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.

1. Introduction and overview

Proximity operators have become an important ingredient of non-smooth 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 that the ℓ^0 penalty.

This paper extends existing characterizations of proximity operators – which are specialized for convex penalties– to the noconvex 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 known as minimum mean square error estimation or MMSE) can be expressed as a proximity operator. This is the object of a companion paper [17] characterizing when certain variational approaches to address inverse problems can in fact be considered as Bayesian approaches.

This work and the companion paper [17] 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, Univ Rennes, Inria, CNRS, IRISA, remi.gribonval@inria.fr;
M. Nikolova, CMLA, CNRS and Ecole Normale Supérieure de Cachan, Université Paris-Saclay, 94235 Cachan, France.

1.1. **Characterization of proximity operators.** The proximity operator of a function $\varphi : \mathbb{R}^n \to \mathbb{R}$ is a mapping from $y \in \mathbb{R}^n$ to the set of solutions of a penalized least-squares problem

$$y \mapsto \operatorname{prox}_\varphi(y) := \arg\min_{x\in\mathbb{R}^n} \left\{ \tfrac{1}{2}\|y - x\|^2 + \varphi(x) \right\}$$

Formally, a proximity operator is set-valued as there may be several solutions to this problem, or the set of solutions may be empty. Informally, a function f is often said to be "the" proximity operator of φ if $f(y) \in \operatorname{prox}_\varphi(y)$ for any y. A primary example is soft-thresholding $f(y) := y(1 - 1/y)_+$, $y \in \mathbb{R}$, which is the proximity operator of the absolute value function $\varphi(x) := |x|$.

Proximity operators can be defined for certain generalized functions $\varphi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ where $\mathcal{H}$ is a Hilbert space equipped with a norm $\|\cdot\|$. A particular example is the projection onto a given convex set $C \subset \mathcal{H}$, which can be written as $\operatorname{proj}_C = \operatorname{prox}_\varphi$ with φ the indicator function of C, i.e., $\varphi(x) = 0$ if $x \in C$, $\varphi(x) = +\infty$ otherwise.

A characterization of proximity operators *of convex lower semi-continuous (lsc) functions* is due to Moreau. It involves the subdifferential $\partial\theta(x)$ of a convex lsc function θ at x, i.e., the set of all its subgradients at x [13, Chapter III.2][1].

PROPOSITION 1. [25, Corollary 10.c] *A function $f : \mathcal{H} \to \mathcal{H}$ defined everywhere is the proximity operator of a* proper convex lsc *function φ if, and only if the following conditions hold jointly:*

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

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

We establish the following extension to possibly nonconvex functions φ on subdomains of $\mathcal{H}$.

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 φ (i.e. $f(y) \in \operatorname{prox}_\varphi(y)$ for any $y \in \mathcal{Y}$) if, and only if there exists a convex lsc function $\psi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ such that for any $y \in \mathcal{Y}$, $f(y) \in \partial\psi(y)$.*

Theorem 1 is proved as a corollary of Theorem 3 (Section 2) characterizing functions such that $f(y) \in \arg\min_{x\in\mathcal{H}}\{D(x, y) + \varphi(x)\}$ for certain types of data-fidelity terms $D(x, y)$. The proof uses an adaptation of the usual notion of subgradient / subdifferential in the context of nonconvex functions. When the domain $\mathcal{Y}$ is convex, Theorem 3 implies that there is a number $K \in \mathbb{R}$ such that the functions f, φ and ψ in Theorem 1 satisfy

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

Among others, the data-fidelity terms covered by Theorem 3 include

- the least squares data fidelity term $\frac{1}{2}\|y - x\|^2$;
- its variant $D(x, y) = \frac{1}{2}\|y - Lx\|^2$ with L some linear operator; and
- Bregman divergences [7].

An analog of Theorem 1 characterizes so-called Bregman proximity operators [10] (Corollary 6).

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

A consequence of Theorem 3 is that for the considered data-fidelity terms $D(x,y)$, if a function $f:\mathcal{Y}\to\mathcal{H}$ 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 (Corollary 8)

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

This is reminiscent of observations on convex optimization with nonconvex penalties [27, 30] and on the hidden convexity of conditional mean estimation under additive Gaussian noise [15, 16, 23, 1]. The latter is extended to other noise models in the companion paper [17].

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, as established in Section 2.4 (Corollary 7). This leads to a characterization of *continuous* proximity operators[2].

Corollary 1. *Let $\mathcal{Y}\subset\mathcal{H}$ be open and $f:\mathcal{Y}\to\mathcal{H}$ be C^0. The following are equivalent:*
(a) *f is a proximity operator of a function φ (i.e. $f(y)\in\mathrm{prox}_{\varphi}(y)$ for any $y\in\mathcal{Y}$);*
(b) *there exists a convex $C^1(\mathcal{Y})$ function ψ such that $f(y)=\nabla\psi(y)$ for any $y\in\mathcal{Y}$.*

Next, we characterize C^1 proximity operators on convex domains more explicitly.

Theorem 2. *Let $\mathcal{Y}\subset\mathcal{H}$ be 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 φ (i.e. $f(y)\in\mathrm{prox}_{\varphi}(y)$ for any $y\in\mathcal{Y}$);*
(b) *there exists a convex $C^2(\mathcal{Y})$ function ψ such that $f(y)=\nabla\psi(y)$ for all $y\in\mathcal{Y}$;*
(c) *the differential $Df(y)$ is a* symmetric positive semi-definite *operator[3] for any $y\in\mathcal{Y}$.*

Corollary 2. *Let $\mathcal{Y}\subset\mathcal{H}$ be non-empty[4] and $f:\mathcal{Y}\to\mathcal{H}$ be a proximity operator. If f is C^1 in a neighborhood of $y\in\mathcal{Y}$, 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. □

Differential 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 stationary point of the corresponding optimization problem.

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

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

[4]*not necessarily* convex or open.

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 $\operatorname{prox}_\varphi(y) = \{f(y)\}$, $\forall\, y \in \mathcal{Y}$ and $\operatorname{dom}(\varphi) = \operatorname{Im}(f)$. Moreover, the only stationary point[5] of $x \mapsto \frac{1}{2}\|y - x\|^2 + \varphi(x)$ is $x = f(y)$.*

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 [29, Theorem B].

1.3. **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, beginning with $\mathcal{H} = \mathbb{R}$.

COROLLARY 4. *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. Apply Theorem 1 and observe that a scalar function f belongs to the sub-gradient of a convex function if, and only if, f is non-decreasing. □

EXAMPLE 1 (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 function so that $f(x) = v_i$ if and only if $x \in [x_i, x_{i+1})$, for $0 \leqslant i < q$. Since f is non-decreasing, f is the proximity operator of a function φ. Its image is the discrete set of points $\{v_0, \ldots, v_{q-1}\}$.*

1.3.1. *Separable examples.* Consider 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 $h_i = \operatorname{prox}_{\varphi_i}$ and $f = \operatorname{prox}_\varphi$ 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 [11, Proposition 2.4] that a function $f : \mathbb{R} \to \mathbb{R}$ is the proximity operator of a *convex lsc* penalty φ if, and only if, f is nondecreasing and nonexpansive: $|f(y) - f(y')| \leqslant |y - y'|$ for any $y, y' \in \mathbb{R}$.

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 1.3.2, Theorem 1 also allows to characterize whether certain *block-thresholding rules* [18, 8, 20] are proximity operators.

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

EXAMPLE 2 (Hard-thresholding). *In $\mathcal{Y} = \mathcal{H} = \mathbb{R}$ consider $\lambda > 0$ and the weighted ℓ^0 penalty*

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

[5] u is a stationary point of $\varrho : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ if $\nabla \varrho(u) = 0$; then ϱ is proper on a neighborhood of u.

Its (set-valued) proximity operator is

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

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* $\mathrm{Im}(f) = (-\infty, -\sqrt{2\lambda}] \cup \{0\} \cup [\sqrt{2\lambda}, +\infty)$,

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

The potential ψ *defined by* (1) *with* $K := 0$ *is*

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

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

Our second example 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 3 (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"* $\psi(y) = \frac{C}{2}\left(\max(|y| - 1, 0)\right)^2 = \frac{C}{2}y^2\left(\max(1 - 1/|y|, 0)\right)^2 = \frac{f^2(y)}{2C}$ *which is convex. Moreover, by Corollary 4,* 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.*

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

EXAMPLE 4 (Group-sparsity shrinkage). *Consider a partition of* $[\![1,n]\!]$ *into disjoint sets* $G \in \mathcal{G}$ *called* groups. *Let* x_G *be the restriction of* $x \in \mathbb{R}^n$ *to its entries indexed by* G*, and define the* group ℓ^1 norm*, or* mixed ℓ_{12} norm*, as*

$$\varphi(x) := \sum_{G\in\mathcal{G}} \|x_G\|_2. \tag{2}$$

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

$$\forall i \in G, \quad f_i(y) := y_i \left(1 - \frac{\lambda}{\|y_G\|_2}\right)_+ . \tag{3}$$

The group-LASSO penalty (2) appeared in statistics in the thesis of Bakin [4, Chapter 2]. It was popularized by Yuan and Lin [35] 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. [14]

$$\forall i \in G, \quad f_i(y) := y_i \left(1 - \frac{\lambda^2}{\|y_G\|_2^2}\right)_+ , \tag{4}$$

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

1.4. **Social shrinkage is generally not a proximity operator.** To account for varied types of structured sparsity, [21, 22] empirically introduced the so-called Windowed Group-LASSO. A weighted version for audio applications was further developed in [31] which coins the notion of *persistency*, and the term *social sparsity* was coined in [20] to cover Windowed Group-LASSO, as well as other structured shrinkage operators that take into account (possibly overlapping) *neighborhoods* of a coefficient index i rather than *groups* of indices to decide whether or not to set a coefficient to zero. These are summarized in the definition below.

DEFINITION 1 (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* $\mathrm{supp}(w^i) = N_i$*. Windowed Group Lasso (WG-LASSO) shrinkage is defined as* $f(y) := (f_i(y))_{i=1}^n$ *with*

$$\forall i, \quad f_i(y) := y_i \left(1 - \frac{\lambda}{\|\mathrm{diag}(w^i)y\|_2}\right)_+ \tag{5}$$

and Persistent Empirical Wiener (PEW) shrinkage (see [32] *for the unweighted version) with*

$$\forall i, \quad f_i(y) := y_i \left(1 - \frac{\lambda^2}{\|\mathrm{diag}(w^i)y\|_2^2}\right)_+ . \tag{6}$$

Kowalski et al [20] write "while the classical proximity operators[6] are directly linked to convex regression problems with mixed norm priors on the coefficients, the new, structured, shrinkage

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

operators can not be directly linked to a convex minimization problem". Similarly, Varoquaux et al [33] 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 2 (Generalized social shrinkage). *Consider a family of nonnegative weight vectors $w^i \in \mathbb{R}^n_+$, $i \in [\![1,n]\!]$ with $w^i_i \neq 0$ and define $N_i = \operatorname{supp}(w^i)$ the neighborhood of index i defined by w^i. 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*

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

We let the reader check that the above definition covers Group LASSO (3), Windowed Group-LASSO (5), Group Empirical Wiener (4) and Persistent Empirical Wiener shrinkage (6).

LEMMA 1. *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a generalized social shrinkage operator and $w^i \in \mathbb{R}^n_+$, $i \in [\![1,n]\!]$ be the corresponding family of weight vectors. Consider the partition of $[\![1,n]\!]$ into disjoint groups $G \in \mathcal{G}$ defined by the equivalence relation between indices: for $i,j \in [\![1,n]\!]$, $i \sim j$ if and only if $w^i = w^j$. Denote w^G the weight vector shared all $i \in G$.*

If f is a proximity operator then, for any $G \in \mathcal{G}$, we have $\operatorname{supp}(w^G) = G$.

The proof is postponed to Appendix A.9.

COROLLARY 5. *Consider non-negative weights $\{w^i\}$ as in Definition 2 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* (5)*. There is no penalty φ such that $f = \operatorname{prox}_\varphi$.*
- *Let f be the PEW shrinkage* (6)*. There is no penalty φ such that $f = \operatorname{prox}_\varphi$.*

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 (3) / Group Empirical Wiener (4).

1.5. **Discussion.** In light of our extension to nonconvex penalties of Moreau's characterization of proximity operators of convex (lsc) 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 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., [34]), 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

In this section we prove Theorem 1 advertized in Section 1, using the following theorem. The most techical proofs are postponed to the Appendix.

2.1. **Detailed notations.** 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 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., $\mathrm{dom}(\theta) \neq \varnothing$. It is lower semi-continuous (lsc) if for any $x_0 \in \mathcal{H}$, $\liminf_{x \to x_0} \theta(x) \geqslant \theta(x_0)$, 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[7] 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 [9, p. 327]. The gradient of a C^1 scalar function θ at x is denoted $\nabla\theta(x)$.

2.2. **Main theorem.**

Theorem 3. *Consider $\mathcal{H}$ and $\mathcal{H}'$ two Hilbert spaces[8], 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 $\mathrm{Im}(f)$ 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 all $y \in \mathcal{Y}$;*
   (ii) *there is a convex lsc $g : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ such that $A(f^{-1}(x)) \subset \partial g\,(x)$ for all $x \in \mathrm{Im}(f)$;*

(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*

$$g(x) \;=\; b(x) + \varphi(x) + K, \quad \forall\, x \in \mathcal{C}. \tag{7}$$

(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 all $y \in \mathcal{Y}$;*
   (ii) *there is a convex lsc $\psi : \mathcal{H}' \to \mathbb{R} \cup \{+\infty\}$ such that $B(f(y)) \in \partial\psi(y)$ for all $y \in \mathcal{Y}$.*

(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*

$$\psi(y) \;=\; \langle B(f(y)), y \rangle - b(f(y)) - \varphi(f(y)) + K', \quad \forall\, y \in \mathcal{C}'. \tag{8}$$

The proof of Theorem 3 is posponed to Appendix A.5. As it relies on a variant of the usual notion of subgradients / subdifferentials adapted to possibly nonconvex functions, we introduce this variant together with some relevant notations and useful lemmas in Appendix A.2.

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

[8] 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 $A : \mathcal{Y} \to \mathcal{H}^\star$, $B : \mathcal{H} \to (\mathcal{H}')^\star$.

Before diving into these technical elements we first detail some consequences of Theorem 3 for classical data-fidelity terms, and discuss some examples. This includes a proof of Theorem 1 in Section 2.3.1, and considerations on the possible convexity of the underlying optimization problem even when the penalty φ is nonconvex in Section 2.5. Finally, we consider how the smoothness of the functions f, φ, ψ and g in the theorems are interrelated in Section 2.4 (Corollary 7), and give the proof of Corollary 1 and the outline of that of Theorem 2.

2.3. **Main characterizations - proof of Theorem 1.**

2.3.1. *Standard proximity operators.* Consider $D(x,y) = \widetilde{D}(x,y) = \frac{1}{2}\|y-x\|^2$. Theorem 3 with $\mathcal{H}' = \mathcal{H}$, $a(y) := \frac{1}{2}\|y\|^2$, $b(x) := \frac{1}{2}\|x\|^2$, $A(y) := y$, and $B(x) := x$ yields the following theorem.

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 all $y \in \mathcal{Y}$;*
   (ii) *there is a convex lsc $g : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ such that $f^{-1}(x) \subset \partial g\,(x)$ for all $x \in \operatorname{Im}(f)$;*
   (iii) *there is a convex lsc $\psi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ such that $f(y) \in \partial\psi(y)$ for all $y \in \mathcal{Y}$.*

(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*

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

Theorem 1 immediately follows from the equivalence (ai)-(aiii).

2.3.2. *Extension to Bregman proximity operators.* The squared Euclidean norm is a particular *Bregman divergence*, and Theorem 3 also characterizes generalized proximity operators defined with such divergences. The Bregman divergence, known also as D-function, was introduced in [7] 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 3. *Let $h : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function that is differentiable on* $\operatorname{dom}(h)$. *The Bregman divergence (associated with h) between x and y is defined by*

$$D_h : \mathcal{H} \times \mathcal{H} \to [0, +\infty) : (x,y) \to h(x) - h(y) - \langle \nabla h(y), x - y\rangle. \tag{9}$$

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

$$a(y) := \langle \nabla h(y), y\rangle - h(y) \quad b(x) := h(x) \qquad \text{and} \quad A(y) = \nabla h(y) \tag{10}$$

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 $x \notin \operatorname{dom}(h)$ and, for $x \in \operatorname{dom}(h)$ and any $y \in \mathcal{H}$,

$$a(y) := h(y) \qquad b(x) := \langle \nabla h(x), x\rangle - h(x) \quad \text{and} \quad B(x) = h(x) \tag{11}$$

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 [10]. 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 $y \mapsto \arg\min_{x\in\mathcal{H}}\{D_h(y,x)+\varphi(x)\}$.

COROLLARY 6. *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* $\mathrm{dom}(h)$. *Let D_h read as in* (9).

(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 lsc $g : \mathcal{H} \to \mathbb{R}\cup\{+\infty\}$ s.t. $\nabla h(f^{-1}(x)) \subset \partial g\,(x)$, $\forall\, x \in \mathrm{Im}(f)$;*

(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 lsc $\psi : \mathcal{H} \to \mathbb{R}\cup\{+\infty\}$ such that $\nabla h(f(y)) \in \partial\psi(y)$, $\forall\, y \in \mathcal{Y}$.*

(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 (10). Further, (c) and (d) use (11). □

The Bregman divergence can be symmetric for certain choices of h. In such a case, both characterizations of Corollary 6 hold simultaneously. Consider for example $h(y) := \langle y, Ly\rangle$ with L some symmetric positive semi-definite linear operator, which yield $\nabla h(y) = Ly$ and $D_h(x,y) = D_h(y,x) = \langle x-y, L(x-y)\rangle$. In this case, the Corollary 6-(ai) and Corollary 6-(ci) are trivially equivalent, and Corollary 6-(aii)/(cii) characterizes g and ψ via the properties $Lf^{-1}(x) \subset \partial g(x)$ and $Lf(y) \in \partial\psi(y)$.

2.3.3. *Generalized proximity operators in the context of linear inverse problems.* In the context of linear inverse problems one often encounters optimization problems involving functions expressed as $\frac{1}{2}\|y - Lx\|^2 + \varphi(x)$ with L some linear operator. Such functions also fit into the framework of Theorem 3(a)/(b) using $a(y) := \frac{1}{2}\|y\|^2$, $b(x) := \frac{1}{2}\|Lx\|^2$, $A(y) := L^\star y$, and $B(x) := Lx$, where $L^\star$ is the adjoint of L. Therefore, $f : \mathcal{Y} \to \mathcal{H}$ is a generalized proximity operator of this type for some penalty φ if, and only if, there is a convex lsc ψ such that $Lf(y) \in \partial\psi(y)$ for all $y \in \mathcal{Y}$.

2.4. **Local smoothness of proximity operators – proof of Corollary 1- proofsketch of Theorem 2.** 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 smoothness of these functions are tightly inter-related. The proof is in Section A.6.

COROLLARY 7. *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

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

The characterization of any *continuous* proximity operator f as the gradient of a C^1 convex potential ψ, i.e., $f = \nabla\psi$, stated in Corollary 1, Section 1, is a direct consequence of Corollary 7 and Theorem 1.

For smooth (C^1) proximity operators, an even more explicit characterization holds: the fact that $f = \nabla\psi$ where ψ is convex and C^2 (by Corollary 7) can be characterized through the differential of f, as expressed in Theorem 2.

In the finite-dimensional setting $\mathcal{H} = \mathbb{R}^n$, the proof of Theorem 2 combines known results about C^1 vector fields that are gradient flows (also known as *conservative vector fields*) with the characterization of C^2 convex potentials by the non-negativity of their Hessian, see e.g. [12, Exercice 2A.5] or [5, Section 3.1.4]. Indeed, if $f(y) \in \mathrm{prox}_{\varphi}(y)$ for all $y \in \mathcal{Y}$ then, by Theorem 1, there is a convex potential $\psi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ such that $f(y) \in \partial\psi(y)$ for all $y \in \mathcal{Y}$. Since f is C^1, by Corollary 7 the potential ψ is C^2 and $f(y) = \nabla\psi(y)$, i.e., $f_i(y) = \frac{\partial}{\partial y_i}\psi(y)$, $1 \leqslant i \neq n$, and the Jacobian matrix associated to the differential $Df(y)$ is

$$J(y) = \left(\tfrac{\partial^2}{\partial y_i \partial y_j}\psi(y)\right)_{1\leqslant i,j\leqslant n},$$

i.e., the Jacobian of f is the Hessian matrix of ψ. Since ψ is C^2, by Schwarz's theorem, its Hessian is symmetric. Since ψ is convex, its Hessian is positive semi-definite. This establishes Theorem 2(a)⇒(b)⇒(c).

For the converse, we use that $\mathcal{Y}$ is convex and therefore *simply connected*. It is known[9] that if the Jacobian of a C^1 vector field f is everywhere symmetric positive semi-definite on a simply connected domain, then f is conservative, hence the existence of a C^2 potential ψ such that $f = \nabla\psi$. Finally, since the Jacobian of f matches the Hessian of ψ, semi-definite positiveness means that ψ is convex.

As stated in Theorem 2, the result is in fact valid in an arbitrary Hilbert space. The proof of Theorem 2 in such a setting follows the same steps as in the finite dimensional case, with proper adaptations of each ingredient. As we could not locate proper proofs of these ingredients in infinite dimension, the details are provided in Appendix A.6.

[9]A well known case is in dimension $n = 3$, where the Jacobian matrix $J(y)$ of f is everywhere symmetric if and only if f is curl-free – this is a common equivalent characterization of conservative fields on simply connected domains of $\mathbb{R}^3$.

2.5. **Convexity in proximity operators of nonconvex penalties.** Another 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 8. *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 all $y \in \mathcal{Y}$, with $D(x, y) = a(y) - \langle x, A(y)\rangle + b(x)$ as in Theorem 3*(a). *Then*

(a) *the function $b(x) + \varphi(x)$ is convex on any 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, $\forall\, y \in \mathcal{Y}$.*

Proof. (a) follows from Theorem 3(a)-(b). Claim (b) follows from claim (a) and the definition of D. □

Corollary 8(b) might seem surprising as, given a nonconvex penalty φ, one may expect the optimization problem $\min_x \frac{1}{2}\|y - x\|^2 + \varphi(x)$ to be non-convex. However, as noticed e.g. by [26, 27, 30], there are nonconvex penalties such that this problem is in fact convex. Corollary 8 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 [15] to be proximity operators. Extensions of this phenomenon for conditional mean estimation with other noise models are discussed in the companion paper [17].

APPENDIX A. PROOFS

A.1. **Brief reminders on (Fréchet) differentials and gradients in 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 $\mathcal{H}' = \mathcal{R}$, 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 $L(h) = \langle u, h\rangle,\ \forall\, h$.

A.2. **Subgradients for possibly nonconvex functions.** We adopt a gentle definition which is familiar when θ is a convex function:

DEFINITION 4. *Let $\theta : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a proper function. The* (non-local) subdifferential *$\bar{\partial}\theta(x)$ of θ at x is the set of all $u \in \mathcal{H}$, called* (non-local) subgradient*s of θ at x, such that*

$$\theta(x') - \theta(x) \geqslant \langle u, x' - x\rangle, \quad \forall x' \in \mathcal{H} \tag{12}$$

If $x \notin \mathrm{dom}(\theta)$, then $\bar{\partial}\theta(x) = \varnothing$.

The function θ is (non-locally) subdifferentiable *at $x \in \mathcal{H}$ if $\bar{\partial}\theta(x) \neq \varnothing$.*

Fact 1. *When $\bar{\partial}\theta(x) \neq \varnothing$ the inequality in* (12) *is trivial for any $x' \notin \mathrm{dom}(\theta)$ since it amounts to $+\infty = \theta(x') - \theta(x) \geqslant \langle u, x' - x\rangle$.*

Fact 2. *If θ is a convex function, the set $\bar{\partial}\theta(x)$ described by Definition 4 is the (usual) subdifferential of θ at x and is denoted by $\partial\theta(x)$, i.e., $\bar{\partial}\theta(x) = \partial\theta(x)$* [13, Chapter III.2].

REMARK 1. Definition 4 appears in lecture notes by Boyd et al [6], where $\bar\partial\theta(x)$ is called a subdifferential. This terminology when θ is nonconvex seems nonstandard, and we could not find it used in any reference textbook. We add the "(nonlocal)" prefix and use the $\bar\partial$ notation to highlight that, contrary to a gradient which definition only involves the local behaviour of θ around x, by definition $\bar\partial\theta(x)$ depends on the properties of $\theta(x')$ everywhere in $\mathcal{H}$.

We consider only global minimizers of (possibly nonconvex) proper functions.

DEFINITION 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*

$$\theta(x') - \theta(x) \geqslant 0 \quad \forall\, x' \in \mathcal{H}$$

or equivalently, if and only if

$$\theta(x') - \theta(x) \geqslant \langle 0, x' - x\rangle \quad \forall\, x' \in \mathcal{H}$$

This, together with Definition 4 leads to the following claim[10]:

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*

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

If θ has a global minimizer at x, then by Theorem 5 the set $\bar\partial\theta(x)$ is nonempty. However, the set $\bar\partial\theta(x)$ can be empty, e.g., at local minimizers that are not the global minimizer, as shown by the following example.

EXAMPLE 5. Let $\theta(x) = \frac{1}{2}x^2 - \cos(\pi x)$. The global minimum of θ is reached at $x = 0$ where $\bar\partial\theta(x) = f'(x) = 0$. At $x = \pm 1.7\bar 9$ θ has local minimizers where $\bar\partial\theta(x) = \varnothing$ (even though θ is $\mathcal{C}^\infty$). For $|x| < 0.53$ one has $\bar\partial\theta(x) = \nabla\theta(x)$ with $\theta''(x) \geqslant 0$ and for $0.54 < |x| < 1.91$ $\bar\partial\theta(x) = \varnothing$.

LEMMA 2. *Let $\theta : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a proper function such that (a) $\mathrm{dom}(\theta)$ is convex and (b) $\bar\partial\theta(x) \neq \varnothing$ for any $x \in \mathrm{dom}(\theta)$. Then θ is a convex function and thus $\bar\partial \equiv \partial$ is the usual subdifferential.*

Proof. Let $x, y \in \mathrm{dom}(\theta)$. For $\lambda \in [0,1]$ set $z = x + \lambda(y - x)$. Then $z \in \mathrm{dom}(\theta)$ by (a) and there is $u \in \bar\partial\theta(z) \neq \varnothing$ by (b). Using Definition 4,

$$\begin{aligned} \theta(x) - \theta(z) &\geqslant \langle u, x - z\rangle = -\lambda\langle u, y - x\rangle \\ \theta(y) - \theta(z) &\geqslant \langle u, y - z\rangle = (1-\lambda)\langle u, y - x\rangle \end{aligned}$$

Multiplying the first equation by $(1-\lambda)$ and the second by λ, and summing up yields

$$(1-\lambda)\theta(x) + \lambda\theta(y) \geqslant \theta(z) = \theta((1-\lambda)x + \lambda y)$$

□

[10] *Proof*

(⇐) If $0 \in \bar\partial\theta(z)$ then applying (12) with $u = 0$ yields $\theta(x) - \theta(z) \geqslant 0$ for all $x \in \mathcal{H}$

(⇒) If z is a global minimizer then $z \in \mathrm{dom}(\theta)$ (because θ is proper) and Definition 5 can be rewritten as $\theta(x) - \theta(z) \geqslant \langle 0, x - z\rangle$ for any $x \in \mathcal{H}$ which by Definition 4 means $0 \in \bar\partial\theta(z)$. □

A.3. **Subdifferential and lower convex enveloppe for possibly nonconvex functions.**

DEFINITION 6. *(Lower convex enveloppe of a function)*
Let $\theta : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a proper function for which there exists $x_0 \in \mathcal{H}$ such that $\bar{\partial}\theta(x_0) \neq \varnothing$. Then for any such x_0 and any $u \in \bar{\partial}\theta(x_0)$ the convex continuous function $\varrho(x) := \theta(x_0) + \langle u, x - x_0\rangle$ satisfies $\varrho(z) \leqslant \theta(z)$, $\forall\ z \in \mathcal{H}$ and $\varrho(x_0) = \theta(x_0)$, hence one can define the lower convex enveloppe[11] *$\breve{\theta}$ as the pointwise supremum of all the convex lower-semicontinuous functions minorizing θ*

$$\breve{\theta}(x) := \sup\{\varrho(x) \,|\, \varrho : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}, \varrho \text{ convex lsc},\ \varrho(z) \leqslant \theta(z), \forall\ z \in \mathcal{H}\}, \quad \forall\ x \in \mathcal{H}. \tag{13}$$

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

$$\breve{\theta}(x) \leqslant \theta(x), \forall\ x \in \mathcal{H}, \tag{14}$$

hence it is proper.

PROPOSITION 2. *Let $\theta : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be proper and $x_0 \in \mathcal{H}$. The following properties are equivalent:*
(a) $\bar{\partial}\theta(x_0) \neq \varnothing$
(b) *the convex enveloppe $\breve{\theta}$ is well defined and satisfies $\breve{\theta}(x_0) = \theta(x_0)$*
When they hold we have $\bar{\partial}\theta(x_0) = \partial\breve{\theta}(x_0)$.

Proof. We first prove that (a) implies both (b) and $\bar{\partial}\theta(x_0) \subset \partial\breve{\theta}(x_0)$. We then prove that (b) similarly implies (a) and $\bar{\partial}\theta(x_0) \supset \partial\breve{\theta}(x_0)$. Combining both result shows that (a) is equivalent to (b), and that when these properties hold we have $\bar{\partial}\theta(x_0) = \partial\breve{\theta}(x_0)$.

(a) implies (b) and $\bar{\partial}\theta(x_0) \subset \partial\breve{\theta}(x_0)$. By (a) we can consider $u \in \bar{\partial}\theta(x_0)$ and define $\varrho(x) := \theta(x_0) + \langle u, x - x_0\rangle$. As ϱ is convex lsc and $\varrho \leqslant \theta$ on $\mathcal{H}$, by the definition (13) of $\breve{\theta}$ and (14) we have $\varrho(x) \leqslant \breve{\theta}(x) \leqslant \theta(x)$, $\forall\ x \in \mathcal{H}$. Since $\varrho(x_0) = \theta(x_0)$, this inequality with $x := x_0$ yields $\breve{\theta}(x_0) = \theta(x_0)$. This establishes (b). Moreover, the equality $\breve{\theta}(x_0) = \theta(x_0)$, together with the inequality $\varrho(x) \leqslant \breve{\theta}(x) \leqslant \theta(x)$, $\forall\ x \in \mathcal{H}$ and the definition of ϱ, shows that

$$\forall\ x \in \mathcal{H}, \qquad \breve{\theta}(x) - \breve{\theta}(x_0) = \breve{\theta}(x) - \theta(x_0) \geqslant \varrho(x) - \theta(x_0) = \langle u, x - x_0\rangle$$

hence $u \in \partial\breve{\theta}(x_0)$. As this holds for any $u \in \bar{\partial}\theta(x_0)$ we obtain $\bar{\partial}\theta(x_0) \subset \partial\breve{\theta}(x_0)$.

(b) implies (a) and $\partial\breve{\theta}(x_0) \subset \bar{\partial}\theta(x_0)$. Since $\breve{\theta}$ is convex, we can consider $u' \in \partial\breve{\theta}(x_0)$. By (14) and (b), for any $x \in \mathcal{H}$, $\theta(x) - \theta(x_0) \geqslant \breve{\theta}(x) - \theta(x_0) = \breve{\theta}(x) - \breve{\theta}(x_0) \geqslant \langle u', x - x_0\rangle$, hence $u' \in \bar{\partial}\theta(x_0)$. This establishes (a). As this holds for any $u' \in \partial\breve{\theta}(x_0)$ we get $\partial\breve{\theta}(x_0) \subset \bar{\partial}\theta(x_0)$. □

PROPOSITION 3. *If $\bar{\partial}\theta(x) \neq \varnothing$ and θ is (Fréchet) differentiable at x then $\bar{\partial}\theta(x) = \{\nabla\theta(x)\}$.*

Proof. Consider $u \in \bar{\partial}\theta(x)$. 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 any $h \in \mathcal{B}$. For any $h \in \mathcal{B}$, Definition 4 yields

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

[11] also known as convex hull, [28, p. 57],[19, Definition 2.5.3]

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.4. **Characterizing functions with a given subdifferential.**

LEMMA 3. *Let a_0, a_1 be two convex functions on $[0,1]$ such that $\partial a_0(t) \cap \partial a_1(t) \neq \varnothing$ 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 on $[0,1]$, is it continuous and differentiable except on a countable set $B_i \subset [0,1]$ [19, Theorem 4.2.1 (ii)]. By Proposition 3, $\partial a_i(t) = \{a_i'(t)\}$ for any $t \in [0,1]\backslash B_i$ hence the function $\delta := a_1 - a_0$ is continuous and differentiable on $[0,1]\backslash(B_0 \cup B_1)$. For $t \in [0,1]\backslash(B_0 \cup B_1)$, $\{a_0'(t)\} \cap \{a_1'(t)\} = \partial a_0(t) \cap \partial a_1(t) \neq \varnothing$, hence $a_0'(t) = a_1'(t)$ and $\delta'(t) = 0$. As $B_0 \cup B_1$ is countable it follows[12] that δ is constant. □

COROLLARY 9. *Let $\theta_0, \theta_1 : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be proper and $\mathcal{C} \subset \mathcal{H}$ be a non-empty set. Assume that $\mathcal{C}$ is polygonally connected*[13]*. The following properties are equivalent:*

(a) *For any $z \in \mathcal{C}$, $\bar\partial\theta_0(z) \cap \bar\partial\theta_1(z) \neq \varnothing$;*
(b) *There is a constant $K \in \mathbb{R}$ such that $\theta_1(x) - \theta_0(x) = K$, $\forall\, x \in \mathcal{C}$.*

Proof. (b) ⇒ (a) is trivial. The proof of (a) ⇒ (b) 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 any $t \in [0,1]$, and $a_i(t) = +\infty$ if $t \notin [0,1]$. As $\mathcal{C}$ is convex, $z_t := x^* + t(x-x^*) \in \mathcal{C}$ hence by(a) for any $t \in [0,1]$ there exists $u_t \in \bar\partial\theta_0(z_t) \cap \bar\partial\theta_1(z_t)$. By Definition 4 for any $t, t' \in [0,1]$,

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

This shows that $\langle u_t, x-x^*\rangle \in \bar\partial a_i(t)$, $i = 0, 1$. Thus $\bar\partial a_i(t) \neq \varnothing$ for any $t \in [0,1]$, so by Lemma 2 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 any $t \in [0,1]$. By Lemma 3, there exists $K \in \mathbb{R}$ such that $a_1(t) - a_0(t) = K$ for any $t \in [0,1]$. 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 any $x \in \mathcal{C}$, we have established (a) ⇒ (b) as soon as $\mathcal{C}$ is convex.

(ii) Now we prove that (a) ⇒ (b) when $\mathcal{C}$ is polygonally connected. Fix some $x^* \in \mathcal{C}$. 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 $0 \leqslant j < n$. Invoking Lemma 3 and

[12] cf for example `https://fr.wikipedia.org/wiki/Lemme_de_Cousin` section 4.9

[13] 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. There are connected sets that are not path-connected. Every *open* connected set is polygonally-connected, see e.g. `https://math.stackexchange.com/questions/2161210/polygonally-connected-open-sets`.

the fact that $\mathcal{C}$ is polygonally connected shows that $\theta_1(x) - \theta_0(x) = \theta_1(x^*) - \theta_0(x^*) = K$. This establishes (b). □

REMARK 2. Corollary 9 generalizes [25, Proposition 8.b] on the characterization of convex functions by their subdifferential. It shows that one only needs the subdifferentials to intersect.

A.5. **Proof of Theorem 3.** The indicator function of a set $\mathcal{S}$ is denoted

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

Fact 3. *Let $\theta : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a proper function and $\mathcal{S} \neq \varnothing$ be a subset of $\mathcal{H}$. The function $\varrho(x) = \theta(x) + \chi_{\mathcal{S}}(x)$ satisfies $\operatorname{dom}(\varrho) \subset \mathcal{S}$ and*[14] *$\arg\min_x \varrho(x) \subset \mathcal{S}$.*

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

$$\theta := b + \varphi + \chi_{\operatorname{Im}(f)}. \tag{15}$$

Consider $x \in \operatorname{Im}(f)$. 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

$$\forall\, x' \in \mathcal{H},\ -\langle A(y), x'\rangle + \underbrace{b(x') + \varphi(x') + \chi_{\operatorname{Im}(f)}(x')}_{=\theta(x')} \geqslant -\langle A(y), x\rangle + \underbrace{b(x) + \varphi(x) + \chi_{\operatorname{Im}(f)}(x)}_{=\theta(x)} \tag{16}$$

which is equivalent to

$$\forall\, x' \in \mathcal{H} \qquad \theta(x') - \theta(x) \geqslant \langle A(y), x' - x\rangle \tag{17}$$

meaning that $A(y) \in \bar{\partial}\theta\left(x\,|_{x=f(y)}\right)$. As this holds for all $y \in \mathcal{Y}$ such that $f(y) = x$, we get $A(f^{-1}(x)) \subset \bar{\partial}\theta(x)$. Consider $g_1 := \breve{\theta}$ according to Definition 6. The function g_1 is convex lsc and

$$\forall\, x \in \operatorname{Im}(f), \bar{\partial}\theta(x) \neq \varnothing. \tag{18}$$

Hence, by Proposition 2, $\bar{\partial}\theta(x) = \partial g_1(x)$ for any $x \in \operatorname{Im}(f)$. This establishes (aii) with $g := g_1 = \breve{\theta}$.

(aii) $\Rightarrow$ (ai). Set $\theta_1 := g + \chi_{\operatorname{Im}(f)}$. By (aii), $\partial g(x) \neq \varnothing$ for any $x \in \operatorname{Im}(f)$. Noticing also that for a convex function g one has $\operatorname{dom}(g) \supset \operatorname{dom}(\partial g) := \{x \in \mathcal{H}, \partial g(x) \neq \varnothing\}$ we infer that $\operatorname{dom}(g) \supset \operatorname{Im}(f)$, and consequently

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

Consider $y \in \mathcal{Y}$ and $x := f(y)$ so that $x \in \operatorname{Im}(f)$, 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 any $(x, x') \in (\operatorname{Im}(f), \mathcal{H})$ one has

$$\theta_1(x') - \theta_1(x) = \theta_1(x') - g(x) \geqslant g(x') - g(x) \geqslant \langle A(y), x' - x\rangle,$$

showing that $A(y) \in \bar{\partial}\theta_1(x)$. This is equivalent to (17) with $\theta := \theta_1$, and since $\operatorname{dom}(\theta_1) = \operatorname{Im}(f)$, the inequality in (16) holds with $\varphi(x) := \theta_1(x) - b(x)$, i.e., x is a global minimizer of $D(x', y) +$

[14] We recall that $\arg\min_x \varrho(x)$ is the set of all global minimizers. As the empty set is a subset of any set, the last inclusion in $\mathcal{S}$ remains valid if $\arg\min_x \varrho(x) = \varnothing$.

$\varphi(x')$. Since this holds for any $y \in \mathcal{Y}$, 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[15] $g_1 := \breve{\theta}$ with θ defined in (15). Following the arguments of (ai) $\Rightarrow$ (aii) we obtain that g_1 (just as g) satisfies (aii). For any $x \in \mathcal{C}$ we thus have $\partial g(x) \cap \partial g_1(x) \supset A(f^{-1}(x)) \neq \varnothing$ with g, g_1 convex lsc 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 (7) between g and φ we now show that $g_1(x) = b(x) + \varphi(x)$ on $\mathcal{C}$. By (18) and Proposition 2 we have $\breve{\theta}(x) = \theta(x)$ for any $x \in \mathrm{Im}(f)$, hence as $\mathcal{C} \subset \mathrm{Im}(f)$ we obtain $g_1(x) := \breve{\theta}(x) = \theta(x) = b(x) + \varphi(x)$ for any $x \in \mathcal{C}$. This establishes (7).

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

$$\varrho(y) \;:=\; \begin{cases} +\infty, & \forall y \notin \mathcal{Y} \\ \langle B(f(y)), y\rangle - b(f(y)) - \varphi(f(y)), & \forall y \in \mathcal{Y}. \end{cases} \tag{19}$$

Consider $y \in \mathcal{Y}$. From (ci), for any 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

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

This shows that

$$B(f(y)) \in \bar{\partial}\varrho(y). \tag{20}$$

Set $\psi_1 := \breve{\varrho}$ according to Definition 6. Then the function ψ_1 is convex lsc and for any $y \in \mathcal{Y}$ the function $B(f(y))$ is well defined, so $\bar{\partial}\varrho(y) \neq \varnothing$. Hence, by Proposition 2, $B(f(y)) \in \bar{\partial}\varrho(y) = \partial\breve{\varrho}(y) = \partial\psi_1(y)$ for any $y \in \mathcal{Y}$. This establishes (cii) with $\psi := \psi_1$.

(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 4 to $\partial\psi$ yields $\psi(y) - \psi(y') \geqslant \langle y - y', B(f(y'))\rangle$. Using this inequality, one has

$$\begin{array}{rcl} \forall\, y, y' \in \mathcal{Y} \qquad h(y') - h(y) & = & \langle B(f(y')), y'\rangle - \psi(y') - \langle B(f(y)), y\rangle + \psi(y) \\ & \geqslant & \langle B(f(y')), y'\rangle - \langle B(f(y)), y\rangle + \langle B(f(y')), y - y'\rangle \\ & = & \big\langle B(f(y') - B(f(y)),\ y\big\rangle \end{array} \tag{21}$$

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

Noticing that for any $x \in \mathrm{Im}(f)$ 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

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

For $x \in \mathrm{Im}(f)$, as $f(y) = f(y') = x$ for any $y, y' \in f^{-1}(x)$, applying (21) 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 (21) 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 all $x \in \mathcal{H}$, 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 all $y \in \mathcal{Y}$, 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 (19) satisfies (cii). As ψ and ψ_1 both satisfy (cii), for any $y \in \mathcal{C}'$ we have $\partial\psi(y) \cap \partial\psi_1(y) \supset B(f(y)) \neq \varnothing$ with ψ, ψ_1 convex lsc 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 (20), $\bar{\partial}\varrho(y) \neq \varnothing$ for any $y \in \mathcal{Y}$, hence by Proposition 2 we have $\breve{\varrho}(y) = \varrho(y)$ for any $y \in \mathcal{Y}$. As $\mathcal{C}' \subset \mathcal{Y}$, it follows that $\psi_1(y) = \breve{\varrho}(y) = \varrho(y)$ for any $y \in \mathcal{C}'$. This establishes (8).

A.6. **Proof of Corollary 7.** The proof relies on the following lemma which we prove first.

Lemma 4. *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 (non-locally) subdifferentiable at any $x \in \mathcal{X}$ and that*

$$\forall\, x \in \mathcal{X} \qquad \varrho(x) \in \bar{\partial}\theta(x) \tag{22}$$

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, $\{\varrho(x)\} = \{\nabla\theta(x)\} = \bar{\partial}\theta(x)$ for any $x \in \mathcal{X}$.

Proof. Applying Definition 4 ((non-local) subdifferential) to any $x, x' \in \mathcal{X}$, one has

$$\langle \varrho(x), x' - x\rangle \leqslant \theta(x') - \theta(x) \leqslant \langle \varrho(x'), x' - x\rangle$$

For $x' = x + \epsilon h$, $\epsilon > 0$ and dividing by ϵ we obtain

$$\langle \varrho(x), h\rangle \leqslant \frac{\theta(x + \epsilon h) - \theta(x)}{\epsilon} \leqslant \langle \varrho(x + \epsilon h), h\rangle \tag{23}$$

(a)⇒(b). Using that ϱ is continuous at x, the limit in (23) when ϵ goes to zero yields the Gâteaux differential $\delta\theta(x;h)$ of θ in the direction of h [24, Section 7.2]

$$\forall\, h \in \mathcal{H} \qquad \delta\theta(x;h) := \lim_{\epsilon\to 0} \frac{\theta(x+\epsilon h)-\theta(x)}{\epsilon} = \langle \varrho(x), h\rangle$$

which holds for any $h \in \mathcal{H}$ because $\mathcal{X}$ is open. Since ϱ is continuous at x, for any $\varepsilon > 0$ there is $\eta > 0$ such that for any $x' \in \mathcal{X}$,

$$\|x'-x\| \;\leqslant\; \eta \quad\Rightarrow\quad \|\varrho(x')-\varrho(x)\| \;\leqslant\; \varepsilon$$

Further, given h with $\|h\| \leqslant \eta$ small enough, we have $x+th \in \mathcal{X}$ for any $0 \leqslant t \leqslant 1$ and by the one-dimensional mean value theorem there is $\widetilde{x} \in \mathcal{X}$ with $\|\widetilde{x}-x\| \leqslant \|h\| \leqslant \eta$ such that

$$\theta(x+h)-\theta(x) = \langle \varrho(\widetilde{x}), h\rangle \tag{24}$$

Using (24) and Cauchy-Schwarz inequality,

$$|\theta(x+h)-\theta(x)-\delta\theta(x;h)| \;=\; |\langle \varrho(\widetilde{x})-\varrho(x),\, h\rangle| \leqslant \|\varrho(\widetilde{x})-\varrho(x)\|\,\|h\| \;\leqslant\; \varepsilon\|h\|$$

This, together with the facts that $h \mapsto \delta\theta(x;h)$ is linear and continuous shows that θ is Fréchet differentiable at x and thus $\nabla\theta(x) = \varrho(x)$ where ϱ is continuous on $\mathcal{X}$. This establishes (b).

(b)⇒(a). Consider $x \in \mathcal{X}$. As $\bar{\partial}\theta(x) \neq \varnothing$ by (22) and θ is differentiable at x by (b), Proposition 3 yields $\bar{\partial}\theta(x) = \{\nabla\theta(x)\}$, hence $\varrho(x) = \nabla\theta(x)$ by (22). As this holds for any $x \in \mathcal{X}$, and as by (b), $\nabla\theta$ is continuous on $\mathcal{X}$, this proves that ϱ is continuous on $\mathcal{X}$, thus establishing (a).

□

Proof. [Proof of Corollary 7] (ai) ⇔ (aii) By assumption ψ satisfies Theorem 4(cii), i.e., $f(y) \in \partial\psi(y)$, $\forall\, y \in \mathcal{V}$. By Lemma 4 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) ⇔ (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 all $x' \in \mathcal{B}_x$. 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 any $x \in \mathcal{V}$, the equivalence holds on $\mathcal{V}$.

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

$$y = \nabla g(f(y)) \quad \forall\, y \in f^{-1}(\mathcal{X}). \tag{25}$$

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 (25) yields $\widetilde{f}^{-1}(x) = \nabla g(x)$ for any $x \in \mathcal{X}$. Then, since g is $C^{k+1}(\mathcal{X})$, it follows that $\widetilde{f}^{-1}$ is $\mathcal{C}^k(\mathcal{X})$.

(biii) ⇒ (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 4 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 $C^{k+1}(\mathcal{X})$. □

A.7. **Proof of Theorem 2.** The equivalence between (a) and (b) is a direct consequence of Corollary 1. We now focus on the equivalence between (b) and (c).
First we prove (b) ⇒ (c). By (b) we have $f(y) = \nabla\psi(y)$ for any $y \in \mathcal{Y}$, and as f is $C^1(\mathcal{Y})$ this implies that ψ is $C^2(\mathcal{Y})$. Consider an arbitrary $y \in \mathcal{Y}$ and $u, v \in \mathcal{H}$. As $\mathcal{Y}$ is open, there is an open subset $\mathcal{V} \subset \mathbb{R}^2$ containing $(0,0)$ such that for any $(a,b) \in \mathcal{V}$, $\varrho(a,b) := \psi(y+au+bv)$ is well defined. As ψ is $C^2(\mathcal{Y})$, ϱ is $C^2(\mathcal{V})$, with $\frac{\partial\varrho}{\partial a}(a,b) = \langle\nabla\psi(y+au+bv), u\rangle = \langle f(y+au+bv), u\rangle$ and $\frac{\partial\varrho}{\partial b}(a,b) = \langle\nabla\psi(y+au+bv), v\rangle = \langle f(y+au+bv), v\rangle$. By Schwarz's theorem its mixed derivatives are equal hence

$$\langle Df(y)\cdot u, v\rangle = \tfrac{\partial^2\varrho}{\partial a\partial b}(0,0) = \tfrac{\partial^2\varrho}{\partial b\partial a}(0,0) = \langle Df(y)\cdot v, u\rangle$$

Moreover, as ψ is convex, $a \mapsto \varrho(a,0)$ is also convex, and its second derivative $\frac{\partial^2}{\partial a^2}\varrho(a,0) = \langle Df(y)\cdot u, u\rangle$ is non-negative. We have just shown that $\langle Df(y)\cdot u, v\rangle = \langle Df(y)\cdot v, u\rangle$ and $\langle Df(y)\cdot u, u\rangle \geqslant 0$, $\forall\, u, v \in \mathcal{H}$, i.e. $Df(y)$ is symmetric positive semi-definite.
We now prove (c) ⇒ (b).
Overwiew of the proof. if indeed $\nabla\psi = f$ then given an arbitrary y^*, for any y we have $\psi(y) - \psi(y^*) = \int_0^1 \langle\nabla\psi(y^* + t(y-y^*)), y-y^*\rangle dt = \int_0^1 \langle f(y^* + t(y-y^*)), y-y^*\rangle dt$. We will use this to define ψ and check that indeed 1) the definition makes sense; 2) the resulting ψ is convex; 3) $f(y) \in \partial\psi(y)$ for all $y \in \mathcal{Y}$; 4) $f = \nabla\psi$.

Choose an arbitrary $y^* \in \mathcal{Y}$. As $\mathcal{Y}$ is convex, $(1-t)y^* + ty \in \mathcal{Y}$ for any $0 \leqslant t \leqslant 1$ and $y \in \mathcal{Y}$, hence $\theta(y,t) := \langle f\left((1-t)y^* + ty\right), y - y^*\rangle$ is well defined. Since f is continuous, for any $y \in \mathcal{Y}$ the function $t \mapsto \theta(y,t)$ is continuous on $[0,1]$ hence integrable and we can define

$$\psi(y) := \int_0^1 \theta(y,t)dt.$$

Given $h \in \mathcal{H}$, let $\varrho(t) := t\langle f(y^* + t(y-y^*)), h\rangle$, $t \in [0,1]$. As f is differentiable, ϱ is differentiable with

$$\varrho'(t) = t\langle Df(y^* + t(y-y^*))\cdot(y-y^*), h\rangle + \langle f(y^* + t(y-y^*)), h\rangle.$$

Denoting $D_y\theta(y,t) : \mathcal{H} \to \mathbb{R}$ the differential of $y \mapsto \theta(y,t)$ and $\nabla_y\theta(y,t) \in \mathcal{H}$ its gradient, as $Df(y^* + t(y-y^*))$ is symmetric we obtain

$$\langle\nabla_y\theta(y,t), h\rangle = D_y\theta(y,t)\cdot h = t\langle Df(y^* + t(y-y^*))\cdot h, (y-y^*)\rangle + \langle f(y^* + t(y-y^*)), h\rangle = \varrho'(t).$$

As f is C^1, $\varrho'(t)$ is continuous on $[0,1]$ hence it is integrable and

$$\langle f(y), h\rangle = \varrho(1) - \varrho(0) = \int_0^1 \varrho'(t)dt = \int_0^1 \langle\nabla_y\theta(y,t), h\rangle\, dt$$

As $\mathcal{Y}$ is open, when $\|h\|$ is small enough we have $y + h \in \mathcal{Y}$ and the above computations yield

$$\psi(y+h) - \psi(y) - \langle f(y), h\rangle = \int_0^1 \left(\theta(y+h,t) - \theta(y,t) - \langle\nabla_y\theta(y,t), h\rangle\right) dt. \tag{26}$$

Now consider $y_0, y_1 \in \mathcal{Y}$ and define $\varphi : \alpha \in [0,1] \mapsto \psi((1-\alpha)y_0 + \alpha y_1)$.
For $\alpha \in [0,1]$ and $\epsilon \in [-\alpha, 1-\alpha]$, by (26) with $y := (1-\alpha)y_0 + \alpha y_1$ and $h := \epsilon(y_1 - y_0)$ we obtain

$$\frac{\varphi(\alpha+\epsilon) - \varphi(\alpha)}{\epsilon} - \epsilon\langle f(y), y_1 - y_0\rangle = \int_0^1 \frac{\theta(y + \epsilon(y_1-y_0), t) - \theta(y,t) - \langle \nabla_y \theta(y,t), \epsilon(y_1-y_0)\rangle}{\epsilon} dt.$$

As f is C^1, $(\epsilon, t) \to \|\nabla_y \theta(y + \epsilon(y_1 - y_0), t)\|$ is continuous. As $[-\alpha, 1-\alpha] \times [0,1]$ is compact, we get

$$B := \max_{\epsilon \in [-\alpha, 1-\alpha], t \in [0,1]} \|\nabla_y \theta(y + \epsilon(y_1 - y_0), t)\| < \infty.$$

Consider $t \in [0,1]$. Denoting $g(u) := \theta(y + u\epsilon(y_1 - y_0), t)$ for $u \in [0,1]$, we have $\theta(y + \epsilon(y_1 - y_0), t) - \theta(y,t) = g(1) - g(0)$ hence by the mean value theorem there exists $c \in (0,1)$ such that $\theta(y + \epsilon(y_1 - y_0), t) - \theta(y,t) = g'(c) = \langle \nabla_y \theta(y + c\epsilon(y_1 - y_0), t), \epsilon(y_1 - y_0)\rangle$. As $c\epsilon \in [-\alpha, 1-\alpha]$ we have

$$\frac{|\theta(y + \epsilon(y_1 - y_0), t) - \theta(y,t) - \langle \nabla_y \theta(y,t), \epsilon(y_1 - y_0)\rangle|}{\epsilon} \leqslant \|\nabla_y \theta(y + c\epsilon(y_1 - y_0), t)\| + \|\nabla_y \theta(y,t)\| \leqslant 2B.$$

As f is differentiable we also have for any $t \in [0,1]$ the pointwise convergence

$$\lim_{\|h\| \to 0} \frac{\theta(y + \epsilon(y_1 - y_0), t) - \theta(y,t) - \langle \nabla_y \theta(y,t), \epsilon(y_1 - y_0)\rangle}{\epsilon} = 0$$

By the dominated convergence theorem we conclude that

$$\lim_{\epsilon \to 0, \epsilon \in [-\alpha, 1-\alpha]} \frac{\varphi(\alpha + \epsilon) - \phi(\alpha)}{\epsilon} - \langle f(y), y_1 - y_0\rangle = 0$$

i.e., φ is differentiable on $[0,1]$ with $\varphi'(\alpha) = \langle f(y), y_1 - y_0\rangle = \langle f(y_0 + \alpha(y_1 - y_0), y_1 - y_0\rangle$. This holds for any $\alpha \in [0,1]$ (considering the one-sided derivative for $\alpha \in \{0,1\}$). As f is C^1, φ' is also C^1 and

$$\varphi''(\alpha) = \langle Df(y_0 + \alpha(y_1 - y_1) \cdot (y_1 - y_0), y_1 - y_0\rangle$$

Since $Df(y_0 + \alpha(y_1 - y_0)) \succeq 0$ we obtain $\varphi''(\alpha) \geqslant 0$. This shows that φ is convex on $[0,1]$ hence

$$\psi((1-\alpha)y_0 + \alpha y_1) = \varphi(\alpha) \leqslant (1-\alpha)\varphi(0) + \alpha\varphi(1) = (1-\alpha)\psi(y_0) + \alpha\psi(y_1), \forall\, \alpha \in [0,1]$$

The convexity of φ also implies that $\varphi'(0) \in \partial\varphi(0)$ hence

$$\psi(y_1) - \psi(y_0) = \varphi(1) - \varphi(0) \geqslant \varphi'(0) = \langle f(y_0), y_1 - y_0\rangle$$

As the above relations hold for any $y_0, y_1 \in \mathcal{Y}$ and $\alpha \in [0,1]$, we conclude that ψ is convex and $f(y_0) \in \partial\psi(y_0)$ for any $y_0 \in \mathcal{Y}$. Since f is continuous, by Lemma 4 it follows that ψ is Fréchet differentiable on $\mathcal{Y}$ with $f(y) = \nabla\psi(y)$ for all $y \in \mathcal{Y}$.
□

A.8. **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 any $y \in \mathcal{Y}$, there is a function φ_0 and a convex lsc function $\psi \in C^2(\mathcal{Y})$ such that $\nabla\psi(y) = f(y) \in \mathrm{prox}_{\varphi_0}(y)$ for any $y \in \mathcal{Y}$. We define $\varphi(x) := \varphi_0(x) + \chi_{\mathrm{Im}(f)}(x)$ and let the reader check that $f(y) \in \mathrm{prox}_{\varphi}(y)$ for any $y \in \mathcal{Y}$.

Uniqueness. Consider $\widetilde{f}$ any function such that $\widetilde{f}(y) \in \mathrm{prox}_{\varphi}(y)$ for all y. This implies that

$$\tfrac{1}{2}\|y - f(y)\|^2 + \varphi(f(y)) = \tfrac{1}{2}\|y - \widetilde{f}(y)\|^2 + \varphi(\widetilde{f}(y)) = \min_{x\in\mathcal{H}}\{\tfrac{1}{2}\|y - x\|^2 + \varphi(x)\}, \quad \forall\, y \in \mathcal{Y}. \tag{27}$$

By Corollary 1 there is a convex lsc function $\widetilde{\psi}$ such that $\widetilde{f}(y) \in \partial\widetilde{\psi}(y)$ for any $y \in \mathcal{Y}$. Since $\mathcal{Y}$ is convex it is polygonally connected hence by Theorem 4(b) and (27) there are $K, K' \in \mathbb{R}$ such that

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

Thus, $\widetilde{\psi}$ is $C^2(\mathcal{Y})$ and $\widetilde{f}(y) \in \partial\widetilde{\psi}(y) = \{\nabla\psi(y)\} = \{f(y)\}$ for any $y \in \mathcal{Y}$. This shows that $\widetilde{f}(y) = f(y)$ for any 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., $\mathrm{prox}_{\varphi}(y) = \{f(y)\}$.

Injectivity. The proof follows that of [15, 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 all 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 $\theta'(t) > 0$.

Global minimum is the unique stationary point. The proof is inspired by that of [15, Theorem 1]. Since x is a stationary point of $\theta : x \mapsto \frac{1}{2}\|y - x\|^2 + \varphi(x)$, the gradient $\nabla\theta(x)$ is well defined and $\nabla\theta(x) = 0$. This implies the existence of a neighborhood $\mathcal{V}$ of x such that $\mathcal{V} \subset \mathrm{dom}(\varphi) = \mathrm{Im}(f)$. As a result $x = f(v)$ for some $v \in \mathcal{Y}$. 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 any $u \in \mathcal{Y}$. On the other hand, for any $u \in \mathcal{Y}$ 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 $Df(v)\ (v - y) = \nabla\theta(v) = Df(v)\ \nabla\theta(x) = 0$. As $Df(v) \succ 0$, this implies $v = y$, hence $x = f(y)$.

A.9. **Proof of Lemma 1.** 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 any open set where it is well defined.

Let f be a generalized social shrinkage operator as described in Lemma 1. For $i \in G$, by Definition 2 we have $i \in N_i = \mathrm{supp}(w^i) = \mathrm{supp}(w^G)$, establishing that[16]

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

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 any $i, j \in [\![1, n]\!]$

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

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 any 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 $w^i_j = 0$.

To conclude we now show that $w^j_i = 0$. As $w^i_j = 0$, 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 $2(w^j_i)^2 y'_j y'_i h_j'\left(\|\mathrm{diag}(w^j)y'\|_2^2\right) = \frac{\partial f_j}{\partial y_i}(y') = \frac{\partial f_i}{\partial y_j}(y') = 0$, hence $w^j_i = 0$. We thus obtain that $w^i_j = w^j_i = 0$ as claimed, establishing (30) and therefore $G = \mathrm{supp}(w^G)$.

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

References

[1] Madhu Advani and Surya Ganguli. An equivalence between high dimensional Bayes optimal inference and M-estimation. *arXiv*, September 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] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2009.

[6] Stephen P Boyd, John Duchi, and Lieven Vandenberghe. Subgradients. Notes for EE364b, Stanford University, Spring 2014-15. `http://web.stanford.edu/class/ee364b/lectures/subgradients_notes.pdf`, April 2015.

[7] 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.

[8] 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.

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

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

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

[12] Asen L Dontchev and R Tyrrell Rockafellar. *Implicit Functions and Solution Mappings*. Springer Series in Operations Research and Financial Engineering. Springer New York, New York, NY, 2014.

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

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

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

[16] 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.

[17] Rémi Gribonval and Mila Nikolova. On bayesian estimation and proximity operators. *arXiv*, July 2018. submitted.

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

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

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

[21] 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.

[22] 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.

[23] 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.

[24] D G Luenberger. *Optimization by vector space methods*. John Wiley & Sons, 1969.

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

[26] 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.

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

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

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

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

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

[32] 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.

[33] Gaël Varoquaux, Matthieu Kowalski, and Bertrand Thirion. Social-sparsity brain decoders - faster spatial sparsity. *arXiv*, stat.ML, 2016.

[34] 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.

[35] 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.