Learning with Optimal Interpolation Norms[*]

Patrick L. Combettes

North Carolina State University, Department of Mathematics
Raleigh, NC 27695-8205, USA
plc@math.ncsu.edu

Andrew M. McDonald

University College London, Department of Computer Science, London WC1E 6BT, UK
a.mcdonald@cs.ucl.ac.uk

Charles A. Micchelli

State University of New York, The University at Albany
Department of Mathematics and Statistics, Albany, NY 12222, USA
charles_micchelli@hotmail.com

Massimiliano Pontil

Istituto Italiano di Tecnologia, 16163 Genoa, Italy
massimiliano.pontil@iit.it
and
University College London, Department of Computer Science, London WC1E 6BT, UK

Abstract

We analyze a class of norms defined via an optimal interpolation problem involving the composition of norms and a linear operator. This construction, known as infimal postcomposition in convex analysis, is shown to encompass various norms which have been used as regularizers in machine learning, signal processing, and statistics. In particular, these include the latent group lasso, the overlapping group lasso, and certain norms used for learning tensors. We establish basic properties of this class of norms and we provide dual norms. The extension to more general classes of convex functions is also discussed. A stochastic block-coordinate version of the Douglas-Rachford algorithm is devised to solve minimization problems involving these regularizers. A prominent feature of the algorithm is that it yields iterates that converge to a solution in the case of non smooth losses and random block updates. Finally, we present numerical experiments with problems employing the latent group lasso penalty.

Key words. Block-coordinate proximal algorithm, Douglas-Rachford splitting, infimal postcomposition, latent group lasso, machine learning, optimal interpolation norm.

---

1 Introduction

In various areas of data analysis such as machine learning, statistics, and signal processing, regularization is a standard tool used to promote known structure in the solutions of an optimization problem. Structured sparsity regularizers have been of particular interest, as a means to encourage specific sparsity patterns in regression vectors, spectra of matrices, gradient of images, and signal decompositions. Important examples include the group lasso [44] and related norms [18, 28, 48], spectral regularizers for low rank matrix learning and multitask learning [1, 19, 36], multiple kernel learning [6, 30], regularizers for learning tensors [34, 39]. Structured sparsity regularizers also arise in speech processing [4], high-dimensional inverse problems [9], image decomposition [22], adaptive image interpolation [23], face reconstruction [37], and hyperspectral imaging [47].

We introduce a general formulation which captures the above norms and allows us to construct new ones via an optimal interpolation problem involving simpler norms and a linear operator. Specifically, given two Banach spaces $\mathcal{X}$ and $\mathcal{Y}$, a norm $\|\cdot\|$ on $\mathcal{Y}$ is constructed as

$$(\forall y \in \mathcal{Y}) \quad \|y\| = \inf_{\substack{x \in \mathcal{X} \\ Lx=y}} |||F(x)|||, \tag{1.1}$$

where $F\colon \mathcal{X} \to \mathbb{R}^m$ is a mapping the components of which are norms, $|||\cdot|||$ is a norm on $\mathbb{R}^m$, and $L\colon \mathcal{X} \to \mathcal{Y}$ is a linear operator. As we shall see, this concise formulation encompasses many classes of regularizers, either via the norm (1.1) or the associated dual norm. In particular, in machine learning, functions of the type described in (1.1) have been investigated in [3, 18, 29, 24, 39]. This formulation also arises in image recovery [8], in inverse problems [11], and in the theory of interpolation spaces [32, 41].

We provide basic properties (cf. Proposition 2.1 and Theorem 2.5) and examples of this construction. We also consider the more general formulation

$$\varphi(y) = \inf_{\substack{x \in \mathcal{X} \\ Lx=y}} h\big(F(x)\big), \tag{1.2}$$

where h and the components of F are convex functions that satisfy certain properties. Such constructs are found for instance in signal recovery formulations [8]. In convex analysis, they are known as infimal postcomposition [7] and their importance was first underlined in [33].

On the numerical side, we shall take advantage of the fact that the mapping F and the operator L can be composed of a large number of "simple" components to create complex structures. Specifically, we present a general method to solve optimization problems involving regularizers of the form (1.1) where $|||\cdot||| = |\cdot|_1$. This method is based on the stochastic block-coordinate Douglas-Rachford iterative framework of [13, 14]. Unlike existing methods, this approach guarantees convergence of the iterates to a solution even when none of the functions present in the model is differentiable and when random coordinate updates are performed.

The paper is organized as follows. In Section 2 we introduce the general class of regularizers and establish some of their basic properties. In Section 3 we construct a number of examples of norms within the proposed framework. In Section 4 we present a random block-coordinate algorithm to solve learning problems involving these regularizers. Finally, in Section 5, we report on numerical experiments with this algorithm using the latent group lasso penalty.

2

Notation. We introduce our notation and recall basic concepts from convex analysis used throughout the paper; for details see [7, 45]. Let $\mathcal{X}$ be a real Banach space, let $\|\cdot\|$ be its norm, let $\mathcal{X}^*$ be its topological dual, and let $\langle\cdot,\cdot\rangle$ be the canonical bilinear form on $\mathcal{X}\times\mathcal{X}^*$. If $\mathcal{X}\neq\{0\}$ and $\mathcal{X}$ is reflexive, it follows from James' theorem that the norm of $\mathcal{X}^*$ is defined by

$$(\forall x^*\in\mathcal{X}^*)\quad \|x^*\|_* = \max_{\substack{x\in\mathcal{X}\\\|x\|=1}} \langle x,x^*\rangle. \tag{1.3}$$

The space of bounded linear operators from $\mathcal{X}$ to a Banach space $\mathcal{Y}$ is denoted by $\mathcal{B}(\mathcal{X},\mathcal{Y})$. Let $\varphi\colon\mathcal{X}\to\,]-\infty,+\infty]$. The domain of φ is $\operatorname{dom}\varphi = \{x\in\mathcal{X}\mid \varphi(x)<+\infty\}$ and the conjugate of φ is $\varphi^*\colon\mathcal{X}^*\to[-\infty,+\infty]\colon x^*\mapsto\sup_{x\in\mathcal{X}}(\langle x,x^*\rangle-\varphi(x))$. $\Gamma_0(\mathcal{X})$ denotes the set of lower semicontinuous convex functions from $\mathcal{X}$ to $\,]-\infty,+\infty]$ with nonempty domain. If $\mathcal{X}$ is a Hilbert space, the proximity operator of $\varphi\in\Gamma_0(\mathcal{X})$ at $x\in\mathcal{X}$ is the unique minimizer, denoted by $\operatorname{prox}_\varphi x$, of $\varphi+\|x-\cdot\|^2/2$. Given $p\in[1,+\infty]$, the ℓ^p norm on $\mathbb{R}^d$ is denoted by $|\cdot|_p$. The associated dual norm is $|\cdot|_q$, where $1/p+1/q=1$. $\mathbb{R}^m_+$ and $\mathbb{R}^m_{++}$ are the positive and strictly positive m-dimensional orthant, respectively, and $\mathbb{R}^m_- = -\mathbb{R}^m_+$.

2 A class of norms

We establish the mathematical foundation of our framework, starting with a scheme to construct convex functions on $\mathcal{Y}$.

Proposition 2.1 *Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ be reflexive real Banach spaces, let K be a nonempty closed convex cone in $\mathcal{Z}$, let $L\colon\mathcal{X}\to\mathcal{Y}$ be linear and bounded, and let $F\colon\mathcal{X}\to\mathcal{Z}$ be K-convex in the sense that*

$$(\forall\alpha\in\,]0,1[)(\forall x\in\mathcal{X})(\forall y\in\mathcal{X})\quad F\big(\alpha x+(1-\alpha)y\big)-\alpha F(x)-(1-\alpha)F(y)\in K. \tag{2.1}$$

Let $h\colon\mathcal{Z}\to\,]-\infty,+\infty]$ be a convex function such that $\operatorname{dom}h\cap\operatorname{ran}F\neq\varnothing$ and

$$(\forall z_1\in\operatorname{ran}F)(\forall z_2\in\operatorname{ran}F)\quad z_1-z_2\in K \quad\Rightarrow\quad h(z_1)\leqslant h(z_2). \tag{2.2}$$

Define

$$\varphi\colon\mathcal{Y}\to[-\infty,+\infty]\colon y\mapsto \inf_{\substack{x\in\mathcal{X}\\Lx=y}} h\big(F(x)\big). \tag{2.3}$$

Then the following hold:

  (i) *φ is convex.*

  (ii) *Suppose that F is continuous, that the cone generated by $\operatorname{ran}L^*-\operatorname{dom}(h\circ F)^*$ is a closed vector subspace of $\mathcal{X}^*$, and that h is lower semicontinuous. Then $\varphi\in\Gamma_0(\mathcal{Y})$.*

Proof. Set $f=h\circ F$.

(i): It is enough to show that f is convex, as this will imply that φ is likewise [7, Proposition 12.36(ii)]. Let $\alpha\in\,]0,1[$, and let x and y be points in $\mathcal{X}$. Combining (2.1) and (2.2) yields

$$h\big(F(\alpha x+(1-\alpha)y)\big)\leqslant h\big(\alpha F(x)+(1-\alpha)F(y)\big). \tag{2.4}$$

3

Therefore, by convexity of h, we obtain

$$(h \circ F)\big(\alpha x + (1 - \alpha)y\big) \leqslant \alpha(h \circ F)(x) + (1 - \alpha)(h \circ F)(y), \tag{2.5}$$

which establishes the convexity of f.

(ii): We first derive from [7, Lemma 1.28] that f is lower semicontinuous. Thus, $f \in \Gamma_0(\mathcal{X})$ and, since the cone generated by $\operatorname{ran} L^* - \operatorname{dom} f^*$ is a closed vector subspace of $\mathcal{X}^*$, it follows from [45, Theorem 2.8.3(vii)] and the same arguments used in the Hilbertian case in [7, Corollary 25.44(i)] that $\varphi \in \Gamma_0(\mathcal{Y})$. $\square$

We are now ready to define a class of norms induced by optimal interpolation, which will be the main focus of the present paper.

Assumption 2.2 $\mathcal{Y}$ *is a real Banach space and* m *is a strictly positive integer. For every* $j \in \{1, \ldots, m\}$, $\mathcal{X}_j$ *is a reflexive real Banach space with norm* $\|\cdot\|_j$. *A generic element in* $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$ *is denoted by* $x = (x_1, \ldots, x_m)$. *Furthermore:*

(i) $F \colon \mathcal{X} \to \mathbb{R}^m \colon x \mapsto (\|x_1\|_1, \ldots, \|x_m\|_m)$.

(ii) $\|\|\cdot\|\|$ *is a norm on* $\mathbb{R}^m$ *which is monotone in the sense that*

$$\big(\forall a \in \mathbb{R}^m_+\big)\big(\forall b \in \mathbb{R}^m_+\big) \quad a - b \in \mathbb{R}^m_- \quad \Rightarrow \quad \|\|a\|\| \leqslant \|\|b\|\|. \tag{2.6}$$

(iii) *For every* $j \in \{1, \ldots, m\}$, $L_j \in \mathcal{B}(\mathcal{X}_j, \mathcal{Y})$, $L \colon \mathcal{X} \to \mathcal{Y} \colon x \mapsto L_1 x_1 + \cdots + L_m x_m$, *and* $\operatorname{ran} L = \mathcal{Y}$.

Set

$$(\forall y \in \mathcal{Y}) \quad \|y\| = \inf_{\substack{x \in \mathcal{X} \\ Lx = y}} \|\|F(x)\|\| = \inf_{\substack{x_1 \in \mathcal{X}_1, \ldots, x_m \in \mathcal{X}_m \\ L_1 x_1 + \cdots + L_m x_m = y}} \|\|\big(\|x_1\|_1, \ldots, \|x_m\|_m\big)\|\|. \tag{2.7}$$

Proposition 2.3 *Consider the setting of Assumption 2.2 and set* $\mathsf{n} = \|\|\cdot\|\| \circ F$. *Then the following hold:*

(i) n *is a norm on* $\mathcal{X}$.

(ii) *The dual norm of* n *at* $x^* \in \mathcal{X}^*$ *is* $\mathsf{n}_*(x^*) = \|\|\big(\|x_1^*\|_{1*}, \ldots, \|x_m^*\|_{m*}\big)\|\|_*$.

Proof. Let $x \in \mathcal{X}$ and $x^* \in \mathcal{X}^*$.

(i): We first deduce from Assumption 2.2(i) that

$$
\begin{aligned}
(\forall \alpha \in \mathbb{R}) \quad \mathsf{n}(\alpha x) &= \|\|\big(\|\alpha x_1\|_1, \ldots, \|\alpha x_m\|_m\big)\|\| \\
&= \|\|\big(|\alpha|\, \|x_1\|_1, \ldots, |\alpha|\, \|x_m\|_m\big)\|\| \\
&= |\alpha|\, \|\|\big(\|x_1\|_1, \ldots, \|x_m\|_m\big)\|\| \\
&= |\alpha|\, \|\|F(x)\|\| \\
&= |\alpha|\, \mathsf{n}(x)
\end{aligned}
\tag{2.8}
$$

4

and that

$$
\begin{aligned}
\mathsf{n}(x) = 0 \;\; &\Leftrightarrow\;\; F(x) = 0 \\
&\Leftrightarrow\;\; (\forall j \in \{1,\dots,m\}) \;\; \|x_j\|_j = 0 \\
&\Leftrightarrow\;\; (\forall j \in \{1,\dots,m\}) \;\; x_j = 0 \\
&\Leftrightarrow\;\; x = 0.
\end{aligned}
\tag{2.9}
$$

To check the triangle inequality, let $z \in \mathcal{X}$. By Assumption 2.2(i), $F(x+z) - F(x) - F(z) \in \mathbb{R}^m_-$. Hence, we derive from (2.6) that

$$
\mathsf{n}(x+z) = |||F(x+z)||| \leqslant |||F(x) + F(z)||| \leqslant |||F(x)||| + |||F(z)||| = \mathsf{n}(x) + \mathsf{n}(z).
\tag{2.10}
$$

(ii): Suppose that $\mathsf{n}(x) = 1$, set $b = (\|x_j^*\|_{j*})_{1 \leqslant j \leqslant m}$, and observe that

$$
\langle x, x^* \rangle = \sum_{j=1}^m \langle x_j, x_j^* \rangle \leqslant \sum_{j=1}^m \|x_j\|_j \|x_j^*\|_{j*} = F(x)^\top b \leqslant |||b|||_*.
\tag{2.11}
$$

Taking the supremum over all such vectors x, we obtain $\mathsf{n}_*(x^*) \leqslant |||b|||_*$. However, by (1.3), since $b \in [0, +\infty[^m$, there exists $a = (\alpha_j)_{1 \leqslant j \leqslant m} \in [0, +\infty[^m$ such that $|||a||| = 1$ and $a^\top b = |||b|||_*$. Likewise, for every $j \in \{1, \dots, m\}$, there exists $z_j \in \mathcal{X}_j$ such that $\|z_j\|_j = 1$ and $\|x_j^*\|_{j*} = \langle z_j, x_j^* \rangle$. Now set $\overline{x} = (\overline{x}_j)_{1 \leqslant j \leqslant m}$, where $(\forall j \in \{1, \dots, m\}) \; \overline{x}_j = \alpha_j z_j$. Then

$$
\mathsf{n}(\overline{x}) = \left|\left|\left|\big(\|\alpha_1 z_1\|_1, \dots, \|\alpha_m z_m\|_m\big)\right|\right|\right| = |||(\alpha_1, \dots, \alpha_m)||| = |||a||| = 1
\tag{2.12}
$$

and therefore

$$
\mathsf{n}_*(x^*) = \sup_{\substack{w \in \mathcal{X} \\ \mathsf{n}(w)=1}} \langle w, x^* \rangle \geqslant \langle \overline{x}, x^* \rangle = \sum_{j=1}^m \langle \overline{x}_j, x_j^* \rangle = \sum_{j=1}^m \alpha_j \langle z_j, x_j^* \rangle = \sum_{j=1}^m \alpha_j \|x_j^*\|_{j*} = a^\top b = |||b|||_*.
\tag{2.13}
$$

We conclude that $\mathsf{n}_*(x^*) = |||b|||_*$. $\square$

Remark 2.4 Let $y \in \mathcal{Y}$ and set $C = \{x \in \mathcal{X} \mid Lx = y\}$. Since $\operatorname{ran} L = \mathcal{Y}$, we have $C \neq \varnothing$. Now let d_C be the distance function to the affine subspace C associated with the norm $\mathsf{n} = ||| \cdot ||| \circ F$ (see Proposition 2.3), that is,

$$
(\forall z \in \mathcal{X}) \quad d_C(z) = \inf_{x \in C} \mathsf{n}(z - x).
\tag{2.14}
$$

It follows from (2.7) that

$$
d_C(0) = \inf_{x \in C} \mathsf{n}(x - 0) = \inf_{x \in C} \mathsf{n}(x) = \|y\|.
\tag{2.15}
$$

Thus, the function $\| \cdot \|$ in (2.7) is defined via a minimal norm interpolation process, that is, the optimization problem underlying (2.7) is that of minimizing the norm n over the affine subspace C. Optimal interpolation and, in particular, the problem of finding a minimal norm interpolant to a finite set of points has a long history in approximation theory; see, e.g., [10] and the references therein.

5

In the next result we show that the construction described in Assumption 2.2 does provide a norm, and we compute its dual norm.

Theorem 2.5 *Consider the setting of Assumption 2.2. Then the following hold:*

(i) $\|\cdot\|$ *is a norm on* $\mathcal{Y}$.

(ii) *Suppose that* $\mathcal{Y}$ *is finite-dimensional. Then the dual norm of* $\|\cdot\|$ *at* $y^* \in \mathcal{Y}^*$ *is*

$$\|y^*\|_* = \left|\|\left(\|L_1^* y^*\|_{1*}, \ldots, \|L_m^* y^*\|_{m*}\right)\right|\|_*. \tag{2.16}$$

Proof. Set $\mathsf{n} = \|\|\cdot\|\| \circ F$ and recall from Proposition 2.3 that n is a norm.

(i): We first note that, since $\operatorname{ran} L = \mathcal{Y}$, $\operatorname{dom} \|\cdot\| = \mathcal{Y}$. Next, we derive from (2.7) that, for every $y \in \mathcal{Y}$ and every $\alpha \in \mathbb{R} \smallsetminus \{0\}$,

$$\|\alpha y\| = \inf_{\substack{x \in \mathcal{X} \\ Lx = \alpha y}} \mathsf{n}(x) = |\alpha| \inf_{\substack{x \in \mathcal{X} \\ L(x/\alpha) = y}} \mathsf{n}(x/\alpha) = |\alpha|\,\|y\|. \tag{2.17}$$

On the other hand, it is clear that F satisfies (2.1) with $K = \mathbb{R}^m$, that $\|\|\cdot\|\|$ satisfies (2.2), and that (2.7) assumes the same form as (2.3). Hence, by Proposition 2.1(i), the function $\|\cdot\|$ is convex. In view of (2.17), we therefore have, for every $(y, z) \in \mathcal{Y} \times \mathcal{Y}$, $\|y + z\| \leqslant \|y\| + \|z\|$. Now let $y \in \mathcal{Y}$ be such that $\|y\| = 0$ and set $C = \{x \in \mathcal{X} \mid Lx = y\}$. Then it follows from (2.15) that $d_C(0) = 0$ and, since C is closed, we get $0 \in C$. Therefore, $y = L0 = 0$. Altogether, $\|\cdot\|$ is a norm.

(ii): Let $y^* \in \mathcal{Y}^*$. Then

$$\begin{aligned}
\|y^*\|_* &= \max\left\{\langle y, y^* \rangle \mid y \in \mathcal{Y},\ \|y\| = 1\right\} \\
&= \max\left\{\langle y, y^* \rangle \,\Big|\, y \in \mathcal{Y},\ \min_{x \in \mathcal{X},\, Lx = y} \mathsf{n}(x) = 1\right\} \\
&= \max\left\{\langle Lx, y^* \rangle \mid x \in \mathcal{X},\ \mathsf{n}(x) = 1\right\} \\
&= \max\left\{\langle x, L^* y^* \rangle \mid x \in \mathcal{X},\ \mathsf{n}(x) = 1\right\} \\
&= \mathsf{n}_*(L^* y^*). \tag{2.18}
\end{aligned}$$

We conclude by applying Proposition 2.3. □

We illustrate the construction (2.7) via two examples.

Example 2.6 In Theorem 2.5 suppose that $m = 2$, that $\mathcal{X}_1$ and $\mathcal{X}_2$ are continuously embedded in the same topological vector space $\mathcal{Y}$, that L_1 and L_2 are the canonical injections, and that $\|\|\cdot\|\| = |\cdot|_1$. Then (2.7) becomes

$$(\forall y \in \mathcal{Y}) \quad \|y\| = \min_{\substack{x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2 \\ x_1 + x_2 = y}} \left(\|x_1\|_1 + \|x_2\|_2\right). \tag{2.19}$$

In other words, $\|\cdot\|$ represents the infimal convolution of the norms $\|\cdot\|_1$ and $\|\cdot\|_2$. This type of construct is central is the theory of interpolation spaces [32, 41]. If we replace the ℓ^1 norm by the ℓ^p norm for some $p \in \,]1, +\infty[$ above, we obtain,

$$(\forall y \in \mathcal{Y}) \quad \|y\| = \min_{\substack{x_1 \in \mathcal{Y}_1,\, x_2 \in \mathcal{Y}_2 \\ x_1 + x_2 = y}} \left(\|x_1\|_1^p + \|x_2\|_2^p\right)^{1/p}. \tag{2.20}$$

This formulation also arises in the area of interpolation spaces [16, 32].

Example 2.7 Let $\mathcal{H}$ be a real Hilbert space with norm $\|\cdot\|_{\mathcal{H}}$, which is identified with its dual. In Theorem 2.5 suppose that $\mathcal{X}_1 = \cdots = \mathcal{X}_m = \mathcal{H}$, let p and q be numbers in $]1, +\infty[$ such that $1/p + 1/q = 1$, and let $\||\cdot\|| = |\cdot|_p$. Then (2.7) becomes

$$(\forall y \in \mathcal{Y}) \quad \|y\| = \left(\min_{\substack{x_1 \in \mathcal{H}, \dots, x_m \in \mathcal{H} \\ \sum_{j=1}^m L_j x_j = y}} \sum_{j=1}^m \|x_j\|_{\mathcal{H}}^p \right)^{1/p}. \tag{2.21}$$

Furthermore, if $\mathcal{H}$ is finite dimensional, the dual norm at $y^* \in \mathcal{Y}$ is given by (2.16) as

$$\|y^*\|_* = \left(\sum_{j=1}^m \|L_j^* y^*\|_{\mathcal{H}}^q \right)^{1/q}. \tag{2.22}$$

This construction is discussed in [24, Theorem 7].

Remark 2.8 Any norm $\|\cdot\|$ on $\mathcal{Y}$ can trivially be written in the form of (2.7) by letting $m = 1$, $\mathcal{X}_1 = \mathcal{Y}$, $\|\cdot\|_1 = \|\cdot\|$, $L_1 = \text{Id}$, and $\||\cdot\|| = |\cdot|$. However, we are interested in exploiting the structure of the construction (2.7) in cases in which the norms $\||\cdot\||$ and $(\|\cdot\|_j)_{1 \leqslant j \leqslant m}$ are chosen from a "simple" class and give rise, via the optimal interpolation problem (2.7), to a "complex" norm $\|\cdot\|$. In particular, when using proximal splitting methods, the computation of $\text{prox}_{\|\cdot\|}$ will typically not be easy whereas that of the operators $(\text{prox}_{\|\cdot\|_j})_{1 \leqslant j \leqslant m}$ will be. This will be exploited in Section 4 to devise an efficient block-coordinate splitting algorithm in the case when $\||\cdot\|| = |\cdot|_1$.

3 Examples

Our objective is to illustrate the fact that the construct presented in Assumption 2.2 contains in a single framework a number of existing regularizers. For simplicity, we focus on the norms captured by Example 2.7. Our main aim here is not to derive new regularizers but, rather, to show that our analysis captures existing ones and to derive their dual norms.

3.1 Latent group lasso

Notation 3.1 The support of $y = (\eta_i)_{1 \leqslant i \leqslant d} \in \mathbb{R}^d$ is $\text{supp}(y) = \{i \in \{1, \dots, d\} \mid \eta_i \neq 0\}$. For every $\varnothing \neq G \subset \{1, \dots, d\}$ and $y \in \mathbb{R}^d$, we set $r = \text{card}\, G$ and let $y|_G$ denote the vector in $\mathbb{R}^r$ obtained by retaining the components of y indexed by G, i.e., $y|_G = (\eta_i)_{i \in G}$. Finally, e_i is the ith standard unit vector in $\mathbb{R}^d$.

The example we consider is known as the latent group lasso (LGL), or group lasso with overlap, which goes back to [18]. For every $j \in \{1, \dots, m\}$, fix $(p_j, q_j) \in [1, +\infty] \times [1, +\infty]$ such that $1/p_j + 1/q_j = 1$. Let $(G_j)_{1 \leqslant j \leqslant m}$ be a covering of $\{1, \dots, d\}$ and define the vector space

$$Z = \big\{ (z_j)_{1 \leqslant j \leqslant m} \mid (\forall j \in \{1, \dots, m\})\ z_j \in \mathbb{R}^d \text{ and } \text{supp}(z_j) \subset G_j \big\}. \tag{3.1}$$

The latent group lasso penalty is defined, for $y \in \mathbb{R}^d$, as

$$\|y\|_{\mathrm{LGL}} = \min \left\{ \sum_{j=1}^{m} |z_j|_{p_j} \ \Big| \ (z_j)_{1 \leqslant j \leqslant m} \in Z, \ \sum_{j=1}^{m} z_j = y \right\}. \tag{3.2}$$

The optimal interpolation problem (3.2) seeks a decomposition of y in terms of vectors $(z_j)_{1 \leqslant j \leqslant m}$ the support sets of which are restricted to the corresponding group of variables in G_j. If the groups overlap then the decomposition is not necessarily unique, and the variational formulation involves those z_j for which $\sum_{j=1}^{m} |z_j|_{p_j}$ is minimal. On the other hand, if the groups are pairwise disjoint, that is $(G_j)_{1 \leqslant j \leqslant m}$ forms a partition of $\{1, \ldots, d\}$, the latent group lasso norm coincides with the "standard" group lasso norm [44], which is defined as

$$\left(\forall y \in \mathbb{R}^d \right) \quad \|y\|_{\mathrm{GL}} = \sum_{j=1}^{m} \left| y_{|G_j} \right|_{p_j}. \tag{3.3}$$

The norms (3.2) and (3.3) are presented in [18] and [44] respectively in the case that $p_1 = \cdots = p_m = p$ and $q_1 = \cdots = q_m = q$. In general, (3.2) has no closed form expression due to the overlapping of the groups. However, in special cases which exhibit additional structure, it can be computed in a finite number of steps. An important example is provided by the (k, p)-support norm [26], in which the groups consist of all subsets of $\{1, \ldots, d\}$ of cardinality no greater than k, for some $k \in \{1, \ldots, d\}$. The case $p = 2$ has been studied in [2] and [25].

Example 3.2 In the above setting, for every $j \in \{1, \ldots, m\}$, set $r_j = \mathrm{card}\, G_j$. The latent group lasso penalty (3.2) is a norm of the form (2.7) with $\mathcal{Y} = \mathbb{R}^d$, $||| \cdot ||| = | \cdot |_1$, and, for every $j \in \{1, \ldots, m\}$, $\mathcal{X}_j = \mathbb{R}^{r_j}$, $\| \cdot \|_j = | \cdot |_{p_j}$ and $L_j = [e_i \mid i \in G_j]$ is a $d \times r_j$ matrix. The change of variables $z_j = L_j x_j$ then yields (3.2). Furthermore, the dual norm is given by

$$\left(\forall y^* \in \mathbb{R}^d \right) \quad \|y^*\|_{\mathrm{LGL}*} = \max_{1 \leqslant j \leqslant m} \|y^*_{|G_j}\|_{q_j}. \tag{3.4}$$

This follows from (2.16) by noting that $||| \cdot |||_* = | \cdot |_\infty$ and, for all $j \in \{1 \ldots, m\}$, $\| \cdot \|_{j*} = | \cdot |_{q_j}$ and $L_j^* y^* = y^*_{|G_j}$.

3.2 Overlapping group lasso

An alternative generalization of the group lasso norm (3.3) is the overlapping group lasso [20, 48], which we denote by $\| \cdot \|_{\mathrm{OGL}}$. It has the same expression as (3.3), except that we drop the restriction that the groups $(G_j)_{1 \leqslant j \leqslant m}$ form a partition of $\{1, \ldots, d\}$. Our next result establishes that the overlapping group lasso penalty is captured by a dual norm of type (2.16). We continue to use Notation 3.1.

Example 3.3 For every $j \in \{1, \ldots, m\}$, let $(p_j, q_j) \in [1, +\infty] \times [1, +\infty]$ be such that $1/p_j + 1/q_j = 1$. Let $(G_j)_{1 \leqslant j \leqslant m}$ be a covering of $\{1, \ldots, d\}$ and let $||| \cdot ||| = | \cdot |_\infty$. For every $j \in \{1, \ldots, m\}$, set $r_j = \mathrm{card}\, G_j$, $\| \cdot \|_j = | \cdot |_{q_j}$, and $L_j = [e_i \mid i \in G_j]$. Then the norm (2.7) evaluated at $y \in \mathbb{R}^d$ is

$$\|y\| = \inf \left\{ \max_{1 \leqslant j \leqslant m} |x_j|_{q_j} \ \Big| \ \sum_{j=1}^{m} L_j x_j = y \right\}. \tag{3.5}$$

In addition, since $||| \cdot |||_* = | \cdot |_1$ and, for every $j \in \{1, \ldots, m\}$, $\| \cdot \|_{j*} = | \cdot |_{p_j}$ and $|L_j^\top y^*|_{p_j} = |y^*_{|G_j}|_{p_j}$, (2.16) yields $\| \cdot \|_* = \| \cdot \|_{\mathrm{OGL}}$.

Remark 3.4 The case $p = +\infty$ corresponds to the iCAP penalty of [48]. We can also consider other choices for the matrices $(L_j)_{1 \leqslant j \leqslant m}$. For example, an appropriate choice gives various total variation penalties [31]. A further example is obtained by choosing $m = 1$ and L_1 to be the incidence matrix of a graph, a setting which has been considered in the context of semi-supervised learning [17]. In particular, for $p = 1$, this corresponds to the fused lasso penalty [38].

Remark 3.5 We have seen that the norm (2.7) captures the latent group lasso when $||| \cdot ||| = | \cdot |_1$, while the dual norm (2.16) captures the overlapping group lasso when $||| \cdot ||| = | \cdot |_\infty$. A natural extension of either setting is to choose $||| \cdot |||$ to be the k Ky-Fan norm for some $k \in \{1, \ldots, m\}$. In other words we choose, for every $x \in \mathbb{R}^m$, $|||x||| = \sum_{j=1}^k |x|_j^\downarrow$, where $|x|^\downarrow \in \mathbb{R}^m$ is the vector obtained by reordering the components of x so that they are decreasing in absolute value. Note that, if $k = 1$, then $||| \cdot ||| = | \cdot |_\infty$ and, if $k = m$, then $||| \cdot ||| = | \cdot |_1$.

3.3 Polyhedral norms

A norm on $\mathcal{Y} = \mathbb{R}^d$ is polyhedral (or is a block-norm) if its closed unit ball B is a polyhedron, i.e., a finite intersection of closed affine half-spaces. Examples of polyhedral norms used in data processing can be found in [26, 46, 48]. In this case, B is bounded, symmetric with respect to the origin, and it has a finite, even number of extreme points. Let us recall a couple of useful facts.

Fact 3.6 [42, Theorem 1] Let $\| \cdot \|$ be a polyhedral norm on $\mathcal{Y} = \mathbb{R}^d$ and let $(b_j)_{1 \leqslant j \leqslant m}$ be the extreme points of its closed unit ball. Then

$$(\forall y \in \mathcal{Y}) \quad \|y\| = \min_{\substack{(\xi_j)_{1 \leqslant j \leqslant m} \in \mathbb{R}^m \\ \sum_{j=1}^m \xi_j b_j = y}} \sum_{j=1}^m |\xi_j|. \tag{3.6}$$

Fact 3.7 [42, Theorem 2] Let $\| \cdot \|$ be a polyhedral norm on $\mathcal{Y} = \mathbb{R}^d$, let B be its closed unit ball, and let $B^\odot = \{y^* \in \mathbb{R}^d \mid (\forall y \in B) \; y^\top y^* \leqslant 1\}$ be the polar set of B. Then $B^\odot$ is a bounded polyhedron and, if $(b_j^*)_{1 \leqslant j \leqslant r}$ denote its extreme points,

$$(\forall y \in \mathbb{R}^d) \quad \|y\| = \max_{1 \leqslant j \leqslant r} y^\top b_j^*. \tag{3.7}$$

It follows from Fact 3.6 that polyhedral norms are special cases of (2.7). Indeed, (3.6) is derived from (2.7) by choosing $||| \cdot ||| = | \cdot |_1$ and $(\forall j \in \{1, \ldots, m\}) \; \mathcal{X}_j = \mathbb{R}$, $\| \cdot \|_j = | \cdot |$, and $L_j \colon \xi \mapsto \xi b_j$. In addition, since a linear function on a nonempty compact convex set attains its maximum at an extreme point of the set [33, Corollary 32.3.2], the dual norm is given by

$$(\forall y^* \in \mathbb{R}^d) \quad \|y^*\|_* = \max_{y \in B} y^\top y^* = \max_{1 \leqslant j \leqslant m} b_j^\top y^*. \tag{3.8}$$

Note that (3.8) is the dual counterpart of (3.7).

3.4 Θ-norms

Assume that $\mathcal{Y} = \mathbb{R}^d$. Families of norms parameterized by a nonempty, convex, and bounded set $\Theta \subset \mathbb{R}^d_{++}$ were considered in [5, 25, 28]. As shown in [27, Proposition 2], the expressions

$$(\forall y \in \mathbb{R}^d) \quad \|y\|_\Theta = \inf_{\theta \in \Theta} \sqrt{\langle y \mid \operatorname{diag}(\theta)^{-1}y\rangle} \tag{3.9}$$

and

$$(\forall y^* \in \mathbb{R}^d) \quad \|y^*\|_{\Theta*} = \sup_{\theta \in \Theta} \sqrt{\langle y^* \mid \operatorname{diag}(\theta)y^*\rangle} \tag{3.10}$$

define dual norms. Examples of norms which are included in this family are the ℓ^p-norms, and the k-support norm mentioned above [2]. Next, we relate Θ-norms to our framework.

Proposition 3.8 *Let $(\theta_j)_{1\leqslant j\leqslant m}$ be vectors in $\mathbb{R}^d_+$ and let Θ be a subset of $\mathbb{R}^d_{++}$ such that $\overline{\Theta} = \operatorname{conv}\{\theta_1,\ldots,\theta_m\}$. Then the norms defined in (3.9) and (3.10) can be written in the form (2.7) and (2.16) respectively, with $|||\cdot||| = |\cdot|_1$ and $(\forall j \in \{1,\ldots,m\}) \|\cdot\|_j = |\cdot|_2$ and $L_j = \operatorname{diag}(\sqrt{\theta_j})$, where the $\sqrt{\cdot}$ operator is understood componentwise.*

Proof. We have $|||\cdot|||_* = |\cdot|_\infty$ and $(\forall j \in \{1,\ldots,m\}) \|\cdot\|_{j*} = |\cdot|_2$. Now set $(\forall j \in \{1,\ldots,m\})$ $L_j = \operatorname{diag}(\sqrt{\theta_j})$ and $\mathcal{X}_j = \mathbb{R}^d$. Then we derive from (2.16) that

$$(\forall y^* \in \mathbb{R}^d) \quad \|y^*\|_*^2 = \max_{1\leqslant j\leqslant m} |L_j^\top y^*|_2^2$$

$$= \max_{1\leqslant j\leqslant m} \langle y^* \mid \operatorname{diag}(\theta_j)y^*\rangle$$

$$= \max_{\theta \in \overline{\Theta}} \langle y^* \mid \operatorname{diag}(\theta)y^*\rangle \tag{3.11}$$

$$= \sup_{\theta \in \Theta} \langle y^* \mid \operatorname{diag}(\theta)y^*\rangle, \tag{3.12}$$

where equality (3.11) results from the fact that, in $\mathbb{R}^d$, a linear function on a nonempty compact convex set attains its maximum at an extreme point of the set [33, Corollary 32.3.2]. This establishes (3.10). As noted above, (3.10) is the dual of (3.9). It follows that (3.9) is of the form (2.7). □

3.5 Tensor norms

A number of regularizers have been proposed to learn low rank tensors; see [15, 34, 35, 40, 43] and the references therein. In this section, we discuss two prominent examples that fit our framework. We first recall some notions from multilinear algebra [21]. Let $\boldsymbol{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_m}$ be an m-mode real tensor, that is,

$$\boldsymbol{Y} = [Y_{i_1,\ldots,i_m}]_{1\leqslant i_1\leqslant d_1,\ldots,1\leqslant i_m\leqslant d_m}. \tag{3.13}$$

Now let $j \in \{1,\ldots,m\}$. A mode-j fiber is a vector composed of elements of $\boldsymbol{Y}$ obtained by fixing all indices except those corresponding to the jth. Set $r_j = \prod_{k\neq j} d_k$. The mode-$j$ matricization $M_j(\boldsymbol{Y})$ of a tensor $\boldsymbol{Y}$ is the $d_j \times r_j$ matrix obtained by arranging the mode-j fibers of $\boldsymbol{Y}$ such that each

of them forms a column of $M_j(\boldsymbol{Y})$. By way of example, a 3-mode tensor $\boldsymbol{Y} \in \mathbb{R}^{3 \times 4 \times 2}$ admits the matricizations: $M_1(\boldsymbol{Y}) \in \mathbb{R}^{3 \times 8}$, $M_2(\boldsymbol{Y}) \in \mathbb{R}^{4 \times 6}$, and $M_3(\boldsymbol{Y}) \in \mathbb{R}^{2 \times 12}$. Note that $M_j \colon \mathbb{R}^{d_1 \times \cdots \times d_m} \to \mathbb{R}^{d_j \times r_j}$ is a linear operator. Its adjoint $M_j^* \colon \mathbb{R}^{d_j \times r_j} \to \mathbb{R}^{d_1 \times \cdots \times d_m}$ is the reverse matricization along mode j.

Recall that the nuclear norm (or trace norm) of a matrix, $\|\cdot\|_{\mathrm{nuc}}$, is the sum of its singular values. Its dual norm is the spectral norm $\|\cdot\|_{\mathrm{sp}}$ which provides the largest singular value. The overlapped nuclear norm [34, 40] is defined as the sum of the nuclear norms of the mode-j matricizations, namely

$$\left(\forall \boldsymbol{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_m} \right) \quad \|\boldsymbol{Y}\|_{\mathrm{ONN}} = \sum_{j=1}^{m} \|M_j(\boldsymbol{Y})\|_{\mathrm{nuc}}. \tag{3.14}$$

Example 3.9 Let $\mathcal{Y} = \mathbb{R}^{d_1 \times \cdots \times d_m}$ and $||| \cdot ||| = | \cdot |_\infty$. In addition, for every $j \in \{1, \ldots, m\}$, set $\mathcal{X}_j = \mathbb{R}^{d_j \times r_j}$, $\|\cdot\|_j = \|\cdot\|_{\mathrm{sp}}$, and $L_j = M_j^*$. Then (2.7) becomes

$$\left(\forall \boldsymbol{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_m} \right) \quad \|\boldsymbol{Y}\| = \inf \left\{ \max_{1 \leqslant j \leqslant m} \|X_j\|_{\mathrm{sp}} \ \Big| \ \sum_{j=1}^{m} M_j^* X_j = \boldsymbol{Y} \right\}. \tag{3.15}$$

In addition, since $||| \cdot |||_* = | \cdot |_1$ and, for every $j \in \{1, \ldots, m\}$, $\|\cdot\|_{j*} = \|\cdot\|_{\mathrm{nuc}}$ and $L_j^* = M_j$, equation (2.16) yields $\|\cdot\|_* = \|\cdot\|_{\mathrm{ONN}}$.

Now let $\{\alpha_j\}_{1 \leqslant j \leqslant m} \subset \ ]0, +\infty[$. The scaled latent nuclear norm is defined by (see [39] and [43] for special cases)

$$\left(\forall \boldsymbol{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_m} \right) \quad \|\boldsymbol{Y}\|_{\mathrm{LNN}} = \inf \left\{ \sum_{j=1}^{m} \frac{1}{\alpha_j} \|X_j\|_{\mathrm{nuc}} \ \Big| \ \sum_{j=1}^{m} M_j^* X_j = \boldsymbol{Y} \right\}. \tag{3.16}$$

Our next example captures this norm.

Example 3.10 The latent nuclear norm (3.16) is of the form (2.7) with $\mathcal{Y} = \mathbb{R}^{d_1 \times \cdots \times d_m}$, $||| \cdot ||| = | \cdot |_1$ and, for every $j \in \{1, \ldots, m\}$, $\mathcal{X}_j = \mathbb{R}^{d_j \times r_j}$, $\|\cdot\|_j = \|\cdot\|_{\mathrm{nuc}}/\alpha_j$, and $L_j = M_j^*$. Furthermore, (2.16) yields the dual norm

$$\left(\forall \boldsymbol{Y}^* \in \mathbb{R}^{d_1 \times \cdots \times d_m} \right) \quad \|\boldsymbol{Y}^*\|_{\mathrm{LNN}*} = \max_{1 \leqslant j \leqslant m} \alpha_j \|M_j(\boldsymbol{Y}^*)\|_{\mathrm{sp}}. \tag{3.17}$$

4 Random block-coordinate algorithm

4.1 Overview

The purpose of this section is to address some of the numerical aspects associated with the class of norms introduced in Assumption 2.2 in the case when $||| \cdot ||| = | \cdot |_1$, which reduces (2.7) to

$$(\forall y \in \mathcal{Y}) \quad \|y\| = \inf_{\substack{x_1 \in \mathcal{X}_1, \ldots, x_m \in \mathcal{X}_m \\ L_1 x_1 + \cdots + L_m x_m = y}} \|x_1\|_1 + \cdots + \|x_m\|_m. \tag{4.1}$$

Since such norms are nonsmooth convex functions, they could in principle be handled via their proximity operators in the context of proximal splitting algorithms [7, 12]. However, the proximity operator of the composite norm $\|\cdot\|$ in (4.1) is usually intractable, which makes this direct approach unviable. We circumvent this problem by formulating the problem in such a way that it involves only the proximity operators of the norms $(\|\cdot\|_j)_{1\leqslant j\leqslant m}$, which will typically be available in closed form.

The main features of the algorithmic approach we propose are the following:

- It can handle general nonsmooth formulations: the functions present in the model need not be differentiable.

- It adapts the recent approach proposed in [13, 14] to devise a block-coordinate algorithm which allows us to select arbitrarily the blocks of norms $(\|\cdot\|_j)_{1\leqslant j\leqslant m}$ to be activated over the course of the iterations. This makes the method amenable to the processing of very large data sets in a flexible manner by adapting the computational load of each iteration to the available computing resources.

- The computations are broken down to the evaluation of simple proximity operators of the norms $(\|\cdot\|_j)_{1\leqslant j\leqslant m}$ and of those appearing in the loss function, while the linear operators $(L_j)_{1\leqslant j\leqslant m}$ are applied separately.

- Knowledge of the norms of the operators $(L_j)_{1\leqslant j\leqslant m}$ is not required.

- Convergence of the iterates to a solution of the minimization problem under consideration is guaranteed.

4.2 Problem formulation

We consider the standard linear problem in which a vector $\overline{y}$ in a real Hilbert space $\mathcal{Y}$ is to be inferred from n noisy linear observations

$$
\begin{cases}
\beta_1 & = \langle \overline{y}, a_1^* \rangle + \zeta_1 \\
\quad \vdots \\
\beta_n & = \langle \overline{y}, a_n^* \rangle + \zeta_n,
\end{cases}
\tag{4.2}
$$

where $(a_i^*)_{1\leqslant i\leqslant n} \in (\mathcal{Y}^*)^n$ are known and $(\zeta_i)_{1\leqslant i\leqslant n} \in \mathbb{R}^n$ models unknown perturbations. This model captures various problems in supervised learning and in inverse problems. A common variational formulation associated with (4.2) is the regularized convex minimization problem

$$
\underset{y\in\mathcal{Y}}{\text{minimize}} \quad \sum_{i=1}^{n} \ell_i(\langle y, a_i^* \rangle, \beta_i) + \lambda \|y\|,
\tag{4.3}
$$

where $\|\cdot\|$ is a norm fulfilling Assumption 2.2 with $\|\|\cdot\|\| = |\cdot|_1$ (see (4.1)), $\lambda \in \,]0,+\infty[$ is a regularization parameter, and, for every $i \in \{1,\ldots,n\}$ and every $\beta \in \mathbb{R}$, $\ell_i(\cdot,\beta) \in \Gamma_0(\mathbb{R})$. We also assume that each $\mathcal{X}_j$ is finite-dimensional and can be equipped with a norm $\|\cdot\|_j$ that makes it a

Euclidean space. We designate by $\mathcal{H}$ the Euclidean space obtained by renorming $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$ with the norm $x = (x_j)_{1 \leqslant j \leqslant m} \mapsto \sqrt{\sum_{j=1}^m \|x_j\|_j^2}$. In this setting, (4.3) becomes

$$\underset{\substack{y \in \mathcal{Y} \\ x_1 \in \mathcal{X}_1, \ldots, x_m \in \mathcal{X}_m \\ \sum_{j=1}^m L_j x_j = y}}{\text{minimize}} \quad \sum_{i=1}^n \ell_i(\langle y, a_i^* \rangle, \beta_i) + \lambda \sum_{j=1}^m \|x_j\|_j. \tag{4.4}$$

One can therefore first obtain a solution $(x_j)_{1 \leqslant j \leqslant m}$ to the problem

$$\underset{x_1 \in \mathcal{X}_1, \ldots, x_m \in \mathcal{X}_m}{\text{minimize}} \quad \sum_{i=1}^n \ell_i \left(\sum_{j=1}^m \langle L_j x_j, a_i^* \rangle, \beta_i \right) + \lambda \sum_{j=1}^m \|x_j\|_j \tag{4.5}$$

and then

$$y = \sum_{j=1}^m L_j x_j \text{ solves (4.4).} \tag{4.6}$$

To make the structure of (4.5) more apparent, let us introduce the functions

$$\Phi \colon \mathcal{X} \to]-\infty, +\infty] : x \mapsto \lambda \sum_{j=1}^m \|x_j\|_j \tag{4.7}$$

and

$$\Psi \colon \mathbb{R}^m \to]-\infty, +\infty] : (\eta_1, \ldots, \eta_n) \mapsto \sum_{i=1}^n \psi_i(\eta_i), \tag{4.8}$$

where, for every $i \in \{1, \ldots, n\}$,

$$\psi_i \colon \mathbb{R} \to]-\infty, +\infty] : \eta_i \mapsto \ell_i(\eta_i, \beta_i). \tag{4.9}$$

Let us also define

$$A \colon \mathcal{Y} \to \mathbb{R}^n : y \mapsto (\langle y, a_i^* \rangle)_{1 \leqslant i \leqslant n} \quad \text{and} \quad B \colon \mathcal{H} \to \mathbb{R}^n : x \mapsto \sum_{j=1}^m B_j x_j, \tag{4.10}$$

where, for every $j \in \{1, \ldots, m\}$,

$$B_j = A L_j \in \mathcal{B}(\mathcal{X}_j, \mathbb{R}^n). \tag{4.11}$$

Then, recalling from Assumption 2.2 that $L \colon \mathcal{H} \to \mathcal{Y} : x \mapsto \sum_{j=1}^m L_j x_j$, we have $B = AL \in \mathcal{B}(\mathcal{H}, \mathbb{R}^n)$ and we can thus rewrite (4.4) as

$$\underset{x \in \mathcal{H}}{\text{minimize}} \quad \Phi(x) + \Psi(Bx). \tag{4.12}$$

Note that our hypotheses imply that $\Phi \in \Gamma_0(\mathcal{H})$ and $\Psi \in \Gamma_0(\mathbb{R}^n)$.

4.3 Douglas-Rachford splitting in a product space

We work in the direct Hilbert sum $\boldsymbol{\mathcal{H}} = \mathcal{H} \oplus \mathbb{R}^n$. Let us introduce the functions

$$\begin{cases} \boldsymbol{F} \colon \boldsymbol{\mathcal{H}} \to \,]{-}\infty, +\infty] : (x, r) \mapsto \Phi(x) + \Psi(r) \\ \boldsymbol{G} = \iota_{\boldsymbol{V}}, \quad \text{where} \quad \boldsymbol{V} = \operatorname{gra} B = \big\{ (x, r) \in \boldsymbol{\mathcal{H}} \ \big| \ Bx = r \big\}. \end{cases} \tag{4.13}$$

Using the variable $\boldsymbol{x} = (x, r)$, we reduce (4.12) to the problem

$$\underset{\boldsymbol{x} \in \boldsymbol{\mathcal{H}}}{\text{minimize}} \ \ \boldsymbol{F}(\boldsymbol{x}) + \boldsymbol{G}(\boldsymbol{x}) \tag{4.14}$$

involving the sum of two functions in $\Gamma_0(\boldsymbol{\mathcal{H}})$ and which can be solved with the Douglas-Rachford algorithm [7, Section 27.2]. Let $\boldsymbol{y}_0 \in \boldsymbol{\mathcal{H}}$, let $\gamma \in \,]0, +\infty[$, and let $(\mu_k)_{k \in \mathbb{N}}$ be a sequence in $]0, 2[$ such that $\sum_{k \in \mathbb{N}} \mu_k (2 - \mu_k) = +\infty$. The Douglas-Rachford algorithm

$$\begin{array}{l} \text{for } k = 0, 1, \dots \\ \left\lfloor \begin{array}{l} \boldsymbol{x}_k = \operatorname{prox}_{\gamma \boldsymbol{G}} \boldsymbol{y}_k \\ \boldsymbol{z}_k = \operatorname{prox}_{\gamma \boldsymbol{F}} (2\boldsymbol{x}_k - \boldsymbol{y}_k) \\ \boldsymbol{y}_{k+1} = \boldsymbol{y}_k + \mu_k (\boldsymbol{z}_k - \boldsymbol{x}_k) \end{array} \right. \end{array} \tag{4.15}$$

produces a sequence $(\boldsymbol{x}_k)_{k \in \mathbb{N}}$ which converges to a solution to (4.14) [7, Corollary 27.4]. However, by [7, Proposition 24.11 and Example 29.19(i)],

$$\operatorname{prox}_{\boldsymbol{F}} \colon (x, r) \mapsto \big(\operatorname{prox}_{\Phi} x, \operatorname{prox}_{\Psi} r \big), \tag{4.16}$$

and

$$\operatorname{prox}_{\boldsymbol{G}} \colon (u, s) \mapsto (x, Bx), \quad \text{where} \ \ x = u - B^*(\operatorname{Id} + BB^*)^{-1}(Bu - s) \tag{4.17}$$

is the projection operator onto $\boldsymbol{V}$. Hence, upon setting $R = B^*(\operatorname{Id} + BB^*)^{-1}$, we can rewrite (4.15) as

$$\begin{array}{l} \text{for } k = 0, 1, \dots \\ \left\lfloor \begin{array}{l} q_k = Bu_k - s_k \\ x_k = u_k - Rq_k \\ r_k = Bx_k \\ v_k = \operatorname{prox}_{\gamma \Phi}(2x_k - u_k) \\ t_k = \operatorname{prox}_{\gamma \Psi}(2r_k - s_k) \\ u_{k+1} = u_k + \mu_k(v_k - x_k) \\ s_{k+1} = s_k + \mu_k(t_k - r_k), \end{array} \right. \end{array} \tag{4.18}$$

where we have set $\boldsymbol{x}_k = (x_k, r_k)$, $\boldsymbol{y}_k = (u_k, s_k)$, and $\boldsymbol{z}_k = (v_k, t_k)$. It follows from the above result that $(x_k)_{k \in \mathbb{N}}$ converges to a solution to (4.12). Let us now express (4.18) in terms of the original variables of problem (4.5). To this end set, for every $j \in \{1, \dots, m\}$,

$$R_j = B_j^*(\operatorname{Id} + BB^*)^{-1} = B_j^* \left(\operatorname{Id} + \sum_{j=1}^m B_j B_j^* \right)^{-1}. \tag{4.19}$$

Moreover, let us denote by $x_{j,k} \in \mathcal{X}_j$ the jth component of x_k, and by $u_{j,k} \in \mathcal{X}_j$ the jth component of u_k. Furthermore, we denote by $\rho_{i,k} \in \mathbb{R}$ the ith component of r_k, and by $\sigma_{i,k} \in \mathbb{R}$ the ith component of s_k. Then (4.18) becomes

$$
\begin{array}{l}
\text{for } k = 0, 1, \ldots \\
\left\lfloor
\begin{array}{l}
q_k = \sum_{j=1}^m B_j u_{j,k} - s_k \\
\text{for } j = 1, \ldots, m \\
\quad \left\lfloor
\begin{array}{l}
x_{j,k} = u_{j,k} - R_j q_k \\
u_{j,k+1} = u_{j,k} + \mu_k \big(\operatorname{prox}_{\gamma\lambda \|\cdot\|_j} (2x_{j,k} - u_{j,k}) - x_{j,k} \big)
\end{array}
\right. \\
r_k = \sum_{j=1}^m B_j x_{j,k} \\
\text{for } i = 1, \ldots, n \\
\quad \left\lfloor \sigma_{i,k+1} = \sigma_{i,k} + \mu_k \big(\operatorname{prox}_{\gamma\psi_i} (2\rho_{i,k} - \sigma_{i,k}) - \rho_{i,k} \big).
\right.
\end{array}
\right.
\end{array}
\tag{4.20}
$$

In large-scale problems, a possible drawback of this approach is that $m+n$ proximity operators must be evaluated at each iteration, which can lead to impractical implementations in terms of computations and/or memory requirements. The analysis of [13, Corollary 5.5] shows that the proximity operators in (4.20) can be sampled by sweeping through the indices in $\{1, \ldots, m\}$ and $\{1, \ldots, n\}$ randomly while preserving the convergence of the iterates. This results in partial updates of the variables which lead to significantly lighter iterations and remarkable flexibility in the implementation of the algorithm. Thus, a variable $u_{j,k}$ is updated at iteration k depending on whether a random activation variable $\varepsilon_{j,k}$ takes on the value 1 or 0 (each component $\sigma_{i,k}$ of the vector s_k is randomly updated according to the same strategy). The method resulting from this random sampling scheme is presented in the next theorem.

Theorem 4.1 *Let* $\mathsf{D} = \{0,1\}^{m+n} \smallsetminus \{0\}$, *let* $\gamma \in \left]0, +\infty\right[$, *let* $(\mu_k)_{k \in \mathbb{N}}$ *be a sequence in* $\left]0, 2\right[$ *such that* $\inf_{k \in \mathbb{N}} \mu_k > 0$ *and* $\sup_{k \in \mathbb{N}} \mu_k < 2$, *let* $(u_{j,0})_{1 \leqslant j \leqslant m} \in \mathcal{H}$, *let* $s_0 = (\sigma_{i,0})_{1 \leqslant i \leqslant n} \in \mathbb{R}^n$, *and let* $(\varepsilon_k)_{k \in \mathbb{N}} = (\varepsilon_{1,k} \ldots, \varepsilon_{m+n,k})_{k \in \mathbb{N}}$ *be identically distributed* D-*valued random variables such that, for*

every $i \in \{1, \ldots, m+n\}$, $\mathrm{Prob}[\varepsilon_{i,0} = 1] > 0$. Iterate

$$
\begin{array}{l}
\textit{for } j = 1, \ldots, m \\
\quad \lfloor \; b_{j,0} = B_j u_{j,0} \\
\textit{for } k = 0, 1, \ldots \\
\quad \Big\lfloor \; q_k = \sum_{j=1}^{m} b_{j,k} - s_k \\
\qquad \textit{for } j = 1, \ldots, m \\
\qquad \Big\lfloor \; x_{j,k} = u_{j,k} - R_j q_k \\
\qquad\quad \textit{if } \varepsilon_{j,k} = 1 \\
\qquad\quad \Big\lfloor \; u_{j,k+1} = u_{j,k} + \mu_k \big(\mathrm{prox}_{\gamma\lambda\|\cdot\|_j}(2x_{j,k} - u_{j,k}) - x_{j,k} \big) \\
\qquad\qquad\quad b_{j,k+1} = B_j u_{j,k+1} \\
\qquad\quad \textit{if } \varepsilon_{j,k} = 0 \\
\qquad\quad \Big\lfloor \; u_{j,k+1} = u_{j,k} \\
\qquad\qquad\quad b_{j,k+1} = b_{j,k} \\
\qquad r_k = \sum_{j=1}^{m} B_j x_{j,k} \\
\qquad \textit{for } i = 1, \ldots, n \\
\qquad \Big\lfloor \; \textit{if } \varepsilon_{m+i,k} = 1 \\
\qquad\quad \lfloor \; \sigma_{i,k+1} = \sigma_{i,k} + \mu_k \big(\mathrm{prox}_{\gamma\psi_i}(2\rho_{i,k} - \sigma_{i,k}) - \rho_{i,k} \big) \\
\qquad\quad \textit{if } \varepsilon_{m+i,k} = 0 \\
\qquad\quad \lfloor \; \sigma_{i,k+1} = \sigma_{i,k}.
\end{array}
\tag{4.21}
$$

Suppose that the random sequences $(\varepsilon_k)_{k\in\mathbb{N}}$ and $(u_k, s_k)_{k\in\mathbb{N}}$ are independent. Then, for every $j \in \{1, \ldots, m\}$, $(x_{j,k})_{k\in\mathbb{N}}$ converges almost surely to a vector x_j and $y = \sum_{j=1}^{m} L_j x_j$ is a solution to (4.3).

Proof. It follows from [13, Corollary 5.5] that $(x_{1,k}, \ldots, x_{m,k})_{k\in\mathbb{N}}$ converges almost surely to a solution to (4.5). In turn, as seen in (4.6), y solves (4.4) and hence (4.3). □

Remark 4.2 The operators $(R_j)_{1\leqslant j\leqslant m}$ of (4.19) are computed off-line only once and they intervene in algorithm (4.21) only via matrix-vector multiplications.

Remark 4.3 It follows from the result of [13] that, under suitable qualification conditions, the conclusions of Theorem 4.1 remain true for a general choice of the functions $f_j \in \Gamma_0(\mathcal{X}_j)$ instead of $\|\cdot\|_j$, and when L does not have full rank. This allows us to solve the more general versions of (4.4) in which the regularizer is not a norm but a function of the form (2.3).

5 Numerical experiments

In this section we present numerical experiments applying the random sweeping stochastic block algorithm outlined in Section 4 to sparse problems in which the regularization penalty is a norm fitting our framework, as described in Assumption 2.2. The goal of these experiments is to show concrete applications of the class of norms discussed in the paper and to illustrate the behavior of the proposed random block-iterative proximal splitting algorithm. Let us stress that these appear to be the first numerical experiments on this kind of block-coordinate method for completely nonsmooth optimization problems with converging sequences of iterates.

The setting we consider is binary classification with the hinge loss and a latent group lasso penalty [18]. Each data matrix $A \in \mathbb{R}^{n \times d}$ is generated with i.i.d. Gaussian entries and each row a_i of A is normalized to have unit ℓ^2 norm. Similarly, the true model vector $\overline{y} \in \mathbb{R}^d$ is sparse and its nonzero entries are generated randomly on the unit ℓ^2 sphere. The n observations are then obtained as $\beta_i = \operatorname{sign}(\langle y \mid a_i \rangle)$. To induce classification errors, a randomly chosen subset of the observations have their sign reversed, with the value of the noise determining the size of the subset, expressed as a percentage of the total observations. For the implementation of the algorithm, we require the proximity operators of the functions $\|\cdot\|_j$ and ψ_i in Theorem 4.1. In this case, these are the ℓ^2-norm and the hinge loss, which have straightforward proximity operators [7]. More precisely, (4.4) assumes the form

$$\underset{\substack{y \in \mathcal{Y} \\ x_1 \in \mathcal{X}_1, \ldots, x_m \in \mathcal{X}_m \\ \sum_{j=1}^{m} L_j x_j = y}}{\operatorname{minimize}} \quad \sum_{i=1}^{n} \max\left\{0, 1 - \beta_i \langle y \mid a_i \rangle\right\} + \lambda \sum_{j=1}^{m} \|x_j\|_j. \tag{5.1}$$

In large-scale applications it is not possible to activate all the functions and all the blocks at each iteration due to computing and memory limitations. The random sweeping algorithm (4.21) allows us to activate only some of the blocks by toggling the activation variables $(\varepsilon_{j,k})_{1 \leqslant j \leqslant m}$ and $(\varepsilon_{m+i,k})_{1 \leqslant i \leqslant n}$. The vectors are updated only when the corresponding activation variable is equal to 1; otherwise it is equal to 0 and no update takes place. In our experiments we always activate the variables $(\varepsilon_{m+i,k})_{1 \leqslant i \leqslant n}$ as they are associated to the set of training points which is small in sparsity regularization problems. On the other hand, only a fraction α of the variables $(\varepsilon_{j,k})_{1 \leqslant j \leqslant m}$ are activated, which is achieved by sampling, at each iteration k, a subset of $\lceil m\alpha \rceil$ distinct indices in $\{1, \ldots, m\}$. In light of Theorem 4.1, convergence of the iterates is guaranteed for every $\alpha \in \,]0,1]$. It is natural to ask to what extent these partial updates slow down the algorithm with respect to the hypothetical fully updated version in which sufficient computational power and memory are available.

To investigate this question, in our first experiment $A \in \mathbb{R}^{1000 \times 10000}$, the true model vector $\overline{y}$ has sparsity 95%, and we apply a 25% classification error rate. The relaxation parameters $(\mu_k)_{k \in \mathbb{N}}$ are all set to 1.99, the proximal parameter γ is set to 0.1, and the regularization parameter λ is set to 0.1. As a stopping rule for the algorithm we use

$$\frac{|x_{k+1} - x_k|_2}{|x_k|_2} \leqslant 10^{-6}. \tag{5.2}$$

We employ the chain latent group lasso penalty, whereby the groups define contiguous sequences of length 10, with an overlap of length 3, and the number of groups is 1429. Table 1 presents the time, number of iterations, and number of iterations normalized by the activation rate for the hinge loss and latent group lasso penalty for values of the activation parameter α in $\{0.1, 0.2, \ldots, 1.0\}$. These quantities have been averaged over 20 independent trials. The normalized iteration numbers are obtained by multiplying the actual iteration number by the activation rate α in order to fairly quantify the global computational effort. Indeed, scaling the iterations by the activation rate allows for a fair comparison between regimes since the computational load of the algorithm per iteration is proportional to the number of activated blocks. We observe that, while the absolute number of iterations naturally increases as the activation rate decreases, the normalized number of iterations is remarkably stable across the different regimes. Thus, in large scale problems in which memory

Table 1: Time, iterations, and normalized iterations averaged over 20 runs for hinge loss classification with the latent group lasso. $A \in \mathbb{R}^{1000 \times 10000}$ and $m = 1429$.

activation rate	time (s)	actual iterations	normalized iterations
1.0	24912	14515	14515
0.9	24517	16047	14443
0.8	26124	18080	14464
0.7	28975	20633	14443
0.6	28223	23872	14323
0.5	29304	28308	14154
0.4	36983	34392	13757
0.3	38100	44080	13224
0.2	50484	62664	12533
0.1	62213	100829	10083

Table 2: Time, iterations, and normalized iterations averaged over 20 runs for hinge loss classification with the k-support norm. $A \in \mathbb{R}^{20 \times 25}$, $k = 4$, and $m = 12650$.

activation rate	time (s)	actual iterations	normalized iterations
1.0	388	463	463
0.8	446	681	544
0.6	454	894	536
0.4	482	1281	512
0.2	423	2557	511
0.1	469	4851	485
0.05	1054	9402	470
0.010	1625	41633	416
0.005	1907	77066	385

space and processing power are limited, the standard optimization algorithm (4.20) with full activation rate is not suitable, whereas our random sweeping procedure can be easily implemented. Interestingly, Table 1 indicates that the normalized number of iterations is not affected. Figure 1 depicts (top) the objective values for the problem and (bottom) the distance to the limiting solution for various activation rates. We note that the paths are similar for all activation rates, and the convergence is similarly fast. This reinforces our findings that partial activation of the blocks does not lead to any deterioration in normalized performance.

As a second experiment, we revisit the above problem using the k-support norm penalty of [2], which is a special case of Example 3.2. Here, $A \in \mathbb{R}^{20 \times 25}$, $k = 4$, and $\alpha \in \{0.005, 0.01, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0\}$. The number of groups is $m = d!/(k!(d-k)!) = 12650$. Even in this relatively small size problem, the number of groups m considerably exceeds both d and n. Table 2 shows the same metrics as the first experiment. We again observe that performance is stable as α varies.
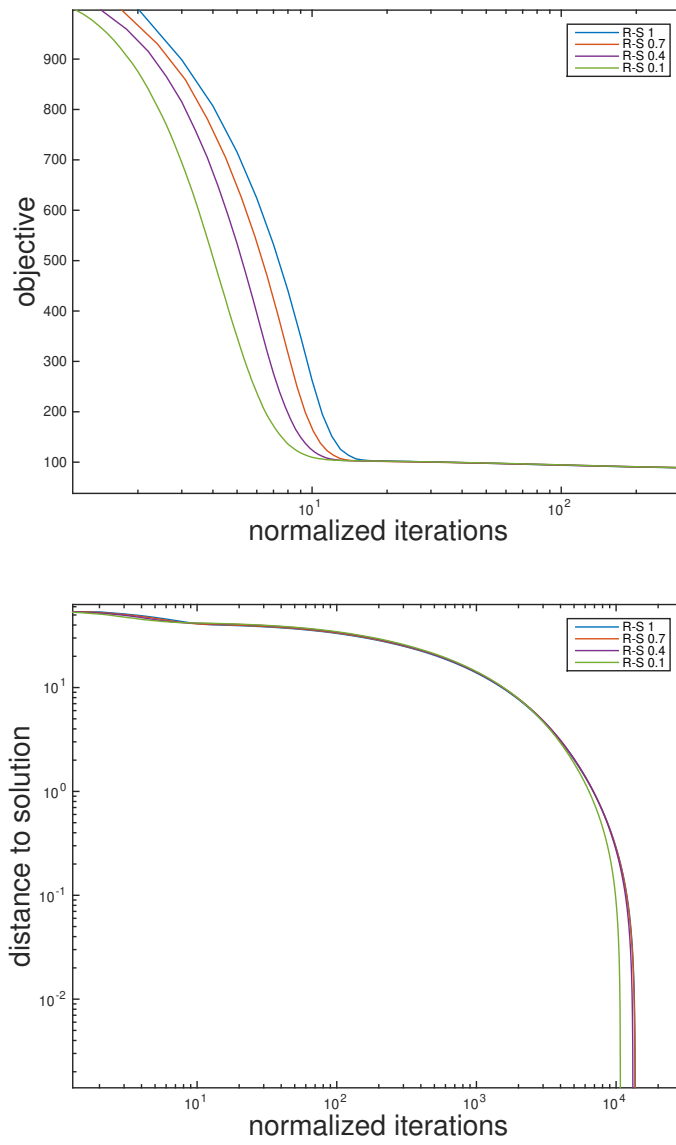
Figure 1: Objective for hinge loss classification with the latent group lasso (top), and distance to solution for the same (bottom).

6 Conclusion

We have investigated a general class of norms induced via an infimal postcomposition operation involving simpler norms and linear operators. Using tools from convex analysis, we have described properties of this class of regularizers and highlighted them with several concrete examples which have been used in different fields of data analysis. In terms of directions of future research, it would be interesting to further study conditions under which our framework can be extended to generate regularizers by means of more general functions and/or possibly nonlinear operators such as spectral functions. Algorithms to address this extended setting should also be investigated. From a statistical point of view, it would be valuable to characterize complexity measures for our regularizers, such as their Rademacher complexity, as this plays a central role in the study of estimation bounds.

References

[1] A. Argyriou, T. Evgeniou, and M. Pontil, Convex multi-task feature learning, *Machine Learn.*, vol. 73, pp. 243–272, 2008.

[2] A. Argyriou, R. Foygel, and N. Srebro, Sparse prediction with the k-support norm, in: *Advances in Neural Information Processing Systems*, vol. 25, pp. 1466–1474, 2012.

[3] A. Argyriou, C. A. Micchelli, and M. Pontil, On spectral learning, *J. Mach. Learn. Res.,* vol. 11, pp. 935–953, 2010.

[4] A. Asaei, M. Golbabaee, H. Bourlard, and V. Cevher, Structured sparsity models for reverberant speech separation, *IEEE/ACM Trans. Audio Speech Language Process.,* vol. 22, pp. 620–633, 2014.

[5] F. R. Bach, R. Jenatton, J. Mairal, and G. Obozinski, Optimization with sparsity-inducing penalties, *Found. Trends Machine Learn.*, vol. 4, pp. 1–106, 2012.

[6] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, Multiple kernel learning, conic duality, and the SMO algorithm, *Proc. 21st Int. Conf. Machine Learn.*, pp. 6–15, 2004.

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

[8] S. R. Becker and P. L. Combettes, An algorithm for splitting parallel sums of linearly composed monotone operators, with applications to signal recovery, *J. Nonlin. Convex Anal.*, vol. 15, pp. 137–159, 2014.

[9] A. Bourrier, M. E. Davies, T. Peleg, P. Pérez, and R. Gribonval, Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems, *IEEE Trans. Inform. Theory,* vol. 60, pp. 7928–7947, 2014.

[10] W. Cheney and W. Light, *A Course in Approximation Theory*. American Mathematical Society, Providence, RI, 2000.

[11] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. Willsky, The convex geometry of linear inverse problems, *Found. Comput. Math.*, vol. 12, pp. 805–849, 2012.

[12] P. L. Combettes, Systems of structured monotone inclusions: Duality, algorithms, and applications, *SIAM J. Optim.,* vol. 23, pp. 2420–2447, 2013.

[13] P. L. Combettes and J.-C. Pesquet, Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping, *SIAM J. Optim.,* vol. 25, pp. 1221–1248, 2015.

[14] P. L. Combettes and J.-C. Pesquet. Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping II: Mean-square and linear convergence, *Math. Programming,* published on line 2018-05-26.

[15] S. Gandy, B. Recht, and I. Yamada, Tensor completion and low-n-rank tensor recovery via convex optimization, *Inverse Problems,* vol. 27, art. 025010, 2011.

[16] C. Goulaouic, Prolongements de foncteurs d'interpolation et applications, *Ann. Inst. Fourier,* vol. 18, pp. 1–98, 1968.

[17] M. Herbster and G. Lever, Predicting the labelling of a graph via minimum p-seminorm interpolation, *Proc. 22nd Annual Conf. Learn. Theory,* 2009.

[18] L. Jacob, G. Obozinski, and J.-Ph. Vert, Group lasso with overlap and graph lasso, *Proc. 26th Int. Conf. Machine Learn.,* pp. 433–440, 2009.

[19] M. Jaggi and M. Sulovsky, A simple algorithm for nuclear norm regularized problems, *Proc. 27th Int. Conf. Machine Learn.,* pp. 471–478, 2010.

[20] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach, Proximal methods for hierarchical sparse coding, *J. Mach. Learn. Res.,* vol. 12, pp. 2297–2334, 2011.

[21] T. G. Kolda and B. W. Bader, Tensor decompositions and applications, *SIAM Rev.,* vol. 51, pp. 455–500, 2009.

[22] X. Liu, G. Zhao, J. Yao, and C. Qi, Background subtraction based on low-rank and structured sparse decomposition, *IEEE Trans. Image Process.,* vol. 24, pp. 2502–2514, 2015.

[23] S. Mallat and G. Yu, Super-resolution with sparse mixing estimators, *IEEE Trans. Image Process.,* vol. 19, pp. 2889–2900, 2010.

[24] A. Maurer and M. Pontil, Structured sparsity and generalization, *J. Mach. Learn. Res.,* vol. 13, pp. 671–690, 2012.

[25] A. M. McDonald, M. Pontil, and D. Stamos, Spectral k-support norm regularization, in: *Advances in Neural Information Processing Systems*, vol. 27 pp. 3644–3652, 2014.

[26] A. M. McDonald, M. Pontil, and D. Stamos, Fitting spectral decay with the k-support norm, *Proc. 19th Int. Conf. Artificial Intell. Stat. Machine Learn. Res.,* pp. 1061–1069, 2016.

[27] A. M. McDonald, M. Pontil, and D. Stamos, New perspectives on k-support and cluster norms, *J. Machine Learn. Res.,* vol. 17, pp. 1–38, 2016.

[28] C. A. Micchelli, J. M. Morales, and M. Pontil, Regularizers for structured sparsity, *Adv. Comput. Math.*, vol. 38, pp. 455–489, 2013.

[29] C. A. Micchelli and M. Pontil, Learning the kernel function via regularization, *J. Mach. Learn. Res.*, vol. 6, pp. 1099–1125, 2005.

[30] C. A. Micchelli and M. Pontil, Feature space perspectives for learning the kernel, *Machine Learn.*, vol. 66, pp. 297–319, 2007.

[31] C. A. Micchelli, L. Shen, and Y. Xu, Proximity algorithms for image models: denoising, *Inverse Problems*, vol. 27, art. 045009, 2011.

[32] J. Peetre, A new approach in interpolation spaces, *Studia Math.,* vol. 34, pp. 23–42, 1970.

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

[34] B. Romera-Paredes, H. Aung, N. Bianchi-Berthouze, and M. Pontil, Multilinear multitask learning, *Proc. 30th Int. Conf. Machine Learn.*, pp. 1444–1452, 2013.

[35] M. Signoretto, Q. T. Dinh, L. De Lathauwer, and J. A. K. Suykens, Learning with tensors: a framework based on convex optimization and spectral regularization, *Machine Learn.*, vol. 94, pp. 303–351, 2014.

[36] N. Srebro, J. D. M. Rennie, and T. S. Jaakkola, Maximum-margin matrix factorization, in: *Advances in Neural Information Processing Systems*, vol. 17, pp. 1329–1336, 2005.

[37] Y. Sun, X. Tao, Y. Li, and J. Lu, Robust 2D principal component analysis: A structured sparsity regularized approach, *IEEE Trans. Image Process.,* vol. 24, pp. 2515–2526, 20015.

[38] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, Sparsity and smoothness via the fused lasso, *J. Roy. Stat. Soc.*, vol. B67, pp. 91–108, 2005.

[39] R. Tomioka and T. Suzuki, Convex tensor decomposition via structured Schatten norm regularization, in: *Advances in Neural Information Processing Systems*, vol. 25, pp. 1331–1339, 2013.

[40] R. Tomioka, T. Sukuki, K. Hayashi, and H. Kashima, Statistical performance of convex tensor decomposition, in: *Advances in Neural Information Processing Systems*, vol. 23, pp. 972–980, 2011.

[41] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators.* North-Holland, New York, 1978.

[42] J. E. Ward and R. E. Wendell, Using block norms for location modeling, *Oper. Res.,* vol. 33, pp. 1074–1090, 1985.

[43] K. Wimalawarne, M. Sugiama, and R. Tomioka, Multitask learning meets tensor factorization: task imputation via convex optimization, in: *Advances in Neural Information Processing Systems*, vol. 26, pp. 2825–2833, 2014.

[44] M. Yuan and Y. Lin, Model selection and estimation in regression with grouped variables, *J. Roy. Stat. Soc.*, vol. B68, pp. 49–67, 2006.

[45] C. Zălinescu, *Convex Analysis in General Vector Spaces.* World Scientific, River Edge, NJ, 2002.

[46] X. Zeng and M. A. T. Figueiredo, Decreasing weighted sorted ℓ_1 regularization, *IEEE Signal Process. Lett.*, vol. 21, pp. 1240–1244, 2014.

[47] L. Zhang, W. Wei, C. Tian, F. Li, and Y. Zhang, Exploring structured sparsity by a reweighted Laplace prior for hyperspectral compressive sensing, *IEEE Trans. Image Process.*, vol. 25, pp. 4974–4988, 2016.

[48] P. Zhao, G. Rocha, and B. Yu, The composite absolute penalties family for grouped and hierarchical variable selection, *Ann. Stat.*, vol. 37, pp. 3468–3497, 2009.