Thresholding gradient methods in Hilbert spaces: support identification and linear convergence

Guillaume Garrigos, Lorenzo Rosasco, Silvia Villa

THRESHOLDING GRADIENT METHODS IN HILBERT SPACES: SUPPORT IDENTIFICATION AND LINEAR CONVERGENCE

GUILLAUME GARRIGOS, LORENZO ROSASCO, AND SILVIA VILLA

ABSTRACT. We study ℓ^1 regularized least squares optimization problem in a separable Hilbert space. We show that the iterative soft-thresholding algorithm (ISTA) converges linearly, without making any assumption on the linear operator into play or on the problem. The result is obtained combining two key concepts: the notion of *extended support*, a finite set containing the support, and the notion of *conditioning over finite dimensional sets*. We prove that ISTA identifies the solution extended support after a finite number of iterations, and we derive linear convergence from the conditioning property, which is always satisfied for ℓ^1 regularized least squares problems. Our analysis extends to the the entire class of thresholding gradient algorithms, for which we provide a conceptually new proof of strong convergence, as well as convergence rates.

KEYWORDS. Forward-Backward method, support identification, conditioning, convergence rates.

MSC. 49K40, 49M29, 65J10, 65J15, 65J20, 65J22, 65K15, 90C25, 90C46.

1. INTRODUCTION

Recent works show that, for many problems of interest, favorable geometry can greatly improve theoretical results with respect to more general, worst-case perspective [1, 16, 5, 20]. In this paper, we follow this perspective to analyze the convergence properties of threshold gradient methods in separable Hilbert spaces. Our starting point is the now classic iterative soft thresholding algorithm (ISTA) to solve the problem

$$f(x) = \|x\|_1 + \frac{1}{2}\|Ax - y\|^2, \tag{1}$$

defined by an operator A on $\ell^2(\mathbb{N})$ and where $\|\cdot\|_1$ is the ℓ^1 norm.

From the seminal work [11], it is known that ISTA converges strongly in $\ell^2(\mathbb{N})$. This result is generalized in [9] to a wider class of algorithms, the so-called thresholding gradient methods, noting that these are special instances of the Forward-Backward algorithm, where the proximal step reduces to a thresholding step onto an orthonormal basis (Section 2). Typically, strong convergence in Hilbert spaces is the consequence of a particular structure of the considered problem. Classic examples being even functions, functions for which the set of minimizers has a nonempty interior, or strongly convex functions [30]. Further examples are uniformly convex functions, or functions presenting a favorable geometry around their minimizers, like conditioned functions or Lojasiewicz functions, see e.g. [4, 20]. Whether the properties of ISTA, and more generally threshold gradient methods, can be explained from this perspective is not apparent from the analysis in [11, 9].

Our first contribution is revisiting these results providing such an explanation: for these algorithms, the whole sequence of iterates is fully contained in a specific finite-dimensional

G. GARRIGOS ✉
CNRS, École Normale Supérieure (DMA), 75005 Paris, France
guillaume.garrigos@ens.fr
L. ROSASCO
LCSL, Istituto Italiano di Tecnologia and Massachusetts Institute of Technology, Cambridge, MA 02139, USA
Università degli Studi di Genova (DIBRIS), 16146 Genova, Italy
lrosasco@mit.edu
S. VILLA
Politecnico di Milano (Dipartimento di Matematica), 20133 Milano, Italy
silvia.villa@polimi.it

subspace, ensuring automatically strong convergence. The key argument in our analysis is that after a finite number of iterations, the iterates identify the so called *extended support* of their limit. This set coincides with the active constraints at the solution of the dual problem, and reduces to the support, if some qualification condition is satisfied.
Going further, we tackle the question of convergence rates, providing a unifying treatment of finite and infinite dimensional settings. In finite dimensions, it is clear that if A is injective, then f becomes a strongly convex function, which guarantees a linear convergence rate. In [22], it is shown, still in a finite dimensional setting, that the linear rates hold just assuming A to be injective on the extended support of the problem. This result is generalized in [8] to a Hilbert space setting, assuming A to be injective on any subspace of finite support. Linear convergence is also obtained by assuming the limit solution to satisfy some nondegeneracy condition [8, 26]. In fact, it was shown recently in [6] that, in finite dimension, no assumption at all is needed to guarantee linear rates. Using a key result in [25], the function f was shown to be 2-conditioned on its sublevel sets, and 2-conditioning is sufficient for linear rates [2]. Our identification result, mentioned above, allows to easily bridge the gap between the finite and infinite dimensional settings. Indeed, we show that in any separable Hilbert space, linear rates of convergence always hold for the soft-thresholding gradient algorithm under no further assumptions. Once again, the key argument to obtain linear rates is the fact that the iterates generated by the algorithm identify, in finite time, a set on which we know the function to have a favorable geometry.

The paper is organized as follows. In Section 2 we describe our setting and introduce the thresholding gradient method. We introduce the notion of extended support in Section 3, in which we show that the thresholding gradient algorithm identifies this extended support after a finite number of iterations (Theorem 3.9). In Section 4 we present some consequences of this result on the convergence of the algorithm. We first derive in Section 4.1 the strong convergence of the iterates, together with a general framework to guarantee rates. We then specify our analysis to the function (1) in Section 4.2, and show the linear convergence of ISTA (Theorem 4.8). We also consider in Section 4.3 an elastic-net modification of (1), by adding an ℓ^p regularization term, and provide rates as well, depending on the value of $p \in]1, +\infty[$.

2. Thresholding gradient methods

Notation. We introduce some notation we will use throughout this paper. $\mathcal{N}$ is a subset of $\mathbb{N}$. Throughout the paper, X is a separable Hilbert space endowed with the scalar product $\langle \cdot, \cdot \rangle$, and $(e_k)_{k\in\mathcal{N}}$ is an orthonormal basis of X. Given $x \in X$, we set $x_k = \langle x, e_k \rangle$. The support of x is $\operatorname{supp}(x) = \{k \in \mathcal{N} \mid x_k \neq 0\}$. Analogously, given $C \subset X$, $C_k = \{\langle x, e_k \rangle \ : \ x \in C\}$. Given $J \subset \mathbb{N}$, the subspace supported by J is denoted by $X_J = \{x \in X \mid \operatorname{supp}(x) \subset J\}$ and the subset of finitely supported vectors $c_{00} = \{x \in X \ : \ \operatorname{supp}(x) \text{ is finite } \}$. Given a collection of intervals $\{I_k\}_{k\in\mathcal{N}}$ of the real line, with a slight abuse of notation, we define, for every $k \in \mathcal{N}$,

$$\mathbb{B}_{\infty,\mathcal{I}} = \bigoplus_{k\in\mathcal{N}} I_k = \{x \in X \ : \ x = \sum_{k\in\mathcal{N}} t_k e_k, \text{ with } t_k \in I_k \text{ for every } k \in \mathcal{N}\}.$$

Note that $\bigoplus_{k\in\mathcal{N}} I_k$ is a subspace of X. Therefore, the components of each element of $\bigoplus_{k\in\mathcal{N}} I_k$ must be square summable. The closed ball of center $x \in X$ and radius $\delta \in]0, +\infty[$ is denoted by $\mathbb{B}_X(x, \delta)$. Let $C \subset X$ be a closed convex set. Its indicator and support functions are denoted δ_C and σ_C, respectively, and the projection onto C is proj_C. Moreover, $\operatorname{int} C$, $\operatorname{bd} C$, $\operatorname{ri} C$, and $\operatorname{qri} C$ will denote respectively the interior, the boundary, the relative interior, and the quasi relative interior of C [4, Section 6.2]. The set of proper convex lower semi-continuous functions from X to $\mathbb{R} \cup \{+\infty\}$ is denoted by $\Gamma_0(X)$. Let $f \in \Gamma_0(X)$ and let $r \in]0, +\infty[$. The sublevel set of f is $S_f(r) = \{x \in X \mid f(x) - \inf f < r\}$. The proximity operator of f is defined as

$$(\forall \lambda \in]0, +\infty[) \qquad \operatorname{prox}_{\lambda f}(x) = \operatorname{argmin} \{y \in X \mid f(y) + \frac{1}{2\lambda}\|y - x\|^2\}.$$

Let $I \subset \mathbb{R}$ be a closed interval. Then, $\operatorname{prox}_{\sigma_I} = \operatorname{soft}_I$, where

$$(\forall t \in \mathbb{R}) \quad \operatorname{soft}_I(t) = \begin{cases} t - \inf I & \text{if } t < \inf I \\ 0 & \text{if } t \in I \\ t - \sup I & \text{if } t > \sup I, \end{cases}$$

is the soft-thresholder corresponding to I.

Problem and main hypotheses. We consider the general optimization problem

$$\text{(P)} \qquad \min_{x \in X} f(x), \qquad f = g + h,$$

where typically h will play the role of a smooth data fidelity term, and g will be a nonsmooth sparsity promoting regularizer. More precisely, we will make the following assumption:

$$\text{(H)} \quad \begin{cases} h \in \Gamma_0(X) \text{ is bounded from below,} \\ h \text{ is differentiable, and } \nabla h \text{ is } L\text{-Lipschitz continuous on } X,\ L \in \left]0, +\infty\right[, \\ g = \sum\limits_{k \in \mathcal{N}} g_k(\langle \cdot, e_k \rangle), \text{ with } g_k = \psi_k + \sigma_{I_k}, \text{ where:} \\ \quad \bullet \text{ for all } k \in \mathcal{N},\ I_k \text{ is a proper closed interval of } \mathbb{R}, \text{ and } \mathcal{I} = \{I_k\}_{k \in \mathcal{N}}, \\ \quad \bullet \text{ for all } k \in \mathcal{N},\ (\exists \omega > 0) \quad [-\omega, \omega] \subset I_k, \\ \quad \bullet \text{ for all } k \in \mathcal{N},\ \psi_k \in \Gamma_0(\mathbb{R}) \text{ is differentiable at 0 with } \psi_k(0) = 0 \text{ and } \psi_k'(0) = 0. \end{cases}$$

As stated in the above assumption, in this paper we focus on a specific class of functions g. They are given by the sum of a weighted ℓ^1 norm and a positive smooth function minimized at the origin, namely:

$$\|\cdot\|_{1,\mathcal{I}} = \sum_{k \in \mathcal{N}} \sigma_{I_k}, \qquad \psi = \sum_{k \in \mathcal{N}} \psi_k.$$

In [9] the following characterization has been proved: the proximity operators of such functions g are the monotone operators $T\colon X \to X$ such that for all $x \in X$, $T(x) = (T_k(x_k))_{k \in \mathcal{N}}$, for some $T_k\colon \mathbb{R} \to \mathbb{R}$ which satisfies

$$(\forall k \in \mathcal{N}) \quad T_k(x_k) = 0 \iff x_k \in I_k.$$

A few examples of such, so called, *thresholding operators* are shown in Figure 1, and a more in-depth analysis can be found in [9].

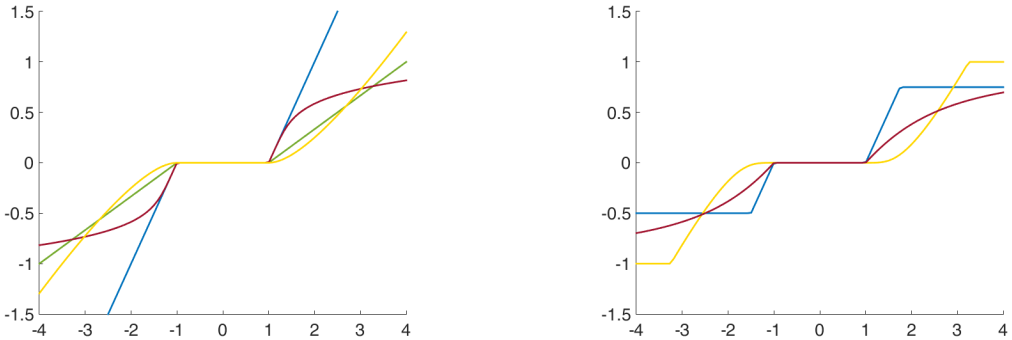

FIGURE 1. Some examples of thresholding proximal operators in $\mathbb{R}$. On the left: $|\cdot| = \sigma_{[-1,1]}$ (blue), $|\cdot| + |\cdot|^{1.5}$ (yellow), $|\cdot| + |\cdot|^2$ (green), $|\cdot| + |\cdot|^6$ (red). On the right: $|\cdot| + \delta_{[-0.5,0.75]}$ (blue), $|\cdot| + |\cdot|^{1.5} + \delta_{[-1,1]}$ (yellow), and $|\cdot| - \ln(1 - |\cdot|) + \delta_{]-1,1[}$ (red). Observe that here the range of prox_g is equal to the domain of $\partial\psi$.

A well-known approach to approximate solutions of (P) is the Forward-Backward algorithm [4]

$$\text{(FB)} \qquad x^0 \in X, \quad \lambda \in]0, 2L^{-1}[, \quad x^{n+1} = \operatorname{prox}_{\lambda g}(x^n - \lambda \nabla h(x^n)).$$

In our setting, (FB) is well-defined and specializes to a *thresholding gradient method*. The Proposition below gathers some basic properties of g and f following from assumption (H).

Proposition 2.1. The following hold.

(i) $\|\cdot\|_{1,\mathcal{I}}$ is the support function of $\mathbb{B}_{\infty,\mathcal{I}} = \bigoplus_{k\in\mathcal{N}} I_k$,
(ii) $\operatorname{dom}\partial\|\cdot\|_{1,\mathcal{I}} = c_{00}$,
(iii) $g \in \Gamma_0(X)$ and it is coercive,
(iv) f is bounded from below and $\operatorname{argmin} f$ is nonempty,
(v) the dual problem

$$\text{(D)} \qquad \min_{u\in X} \; g^*(u) + h^*(-u),$$

admits a unique solution $\bar{u} \in X$, and for all $\bar{x} \in \operatorname{argmin} f$, $\bar{u} = -\nabla h(\bar{x})$.
(vi) for all $x \in X$ and all $\lambda > 0$, the proximal operator of g can be expressed as

$$\operatorname{prox}_{\lambda g}(x) = \sum_{k\in\mathcal{N}} \operatorname{prox}_{\lambda\psi_k}(\operatorname{soft}_{\lambda I_k}(x_k))\, e_k.$$

Proof. (i): see Proposition A.5(ii).
(ii): see Proposition A.5(iii).
(iii): see Proposition A.5(ii).
(iv): it is a consequence of the coercivity of g and the fact that both h and g are bounded from below.
(v): the smoothness of h implies the strong convexity of h^*, and the existence and uniqueness of $\bar{u}$, see [4, Theorems 15.13 and 18.15]. The equality $\bar{u} = -\nabla h(\bar{x})$ follows from [4, Proposition 26.1(iv)(b)].
(vi): it follows from A.5(iv) together with [9, Proposition 3.6]. ■

3. Extended support and finite identification

3.1. **Definition and basic properties.** We introduce the notion of extended support of a vector and prove some basic properties of the support of solutions of problem (P).

Definition 3.1. Let $x \in X$. The *extended support* of x is

$$\operatorname{esupp}(x) = \operatorname{supp}(x) \cup \{k \in \mathcal{N} \mid -\nabla h(x)_k \in \operatorname{bd} I_k\}.$$

It is worth noting that the notion of extended support depends on the problem (P), since its definition involves h (see Remark 3.4 for more details). It appears without a name in [22], and also in [14, 15, 17] for regularized least squares problems. Below we gather some results about the support and the extended support.

Proposition 3.2. Let $x \in \operatorname{dom}\partial f$, then $\operatorname{supp}(x)$ and $\operatorname{esupp}(x)$ are finite.

Proof. Let $x \in \operatorname{dom}\partial f = \operatorname{dom}\partial g$, and let $u \in \partial g(x)$ and let us start by verifying that $\operatorname{supp}(x)$ is finite. Let $x^* \in \partial g(x)$, and let $y = x + x^*$. Proposition 2.1(vi) implies that for all $k \in \operatorname{supp}(x)$, $\operatorname{prox}_{\psi_k} \circ \operatorname{soft}_{I_k}(y_k) \neq 0$. Lemma A.4 and the definition of $\operatorname{soft}_{I_k}$ imply that $y_k \notin I_k$, and in particular that $|y_k| \geqslant \omega$ for all $k \in \operatorname{supp}(x)$. Then we derive that

$$|\operatorname{supp}(x)| = \omega^{-2} \sum_{k\in\operatorname{supp}(x)} \omega^2 \leqslant \omega^{-2} \sum_{k\in\operatorname{supp}(x)} |y_k|^2 \leqslant \omega^{-2}\|y\|^2 < +\infty.$$

Next, we have to verify that J is finite, where $J = \{k \in \mathcal{N} \mid -\nabla h(x)_k \in \operatorname{bd} I_k\}$. If $\mathcal{N}$ is finite, this is trivial. Otherwise, we observe that $(\nabla h(x)_k)_{k\in\mathcal{N}} \in \ell^2(\mathcal{N})$, which both implies that $\nabla h(x)_k$ tends to 0 when $k \to +\infty$ in $\mathcal{N}$. Since $[-\omega, \omega] \subset I_k$, we deduce that J must be finite. ■

The following proposition clarifies the relationship between the support and the extended support for minimizers.

Proposition 3.3. Let $\bar{x} \in \operatorname{argmin} f$.

(i) If $0 \in \operatorname{qri} \partial f(\bar{x})$ then $\operatorname{esupp}(\bar{x}) = \operatorname{supp}(\bar{x})$.

Assume that ψ_k is differentiable on $\operatorname{dom} \partial\psi_k$, for all $k \in \mathcal{N}$. Then

(ii) $\operatorname{esupp}(\bar{x}) = \operatorname{supp}(\bar{x}) \Leftrightarrow 0 \in \operatorname{qri} \partial f(\bar{x})$.

Assume moreover that $\psi \equiv 0$. Then

(iii) $\operatorname{esupp}(\bar{x}) = \{k \in \mathcal{N} \mid -\nabla h(\bar{x})_k \in \operatorname{bd} I_k\}$.

(iv) There exists $J \subset \mathcal{N}$ such that $J = \operatorname{esupp}(\bar{x})$ for every $\bar{x} \in \operatorname{argmin} f$.

(v) $\operatorname{esupp}(\bar{x}) = \cup\{\operatorname{supp}(x) \mid x \in \operatorname{argmin} f\} \Leftrightarrow (\exists x \in \operatorname{argmin} f)\ 0 \in \operatorname{qri} \partial f(x)$.

Proof of Proposition 3.3. Since $\bar{x} \in \operatorname{argmin} f \subset \operatorname{dom} \partial g$, it follows from Proposition 3.2 that $\operatorname{supp}(\bar{x})$ is finite. Moreau-Rockafellar's sum rule [29, Theorem 3.30], Proposition A.5(iii), Proposition A.1(i) then yield

$$\partial f(\bar{x})_k = \nabla h(\bar{x})_k + \begin{cases} \partial\psi_k(\bar{x}_k) + \partial\sigma_{I_k}(\bar{x}_k) & \text{if } k \in \operatorname{supp}(\bar{x}) \\ I_k & \text{if } k \notin \operatorname{supp}(\bar{x}). \end{cases} \tag{2}$$

Since $\operatorname{supp}(\bar{x})$ is finite and $\partial\psi_k(\bar{x}_k) + \partial\sigma_{I_k}(\bar{x}_k)$ is a closed interval of $\mathbb{R}$, Proposition A.3 and Proposition A.1(iii) imply

$$(\forall k \in \mathcal{N}) \qquad (\operatorname{qri} \partial f(\bar{x}))_k = \nabla h(\bar{x})_k + \begin{cases} \operatorname{ri}\left(\partial\psi_k(\bar{x}_k) + \partial\sigma_{I_k}(\bar{x}_k)\right) & \text{if } k \in \operatorname{supp}(\bar{x}) \\ \operatorname{int} I_k & \text{if } k \notin \operatorname{supp}(\bar{x}). \end{cases} \tag{3}$$

(i): observe that

$$\begin{aligned} 0 \in \operatorname{qri} \partial f(\bar{x}) &\Rightarrow (\forall k \notin \operatorname{supp}(\bar{x})) \quad -\nabla h(\bar{x})_k \in \operatorname{int} I_k \\ &\Leftrightarrow \{k \in \mathcal{N} \mid x_k = 0 \text{ and } -\nabla h(\bar{x})_k \in \operatorname{bd} I_k\} = \emptyset \\ &\Leftrightarrow \operatorname{esupp}(\bar{x}) = \operatorname{supp}(\bar{x}). \end{aligned} \tag{4}$$

(ii): note that from $0 \in \partial f(\bar{x})$ and (2), we have $-\nabla h(\bar{x})_k \in \partial\psi_k(\bar{x}_k) + \partial\sigma_{I_k}(\bar{x}_k)$ for all $k \in \operatorname{supp}(\bar{x})$. But both ψ_k and σ_{I_k} are differentiable at $\bar{x}_k \neq 0$, so for all $k \in \operatorname{supp}(\bar{x})$, $\operatorname{qri}\left(\partial\psi_k(\bar{x}_k) + \partial\sigma_{I_k}(\bar{x}_k)\right) = \partial\psi_k(\bar{x}_k) + \partial\sigma_{I_k}(\bar{x}_k)$ holds. So we deduce from (4) that item (ii) holds.

(iii): observe that, via (2) and Proposition A.1(ii), for all $k \in \operatorname{supp}(\bar{x})$, $-\nabla h(\bar{x})_k \in \operatorname{bd} I_k$, meaning that indeed $\operatorname{esupp}(\bar{x}) = \{k \in \mathbb{N} \mid -\nabla h(\bar{x})_k \in \operatorname{bd} I_k\}$.

(iv): it follows from the uniqueness of $\nabla h(\bar{x})$, see Proposition 2.1(v).

(v): if there is some $x \in \operatorname{argmin} f$ such that $0 \in \operatorname{qri} \partial f(x)$, we derive from (ii) and (iv) that $\operatorname{esupp}(\bar{x}) = \operatorname{supp}(x)$. So, the inclusion $\operatorname{esupp}(\bar{x}) \subset \cup\{\operatorname{supp}(x') \mid x' \in \operatorname{argmin} f\}$ holds. The reverse inclusion comes directly from the definition of $\operatorname{esupp}(\bar{x})$ and (iv). For the reverse inclusion, assume that $\operatorname{esupp}(\bar{x}) = \cup\{\operatorname{supp}(x) \mid x \in \operatorname{argmin} f\}$ holds, and use the fact that $\operatorname{esupp}(\bar{x})$ is finite to apply Lemma A.9, and obtain some $x \in \operatorname{argmin} f$ such that $\operatorname{supp}(x) = \operatorname{esupp}(\bar{x})$. We then conclude that $0 \in \operatorname{qri} \partial f(x)$ using (iv) and (ii). ∎

Remark 3.4 (Extended support and active constraints)**.** Assume that $\psi = 0$. Since g^* is the indicator function of $\mathbb{B}_{\infty,\mathcal{I}}$, in this case, the dual problem (D) introduced in Proposition 2.1(v) can be rewritten as

$$\min_{\substack{u \in X \\ (\forall k \in \mathcal{N})\ u_k \in I_k}} \quad h^*(-u). \tag{D'}$$

This problem admits a unique solution $\bar{u} \in \mathbb{B}_{\infty,\mathcal{I}}$, and the set of active constraints at $\bar{u}$ is

$$\{k \in \mathcal{N} \mid \bar{u}_k \in \operatorname{bd} I_k\}.$$

Since $\bar{u} = -\nabla h(\bar{x})$ for any $\bar{x} \in \operatorname{argmin} f$ by Proposition 2.1(v), Proposition 3.3(iii) implies that the extended support for the solutions of (P) is in that case nothing but the set of active constraints for the solution of (D').

Remark 3.5 (Maximal support and interior solution)**.** If $\psi = 0$ and the following (weak) qualification condition holds

$$\text{(w-CQ)} \qquad (\exists x \in \operatorname{argmin} f)\ 0 \in \operatorname{qri} \partial f(x),$$

then, thanks to Lemma A.9 the extended support is the maximal support to be found among the solutions. If for instance h is the least squares loss on a finite dimensional space, it can be shown that the solutions having a maximal support are the ones belonging to the relative interior of the solution set [3, Theorem 2]. However, there are problems for which (w-CQ) does not hold. In such a case Proposition 3.3 implies that the extended support will be strictly larger than the maximal support (see Example 3.7). The gap between the maximal support and the extended support is equivalent to the lack of duality between (P) and (D).

Example 3.6. Let $g\colon \mathbb{R}^2 \to \mathbb{R} : x \mapsto \|x\|_1$ and $h\colon \mathbb{R}^2 \to \mathbb{R}\colon x \mapsto (x_1 - x_2 - 1)^2$. In this case, $\operatorname{argmin} f = [\bar{x}^1, \bar{x}^2]$, where $\bar{x}^1 = (0.5, 0)$ and $\bar{x}^2 = (0, -0.5)$, as can be seen in Figure 2. The solutions $\bar{x} \in]\bar{x}^1, \bar{x}^2[$ are the ones having the maximal support, since $\operatorname{supp}(\bar{x}) = \{1, 2\}$, and also satisfy $0 \in \operatorname{ri} \partial f(\bar{x})$. Instead, on the relative boundary of $\operatorname{argmin} f$ we have $\operatorname{supp}(\bar{x}^i) = \{i\}$ and $0 \notin \operatorname{ri} \partial f(\bar{x}^i)$ for $i \in \{1, 2\}$. This example is a one for which the extended support is the maximal support among the solutions.

Example 3.7. Let $g\colon \mathbb{R} \to \mathbb{R} : x \mapsto |x|$ and $h\colon \mathbb{R} \to \mathbb{R}\colon x \mapsto (x-1)^2/2$. Then $\operatorname{argmin} f = \{\bar{x}\}$, with $\bar{x} = 0$, as can be seen in Figure 2. The support of $\bar{x}$ is empty, and $0 \notin \operatorname{ri} \partial f(\bar{x}) = [-2, 0]$. In this case, condition (w-CQ) does not hold. This can also be seen from the dual problem $\min_{u \in [-1,1]} u^2/2 - u$, whose unique constraint is active at the solution $\bar{u} = -\nabla h(\bar{x}) = 1$, meaning that $\operatorname{esupp}(\bar{x}) = \{1\} \neq \operatorname{supp}(\bar{x})$.

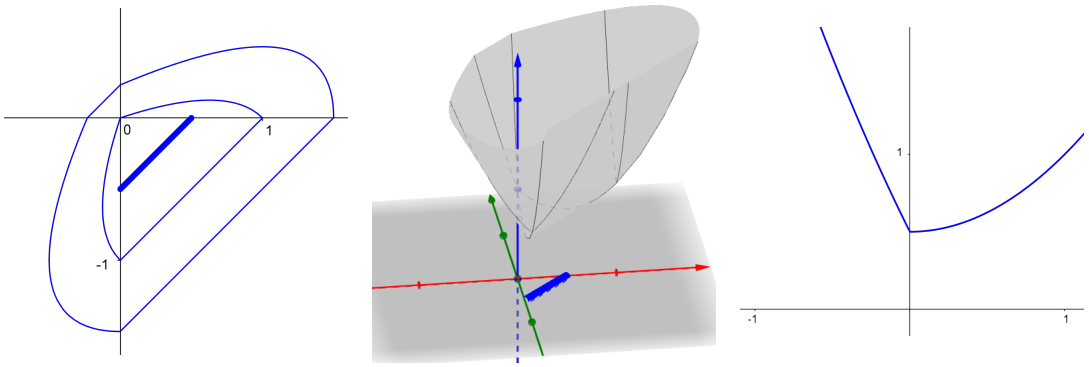

FIGURE 2. Left and center: respectively level sets and graph of f in Example 3.6, with $\operatorname{argmin} f$ in thick. Right: graph of f in Example 3.7.

3.2. **Finite identification.** A sparse solution $\bar{x}$ of problem (P) is usually approximated by means of an iterative procedure $(x^n)_{n \in \mathbb{N}}$. To obtain an interpretable approximation, a crucial property is that, after a finite number of iterations, the support of x^n stabilizes and is included in the support of $\bar{x}$. In that case, we say that the sequence $(x^n)_{n \in \mathbb{N}}$ *identifies* $\operatorname{supp}(\bar{x})$. The support identification property has been the subject of active research in the last years [22, 15, 26, 18, 17], and roughly speaking, in finite dimension it is known that support identification holds whenever $\bar{x}$ satisfies the qualification condition $0 \in \operatorname{qri} \partial f(\bar{x})$. But this assumption is often not satisfied in practice, in particular for noisy inverse problems (see e.g. [18]). In [22, 14], the case $g(x) = \|x\|_1$ is studied in finite dimension and it is shown that the extended support of $\bar{x}$ is identified even if the qualification condition does not hold. Thus, the qualification condition $0 \in \operatorname{qri} \partial f(\bar{x})$ is only used to ensure that the extended support coincides with the support (see Proposition 3.3).

In this section we extend these ideas to the setting of thresholding gradient methods in separable Hilbert spaces, and we show in Theorem 3.9 that indeed the extended support is

always identified after a finite number of iterations. For this, we need to introduce a quantity, which measures the stability of the dual problem (D).

Definition 3.8. We define the function $\rho\colon X \longrightarrow \mathbb{R}$ as follows:

$$(\forall u \in X) \quad \rho(u) = \inf_{u_k \in \operatorname{int} I_k} \operatorname{dist}(u_k, \operatorname{bd} I_k). \tag{5}$$

Also, given any $\bar{x} \in \operatorname{argmin} f$, we define $\rho_{\mathrm{sol}} = \rho(-\nabla h(\bar{x}))$.

It can be verified that $\rho(u) \in]0, +\infty[$ for all $u \in X$ (this is left in the Annex, see Proposition A.2). Moreover, ρ_{sol} is uniquely defined, thanks to Proposition 2.1(v).

Theorem 3.9 (Finite identification of the extended support)**.** *Let $(x^n)_{n\in\mathbb{N}}$ be generated by the Forward-Backward algorithm (FB), and let $\bar{x}$ be any minimizer of f. Then, the number of iterations for which the support of x^n is not included in $\operatorname{esupp}(\bar{x})$ is finite, and cannot exceed $\rho_{\mathrm{sol}}^{-2}\lambda^{-2}\|x^0 - \bar{x}\|^2$.*

Remark 3.10 (Optimality of the identification result)**.** Theorem 3.9 implies that after some iterations the inclusion $\operatorname{supp}(x^n) \subset \operatorname{esupp}(\bar{x})$ holds. Let us verify that it is impossible to improve the result, i.e. that in general we cannot identify a set smaller than $\operatorname{esupp}(\bar{x})$. In other words, is it true that

$$(\exists x^0 \in X)(\exists \bar{x} \in \operatorname{argmin} f)(\forall n \in \mathbb{N}) \quad \operatorname{supp}(x^n) = \operatorname{esupp}(\bar{x})? \tag{6}$$

If (w-CQ) holds, the answer is yes. Indeed, if there is $\bar{x} \in \operatorname{argmin} f$ such that $0 \in \operatorname{qri} \partial f(\bar{x})$, we derive from Proposition 3.3(i) that $\operatorname{esupp}(\bar{x}) = \operatorname{supp}(\bar{x})$. So by taking $x^0 = \bar{x}$, and using the fact that it is a fixed point for the Forward-Backward iterations, we conclude that $\operatorname{supp}(x^n) \equiv \operatorname{esupp}(\bar{x})$. If (w-CQ) does not hold, then this argument cannot be used, and it is not clear in general if there always exists an initialization which produces a sequence verifying (6). Consider for instance the function in Example 3.7. Taking $x^0 \in]0, +\infty[$ and a stepsize $\lambda \in]0, 1[$, the iterates are defined by $x^{n+1} = (1-\lambda)x^n$, meaning that for all $n \in \mathbb{N}$, $\operatorname{supp}(x^n) \equiv \{1\}$, which is exactly $\operatorname{esupp}(\bar{x})$. So in that case (6) holds true.

Proof. Let $\bar{x} \in \operatorname{argmin} f$, and let $E = X_{\operatorname{esupp}(\bar{x})}$ be the finite dimensional subspace of X supported by $\operatorname{esupp}(\bar{x})$. First define the "gradient step" operator

$$T_{\lambda h} = \mathrm{Id} - \lambda \nabla h,$$

so that the Forward-Backward iteration can be rewritten as $x^{n+1} = \operatorname{prox}_{\lambda g}(T_{\lambda h}(x^n))$. Proposition 2.1(vi) implies that for all $k \in \mathcal{N}$ and all $n \in \mathbb{N}^*$,

$$x_k^n = \operatorname{prox}_{\lambda \psi_k} \circ \operatorname{soft}_{\lambda I_k}(T_{\lambda h}(x^{n-1})_k). \tag{7}$$

Since $\bar{x}$ is a fixed point for the forward-backward iteration [4, Proposition 26.1(iv)], we also have

$$\bar{x}_k = \operatorname{prox}_{\lambda \psi_k} \circ \operatorname{soft}_{\lambda I_k}(T_{\lambda h}(\bar{x})_k). \tag{8}$$

Using the fact that $\operatorname{prox}_{\lambda\psi_k}$ is nonexpansive, and that $\operatorname{soft}_{\lambda I_k}$ is firmly non-expansive [4, Proposition 12.28], we derive

$$\begin{aligned}
\|x^n - \bar{x}\|^2 &= \sum_{k\in\mathcal{N}} |x_k^n - \bar{x}_k|^2 \leqslant \sum_{k\in\mathcal{N}} |\operatorname{soft}_{\lambda I_k}(T_{\lambda h}(x^{n-1})_k) - \operatorname{soft}_{\lambda I_k}(T_{\lambda h}(\bar{x})_k)|^2 \\
&\leqslant \sum_{k\in\mathcal{N}} |T_{\lambda h}(x^{n-1})_k - T_{\lambda h}(\bar{x})_k|^2 - |(\mathrm{Id} - \operatorname{soft}_{\lambda I_k})(T_{\lambda h}(x^{n-1})_k) - (\mathrm{Id} - \operatorname{soft}_{\lambda I_k})(T_{\lambda h}(\bar{x})_k)|^2 \\
&\leqslant \|T_{\lambda h}(x^{n-1}) - T_{\lambda h}(\bar{x})\|^2 - \sigma_{n,k}^2,
\end{aligned}$$

where

$$\sigma_{n,k} = |(\mathrm{Id} - \operatorname{soft}_{\lambda I_k})(T_{\lambda h}(x^{n-1})_k) - (\mathrm{Id} - \operatorname{soft}_{\lambda I_k})(T_{\lambda h}(\bar{x})_k)|.$$

Moreover, the gradient step operator T^G is non-expansive since $\lambda \in \left]0, 2L^{-1}\right[$ (see e.g. [24, Lemma 3.2]), so we end up with

$$(\forall n \in \mathbb{N}^*)(\forall k \in \mathcal{N}) \quad \|x^n - \bar{x}\|^2 \leqslant \|x^{n-1} - \bar{x}\|^2 - \sigma_{n,k}^2. \tag{9}$$

The key point of the proof is to get a nonnegative lower bound for $\sigma_{n,k}$ which is independent of n, when $x^n \notin E$.

Assume that there is some $n \in \mathbb{N}^*$ such that $x^n \notin E$. This means that there exists $k \in \mathcal{N} \setminus \operatorname{esupp}(\bar{x})$ such that $x^n_k \neq 0$. Also, since $\operatorname{supp}(\bar{x}) \subset \operatorname{esupp}(\bar{x})$, we must have $\bar{x}_k = 0$, meaning that $T_{\lambda h}(\bar{x})_k = -\lambda \nabla h(\bar{x})_k$. We deduce from (7), (8), and Lemma A.4, that

$$T_{\lambda h}(x^{n-1})_k \notin \lambda I_k \text{ and } T_{\lambda h}(\bar{x})_k \in \operatorname{int} \lambda I_k. \tag{10}$$

Since $\mathrm{Id} - \mathrm{soft}_{\lambda I_k}$ is the projection on λI_k, we derive from (10) that

$$\sigma_{n,k} = |\mathrm{proj}_{\lambda I_k}(T_{\lambda h}(x^{n-1})_k) - T_{\lambda h}(\bar{x})_k|.$$

Moreover $\mathrm{proj}_{\lambda I_k}(T_{\lambda h}(x^{n-1})_k) \in \operatorname{bd} I_k$, therefore by Definition 3.8 and (10), we obtain that

$$\sigma_{n,k} \geqslant \lambda \operatorname{dist}(\lambda^{-1} T_{\lambda h}(\bar{x})_k, \operatorname{bd} I_k) \geqslant \lambda \rho(\lambda^{-1} T_{\lambda h}(\bar{x})_k) = \lambda \rho(-\nabla h(\bar{x})_k) = \lambda \rho_{\mathrm{sol}}.$$

Plugging this into (9), we obtain

$$\forall n \in \mathbb{N}^*,\ x^n \notin E \ \Rightarrow\ \|x^n - \bar{x}\|^2 \leqslant \|x^{n-1} - \bar{x}\|^2 - \rho_{\mathrm{sol}}^2 \lambda^2. \tag{11}$$

Next note that the sequence $(x^n)_{n\in\mathbb{N}}$ is Féjer monotone with respect to the minimizers of f (see e.g. [20, Theorem 2.2]) — meaning that $(\|x^n - \bar{x}\|)_{n\in\mathbb{N}}$ is a decreasing sequence. Therefore the inequality (11) cannot hold an infinite number of times. More precisely, $x_n \notin E$ can hold for at most $\lambda^{-2}\rho_{\mathrm{sol}}^{-2}\|x^0 - \bar{x}\|^2$ iterations. ■

4. Strong convergence and rates

4.1. **General results for thresholding gradient methods.** Strong convergence of the iterates for the thresholding gradient algorithm was first stated in [11, Section 3.2] for $g = \|\cdot\|_1$, and then generalized to general thresholding gradient methods in [9, Theorem 4.5]. We provide a new and simple proof for this result, exploiting the "finite-dimensionality" provided by the identification result in Theorem 3.9.

Corollary 4.1 (Finite dimensionality for thresholding gradient methods)**.** Let $(x^n)_{n\in\mathbb{N}}$ be generated by a thresholding gradient algorithm. Then:

(i) There exists a finite set $J \subset \mathcal{N}$ such that $x^n \in X_J$ for all $n \in \mathbb{N}^*$.
(ii) x^n converges strongly to some $\bar{x} \in \operatorname{argmin} f$.

Proof. (i): let $x \in \operatorname{argmin} f$ and let

$$J = \operatorname{esupp}(x) \bigcup \{\operatorname{supp}(x^n) \mid n \in \mathbb{N}^*, x^n \notin X_{\operatorname{esupp}(x)}\},$$

and observe that it is finite, as a finite union of finite sets (see Proposition 3.2 and Theorem 3.9).

(ii): it is well known that $\operatorname{argmin} f \neq \emptyset$ implies that $(x^n)_{n\in\mathbb{N}}$ converges weakly towards some $\bar{x} \in \operatorname{argmin} f$ (see e.g. [20, Theorem 2.2]). In particular, $(x^n)_{n\in\mathbb{N}}$ is a bounded sequence in X. Moreover, (i) implies that $(x^n)_{n\in\mathbb{N}^*}$ belongs to X_J, which is finite dimensional. This two facts imply that $(x^n)_{n\in\mathbb{N}^*}$ is contained in a compact set of X with respect to the strong topology, and thus converges strongly. ■

Next we discuss the rate of convergence for the thresholding gradient methods. Beforehand, we briefly recall how the geometry of a function around its minimizers is related to the rates of convergence of the Forward-Backward algorithm.

Definition 4.2. Let $p \in [2, +\infty[$ and $\Omega \subset X$. We say that $\phi \in \Gamma_0(X)$ is p-conditioned on Ω if

$$(\exists \gamma_{\phi,\Omega} > 0)(\forall x \in \Omega) \quad \frac{\gamma_{\phi,\Omega}}{p} \operatorname{dist}(x, \operatorname{argmin} \phi)^p \leqslant \phi(x) - \inf \phi.$$

A p-conditioned function is a function which somehow behaves like $\operatorname{dist}(\cdot, \operatorname{argmin} \phi)^p$ on a set. For instance, strongly convex functions are 2-conditioned on $\Omega = X$, and the constant $\gamma_{\phi,X}$ is nothing but the constant of strong convexity. But the notion of p-conditioning is more general and also describes the geometry of functions having more than one minimizer. For instance in finite dimension, any positive quadratic function is 2-conditioned on $\Omega = X$, in

which case the constant $\gamma_{\phi,X}$ is the smallest *nonzero* eigenvalue of the hessian. This notion is interesting since it allows to get precise convergence rates for some algorithms (including the Forward-Backward one) [2]:

- sublinear rates if $p > 2$,
- linear rates if $p = 2$.

For more examples, related notions and references, we refer the interested reader to [16, 5, 20].

Corollary 4.1 highlights the fact that the behavior of the thresholding gradient method essentially depends on the conditioning of f on finitely supported subspaces. It is then natural to introduce the following notion of finite uniform conditioning.

Definition 4.3. Let $p \in [2, +\infty[$. We say that a function $\phi \in \Gamma_0(X)$ satisfies the *finite uniform conditioning property of order* p if, for every finite $J \subset \mathcal{N}$, $\forall \bar{x} \in \operatorname{argmin} \phi$, $\forall (\delta, r) \in]0, +\infty[^2$, ϕ is p-conditioned on $X_J \cap \mathbb{B}_X(\bar{x}, \delta) \cap S_\phi(r)$.

Remark 4.4. In this definition, we only need information about ϕ over supports J satisfying $\operatorname{argmin} \phi \cap X_J \neq \emptyset$. Indeed, if $\operatorname{argmin} \phi \cap X_J = \emptyset$, then ϕ is p-conditioned on $X_J \cap \mathbb{B}_X(\bar{x}, \delta) \cap S_\phi(r)$ for any (δ, r) and for all $p \in [2, +\infty[$ according to [20, Proposition 3.4].

In the following theorem, we illustrate how finite uniform conditioning guarantees global rates of convergence for the threshold gradient methods: linear rates if $p = 2$, and sublinear rates for $p > 2$. Note that these sublinear rates are better than the $O(n^{-1})$ rate guaranteed in the worst case.

Theorem 4.5 (Convergence rates for threshold gradient methods)**.** *Let $(x^n)_{n\in\mathbb{N}}$ be generated by the Forward-Backward algorithm (FB), and let $\bar{x} \in \operatorname{argmin} f$ be its (weak) limit. Then the following hold.*

(i) If f satisfies the finite uniform conditioning property of order 2, then there exist $\varepsilon \in]0, 1[$ and $C \in]0, +\infty[$, depending on (λ, f, x^0), such that

$$(\forall n \geqslant 1) \quad f(x^n) - \inf f \leqslant \varepsilon^n (f(x^0) - \inf f) \quad \text{and} \quad \|x^{n+1} - \bar{x}\| \leqslant C\sqrt{\varepsilon}^n.$$

(ii) If f satisfies the finite uniform conditioning property of order $p > 2$, then there exist $(C_1, C_2) \in]0, +\infty[^2$, depending on (λ, f, x^0), such that

$$(\forall n \geqslant 1) \quad f(x^n) - \inf f \leqslant C_1 n^{-\frac{p}{p-2}} \quad \text{and} \quad \|x_{n+1} - x_\infty\| \leqslant C_2 n^{-\frac{1}{p-2}}.$$

Proof. According to Corollary 4.1, there exists a finite set $J \subset \mathcal{N}$ such that for all $n \geqslant 1$, $x^n \in X_J$, and x^n converges strongly to $\bar{x} \in \operatorname{argmin} f$. Also, the decreasing and Féjer properties of the Forward-Backward algorithm (see e.g. [20, Theorem 2.2]) tells us that for all $n \in \mathbb{N}$, $x^n \in \mathbb{B}_X(\bar{x}, \delta) \cap S_f(r)$, by taking $\delta = \|x^0 - \bar{x}\|$ and $r = f(x^0) - \inf f$. Therefore, thanks to the finite uniform conditioning assumption, we can apply [20, Theorem 4.2] to the sequence $(x^{n+1})_{n\in\mathbb{N}} \subset \Omega = X_J \cap \mathbb{B}_X(\bar{x}, \delta) \cap S_f(r)$ and conclude. ∎

4.2. ℓ^1 **regularized least squares.** Let $A \colon X \to Y$ be a linear operator from X to a separable Hilbert space Y, and let $y \in Y$. In this section, we discuss the particular case when $h(x) = \frac{1}{2}\|Ax - y\|_Y^2$ and $\psi \equiv 0$. The function in (P) then becomes

$$X \ni x \mapsto f(x) = \|x\|_{1,\mathcal{I}} + \frac{1}{2}\|Ax - y\|_Y^2,$$

and the Forward-Backward algorithm specializes to the iterative soft-thresholding algorithm (ISTA). In this special case, linear convergence rates have been studied under additional assumptions on the operator A. A common one is injectivity of A or, more generally, the so-called Finite Basis Injectivity property (FBI) [8]. The FBI requires A to be injective once restricted to X_J, for any finite $J \subset \mathcal{N}$. It is clear that the FBI property implies that h is a strongly convex function once restricted to each X_J, meaning that the finite uniform conditioning of order 2 holds. So, the linear rates obtained in [8, Theorem 1] under the FBI assumption can be directly derived from Theorem 4.5. However, as can be seen in Theorem 4.5 , strong convexity is not necessary to get linear rates, and the finite uniform 2-conditioning is a sufficient condition (and

it is actually necessary, see [20, Proposition 4.18]). By using Li's Theorem on convex piecewise polynomials [25, Corollary 3.6], we show in Proposition 4.7 below that f satisfies a finite uniform conditioning of order 2 on finitely supported subsets, without doing *any assumption* on the problem. First, we need a technical Lemma which establishes the link between the conditioning of a function on a finitely supported space and the conditioning of its restriction to this space.

Lemma 4.6. Let $\phi \in \Gamma_0(X)$, let $m \in \mathbb{N}^*$ and let $J = \{k_1, \dots, k_m\} \subset \mathcal{N}$. Suppose that $\bar{x} \in \operatorname{argmin} \phi \cap X_J$. Let $\Xi : \mathbb{R}^m \to X_J \colon (u_1, ..., u_m) \mapsto \sum_{i=1}^m u_i e_{k_i}$. Assume that, for every $(\delta, r) \in]0, +\infty[^2$,

$$\phi_J = \phi \circ \Xi \in \Gamma_0(\mathbb{R}^m) \text{ is } p\text{-conditioned on } \mathbb{B}_{\mathbb{R}^m}(\Xi^{-1}(\bar{x}), \delta) \cap S_{\phi_J}(r)$$

Then ϕ is p-conditioned on X_J.

Proof. Assume without loss of generality that $k_1 < ... < k_m$. Also, observe that $\bar{x} \in X_J \cap \operatorname{argmin} \phi$ implies that $\bar{u} = \Xi^{-1}(\bar{x})$ is well-defined. By definition, $\inf \phi \leqslant \inf \phi_J$, and

$$\inf \phi = \phi(\bar{x}) = \phi \circ \Xi(\bar{u}) = \phi_J(\bar{u}) \geqslant \inf \phi_J,$$

which implies $\inf \phi = \inf \phi_J$. Also, we have

$$x \in \Xi(\operatorname{argmin} \phi_J) \Leftrightarrow x = \Xi(u) \text{ and } \phi_J(u) = \inf \phi_J \Leftrightarrow x \in X_J \text{ and } \phi(x) = \inf \phi,$$

meaning that $\Xi(\operatorname{argmin} \phi_J) = \operatorname{argmin} \phi \cap X_J$. Let $(\delta, r) \in]0, +\infty[^2$, and let $\Omega = X_J \cap \mathbb{B}_X(\bar{x}, \delta) \cap S_\phi(r)$. Since ϕ_J is p-conditioned on $\mathbb{B}_{\mathbb{R}^m}(\Xi^{-1}(\bar{x}), \delta) \cap S_{\phi_J}(r)$ there exists $\gamma \in]0, +\infty[$ such that

$$(\forall u \in \mathbb{B}_{\mathbb{R}^m}(\bar{u}, \delta) \cap S_{\phi_J}(r)) \quad \frac{\gamma}{p} \operatorname{dist}(u, \operatorname{argmin} \phi_J)^p \leqslant \phi_J(u) - \inf \phi_J. \tag{12}$$

Let $x = \Xi(u)$ in (12). Since $\|\Xi\| = 1$, it is easy to see that $\|x - \bar{x}\| \leqslant \delta$ and $\phi(x) - \inf \phi < r$. So we can rewrite (12) as:

$$(\forall x \in \Omega) \quad \frac{\gamma}{p} \operatorname{dist}(\Xi^{-1}x, \operatorname{argmin} \phi_J)^p \leqslant \phi(x) - \inf \phi.$$

It follows from $\Xi(\operatorname{argmin} \phi_J) = \operatorname{argmin} \phi \cap X_J$ that

$$(\forall x \in \Omega) \quad \phi(x) - \inf \phi \geqslant \frac{\gamma}{p} \operatorname{dist}(x, \operatorname{argmin} \phi \cap X_J)^p \geqslant \frac{\gamma}{p} \operatorname{dist}(x, \operatorname{argmin} f)^p.$$

Therefore ϕ is p-conditioned on Ω. ∎

Proposition 4.7 (Conditioning of ℓ^1 regularized least squares)**.** Let $(Y, \|\cdot\|_Y)$ be a separable Hilbert space, let $y \in Y$ and let $A \colon X \to Y$ be a bounded linear operator. In assumption (H) suppose that for every $k \in \mathcal{N}$, $I_k \in \mathcal{I}$ is bounded. Then $X \ni x \mapsto f(x) = \|x\|_{1,\mathcal{I}} + \frac{1}{2}\|Ax - y\|_Y^2$ has a finite uniform conditioning of order 2.

Proof. Let $J \subset \mathcal{N}$, $J = \{k_1, \dots, k_m\}$, with $k_1 < \dots < k_m$, and suppose that $\operatorname{argmin} f \cap X_J \neq \varnothing$. Define, using the same notation as in Lemma 4.6

$$h_J \colon \mathbb{R}^m \to \mathbb{R} \colon u \mapsto \frac{1}{2}\|A\Xi u - y\|_Y^2.$$

Define $A_J = A\Xi : \mathbb{R}^m \to Y$, and let $S_J = (A_J^* A_J)^{1/2}$, which verifies $R(S_J^*) = R(A_J^*)$. Thus, there exists $y_J \in \mathbb{R}^m$ such that $A_J^* y = S_J^* y_J$, so that we can rewrite

$$h_J(u) = \frac{1}{2}\|A_J u\|_{\mathbb{R}^m}^2 + \frac{1}{2}\|y\|_Y^2 - \langle A_J u, y\rangle_Y = \frac{1}{2}\|S_J u - y_J\|_{\mathbb{R}^m}^2 + \frac{1}{2}(\|y\|_Y^2 - \|y_J\|_{\mathbb{R}^m}^2). \tag{13}$$

Set $s_k = S_J e_k \in Y$. Then, (13) yields

$$f_J(u) = \sum_{i=1}^m \sigma_{I_{k_i}}(u_i) + \frac{1}{2}\sum_{i,j=1}^m \langle s_{k_i}, s_{k_j}\rangle_Y u_i u_j - \sum_{i=1}^m (S_J^* y_J)_i u_i + \frac{1}{2}\|y\|_Y^2.$$

Since the intervals I_k are bounded, their support functions are finite valued and piecewise linear, so f_J is a piecewise polynomial of degree two in $\mathbb{R}^m$. We then apply [25, Corollary 3.6]

to derive that f_J is 2-conditioned on $S_{f_J}(r)$, for any $r \in]0, +\infty[$. We conclude by using Lemma 4.6. ■

Combining Theorem 4.5 and Proposition 4.7, we can now state our main result concerning the linear rates of ISTA.

Theorem 4.8 (Linear convergence for the iterative soft thresholding)**.** *Under the assumptions of Proposition 4.7, let $(x^n)_{n\in\mathbb{N}}$ be the sequence generated by the forward-backward algorithm applied to f. Then $(x^n)_{n\in\mathbb{N}}$ converges strongly to some $\bar{x} \in \operatorname{argmin} f$, and there exists two constants $\varepsilon \in]0,1[$ and $C \in]0, +\infty[$, depending on $(\lambda, L, x^0, \mathcal{I}, A, y)$, such that*

$$(\forall n \geqslant 1) \quad f(x^n) - \inf f \leqslant \varepsilon^n (f(x^0) - \inf f) \quad \text{and} \quad \|x^{n+1} - \bar{x}\| \leqslant C\sqrt{\varepsilon}^n.$$

Remark 4.9 (On the linear rates)**.** The convergence rate for the iterative soft-thresholding has been a subject of interest since years, and have been obtained only under additional assumptions on A [8]. Theorem 4.8 closes the question about the linear rates, by proving that they always hold. However, there are still several open problems, related to the estimation of the constant appearing in these linear rates. This is related to the estimation of the constant $\gamma_{f,\Omega}$ in Definition 4.2, when $\Omega = S_f(r) \cap X_J$ for some finite $J \subset \mathcal{N}$. Up to now, the only available result is based on Hoffman's lemma, which doesn't allow for explicit lower bounds on $\gamma_{f,\Omega}$ [28, 6]. Having a tight lower bound for $\gamma_{f,\Omega}$, depending on A restricted to X_J, would be of interest to go in this direction.

4.3. $\ell^1 + \ell^p$ **regularized least squares.** We are now interested in $\ell^1 + \ell^p$-regularizers, i.e. when

$$g(x) = \|x\|_{1,\mathcal{I}} + \frac{1}{p}\|x\|^p, \text{ with } \|x\|^p = \sum_{k\in\mathbb{N}} |x_k|^p, \quad p > 1.$$

The case $p = 2$ is also known as *elastic net regularization* and has been proposed in [34]. The elasitc-net penalty has been studied by the statistical machine learning community as an alternative to the ℓ^1 regularization in variable selection problems where there are highly correlated features and all the relavant ones have to be identified [13]. See also [10] for the case $p < 2$. Note that the proximal operator of such g can be computed explicitly when $p \in \{4/3, 3/2, 2, 3, 4\}$ (see [9]).

Proposition 4.10 (Geometry of $(\ell^1 + \ell^p)$ regularized least squares)**.** Let $p \in]1, +\infty[$, let $(Y, \|\cdot\|_Y)$ be a separable Hilbert space, let $y \in Y$ and let $A\colon X \to Y$ be a bounded linear operator. In assumption (H) suppose that for every $k \in \mathcal{N}$, $I_k \in \mathcal{I}$ is bounded. Then $f\colon X \to \mathbb{R}\colon x \mapsto \|x\|_{1,\mathcal{I}} + \frac{1}{p}\|x\|_p^p + \frac{1}{2}\|Ax - y\|_Y^2$ has a finite uniform conditioning of order $\max\{2, p\}$.

Proof. Let $J \subset \mathcal{N}$, $m = |J|$ and $p' = \max\{p, 2\}$. We define, by using the same notation as in Lemma 4.6,

$$g_J(u) = \sum_{i=1}^m \sigma_{I_{k_i}}(u_i) + \frac{1}{p}|u_i|^p \text{ and } h_J(u) = \frac{1}{2}\|A\Xi u - y\|_Y^2.$$

We are going to prove that $f_J = g_J + h_J$ is p'-conditioned on $\mathbb{B}_{\mathbb{R}^m}(\bar{u}, \delta)$, for any $\delta > 0$. To do so, we will apply to f_J the sum rule in Theorem A.7, which requires two hypotheses. We must verify that the functions g_J and h_J are conditioned up to linear perturbations (see equation (18)), and that the qualification condition in (19) holds, namely (since $\mathbb{R}^m$ is finite dimensional the strong relative interior coincides with the relative interior):

$$0 \in \operatorname{ri}\left(\partial g_J^*(-\nabla h_J(\bar{u})) - \partial h_J^*(\nabla h_J(\bar{u}))\right). \tag{14}$$

According to (13), $\partial h_J^*(\nabla h_J(\bar{u})) = \bar{u} + \ker S_J$. Also, according to [4, Proposition 13.30 & Example 13.27(iii)], we have, for every $v \in \mathbb{R}^m$, $g_J^*(v) = \sum_{i=1}^m \frac{1}{q}\operatorname{dist}(v_i, I_{k_i})^q$, with $q = p/(p-1)$. Since $t \mapsto |t|^q$ is continuously differentiable on $\mathbb{R}$, [4, Example 17.33 and Proposition 17.31(ii)]) imply that g_J^* is Gâteaux differetiable on $\mathbb{R}^m$. This, together with the fact that $\operatorname{ri} \ker S_J = \ker S_J$, means that (14) is equivalent to

$$0 \in \operatorname{ri}\left(\nabla g_J^*(-\nabla h_J(\bar{u})) - \bar{u} + \ker S_J\right) = \nabla g_J^*(-\nabla h_J(\bar{u})) - \bar{u} + \ker S_J. \tag{15}$$

The latter inclusion holds true, since $\bar{u} \in \operatorname{argmin} f_J$ is equivalent to $\bar{u} \in \nabla g_J^*(-\nabla h_J(\bar{u}))$. Thus, it only remains to prove that $\bar{h}_J = h_J - \langle \ell, \cdot \rangle$ and $\bar{g}_J = g_J - \langle \ell, \cdot \rangle$ are respectively 2 and p'-conditioned on $\mathbb{B}_{\mathbb{R}^m}(\bar{u}, \delta)$, for ℓ being respectively in $R(\nabla h_J)$ and $R(\partial g_J)$.

Let us start with $\bar{h}_J$. According to (13), $\bar{h}_J$ is a positive quadratic function being bounded from below, so it is 2-conditioned on $\mathbb{R}^m$, with $\gamma_{\bar{h}_J, \mathbb{R}^m}$ being the smallest nonzero eigenvalue of S_J. Next, $\ell \in R(\partial g_J)$ implies that there exists $v \in X$ such that $\ell \in \partial \bar{g}_J(v)$. Then, $0 \in \partial \bar{g}_J((v))$, and this implies that v is a minimizer of $\bar{g}_J$. It is also unique since g_j is strictly convex. If $v \notin \mathbb{B}_{\mathbb{R}^m}(\bar{u}, \delta)$, then $\bar{g}_J$ is automatically p'-conditioned on $\mathbb{B}_{\mathbb{R}^m}(\bar{u}, \delta)$, see for instance [20, Proposition 3.3]. Assume then that $v \in \mathbb{B}_{\mathbb{R}^m}(\bar{u}, \delta)$, and use [10, Proposition A.9] to obtain the existence of $\gamma \in]0, +\infty[$ such that

$$(\forall u \in \mathbb{B}_{\mathbb{R}^m}(\bar{u}, \delta))(\forall i \in \{1, ..., m\}) \quad \frac{\gamma}{p'}|u_i - v_i|^{p'} \leqslant \frac{1}{p}|u_i|^p - \frac{1}{p}|v_i|^p - (u_i - v_i)\operatorname{sgn}(v_i)|v_i|^{p-1}.$$

Summing the above inequality over i, and using the fact that $\|\cdot\|_2^{p'} \leqslant \max\{1, m^{(p-2)/2}\}\|\cdot\|_{p'}^{p'}$, we derive by taking $\gamma' = \gamma \max\{1, m^{(p-2)/2}\}^{-1}$ that for all $u \in \mathbb{B}_{\mathbb{R}^m}(\bar{u}, \delta)$:

$$\frac{\gamma'}{p'}\operatorname{dist}(u, \operatorname{argmin} \bar{g}_J)^{p'} \leqslant \sum_{i=1}^m \frac{1}{p}|u_i|^p - \frac{1}{p}|v_i|^p - (u_i - v_i)\operatorname{sgn}(v_i)|v_i|^{p-1}. \tag{16}$$

Introduce the following constant: $\omega_i = \sup I_{k_i}$ if $\ell_i > \sup I_{k_i}$, $\omega_i = |\ell_i|$ if $\ell_i \in I_{k_i}$, and $w_i = -\inf I_{k_i}$ if $\ell_i < \inf I_{k_i}$. By making use of the first order condition at $v = \operatorname{argmin} \bar{g}_J$, it can be verified that

$$(\forall_i \in \{1, ..., m\}) \quad |v_i|^{p-1} = |\ell_i| - \omega_i,\ \operatorname{sgn}(v_i) = \operatorname{sgn}(\ell_i) \text{ and } \sigma_{I_{k_i}}(v_i) = \omega_i|v_i|.$$

So we can deduce that

$$\begin{aligned} & \frac{1}{p}|u_i|^p - \frac{1}{p}|v_i|^p - (u_i - v_i)\operatorname{sgn}(v_i)|v_i|^{p-1} \\ = \ & \frac{1}{p}|u_i|^p - \frac{1}{p}|v_i|^p - (u_i - v_i)(\ell_i - \operatorname{sgn}(v_i)\omega_i) \\ = \ & \frac{1}{p}|u_i|^p - \frac{1}{p}|v_i|^p - u_i\ell_i + v_i\ell_i - \sigma_{I_{k_i}}(v_i) + u_i\operatorname{sgn}(\ell_i)\omega_i. \end{aligned}$$

This, combined with (16), leads to

$$\frac{\gamma'}{p'}\operatorname{dist}(u, \operatorname{argmin} \bar{g}_J)^{p'} \leqslant \bar{g}_J(u) - \inf \bar{g}_J + \sum_{i=1}^m -\sigma_{I_{k_i}}(u_i) + u_i\operatorname{sgn}(\ell_i)\omega_i \leqslant \bar{g}_J(u) - \inf \bar{g}_J,$$

where the last inequality comes from the fact that $\operatorname{sgn}(\ell_i)\omega_i \in I_{k_i}$. So we proved that $\bar{g}_J$ is p'-conditioned on $\mathbb{B}_{\mathbb{R}^m}(\bar{u}, \delta)$. Theorem A.7 then yields that f_J is p'-conditioned on $\mathbb{B}_{\mathbb{R}^m}(\bar{u}, \delta)$. We conclude the proof applying Lemma 4.6. ■

Combining Theorem 4.5 and Proposition 4.10, be obtain rates for the corresponding thresholding gradient method.

Theorem 4.11. *Under the assumptions of Proposition 4.10, let $(x^n)_{n\in\mathbb{N}}$ be the sequence generated by the forward-backward algorithm applied to f. Then $(x^n)_{n\in\mathbb{N}}$ converges strongly to some $\bar{x} \in \operatorname{argmin} f$. If $p \in]1, 2[$, there exists two constants $\varepsilon \in]0, 1[$ and $C \in]0, +\infty[$, depending on $(\lambda, L, x^0, \mathcal{I}, A, y, p)$, such that*

$$(\forall n \geqslant 1) \quad f(x^n) - \inf f \leqslant \varepsilon^n(f(x^0) - \inf f) \quad \text{and} \quad \|x^{n+1} - \bar{x}\| \leqslant C\sqrt{\varepsilon}^n.$$

If $p \in]2, +\infty[$, there exists two constants $(C_1, C_2) \in]0, +\infty[^2$, depending on $(\lambda, L, x^0, \mathcal{I}, A, y, p)$, such that

$$(\forall n \geqslant 1) \quad f(x^n) - \inf f \leqslant C_1 n^{-\frac{p}{p-2}} \quad \text{and} \quad \|x_{n+1} - x_\infty\| \leqslant C_2 n^{-\frac{1}{p-2}}.$$

5. Conclusion and perspectives

In this paper we study and highlight the importance of the notion of extended support for minimization problems with sparsity inducing separable penalties. An identification result, together with uniform conditioning on finite dimensional sets, allow us to generalize and revisit classic convergence results for thresholding gradient methods, from a novel and different perspective, while further providing new convergence rates.

An interesting direction for future research would be to go beyond separable penalties, in particular extending our results to regularizers promoting structured sparsity [27], such as group lasso. A reasonable approach would be to extend the primal-dual arguments in [18] to the infinite-dimensional setting. A more challenging research direction seems the extension of our results to gridless problems [17]. Indeed, our analysis relies on the fact that the variables (signals) we consider are supported on a grid (indexed by $\mathcal{N} \subset \mathbb{N}$), which allows to use finite-dimensional arguments. Such an extension would require to work on Banach spaces of functions or of measures, and seems an interesting venue for future research.

Acknowledgements

This material is supported by the Center for Brains, Minds and Machines, funded by NSF STC award CCF-1231216, and the Air Force project FA9550-17-1-0390. G. Garrigos is supported by the European Research Council (ERC project NORIA), and part of his work was done while being a postdoc at the LCSL-IIT@MIT. L. Rosasco acknowledges the financial support of the Italian Ministry of Education, University and Research FIRB project RBFR12M3AC. S. Villa is supported by the INDAM GNAMPA research project 2017 Algoritmi di ottimizzazione ed equazioni di evoluzione ereditarie. L. Rosasco and S. Villa acknowledge the financial support from the EU project 777826 - NoMADS.

Appendix A. Annex

A.1. **Closure, interior, boundary.**

Proposition A.1. Let $C \subset X$ be a closed convex set. Then:

(i) $\partial\sigma_C(0) = C$.

(ii) For all $d \in X \setminus \{0\}$, $\partial\sigma_C(d) \subset \operatorname{bd} C$.

Assume moreover that $\operatorname{int} C \neq \emptyset$.

(iii) $\operatorname{int} C = \operatorname{qri} C$.

Proof. (i): see [4, Example 16.34].
(ii): see [4, Proposition 7.3 & Theorem 7.4].
(iii): see [4, Fact 6.14]. ∎

Proposition A.2. Let $\mathcal{I} = (I_k)_{k\in\mathcal{N}}$ be a collection of closed proper intervals of $\mathbb{R}$, and suppose that $[-\omega, \omega] \subset I_k$ for all $k \in \mathcal{N}$. For every $x \in X$, let

$$\rho(x) = \inf_{x_k \in \operatorname{int} I_k} \operatorname{dist}(x_k, \operatorname{bd} I_k). \tag{17}$$

Then the following hold

(i) For every $x \in X$, $\quad \rho(x) \in]0, +\infty[$;

(ii) $\operatorname{int}\left(\bigoplus_{k\in\mathcal{N}} I_k\right) = \bigoplus_{k\in\mathcal{N}} \operatorname{int} I_k$;

Proof. (i) Let $x \in X$. If $\mathcal{N}$ is finite the statement follows immediately. If $\mathcal{N}$ is infinite, since $|x_k|$ tends to 0 when $k \to +\infty$, there exists $K \in \mathcal{N}$ such that for all $k \geqslant K$, $|u_k| \leqslant \omega/2$. Now, consider the following subsets of $\mathcal{N}$

$$J = \{k \in \mathcal{N} \mid u_k \in \operatorname{int} I_k\},\ J_F = J \cap \{0, \dots, K-1\},\ J_\infty = J \setminus J_F,$$

which are defined in such a way that $\rho(x) = \inf_{k\in J} \operatorname{dist}(u_k, \operatorname{bd} I_k)$ and $J = J_F \sqcup J_\infty$. Observe that $\rho(x) \leqslant \operatorname{dist}(x_K, \operatorname{bd} I_K) < +\infty$ since $\operatorname{bd} I_K \neq \emptyset$, so we only need to show that $\rho(x) > 0$. Since J_F is finite and $x_k \in \operatorname{int} I_k$ for all $k \in J_F$, we have $\operatorname{dist}(x_k, \operatorname{bd} I_k) > 0$ for all $k \in J_F$. So we

deduce that $\inf_{k\in J_F} \operatorname{dist}(x_k, \operatorname{bd} I_k) > 0$. On the other hand, for any $k \in J_\infty$, we have $|u_k| \leqslant \omega/2$, while $[-\omega, \omega] \subset I_k$, therefore $\operatorname{dist}(x_k, \operatorname{bd} I_k) \geqslant \omega/2$, and $\rho(x) = \inf_{k\in J} \operatorname{dist}(x_k, \operatorname{bd} I_k) > 0$.

(ii): let $x \in \operatorname{int} \bigoplus_{k\in\mathcal{N}} I_k$. We are going to show that $x_k \in \operatorname{int} I_k$ for all $k \in \mathcal{N}$. By assumption, there exists $\delta \in]0, +\infty[$ such that $\mathbb{B}_X(x, \delta) \subset \bigoplus_{k\in\mathcal{N}} I_k$. Let $k \in \mathcal{N}$, and let us show that $[x_k - \delta, x_k + \delta] \subset I_k$. Let $y_k \in [x_k - \delta, x_k + \delta]$, and define $\bar{x} \in X$ such that $\bar{x}_k = y_k$ and $\bar{x}_i = x_i$ for every $i \neq k$. Then we derive $\|x - \bar{x}\| = |x_k - y_k| = \delta$, whence $\bar{x} \in \mathbb{B}(x, \delta) \subset \bigoplus_{k\in\mathcal{N}} I_k$. This implies that $y_k \in I_k$, which proves that $x_k \in \operatorname{int} I_k$. Now, we let $x \in \bigoplus_{k\in\mathcal{N}} \operatorname{int} I_k$, and we show that $x \in \operatorname{int}(\bigoplus_{k\in\mathcal{N}} I_k)$. By (i), $\rho(x) > 0$ and, for every $k \in \mathcal{N}$, $x_k \in \operatorname{int} I_k$ by assumption. Let $\eta \in]0, \rho[$. Since $\operatorname{dist}(x_k, \operatorname{bd} I_k) \geqslant \rho(x)$, we derive $[x_k - \eta, x_k + \eta] \subset I_k$. On the other hand, the non-expansiveness of the projection implies that $\|x_k - p_k\| \leqslant \|x_k - y_k\| \leqslant \eta < \rho$, which leads to a contradiction. Therefore $\mathbb{B}_X(x, \eta) \subset \bigoplus_{k\in\mathcal{N}} [x_k - \eta, x_k + \eta] \subset \bigoplus_{k\in\mathcal{N}} I_k$. This yields $x \in \operatorname{int} \bigoplus_{k\in\mathcal{N}} I_k$. ■

Proposition A.3 (Quasi relative interior of infinite products)**.** Let $\mathcal{I} = (I_k)_{k\in\mathbb{N}}$ be a collection of closed intervals of $\mathbb{R}$. Let $J \subset \mathcal{N}$ be a finite set, and suppose that $[-\omega, \omega] \subset I_k$ for all $k \in \mathcal{N} \setminus J$. Then

$$\operatorname{qri} \bigoplus_{k\in\mathcal{N}} I_k = \bigoplus_{k\in\mathcal{N}} \operatorname{ri} I_k.$$

Proof. Assume $\mathcal{N}$ is infinite and set $J_\infty = \mathcal{N} \setminus J$. We can then write

$$\begin{aligned}
\operatorname{qri} \bigoplus_{k\in\mathcal{N}} I_k &= \operatorname{qri}\Big((\bigoplus_{k\in J_\infty} I_k) \oplus (\bigoplus_{k\in J} I_k) \Big) && \text{because } J \text{ is finite,} \\
&= \operatorname{qri} (\bigoplus_{k\in J_\infty} I_k) \oplus (\bigoplus_{k\in J} \operatorname{qri} I_k) && \text{by [7, Proposition 2.5],} \\
&= (\bigoplus_{k\in J_\infty} \operatorname{int} I_k) \oplus (\bigoplus_{k\in J} \operatorname{qri} I_k) && \text{by Proposition A.2,} \\
&= \bigoplus_{k\in\mathcal{N}} \operatorname{ri} I_k && \text{by Proposition A.1(iii).}
\end{aligned}$$

■

A.2. **Functions.**

Lemma A.4. Let $\psi \in \Gamma_0(X)$ be differentiable at $0 \in \operatorname{argmin} \psi$ and let $x \in X$. Then

$$x = 0 \Leftrightarrow \operatorname{prox}_\psi(x) = 0.$$

Proof. $\operatorname{prox}_\psi(x) = 0 \Leftrightarrow (Id + \partial\psi)^{-1}(x) = 0 \Leftrightarrow x \in 0 + \partial\psi(0) \Leftrightarrow x = \nabla\psi(0) \Leftrightarrow x = 0.$ ■

Proposition A.5. Let $g_k \in \Gamma_0(\mathbb{R})$ with $\inf g_k = g_k(0) = 0$ for all $k \in \mathcal{N}$. Define $g\colon X \to \mathbb{R} \cup \{+\infty\}\colon x \mapsto \sum_{k\in\mathcal{N}} g_k(x_k)$. Then:

(i) $g \in \Gamma_0(X)$.
(ii) $\operatorname{dom} \partial g = \{x \in X \mid \bigoplus_{k\in\mathcal{N}} \partial g_k(x_k) \neq \emptyset\}$.
(iii) For all $x \in \operatorname{dom} \partial g$, $\partial g(x) = \bigoplus_{k\in\mathcal{N}} \partial g_k(x_k)$.
(iv) For all $x \in X$, $\operatorname{prox}_g(x) = \sum_{k\in\mathcal{N}} \operatorname{prox}_{g_k}(x_k) e_k$.

Proof. (i): g is convex by definition. It is proper because $g(0) = 0$ and $g \geqslant 0$. Fatou's lemma implies that g is lower semicontinuous.

(ii)-(iii): follow directly from the fact that

$$\begin{aligned}
(\forall (x^*, x) \in X^2) \quad x^* \in \partial g(x) \quad &\Leftrightarrow \quad (\forall y \in X) \quad g(y) - g(x) - \langle x^*, y - x\rangle \geqslant 0 \\
&\Leftrightarrow \quad (\forall y \in X) \quad \sum_{k\in\mathbb{N}} g_k(y_k) - g_k(x_k) - \langle x_k^*, y_k - x_k\rangle \geqslant 0 \\
&\Leftrightarrow \quad (\forall k \in \mathcal{N}) \quad x_k^* \in \partial g_k(x_k),
\end{aligned}$$

where the last equivalence holds by taking for all $k \in \mathcal{N}$ $y_i = x_i$ if $i \neq k$.

(iv): let $(x,p) \in X^2$. It follows from (iii) that

$$\begin{aligned} p = \operatorname{prox}_g(x) &\iff p - x \in \partial g(p) \\ &\implies (\forall k \in \mathcal{N}) \quad p_k - x_k \in \partial g_k(p_k) \\ &\iff (\forall k \in \mathcal{N}) \quad p_k = \operatorname{prox}_{g_k}(x_k). \end{aligned}$$

■

Proposition A.6. Let $\mathcal{I} = (I_k)_{k\in\mathbb{N}}$ is a family of proper closed interval of $\mathbb{R}$. Let, for every $x \in X$, $g(x) = \sum_{k\in\mathcal{N}} \sigma_{I_k}(x_k)$. Then the following hold.

(i) g is coercive if and only if $0 \in \operatorname{int} I_k$ for all $k \in \mathcal{N}$.

Assume moreover that there exists $\omega > 0$ such that $[-\omega,\omega] \subset I_k$ for all $k \in \mathcal{N}$. Then

(ii) $g \in \Gamma_0(X)$ is coercive and g is the support function of $\mathbb{B}_{\infty,\mathcal{I}} = \bigoplus_{k\in\mathcal{N}} I_k$,

(iii) $\operatorname{dom}\partial g = c_{00}$ and $\operatorname{dom}\partial g^* = \mathbb{B}_{\infty,\mathcal{I}}$,

(iv) for every $x \in X$, and for every $\lambda > 0$, $\operatorname{prox}_{\lambda g}(x) = \Big(x_k - \lambda \operatorname{proj}_{I_k}(\lambda^{-1}x_k)\Big)_{k\in\mathcal{N}}$.

Proof. (i): observe that

$$\begin{aligned} g \text{ is coercive} &\Leftrightarrow (\forall k \in \mathcal{N}) \quad \sigma_{I_k} \text{ is coercive} \quad (\text{take } x_k = 0 \text{ except for one index } k) \\ &\Leftrightarrow (\forall k \in \mathcal{N}) \quad 0 \in \operatorname{int} \operatorname{dom} \sigma_{I_k}^* \quad \text{by [4, Proposition 14.16]} \\ &\Leftrightarrow (\forall k \in \mathcal{N}) \quad 0 \in \operatorname{int} I_k \quad \text{since } \sigma_{I_k}^* = \delta_{I_k}. \end{aligned}$$

(ii): assume that $\mathcal{N}$ is infinite. Item (i) implies that $g \in \Gamma_0(X)$ and is coercive. To prove that g is the support function of $\mathbb{B}_{\infty,\mathcal{I}}$, we will show that g^* is its indicator function. Let $x^* \in X$. Then

$$\begin{aligned} g^*(x^*) = \sup_{x\in X} \langle x^*, x\rangle - g(x) &= \sup_{x\in X} \sum_{k\in\mathcal{N}} \langle x_k^*, x_k\rangle - \sigma_{I_k}(x_k) \\ &\leqslant \sum_{k\in\mathcal{N}} \sup_{x_k\in X_k} \langle x_k^*, x_k\rangle - \sigma_{I_k}(x_k) = \sum_{k\in\mathcal{N}} \sigma_{I_k}^*(x_k^*) = \delta_{\mathbb{B}_{\infty,\mathcal{I}}}(x^*) \end{aligned}$$

To prove the converse inequality, since $x^* \in X$, there exists some $K \in \mathcal{N}$ such that for all $k \geqslant K$, $\|x_k^*\| < \omega$, meaning that $x_k^* \in I_k$, and therefore $\delta_{I_k}(x_k^*) = 0$. Let $J_K = \{0,\dots,K-1\}$. Since we deal with a finite sum,

$$\delta_{\mathbb{B}_{\infty,\mathcal{I}}}(x^*) \quad = \quad \sum_{k\in J_K} \delta_{I_k}(x_k^*) = \sum_{k\in J_K} \sup_{x_k\in\mathbb{R}} \langle x_k^*, x_k\rangle - \sigma_{I_k}(x_k) = \sup_{x\in X_{J_K}} \sum_{k\in J_K} \langle x_k^*, x_k\rangle - \sigma_{I_k}(x_k).$$

Moreover, setting $J_\infty = \mathcal{N} \setminus J_K$:

$$\sup_{x\in X_{J_\infty}} \sum_{k\in J_\infty} \langle x_k^*, x_k\rangle - \sigma_{I_k}(x_k) \geqslant 0,$$

and this yields

$$\delta_{\mathbb{B}_{\infty,\mathcal{I}}}(x^*) \leqslant \sup_{x\in X_{J_K}} \sum_{k\in J_K} \langle x_k^*, x_k\rangle - \sigma_{I_k}(x_k) + \sup_{x\in X_{J_\infty}} \sum_{k\in J_\infty} \langle x_k^*, x_k\rangle - \sigma_{I_k}(x_k) = g^*(x^*).$$

(iii): assume that $\mathcal{N}$ is infinite. The equality $\operatorname{dom}\partial g^* = \mathbb{B}_{\infty,\mathcal{I}}$ follows from (ii). It remains to show that $\operatorname{dom}\partial g = c_{00}$. Let $x \in \operatorname{dom}\partial g$. By Proposition A.5(ii) there exists $x^* \in \bigoplus_{k\in\mathcal{N}} \partial\sigma_{I_k}(x_k)$. For all $k \in \mathcal{N}$, Proposition A.1(i)-(ii) yields that $\partial\sigma_{I_k}(x_k) = I_k$ if $x_k = 0$, and $\partial\sigma_{I_k}(x_k) \subset \operatorname{bd} I_k$ if $x_k \neq 0$. Assume by contradiction that $x \notin c_{00}$, i.e. there exists $k_n \to +\infty$ such that $x_{k_n} \neq 0$ for all $n \in \mathbb{N}$. Then, it follows that $x_{k_n}^* \in \operatorname{bd} I_{k_n}$ for all $n \in \mathbb{N}$, and therefore $\|x_{k_n}^*\| \geqslant \omega$, which contradicts the fact that $x^* \in X$. Now, let $x \in c_{00}$ and let $K \in \mathcal{N}$ be such that $x_k = 0$ for all $k \geqslant K$ and let $\mathcal{N}_K = \mathcal{N} \cap \{0,\dots,K\}$. By Proposition A.1(i) $\partial\sigma_{I_k}(x_k) = I_k \ni 0$ for all $k \geqslant K$, therefore

$$\emptyset \neq \bigoplus_{k\in\mathcal{N}_K} \partial\sigma_{I_k}(x_k) \subset \partial g(x).$$

(iv): is a direct consequence of Proposition A.5(iv) and Moreau's identity [4, Theorem 14.3.ii]. ■

We recall a sum rule for conditioning, obtained in [20, Theorem 3.1].

Theorem A.7. *Let $f = g + h$, where $g \in \Gamma_0(X)$ and $h \in \Gamma_0(X)$ is of class C^1. Let $\Omega \subset X$. Assume that there exists $\bar{x} \in \operatorname{argmin} f$ such that, for $\bar{v} = -\nabla h(\bar{x})$ and $(p_1, p_2) \in [1, +\infty[^2$,*

$$g - \langle \bar{v}, \cdot \rangle \text{ is } p_1\text{-conditioned on } \Omega \text{ and } h + \langle \bar{v}, \cdot \rangle \text{ is } p_2\text{-conditioned on } \Omega. \tag{18}$$

Suppose that

$$0 \quad \in \quad \operatorname{sri}(\partial g^*(\bar{v}) - \partial h^*(-\bar{v})), \tag{19}$$

and let $p = \max\{p_1, p_2\}$. Then, for any $\delta \in]0, +\infty[$, f is p-conditioned on $\Omega \cap \mathbb{B}_X(0, \delta)$.

A.3. **Auxiliary results.**

Lemma A.8. Let $\{x^1, ..., x^N\} \subset X$ be a finite family. Then there exists $\bar{x} \in \operatorname{co}\{x^1, ..., x^N\}$ such that $\operatorname{supp}(\bar{x}) = \cup\{\operatorname{supp}(x^i) \mid i \in \{1, ..., N\}\}$.

Proof. We proceed by induction. If $N = 1$ this is trivially true. Let us turn on the $N = 2$ case, by considering $\{x^1, x^2\}$ in X. If $\operatorname{supp}(x^1) = \operatorname{supp}(x^2)$, then it is enough to take $\bar{x} = x^1$ or $\bar{x} = x^2$. Assume that $\operatorname{supp}(x^1) \neq \operatorname{supp}(x^2)$. Define

$$\Lambda = \left\{ \frac{|x_k^2|}{|x_k^2 - x_k^1|} \mid k \in \mathcal{N},\ x_k^1 x_k^2 < 0 \right\}.$$

Λ is well defined because $x_k^1 x_k^2 < 0$ implies that $x_k^2 - x_k^1 \neq 0$. Moreover, $\Lambda \subset]0, 1[$, and is at most countable. Let $\lambda \in]0, 1[\setminus \Lambda$, and define $\bar{x} = \lambda x^1 + (1 - \lambda)x^2$. By definition we have $\bar{x} \in \operatorname{co}\{x^1, x^2\}$, so it remains to check that $\operatorname{supp}(\bar{x}) = \operatorname{supp}(x^1) \cup \operatorname{supp}(x^2)$. To prove this, first assume that $k \in \operatorname{supp}(\bar{x})$. If $k \in \operatorname{supp}(x^1)$ it is trivial, so assume that $k \notin \operatorname{supp}(x^1)$. In that case $\bar{x}_k = \lambda \cdot 0 + (1 - \lambda)x_k^2$, where $\lambda \neq 1$ and $\bar{x}_k \neq 0$, from which we deduce that $k \in \operatorname{supp}(x^2)$. This shows that $\operatorname{supp}(\bar{x}) \subset \operatorname{supp}(x^1) \cup \operatorname{supp}(x^2)$ Now, take $k \in \operatorname{supp}(x^1) \cup \operatorname{supp}(x^2)$, and assume by contradiction that $\bar{x}_k = 0$. Then

$$x_k^1 \neq 0, x_k^2 \neq 0, x_k^1 = (1 - \lambda^{-1})x_k^2, \text{ and } \lambda = \frac{|x_k^2|}{|x_k^2 - x_k^1|},$$

which contradicts the fact that $\lambda \notin \Lambda$. Therefore $\operatorname{supp}(x^1) \cup \operatorname{supp}(x^2) \subset \operatorname{supp}(\bar{x})$. Assume now that the statement holds for $N \geqslant 2$, and let us prove it for $N + 1$. Let $\{x^1, ..., x^N, x^{N+1}\} \subset X$ be a finite family. By inductive hypotheses we can find some $\bar{x}^1 \in \operatorname{co}\{x^1, ..., x^N\}$ such that $\operatorname{supp}(\bar{x}^1) = \cup\{\operatorname{supp}(x^i) \mid i \in \{1, ..., N\}\}$. Moreover, the inductive hypotheses guarantees the existence of some $\bar{x} \in \operatorname{co}\{\bar{x}^1, x^{N+1}\}$ such that $\operatorname{supp}(\bar{x}) = \operatorname{supp}(\bar{x}^1) \cup \operatorname{supp}(x^{N+1})$. We derive from the definition of $\bar{x}^1$ that $\operatorname{supp}(\bar{x}) = \cup\{\operatorname{supp}(x^i) \mid i \in \{1, ..., N + 1\}\}$. Also, $\bar{x}^1 \in \operatorname{co}\{x^1, ..., x^N\}$ and $\bar{x} \in \operatorname{co}\{\bar{x}^1, x^{N+1}\}$ imply that $\bar{x} \in \operatorname{co}\{x^1, ..., x^N, x^{N+1}\}$, which ends the proof. ■

Lemma A.9. Let $C \subset X$ be a convex nonempty set, and $J = \cup\{\operatorname{supp}(x) \mid x \in C\}$. If J is finite, then there exists $\bar{x} \in C$ such that $\operatorname{supp}(\bar{x}) = J$.

Proof. Since J is finite, there exists a finite family $\{x^1, ..., x^N\} \subset C$ such that $J = \cup\{\operatorname{supp}(x^i) \mid i \in \{1, ..., N\}\}$. It suffices then to apply the previous lemma to obtain such $\bar{x} \in \operatorname{co}\{x^1, ..., x^N\} \subset C$. ■

References

[1] H. Attouch and J. Bolte, *On the convergence of the proximal algorithm for nonsmooth functions involving analytic features*, Math. Program. Ser. B **116**, pp. 5–16, 2009.

[2] H. Attouch, J. Bolte and B.F. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods*, Math. Program., **137**(1-2), pp. 91–129, 2013.

[3] A. Barbara, A. Jourani and S. Vaiter, Maximal Solutions of Sparse Analysis Regularization, arXiv:1703.00192, 2017.

[4] H.H. Bauschke and P. Combettes, *Convex analysis and monotone operator theory*, 2nd edition, CMS books in Mathematics, Springer, 2017.

[5] A. Blanchet and J. Bolte, *A family of functional inequalities: lojasiewicz inequalities and displacement convex functions*, preprint on arXiv:1612.02619, 2016.

[6] J. Bolte, T.-P. Nguyen, J. Peypouquet and B. Suter, *From error bounds to the complexity of first-order descent methods for convex functions*, To appear in Mathematical Programming, DOI:10.1007/s10107-016-1091-6, 2016.

[7] J.M. Borwein, and A. Lewis, *Partially finite convex programming, part I: Quasi relative interiors and duality theory*, Mathematical Programming, **57**(1), pp. 15–48, 1992.

[8] K. Bredies, and D.A. Lorenz, *Linear convergence of iterative soft-thresholding. Journal of Fourier Analysis and Applications*, **14**(5-6), pp. 813–837, 2008.

[9] P.L. Combettes and J.-C. Pesquet, *Proximal Thresholding Algorithm for Minimization over Orthonormal Bases*, SIAM Journal on Optimization, **18**(4), pp. 1351–1376, 2007.

[10] P.L. Combettes, S. Salzo, and S. Villa, *Consistency of Regularized Learning Schemes in Banach Spaces*, Mathematical Programming, published online 2017-03-25.

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

[12] D. Davis and W. Yin, *Convergence rate analysis of several splitting schemes*, in Splitting Methods in Communication, Imaging, Science, and Engineering, Springer International Publishing, pp. 115–163, 2016.

[13] C. De Mol, E. De Vito, L. Rosasco, *Elastic-net regularization in learning theory*, Journal of complexity, **25**, 201–230, 2009.

[14] K. Degraux, G. Peyré, J. Fadili, and L. Jacques, *Sparse Support Recovery with Non-smooth Loss Functions*, in Advances in Neural Information Processing Systems, pp. 4269-4277, 2016.

[15] C. Dossal, *A necessary and sufficient condition for exact recovery by l1 minimization*, Comptes Rendus Mathématique **350**(1) 117–120, 2011.

[16] D. Drusvyatskiy and A.D. Lewis, *Error bounds, quadratic growth, and linear convergence of proximal methods*, preprint available on arXiv:1602.06661, 2016.

[17] V. Duval and G. Peyré. *Sparse Spikes Super-resolution on Thin Grids I: the LASSO*, Inverse Problems, **33**(5), 055008, 2017.

[18] J. Fadili, J. Malick, and G. Peyré, *Sensitivity Analysis for Mirror-Stratifiable Convex Functions*, preprint on arXiv:1707.03194, 2017.

[19] P. Frankel, G. Garrigos and J. Peypouquet, *Splitting methods with variable metric for Kurdyka-Łojasiewicz functions and general convergence rates*, Journal of Optimization Theory and Applications, **165**(3), pp. 874–900, 2015.

[20] G. Garrigos, L. Rosasco, and S. Villa, *Convergence of the Forward-Backward Algorithm: Beyond the Worst Case with the Help of Geometry*, preprint on arXiv:1703.09477, 2017.

[21] W.L. Hare and A.S. Lewis, *Identifying Active Constraints via Partial Smoothness and Prox-Regularity*, Journal of Convex Analysis, **11**(2), pp. 251–266, 2004.

[22] E.T. Hale, W. Yin and Y. Zhang, *Fixed-Point Continuation for ℓ_1-Minimization: Methodology and Convergence*, SIAM Journal on Optimization 19(3), pp. 1107-1130, 2008.

[23] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex analysis and minimization algorithms I: Fundamentals*, Springer Science & Business Media, 1993.

[24] B. Lemaire, *Stability of the iteration method for non expansive mappings*, Serdica Mathematical Journal, **22**(3), pp. 331–340, 1996.

[25] G. Li, *Global error bounds for piecewise convex polynomials*, Mathematical Programming, **137**(1-2), Ser. A, pp. 37–64, 2013.

[26] J. Liang, J. Fadili and G. Peyré, *Local linear convergence of Forward–Backward under partial smoothness*, in Advances in Neural Information Processing Systems, pp. 1970–1978, 2014.

[27] S. Mosci , L. Rosasco , M. Santoro , A. Verri and S. Villa *Solving Structured Sparsity Regularization with Proximal Methods*, In: Balcázar J.L., Bonchi F., Gionis A., Sebag M. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2010. Lecture Notes in Computer Science, vol 6322. Springer, Berlin, Heidelberg

[28] I. Necoara, Y. Nesterov and F. Glineur, *Linear convergence of first order methods for non-strongly convex optimization*, preprint on arXiv:1504.06298, 2015.

[29] J. Peypouquet, *Convex optimization in normed spaces. Theory, methods and examples.*, Springer Science & Buisness media, 2015.

[30] J. Peypouquet and S. Sorin, *Evolution equations for maximal monotone operators: asymptotic analysis in continuous and discrete time*, Journal of Convex Analysis, **17**(3-4), pp. 1113–1163, 2010.

[31] R.T. Rockafellar, *Convex analysis*, Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[32] S. Weis, *A note on touching cones and faces*, Journal of Convex Analysis, **19**(2), pp. 323–353, 2012.

[33] C. Zalinescu, *Convex analysis in general vector spaces*, World Scientific, 2002.

[34] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society Series B, **67**(2):301–320, 2005