $\mathcal{B}_x$. As this holds for each $x \in \mathcal{V}$, the equivalence holds on $\mathcal{V}$.

(bii) $\Rightarrow$ (biii) By (bii), g is $C^{k+1}(\mathcal{X})$ hence $\partial g(x) = \{\nabla g(x)\}$ for each $x \in \mathcal{X}$. By Theorem 4(aii), $f^{-1}(x) \subset \partial g(x)$ for each $x \in \mathrm{Im}(f)$. Combining both facts yields

$$y = \nabla g(f(y)) \quad \forall\, y \in f^{-1}(\mathcal{X}). \tag{10}$$

Consider $y, y' \in f^{-1}(\mathcal{X})$ such that $f(y) = f(y')$. Then $y = \nabla g(f(y)) = \nabla g(f(y')) = y'$, which shows that f is injective on $f^{-1}(\mathcal{X})$. Consequently, $\widetilde{f}$ is a bijection between $f^{-1}(\mathcal{X})$ and $\mathcal{X}$, hence the inverse function $(\widetilde{f})^{-1}$ is well defined. Inserting $y = (\widetilde{f})^{-1}(x)$ into (10) yields $(\widetilde{f})^{-1}(x) = \nabla g(x)$ for each $x \in \mathcal{X}$. Then, since g is $C^{k+1}(\mathcal{X})$, it follows that $(\widetilde{f})^{-1}$ is $\mathcal{C}^k(\mathcal{X})$.

(biii) $\Rightarrow$ (bii) Consider $x \in \mathcal{X}$. As $\widetilde{f}$ is injective on $f^{-1}(\mathcal{X})$ by (biii), there is a unique $y \in f^{-1}(\mathcal{X})$ such that $x = f(y)$. Using that $f^{-1}(x) \subset \partial g(x)$ by Theorem 4(aii) shows that $(\widetilde{f})^{-1}(x) = y \in \partial g(x)$. Since $(\widetilde{f})^{-1}$ is $C^k(\mathcal{X})$, using Lemma 1 with $\varrho := (\widetilde{f})^{-1}$ and $\theta := g$ proves that

$$(\widetilde{f})^{-1}(x) = \nabla g(x) \quad \forall\, x \in \mathcal{X}$$

Since $(\widetilde{f})^{-1}$ is $C^k(\mathcal{X})$ it follows that g is $C^{k+1}(\mathcal{X})$. □

2.7. **Illustration using classical examples.** Theorem 1 and its corollaries characterize whether a function f is a proximity operator. This is particularly useful when f is not *explicitly built* as a proximity operator. We illustrate this with a few examples. We begin with $\mathcal{H} = \mathbb{R}$, where proximity operators happen to have a particularly simple characterization.

COROLLARY 7. *Let* $\mathcal{Y} \subset \mathbb{R}$ *be non-empty. A function* $f : \mathcal{Y} \to \mathbb{R}$ *is the proximity operator of some penalty* φ *if, and only if,* f *is nondecreasing.*

Proof. By Theorem 1 we just need to prove that a scalar function $f : \mathcal{Y} \to \mathbb{R}$ belongs to the sub-gradient of a convex function if, and only if, f is non-decreasing. When f is continuous and $\mathcal{Y}$ is an open interval, a primitive ψ of f is indeed convex if, and only if, $\psi' = f$ is non-decreasing [6, Proposition 17.7]. We now prove the result for more general $\mathcal{Y}$ and f. First, if $f(y) \in \partial\psi(y)$ for each $y \in \mathcal{Y}$ where $\psi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is convex, then by [21, Theorem 4.2.1 (i)] f is non-decreasing. To prove the converse define $a := \inf\{y : y \in \mathcal{Y}\}$, $I := (a, \infty)$ if $a \notin \mathcal{Y}$ (resp. $I := [a, \infty)$ if $a \in \mathcal{Y}$), and set $\bar{f}(x) := \sup_{y \in \mathcal{Y}, y \leqslant x} f(y) \in \mathbb{R} \cup \{+\infty\}$ for each $x \in I$, $\bar{f}(x) = +\infty$, for $x \notin I$. By construction $\bar{f}$ is non-decreasing. If f is non-decreasing on $\mathcal{Y}$ then $\bar{f}(y) = f(y)$ for each $y \in \mathcal{Y}$ hence $\mathcal{Y} \subset \mathrm{dom}(\bar{f}) \subset I$ and $\mathrm{dom}(\bar{f})$ is an interval. Choose an arbitrary $b \in \mathcal{Y}$. As $\bar{f}$ is monotone it is integrable on each bounded interval one can define $\psi(x) := \int_b^x \bar{f}(t)dt$ for each $x \in \mathrm{dom}(\bar{f})$ (with the usual convention that if $x < b$ then $\int_b^x = -\int_x^b$) and $\psi(x) := +\infty$ for $x \notin \mathrm{dom}(\bar{f})$. Consider $x \in \mathrm{dom}(\bar{f})$. Since $\bar{f}$ is non-increasing for $h \geqslant 0$ such that $x+h \in \mathrm{dom}(\bar{f})$ we have $\psi(x+h) - \psi(x) = \int_x^{x+h} \bar{f}(t)dt \geqslant \bar{f}(x)h$; similarly for $h \leqslant 0$ such that $x - h \in \mathrm{dom}(\bar{f})$ we have $\psi(x) - \psi(x-h) = \int_{x-h}^{x} \bar{f}(t)dt \leqslant \bar{f}(x)(-h)$, hence $\psi(x-h) - \psi(x) \geqslant \bar{f}(x)h$. Combining both results shows $\psi(y) - \psi(x) \geqslant \bar{f}(x)(y-x)$ for each $x, y \in \mathrm{dom}(\bar{f})$. This establishes that $\bar{f}(x) \in \partial\psi(x)$ for each $x \in \mathrm{dom}(\bar{f})$, hence that ψ is convex on its domain $\mathrm{dom}(\psi) = \mathrm{dom}(\bar{f})$. To conclude, simply observe that for $y \in \mathcal{Y} \subset \mathrm{dom}(\bar{f})$ we have $f(y) = \bar{f}(y) \in \partial\psi(y)$. □

EXAMPLE 2 (Quantization). *In* $\mathcal{Y} = [0,1) \subset \mathbb{R} = \mathcal{H}$*, consider* $0 = x_0 < x_1 < \ldots < x_{q-1} < x_q = 1$ *and* $v_0 \leqslant \ldots \leqslant v_{q-1}$*. Let* f *be the quantization-like function so that* $f(x) = v_i$ *if and only if* $x \in [x_i, x_{i+1})$*, for* $0 \leqslant i < q$*. Quantization traditionally corresponds to the case where* $q \geqslant 2$ *and for each* $0 \leqslant i < q-1$*,* x_{i+1} *is the middle point between* v_i *and* v_{i+1}*. Since* f *is non-decreasing,* f *is the proximity operator of a function* φ*. The image of* f *is the discrete set of points* $\{v_0, \ldots, v_{q-1}\}$*.*

Let us give another example to illustrate the role of the connectedness of the sets $\mathcal{C}$, $\mathcal{C}'$ in Theorem 4.

EXAMPLE 3. *Consider the identity function* $f(y) := y \mapsto y$ *on a subset* $\mathcal{Y} \subset \mathbb{R} = \mathcal{H}$*. Since* f *is increasing it is a proximity operator by Corollary 7. Particular functions satisfying the equivalent properties* (ai), (aii) *and* (aiii) *of Theorem 4 are* $\varphi_0 : x \mapsto 0$*,* $g_0 : x \mapsto x^2/2$ *and* $\psi_0 : y \mapsto y^2/2$*. They further satisfy* (7) *(resp.* (8)*) with* $K = K' = 0$ *on* $\mathbb{R}$*. When* $\mathcal{Y} \subset \mathbb{R}$ *is polygonally connected,* $\mathrm{Im}(f)$ *is also polygonally connected by the continuity of* f *and Theorem 4 implies that* φ_0, g_0, ψ_0 *are, up to global additive constants* K, K'*, the only functions satisfying* (ai), (aii) *and* (aiii)*. Now, consider as a particular example of disconnected set* $\mathcal{Y} = (-\infty, 0) \cup (1, +\infty)$*. We exhibit two other functions* g, ψ *such that* φ_0, g, ψ *also satisfy* (ai), (aii) *and* (aiii)*, but* (7) *fails on the disconnected set* $\mathcal{C} := \mathrm{Im}(f) = \mathcal{Y}$ *(resp.* (8) *fails on the disconnected set* $\mathcal{C}' := \mathcal{Y}$*). Intuitively, what happens is that the presence of a "hole" (the interval* $[0,1]$*) in* $\mathcal{Y}$ *gives some freedom in designing separately the components of these functions on each connected component. For this, consider* $H : [0,1] \to [0,1]$ *any continuous increasing function such that* $H(0) = 0$*,* $H(1) = 1$ *and* $C := \int_0^1 H(t)dt \neq 1/2$*. Observe that the function*

$$h(x) := \begin{cases} \frac{x^2}{2}, x < 0 \\ \int_0^x H(t)dt, 0 \leqslant x \leqslant 1 \\ \int_0^1 H(t)dt + \frac{x^2-1}{2}, x > 1. \end{cases} \quad .$$

is convex and satisfies $\partial h(x) = \{h'(x)\} = \{x\}$ *for each* $x \in \mathcal{Y}$*. As a result the functions* $g := h$ *and* $\psi := h$ *also satisfy properties* (aii) *and* (aiii) *of Theorem 4. Yet on the interval* $(-\infty, 0)$ *we have* $g(x) = g_0(x) = \varphi_0(x) + \frac{x^2}{2} + K_0$ *with* $K_0 = 0$*, while on the interval* $(1, +\infty)$ *we have* $g(x) = g_0(x) + C - 1/2 = \varphi_0(x) + \frac{x^2}{2} + K_1$ *with* $K_1 = C - 1/2 \neq 0 = K_0$*. Similarly* $\psi(x) - \psi_0(x)$ *is not constant on* $\mathcal{Y}$*. This shows that* (7) *(resp.* (8)*) fails to hold on* $\mathcal{C} := \mathrm{Im}(f)$ *(resp.* $\mathcal{C}' := \mathcal{Y}$*).*

Consider now functions $f : \mathbb{R}^n \to \mathbb{R}^n$ given by $f(y) = (f_i(y))_{i=1}^n$. When each f_i can be written as $f_i(y) = h_i(y_i)$, the function is said to be separable. If each h_i is a scalar proximity operator then the function f is also a proximity operator, and vice-versa. This can be seen, e.g., by writing $h_i = \mathrm{prox}_{\varphi_i}$ and $f = \mathrm{prox}_\varphi$ with $\varphi(x) := \sum_{i=1}^n \varphi_i(x_i)$. All examples below hold for the components of separable functions.

As recalled in Proposition 1 it is known [13, Proposition 2.4] that a function $f : \mathbb{R} \to \mathbb{R}$ is the proximity operator of a *convex l.s.c.* penalty φ if, and only if, f is nondecreasing and nonexpansive: $|f(y) - f(y')| \leqslant |y - y'|$ for each $y, y' \in \mathbb{R}$.

A particular example is that of scalar *thresholding rules* which are known [2, Proposition 3.2] to be the proximity operator of a (*continuous positive*) penalty function. As we will see in Section 3, Theorem 1 also allows to characterize whether certain *block-thresholding rules* [20, 10, 22] are proximity operators.

Our next example illustrates the functions appearing in Theorem 1 on the classical hard-thresholding operator, which is the proximity operator of a nonconvex function.

Example 4 (Hard-thresholding). *In* $\mathcal{Y} = \mathcal{H} = \mathbb{R}$ *consider* $\lambda > 0$ *and the weighted* ℓ^0 *penalty*

$$\varphi(x) := \begin{cases} 0, & \text{if } x = 0; \\ \lambda, & \text{otherwise.} \end{cases}$$

Its (set-valued) proximity operator is

$$\mathrm{prox}_\varphi(y) = \left\{ \begin{array}{ll} \{0\} & \text{if } \ |y| < \sqrt{2\lambda} \\ \{0, \sqrt{2\lambda}\} & \text{if } \ y = \sqrt{2\lambda} \\ \{-\sqrt{2\lambda}, 0\} & \text{if } \ y = -\sqrt{2\lambda} \\ \{y\} & \text{if } \ |y| > \sqrt{2\lambda} \end{array} \right.$$

which is discontinuous. Choosing $\pm\sqrt{2\lambda}$ *as the value at* $y = \pm\sqrt{2\lambda}$ *yields a function* $f(y) \in \mathrm{prox}_\varphi(y)$ *with disconnected (hence nonconvex) range* $\mathrm{Im}(f) = (-\infty, -\sqrt{2\lambda}] \cup \{0\} \cup [\sqrt{2\lambda}, +\infty)$*,*

$$f(y) := \begin{cases} 0, & \text{if } |y| < \sqrt{2\lambda} \\ y, & \text{if } |y| \geqslant \sqrt{2\lambda} \end{cases}$$

Since $\mathcal{Y}$ *is convex, the potential* ψ *is characterized by* (1)*. For* $K := 0$ *we get*

$$\psi(y) = yf(y) - \tfrac{1}{2}f^2(y) - \varphi(f(y)) = \begin{cases} 0, & \text{if } |y| < \sqrt{2\lambda} \\ y^2/2 - \lambda, & \text{otherwise} \end{cases} = \max(y^2/2 - \lambda, 0).$$

This is indeed a convex potential, and $f(y) \in \partial\psi(y)$ *for each* $y \in \mathbb{R}$*.*

Our last example of this section is a scaled version of soft-thresholding: it is still a proximity operator, however for $C > 1$ the corresponding penalty is nonconvex, and is even unbounded from below.

Example 5 (Scaled soft-thresholding). *In $\mathcal{Y} = \mathcal{H} = \mathbb{R}$ consider*

$$f(y) := \begin{cases} 0, & \text{if } |y| < 1 \\ C(y-1), & \text{if } y \geqslant 1 \\ C(y+1), & \text{if } y \leqslant -1 \end{cases} = Cy\max(1-1/|y|,0).$$

This function has the same shape as the classical soft-thresholding operator, but is scaled by a multiplicative factor C. When $C = 1$, f is the soft-thresholding operator which is the proximity operator of the absolute value, $\varphi(x) = |x|$, which is convex. For $C > 1$, as f is expansive, by Proposition 1 it cannot be the proximity operator of any convex function. Yet, as f is monotonically increasing, $f(y)$ is a subgradient of its "primitive" $\psi(y) = \frac{C}{2}(\max(|y|-1,0))^2 = \frac{C}{2}y^2(\max(1-1/|y|,0))^2 = \frac{f^2(y)}{2C}$ which is convex. Moreover, by Corollary 7, f is still the proximity operator of some (necessarily nonconvex) function $\varphi(x)$. By (1), up to an additive constant $K \in \mathbb{R}$, φ satisfies

$$\varphi(f(y)) = yf(y) - \tfrac{1}{2}f^2(y) - \psi(y) = yf(y) - \tfrac{1+C}{2C}f^2(y), \forall\, y \in \mathbb{R}$$

For $x > 0$, writing $x = f(y)$ with $y = f^{-1}(x) = 1 + x/C$ yields $\varphi(x) = \varphi(f(y)) = (1 + x/C)x - \frac{1+C}{2C}x^2$. Similar considerations for $x < 0$ and for $x = 0$ show that $\varphi(x) = |x| + \left(\frac{1}{C} - 1\right)\frac{x^2}{2}$. When $C > 1$, φ is indeed not bounded from below, and not convex.

3. When is social shrinkage a proximity operator ?

We conclude this paper by studying so-called social shrinkage operators, which have been introduced to mimic classical sparsity promoting proximity operators when certain types of structured sparsity are targeted. We show that the characterization of proximity operators obtained in this paper provides answers to questions raised by Kowalski et al [22] and by Varoqueaux et al [35] on the link between such non-separable shrinkage operators and proximity operators.

Most proximity operators are indeed not separable. A classical example is the proximity operator associated to mixed ℓ_{12} norms, which enforces group-sparsity.

Example 6 (Group-sparsity shrinkage). *Consider a partition $\mathcal{G} = \{G_1, \ldots, G_p\}$ of $[\![1, n]\!]$, the interval of integers from 1 to n, into disjoint sets called* groups. *Let x_G be the restriction of $x \in \mathbb{R}^n$ to its entries indexed by $G \in \mathcal{G}$, and define the* group ℓ^1 norm, *or* mixed ℓ_{12} norm, *as*

$$\varphi(x) := \sum_{G \in \mathcal{G}} \|x_G\|_2. \tag{11}$$

The proximity operator $f(y) := \mathrm{prox}_{\lambda\varphi}$ is the group-sparsity shrinkage *operator with threshold λ*

$$\forall i \in G, \quad f_i(y) := y_i\left(1 - \frac{\lambda}{\|y_G\|_2}\right)_+. \tag{12}$$

The group-LASSO penalty (11) appeared in statistics in the thesis of Bakin [4, Chapter 2]. It was popularized by Yuan and Lin [37] who introduced an iterative shrinkage algorithm to address the corresponding optimization problem. A generalization is Group Empirical Wiener / Group Non-negative Garrotte, see e.g. [15]

$$\forall i \in G, \quad f_i(y) := y_i \left(1 - \frac{\lambda^2}{\|y_G\|_2^2}\right)_+, \tag{13}$$

see also [2] for a review of thresholding rules, and [3] for a review on sparsity-inducing penalties.

To account for varied types of structured sparsity, [23, 24] empirically introduced the so-called Windowed Group-LASSO. A weighted version for audio applications was further developed in [32] which coins the notion of *persistency*, and the term *social sparsity* was coined in [22] to cover Windowed Group-LASSO, as well as other structured shrinkage operators. As further described in these papers, the main motivation of such social shrinkage operators is to obtain flexible ways of taking into account (possibly overlapping) *neighborhoods* of a coefficient index i rather than *disjoint groups* of indices to decide whether or not to set a coefficient to zero. These are summarized in the definition below.

Definition 3 (Social shrinkage). *Consider a family $N_i \subset [\![1,n]\!]$, $i \in [\![1,n]\!]$ of sets such that $i \in N_i$. The set N_i is called a* neighborhood *of its index i. Consider nonnegative weight vectors $w^i = (w^i_\ell)_{\ell=1}^n$ such that $\operatorname{supp}(w^i) = N_i$. Windowed Group Lasso (WG-LASSO) shrinkage is defined as $f(y) := (f_i(y))_{i=1}^n$ with*

$$\forall i, \quad f_i(y) := y_i \left(1 - \frac{\lambda}{\|\operatorname{diag}(w^i)y\|_2}\right)_+ \tag{14}$$

and Persistent Empirical Wiener (PEW) shrinkage (see [33] for the unweighted version) with

$$\forall i, \quad f_i(y) := y_i \left(1 - \frac{\lambda^2}{\|\operatorname{diag}(w^i)y\|_2^2}\right)_+. \tag{15}$$

Kowalski et al [22] write "*while the classical proximity operators*[6] *are directly linked to convex regression problems with mixed norm priors on the coefficients, [those] new, structured, shrinkage operators cannot be directly linked to a convex minimization problem*". Similarly, Varoquaux et al [35] write that Windowed Group Lasso "*is not the proximal operator of a known penalty*". They leave open the question of whether social shrinkage is the proximity operator of some yet to be discovered penalty. Using Theorem 2, we answer these questions for generalized social shrinkage operators. The answer is negative unless the involved neighborhoods form a partition.

Definition 4 (Generalized social shrinkage). *Consider subsets $N_i \subset [\![1,n]\!]$ and nonnegative weight vectors $w^i \in \mathbb{R}^n_+$ such that $i \in N_i$ and $\operatorname{supp}(w^i) = N_i$ for each $i \in [\![1,n]\!]$. Consider $\lambda > 0$ and a family of $C^1(\mathbb{R}^*_+)$ scalar functions h_i, $i \in [\![1,n]\!]$ such that $h_i'(t) \neq 0$ for $t \in \mathbb{R}^*_+$. A generalized social shrinkage operator is defined as $f(y) := (f_i(y))_{i=1}^n$ with*

$$f_i(y) := \begin{cases} y_i h_i\left(\|\operatorname{diag}(w^i)y\|_2^2\right), & \textit{if } \|\operatorname{diag}(w^i)y\|_2 > \lambda, \\ 0 & \textit{otherwise.} \end{cases}$$

[6]that are explicitly constructed as the proximity operator of a convex l.s.c. penalty, e.g., soft-thresholding.

We let the reader check that the above definition covers Group LASSO (12), Windowed Group-LASSO (14), Group Empirical Wiener (13) and Persistent Empirical Wiener shrinkage (15).

Lemma 2. *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a generalized social shrinkage operator and $N_i \subset [\![1, n]\!]$, $w^i \in \mathbb{R}^n_+$, $i \in [\![1, n]\!]$ be the corresponding families of neighborhoods and weight vectors. If f is a proximity operator then there exists a partition $\mathcal{G} = \{G_p\}_{p=1}^P$ of the set $[\![1, n]\!]$ of indices such that: for each p and all $i, j \in G_p$ we have $w^i = w^j$ and $\mathrm{supp}(w^i) = \mathrm{supp}(w^j) = G_p$. As a consequence for $i \in G_p$, $j \in G_q$ with $p \neq q$, the weight vectors w^i and w^j have disjoint support.*

The proof of Lemma 2 is postponed to Appendix A.7. An immediate consequence of this lemma is that if f is a generalized social shrinkage operator, then the neighborhood system $N_i = \mathrm{supp}(w^i)$ coincides with the groups G from the partition $\mathcal{G}$. In particular, the neighborhood system must form a partition. By contraposition we get the following corollary:

Corollary 8. *Consider non-negative weights $\{w^i\}$ as in Definition 4 and $\{N_i\}$ the corresponding neighborhood system. Assume that there exists i, j such that $N_i \neq N_j$ and $N_i \cap N_j \neq \varnothing$.*

- *Let f be the WG-LASSO shrinkage (14). There is no penalty φ such that $f = \mathrm{prox}_\varphi$.*
- *Let f be the PEW shrinkage (15). There is no penalty φ such that $f = \mathrm{prox}_\varphi$.*

In other words, WG-LASSO / PEW can be a proximity operator *only if* the neighborhood system has *no overlap*, i.e. with "plain" Group-LASSO (12) / Group Empirical Wiener (13).

Acknowledgements

The first author wishes to thank Laurent Condat, Jean-Christophe Pesquet and Patrick-Louis Combettes for their feedback that helped improve an early version of this paper, as well as the anonymous reviewers for many insightful comments that improved it much further.

Appendix A. Proofs

The proofs of technical results of Section 2 are provided in Sections A.4 (Theorem 3), A.5 (Lemma 1), A.6 (Corollary 3) and A.7 (Lemma 2). As a preliminary we give brief reminders on some useful but classical notions in Sections A.1-A.3.

A.1. **Brief reminders on (Fréchet) differentials and gradients in Hilbert spaces.** Consider $\mathcal{H}, \mathcal{H}'$ two Hilbert spaces. A function $\theta : \mathcal{X} \to \mathcal{H}'$ where $\mathcal{X} \subset \mathcal{H}$ is an open domain is (Fréchet) differentiable at x if there exists a continuous linear operator $L : \mathcal{H} \to \mathcal{H}'$ such that $\lim_{h\to 0} \|\theta(x+h) - \theta(x) - L(h)\|_{\mathcal{H}'}/\|h\|_{\mathcal{H}} = 0$. The linear operator L is called the differential of θ at x and denoted $D\theta(x)$. When $\mathcal{H}' = \mathbb{R}$, L belongs to the dual of $\mathcal{H}$, hence there is $u \in \mathcal{H}$ –called the gradient of θ at x and denoted $\nabla\theta(x)$– such that $L(h) = \langle u, h\rangle$, $\forall\, h \in \mathcal{H}$.

A.2. **Subgradients and subdifferentials for possibly nonconvex functions.** We adopt a gentle definition which is familiar when θ is a convex function. Although this is possibly less well-known by non-experts, this definition is also valid when θ is possibly nonconvex, see e.g. [6, Definition 16.1].

Definition 5. *Let $\theta : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a proper function. The* subdifferential $\partial\theta(x)$ *of* θ *at* x *is the set of all* $u \in \mathcal{H}$*, called* subgradients *of* θ *at* x*, such that*

$$\theta(x') \geqslant \theta(x) + \langle u, x' - x\rangle, \quad \forall x' \in \mathcal{H}. \tag{16}$$

If $x \notin \mathrm{dom}(\theta)$, then $\partial\theta(x) = \varnothing$. The function θ is subdifferentiable *at $x \in \mathcal{H}$ if $\partial\theta(x) \neq \varnothing$. The domain of $\partial\theta$ is $\mathrm{dom}(\partial\theta) := \{x \in \mathcal{H}, \partial\theta(x) \neq \varnothing\}$. It satisfies $\mathrm{dom}(\partial\theta) \subset \mathrm{dom}(\theta)$.*

Fact 1. *When $\partial\theta(x) \neq \varnothing$ the inequality in* (16) *is trivial for each $x' \notin \mathrm{dom}(\theta)$ since it amounts to $+\infty = \theta(x') - \theta(x) \geqslant \langle u, x' - x\rangle$.*

Definition 5 leads to the well-known Fermat's rule [6, Theorem 16.3]

Theorem 5. *Let $\theta : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a proper function. A point $x \in \mathrm{dom}(\theta)$ is a global minimizer of θ if and only if*

$$0 \in \partial\theta(x).$$

If θ has a global minimizer at x, then by Theorem 5 the set $\partial\theta(x)$ is non-empty. However, $\partial\theta(x)$ can be empty, e.g., at local minimizers that are not the global minimizer:

Example 7. Let $\theta(x) = \frac{1}{2}x^2 - \cos(\pi x)$. The global minimum of θ is reached at $x = 0$ where $\partial\theta(x) = f'(x) = 0$. At $x = \pm 1.7\bar{9}$ θ has local minimizers where $\partial\theta(x) = \varnothing$ (even though θ is $\mathcal{C}^\infty$). For $|x| < 0.53$ one has $\partial\theta(x) = \nabla\theta(x)$ with $\theta''(x) \geqslant 0$ and for $0.54 < |x| < 1.91$ $\partial\theta(x) = \varnothing$.

The proof of the following lemma is a standard exercice in convex analysis [6, Exercice 16.8].

Lemma 3. *Let $\theta : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a proper function such that (a) $\mathrm{dom}(\theta)$ is convex and (b) $\partial\theta(x) \neq \varnothing$ for each $x \in \mathrm{dom}(\theta)$. Then θ is a convex function.*

Definition 6. *(Lower convex envelope of a function)*
Let $\theta : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be proper with $\mathrm{dom}(\partial\theta) \neq \varnothing$. Its lower convex envelope,[7] *denoted $\breve{\theta}$, is the pointwise supremum of all the convex lower-semicontinuous functions minorizing θ*

$$\breve{\theta}(x) := \sup\{\varrho(x) \,|\, \varrho : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}, \varrho \text{ convex l.s.c.}, \ \varrho(z) \leqslant \theta(z), \forall\, z \in \mathcal{H}\}, \quad \forall\, x \in \mathcal{H}. \tag{17}$$

The function $\breve{\theta}$ is proper, convex and lower-semicontinuous. It satisfies

$$\breve{\theta}(x) \leqslant \theta(x), \forall\, x \in \mathcal{H}. \tag{18}$$

Proposition 3. *Let $\theta : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be proper with $\mathrm{dom}(\partial\theta) \neq \varnothing$. For any $x_0 \in \mathrm{dom}(\partial\theta)$ we have $\breve{\theta}(x_0) = \theta(x_0)$, $\partial\theta(x_0) = \partial\breve{\theta}(x_0)$.*

Proof. As $\partial\theta(x_0) \neq \varnothing$, by [6, Proposition 13.45], $\breve{\theta}$ is the so-called biconjugate θ^{**} of θ [6, Definition 13.1]. Moreover, [6, Proposition 16.5] yields $\theta^{**}(x_0) = \theta(x_0)$ and $\partial\theta^{**}(x_0) = \partial\theta(x_0)$. □

We need to adapt [6, Proposition 17.31] to the case where θ is proper but possibly nonconvex, with a stronger assumption of Fréchet (instead of Gâteaux) differentiability.

Proposition 4. *If $\partial\theta(x) \neq \varnothing$ and θ is (Fréchet) differentiable at x then $\partial\theta(x) = \{\nabla\theta(x)\}$.*

[7] also known as convex hull, [29, p. 57],[21, Definition 2.5.3]

Proof. Consider $u \in \partial\theta(x)$. As θ is differentiable at x there is an open ball $\mathcal{B}$ centered at 0 such that $x+h \in \mathrm{dom}(\theta)$ for each $h \in \mathcal{B}$. For each $h \in \mathcal{B}$, Definition 5 yields

$$\theta(x-h)-\theta(x) \geqslant \langle u,-h\rangle \qquad \text{and} \qquad \theta(x+h)-\theta(x) \geqslant \langle u,h\rangle$$

hence $-(\theta(x-h)-\theta(x)) \leqslant \langle u,h\rangle \leqslant \theta(x+h)-\theta(x)$. Since θ is Fréchet differentiable at x, letting $\|h\|$ tend to zero yields

$$-\left(\langle\nabla\theta(x),-h\rangle+o(\|h\|)\right) \leqslant \langle u,h\rangle \leqslant \langle\nabla\theta(x),h\rangle+o(\|h\|)$$

hence $\langle u-\nabla\theta(x),h\rangle = o(\|h\|)$, $\forall\, h \in \mathcal{B}$. This shows that $u = \nabla\theta(x)$. □

A.3. **Characterizing functions with a given subdifferential.** Corollary 9 below generalizes a result of Moreau [26, Proposition 8.b] characterizing functions by their subdifferential. It shows that one only needs the subdifferentials to intersect. We begin in dimension one.

Lemma 4. *Consider $a_0, a_1 : \mathbb{R} \to \mathbb{R}\cup\{+\infty\}$ convex functions such that* $\mathrm{dom}(a_i) = \mathrm{dom}(\partial a_i) = [0,1]$ *and $\partial a_0(t)\cap\partial a_1(t) \neq \varnothing$ on $[0,1]$. Then there exists a constant $K \in \mathbb{R}$ such that $a_1(t) - a_0(t) = K$ on $[0,1]$.*

Proof. As a_i is convex it is continuous on $(0,1)$ [21, Theorem 3.1.1, p16]. Moreover, by [21, Proposition 3.1.2] we have $a_i(0) \geqslant \lim_{t\to 0, t>0} a_i(t) =: a_i(0_+)$, and since $\partial a_i(0) \neq \varnothing$, there is $u_i \in \partial a_i(0)$ such that $a_i(t) \geqslant a_i(0) + u_i(t-0)$ for each $t \in [0,1]$ hence $a_i(0_+) \geqslant a_i(0)$. This shows that $a_i(0_+) = a_i(0)$, and similarly $\lim_{t\to 1, t<1} a_i(t) = a_i(1)$, hence a_i is continuous on $[0,1]$ relatively to $[0,1]$. In addition, a_i is differentiable on $[0,1]$ except on a countable set $B_i \subset [0,1]$ [21, Theorem 4.2.1 (ii)].

For $t \in [0,1]\backslash(B_0\cup B_1)$ and $i \in \{0,1\}$, Proposition 4 yields $\partial a_i(t) = \{a_i'(t)\}$ hence the function $\delta := a_1 - a_0$ is continuous on $[0,1]$ and differentiable on $[0,1]\backslash(B_0\cup B_1)$. For $t \in I\backslash(B_0\cup B_1)$, $\{a_0'(t)\}\cap\{a_1'(t)\} = \partial a_0(t)\cap\partial a_1(t) \neq \varnothing$, hence $a_0'(t) = a_1'(t)$ and $\delta'(t) = 0$. A classical exercice[8] in real analysis [34, Example 4] is to show that if a function f is continuous on an interval, and differentiable with zero derivative except on a countable set, then f is constant. As $B_0\cup B_1$ is countable it follows δ is constant on $(0,1)$. As it is continuous on $[0,1]$, it is constant on $[0,1]$. □

Corollary 9. *Let $\theta_0, \theta_1 : \mathcal{H} \to \mathbb{R}\cup\{+\infty\}$ be proper and $\mathcal{C} \subset \mathcal{H}$ a non-empty polygonally connected set. Assume that for each $z \in \mathcal{C}$, $\partial\theta_0(z)\cap\partial\theta_1(z) \neq \varnothing$; then there is a constant $K \in \mathbb{R}$ such that $\theta_1(x) - \theta_0(x) = K$, $\forall\, x \in \mathcal{C}$.*

Remark 4. Note that the functions θ_i and the set $\mathcal{C}$ are not assumed to be convex.

Proof. The proof is in two parts.

(i) Assume that $\mathcal{C}$ is convex and fix some $x^* \in \mathcal{C}$. Consider $x \in \mathcal{C}$, and define $a_i(t) := \theta_i(x^* + t(x-x^*))$, for $i = 0,1$ and each $t \in [0,1]$, and $a_i(t) = +\infty$ if $t \notin [0,1]$. As $\mathcal{C}$ is convex,

[8] for a proof see e.g. (in french) `https://fr.wikipedia.org/wiki/Lemme_de_Cousin` section 4.9, version from 13/01/2019.

$z_t := x^* + t(x - x^*) \in \mathcal{C}$ hence for each $t \in [0,1]$ there exists $u_t \in \partial\theta_0(z_t) \cap \partial\theta_1(z_t)$. By Definition 5 for each $t, t' \in [0,1]$,

$$a_i(t') - a_i(t) = \theta_i(x^* + t'(x - x^*)) - \theta_i(x^* + t(x - x^*)) \geqslant \langle u_t, (t'-t)(x-x^*)\rangle = \langle u_t, x - x^*\rangle(t'-t).$$

For $t \in [0,1]$ and $t' \in \mathbb{R}\backslash[0,1]$ since $a_i(t') = +\infty$ the inequality $a_i(t') - a_i(t) \geqslant \langle u_t, x - x^*\rangle(t'-t)$ also obviously holds, hence $\langle u_t, x - x^*\rangle \in \partial a_i(t)$, $i = 0, 1$. Thus $\partial a_i(t) \neq \varnothing$ for each $t \in [0,1]$, so by Lemma 3 a_i is convex on $[0,1]$ for $i = 0,1$, and $\langle u_t, x - x^*\rangle \in \partial a_0(t) \cap \partial a_1(t)$ for each $t \in [0,1]$. By Lemma 4, there exists $K \in \mathbb{R}$ such that $a_1(t) - a_0(t) = K$ for each $t \in [0,1]$. Therefore,

$$\theta_1(x) - \theta_0(x) = a_1(1) - a_0(1) = a_1(0) - a_0(0) = \theta_1(x^*) - \theta_0(x^*) = K.$$

As this holds for each $x \in \mathcal{C}$, we have established the result as soon as $\mathcal{C}$ is convex.

(ii) Now we prove the result when $\mathcal{C}$ is polygonally connected. Fix some $x^* \in \mathcal{C}$ and define $K := \theta_1(x^*) - \theta_0(x^*)$. Consider $x \in \mathcal{C}$: by the definition of polygonal connectedness, there exists an integer $n \geqslant 1$ and $x_j \in \mathcal{C}$, $0 \leqslant j \leqslant n$ with $x_0 = x^*$ and $x_n = x$ such that the (convex) segments $\mathcal{C}_j = [x_j, x_{j+1}] = \{tx_j + (1-t)x_{j+1}, t \in [0,1]\}$ satisfy $\mathcal{C}_j \subset \mathcal{C}$. Since each $\mathcal{C}_j$ is convex, the result established in (i) implies that $\theta_1(x_{j+1}) - \theta_0(x_{j+1}) = \theta_1(x_j) - \theta_0(x_j)$ for $0 \leqslant j < n$. This shows that $\theta_1(x) - \theta_0(x) = \theta_1(x_n) - \theta_0(x_n) = \ldots = \theta_1(x_0) - \theta_0(x_0) = \theta_1(x^*) - \theta_0(x^*) = K$. □

A.4. **Proof of Theorem 3.** The indicator function of a set $\mathcal{S}$ is denoted

$$\chi_{\mathcal{S}}(x) := \begin{cases} 0 & \text{if} \quad x \in \mathcal{S}, \\ +\infty & \text{if} \quad x \notin \mathcal{S}. \end{cases}$$

(ai) ⇒ (aii). We introduce the function $\theta : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ by

$$\theta := b + \varphi + \chi_{\mathrm{Im}(f)}. \tag{19}$$

Consider $x \in \mathrm{Im}(f)$. By definition $x = f(y)$ where $y \in \mathcal{Y}$, hence by (ai) x is a global minimizer of $x' \mapsto \{D(x', y) + \varphi(x')\}$. Therefore, we have

$$\forall\, x' \in \mathcal{H},\ -\langle A(y), x'\rangle + \underbrace{b(x') + \varphi(x') + \chi_{\mathrm{Im}(f)}(x')}_{=\theta(x')} \geqslant -\langle A(y), x\rangle + \underbrace{b(x) + \varphi(x) + \chi_{\mathrm{Im}(f)}(x)}_{=\theta(x)} \tag{20}$$

which is equivalent to

$$\forall\, x' \in \mathcal{H} \qquad \theta(x') \geqslant \theta(x) + \langle A(y), x' - x\rangle \tag{21}$$

meaning that $A(y) \in \partial\theta\,(f(y))$. As this holds for each $y \in \mathcal{Y}$ such that $f(y) = x$, we get $A(f^{-1}(x)) \subset \partial\theta(x)$. Consider $g_1 := \breve{\theta}$ according to Definition 6. Since g_1 is convex l.s.c. and

$$\forall\, x \in \mathrm{Im}(f), \partial\theta(x) \neq \varnothing, \tag{22}$$

by Proposition 3, $\partial\theta(x) = \partial g_1(x)$ and $\theta(x) = g_1(x)$ for each $x \in \mathrm{Im}(f)$. This establishes (aii) with $g := g_1 = \breve{\theta}$.

(aii) $\Rightarrow$ (ai). Set $\theta_1 := g + \chi_{\mathrm{Im}(f)}$. By (aii), $\partial g(x) \neq \varnothing$ for each $x \in \mathrm{Im}(f)$. Since $\mathrm{dom}(\partial g) \subset \mathrm{dom}(g)$ it follows that $\mathrm{Im}(f) \subset \mathrm{dom}(g)$ and consequently

$$\mathrm{dom}(\theta_1) = \mathrm{Im}(f).$$

Consider $y \in \mathcal{Y}$ and $x := f(y)$ so that $x \in \mathrm{Im}(f)$, hence $\theta_1(x) = g(x)$ and $A(y) \in A(f^{-1}(x)) \subset \partial g(x)$ where the inclusion comes from (aii). It follows that for each $(x, x') \in \mathrm{Im}(f) \times \mathcal{H}$ one has

$$\theta_1(x') = g(x') + \chi_{\mathrm{Im}(f)}(x') \geqslant g(x') \geqslant g(x) + \langle A(y), x' - x\rangle = \theta_1(x) + \langle A(y), x' - x\rangle,$$

showing that $A(y) \in \partial\theta_1(x)$. This is equivalent to (21) with $\theta := \theta_1$, and since $\mathrm{dom}(\theta_1) = \mathrm{Im}(f)$, the inequality in (20) holds with $\varphi(x) := \theta_1(x) - b(x)$, i.e., x is a global minimizer of $D(x', y) + \varphi(x')$. Since this holds for each $y \in \mathcal{Y}$, this establishes (ai) with $\varphi := \theta_1 - b = g - b + \chi_{\mathrm{Im}(f)}$.

(b). Consider φ and g satisfying (ai) and (aii), respectively. Let[9] $g_1 := \breve{\theta}$ with θ defined in (19). Following the arguments of (ai) $\Rightarrow$ (aii) we obtain that g_1 (just as g) satisfies (aii). For each $x \in \mathcal{C}$ we thus have $\partial g(x) \cap \partial g_1(x) \supset A(f^{-1}(x)) \neq \varnothing$ with g, g_1 convex l.s.c. functions. Hence, by Corollary 9, since $\mathcal{C}$ is polygonally connected, there is a constant K such that $g(x) = g_1(x) + K$, $\forall\, x \in \mathcal{C}$. To establish the relation (2) between g and φ we now show that $g_1(x) = b(x) + \varphi(x)$ on $\mathcal{C}$. By (22) and Proposition 3 we have $\breve{\theta}(x) = \theta(x)$ for each $x \in \mathrm{Im}(f)$, hence as $\mathcal{C} \subset \mathrm{Im}(f)$ we obtain $g_1(x) := \breve{\theta}(x) = \theta(x) = b(x) + \varphi(x)$ for each $x \in \mathcal{C}$. This establishes (2).

(ci) $\Rightarrow$ (cii). Define

$$\varrho(y) \;:=\; \begin{cases} +\infty, & \forall y \notin \mathcal{Y} \\ \langle B(f(y)), y\rangle - b(f(y)) - \varphi(f(y)), & \forall y \in \mathcal{Y}. \end{cases} \tag{23}$$

Consider $y \in \mathcal{Y}$. From (ci), for each y' the global minimizer of $x \mapsto \widetilde{D}(x, y') + \varphi(x)$ is reached at $x' = f(y')$. Hence, for $x = f(y)$ we have

$$-\langle B(f(y')), y'\rangle + b(f(y')) + \varphi(f(y')) \leqslant -\langle B(x), y'\rangle + b(x) + \varphi(x) = -\langle B(f(y)), y'\rangle + b(f(y)) + \varphi(f(y))$$

Using this inequality we obtain that

$$\begin{aligned} \forall\, y' \in \mathcal{Y},\ \varrho(y') - \varrho(y) \;&=\; -\langle B(f(y)), y\rangle + b(f(y)) + \varphi(f(y)) + \langle B(f(y')), y'\rangle - b(f(y')) - \varphi(f(y')) \\ &\geqslant\; \langle B(f(y)), y'\rangle - \langle B(f(y)), y\rangle \geqslant \langle B(f(y)), y' - y\rangle \end{aligned}$$

This shows that

$$B(f(y)) \in \partial\varrho(y). \tag{24}$$

Set $\psi_1 := \breve{\varrho}$ according to Definition 6. Then the function ψ_1 is convex l.s.c. and for each $y \in \mathcal{Y}$ the function $B(f(y))$ is well defined, so $\partial\varrho(y) \neq \varnothing$. Hence, by Proposition 3, $\partial\varrho(y) = \partial\breve{\varrho}(y) = \partial\psi_1(y)$ and $\varrho(y) = \breve{\varrho}(y) = \psi_1(y)$ for each $y \in \mathcal{Y}$. This establishes (cii) with $\psi := \psi_1 = \breve{\varrho}$.

[9] In general, we may have $g \neq g_1$ as there is no connectedness assumption on $\mathrm{dom}(\theta)$.

(cii) $\Rightarrow$ (ci). Define $h : \mathcal{Y} \to \mathbb{R}$ by

$$h(y) := \langle B(f(y)), y\rangle - \psi(y)$$

Since $B(f(y')) \in \partial\psi(y')$ with ψ convex by (cii), applying Definition 5 to $\partial\psi$ yields $\psi(y)-\psi(y') \geqslant \langle y - y', B(f(y'))\rangle$. Using this inequality, one has

$$\forall\, y, y' \in \mathcal{Y} \qquad \begin{aligned} h(y') - h(y) &= \langle B(f(y')), y'\rangle - \psi(y') - \langle B(f(y)), y\rangle + \psi(y) \\ &\geqslant \langle B(f(y')), y'\rangle - \langle B(f(y)), y\rangle + \langle B(f(y')), y - y'\rangle \\ &= \big\langle B(f(y') - B(f(y)),\ y\big\rangle \end{aligned} \tag{25}$$

Noticing that for each $x \in \mathrm{Im}(f)$ there is $y \in \mathcal{Y}$ such that $x = f(y)$, we can define $\theta : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ obeying $\mathrm{dom}(\theta) = \mathrm{Im}(f)$ by

$$\theta(x) := \begin{cases} h(y) \quad \text{with } y \in f^{-1}(x) & \text{if } x \in \mathrm{Im}(f) \\ +\infty & \text{otherwise} \end{cases}$$

For $x \in \mathrm{Im}(f)$, as $f(y) = f(y') = x$ for each $y, y' \in f^{-1}(x)$, applying (25) yields $h(y')-h(y) \geqslant 0$. By symmetry $h(y') = h(y)$, hence the definition of $\theta(x)$ does not depend of which $y \in f^{-1}(x)$ is chosen.

For $x' \in \mathrm{Im}(f)$ we write $x' = f(y')$. Using (25) and the definition of θ yields

$$\theta(x')-\theta(f(y)) = \theta(f(y'))-\theta(f(y)) = h(y')-h(y) \geqslant \langle B(f(y'))-B(f(y)), y\rangle = \langle B(x')-B(f(y)), y\rangle.$$

that is to say

$$\theta(x') - \langle B(x'), y\rangle \geqslant \theta(f(y)) - \langle B(f(y)), y\rangle, \quad \forall\, x' \in \mathrm{Im}(f).$$

This also trivially holds for $x' \notin \mathrm{Im}(f)$. Setting $\varphi(x) := \theta(x)-b(x)$ for each $x \in \mathcal{H}$, and replacing θ by $b + \varphi$ in the inequality above yields

$$a(y) - \langle B(x'), y\rangle + b(x') + \varphi(x') \geqslant a(y) - \langle B(f(y)), y\rangle + b(f(y)) + \varphi(f(y)), \quad \forall\, x' \in \mathcal{H}$$

showing that $f(y) \in \arg\min_{x'}\{\widetilde{D}(x', y) + \varphi(x')\}$. As this holds for each $y \in \mathcal{Y}$, φ satisfies (ci).

(d). Consider φ and ψ satisfying (ci) and (cii), respectively. Using the arguments of (ci) $\Rightarrow$ (cii), the function $\psi_1 := \breve{\varrho}$ with ϱ defined in (23) satisfies (cii). As ψ and ψ_1 both satisfy (cii), for each $y \in \mathcal{C}'$ we have $\partial\psi(y) \cap \partial\psi_1(y) \supset B(f(y)) \neq \varnothing$ with ψ, ψ_1 convex l.s.c. functions. Hence, by Corollary 9, since $\mathcal{C}'$ is polygonally connected, there is a constant K' such that $\psi(y) = \psi_1(y)+K'$, $\forall\, y \in \mathcal{C}'$. By (24), $\partial\varrho(y) \neq \varnothing$ for each $y \in \mathcal{Y}$, hence by Proposition 3 we have $\breve{\varrho}(y) = \varrho(y)$ for each $y \in \mathcal{Y}$. As $\mathcal{C}' \subset \mathcal{Y}$, it follows that $\psi_1(y) = \breve{\varrho}(y) = \varrho(y)$ for each $y \in \mathcal{C}'$. This establishes (3).

A.5. **Proof of Lemma 1.** *Proof*. Without loss of generality we prove the equivalence for the convex envelope $\breve{\theta}$ instead of θ: indeed by Proposition 3, since $\partial\theta(x) \neq \varnothing$ on $\mathcal{X}$ we have $\breve{\theta}(x) = \theta(x)$ and $\partial\breve{\theta}(x) = \partial\theta(x)$ on $\mathcal{X}$.

(a)$\Rightarrow$(b). By [6, Prop 17.41(iii)$\Rightarrow$(i)], as $\breve{\theta}$ is convex l.s.c. and ϱ is a selection of its subdifferential which is continuous at each $x \in \mathcal{X}$, $\breve{\theta}$ is (Frchet) differentiable at each $x \in \mathcal{X}$. By Proposition 4 we get $\partial\breve{\theta}(x) = \{\nabla\breve{\theta}(x)\} = \{\varrho(x)\}$ on $\mathcal{X}$. Since ϱ is continuous, $x \mapsto \nabla\breve{\theta}(x)$ is continuous on $\mathcal{X}$.

(b)$\Rightarrow$(a). Since $\breve{\theta}$ is differentiable on $\mathcal{X}$, by Proposition 4 we have $\partial\breve{\theta}(x) = \{\nabla\breve{\theta}(x)\}$ on $\mathcal{X}$. By (9) it follows that $\varrho(x) = \nabla\breve{\theta}(x)$ on $\mathcal{X}$. Since $\nabla\breve{\theta}$ is continuous on $\mathcal{X}$, so is ϱ. □

A.6. **Proof of Corollary 3.** By Theorem 2, as $\mathcal{Y}$ is open and convex and f is $C^1(\mathcal{Y})$ with $Df(y)$ symmetric semi-definite positive for each $y \in \mathcal{Y}$, there is a function φ_0 and a convex l.s.c. function $\psi \in C^2(\mathcal{Y})$ such that $\nabla\psi(y) = f(y) \in \text{prox}_{\varphi_0}(y)$ for each $y \in \mathcal{Y}$. We define $\varphi(x) := \varphi_0(x) + \chi_{\text{Im}(f)}(x)$ and let the reader check that $f(y) \in \text{prox}_{\varphi}(y)$ for each $y \in \mathcal{Y}$. By construction, $\text{dom}(\varphi) = \text{Im}(f)$.

Uniqueness of the global minimizer. Consider $\widetilde{f}$ any function such that $\widetilde{f}(y) \in \text{prox}_{\varphi}(y)$ for each y. This implies

$$(26)\quad \tfrac{1}{2}\|y - f(y)\|^2 + \varphi(f(y)) = \tfrac{1}{2}\|y - \widetilde{f}(y)\|^2 + \varphi(\widetilde{f}(y)) = \min_{x\in\mathcal{H}}\{\tfrac{1}{2}\|y - x\|^2 + \varphi(x)\}, \quad \forall\, y \in \mathcal{Y}.$$

By Corollary 1 there is a convex l.s.c. function $\widetilde{\psi}$ such that $\widetilde{f}(y) \in \partial\widetilde{\psi}(y)$ for each $y \in \mathcal{Y}$. Since $\mathcal{Y}$ is convex it is polygonally connected hence by Theorem 4(b) and (26) there are $K, K' \in \mathbb{R}$ such that

$$(27)\quad \psi(y) - K = \tfrac{1}{2}\|y\|^2 - \tfrac{1}{2}\|y - f(y)\|^2 - \varphi(f(y)) = \tfrac{1}{2}\|y\|^2 - \tfrac{1}{2}\|y - \widetilde{f}(y)\|^2 - \varphi(\widetilde{f}(y)) = \widetilde{\psi}(y) - K', \quad \forall\, y \in \mathcal{Y}.$$

Thus, $\widetilde{\psi}$ is also $C^2(\mathcal{Y})$ and $\widetilde{f}(y) \in \partial\widetilde{\psi}(y) = \{\nabla\psi(y)\} = \{f(y)\}$ for each $y \in \mathcal{Y}$. This shows that $\widetilde{f}(y) = f(y)$ for each y, hence $f(y)$ is the unique global minimizer on $\mathcal{H}$ of $x \mapsto \frac{1}{2}\|y - x\|^2 + \varphi(x)$, i.e., $\text{prox}_{\varphi}(y) = \{f(y)\}$.

Injectivity of f. The proof follows that of [17, Lemma 1]. Given $y \neq y'$ define $v := y' - y \neq 0$ and $\theta(t) := \langle f(y + tv), v\rangle$ for $t \in [0, 1]$. As $\mathcal{Y}$ is convex this is well defined. As $f \in \mathcal{C}^1(\mathcal{Y})$ and $Df(y + tv) \succ 0$, the function θ is $C^1([0, 1])$ with $\theta'(t) = \langle Df(y + tv)\ v, v\rangle > 0$ for each t. If we had $f(y) = f(y')$ then by Rolle's theorem there would be $t \in [0, 1]$ such that $\theta'(t) = 0$, contradicting the fact that $\theta'(t) > 0$.

Differentiability of φ. If $Df(y)$ is boundedly invertible for each $y \in \mathcal{Y}$, then by the inverse function theorem $\text{Im}(f)$ is open and $f^{-1} : \text{Im}(f) \to \mathcal{Y}$ is C^1. Given $x \in \text{Im}(f)$, denoting $u := f^{-1}(x)$, (27) yields

$$\varphi(x) = \varphi(f(u)) = -(\psi(u) - K) + \tfrac{1}{2}\|u\|^2 - \tfrac{1}{2}\|u - f(u)\|^2 = -(\psi(f^{-1}(x)) - K) + \tfrac{1}{2}\|f^{-1}(x)\|^2 - \tfrac{1}{2}\|f^{-1}(x) - x\|^2.$$

Since ψ is C^2 and f^{-1} is C^1, it follows that φ is C^1.

Global minimum is the unique critical point. The proof is inspired by that of [17, Theorem 1]. Consider x a critical point of $\theta : x \mapsto \frac{1}{2}\|y - x\|^2 + \varphi(x)$, i.e., since φ is C^1, a point where $\nabla\theta(x) = 0$. Since $\text{dom}(\varphi) = \text{Im}(f)$ there is some $v \in \mathcal{Y}$ such that $x = f(v)$. Moreover, as φ is C^1 on the open set $\text{Im}(f)$, the gradient $\nabla\theta(x)$ is well defined and $\nabla\theta(x) = 0$. On the one hand, denoting $\varrho(u) := (\theta \circ f)(u) = \frac{1}{2}\|y - f(u)\|^2 + \varphi(f(u))$ we have $\nabla\varrho(u) = Df(u)\nabla\theta(f(u))$ for each $u \in \mathcal{Y}$. On the other hand, for each $u \in \mathcal{Y}$, as $f(u) = \nabla\psi(u)$ we also have

$$\begin{aligned}
\varrho(u) &= \tfrac{1}{2}\|y\|^2 + \tfrac{1}{2}\|f(u)\|^2 - \langle y, f(u)\rangle + \varphi(f(u)) \\
&= +\tfrac{1}{2}\|y\|^2 + \langle u - y, f(u)\rangle - (\psi(u) - K), \\
\nabla\varrho(u) &= Df(u)\ (u - y) + f(u) - \nabla\psi(u) = Df(u)\ (u - y)
\end{aligned}$$

For $u = v$ we get $Df(v)\ (v - y) = \nabla\varrho(v) = Df(v)\nabla\theta(f(v)) = Df(v)\ \nabla\theta(x) = 0$. As $Df(v) \succ 0$, this implies $v = y$, hence $x = f(y)$.

A.7. **Proof of Lemma 2.** As a preliminary let us compute the entries of the $n \times n$ matrix associated to $Df(y)$:

$$\forall\, i,j \in [\![1,n]\!] \quad \frac{\partial f_i}{\partial y_j}(y) = \begin{cases} 0 & \text{if } \|\mathrm{diag}(w^i)y\|_2 < \lambda \\ 2(w^i_j)^2 y_i y_j h'_i\left(\|\mathrm{diag}(w^i)y\|_2^2\right) & \text{if } \|\mathrm{diag}(w^i)y\|_2 > \lambda \end{cases} \tag{28}$$

NB: if $\|\mathrm{diag}(w^i)y\|_2 = \lambda$ then f may not be differentiable at y; this case will not be useful below.

The proof exploits Corollary 2 which shows that if f is a proximity operator then $Df(y)$ is symmetric in each open set where it is well defined.

Let f be a generalized social shrinkage operator as described in Lemma 2 and consider $\mathcal{G} = \{G_1, \ldots, G_p\}$ the partition of $[\![1,n]\!]$ into disjoint groups corresponding to the equivalence classes defined by the equivalence relation between indices: for $i,j \in [\![1,n]\!]$, $i \sim j$ if and only if $w^i = w^j$. Given $G \in \mathcal{G}$, denote w^G the weight vector shared all $i \in G$. If f is a proximity operator then we show that for each $G \in \mathcal{G}$, we have $\mathrm{supp}(w^G) = G$.

For $i \in G$, by Definition 4 we have $i \in N_i = \mathrm{supp}(w^i) = \mathrm{supp}(w^G)$, establishing that[10]

$$G \subset \mathrm{supp}(w^G). \tag{29}$$

From now on we assume that f is a proximity operator, and consider a group $G \in \mathcal{G}$. To prove that $G = \mathrm{supp}(w^G)$, we will establish that for each $i,j \in [\![1,n]\!]$

$$\text{if there exists } y \in \mathbb{R}^n \text{ such that } \|\mathrm{diag}(w^j)y\|_2 \neq \|\mathrm{diag}(w^i)y\|_2 \text{ then } w^i_j = 0 \text{ and } w^j_i = 0. \tag{30}$$

To see why it allows to conclude, consider $j \in \mathrm{supp}(w^G)$, and $i \in G$. As $N_i := \mathrm{supp}(w^i) = \mathrm{supp}(w^G)$ we obtain that $j \in N_i$, i.e., $w^i_j \neq 0$. By (30), it follows that $\|\mathrm{diag}(w^j)y\|_2 = \|\mathrm{diag}(w^i)y\|_2$ for each y. As w^i, w^j have non-negative entries, this means that $w^i = w^j$. As $i \in G$, this implies $j \in G$ by the very definition of G as an equivalence class. This shows $\mathrm{supp}(w^G) \subset G$. Using also (29), we conclude that $\mathrm{supp}(w^G) = G$.

Let us now prove (30). Consider a given pair $i,j \in [\![1,n]\!]$. Assume that $\|\mathrm{diag}(w^j)y\|_2 \neq \|\mathrm{diag}(w^i)y\|_2$ for at least one vector y. Without loss of generality assume that $a := \|\mathrm{diag}(w^j)y\|_2 < \|\mathrm{diag}(w^i)y\|_2 =: b$. Rescaling y by a factor $c = 2\lambda/(a+b)$ yields the existence of y such that for the considered pair i,j

$$\|\mathrm{diag}(w^j)y\|_2 < \lambda < \|\mathrm{diag}(w^i)y\|_2. \tag{31}$$

By continuity, perturbing y if needed we can also assume that for this pair i,j we have $y_i y_j \neq 0$.

By (28), as (31) holds in a neighborhood of y, f is C^1 at y and its partial derivatives for the considered pair i,j satisfy

$$\frac{\partial f_i}{\partial y_j}(y) = 2(w^i_j)^2 y_i y_j h'_i\left(\|\mathrm{diag}(w^i)y\|_2^2\right) \quad \text{and} \quad \frac{\partial f_j}{\partial y_i}(y) = 0.$$

Since f is a proximity operator, by Corollary 2 we have $\frac{\partial f_i}{\partial y_j}(y) = \frac{\partial f_j}{\partial y_i}(y)$. It follows that for the considered pair i,j

$$(w^i_j)^2 y_i y_j h'_i\left(\|\mathrm{diag}(w^i)y\|_2^2\right) = 0.$$

As $y_i y_j \neq 0$ and $h'_i(t) \neq 0$ for $t \neq 0$, we obtain $w^i_j = 0$.

[10]The inclusion (29) is true even if f is not a proximity operator.

To conclude we now show that $w_i^j = 0$. As $w_j^i = 0$, f_i is in fact independent of y_j and $\frac{\partial f_i}{\partial y_j}$ is *identically zero* on $\mathbb{R}^n$. By scaling y as needed, we get a vector y' such that $y'_i y'_j \neq 0$ and

$$\lambda < \|\mathrm{diag}(w^j) y'\|_2 < \|\mathrm{diag}(w^i) y'\|_2.$$

Reasoning as above yields $2(w_i^j)^2 y'_j y'_i h'_j \left(\|\mathrm{diag}(w^j)y'\|_2^2\right) = \frac{\partial f_j}{\partial y_i}(y') = \frac{\partial f_i}{\partial y_j}(y') = 0$, hence $w_i^j = 0$. We thus obtain that $w_j^i = w_i^j = 0$ as claimed, establishing (30) and therefore $G = \mathrm{supp}(w^G)$.

References

[1] Madhu Advani and Surya Ganguli. An equivalence between high dimensional Bayes optimal inference and M-estimation. In D D Lee, M Sugiyama, U V Luxburg, I Guyon, and R Garnett, editors, *Advances in Neural Information Processing Systems*, pages 3378–3386. Curran Associates, Inc., 2016.

[2] Anestis Antoniadis. Wavelet methods in statistics: Some recent developments and their applications. *Statistics Surveys*, 1(0):16–55, 2007.

[3] Francis Bach. Optimization with Sparsity-Inducing Penalties. *FNT in Machine Learning*, 4(1):1–106, 2011.

[4] Sergey Bakin. *Adaptive regression and model selection in data mining problems*. PhD thesis, School of Mathematical Sciences, Australian National University, 1999.

[5] Heinz H Bauschke, Jonathan M Borwein, and Patrick L Combettes. Bregman Monotone Optimization Algorithms. *SIAM J. Control and Optimization*, 42(2):596–636, 2003.

[6] Heinz H Bauschke and Patrick L Combettes. Convex Analysis and Monotone Operator Theory in Hilbert Spaces – Second Edition. Springer International Publishing, Cham, 2017.

[7] Thomas Blumensath and Michael E Davies. Iterative hard thresholding for compressed sensing. *Appl. Comp. Harm. Anal.*, 27(3):265–274, November 2009.

[8] Kristian Bredies, Dirk A Lorenz, and Stefan Reiterer. Minimization of Non-smooth, Non-convex Functionals by Iterative Thresholding. *Journal of Optimization Theory and Applications*, 165(1):78–112, July 2014.

[9] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

[10] T Tony Cai and Bernard W Silverman. Incorporating information on neighbouring coefficients into wavelet estimation. *Sankhyà: The Indian Journal of Statistics, Series B (1960-2002)*, 63(Special Issue on Wavelets):127–148, 2001.

[11] H. Cartan. *Cours de calcul différentiel*. Collection Méthodes. Editions Hermann, 1977.

[12] Y. Censor and S. A. Zenios. Proximal minimization algorithm with d-functions. *J. Optim. Theory Appl.*, 73(3):451–464, 1992.

[13] Patrick L Combettes and Jean-Christophe Pesquet. Proximal Thresholding Algorithm for Minimization over Orthonormal Bases. *SIAM J. Optim.*, 18:1351–1376, 2007.

[14] I Ekeland and T Turnbull. *Infinite-dimensional optimization and convexity*. Chicago Lectures in Mathematics. The University of Chicago Press, 1983.

[15] Cédric Févotte and Matthieu Kowalski. Hybrid sparse and low-rank time-frequency signal decomposition. *EUSIPCO*, pages 464–468, 2015.

[16] Antonio Galbis and Manuel Maestre. *Vector Analysis Versus Vector Calculus*. Universitext. Springer US, Boston, MA, 2012.

[17] Remi Gribonval. Should Penalized Least Squares Regression be Interpreted as Maximum A Posteriori Estimation? *IEEE Transactions on Signal Processing*, 59(5):2405–2410, 2011.

[18] Remi Gribonval and Pierre Machart. Reconciling "priors" and "priors" without prejudice? In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 2193–2201, 2013.

[19] Rémi Gribonval and Mila Nikolova. On bayesian estimation and proximity operators. *Applied and Computational Harmonic Analysis*, pages 1–25, 2019.

[20] Peter Hall, Spiridon I Penev, Gerard Kerkyacharian, and Dominique Picard. Numerical performance of block thresholded wavelet estimators. *Statistics and Computing*, 1997.

[21] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and Minimization Algorithms, vol. I.* Springer-Verlag, Berlin, 1996.

[22] Matthieu Kowalski, Kai Siedenburg, and Monika Dörfler. Social Sparsity! Neighborhood Systems Enrich Structured Shrinkage Operators. *IEEE Trans. Signal Processing*, 2013.

[23] Matthieu Kowalski and Bruno Torrésani. Sparsity and persistence: mixed norms provide simple signal models with dependent coefficients. *Signal, Image and Video Processing*, 3(3):251–264, 2009.

[24] Matthieu Kowalski and Bruno Torrésani. Structured Sparsity: from Mixed Norms to Structured Shrinkage. In Rémi Gribonval, editor, *SPARS'09 - Signal Processing with Adaptive Sparse Structured Representations*, Saint Malo, France, April 2009. Inria Rennes - Bretagne Atlantique.

[25] Cécile Louchet and Lionel Moisan. Posterior Expectation of the Total Variation Model: Properties and Experiments. *SIAM J. Imaging Sci.*, 6(4):2640–2684, January 2013.

[26] Jean-Jacques Moreau. Proximité et dualité dans un espace Hilbertien. *Bull. Soc. math. France*, 93:273–299, 1965.

[27] M Nikolova. Estimation of binary images by minimizing convex criteria. In *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269)*, pages 108–112. IEEE Comput. Soc, 1998.

[28] Ankit Parekh and Ivan W Selesnick. Convex Denoising using Non-Convex Tight Frame Regularization. *IEEE Signal Process. Lett.*, 2015.

[29] R Tyrrell Rockafellar and Roger J B Wets. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998.

[30] Ralph Rockafellar. On the maximal monotonicity of subdifferential mappings. *Pacific Journal of Mathematics*, 33(1):209–216, 1970.

[31] Ivan W Selesnick. Sparse Regularization via Convex Analysis. *IEEE Trans. Signal Processing*, 65(17):4481–4494, 2017.

[32] K Siedenburg and M Dörfler. Structured sparsity for audio signals. In *Proc. 14th Int. Conf. on Digital Audio Effects (DAFx-11)*, Paris, 2011.

[33] Kai Siedenburg, Matthieu Kowalski, and Monika Dörfler. Audio declipping with social sparsity. In *ICASSP 2014 - 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1577–1581. IEEE, 2014.

[34] Brian S Thomson. Rethinking the Elementary Real Analysis Course. *The American Mathematical Monthly*, 2007.

[35] Gaël Varoquaux, Matthieu Kowalski, and Bertrand Thirion. Social-sparsity brain decoders: faster spatial sparsity. In *International Workshop on Pattern Recognition in Neuroimaging*, Trento, 2016.

[36] C Villani. *Optimal Transport - Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften - A series of Comprehensive Studies in Mathematics*. Springer-Verlag Berlin Heidelberg, 2009.

[37] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, February 2006.