A unified perspective on convex structured sparsity: Hierarchical, symmetric, submodular norms and beyond

Guillaume Obozinski
Université Paris-Est
Laboratoire d'Informatique Gaspard Monge
Groupe Imagine, Ecole des Ponts - ParisTech
Marne-la-Vallée, France
guillaume.obozinski@enpc.fr

Francis Bach
INRIA - Sierra project-team
Département d'Informatique
de l'Ecole Normale Supérieure
Paris, France
francis.bach@ens.fr

December 9, 2016

Abstract

In this paper, we propose a unified theory for convex structured sparsity-inducing norms on vectors associated with combinatorial penalty functions. Specifically, we consider the situation of a model simultaneously (a) penalized by a set-function defined on the support of the unknown parameter vector which represents prior knowledge on supports, and (b) regularized in ℓ_p-norm. We show that each of the obtained combinatorial optimization problems admits a natural relaxation as an optimization problem regularized by a matching sparsity-inducing norm.

To characterize the tightness of the relaxation, we introduce a notion of *lower combinatorial envelope* of a set-function. Symmetrically, a notion of *upper combinatorial envelope* produces the most concise norm expression. We show that these relaxations take the form of combinatorial latent group Lassos associated with min-cover penalties also known as *block-coding* schemes. For submodular penalty functions, the associated norm, dual norm and the corresponding proximal operator can be computed efficiently using a generic divide-and-conquer algorithm.

Our framework obtains constructive derivations for the Lasso, group Lasso, exclusive Lasso, the OWL, OSCAR and SLOPE penalties, the k-support norm, several hierarchical penalties considered in the literature for chains and tree structures, and produces also new norms. It leads to general efficient algorithms for all these norms, recovering as special cases several algorithms proposed in the literature and yielding improved procedures for some cases.

For norms associated with submodular penalties, including a large number of non-decomposable norms, we generalize classical support recovery and fast rates convergence results based respectively on generalization of the *irrepresentability condition* and the *restricted eigenvalue* condition.

1 Introduction

The last years have seen the emergence of the field of *structured sparsity*, which aims at identifying a model of small complexity given a priori knowledge on its possible structure.

Various regularizations, in particular convex, have been proposed that formalized the notion that prior information can be expressed through functions encoding the set of possible or encouraged supports[1] in the model. Several convex regularizers for structured sparsity arose as generalizations

[1] By support, we mean the set of indices of non-zero parameters.

of the group Lasso (Yuan and Lin, 2006) to the case of overlapping groups (Jacob et al., 2009; Jenatton et al., 2011a; Mairal et al., 2011), in particular to tree-structured groups (Jenatton et al., 2011b; Kim and Xing, 2010; Zhao et al., 2009b). Other formulations have been considered based on variational formulations (Micchelli et al., 2013), the perspective of multiple kernel learning (Bach et al., 2012), submodular functions (Bach, 2010) and norms defined as convex hulls (Chandrasekaran et al., 2012; Obozinski et al., 2011). Non convex approaches were introduced as well, by Baraniuk et al. (2010); He and Carin (2009); Huang et al. (2011). We refer the reader to Huang et al. (2011) for a concise overview and discussion of the related literature and to Bach et al. (2012) for a more detailed tutorial presentation.

In this context, and given a model parametrized by a vector of coefficients $w \in \mathbb{R}^V$ with $V = \{1, \ldots, d\}$, the main objective of this paper is to find an appropriate way to combine together *combinatorial penalties*, that control the structure of a model in terms of the sets of variables allowed or favored to enter the function learned, with *continuous regularizers* — such as ℓ_p-norms, that control the magnitude of their coefficients — into a convex regularization that would control both.

Part of our motivation stems from previous work on regularizers that “convexify” combinatorial penalties. Bach (2010) proposes to consider the tightest convex relaxation of the restriction of a submodular penalty to a unit ℓ_∞-ball in the space of model parameters $w \in \mathbb{R}^d$. However, this relaxation scheme implicitly assumes that the coefficients are in a unit ℓ_∞-ball; then, the obtained relaxation induces clustering artifacts of the values of the learned vector. It would thus seem desirable to propose relaxation schemes that do not assume that coefficients are bounded but rather to control continuously their magnitude and to find alternatives to the ℓ_∞-norm. Finally the class of functions considered is restricted to submodular functions.

Yet another motivation is to follow loosely the principle of two-part or multiple-part codes from minimum description length (MDL) theory (Rissanen, 1978). In particular if the model is parametrized by a vector of parameters w, it is possible to encode (an approximation of) w itself with a two-part code, by encoding first the support $\mathrm{Supp}(w)$ — or set of non-zero values — of w with a code length of the form $F(\mathrm{Supp}(w))$ and by encoding the actual values of w using a code based on a log prior distribution on the vector w that could motivate the choice of an ℓ_p-norm as a surrogate for the code length. This leads naturally to consider penalties of the form $\mu F(\mathrm{Supp}(w)) + \nu \|w\|_p^p$ and to find appropriate notions of relaxation.

In this paper, we therefore consider combined penalties of the form mentioned above and propose first an appropriate convex relaxation in Section 2; first elementary examples are listed in Section 2.1; the properties of general combinatorial functions preserved by the relaxation are captured by the notion of lower combinatorial envelope introduced in Section 2.2. In Section 2.3, we introduce the upper combinatorial envelope, which provides concise representation of the norm and establishes links with atomic norms. Section 3 relates the obtained norms to the latent group Lasso and to set-cover penalties. In Section 4, we provide first examples of instances of the norms, in particular, by considering what we call overlap count Lasso norms; we relate the proposed norms to overlapped ℓ_1/ℓ_p-group norms and with the latent group Lasso in Section 4.1. The exclusive Lasso is presented in Section 4.3. After introducing key variational forms of the norm in Section 5, we discuss the case of submodular functions in Section 6 and propose in particular general algorithms to compute each norm, its dual and its associated proximal operator. Based on this theory, we study more sophisticated examples of the norms in Section 7. In particular, we discuss the case of overlap count Lasso norms in Section 7.1, the case of norms for hierarchical sparsity in Section 7.2 and the case of symmetric norms associated to functions of the cardinality of the support in section 7.3. In Section 8, we extend two statistical results that are classical for the Lasso to all norms associated with submodular functions, namely a result of support recovery based on an irrepresentability condition and fast rates based on a restricted eigenvalue condition. Finally, we present some experiments in Section 9.

Notations. When indexing vectors of $\mathbb{R}^d$ with a set A or B in *exponent*, x^A and $x^B \in \mathbb{R}^d$ refer to two a priori unrelated vectors; by contrast, when using A as an *index*, and given a vector $x \in \mathbb{R}^d$, x_A denotes the vector of $\mathbb{R}^d$ such that $[x_A]_i = x_i,\ i \in A$ and $[x_A]_i = 0,\ i \notin A$. If s is a vector in $\mathbb{R}^d$, we use the shorthand $s(A) := \sum_{i \in A} s_i$ and $|s|$ denotes the vector whose elements are the absolute values $|s_i|$ of the elements s_i in s. For $p \geq 1$, we define q through the relation $\frac{1}{p} + \frac{1}{q} = 1$. The ℓ_q-norm of a vector w will be noted $\|w\|_q = \big(\sum_i w_i^q\big)^{1/q}$. For a function $f : \mathbb{R}^d \to \mathbb{R}$, we will denote by f^* is Fenchel-Legendre conjugate. We will write $\overline{\mathbb{R}}_+$ for $\mathbb{R}_+ \cup \{+\infty\}$. We will denote by $\iota_{x \in S}$ the indicator function of the set S, taking value 0 on the set and $+\infty$ outside. We will write $[\![k_1, k_2]\!]$ to denote the discrete interval $\{k_1, \dots, k_2\}$.

2 Penalties and convex relaxations

Let $V = \{1, \dots, d\}$ and $2^V = \{A \mid A \subset V\}$ its power-set. We will consider positive-valued set-functions of the form $F : 2^V \to \overline{\mathbb{R}}_+$ such that $F(\varnothing) = 0$ and $F(A) > 0$ for all $A \neq \varnothing$. We do not necessarily assume that F is non-decreasing, even if it would a priori be natural for a penalty function of the support. We however assume that the domain of F, defined as $\mathcal{D}_0 := \{A \mid F(A) < \infty\}$, covers V, i.e., satisfies $\cup_{A \in \mathcal{D}_0} A = V$ (if F is non-decreasing, this just implies that it should be finite on singletons).

With the motivations of the previous section, and denoting by $\mathrm{Supp}(w)$ the set of non-zero coefficients of a vector w, we consider a penalty involving both a *combinatorial* function F and ℓ_p-regularization:

$$\mathrm{pen} : w \mapsto \mu\, F(\mathrm{Supp}(w)) + \nu\, \|w\|_p^p, \tag{1}$$

where μ and ν are strictly positive scalar coefficients. Since such non-convex discontinuous penalizations are untractable computationally, we undertake to construct an appropriate convex relaxation. The most natural convex surrogate for a non-convex function, say A, is arguably its *convex envelope* (i.e., its tightest convex lower bound) which can be computed as its Fenchel-Legendre bidual A^{**}. However, one relatively natural requirement for a regularizer is to ask that it be also *positively homogeneous* (p.h.) since this leads to formulations that are invariant by rescaling of the data. Our goal will therefore be to construct the tightest positively homogeneous convex lower bound of the penalty considered.

Now, it is a classical result that, given a function A, its tightest p.h. (but not necessarily convex) lower bound A_h is $A_h(w) = \inf_{\lambda > 0} \frac{A(\lambda w)}{\lambda}$ (see Rockafellar, 1970, p.35). This is instrumental here given the following proposition:

Proposition 1. *Let $A : \mathbb{R}^d \to \mathbb{R}_+$ be a real valued function, A_h defined as above. Then C, the tightest positively homogeneous and convex lower bound of A, is well-defined and $C = A_h^{**}$.*

Proof. The set of convex p.h. lower bounds of A is non-empty (since it contains the constant zero function) and stable by taking pointwise suprema. Therefore it has a unique majorant, which we call C. We have for all $w \in \mathbb{R}^d$, $A_h^{**}(w) \leqslant C(w) \leqslant A(w)$, by definition of C, the fact that A_h is an p.h. lower bound on A and that Fenchel bi-conjugation preserves homogeneity. (It can indeed be checked that the conjugate of a homogeneous function h is the indicator of the polar of $\{w \mid h(w) \leq 1\}$; then, since polar sets are closed convex sets containing the origin, the bi-conjugate function is the support function of this polar set and must therefore be a gauge; finally gauges are homogeneous (see Rockafellar, 1970, for more details)). We thus have for all $\lambda > 0$, $A_h^{**}(\lambda w)\lambda^{-1} \leqslant C(\lambda w)\lambda^{-1} \leqslant A(\lambda w)\lambda^{-1}$, which implies that for all $w \in \mathbb{R}^d$, $A_h^{**}(w) \leqslant C(w) \leqslant A_h(w)$. Since C is convex, we must have $C = A_h^{**}$, hence the desired result. □

Using its definition we can easily compute the tightest positively homogeneous lower bound of the penalization of Eq. (1), which we denote pen_h:

$$\mathrm{pen}_h(w) \quad = \quad \inf_{\lambda>0} \frac{\mu}{\lambda} F(\mathrm{Supp}(w)) + \nu\, \lambda^{p-1}\, \|w\|_p^p.$$

Setting the gradient of the convex objective to 0, one gets that the minimum is obtained for $\lambda = \left(\frac{\mu q}{\nu p}\right)^{1/p} F(\mathrm{Supp}(w))^{1/p}\, \|w\|_p^{-1}$, and that

$$\mathrm{pen}_h(w) = (q\mu)^{1/q}\, (p\nu)^{1/p}\, \Theta(w),$$

where we introduced the notation

$$\Theta(w) := \; F(\mathrm{Supp}(w))^{1/q}\, \|w\|_p.$$

Up to a constant factor depending on the choices of μ and ν, we are therefore led to consider the positively homogeneous penalty Θ we just defined, which combines the two terms *multiplicatively*. Consider the norm Ω_p (or Ω_p^F if a reference to F is needed) whose dual norm[2] is defined as

$$\Omega_p^*(s) := \max_{A \subset V, A \neq \varnothing} \frac{\|s_A\|_q}{F(A)^{1/q}}. \tag{2}$$

We have the following result:

Proposition 2 (Convex relaxation)**.** *The norm Ω_p is the convex envelope of Θ.*

Proof. Denote $\Theta(w) = \|w\|_p\, F(\mathrm{Supp}(w))^{1/q}$, and compute its Fenchel conjugate:

$$\begin{aligned}
\Theta^*(s) \quad &= \quad \max_{w \in \mathbb{R}^d} w^\top s - \|w\|_p\, F(\mathrm{Supp}(w))^{1/q}, \text{ by definition of } \Theta^*, \\
&= \quad \max_{A \subset V} \max_{w_A \in \mathbb{R}_*^{|A|}} w_A^\top s_A - \|w_A\|_p\, F(A)^{1/q} \text{ by decomposing on subsets of } V, \\
&= \quad \max_{A \subset V} \iota_{\{\|s_A\|_q \leqslant F(A)^{1/q}\}} = \iota_{\{\Omega_p^*(s) \leqslant 1\}},
\end{aligned}$$

where $\iota_{\{s \in S\}}$ is the indicator of the set S, that is the function equal to 0 on S and $+\infty$ on S^c. The Fenchel bidual of Θ, i.e., its largest (thus tightest) convex lower bound, is therefore exactly Ω_p. □

Note that the function F is not assumed submodular in the previous result. Since the function Θ depends on w only through $|w|$, by symmetry, the norm Ω_p is also a function of $|w|$; such norms are often called *absolute* (Stewart and Sun, 1990). Given Proposition 1, we have the immediate corollary:

Corollary 1 (Two parts-code relaxation)**.** *Let $p > 1$. The norm $w \mapsto (q\mu)^{1/q}(p\nu)^{1/p}\, \Omega_p(w)$ is the tightest convex* positively homogeneous *lower bound of the function $w \mapsto \mu F(\mathrm{Supp}(w)) + \nu \|w\|_p^p$.*

The penalties and relaxation results considered in this section are illustrated on Figure 1.

[2] The assumptions on the domain $\mathcal{D}_0$ of F and on the positivity of F indeed guarantee that Ω_p^* is a norm.

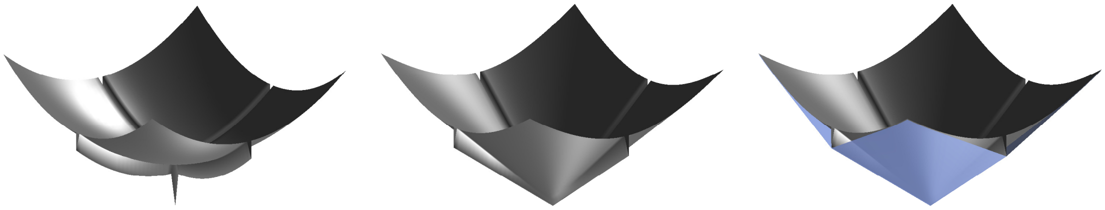


Figure 1: **Penalties in 2D.** Left: graph of the penalty `pen`. Middle: graph of penalty $\texttt{pen}_h$ with $p = 2$. Right: graph of the norm Ω_2^F in blue overlaid over graph of $\texttt{pen}_h$. All of them are for the combinatorial function $F : 2^V \to \mathbb{R}^+$, with $F(\varnothing) = 0$, $F(\{1\}) = F(\{2\}) = 0.65$ and $F(\{1, 2\}) = 1$.

2.1 Special cases

Case $p = 1$. In that case, we have $q = \infty$, and we always have $\Omega_1 = \| \cdot \|_1$, which can be seen from the definition of Θ or from Eq. (2). But regularizing with an ℓ_1-norm leads to estimators that can potentially have all possible sparsity patterns and in that sense an ℓ_1-norm cannot encode hard structural constraints on the patterns. Since this means in other words that the ℓ_1-relaxations essentially lose the combinatorial structure of allowed sparsity patterns possibly encoded in F, we focus, from now on, on the case $p > 1$.

Lasso, group Lasso. Our norm Ω_p instantiates as the ℓ_1, ℓ_p and ℓ_1/ℓ_p-norms for the simplest functions:

- If $F(A) = |A|$, then $\Omega_p(w) = \|w\|_1$, since $\Omega_p^*(s) = \max_{A\subset V} \frac{\|s_A\|_q}{|A|^{1/q}} = \big(\max_{A\subset V} \frac{|s|^q(A)}{|A|}\big)^{1/q} = \|s\|_\infty$. It is interesting that the cardinality function is always relaxed to the ℓ_1-norm for all ℓ_p-relaxations, and that it is not an artifact of the traditional relaxation on an ℓ_∞-ball.
- If $F(A) = 1_{\{A \neq \varnothing\}}$, then $\Omega_p(w) = \|w\|_p$, since $\Omega_p^*(s) = \max_{A\subset V} \|s_A\|_q = \|s\|_q$.
- If $F(A) = \sum_{j=1}^{g} 1_{\{A\cap G_j \neq \varnothing\}}$, for $(G_j)_{j\in\{1,\dots,g\}}$ a partition of V, then $\Omega_p(w) = \sum_{j=1}^{g} \|w_{G_j}\|_p$ is the group Lasso or ℓ_1/ℓ_p-norm (Yuan and Lin, 2006). This result provides a principled derivation for the form of these norms, which did not exist in the literature. For groups which do not form a partition, this identity does in fact not hold in general for $p < \infty$, as we discuss in Section 4.1.

Submodular functions and $p = \infty$. For a submodular function F and in the $p = \infty$ case, the norm Ω_∞^F that we derived actually coincides with the relaxation proposed by Bach (2010), and as showed in that work, $\Omega_\infty^F(w) = f(|w|)$, where f is a function associated with F and called the *Lovász extension* of F. We discuss the case of submodular functions in detail in Section 6.

2.2 Lower combinatorial envelope

The fact that, when F is a submodular function, Ω_∞^F is equal to the Lovász extension f on the positive orthant provides a guarantee on the tightness of the relaxation. Indeed f is called an "extension" because $\forall A \subset 2^V$, $f(1_A) = F(A)$, so that f can be seen to extend the function F to $\mathbb{R}^d$ (set-functions are naturally defined as functions on the vertices of the hypercube, that is, $\{0, 1\}^d$, and thus f extends this representation of set-functions).

As a consequence, when F is submodular, $\Omega_\infty^F(1_A) = f(1_A) = F(A)$, which means that the relaxation is tight for all w of the form $w = c\,1_A$, for any scalar constant $c \in \mathbb{R}$ and any set $A \subset V$. If F is not submodular, this property does not necessarily hold, thereby suggesting that the relaxation could be less tight in general. To characterize to which extend this is true, we introduce a couple of new concepts.

Many of the properties of Ω_p, for any $p > 1$, are captured by the unit ball of Ω_∞^* or its intersection with the positive orthant. In fact, as we will see in the sequel, the ℓ_∞-relaxation plays a particular role, to establish properties of the norm, to construct algorithms and for the statistical analysis, since it it reflects most directly the combinatorial structure of the function F.

We define the *canonical polyhedron*[3] associated with the combinatorial function as the polyhedron $\mathcal{P}_F$ defined by

$$\mathcal{P}_F = \big\{s \in \mathbb{R}_+^d,\ \forall A \subset V,\ s(A) \le F(A)\big\}.$$

By construction, it is immediate that the unit ball of Ω_∞^* is $\{s \in \mathbb{R}^d \mid |s| \in \mathcal{P}_F\}$.

From this polyhedron, we construct a new set-function which reflects the features of F that are captured by $\mathcal{P}_F$:

Definition 2 (Lower combinatorial envelope)**.** *Define the* lower combinatorial envelope *(LCE) of F as the set-function F_- defined by:*

$$F_-(A) = \max_{s \in \mathcal{P}_F} s(A) = \max_{s \in \mathbb{R}_+^d,\ \forall B \subset V, s(B) \leqslant F(B)} s(A).$$

By construction, (a) for any $A \subset V$, $F_-(A) \leqslant F(A)$ and, (b) even when F is not monotonic, F_- is always non-decreasing (because $\mathcal{P}_F \subset \mathbb{R}_+^d$).

One of the key properties of the lower combinatorial envelope is that, as shown in the next lemma, Ω_∞^F is an extension of F_- (and not of F in general), in the same way that the Lovász extension is an extension of F when F is submodular.

Lemma 1 (Extension property)**.** *For any $A \subset V$, we have $\Omega_\infty^F(1_A) = F_-(A)$.*

Proof. From the definitions of $\mathcal{P}_F$ and F_-, we get: $\Omega_\infty^F(1_A) = \max_{[\Omega_\infty^F]^*(s) \le 1} 1_A^\top s = \max_{s \in \mathcal{P}_F} s^\top 1_A = F_-(A)$. □

A second important property is that a function F and its LCE F_- share the same canonical polyhedron $\mathcal{P}_F$.

Lemma 2 (Equality of canonical polyhedra)**.** $\mathcal{P}_F = \mathcal{P}_{F_-}$.

Proof. Since $F_- \le F$, any $s \in \mathcal{P}_{F_-}$ is such that $s(A) \le F_-(A) \le F(A)$ for any A so that clearly $\mathcal{P}_{F_-} \subset \mathcal{P}_F$. Now conversely, for any $s \in \mathcal{P}_F$, any for any A, we have $s(A) \le \max_{s' \in \mathcal{P}_F} s'(A) = F_-(A)$, so that $s \in \mathcal{P}_{F_-}$ which implies $\mathcal{P}_F \subset \mathcal{P}_{F_-}$. □

But the sets $\{w \in \mathbb{R}^d \mid |w| \in \mathcal{P}_F\}$ and $\{w \in \mathbb{R}^d \mid |w| \in \mathcal{P}_{F_-}\}$ are respectively the unit balls of Ω_∞^F and $\Omega_\infty^{F_-}$. As a direct consequence, we have:

Lemma 3 (Equality of norms)**.** *For all $p \ge 1$,* $\quad \Omega_p^F = \Omega_p^{F_-}$.

[3] The reader familiar with submodular functions will recognize that for these functions the canonical polyhedron is the intersection of the submodular polyhedron with the positive orthant.

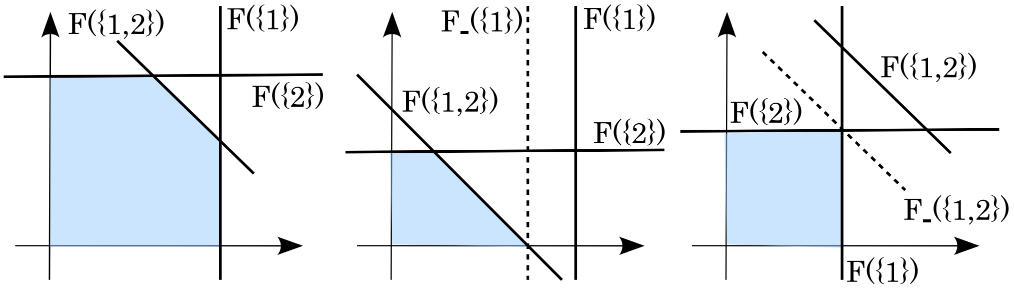


Figure 2: Intersection of the canonical polyhedron with the positive orthant for three different functions F. Full lines materialize the inequalities $s(A) \leq F(A)$ that define the polyhedron. Dashed lines materialize the induced constraints $s(A) \leq F_-(A)$ that results from all constraints $s(B) \leq F(B)$, $B \in 2^V$. From left to right: (i) submodular case, that is, $\mathcal{D}_F = 2^V$ and $F_- = F = F_+$; (ii) $\mathcal{D}_F = \{\{2\}, \{1, 2\}\}$ and $F_-(\{1\}) < F(\{1\})$; (iii) $\mathcal{D}_F = \{\{1\}, \{2\}\}$ corresponding to a weighted ℓ_1-norm.

Lemma 4 (Lower envelope properties)**.** *The operator $\mathcal{L} : F \mapsto F_-$ is order-preserving (i.e., if $G \leq F$ then $G_- \leq F_-$), idempotent (i.e., $F_{--} = F_-$), and F_- is the unique pointwise smallest combinatorial function among all functions G such that $\mathcal{P}_F = \mathcal{P}_G$.*

Proof. To see that $\mathcal{L}$ is order preserving, note that if $G \leq F$, then $\mathcal{P}_G \subset \mathcal{P}_F$ so that $G_-(A) = \max_{s \in \mathcal{P}_G} s(A) \leq \max_{s \in \mathcal{P}_F} s(A) = F_-(A)$. Idempotence follows from Lemma 2: indeed, since $\mathcal{P}_F = \mathcal{P}_{F_-}$, we have $F_{--}(A) = \max_{s \in \mathcal{P}_{F_-}} s(A) = \max_{s \in \mathcal{P}_F} s(A) = F_-(A)$, which shows the result. Finally, if $\mathcal{P}_F = \mathcal{P}_G$ we have $G_- = F_-$, in particular $F_- \leq G$. Since F_- itself satisfies the property that $\mathcal{P}_F = \mathcal{P}_{F_-}$, this shows that this is indeed the smallest element in that set. □

Note that this shows that F_- is really a *combinatorial counterpart of the convex envelope*. Indeed, the operator which maps the function f to its convex envelope is also order-preserving and idempotent, and while the convex envelope of f provides a lower bound of f which is the pointwise infimum of all the functions that are above all the affine functions smaller than f, the LCE is a lower bound of F which is the pointwise infimum of all the function that are greater than all the non-decreasing modular functions smaller than F.

Figure 2 illustrates the fact that F and F_- share the same canonical polyhedron and that the value of $F_-(A)$ is determined by the values that F takes on other sets. This figure also suggests that some constraints $\{\, s(A) \leq F(A) \,\}$ can never be active and could therefore be removed. This will be formalized in Section 2.3.

To illustrate the relevance of the concept of lower combinatorial envelope, we compute it for a few examples.

Example 1 (Basic functions)**.** *For $A \mapsto |A|$, we have $|A|_- = |A|$ because by the extension property $|A|_- = \Omega_\infty^{|\cdot|}(1_A) = \|1_A\|_1 = |A|$. Likewise, for $F : A \mapsto 1_{\{A \neq \varnothing\}}$, $F_-(A) = \|1_A\|_\infty = F(A)$ and for the combinatorial function associated with the group Lasso and defined by $F(A) := \sum_{B \in \mathcal{G}} 1_{\{A \cap B \neq \varnothing\}}$, with $\mathcal{B}$ a partition of V, we have $F_-(A) = \sum_{B \in \mathcal{G}} \|[1_A]_B\|_\infty = \sum_{B \in \mathcal{G}} \|[1_{A \cap B}\|_\infty = F(A)$. In fact, since all these functions are submodular we have $\Omega_\infty^F(w) = f(|w|)$ for f the Lovász extension of F, which satisfies $f(1_A) = F(A)$, so that we necessarily have $F_-(A) = f(1_A) = F(A)$.*

Example 2 (Range function)**.** *Consider, on $V = [\![1, d]\!]$, the range function $F : A \mapsto \max(A) - \min(A) + 1$ where $\min(A)$ (resp. $\max(A)$) is the smallest (resp. largest) element in A. A motivation to consider this function is that it induces the selection of supports that are exactly intervals. Since the range is always larger than the cardinality we have $F(A) \geq |A|$ for all A and so since taking LCEs is order-preserving and using that $|A|_- = |A|$ we have $F_-(A) \geq |A|_- = |A|$. On the other hand, $F_-(A) = \max_{s \in \mathcal{P}_F} s(A) \leq \sum_{i \in A} s_i \leq |A|$ because $s_i \leq F(\{i\}) = 1$. Combining these inequalities proves that $F_-(A) = |A|$. As an immediate consequence $\Omega_p^F = \| \cdot \|_1$ which does not tend to favor supports that are intervals. In this case, the structure encoded in the combinatorial function is lost in the relaxation...*

To summarize, the LCE of a function F is the combinatorial function that is actually extended by the norm Ω_p^F. It thus essentially worth considering only combinatorial functions that are equal to their LCE.

2.3 Upper combinatorial envelope

Let F be a set-function and $\mathcal{P}_F$ its canonical polyhedron. In this section, we follow an intuition conveyed by Figure 2 and find a compact representation of F: the polyhedron $\mathcal{P}_F$ has in many cases a number of faces which much smaller than 2^d. We formalize this in the next lemma.

Lemma 5 (Core set)**.** *There exists a unique minimal subset $\mathcal{D}_F$ of 2^V such that for $s \in \mathbb{R}_+^d$,*

$$s \in \mathcal{P}_F \Leftrightarrow (\forall A \in \mathcal{D}_F,\ s(A) \leq F(A)).$$

Proof. If $\mathcal{C}_F$ is the convex hull of $\{0\} \cup \{F(A)^{-1} 1_A\}_{A \subset V, A \neq \varnothing}$ and $\mathcal{A}_F$ the set of vertices of the polytope $\mathcal{C}_F$ that are different from 0, then, for $s \in \mathbb{R}_+^d$ we have

$$\big(s \in \mathcal{P}_F\big) \Leftrightarrow \big(\max_{\varnothing \neq A \subset V} \langle s, F(A)^{-1}\, 1_A \rangle \leq 1\big) \Leftrightarrow \big(\max_{c \in \mathcal{C}_F} \langle s, c \rangle \leq 1\big) \Leftrightarrow \big(\max_{a \in \mathcal{A}_F} \langle s, a \rangle \leq 1\big).$$

But we must have $\mathcal{A}_F \subset \{F(A)^{-1} 1_A\}_{A \subset V, A \neq \varnothing}$ and so there exists a set $\mathcal{D}_F$ such that $\mathcal{A}_F = \{F(A)^{-1} 1_A\}_{A \in \mathcal{D}_F}$. This set satisfies the property announced in the lemma and is clearly minimal, because removing a vertex would lead to a convex hull strictly included in $\mathcal{C}_F$ whose polar would strictly include $\mathcal{P}_F$. □

We call $\mathcal{D}_F$ the *core set* of F. It corresponds to the set of faces of dimension $d - 1$ of $\mathcal{P}_F$. Note that the set $\mathcal{A}_F$ is almost the set of atoms characterizing the norm in the sense of Chandrasekaran et al. (2012). More precisely, since the norm Ω_∞^F is such that $\Omega_\infty^F(w) = \Omega_\infty^F(|w|)$, i.e. the norm is an *absolute norm* (Bach et al., 2012, p. 27), it follows from the previous result that Ω_∞^F is the *atomic norm* in the sense of Chandrasekaran et al. (2012) associated with the collection of atoms $\mathcal{A}_F^{\text{sym}} := \big\{a \in \{-1, 0, 1\}^d \mid |a| \in \mathcal{A}_F\big\}$. Similarly, it is easy to show that Ω_p^F is the atomic norm associated with the following set of atoms $\{u \in \mathbb{R}^d,\ \|u\|_p = 1, u_{A^c} = 0 \text{ for some } A \in \mathcal{D}_F\}$. This is illustrated in Figure 4 and 5.

This notion motivates the definition of a new set-function:

Definition 3 (Upper combinatorial envelope)**.** *We call* upper combinatorial envelope *(UCE) the function F_+ defined by $F_+(A) = F(A)$ for $A \in \mathcal{D}_F$ and $F_+(A) = \infty$ otherwise.*

As the reader might expect at this point, F_+ provides a compact representation which captures all the information about F that is preserved in the relaxation:

Proposition 3 (Equality of canonical polyhedra)**.** *F, F_- and F_+ all define the same canonical polyhedron $\mathcal{P}_{F_-} = \mathcal{P}_F = \mathcal{P}_{F_+}$ and share the same core set $\mathcal{D}_F$. Moreover, $\forall A \in \mathcal{D}_F$, $F_-(A) = F(A) = F_+(A)$.*

Proof. To show that $\Omega_p^{F_+} = \Omega_p^F$ we just need to show $\mathcal{P}_{F_+} = \mathcal{P}_F$. By the definition of F_+ we have $\mathcal{P}_{F_+} = \{s \in \mathbb{R}^d \mid s(A) \leq F(A),\ A \in \mathcal{D}_F\}$ but the previous lemma precisely states that the last set is equal to $\mathcal{P}_F$.

We now argue that, for all $A \in \mathcal{D}_F$, $F_-(A) = F(A) = F_+(A)$. Indeed, the equality $F(A) = F_+(A)$ holds by definition, and, for all $A \in \mathcal{D}_F$, we need to have $F(A) = F_-(A)$: by polarity, and with notations of Lemma 5, the fact that $\mathcal{P}_F = \mathcal{P}_{F_-}$ entails that $\mathcal{C}_F = \mathcal{C}_{F_-}$, so that $F_-(A)^{-1}1_A \in \mathcal{C}_F$, and, if we had $F_-(A) < F(A)$ then $F(A)^{-1}1_A$ would be a strict convex combination of the origin and $F_-(A)^{-1}1_A$, which contradicts the fact that $F(A)^{-1}1_A$ is an extreme point of $\mathcal{C}_F$. □

Finally, the term "upper combinatorial envelope" is motivated by the following lemma:

Lemma 6 (Upper envelope property). *F_+ is the pointwise supremum of all the set-functions H such that $\mathcal{P}_H = \mathcal{P}_F$.*

Proof. If $\mathcal{P}_F = \mathcal{P}_H$ then we must have $\mathcal{C}_{F_+} = \mathcal{C}_H$, which is only possible if $F(A)^{-1}1_A \in \mathcal{C}_H$ for all A; in particular, for all $A \in \mathcal{D}_F$, since $F(A)^{-1}1_A$ is an extreme point of $\mathcal{C}_{F_+}$ it must also be an extreme point of $\mathcal{C}_H$ because of the inclusion $\mathcal{C}_H \subset \mathcal{C}_{F_+}$, so that we must have $H(A) = F(A) = F_+(A)$ for all $A \in \mathcal{D}_F$. For any set $A \notin \mathcal{D}_F$, we clearly have $H(A) \leq F_+(A)$ since $F_+(A) = +\infty$. Finally, we proved in 3 that $\mathcal{P}_{F_+} = \mathcal{P}_F$ so that F_+ is indeed the largest element in the above defined set of functions. □

Example 3. *(Basic functions)*

- *For $F = |\cdot|$, we have $(\Omega_\infty^F)^* = \|\cdot\|_\infty$ so that $\mathcal{P}_F = [0,1]^d$. This shows that $\mathcal{D}_F$ is the set of singletons $\mathcal{D}_F = \big\{\{1\}, \dots, \{d\}\big\}$.*
- *For $F = 1_{\{A \neq \varnothing\}}$, since $(\Omega_\infty^F)^* = \|\cdot\|_1$, we have $\mathcal{P}_F = \{s \in \mathbb{R}_+^d \mid s(V) \leq F(V)\}$ so that the coreset is $\mathcal{D}_F = \{V\}$.*
- *For the group Lasso with $\mathcal{G}$ a partition of V, we have $(\Omega_\infty^F)^*(s) = \max_{B \in \mathcal{G}} \|s(B)\|_1$, so that $\mathcal{P}_F = \{s \in \mathbb{R}_+^d \mid s(B) \leq F(B), B \in \mathcal{G}\}$. Clearly, given that $\mathcal{G}$ is a partition, none of the constraints indexed by $\mathcal{G}$ can be removed so that $\mathcal{D}_F = \mathcal{G}$.*

The picture that emerges at this point from the results above is rather simple: any combinatorial function F defines a polyhedron $\mathcal{P}_F$ whose faces of dimension $d-1$ are indexed by a set $\mathcal{D}_F \subset 2^V$ that we called the *core set*. In symbolic notation: $\mathcal{P}_F = \{s \in \mathbb{R}^d \mid s(A) \leq F(A),\ A \in \mathcal{D}_F\}$. All the combinatorial functions which are equal to F on $\mathcal{D}_F$ and which otherwise take values that are larger than its lower combinatorial envelope F_-, have the same ℓ_p tightest positively homogeneous convex relaxation Ω_p^F for all $p > 1$, the smallest such function being F_- and the largest F_+. Moreover $F_-(A) = \Omega_\infty^F(A)$, so that Ω_∞^F is an extension of F_-. By construction, and even if F is a non-decreasing function, F_- is non-decreasing, while F_+ is obviously not a non-decreasing function, even though its restriction to $\mathcal{D}_F$ is. It might therefore seem an odd set-function to consider; however if $\mathcal{D}_F$ is a small set, since $\Omega_p^F = \Omega_p^{F_+}$, it provides a potentially much more compact representation of the norm, which we now relate to a norm previously introduced in the literature.

3 Latent group Lasso, block-coding and set-cover penalties

The norm Ω_p is actually not a new norm. It was introduced from a different point of view by Jacob et al. (2009) (see also Obozinski et al., 2011) as one of the possible generalizations of the group Lasso to the case where groups overlap.

To establish the connection, we now provide a more explicit form for Ω_p, which is different from the definition via its dual norm which we have exploited so far.

We consider models that are parameterized by a vector $w \in \mathbb{R}^V$ and associate to them latent variables that are tuples of vectors of $\mathbb{R}^V$ indexed by the power-set of V. Precisely, with the notation

$$\mathcal{V} = \big\{v = (v^A)_{A \subset V} \in \big(\mathbb{R}^V\big)^{2^V} \;\text{ s.t. }\; \mathrm{Supp}(v^A) \subset A\big\},$$

given a set function $F : 2^V \to \bar{\mathbb{R}}_+$, we define the norms Ω_p as (see an illustration in Figure 3):

$$\Omega_p(w) = \min_{v \in \mathcal{V}} \sum_{A \subset V} F(A)^{\frac{1}{q}} \, \|v^A\|_p \;\text{ s.t. }\; w = \sum_{A \subset V} v^A. \tag{3}$$

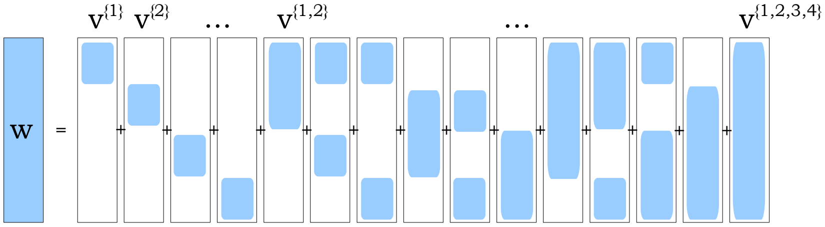


Figure 3: Illustration of the decomposition of w into $w = \sum_{A \subset V} v^A$.

As suggested by notations and as first proved for $p = 2$ by Jacob et al. (2009), we have:

Lemma 7. *Ω_p and Ω_p^* are dual to each other.*

An elementary proof of this result is provided by Obozinski et al. (2011)[4]. We propose a slightly more abstract proof of this result in appendix A using explicitly the fact that Ω_p is defined as an infimal convolution.

We will refer to this norm Ω_p as the *latent group Lasso* since it is defined by introducing latent variables v^A that are themselves regularized instead of the original model parameters. We refer the reader to Obozinski et al. (2011) for a detailed presentation of this norm, some of its properties and some support recovery results in terms of the support of the latent variables. In Jacob et al. (2009) the expansion (3) did not involve all terms of the power-set but only a subcollection of sets $\mathcal{G} \subset 2^V$. The notion of core set discussed in Section 2.3 is dual to the notion of redundant set introduced by Obozinski et al. (2011, Sec. 8.1).

The motivation of Jacob et al. (2009) was to find a convex regularization which would induce sparsity patterns that are unions of groups in $\mathcal{G}$ and explain the estimated vector w as a combination of a small number of latent components, each supported on one group of $\mathcal{G}$. The motivation is very similar in Huang et al. (2011) who consider an ℓ_0-type penalty they call *block coding*, where each support is penalized by the minimal sum of the *coding complexities* of a certain number of elementary sets called "blocks" which *cover* the support. In both cases the underlying combinatorial penalty is the *minimal weighted set cover* defined for a set $B \subset V$ by:

$$\widetilde{F}(B) \quad = \quad \min_{(\delta^A)_{A \subset V}} \sum_{A \subset V} F(A)\, \delta^A \qquad \text{s.t.} \qquad \sum_{A \subset V} \delta^A 1_A \geq 1_B, \qquad \delta^A \in \{0,1\},\; A \subset V.$$

While the norm proposed by Jacob et al. (2009) can be viewed as a form of "relaxation" of the cover-set problem, a rigorous link between the ℓ_0 and convex formulation is missing. We will make this statement rigorous through a new interpretation of the lower combinatorial envelope of F.

[4] The proof in Obozinski et al. (2011) addresses the $p = 2$ case but generalizes immediately to other values of p.

Indeed, assume w.l.o.g. that $w \in \mathbb{R}_+^d$. For $x, y \in \mathbb{R}^V$, we write $x \geq y$ if $x_i \geq y_i$ for all $i \in V$. Then,

$$\begin{aligned}\Omega_\infty(w) &= \min_{v\in\mathcal{V}} \sum_{A\subset V} F(A)\|v^A\|_\infty \qquad \text{s.t.} \qquad \sum_{A\subset V} v^A \geq w \\ &= \min_{\delta^A\in\mathbb{R}_+} \sum_{A\subset V} F(A)\,\delta^A \qquad \text{s.t.} \qquad \sum_{A\subset V} \delta^A 1_A \geq w,\end{aligned}$$

since if $(v^A)_{A\subset V}$ is a solution so is $(\delta^A 1_A)_{A\subset V}$ with $\delta^A = \|v^A\|_\infty$. We then have

$$F_-(B) = \min_{(\delta^A)} \sum_{A\subset V} F(A)\,\delta^A, \qquad \text{s.t.} \qquad \sum_{A\subset V} \delta^A 1_A \geq 1_B, \quad \delta^A \in [0,1],\ A \subset V, \tag{4}$$

because constraining δ to the unit cube does not change the optimal solution, given that $1_B \leq 1$. But the optimization problem in (4) is exactly the *fractional weighted set-cover problem* (Lovász, 1975), a classical relaxation of the *weighted cover set problem* in Eq. (4), where $\delta \in \{0,1\}$ is replaced by $\delta \in [0,1]$.

Combining Proposition 2 with the fact that $F_-(A)$ is the fractional weighted set-cover, now yields:

Theorem 4. *$\Omega_p(w)$ is the tightest convex relaxation of the function $w \mapsto \|w\|_p\,\widetilde{F}(\mathrm{Supp}(w))^{1/q}$ where $\widetilde{F}(\mathrm{Supp}(w))$ is the* weighted set-cover *of the support of w.*

Proof. We have $F_-(A) \leq \widetilde{F}(A) \leq F(A)$ so that, since F_- is the lower combinatorial envelope of F, it is also the lower combinatorial envelope of $\widetilde{F}$, and therefore $\Omega_p^{F_-} = \Omega_p^{\widetilde{F}} = \Omega_p^F$. □

This proves that the norm Ω_p^F proposed by Jacob et al. (2009) is indeed in a rigorous sense a relaxation of the block-coding or set-cover penalty.

Example 4. *To illustrate the above results consider the block-coding scheme for subsets of $V = \{1,2,3\}$ with blocks consisting only of pairs, i.e., chosen from the collection $\mathcal{A} := \{\{1,2\},\{2,3\},\{1,3\}\}$ with costs all equal to 1. The following table lists the values of F, F_- and $\widetilde{F}$:*

	$\varnothing$	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{2,3\}$	$\{1,3\}$	$\{1,2,3\}$
F	0	∞	∞	∞	1	1	1	∞
$\widetilde{F}$	0	1	1	1	1	1	1	2
F_-	0	1	1	1	1	1	1	3/2

Here, F is equal to its UCE (except that $F_+(\varnothing) = \infty$) and takes therefore non trivial values only on the core set $\mathcal{D}_F = \mathcal{A}_0$. All non-empty sets except V can be covered by exactly one set, which explains the cases where F_- and $\widetilde{F}$ take the value one. $\widetilde{F}(V) = 2$ since V is covered by any pair of blocks and a slight improvement is obtained if fractional covers is allowed since for $\delta_1 = \delta_2 = \delta_3 = \frac{1}{2}$, we have $1_V = \delta_1\,1_{\{2,3\}} + \delta_2\,1_{\{3,1\}} + \delta_3\,1_{\{1,2\}}$ and therefore $F_-(V) = \delta_1 + \delta_2 + \delta_3 = \frac{3}{2}$.

The interpretation of the LCE as the value of a minimum fractional weighted set cover suggests a new interpretation of F_+ (or equivalently of $\mathcal{D}_F$) as defining the smallest set of blocks ($\mathcal{D}_F$) and their costs, that induce a fractional set over problem with the same optimal value.

It is interesting to note that it is Lovász (1975) who introduced the concept of optimal fractional weighted set cover, while we just showed that the value of that cover is precisely F_-, i.e., the combinatorial function which is the restriction on indicators of sets of the function $\Omega_\infty^{F_+} = \Omega_\infty^{F_-} = f \circ |\cdot|$, where, if F_- is submodular, f is its Lovász extension. As an immediate consequence, if F_+ is submodular, $F_+ = F_-$ is equal itself to its associated fractional weighted set cover.

The interpretation of F_- as the value of a minimum fractional weighted cover set problem allows us also to show a result which is dual to the property of LCEs, and which we now present.

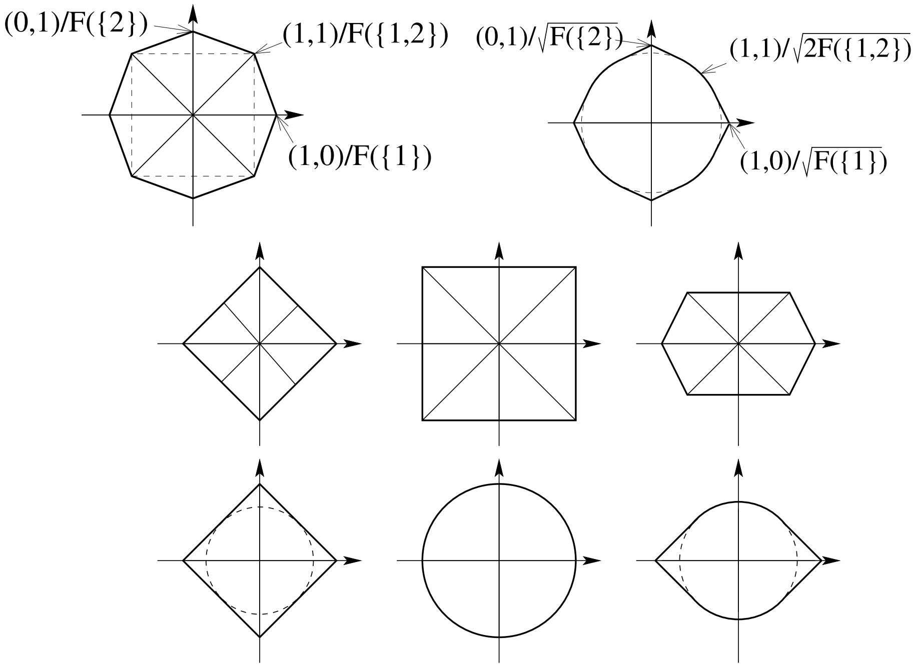
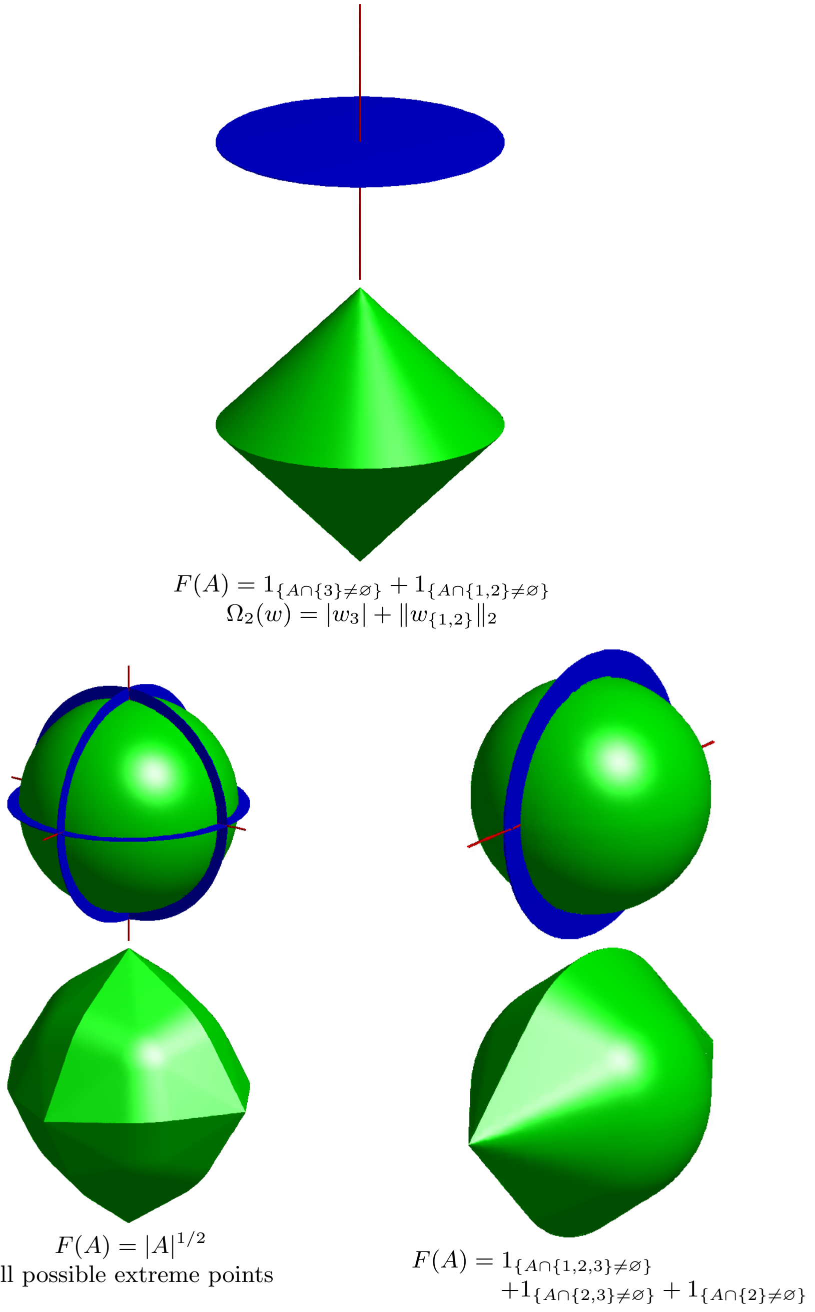


Figure 4: Unit balls in $\mathbb{R}^2$ for four combinatorial functions (actually all submodular) on two variables. Top left and middle row: $p = \infty$; top right and bottom row: $p = 2$. Changing values of F may make some of the extreme points disappear. All norms are hulls of a disk and points along the axes, whose size and position is determined by the values taken by F. On top row: $F(A) = F_-(A) = |A|^{1/2}$ (all possible extreme points); and from left to right on the middle and bottom rows: $F(A) = |A|$ (leading to $\|\cdot\|_1$), $F(A) = F_-(A) = \min\{|A|, 1\}$ (leading to $\|\cdot\|_p$), $F(A) = F_-(A) = \frac{1}{2}1_{\{A\cap\{2\}\neq\varnothing\}} + 1_{\{A\neq\varnothing\}}$.



$F(A) = 1_{\{A\cap\{3\}\neq\varnothing\}} + 1_{\{A\cap\{1,2\}\neq\varnothing\}}$
$\Omega_2(w) = |w_3| + \|w_{\{1,2\}}\|_2$

$F(A) = |A|^{1/2}$
all possible extreme points

$F(A) = 1_{\{A\cap\{1,2,3\}\neq\varnothing\}}$
$+1_{\{A\cap\{2,3\}\neq\varnothing\}} + 1_{\{A\cap\{2\}\neq\varnothing\}}$

Figure 5: Unit balls for structured sparsity-inducing norms, with the corresponding submodular functions and the associated norm, for ℓ_2-relaxations. For each example, we plot on top the sets D_A and on the bottom the convex hull of their union.

3.1 Largest convex positively homogeneous extension

By symmetry with the characterization of the *lower combinatorial envelope* as the smallest combinatorial function that has the same tightest convex and positively homogeneous (p.h.) relaxation as a given combinatorial function F, we can, given a convex positively homogeneous function g, define the combinatorial function $F : A \mapsto g(1_A)$, which by construction, is the combinatorial function which g *extends* (in the sense of Lovász) to $\mathbb{R}^d_+$, and ask if there exists a largest convex and p.h. function g^+ among all such functions. It turns out that this problem is well-posed if the question is restricted to functions that are also coordinate-wise non-decreasing. Perhaps not surprisingly, it is then the case that the largest convex p.h. function extending the same induced combinatorial function is precisely Ω^F_∞, as we show in the next lemma.

Lemma 8 (Largest convex positively homogeneous extension)**.** *Let g be a convex, p.h. and coordinate-wise non-decreasing function defined on $\mathbb{R}^d_+$. Define F as $F : A \mapsto g(1_A)$ and denote by F_- its lower combinatorial envelope. Then $F = F_-$ and $\forall w \in \mathbb{R}^d$, $g(|w|) \leq \Omega^F_\infty(w)$.*

Proof. From Equation (4), we know that F_- can be written as the value of a minimal weighted fractional set-cover. But if $1_B \leq \sum_{A \subset V} \delta^A 1_A$, we have

$$\sum_{A \subset V} \delta^A g(1_A) \geq g\big(\textstyle\sum_{A \subset V} \delta^A\big) \geq g(1_B),$$

where the first inequality results from the convexity and homogeneity of g, and the second from the assumption that it is coordinate-wise non-decreasing. As a consequence, injecting the above inequality in (4), we have $F_-(B) \geq F(B)$. But since, we always have $F_- \leq F$, this proves the equality.

For the second statement, using the coordinate-wise monotonicity of g and its homogeneity, we have $g(|w|) \leq \|w\|_\infty g(1_{\mathrm{Supp}(w)}) = \|w\|_\infty F(\mathrm{Supp}(w))$. Then, taking the convex envelope of functions on both sides of the inequality we get $g(|\cdot|)^{**} \leq \big(\|\cdot\|_\infty F(\mathrm{Supp}(\cdot))\big)^{**} = \Omega^F_\infty$, where $(\cdot)^*$ denotes the Fenchel-Legendre transform. □

4 Examples

In this section, we present the main examples of existing and new norms that fall into our framework. For more advanced examples, see Section 7.

4.1 Overlap count functions, their relaxations and the ℓ_1/ℓ_p-norms

A natural family of set functions to consider are the functions that, given a collection of sets $\mathcal{G} \subset 2^V$ are defined as the (weighted with positive weights d_B, $B \in \mathcal{G}$) number of these sets that are intersected by the support:

$$F_\cap(A) = \sum_{B \in \mathcal{G}} d_B 1_{\{A \cap B \neq \varnothing\}}. \tag{5}$$

Since $A \mapsto 1_{\{A \cap B \neq \varnothing\}}$ is clearly submodular and since submodular functions form a positive cone, all these functions are submodular, which implies that $\Omega^{F_\cap}_p$ is a tight relaxation of $F_\cap$. We call them overlap count functions.

Overlap count functions *vs.* set-covers. As mentioned in Section 2.1, if $\mathcal{G}$ is a partition, the norm $\Omega_p^{F_\cap}$ is the ℓ_1/ℓ_p-norm; in this special case, $F_\cap$ is actually the value of the minimal (integer-valued) weighted set-cover associated with the sets in $\mathcal{G}$ and the weights d_G.

However, it should be noted that, in general, the values of functions of the form $F_\cap$ are quite different from values of a minimal weighted set-covers. It has rather the flavor of some sort of "maximal weighted set-cover" in the sense that any set that has a non-empty intersection in the support would be included in the cover.

ℓ_p-relaxations of $F_\cap$ *vs.* ℓ_1/ℓ_p-norms. In the case where $p = \infty$, Bach (2010) showed that even when groups overlap we have $\Omega_\infty^{F_\cap}(w) = \sum_{B\in\mathcal{G}} d_B \|w_G\|_\infty$, since the Lovász extension of a sum of submodular functions is just the sum of the Lovász extensions of the terms in the sum, and given that on the positive orthant the Lovász extension of $A \mapsto d_B\, 1_{\{A\cap B \neq \varnothing\}}$ (which is a submodular function) coincides with $w \mapsto d_B \|w_B\|_\infty$.

The situation is more subtle when $p < \infty$: in that case, and perhaps surprisingly, $\Omega_p^{F_\cap}$ is not the *weighted ℓ_1/ℓ_p norm with overlap* (Jenatton et al., 2011a), also referred to as the *overlapping group Lasso* (which should clearly be distinguished from the *latent group Lasso*) and which is the norm defined as $\widetilde{\Omega}_p : w \mapsto \sum_{B\in\mathcal{G}} d'_B \|w_B\|_p$. The differences between the norm $\Omega_p^{F_\cap}$ and $\widetilde{\Omega}_p$ is illustrated in Example 5, Table 1 and Figure 6. The norm $\Omega_p^{F_\cap}$ does not have a simple closed form in general. In terms of sparsity patterns induced however, $\Omega_p^{F_\cap}$ behaves like $\Omega_\infty^{F_\cap}$, and as a result the sparsity patterns allowed by $\Omega_p^{F_\cap}$ are the same as those allowed by the corresponding *weighted ℓ_1/ℓ_p norm with overlap*. However, the definition of $\Omega_p^{F_\cap}$ as a convex relaxation leads to fewer overpenalization artefacts than the ℓ_1/ℓ_p-norm with overlap (see Section 9).

ℓ_p-relaxation of $F_\cap$ *vs.* latent group Lasso based on $\mathcal{G}$. It should be clear as well that $\Omega_p^{F_\cap}$ is not itself the *latent group Lasso* associated with the collection $\mathcal{G}$ and the weights d_G in the sense of Jacob et al. (2009). Indeed, the latter corresponds to the function $F_\cup$ such that $F_\cup(A) = d_A$ for $A \in \mathcal{G}$ and $F_\cup(A) = \infty$ otherwise, and whose LCE is the value of the minimal fractional weighted set cover by elements in $\mathcal{G}$ and with the weights $(d_G)_{G\in\mathcal{G}}$. Clearly, $F_\cup$ is in general strictly smaller than $F_\cap$ and since the relaxation of the latter is tight, it cannot be equal to the relaxation of the former, if the combinatorial functions are themselves different. Obviously, the function $\Omega_p^{F_\cap}$ is still (see Table 1) another latent group Lasso corresponding to a fractional weighted set cover and involving a larger number of sets that the ones in $\mathcal{G}$ (possibly all of 2^V). This last statement leads us to what might appear to be a paradox, which we discuss next.

Example 5. *To illustrate the difference between the norms $\Omega^{F_\cup}$, $\Omega^{F_\cap}$ and the weighted ℓ_1/ℓ_p-norm associated with a given set of groups $\mathcal{G}$ with associated weights $(d_G)_{G\in\mathcal{G}}$, consider the case where $\mathcal{G} = \{\{1,2\},\{2,3\}\}$ and all weights equal 1. By definition $F_\cap(A) = 1_{\{A\cap\{1,2\}\neq\varnothing\}} + 1_{\{A\cap\{2,3\}\neq\varnothing\}}$, $F_\cup(A) = F_{\cup,+}(A) = 1$ for $A \in \mathcal{G}$ and ∞ otherwise, and $F_{\cup,-}(A) = \min_{\delta,\delta'} \big\{\delta + \delta' \mid 1_A \leq \delta\, 1_{\{1,2\}} + \delta'\, 1_{\{2,3\}}\big\}$. We have the table below:*

	$\varnothing$	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{2,3\}$	$\{1,3\}$	$\{1,2,3\}$
$F_\cup$	∞	∞	∞	∞	1	1	∞	∞
$F_{\cup,-}$	0	1	1	1	1	1	2	2
$F_\cap$	0	1	2	1	2	2	2	2
$F_{\cap,+}$	∞	1	∞	1	∞	∞	∞	2

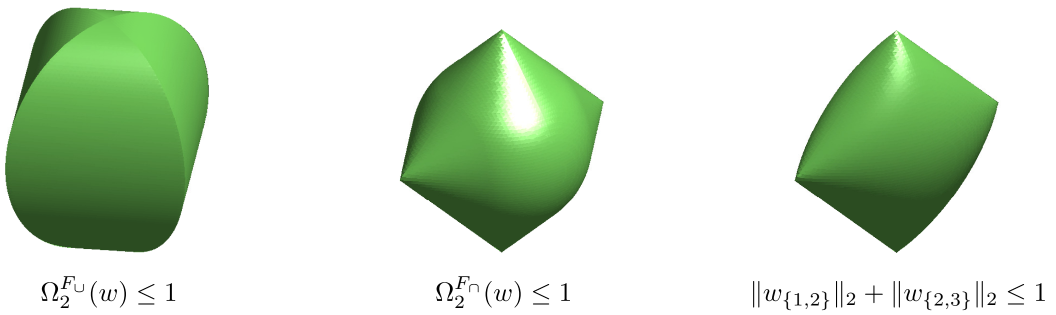
The two set functions $F_{\cup,-}$ and $F_\cap$ are clearly different. In fact, we have $F_\cup(A) = \max(|A\cap\{1,3\}|, |A\cap\{2\}|)$ and $F_\cap$ is the value of the set cover associated with $\mathcal{G}'= \{\{1\},\{3\},\{1,2,3\}\}$ with weights $(1,1,2)$. The corresponding unit balls are represented on Figure 6 together with the unit ball

	Latent group Lasso Jacob et al. (2009)	Overlap count Lasso (new)	ℓ_1/ℓ_p with overlap Jenatton et al. (2011a)
F	$F_\cup(A)$	$F_\cap(A) = f_A$	-
Def.	$\min_{\delta\in[0,1]^{\lvert\mathcal{G}\rvert}} \Big\{ \sum_{B\in\mathcal{G}} d_B\,\delta^B \mid \sum_{B\in\mathcal{G}} 1_B\,\delta^B \geq 1_A \Big\}$	$f_A := \sum_{B\in\mathcal{G}} d_B\,1_{\{A\cap B\neq\varnothing\}}$	-
$\Omega_p(w)$	$\min_{v\in\mathcal{V}(w,\mathcal{G})} \sum_{B\in\mathcal{G}} d_B^{1/q}\lVert v^B\rVert_p$	$\min_{v\in\mathcal{V}(w)} \sum_{B\subset V} f_B^{1/q}\lVert v^B\rVert_p$	$\sum_{B\in\mathcal{G}} d_B^{1/q}\,\lVert w_B\rVert_p$
$\Omega_p^*(s)$	$\max_{B\in\mathcal{G}} d_B^{-1/q}\,\lVert s_B\rVert_q$	$\max_{B\subset V} f_B^{-1/q}\,\lVert s_B\rVert_q$	$\min_{z\in\mathcal{V}(s,\mathcal{G})} \max_{B\in\mathcal{G}} d_B^{-1/q}\lVert z^B\rVert_q$

Table 1: Three norms naturally associated with a set of blocks $B \in \mathcal{G}$ with associated weights d_B either via minimal-set cover, or "overlap count". For the two first norms that are tight relaxations of a combinatorial function the latter is given in the first and second rows. The notation used is $\mathcal{V}(w,\mathcal{G}) = \big\{v \in \mathcal{V} \mid w = \sum_{B\in\mathcal{G}} v^B\big\}$ and $\mathcal{V}(w) = \mathcal{V}(w, 2^V)$, with $\mathcal{V}$ defined in Section 3. When $p = \infty$, the norms of the two last columns are equal, with the correspondence between d_B and f_B given by the definition of $f_A := F_\cap(A)$. See appendix B for a proof of the form of the dual norm of the ℓ_1/ℓ_p-norm with overlap.

of the ℓ_1/ℓ_2-norm with overlap. As can be seen on the figure, the non-trivial supports induced by $\Omega_2^{F_\cup}$ are $\{1,2\},\{2,3\}$, while the nontrivial supports induced by the other norms are $\{3\} = \{1,2\}^c$ and $\{1\} = \{2,3\}^c$.

Supports *stable by intersection* vs. *formed as unions*. Jenatton et al. (2011a) have shown that the family of norms they considered induces possible supports which form a family that is *stable by intersection*, in the sense that the intersection of any two possible support is also a possible support. But since as mentioned above they have the same support as the norms $\Omega_p^{F_\cap}$, for $1 < p \leq \infty$, which are latent group Lasso norms, and since Jacob et al. (2009) have discussed the fact that the supports induced by any norm Ω_p are formed by *unions* of elements of the core set $\mathcal{A}$, this might appear paradoxical that the allowed support can be described at the same time as intersections and as unions. There is in fact no contradiction because in general the set of supports that are induced



$\Omega_2^{F_\cup}(w) \leq 1$

$\Omega_2^{F_\cap}(w) \leq 1$

$\|w_{\{1,2\}}\|_2 + \|w_{\{2,3\}}\|_2 \leq 1$

Figure 6: Units balls for $\Omega_2^{F_\cup}$, $\Omega_2^{F_\cap}$ and ℓ_1/ℓ_2 with overlap for the groups $\mathcal{G} = \{\{1,2\}\{2,3\}\}$ in $\mathbb{R}^3$.

by the latent group Lasso are in fact not necessarily *stable by union*: for some set A obtained exactly as a union it is possible to have another set B with $A \subsetneq B$ and $F_(A) = F_(B)$.

Three different norms. To conclude, we must, given a set of groups $\mathcal{G}$ and a collection of weights $(d_G)_{G\in\mathcal{G}}$, distinguish three norms that can be defined from it, the weighted ℓ_1/ℓ_p-norm with overlap, the norm $\Omega_p^{F_\cap}$ obtained as the ℓ_p relaxation of the submodular penalty $F_\cap$, and finally, the norm $\Omega_p^{F_\cup}$ obtained as the relaxation of the set-cover or block-coding penalty with the weights d_G. For sets of groups that form a partition, they are all equal, but not in general.

Some of the advantages of using a tight relaxation still need to be assessed empirically and theoretically, but the possibility of using ℓ_p-relaxation for $p < \infty$ removes the artifacts that were specific to the ℓ_∞ case.

4.2 Submodular range function

The weighted ℓ_1/ℓ_p-norm with overlap has been, among others, used to induce interval patterns on chains and rectangular or convex patterns on grids (Jenatton et al., 2011a). In particular, one of the norms considered by Jenatton et al. (2011a) provides a nice example of an overlap count function, which it is worth presenting.

Example 6 (Modified range function)**.** *A shown in Example 2 in Section 2.2, the natural range function on a sequence leads to a trivial LCE. Consider now the penalty with the form of Eq. (5) with $\mathcal{G}$ the set of groups defined as*

$$\mathcal{G} = \big\{[\![1,k]\!] \mid 1 \leq k \leq p\big\} \cup \big\{[\![k,p]\!] \mid 1 \leq k \leq p\big\}.$$

A simple calculation shows that $F_\cap(\varnothing) = 0$ and that for $A \neq \varnothing$, $F_\cap(A) = d - 1 + range(A)$. This function is submodular as a sum of submodular functions, and thus equal to it lower combinatorial envelope, which implies that the relaxation retains the structural a prior encoded by the combinatorial function itself. We will consider the ℓ_2 relaxation of this submodular function in the experiments (see Section 9) and compare it with the ℓ_1/ℓ_2-norm with overlap of Jenatton et al. (2011a).

4.3 Exclusive Lasso

The exclusive Lasso is a formulation proposed by Zhou et al. (2010) which considers the case where a partition $\mathcal{G} = \{G_1, \ldots, G_k\}$ of V is given and the sparsity imposed is that w should have at most one non-zero coefficient in each group G_j. The regularizer proposed by Zhou et al. (2010) is the ℓ_p/ℓ_1-norm defined[5] by $\|w\|_{\ell_p/\ell_1} = (\sum_{G\in\mathcal{G}} \|w_G\|_1^p)^{1/p}$. Is this the tightest relaxation?

A natural combinatorial function corresponding to the desired constraint is the function $F(A)$ defined by $F(A) = 1$ if $\max_{G\in\mathcal{G}} |A \cap G| = 1$ and $F(A) = \infty$ otherwise.

[5] The Exclusive Lasso norm which is ℓ_p/ℓ_1 should not be confused with the group Lasso norm which is ℓ_1/ℓ_p.

To characterize the corresponding Ω_p we can compute explicitly its dual norm Ω_p^*:

$$\begin{aligned}
\big(\Omega_p^*(w)\big)^q &= \max_{A\subset V} \frac{\|s_A\|_q^q}{F(A)}\\
&= \max_{A\subset V} \|s_A\|_q^q \quad \text{s.t.} \quad |A\cap G|\leq 1,\, G\in\mathcal{G}\\
&= \max_{i_j\in G_j,\, 1\leq j\leq k} \sum_{j=1}^k |s_{i_j}|^q \;=\; \sum_{j=1}^k \max_{i\in G_j} |s_i|^q \;=\; \sum_{j=1}^k \|s_{G_j}\|_\infty^q,
\end{aligned}$$

which shows that Ω_p^* is the ℓ_q/ℓ_∞-norm or equivalently that Ω_p is the ℓ_p/ℓ_1-norm and provides a theoretical justification for the choice of this norm: it is indeed the tightest relaxation! It is interesting to compute the lower combinatorial extension of F which is $F_-(A) = \Omega_\infty^F(1_A) = \|1_A\|_{\ell_\infty/\ell_1} = \max_{G\in\mathcal{G}} |A\cap G|$. This last function is also a natural combinatorial function to consider; by the previous result F_- has the same convex relaxation as F, but it would be however less obvious to show directly that $\Omega_p^{F_-}$ is the ℓ_p/ℓ_1-norm (see appendix C for a direct proof which uses Lemma 8). It is easy to check that F_- is not submodular.

5 A variational forms of the norm

Several results on Ω_p rely on the fact that it can be related variationally to Ω_∞.

Lemma 9 (variational formulation)**.** Ω_p *admits the two following variational formulations:*

$$\begin{aligned}
\Omega_p(w) &= \max_{\kappa\in\mathbb{R}_+^d} \sum_{i\in V} \kappa_i^{1/q}|w_i| \quad \textit{s.t.} \quad \forall A\subset V,\ \kappa(A)\leq F(A) &(6)\\
&= \min_{\eta\in\mathbb{R}_+^d} \sum_{i\in V} \frac{1}{p}\frac{|w_i|^p}{\eta_i^{p-1}} + \frac{1}{q}\Omega_\infty(\eta). &(7)
\end{aligned}$$

Proof. Using Fenchel duality, we have:

$$\begin{aligned}
\Omega_p(w) &= \max_{s\in\mathbb{R}^d} s^\top w \quad \text{s.t.} \quad \Omega_p^*(w)\leq 1\\
&= \max_{s\in\mathbb{R}^d} s^\top w \quad \text{s.t.} \quad \forall A\subset V,\ \|s_A\|_q^q\leq F(A) \text{ by definition of } \Omega_p^*,\\
&= \max_{s\in\mathbb{R}_+^d} s^\top |w| \quad \text{s.t.} \quad \forall A\subset V,\ s^q(A)\leq F(A)\\
&= \max_{\kappa\in\mathbb{R}_+^d} \sum_{i\in V} \kappa_i^{1/q}|w_i| \quad \text{s.t.} \quad \forall A\subset V,\ \kappa(A)\leq F(A) \text{ by a change of variable.}
\end{aligned}$$

But it is easy to verify that $\kappa_i^{1/q}|w_i| = \min_{\eta_i\in\mathbb{R}_+} \frac{1}{p}\frac{|w_i|^p}{\eta_i^{p-1}} + \frac{1}{q}\eta_i\kappa_i$ with the minimum attained for $\eta_i = \frac{|w_i|}{\kappa_i^{1/p}}$. We therefore get:

$$\begin{aligned}
\Omega_p(w) &= \max_{\kappa\in\mathbb{R}_+^d} \min_{\eta\in\mathbb{R}_+^d} \sum_{i\in V} \frac{1}{p}\frac{|w_i|^p}{\eta_i^{p-1}} + \frac{1}{q}\eta^\top\kappa \quad \text{s.t.} \quad \forall A\subset V,\ \kappa(A)\leq F(A)\\
&= \min_{\eta\in\mathbb{R}_+^d} \max_{\kappa\in\mathbb{R}_+^d} \sum_{i\in V} \frac{1}{p}\frac{|w_i|^p}{\eta_i^{p-1}} + \frac{1}{q}\eta^\top\kappa \quad \text{s.t.} \quad \forall A\subset V,\ \kappa(A)\leq F(A)\\
&= \min_{\eta\in\mathbb{R}_+^d} \sum_{i\in V} \frac{1}{p}\frac{|w_i|^p}{\eta_i^{p-1}} + \frac{1}{q}\Omega_\infty(\eta),
\end{aligned}$$

where we could exchange minimization and maximization since the function is convex-concave in η and κ, and where we eliminated formally κ by introducing the value of the dual norm $\Omega_\infty(\eta) = \max_{\kappa \in \mathcal{P}_F} \kappa^\top \eta$. □

Since Ω_∞ is convex, the last formulation is actually jointly convex in (w, η) since $(x, z) \mapsto \frac{1}{p} \frac{\|x\|_p^p}{z^{p-1}} + \frac{1}{q} z$ is convex, as $(x, z) \mapsto \frac{\|x\|_p^p}{z^{p-1}}$ is the perspective function of $x \mapsto \|x\|_p^p$ (see Boyd and Vandenberghe, 2004, p. 89).

It should be noted that the norms Ω_p therefore belong to the broad family of H-norms as defined[6] in Bach et al. (2012, Sec. 1.4.2.) and studied by Micchelli et al. (2013).

The above result is particularly interesting if F is submodular since Ω_∞ is then equal to the Lovász extension of F on the positive orthant (Bach, 2010). In this case in particular, it is possible, as we will see in the next section to propose efficient algorithms to compute Ω_p and Ω_p^*, the associated proximal operators, and algorithms to solve learning problems regularized with Ω_p, thanks to the above variational form.

Using the variational form to compute the proximal operator of the norm. Consider the proximal problem $\min_{w \in \mathbb{R}^d} \frac{1}{2}\|w - u\|_2^2 + \lambda \Omega_2(w)$. Expressing the norm, with the variational form (16) and minimizing with respect to w shows that the solution satisfies $w_i^\star = \big(1 + \frac{\lambda}{\eta_i^\star}\big)^{-1} u_i$, with $\eta^\star$ the solution of the optimization problem in which w has been eliminated and which after some algebra takes the form

$$\min_{\eta \in \mathbb{R}_+^d} \sum_{i \in V} \frac{u_i^2}{\eta_i + \lambda} + \Omega_\infty(\eta). \tag{8}$$

For submodular functions, these variational forms are also the basis for the *local decomposability* result of Section 8.1 which is key to establish support recovery in Section 8.2.

6 The case of submodular penalties

In this section, we focus on the case where the combinatorial function F is submodular.

Specifically, we will consider a function F defined on the power set 2^V of $V = \{1, \dots, d\}$, which is *nondecreasing* and *submodular*, meaning that it satisfies respectively

$$\begin{aligned} &\forall A, B \subset V, && A \subset B \Rightarrow F(A) \leqslant F(B), \\ &\forall A, B \subset V, && F(A) + F(B) \geqslant F(A \cap B) + F(A \cup B). \end{aligned}$$

Moreover, we assume that $F(\varnothing) = 0$. These set-functions are often referred to as *polymatroid set-functions* (Edmonds, 2003; Fujishige, 2005). Also, without loss of generality, we assume that F is strictly positive on singletons, i.e., for all $k \in V$, $F(\{k\}) > 0$. Indeed, if $F(\{k\}) = 0$, then by submodularity and monotonicity, if $A \ni k$, $F(A) = F(A \backslash \{k\})$ and thus we can simply consider $V \backslash \{k\}$ instead of V.

Classical examples are the cardinality function and, given a partition of V into $G_1 \cup \dots \cup G_k = V$, the set-function $A \mapsto F(A)$ which is equal to the number of groups $G_1, \dots, G_k$ with non empty intersection with A, which, as mentioned in Section 2.1 leads to the grouped ℓ_1/ℓ_p-norm.

[6] Note that H-norms are in these references defined for $p = 2$ and that the variational formulation proposed here generalizes this to other values of $p \in (1, \infty)$.

With a slightly different perspective than the approach of this paper, Bach (2010) studied the special case of the norm Ω_p^F when $p = \infty$ and F is submodular. As mentioned previously, he showed that in that case the norm Ω_∞^F is the Lovász extension of the submodular function F, which is a well studied mathematical object.

Before presenting results on ℓ_p-relaxations of submodular penalties, we review a certain number of relevant properties and concepts from submodular analysis. For more details, see, e.g., Fujishige (2005), and, for a review with proofs derived from classical convex analysis, see Bach (2013).

6.1 Review of submodular function theory

Lovász extension. Given any set-function F, one can define its *Lovász extension* $f : \mathbb{R}_+^d \to \mathbb{R}$, as follows: given $w \in \mathbb{R}_+^d$, we can order the components of w in decreasing order $w_{j_1} \geqslant \cdots \geqslant w_{j_p} \geqslant 0$, the value $f(w)$ is then defined as

$$\begin{aligned} f(w) &= \sum_{k=1}^{p-1} (x_{j_k} - x_{j_{k+1}}) F(\{j_1, \ldots, j_k\}) + x_{j_p} F(\{j_1, \ldots, j_p\}) && (9) \\ &= \sum_{k=1}^{p} w_{j_k} [F(\{j_1, \ldots, j_k\}) - F(\{j_1, \ldots, j_{k-1}\})]. && (10) \end{aligned}$$

We will refer to this formula as the Choquet integral form of the function. The Lovász extension f is always piecewise-linear, and when F is submodular, it is also convex—see, e.g., Bach (2013); Fujishige (2005). Moreover, for all $\delta \in \{0,1\}^d$, $f(\delta) = F(\mathrm{Supp}(\delta))$ and f is in that sense an extension of F from vectors in $\{0,1\}^d$ (which can be identified with indicator vectors of sets) to all vectors in $\mathbb{R}_+^d$. Moreover, it turns out that minimizing F over subsets, i.e., minimizing f over $\{0,1\}^d$ is equivalent to minimizing f over $[0,1]^d$ (Edmonds, 2003).

Canonical polyhedron and norm. For consistency with notations, we denote by $\mathcal{P}_F$ the *canonical polyhedron* which we define as the set of $s \in \mathbb{R}_+^d$ such that for all $A \subset V$, $s(A) \leqslant F(A)$, i.e., $\mathcal{P}_F = \{s \in \mathbb{R}_+^d,\ \forall A \subset V,\ s(A) \leqslant F(A)\}$, where we use the notation $s(A) = \sum_{k \in A} s_k$. The *submodular polyhedron* $\overline{\mathcal{P}}_F = \{s \in \mathbb{R}^d,\ \forall A \subset V,\ s(A) \leqslant F(A)\}$, is a classical polyhedron considered in submodular theory (Fujishige, 2005). Our canonical polyhedron is thus $\mathcal{P}_F = \overline{\mathcal{P}}_F \cap \mathbb{R}_+^d$, which is also called the positive submodular polyhedron. One important result in submodular analysis is that, if F is a nondecreasing submodular function, then we have a representation of f as a maximum of linear functions (Bach, 2013; Fujishige, 2005). In particular, for all $w \in \mathbb{R}_+^d$,

$$f(w) = \max_{s \in \mathcal{P}_F} \ w^\top s. \tag{11}$$

We recognize here that the Lovász extension of a submodular function F is directly related to the norm Ω_∞^F in that $f(|w|) = \Omega_\infty^F(w)$ for all $w \in \mathbb{R}^d$. A striking consequence of submodularity is that the extension f can be computed in closed form (via the Choquet integral).

Greedy algorithm. Instead of solving a linear program with $d + 2^d$ constraints, a solution s to (11) may be obtained by the following algorithm (a.k.a. "greedy algorithm"): order the components of w in decreasing order $w_{j_1} \geqslant \cdots \geqslant w_{j_d}$, and then take for all $k \in V$, $s_{j_k} = F(\{j_1, \ldots, j_k\}) - F(\{j_1, \ldots, j_{k-1}\})$. Moreover, if $w \in \mathbb{R}^d$ has some negative components, then, to obtain a solution to $\max_{s \in \mathcal{P}} \ w^\top s$, we can take s_{j_k} to be simply equal to zero for all k such that w_{j_k} is negative (Edmonds, 2003).

Contraction and restriction of a submodular function. Given a submodular function F and a set J, two related functions, which are submodular as well, will play a crucial role both algorithmically and for the theoretical analysis of the norm. Those are the *restriction* of F to a set J, denoted F_J, and the *contraction* of F on J, denoted F^J. They are defined respectively as

$$F_J : A \mapsto F(A \cap J) \qquad \text{and} \qquad F^J : A \mapsto F(A \cup J) - F(A).$$

Both F_J and F^J are submodular if F is.

In particular the norms $\Omega_p^{F_J} : \mathbb{R}^J \to \mathbb{R}_+$ and $\Omega_p^{F^J} : \mathbb{R}^{J^c} \to \mathbb{R}_+$ associated respectively with F_J and F^J will be useful to "decompose" Ω_p^F in the sequel. We will denote these two norms by Ω_J and Ω^J for short. Note that their domains are not $\mathbb{R}^d$ but the vectors with support in J and J^c respectively.

Stable sets. Another concept which will be key in this section is that of *stable set.* A set A is said *stable* if it cannot be augmented without increasing F, i.e., if for all sets $B \supset A$, $B \neq A \Rightarrow F(B) > F(A)$. If F is strictly increasing (such as for the cardinality), then all sets are stable. The set of stable sets is closed by intersection. In the case $p = \infty$, Bach (2013) has shown that these stable sets were the only allowed sparsity patterns.

Separable sets. A set A is separable if we can find a partition of A into $A = B_1 \cup \cdots \cup B_k$ such that $F(A) = F(B_1) + \cdots + F(B_k)$. A set A is inseparable if it is not separable. As shown by Edmonds (2003), the submodular polytope $\mathcal{P}_F$ has full dimension d as soon as F is strictly positive on all singletons, and its faces are exactly the sets $\{s(A) = F(A)\}$ for stable *and* inseparable sets A. With the terminology that we introduced in Section 2.3, this means that the core set $\mathcal{D}_F$ of F is here exactly the set of its stable and inseparable sets. The core set will clearly play a role when deriving concentration inequalities in Section 8.2. For the cardinality function, stable and inseparable sets are singletons.

6.2 Submodular function and lower combinatorial envelope

A few comments are in order to confront submodularity to the previously introduced notions associated with cover-sets, and lower and upper combinatorial envelopes. We have showed that $F_-(A) = \Omega_\infty(1_A)$. But for a submodular function $\Omega_\infty(1_A) = f(1_A) = F(A)$ since f is the Lovász extension of F. This shows that a submodular function is its own lower combinatorial envelope. However the converse is not true: a lower combinatorial envelope is not submodular in general. E.g., in Example 4, we have $F_-(\{1,2\}) + F_-(\{2,3\}) \not\geq F_-(\{2\}) + F_-(\{1,2,3\})$.

The core set of a submodular function is the set $\mathcal{D}_F$ of its stable and inseparable sets, which implies that F can be retrieved as the value of the minimal fractional weighted set cover the sets $A \in \mathcal{D}_F$ with weights $F(A)$.

6.3 Optimization algorithms for the submodular case

In the context of sparsity and structured sparsity, *proximal methods* have emerged as methods of choice to design efficient algorithm to minimize objectives of the form $f(w) + \lambda\Omega_p(w)$, where f is a smooth function with Lipschitz gradients and Ω_p is a proper convex function (Bach et al., 2012). In a nutshell, their principle is to linearize f at each iteration and to solve the problem

$$\min_{w \in \mathbb{R}^d} \nabla f(w_t)^\top (w - w_t) + \frac{L}{2}\|w - w_t\|^2 + \lambda\Omega_p(w),$$

for some constant L. Setting $\lambda' = \frac{\lambda}{L}$ This problem is a special case of the so-called *proximal problem*:

$$\min_{w\in\mathbb{R}^d} \frac{1}{2}\|w-z\|_2^2 + \lambda'\Omega_p(w). \tag{12}$$

The function mapping z to the solution of the above problem is called *proximal operator*. If this proximal operator can be computed efficiently, then proximal algorithm provide good rates of convergence especially for strongly convex objectives. We show in this section that the structure of submodular functions can be leveraged to compute efficiently Ω_p, Ω_p^* and the proximal operator.

Computation of Ω_p and Ω_p^*. A simple approach to compute the norm is to maximize in κ in the variational formulation (8). This can be done efficiently using for example a *conditional gradient* algorithm, given that maximizing a linear form over the submodular polyhedron is done easily with the *greedy algorithm* (see Section 6.1).

We will propose another algorithm to compute the norm based on the so-called *decomposition algorithm*, which is a classical algorithm of the submodular analysis literature that makes it possible to minimize a separable convex function over the submodular polytope efficiently (see, e.g., Bach, 2013, Section 8.6).

As we show in the following proposition, we can also compute $\Omega_p^*(s)$ using Algorithm 1.

Algorithm 1 Dual norm computation algorithm

```
1: Initialization: $\lambda_0 = 0,\ t = 0$
2: while $\varphi(\lambda_t) \neq 0$ do
3:     $\mathcal{S}_t \leftarrow \text{Argmax}_{A\subset V}\left[\|s_A\|_q^q - \lambda_t F(A)\right]$
4:     $A_t \leftarrow \text{argmin}_{A\in\mathcal{S}_t} F(A)$
5:     $\lambda_{t+1} \leftarrow \frac{\|s_A\|_q^q}{F(A)}$
6:     $t \leftarrow t+1$
7: end while
8: return $\lambda_t$
```

Proposition 4. *The sequence $(\lambda_t)_t$ generated by Algorithm 1 is monotonically increasing and converges in a finite number of iterations to Ω_p^*.*

Proof. As the maximum of a finite number non-increasing linear functions of a scalar argument, the function $\varphi : \lambda \mapsto \max_{A\subset V}\left[\|s_A\|_q^q - \lambda F(A)\right]$ is a non-increasing, continuous, piecewise linear convex function. It is also non negative because $\|s_\varnothing\|_p^p = 0 = F(\varnothing)$. It is immediate to check that $\lambda^* := \min\{\lambda \mid \varphi(\lambda) = 0\} = \max_{\varnothing\subset A\subset V} \frac{\|s_A\|_q^q}{F(A)}$. At each iteration, if $\varphi(\lambda_t) \neq 0$, we must have $\lambda_{t+1} > \lambda_t$, because the function $\lambda \mapsto \|s_{A_t}\|_q^q - \lambda F(A_t)$ is strictly positive for $\lambda = \lambda_t$ and equal to 0 for $\lambda = \lambda_{t+1}$. Moreover by construction, the sets A_t are all distinct, as long as $\varphi(\lambda_t) \neq 0$. As a consequence we must reach $\varphi(\lambda_T) = 0$ after a finite number of iterations T. At the end of the algorithm, $\varphi(\lambda_T) = 0$ entails that $\forall A \subset V$, $\|s_A\|_p^p \leq \lambda_T F(A)$, which entails that for all $A \neq \varnothing$, $F(A)^{-1}\|s_A\|_p^p \leq \lambda_T = F(A_{T-1})^{-1}\|s_{A_{T-1}}\|_p^p$. This shows that $\lambda_T = \Omega^*(s)$. This concludes the proof. The choice of taking the maximizer with smallest value of $F(A)$ on line 4 of the algorithm is not key to ensure convergence of the algorithm, but aims at (a) computing the right-derivative which maximizes the step size in λ, and simultaneously (b) obtaining a maximizing set as sparse as possible. □

Note that this algorithm is closely related to the algorithm of Dinkelbach (1967) to maximize a ratio of functions, and in fact applies to all functions F; but step 3 in the algorithm requires to minimize a function $(A \mapsto \lambda F(A) - \|s_A\|_p^p)$ which can be done in polynomial time for submodular functions. Moreover, for submodular functions, the number of iterations may be bounded by d, because the algorithm may be reinterpreted as the divide-and-conquer algorithm for a certain separable function (see Bach, 2013, p. 160); for the general case, it may only be bounded in general by 2^d.

Computation of the proximal operator. Using Eq. (8), we can reformulate problem (12) as

$$\begin{aligned}
\min_{w \in \mathbb{R}^d} \frac{1}{2}\|w - z\|_2^2 + \lambda \Omega_p(w) &= \min_{w \in \mathbb{R}^d} \max_{\kappa \in \mathbb{R}_+^d \cap \mathcal{P}} \frac{1}{2}\|w - z\|_2^2 + \lambda \sum_{i \in V} \kappa_i^{1/q} |w_i| \\
&= \max_{\kappa \in \mathbb{R}_+^d \cap \mathcal{P}} \sum_{i \in V} \min_{w_i \in \mathbb{R}} \left\{ \frac{1}{2}(w_i - z_i)^2 + \lambda \kappa_i^{1/q} |w_i| \right\} \\
&= \max_{\kappa \in \mathbb{R}_+^d \cap \mathcal{P}} \sum_{i \in V} \psi_i(\kappa_i),
\end{aligned}$$

with $\psi_i : \kappa_i \mapsto \min_{w_i \in \mathbb{R}} \left\{ \frac{1}{2}(w_i - z_i)^2 + \lambda \kappa_i^{1/q} |w_i| \right\}$.

Thus, solving the proximal problem is equivalent to maximizing a concave separable function $\sum_i \psi_i(\kappa_i)$ over the submodular polytope. For a submodular function, this can be solved as well using the divide-and-conquer algorithm. More precisely, this algorithm also called *decomposition algorithm* involves a sequence of submodular function minimizations (see Bach, 2013; Groenevelt, 1991). This yields an algorithm which finds a decomposition of the norm and applies recursively the proximal algorithm to the two parts of the decomposition corresponding respectively to a *restriction* and a *contraction* of the submodular function. We explicit this algorithm as Algorithm 2 for the case $p = 2$.

Algorithm 2 Computation $x = \mathrm{Prox}_{\lambda \Omega_2^F}(z)$

```
Require: $z \in \mathbb{R}^d$, $\lambda > 0$
Let $A = \{j \mid z_j \neq 0\}$
if $A \neq V$ then
    Set $x_A = \mathrm{Prox}_{\lambda \Omega_2^{F_A}}(z_A)$
    Set $x_{A^c} = 0$
    return $x$ by concatenating $x_A$ and $x_{A^c}$
end if
Let $t \in \mathbb{R}^d$ with $\kappa_i = \frac{z_i^2}{\|z\|_2^2} F(V)$
Find $A$ minimizing the submodular function $F - t$
if $A = V$ then
    return $x = \left(\|z\|_2 - \lambda\sqrt{F(V)}\right)_+ \frac{z}{\|z\|_2}$
end if
Let $x_A = \mathrm{Prox}_{\lambda \Omega_2^{F_A}}(z_A)$
Let $x_{A^c} = \mathrm{Prox}_{\lambda \Omega_2^{F^A}}(z_{A^c})$
return $x$ by concatenating $x_A$ and $x_{A^c}$
```

The derivation of this algorithm and the general form of the algorithm for the ℓ_p-case can be found in appendix F.1. It is possible to construct a similar decomposition algorithm, namely Algorithm 5 in appendix F.2, to compute the norm itself.

Name	$F(A)$	Norm Ω_p
cardinality	$\|A\|$	Lasso (ℓ_1)
nb of groups	$\sum_{B\in\mathcal{G}} 1_{\{A\cap B\neq\varnothing\}}$	Group Lasso (ℓ_1/ℓ_p)
max. nb of el./group	$\max_{B\in\mathcal{G}} \|A\cap B\|$	Exclusive Lasso (ℓ_p/ℓ_1)
constant	$1_{\{A\neq\varnothing\}}$	ℓ_p-norm
sublinear f. of cardinality	$h(\|A\|)$, h sublinear	
	$1_{\{A\neq\varnothing\}} \vee \frac{\|A\|}{k}$	k-support norm ($p=2$)
concave f. of cardinality	$h(\|A\|)$, h concave	OWL ($p=\infty$)
	$\lambda_1\|A\| + \lambda_2\Big[\binom{d}{k} - \binom{d-\|A\|}{k}\Big]$	OSCAR ($p=\infty, k=2$)
	$\sum_{i=1}^{\|A\|} \Phi^{-1}\big(1-\frac{qi}{2d}\big)$	SLOPE ($p=\infty$)
chain length	$h(\max(A))$	wedge penalty
tree leaf volume	$\sum_{i\in T_B} f_i$	
graphical hull volume	$\sum_{i\in A_B} d_i$	

Table 2: **Combinatorial functions and the corresponding norms.** All the norms in this table are instances of the family of norms we study in this paper. See section 4.3 for the Exclusive lasso, section 7.3 for functions of the cardinality, and the current section for tree and graph penalties.

7 More examples

Having presented some elements of the theory of submodular functions and presented some general results, we are in position to develop more sophisticated examples, namely combinatorial functions inducing hierarchical sparsity patterns and leading to norms such as the wedge penalty considered by Micchelli et al. (2013) (see also Yan and Bien, 2015), the ℓ_∞-version of the tree-structured norm considered by Jenatton et al. (2011b); Zhao et al. (2009a) as well as tighter relaxations for the ℓ_p-case, and more general functions of the cardinality which lead to the k-support norm of Argyriou et al. (2012), the dual of the vector Ky-Fan k-norm, and the OWL penalties of Figueiredo and Nowak (2014) with as particular cases the OSCAR penalty (Bondell and Reich, 2008) and the SLOPE penalty (Bogdan et al., 2015), but also in each case to a number of new norms with algorithms to compute them as well as the corresponding proximal operators.

7.1 Overlap count Lasso

Mairal et al. (2011) studied regularization with overlapped group ℓ_1/ℓ_∞-norms; they showed in particular that the proximal problem could be solved efficiently by reformulating it as a quadratic min-cost flow problem and using an efficient divide-and-conquer algorithm proposed by Hochbaum and Hong (1995) and Gallo et al. (1989). We provide an interpretation of this result in the light of the theory developed in this paper. As discussion in Section 4.1, the function $F_\cap(A)=\sum_{B\in\mathcal{G}} d_B 1_{\{A\cap B\neq\varnothing\}}$ is submodular as a positive combination of simple submodular functions. For any $v\in\mathbb{R}^d_+$, its Lovász extension satisfies $f(v)=\sum_{B\in\mathcal{G}} d_B \max_{j\in B} v_j$. The corresponding norm $\Omega_\infty^{F_\cap}$ is thus equal to the overlapped ℓ_1/ℓ_∞-norm $\Omega_\infty^{F_\cap}(w)=\sum_{B\in\mathcal{G}} d_B\|w_B\|_\infty$ studied by Mairal et al. (2011). However, for $p<\infty$, $\Omega_p^{F_\cap}(w)\neq\sum_{B\in\mathcal{G}} d_B\|w_B\|_p$. To work with a given submodular function it is key to be able to solve $\min_A \lambda F(A) - s(A)$ for $s \in \mathbb{R}_+^d$, but this problem is equivalent to solving $\min_{w\in[0,1]^d} \lambda f(w) - \langle s, w\rangle$. Yet, $\lambda f(w) = \max_{\kappa:\, \Omega_\infty^*(\kappa)\leq\lambda} \langle \kappa, w\rangle$ so that by duality the initial submodular minimization is equivalent to

$$\max_{\kappa\in\mathbb{R}^d} \quad -\sum_{i=1}^d (s_i - \kappa_i)_+ \quad \text{s.t.} \quad \Omega_\infty^*(\kappa) \leq \lambda,$$

with

$$\Omega_\infty^*(\kappa) = \inf_\xi \Big\{ \max_{B\in\mathcal{G}} d_B^{-1} \|\xi^{(B)}\|_1 \mid \kappa = \sum_{B\in\mathcal{G}} \xi^{(B)},\ \big(\forall B \in \mathcal{G},\ \xi^{(B)}_{B^c} = 0\big) \Big\}.$$

Since $s \geq 0$, we can let $\kappa, \xi \geq 0$, and we can rewrite the previous problem as

$$\max_{0\leq\kappa\leq s} \sum_{i=1}^d \kappa_i - s(V) \quad \text{s.t.} \quad \forall i,\ \kappa_i = \sum_{B\ni i} \xi_i^{(B)}, \quad \text{and} \quad \forall B \in \mathcal{G},\ \sum_{j\in B} \xi_j^{(B)} \leq d_B \lambda.$$

This last problem can be interpreted as a max-flow problem with the following structure: Let σ and τ be respectively a source and a sink, and consider the directed graph with nodes $\{\sigma, \tau\} \cup [\![1, d]\!] \cup \mathcal{G}$ and with the following set of edges

$$\begin{cases} \forall B \in \mathcal{G}, & (\sigma, B) \quad \text{with capacity } d_B\lambda \\ \forall B \in \mathcal{G},\ i \in B, & (B, i) \quad \text{with unlimited capacity} \\ \forall i \in [\![1, d]\!], & (i, \tau) \quad \text{with capacity } s_i. \end{cases}$$

Then $\xi_i^{(B)}$ and κ_i are respectively interpreted as the flows on the edges (B, i) and (i, τ), and the previous optimization problem is equivalent to the maximization of the flow between σ and τ. Mairal et al. (2011) write the counterpart of this formulation for the proximal problem of $\Omega_\infty^{F_\cap}$ which involves the same graph, but with additional variables u_i related to the quadratic term. By reformulating directly the submodular minimization as a max-flow problem, we can extend the results for $p = \infty$ to $p < \infty$ and compute efficiently all norms $\Omega_p^{F_\cap}$ (with Algorithm 5) and their proximal operators (with e.g. Algorithm 2 for $p = 2$ and Algorithm 4 for general p). If some groups are nested the max-flow formulation can be simplified to some extent: see Mairal et al. (2011) for more details. It is interesting to note that other submodular functions such as cut functions that lead to extensions that are variant of the total variation can take advantage of the same divide-and-conquer algorithm with other max-flow formulations (Chambolle and Darbon, 2009; Luss and Rosset, 2014).

7.2 Hierarchical sparsity

In a number of applications, the variables or group of variables are naturally organized on a chain, a tree or more generally a directed acyclic graph $G = (V, E)$, in a hierarchical fashion. Obtaining sparsity patterns that satisfy hierarchical relations is however not easy and has been the focus of a number of papers (Bien et al., 2013; Jenatton et al., 2011b; Mairal et al., 2011; Yan and Bien, 2015; Yuan et al., 2009; Zhao et al., 2009a).

With the usual terminology of graph theory, the variable associated to node i has therefore a set of descendants D_i (the set of nodes j such that there exists a directed path from i to j, including by convention the node i itself) and ascendants A_i (the set of nodes j such that $i \in D_j$). As usual, the set of immediate descendants is called the set of children and denoted C_i and the set of immediate ancestors is called the set of parents and denoted Π_i. For trees, π_i will denote the only parent of a node i which is not the root. We will call the *hull* of B the set A_B of ancestors of B, that is the set

$A_B := \cup_{i \in B} A_i$. We will call set of terminal nodes of B the set T_B of nodes of B that do not have any descendant in B (except themselves).

In this type of setting, functions of the form $F_\cap$ and $F_\cup$ are naturally associated to the graph by choosing $\mathcal{G}$ to be either the collection of ancestor sets $\mathcal{G} = \{A_i\}_{i \in V}$ or the collection of descendant sets $\mathcal{G} = \{D_i\}_{i \in V}$. Indeed, given non-negative weights d_i and f_i, it is natural to define the counting function

$$F_\cap(B) := \sum_{i \in V} d_i \, 1_{\{B \cap D_i \neq \varnothing\}} = \sum_{i \in A_B} d_i, \tag{13}$$

and the function defined as the weighted set-cover by the ancestor sets $(A_i)_{i \in V}$

$$F_\cup(B) := \inf_{I \subset V} \Big\{ \sum_{i \in I} f_i \;\mid\; B \subset \bigcup_{i \in I} A_i \Big\}. \tag{14}$$

Obviously, the role of the descendant sets and the ascendant sets in both functions can be exchanged by considering the graph with flipped edges. Note that without loss of generality the weights f_i can be assumed non-decreasing w.r.t. to the graph (i.e., such that $(j \in D_i) \Rightarrow (f_i \leq f_j)$ since they can be modified to satisfy this property without changing the function $F_\cup$: to see this, note that $F_\cup$ is a weighted min-cover, and that $j \in D_i$ is equivalent to $A_i \subset A_j$, therefore if $f_i > f_j$, $F(A_i) > F_-(A_i)$ and so A_i will never enter the cover; therefore, decreasing the value of f_i to f_j does not change anything (A_i is still never selected). Given this argument, using the weights $f_i^- := \min_{j \in D_i} f_j$ yields the same function $F_\cup$. In fact, $f_i^- = F_{\cup,-}(A_i)$.

If $d_i = 1$ for all i, $F_\cap$ reduces to $B \mapsto |A_B|$, the size of the hull of B.

7.2.1 Special cases and comparison between $F_\cap$ and $F_\cup$.

To illustrate the relevance of combinatorial function and norms defined on a graph, we consider the special case when the graph is either a chain or a tree.

Case of the chain. In a chain on p nodes, oriented from left to right, we have $D_i = [\![i, p]\!]$, $A_B = [\![1, \max(B)]\!]$ and $T_B = \{\max(B)\}$.

So that

$$F_\cap(B) = \sum_{i=1}^{\max(B)} d_i \qquad \text{and} \qquad F_\cup(B) = f_{\max(B)}.$$

These two functions are thus equal if and only if, for $i \in [\![1, p]\!]$, $d_i = f_i - f_{i-1}$ with $f_0 = F(\varnothing) = 0$. The counting and set-cover functions thus define here the same family of combinatorial functions.

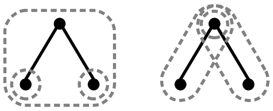
In the ℓ_2-case the variational form of the norm

$$\Omega_2(w) = \min_{\eta \in \mathbb{R}_+^d} \frac{1}{2} \sum_{i=1}^{n} \Big[\frac{w_i^2}{\eta_i} + \eta_i d_i \Big] \qquad \text{s.t.} \qquad \forall i > 1,\ \eta_i \leq \eta_{i-1}$$

shows that this norm is the wedge penalty considered by Micchelli et al. (2013). We will show in Corollaries 5 and 6 that this norm and its proximal operators can be computed very efficiently, in fact in linear time, using the PAV algorithm (Best and Chakravarti, 1990). Yan and Bien (2015) compared the norms $\Omega_2^{F_\cup}$ with the norms $\widetilde{\Omega}_2$ of Jenatton et al. (2011b) (see Sec. 4 and the next paragraph on trees) and concluded that, even in the chain case these norms are different, and that the norm $\widetilde{\Omega}_2$ over-penalizes elements at the ends of the chain; in the light of our work, this is not surprising since the norm $\widetilde{\Omega}_p$ do not provide a tight relaxation of $F_\cup$ as opposed to $\Omega_2^{F_\cap}$.

Case of a tree. In the case of a tree, we first show that the two families of functions are not equivalent. Indeed, consider the tree consisting of a root 1 with two children 2 and 3. $F_\cap$ and $F_\cup$ are defined respectively as weighted intersection counts with descendants set D_i and by minimum weight set-cover by collections of ancestor sets A_i, with weights associated to sets and resulting values reported in Figure 7 below.

	set	weight
$F_\cap$	$D_1 = \{1,2,3\}$	d_1
	$D_2 = \{2\}$	d_2
	$D_3 = \{3\}$	d_2



	set	weight
$F_\cup$	$A_1 = \{1\}$	f_1
	$A_2 = \{1,2\}$	f_2
	$A_3 = \{1,3\}$	f_2

	$\varnothing$	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
$F_\cap$	0	d_1	$d_1 + d_2$	$d_1 + d_2$	$d_1 + d_2$	$d_1 + d_2$	$d_1 + 2d_2$	$d_1 + 2d_2$
$F_\cup$	0	f_1	f_2	f_2	f_2	f_2	$2f_2$	$2f_2$

Figure 7: (top center) Descendant and ascendant sets defining respectively $F_\cap$ and $F_\cup$, (top left and right) tables defining weights associated with sets, (bottom) table of values assigned by $F_\cap$ and $F_\cup$ to all subsets of $\{1, 2, 3\}$.

For the two functions to be equal, we would need to have $d_1 = f_1 = 0$. This shows that the families of functions are in general distinct for trees. Furthermore, we can only have the inequality $F_\cup(\{1,2\}) + F_\cup(\{1,3\}) = 2f_2 \geq f_1 + 2f_2 = F_\cup(\{1\}) + F_\cup(\{1,2,3\})$ if $f_1 = 0$ which shows that $F_\cup$ is not submodular[7].

For the function $F_\cap$, we have

$$\Omega_\infty^{F_\cap}(w) = \sum_{i \in V} d_i \|w_{D_i}\|_\infty.$$

Clearly, this norm is an instance of the weighted ℓ_1/ℓ_p-norms of the form

$$\widetilde{\Omega}_p(w) = \sum_{i \in V} d_i \|w_{D_i}\|_p, \quad p \in \{2, \infty\},$$

that were considered by Jenatton et al. (2011b). It should be noted however that $\Omega_p^{F_\cap} \neq \widetilde{\Omega}_p$ for any $p \in (1, \infty)$, with $\Omega_p^{F_\cap}$ having no simple established closed form; the only value of p for which the two norm coincide is $p = \infty$. Note that for $p = \infty$ and $p = 2$, the proximal operator for $\widetilde{\Omega}_p$ can be computed efficiently in closed form, as a composition of proximal operators for groups of descendants starting from the leaves (Jenatton et al., 2010).

For $F_\cup$, if the f_i are assumed non decreasing w.r.t. to the tree (i.e. such that $\forall i \in V$, $f_{\pi_i} \leq f_i$), then if we call T_B the set of terminal nodes (or leaves) of the tree induced on a set B of nodes, that is, the subset of nodes i of B such that $D_i \cap B = \{i\}$, then we have $F_\cup(B) = \sum_{i \in T_B} f_i$. In particular, if $f_i = 1$ for all i, then $F_\cup(B) = |T_B|$. Note however that in that last case, the only possible supports are unions of paths from the root to a leaf of the tree: in order to obtain a penalty that allows as possible sparsity patterns all rooted subtrees it is necessary to impose that $i \mapsto f_i$ is *strictly increasing* along the graph.

7.2.2 Computations of norms and proximal operators for the hierarchical $F_\cap$

The following lemma shows that the norms Ω_∞ can be computed in linear time and the norms Ω_p and Prox_{Ω_2} can be computed by solving a general isotonic regression problem on the graph (V, E).

[7] except in very degenerate cases.

Lemma 10 (Computation of Ω_p, Ω_p^* and Prox_{Ω_2} for $F_\cap$)**.** *For the function $F : B \mapsto \sum_{i \in A_B} d_i$, with A_B the set of ancestors of B:*

1. *When $p = \infty$, we have $\Omega_\infty(w) = \sum_{i \in V} d_i \|w_{D_i}\|_\infty$ so that $\Omega_\infty(w)$ is computed recursively (in reverse topological order on the graph) in linear time.*

2. *For any $1 < p \leq \infty$, we have $\Omega_p^*(s) = \max_{B \subset V} F(A_B)^{-1/q} \|s_{A_B}\|_q$. The norm Ω_p^* can be computed using Algorithm 1 via a sequence of minimization of functions of the form $A \mapsto \lambda F(A) - \|s_A\|_q^q$.*

3. *When $1 < p < \infty$, $\Omega_p(w) = \min_{\eta \in \mathbb{R}_+^d} \frac{1}{p} \frac{w_i^p}{\eta_i^{p-1}} + \frac{1}{q} \sum_{i=1}^n d_i \eta_i$ s.t. $\forall (i,j) \in E,\ \eta_i \geq \eta_j$.*

4. *The proximal operator $Prox_{\Omega_2}$ satisfies $[Prox_{\Omega_2}(u)]_i = \big(1 + \frac{\lambda}{\eta_i^\star}\big)^{-1} u_i$ where $\eta^\star$ is the solution of*
$$\min_{\eta \in \mathbb{R}_+^d} \sum_{i \in V} \left(\frac{u_i^2}{\eta_i + \lambda} + d_i \eta_i \right) \qquad \textit{s.t.}\ \ \forall (i,j) \in E,\ \eta_i \geq \eta_j. \tag{15}$$

Proof.

1. The form of Ω_∞ follows from the fact that F is a counting function.

2. The form of the dual norm stems from the fact that the core set $\mathcal{D}_F$ consists of the sets that are hulls. We discuss in section 7.2.4 that, for a tree, the minimization of $A \mapsto \lambda F(A) - s(A)$ for $s \in \mathbb{R}_+^n$ can be done in $O(n)$. For a more general DAG, the general max-flow formulation of Section 7.1 can be used, but unfortunately the DAG structure cannot a priori be easily leveraged to obtain a formulation scaling linearly with the number of nodes or edges.

3. We have for $\eta \in \mathbb{R}_+^d$, $\Omega_\infty(\eta) = \sum_{i=1}^d d_i \max_{j \in D_i} \eta_j$. As a consequence, using the variational formula (6) we have
$$\begin{aligned} \Omega_p(w) &= \min_{\eta \in \mathbb{R}_+^d} \frac{1}{p} \frac{w_i^p}{\eta_i^{p-1}} + \frac{1}{q} \sum_{i=1}^n d_i \max_{j \in D_i} \eta_j \\ &= \min_{\eta \in \mathbb{R}_+^d} \frac{1}{p} \frac{w_i^p}{\eta_i^{p-1}} + \frac{1}{q} \sum_{i=1}^n d_i \eta_i \quad \text{s.t. } \forall (i,j) \in E,\ \eta_i \geq \eta_j, \end{aligned} \tag{16}$$
where the second equality stems from the fact that $\eta_i \mapsto \frac{w_i^2}{\eta_i}$ is non-increasing so that at the optimum, we should have $\eta_i = \max_{j \in D_i} \eta_j$ and thus $\eta_i \geq \eta_j$ for all $j \in D_i$.

4. The proof for the proximal problem when $p = 2$ uses the same argument by rewriting (8).

□

Since Eq. (15) is the minimization of the separable convex function subject to isotonic constraints, it can be solved using the divide-and-conquer algorithm for minimizing $\sum_{i \in V} \big(\frac{u_i^2}{\eta_i + \lambda} + d_i \eta_i\big) + h(\eta)$, for $h(\eta)$ the Lovász extension $h(\eta) = M \sum_{(i,j) \in E} (\eta_j - \eta_i)_+$ of a cut funtion, and for M sufficiently large (Bach, 2013, Section 9.1); see also Luss and Rosset (2014).

Note that the variational formulations show clearly that the sparsity pattern obtained have a hierarchical structure. Indeed, the inequality constraint $\eta_i \geq \eta_j$ for $(i,j) \in E$ enforces that if $\eta_i = 0$ then $\eta_j = 0$ for all $j \in D_i$, and since $(\eta_j = 0) \Rightarrow (w_j = 0)$ this entails as well that $w_{D_i} = 0$. Note however, that the norm does not impose the type of constraint $|w_i| \geq |w_j|$ introduced in some of the previous literature (Bien et al., 2013; Yuan et al., 2009) which imposed that the estimated coefficient where decreasing in magnitude.

7.2.3 Computations of norms and proximal operators for the chain case

In this section, we show that in the chain case (considered in Micchelli et al. (2013) and Yan and Bien (2015)) the optimization problems from Lemma 10 defining respectively Ω_p and the proximal operator for Ω_2 can be both solved as a general isotonic regression problem with a total order.

Consider the following form of the classical isotonic regression with a total order

$$\min_{x\in\mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{d} \omega_i (x_i - y_i)^2 \quad \text{s.t.} \quad x_1 \leq \ldots \leq x_d \leq b, \tag{IRC(ω, y, b)}$$

where $\omega_i > 0$ for all $i \in V = \{1, \ldots, d\}$ and $b \in \overline{\mathbb{R}}$. Note that the objective being strongly convex, the problem has a unique solution. This optimization problem is known to be solved efficiently by the pooled adjacent violators (PAV) algorithm (Best and Chakravarti, 1990).

Following Barlow and Brunk (1972), we show that a form of generalized isotonic regression problem with total order can be reduced to solving a classical isotonic regression, and thus benefits also from the efficiency of that algorithm.

Lemma 11. *For $c \in \mathbb{R}^d_+, z \in \mathbb{R}^d_+$ with $z_1 > 0$ and ψ a nonnegative differentiable, decreasing and strictly convex function, consider the optimization problem:*

$$\min_{\eta\in\mathbb{R}^d} \sum_{i=1}^{d} c_i \psi(\eta_i) + z_i \eta_i \quad \textit{s.t.} \quad b' \leq \eta_d \leq \ldots \leq \eta_1. \tag{GIRC(c, z, b')}$$

If $\forall i,\ c_i \neq 0$, then if x^* is the solution of IRC(ω, y, b) with $\omega_i = c_i$, $y_i = \frac{z_i}{c_i}$ and $b = -\lim_{\eta\to b'} \psi'(\eta)$, then the vector η^* with components $\eta^*_i = (\psi')^{-1}(-x^*_i)$ is the unique solution to GIRC(c, z, b'). If for some indices i, $c_i = 0$, the problem reduces to the previous one after clustering or removing some of the variables η_i.

A more detailed version of this lemma and a proof are provided in appendix F.4.

Corollary 5. *For a chain, problem (16) can be solved by applying Lemma 11 with*

$$\psi(\eta) = \tfrac{q}{p}\eta^{1-p},\ b' = 0,\ c_i = w_i^p,\ z_i = d_i, \quad \textit{so that} \quad b = +\infty \quad \textit{and} \quad \forall i \in I,\ \omega_i = w_i^p,\ \bar{y}_i = \bar{z}_i\, w_i^{-p}$$

and $\eta^*_i = x^{*\,-1/p}_i$, where x^* is the solution of IRC($\omega, \bar{y}, +\infty$).

Corollary 6. *For a chain, problem (15) can be solved by applying Lemma 11 with*

$$\psi(\eta) = (\eta+\lambda)^{-1}, b' = 0,\ c_i = u_i^2,\ z_i = d_i, \quad \textit{so that} \quad b = \lambda^{-2} \quad \textit{and} \quad \forall i \in I,\ \omega_i = u_i^2,\ \bar{y}_i = \bar{z}_i\, u_i^{-2}$$

and $\eta^*_i = x^{*\,-1/2}_i - \lambda$, where x^* is the solution of IRC($\omega, \bar{y}, \lambda^{-2}$).

As a consequence, for chains, problems (15) and (16) can be solved efficiently using the PAV algorithm. Yan and Bien (2015) propose an algorithm to compute the proximal operator in the chain case, but its complexity is quadratic in the length of the chain, while PAV is linear.

7.2.4 Computations for $F_\cap$ on a tree

If the graph is a tree and if the nodes are indexed in topological order, Jenatton et al. (2011b) showed that, for $p \in \{2, \infty\}$, the proximal operator of the norm $\widetilde{\Omega}_p : z \mapsto \sum_{i=1}^n d_i \|z_{D_i}\|_p$ is computed as

$$\mathrm{Prox}_{\widetilde{\Omega}_p} = \mathrm{Prox}_p^{(1)} \circ \ldots \circ \mathrm{Prox}_p^{(n)} \qquad \text{with} \qquad \mathrm{Prox}_p^{(i)}(z) = \arg\min_x \frac{1}{2}\|x - z\|_2^2 + \lambda d_i \|x_{D_i}\|_p.$$

Since $\Omega_\infty^{F_\cap} = \widetilde{\Omega}_\infty$, this provides an efficient algorithm to compute the proximal operator in that case. Jenatton et al. (2011b) show (see their Lemma 7) that when $p = \infty$ this algorithm can be implemented with a complexity of $O(hn)$, where h is the h is the height of the tree. This suggests that its complexity is similar to that of the divide-and-conquer algorithm (which is however likely to be more efficient for tall thin trees).

In the case $p < \infty$ and in particular when $p = 2$, $\Omega_p^{F_\cap} \neq \widetilde{\Omega}_p$, whether it is possible to compute the norm or the proximal operator with similar dynamic programs remains open. Nevertheless, for a tree, the divide-and-conquer algorithms to compute the norm and the proximal operator (Alg. 2) are efficient, because the submodular function of the form $A \mapsto \lambda F(A) - s(A)$ for $s \in \mathbb{R}_+^n$ can be minimized in linear time. The minimizer has to be a stable set, thus here a rooted subtree, and the optimal one is computed by Algorithm 3 (see Appendix F.3 for a proof). Moreover the restriction F_A and the contraction F^A are both themselves of the same form $F_\cap$ for a tree/forest graph: F_A is of the same form on the tree induced by the restriction on A and F^A is of the same form on the forest induced on the complement of A.

Whether it is possible to leverage efficient algorithms for isotonic regression that have been proposed for trees (Pardalos and Xue, 1999) or under other assumptions (Stout, 2013) to solve problems (16) and (15) with more efficient algorithms for more general graphs is left open.

Algorithm 3 Minimizing $\lambda F(A) - s(A)$ for $s \in \mathbb{R}_+^n$

```
Require: Nodes indexed in topological order, $(\pi_i)_i$ parents, $(C_i)_i$ children sets.
for $i = n$ to 1 do
    $s_{\pi_i} \leftarrow s_{\pi_i} + (s_i - \lambda)_+$
    $u_i = 1_{\{s_i > \lambda\}}$
end for
$A \leftarrow$rectree$(1, u)$
return $A$
```

with

```
function rectree$(k, u)$
$A \leftarrow \varnothing$
if $u_k = 1$ then
    $A \leftarrow \{k\} \cup$ rectree$(j_1, u) \cup \ldots \cup$ rectree$(j_\kappa, u)$ with $\{j_1, \ldots, j_\kappa\} := C_k$
end if
return $A$
```

7.2.5 Computations of Ω_∞ for $F_\cup$ on a tree

The construction of norms associated to $F_\cup$ on a DAG (as defined in Equation (14)) has been recently discussed in Yan and Bien (2015). For trees, the dual norms $(\Omega_p^{F_\cup})^*$ can clearly be computed efficiently by dynamic programming. Unfortunately, even for a tree $F_\cup$ is clearly not submodular as discussed after Figure 7. It is however possible to compute efficiently the primal norm $\Omega_\infty^{F_\cup}$ with dynamic programming.

Proposition 5. *For the function $F_\cup$ defined as the minimal weighted set cover by the ancestor sets $(A_i)_{i \in V}$ with weights f_i, if C_i denotes the set of children of node i, π_i denotes the parent of node i and with $d_i := f_i - f_{\pi_i}$, the associated norm $\Omega_\infty^{F_\cup}$ is computed as*

$$\Omega_\infty^{F_\cup}(w) = \sum_{i=1}^{d} d_i \zeta_i \quad \textit{with} \quad \zeta_i \textit{ defined by the recursion} \quad \zeta_i = \max\Big(|w_i|, \sum_{j \in C_i} \zeta_j\Big).$$

Proof.

$$\begin{aligned}
\Omega_\infty(w) &= \max_{\kappa \in \mathbb{R}_+^d} \kappa^\top |w| \quad \text{s.t.} \quad \forall j \in V, \sum_{i \in A_j} \kappa_i \leq f_j \\
&= \min_{\mu \in \mathbb{R}_+^d} \max_{\kappa \in \mathbb{R}_+^d} \kappa^\top |w| - \sum_{j \in V} \mu_j \Big[\sum_{i \in A_j} \kappa_i - f_j \Big] \\
&= \min_{\mu \in \mathbb{R}_+^d} \max_{\kappa \in \mathbb{R}_+^d} \sum_{i \in V} \Big[\kappa_i |w_i| - \Big(\sum_{j \in D_i} \mu_j \Big) \kappa_i + \mu_i f_i \Big] \\
&= \min_{\mu \in \mathbb{R}_+^d} \sum_{i \in V} \mu_i f_i \quad \text{s.t.} \quad \forall i \in V, \ |w_i| \leq \sum_{j \in D_i} \mu_j. \\
&= \sum_{i \in V} \zeta_i \big(f_i - f_{\pi_i} \big) \ \text{ s.t. } \ \forall i \in V, \ |w_i| \leq \zeta_i, \ \zeta_i \geq \sum_{j \in C_i} \zeta_j.
\end{aligned}$$

Hence the result by minimizing recursively over ζ_i in reverse topological order. □

The possibility of designing efficient optimization schemes for the computation of Ω_p, Ω_p^* and the corresponding proximal operators for the case of functions $F_\cup$ on a tree, let alone on a general graph, remains an open problem.

7.3 Functions of the cardinality

Another particular instance of combinatorial functions are functions that only depend on the cardinality of the set, i.e., functions of the form

$$F(A) = \sum_{k=1}^{d} f_k \, 1_{\{|A|=k\}}. \tag{17}$$

We have already discussed the cardinality function which is relaxed into the ℓ_1-norm and the function $B \mapsto 1_{\{B \neq \varnothing\}}$, whose ℓ_p relaxation is simply the ℓ_p-norm, but as we will see other functions are also of interest. To consider more elaborate examples and since we are interested by the convex relaxation of these functions, only the LCEs of this type should retain our attention. Given the interpretation of the LCE in terms of fractional weighted set-cover, we can essentially restrict ourselves to functions that are *non-decreasing* and *sublinear*, where sublinearity follows from the fact that for any sets A and B, we must have $F(A \cup B) \leq F(A) + F(B)$ which implies that $f_{k+l} \leq f_k + f_l$. Note that the function $k \mapsto f_k$ is concave if and only if F is submodular (see Bach, 2013, Section 9.1). As illustrated by Example 4, LCEs depending only on the cardinality are not necessarily submodular.

In general the dual norm can be computed in linear time since we have

$$\big(\Omega_p^*(w)\big)^q = \max_{1 \leq j \leq d} \frac{1}{f_j} \sum_{i=1}^{j} |s|_{(i)}^q,$$

Now, if F is submodular (i.e. $k \mapsto f_k$ is the restriction of a concave function), Ω_∞ takes the very simple form

$$\Omega_\infty(w) = \sum_{i=1}^{d} (f_i - f_{i-1}) \, |w|_{(i)}, \tag{18}$$

where $|w|_{(i)}$ is the ith largest order statistic of the vector $|w|$, and with $f_0 = 0$. We thus obtain the family of ordered weighted ℓ_1-norms, introduced as the ordered weighted Lasso (OWL) penalties by Figueiredo and Nowak (2014).

Furthermore, in this submodular situation, the reformulation of the proximal problem provided by (8) provides a way to compute the norm Ω_p for all $p \in (1, \infty)$ and the proximal operator for $p \in \{2, \infty\}$ in $O(n)$ using the Pooled Adjacent Violators algorithm (PAV) via a reduction to the case of the chain.

Indeed, for a function F which depends only on the cardinality, the function f is a symmetric function of its arguments and so for any $\eta \in \mathbb{R}^d_+$ and any permutation σ, we have $f(\eta) = f(\eta^{(\sigma)})$ with $\eta^{(\sigma)} = (\eta_{\sigma(1)}, \dots, \eta_{\sigma(d)})$.

Proximal operator when F is submodular and $p = \infty$. When $p = \infty$, the proximal problem takes the form:

$$\min_{w \in \mathbb{R}^d} \frac{1}{2}\|w - u\|^2 + \lambda \Omega_\infty(w).$$

Since the norms Ω_p are *absolute norms* (Bauer et al., 1961) (i.e. $\Omega_p(w) = \Omega_p(|w|)$) we have $[\mathrm{Prox}_{\Omega_p}(u)]_i = [\mathrm{Prox}_{\Omega_p}(|u|)_i \,\mathrm{sign}(u_i)$. Without loss of generality, we can thus assume that $u \in \mathbb{R}^d_+$. Since the norm is also symmetric, we can assume $u_1 \geq \dots \geq u_d$. But because of symmetry, the components of the solution $w^\star$ of the proximal problem must then be in the same order as u: indeed, first, if $u_i = u_j$ then $w^\star_i = w^\star_j$ by symmetry, and second, since $\Omega_p(w)$ does not depend on the order and since we have $\big[(u_1 - w_2)^2 + (u_2 - w_1)^2\big] - \big[(u_1 - w_1)^2 + (u_2 - w_2)^2\big] = 2(w_1 - w_2)(u_1 - u_2)$ which is negative if $u_1 > u_2$ and $w_2 > w_1$, the objective is decreased by any transposition which brings $\{u_i, u_j\}$ and $\{w_i, w_j\}$ in the same order. So for $u_1 \geq \dots \geq u_d \geq 0$, using the Choquet integral representation of f, the proximal problem is equivalent to

$$\min_{w \in \mathbb{R}^d_+} \frac{1}{2}\|w - u\|^2 + \lambda \sum_{i=1}^{d} w_i (f_i - f_i - 1), \quad \text{ s.t. } \; w_1 \geq \dots \geq w_d,$$

which is a classical isotonic regression problem with total order. We recover the algorithm of Figueiredo and Nowak (2014).

Proximal operator when F is submodular and $p = 2$. Consider the computation of the proximal operator for a vector u which w.l.o.g. satisfies $u_1 \geq \dots \geq u_d$; if $\eta^\star$ is the solution of (8) then we must have $\eta^\star_1 \geq \dots \geq \eta^\star_d$. Indeed, if this is not the case then consider the vector $\eta^{(\sigma)}$ obtained by sorting the components of η in decreasing order: it leads to a smaller value of the term involving u and does not change the value of $f(\eta)$. This implies (assuming $u_1 \geq \dots \geq u_d$) that (8) is equivalent to

$$\min_{\eta \in \mathbb{R}^d_+} \sum_{i \in V} \frac{u_i^2}{\eta_i + \lambda} + f(\eta) \quad s.t. \quad \eta_1 \geq \dots \geq \eta_d.$$

Now, given that the order of the coefficients of η is fixed, by the Choquet integral representation of f the latter is linear, the problem is thus

$$\min_{\eta \in \mathbb{R}^d_+} \sum_{i \in V} \left[\frac{u_i^2}{\eta_i + \lambda} + \eta_i (f_i - f_{i-1})\right] \quad s.t. \quad \eta_1 \geq \dots \geq \eta_d.$$

which is the same as in the chain case and can be solved thanks to Corollary 6 using a PAV algorithm.

As in Lemma 10, similarly, for any $p < \infty$, the computation of the norm reduces to a problem of the form (16) which can be solved efficiently by a PAV algorithm thanks to Corollary 5.

Computation of proximal operator when the function is not submodular. It is important to stress that these reductions to generalized isotonic formulation is not possible when the function is not submodular because in that case $f(\eta)$ is not linear given the ordering constraints. It is however possible to propose an efficient algorithm to compute the proximal operator for some of these functions. The k-support norm is an example of such as case.

Illustration 1: The k-support norm and the vector Ky-Fan k-norm. One of the simplest functions that depends only on the cardinality is the function

$$F : A \mapsto \begin{cases} 0 & \text{if} \quad A = \varnothing \\ 1 & \text{if} \quad |A| = k, \\ \infty & \text{if} \quad |A| \notin \{0, k\}. \end{cases}$$

The norms Ω_p associated with this function are naturally of the form of a Latent group Lasso, since the domain of F is restricted to sets of of cardinality k or 0. Clearly, the dual norm Ω_p^* satisfies $\Omega_p^*(s) = \max_{A:|A|=k} \|s_A\|_q$. This shows first that Ω_2 is the *k-support norm* introduced by Argyriou et al. (2012). It also implies that $\Omega_\infty^*(s) = \max_{A:|A|=k} \|s_A\|_1 = |s_{(1)}| + \ldots + |s_{(k)}|$, where $|s_{(1)}|, \ldots, |s_{(d)}|$ are the order statistics of $(|s_1|, \ldots, |s_d|)$, so that Ω_∞^* is the vector Ky-Fan k-norm[8]. The LCE of F is the function $F_-(B) = 1_{\{B \neq \varnothing\}} \vee \frac{|B|}{k}$. It is immediate to check that F_- is not submodular by considering a pair of sets of cardinality k. Extensions of k-support norms considered in McDonald et al. (2015) could also be cast in this framework.

Illustration 2: The SLOPE penalty introduced in Bogdan et al. (2015) is clearly of the form of (18) with $f_i - f_{i-1} = \Phi^{-1}(1 - \frac{iq}{2d})$, where Φ is the cumulative density function the standard Gaussian distribution. Since Φ in an increasing function, $f_i - f_{i-1}$ is positive and decreasing, which shows that $i \mapsto f_i$ an increasing concave function. It is therefore submodular and the theory develops in this section applies. In particular it retrieves the algorithm of Bogdan et al. (2015) to compute the proximal operator, and propose ℓ_p variants of SLOPE.

Illustration 3: The OSCAR penalty. A norm of the form $w \mapsto \lambda_1 \|w\|_1 + \lambda_2 \sum_{i<j} \max\big(|w_i|, |w_j|\big)$ was introduced in Bondell and Reich (2008) because its non-differentiabilities when $|w_i| = |w_j|$ induce some clustering of the amplitudes of the coefficients. Clearly, the second term in the definition of the OSCAR penalty is of the form $\Omega_\infty(w) = \sum_{A:|A|=k} \|w_A\|_\infty$. This is a particular instance of an Overlap Count Lasso. The LCE of the combinatorial functions associated with Ω_∞ is the counting (and thus submodular) function $F(B) = \sum_{A:|A|=k} 1_{\{A \cap B \neq \varnothing\}}$. Clearly,

$$F(B) = \big|\{A : |A| = k, A \nsubseteq B^c\}\big| = f_l := \binom{d}{k} - \binom{d - |B|}{k},$$

and we have

$$\Omega_\infty^*(s) = \max_{1 \leq l \leq d} \frac{1}{f_l} \sum_{i=1}^{l} |s_{(i)}|.$$

[8] The vector Ky-Fan k-norm is the vector counterpart of the matrix Ky-Fan norm, the latter being computed as the k-norm of the singular values of the matrix.

As shown in Section 6, since F is submodular, Ω_∞ is its Lovász extension and using the so-called Choquet integral representation (10) of F, and since F depends only on the cardinality, we have

$$\begin{aligned}\Omega_\infty(w) &= \sum_{l=1}^{d} |w_{(l)}| \left[F(\{1,\dots,l\}) - F(\{1,\dots,l-1\})\right] \\ &= \sum_{l=1}^{d} |w_{(l)}| \left[\binom{d-l}{k} - \binom{d-l-1}{k}\right] \\ &= \sum_{l=1}^{d} |w_{(l)}| \binom{d-l-1}{k-1}.\end{aligned}$$

It should be noted that essentially any submodular function of the form (17), with a sequence $(f_k)_k$ that is strictly increasing can be considered as a possible alternative to the OSCAR penalty, since it provides, like the latter, a norm whose core set contains all the subsets of $\{1,\dots,d\}$ and therefore has sharp faces of dimension l for all groups of size of coefficients with equal amplitude. Moreover for any such norm the proximal operator is computed efficiently, as shown already by Figueiredo and Nowak (2014).

In particular, the algorithm proposed by Zhong and Kwok (2012) to compute the proximal operator of the OSCAR penalty is a special instance of the algorithm proposed above.

8 Statistical analysis for submodular functions

In this section, we show that two classical theoretical results that can be proved for the Lasso and more generally for problems regularized by *decomposable norms* (Negahban et al., 2012), can be extended to the family of norms we considered in this paper when the associated function F is submodular. Namely, if the data is generated from a sparse linear model, it is possible to show that (a) under a generalization of the usual *irrepresentability condition*, the smallest stable subset containing the true support is identified with high probability for n sufficiently large, (b) under a generalization of the *restricted eigenvalue condition* the estimator is consistent in prediction error with so-called *fast rates* of convergence.

8.1 Weak and local decomposability of the norm for submodular functions.

The work of Negahban et al. (2012) has shown that when a norm is *decomposable with respect to a pair of subspaces* A and B, meaning that for all $\alpha \in A$ and $\beta \in B^\perp$ we have $\Omega(\alpha+\beta) = \Omega(\alpha)+\Omega(\beta)$, then common proof schemes allows to (a) show support recovery results and (b) fast rates of convergence in prediction error. For the norms we are considering, this type of assumption would be too strong.

However, based on a notion of weak decomposability Bach (2010), tackled the $p = \infty$. case. Weak decomposability was also proposed in van de Geer (2014), who obtained sparsity oracle inequalities based on an analysis that is similar to the one we develop, and applied it in particular to the norms proposed in Micchelli et al. (2013).

For the norms we consider, we use the notions of *weak* and *local decomposability* with decompositions that involve Ω_J and Ω^J, that are respectively the norms associated with the *restriction* and the *contraction* of the submodular function F to or on the set J.

Concretely, let $c = \frac{\tilde{m}}{M}$ with $M = \max_{k \in V} F(\{k\})$ and

$$\tilde{m} = \min_{A,k} F(A \cup \{k\}) - F(A) \text{ s.t. } F(A \cup \{k\}) > F(A).$$

Then we have:

Proposition 6. *(Weak and local decomposability)*

Weak decomposability. *For any set J and any $w \in \mathbb{R}^d$, we have*

$$\Omega(w) \geq \Omega_J(w_J) + \Omega^J(w_{J^c}).$$

Local decomposability. *Let $K = Supp(w)$ and J the smallest stable set containing K, if $\|w_{J^c}\|_p \leq c^{1/p} \min_{i \in K} |w_i|$, then*

$$\Omega(w) = \Omega_J(w_J) + \Omega^J(w_{J^c}).$$

Note that when $p = \infty$, if $J = K$, the condition becomes $\min_{i \in J} |w_i| \geqslant \max_{i \in J^c} |w_i|$, and we recover exactly the corresponding result from Bach (2010).

This proposition shows that a sort of reverse triangular inequality involving the norms Ω, Ω_J and Ω^J always holds and that if there is a sufficiently large positive gap between the values of w on J and on its complement then Ω can be written as a separable function on J and J^c.

8.2 Theoretical analysis for submodular functions

In this section, we consider a fixed design matrix $X \in \mathbb{R}^{n \times p}$ and $y \in \mathbb{R}^n$ a vector of random responses. Given $\lambda > 0$, we define $\hat{w}$ as a minimizer of the regularized least-squares cost:

$$\min_{w \in \mathbb{R}^d} \tfrac{1}{2n} \|y - Xw\|_2^2 + \lambda \Omega(w). \tag{19}$$

We study the sparsity-inducing properties of solutions of (19), i.e., we determine which patterns are allowed and which sufficient conditions lead to correct estimation.

We assume that the linear model is well-specified and extend results from Zhao and Yu (2006) for sufficient support recovery conditions and from Negahban et al. (2012) for estimation consistency, which were already derived by Bach (2010) for $p = \infty$. The following propositions allow us to retrieve and extend well-known results for the ℓ_1-norm.

Denote by ρ the following constant:

$$\rho = \min_{A \subset B, F(B) > F(A)} \frac{F(B) - F(A)}{F(B \backslash A)} \in (0, 1].$$

The following proposition extends results based on support recovery conditions (Zhao and Yu, 2006):

Proposition 7 (Support recovery). *Assume that $y = Xw^* + \sigma\varepsilon$, where ε is a standard multivariate normal vector. Let $Q = \frac{1}{n} X^\top X \in \mathbb{R}^{d \times d}$. Denote by J the smallest stable set containing the support $Supp(w^*)$ of w^*. Define $\nu = \min_{j, w_j^* \neq 0} |w_j^*| > 0$ and assume $\kappa = \lambda_{\min}(Q_{JJ}) > 0$.*

If the following ***generalized Irrepresentability Condition*** *holds:*

$$\exists \eta > 0, \qquad (\Omega^J)^* \Big(\big(\Omega_J(Q_{JJ}^{-1} Q_{Jj}) \big)_{j \in J^c} \Big) \leqslant 1 - \eta,$$

then, if $\lambda \leqslant \frac{\kappa \nu}{2|J|^{1/p} F(J)^{1-1/p}}$, the minimizer $\hat{w}$ is unique and has support equal to J, with probability larger than $1 - 3\,\mathbb{P}\big(\Omega^*(z) > \frac{\lambda \eta \rho \sqrt{n}}{2\sigma}\big)$, where z is a multivariate normal with covariance matrix Q.

In terms of prediction error the next proposition extends results based on restricted eigenvalue conditions (see, e.g. Negahban et al., 2012).

Proposition 8 (Consistency). *Assume that $y = Xw^* + \sigma\varepsilon$, where ε is a standard multivariate normal vector. Let $Q = \frac{1}{n}X^\top X \in \mathbb{R}^{d\times d}$. Denote by J the smallest stable set containing the support Supp(w^*) of w^*.*

*If the following Ω_J-**Restricted Eigenvalue condition** holds:*

$$\forall \Delta \in \mathbb{R}^d, \qquad \big(\,\Omega^J(\Delta_{J^c}) \leqslant 3\Omega_J(\Delta_J)\,\big) \quad \Rightarrow \quad \big(\,\Delta^\top Q \Delta \geqslant \kappa\,\Omega_J(\Delta_J)^2\,\big),$$

then we have

$$\Omega(\hat{w} - w^*) \leqslant \frac{24^2\lambda}{\kappa\rho^2} \qquad \text{and} \qquad \frac{1}{n}\|X\hat{w} - Xw^*\|_2^2 \leqslant \frac{36\lambda^2}{\kappa\rho^2},$$

with probability larger than $1 - \mathbb{P}\big(\Omega^*(z) > \frac{\lambda\rho\sqrt{n}}{2\sigma}\big)$ where z is a multivariate normal with covariance matrix Q.

The concentration of the values of $\Omega^*(z)$ for z is a multivariate normal with covariance matrix Q can be controlled via the following result that implies that if λ is larger then a constant times $\sqrt{\log|\mathcal{D}_F|}$, then the probability in the proposition is close to one. We thus recover known results for the Lasso (where $|\mathcal{D}_F| = d$) and the group Lasso (Negahban and Wainwright, 2008).

Proposition 9. *Let z be a normal variable with covariance matrix Q that has unit diagonal. Let $\mathcal{D}_F$ be the set of stable inseparable sets. Then*

$$\mathbb{P}\bigg(\Omega^*(z) \geqslant 4\sqrt{q\log(2|\mathcal{D}_F|)}\max_{A\in\mathcal{D}_F}\frac{|A|^{1/q}}{F(A)^{1/q}} + u\max_{A\in\mathcal{D}_F}\frac{|A|^{(1/q-1/2)_+}}{F(A)^{1/q}}\bigg) \leqslant e^{-u^2/2}. \tag{20}$$

9 Experiments

We illustrate the use of the theory presented in this paper by an application to the estimation of the parameter vector of linear least-squares regression, when this parameter vector is either supported on an interval on or a rectangular region in a two dimensional grid. In particular, we compare the performance on synthetic data of the estimators obtained, using different norms either classical or particularly tailored to the problem considered, both in terms of error in support estimation in Hamming distance and in ℓ_2-error.

9.1 Setting

To illustrate the results presented in this paper we consider the problem of estimating the support of a parameter vector $w \in \mathbb{R}^d$, when its support is assumed either

(i) to form an *interval* in $[\![1, d]\!]$, or

(ii) to form a *rectangle* $[\![k_{\min}, k_{\max}]\!] \times [\![k'_{\min}, k'_{\max}]\!] \subset [\![1, d_1]\!] \times [\![1, d_2]\!]$, with $d = d_1 d_2$.

These two settings were considered by Jenatton et al. (2011a). These authors showed that, for both types of supports, it was possible to construct an ℓ_1/ℓ_2-norm with overlap based on a well-chosen collection of overlapping groups, so that the obtained estimators almost surely have a support of

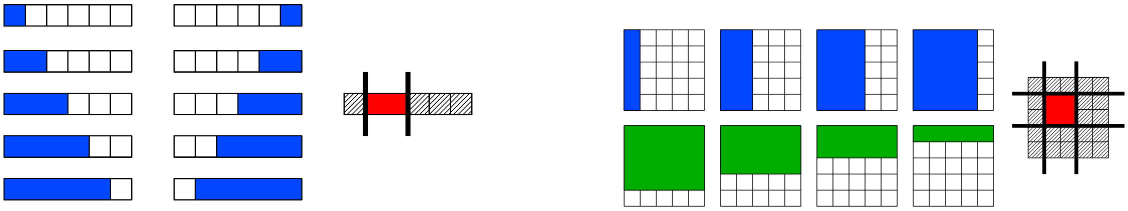


Figure 8: Set $\mathcal{G}$ of overlapping groups defining the norm proposed by Jenatton et al. (2011a) (set in blue or green and their complements) and an example of corresponding induced sparsity patterns (in red), respectively for interval patterns in 1D (left) and for rectangular patterns in 2D (right).

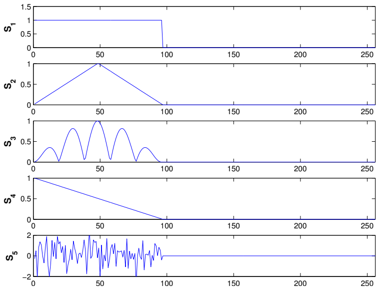


Figure 9: Examples of the shape of the signals used to define the amplitude of the coefficients of w on the support. Each plot represents the value of w_i as a function of i. The first (w constant on the support), third ($w_i = g(c\,i)$ with $g : x \mapsto |\sin(x)\sin(5x)|$) and last signal ($w_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$) are the ones used in reported results.

the correct form. Specifically, it was shown in Jenatton et al. (2011a) that norms of the form $w \mapsto \sum_{B\in\mathcal{G}} \|w_B\|_2$ induce sparsity patterns that are exactly intervals of $V = \{1,\dots,p\}$ if

$$\mathcal{G} = \big\{[1,k] \mid 1 \le k \le p\big\} \cup \big\{[k,p] \mid 1 \le k \le p\big\},$$

and induce rectangular supports on $V = V_1 \times V_2$ with $V_1 := \{1,\dots,p_1\}$ and $V_2 := \{1,\dots,p_2\}$ if

$$\begin{aligned}\mathcal{G} \quad = \quad & \big\{[\![1,k]\!] \times V_2 \mid 1 \le k \le p_1\big\} \cup \big\{[\![k,p_1]\!] \times V_2 \mid 1 \le k \le p_1\big\} \\ & \cup \big\{V_1 \times [\![1,k]\!] \mid 1 \le k \le p_2\big\} \cup \big\{V_1 \times [\![k,p_2]\!]\} \mid 1 \le k \le p_2\big\}.\end{aligned}$$

These sets of groups are illustrated on Figure 8, and, for the first case, the set $\mathcal{G}$ has already discussed in Example 6 to define a modified range function which is submodular.

Moreover, the authors showed that with a weighting scheme introduced *inside* the groups and leading to a norm of the form $w \mapsto \sum_{B\in\mathcal{G}} \|w_B \circ d^B\|$, where $\circ$ denotes the Hadamard product and $d^B \in \mathbb{R}^d_+$ is a certain vector of weights designed specifically for these case[9] it is possible to obtain compelling empirical results in terms of support recovery, especially in the 1D case.

Interval supports. From the point of view of our work, that is, approaching the problem in terms of combinatorial functions, for supports constrained to be intervals, it is natural to consider the range function as a possible form of penalty: $F_0(A) := \mathrm{range}(A) = i_{\max}(A) - i_{\min}(A) + 1$. Indeed the range function assigns the same penalty to sets with the same range, regardless of whether these sets are connected or have "holes"; this clearly favors intervals since they are exactly the sets with

[9] We refer the reader to the paper for the details.

the largest support for a given value of the penalty. Unfortunately, as discussed in the Example 2 of Section 2.2, the combinatorial lower envelope of the range function is $A \mapsto |A|$, the cardinality function, which implies that $\Omega_p^{F_0}$ is just the ℓ_1-norm: in this case, the structure implicitly encoded in F_0 is lost through the convex relaxation.

However, as mentioned by Bach (2010) and discussed in Example 6 the function F_r defined by $F_r(A) = d - 1 + \mathrm{range}(A)$ for $A \neq \varnothing$ and $F(\varnothing) = 0$ is submodular, which means that $\Omega_p^{F_r}$ is a tight relaxation and that regularizing with it leads to tractable convex optimization problems.

Rectangular supports. For the case of rectangles on the grid, a good candidate is the function F_2 with $F_2(A) = F_r(\Pi_1(A)) + F_r(\Pi_2(A))$ with $\Pi_i(A)$ the projection of the set A along the ith axis of the grid.

This makes of $\Omega_p^{F_r}$ and $\Omega_p^{F_2}$ two good candidates to estimate a vector w whose support matches respectively the two described a priori.

9.2 Methodology

We consider a simple regression setting in which $w \in \mathbb{R}^d$ is a vector such that $\mathrm{Supp}(w)$ is either an interval on $[1, d]$ or a rectangle on a fixed $2D$ grid. We draw the design matrix $X \in \mathbb{R}^{n \times d}$ and a noise vector $\epsilon \in \mathbb{R}^n$ both with i.i.d. standard Gaussian entries and compute $y = Xw + \epsilon$. We then solve problem (19), with Ω chosen in turn to be the ℓ_1-norm (Lasso), the elastic net, the norms Ω_p^F for $p \in \{2, \infty\}$ and F chosen to be F_r or F_2 in $1D$ and $2D$ respectively; we consider also the overlapping ℓ_1/ℓ_2-norm proposed by Jenatton et al. (2011a) and the weighted overlapping ℓ_1/ℓ_2-norm proposed by the same authors, i.e., $\Omega(w) = \sum_{B \in \mathcal{G}} \|w_B \circ d^B\|_2$ with the same notations as before[10].

We assess the estimators obtained through the different regularizers both in terms of support recovery and in terms of mean-squared error in the following way: assuming that held out data permits to choose an optimal point on the regularization path obtained with each norm, we determine along each such path, the solution which either has a support with minimal Hamming distance to the true support or the solution which as the best ℓ_2 distance, and we report the corresponding distances as a function the sample size on Figures 10 and 11 respectively for the 1D and the 2D case.

Finally, we assess the incidence of the fluctuation in amplitude of the coefficients in the vector w generating the data: we consider different cases among which:

(i) the case where w has a constant value on the support,

(ii) the case where w_i varies as a modulated cosine, with $w_i = g(c \cdot i)$ for c a constant scaling and $g : x \mapsto |\cos(x)\cos(5x)|$

(iii) the case where w_i is drawn i.i.d. from a standard normal distribution.

These cases (and two others for which we do not report results) are illustrated on Figure 9.

9.3 Results

Results reported for the Hamming distances in the left columns of Figures 10 and 11 show that the norms $\Omega_2^{F_r}$ and $\Omega_2^{F_2}$ perform quite well for support recovery overall and tend to outperform

[10] Note that we do not need to compare with an $\ell_i nfty$ counterpart of the unweighted norm considered by Jenatton et al. (2011a) since for $p = \infty$ the unweighted ℓ_1/ℓ_∞ norm defined with the same collection $\mathcal{G}$ is exactly the norm $\Omega_\infty^{F_r}$: this follows from the form of F_r as defined in Example 6 and the preceding discussion.

significantly their ℓ_∞ counterpart in most cases. In 1D, several norms achieve reasonably small Hamming distance, including the ℓ_1-norm, the norm $\Omega_2^{F_r}$ and the weighted overlapping ℓ_1/ℓ_2-norm although the latter clearly dominates for small values of n.

In 2D, $\Omega_2^{F_2}$ leads clearly to smaller Hamming distances than other norms for the larger values of n, while is outperformed by the ℓ_1-norm for small sample sizes. It should be noted that neither $\Omega_\infty^{F_2}$ nor the weighted overlapping ℓ_1/ℓ_2-norm that performed so well in 1D achieve good results.

The performance of the ℓ_2-relaxation tends to be comparatively better when the vector of parameter w has entries that vary a lot, especially when compared to the ℓ_∞-relaxation. Indeed, the choice of the value of p for the relaxation can be interpreted as encoding a prior on the joint distribution of the amplitudes of the w_i: as discussed before, and as illustrated in Bach (2010) the unit balls for the ℓ_∞ relaxations display additional "edges and corners" that lead to estimates with clustered values of $|w_i|$, corresponding to an priori that many entries in w have identical amplitudes. More generally, large values of p correspond to the prior that the amplitude varies little while their vary more significantly for small p.

The effect of this other type of a priori encoded in the regularization is visible when considering the performance in terms of ℓ_2 error. Overall, both in 1D and 2D all methods perform similarly in ℓ_2-error, except that when w is constant on the support, the ℓ_∞-relaxations $\Omega_\infty^{F_r}$ and $\Omega_\infty^{F_2}$ perform significantly better, and this is the case most likely because the additional "corners" of these norms induce some pooling of the estimates of the value of the w_i, which improves their estimation. By contrast it can be noted that when w is far from constant the ℓ_∞-relaxations tend to have slightly larger least-square errors, while, on contrary, the ℓ_1-regularisation tends to be among the better performing methods.

10 Conclusion

We proposed a family of convex norms defined as relaxations of penalizations that combine a combinatorial set-function with an ℓ_p-norm. Our formulation allows to recover in a principled way a number of sparsity inducing regularizations that have appeared in the literature such as the ℓ_1-norm, the group Lasso, the exclusive Lasso, the k-support norm, the OWL penalties (including OSCAR and SLOPE penalties), that are all specific instances. In addition, this formulation establishes that the latent group Lasso is the tightest relaxation of block-coding penalties. We discuss the use of the proposed formulation for the construction of relaxation for different hierarchical penalties on a DAG, and recover both new and existing norms.

There are several directions for future research. First, it would be of interest to determine for which combinatorial functions beyond submodular ones, efficient algorithms and consistency results can be established. Then a sharper analysis of the relative performance of the estimators using different levels of a priori would be needed to answer question such as: When is using a structured a priori likely to yield better estimators? When could it degrade the performance? What is the relation to the performance of an oracle given a specified structured a priori?

Acknowledgements

The authors acknowledge funding from the European Research Council grant SIERRA (project 239993), and would like to thank Rodolphe Jenatton and Julien Mairal for stimulating discussions. Guillaume Obozinski acknowledges funding from the ANR CHORUS research grant 13-MONU-0005-10.

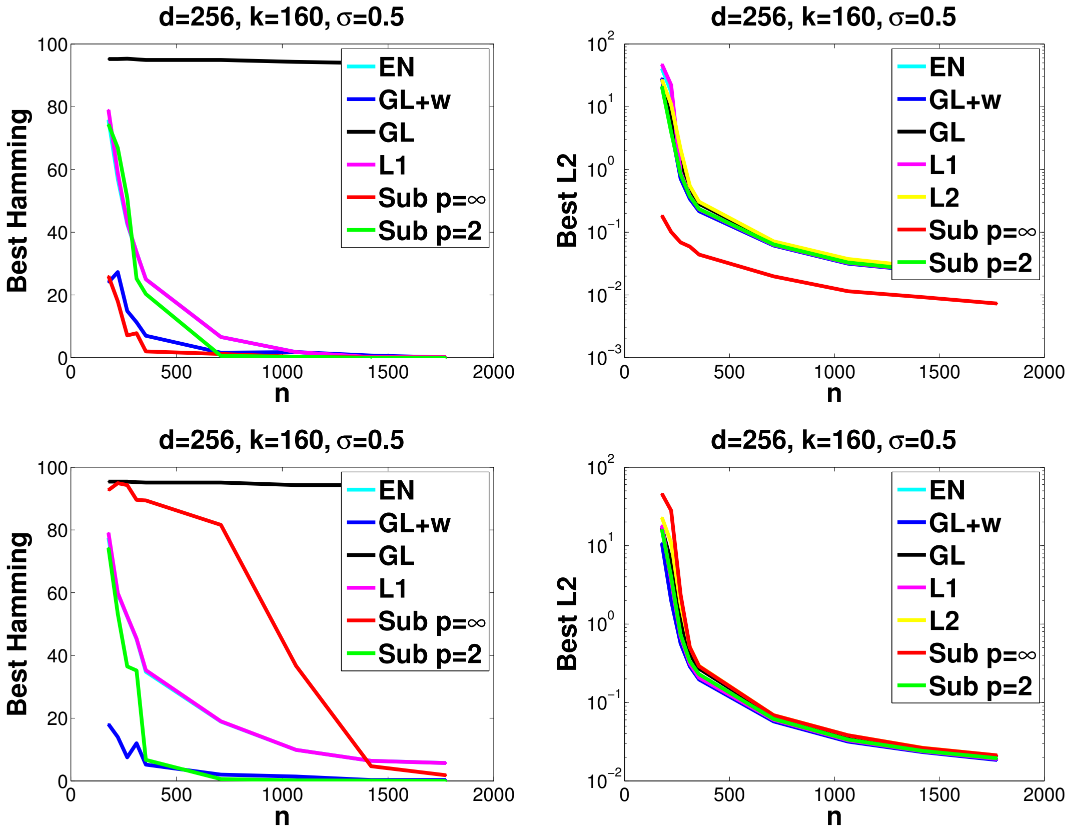


Figure 10: Best Hamming distance (left column) and best least square error (right column) to the true parameter vector w^*, among all vectors along the regularization path of a least square regression regularized with a given norm, for different patterns of values of w^*. The different regularizers compared include the Lasso (L1), Ridge (L2), the elastic net (EN), the unweighted (GL) and weighted (GL+w) ℓ_1/ℓ_2 regularizations proposed by Jenatton et al. (2011a), the norms Ω_2^F (Sub $p = 2$) and Ω_∞^F (Sub $p = \infty$) for a specified function F. (first row) Constant signal supported on an interval, with an a priori encoded by the combinatorial function $F : A \mapsto d - 1 + \text{range}(A)$. (second row) Same setting with a signal w^* supported by an interval consisting of coefficients w_i^* drawn from a standard Gaussian distribution. In each case, the dimension is $d = 256$, the size of the true support is $k = 160$, the noise level is $\sigma = 0.5$ and signal amplitude $\|w\|_\infty = 1$.

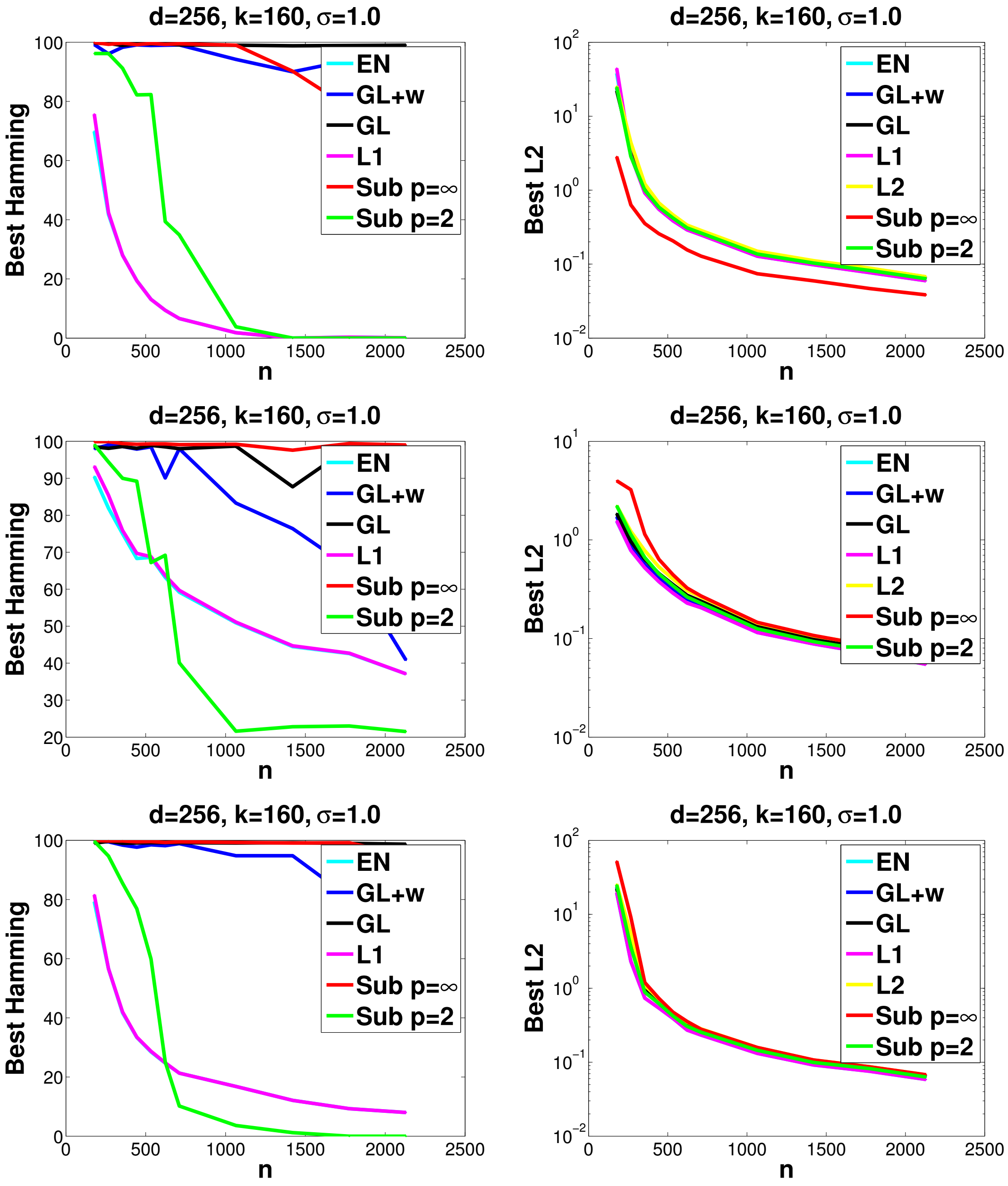


Figure 11: Best Hamming distance (left column) and best least square error (right column) to the true parameter vector w^*, among all vectors along the regularization path of a least square regression regularized with a given norm, for different patterns of values of w^*. The regularizations compared include the Lasso (L1), Ridge (L2), the elastic net (EN), the unweighted (GL) and weighted (GL+w) ℓ_1/ℓ_2 regularizations proposed by Jenatton et al. (2011a), the norms Ω_2^F (Sub $p = 2$) and Ω_∞^F (Sub $p = \infty$) for a specified function F. Parameter vectors w^* considered here have coefficients that are supported by a rectangle on a grid with size $d_1 \times d_2$ with $d = d_1 d_2$. (first row) Constant signal supported on a rectangle with an a priori encoded by the combinatorial function $F : A \mapsto d_1 + d_2 - 4 + \text{range}(\Pi_1(A)) + \text{range}(\Pi_2(A))$. (second row) Same setting with coefficients of w on the support given as $w^*_{i_1 i_2} = g(c\, i_1)g(c\, i_2)$ for c a positive constant and $g : x \mapsto |\cos(x)\cos(5x)|$. (third row) Same setting with coefficients $w^*_{i_1 i_2}$ drawn from a standard Gaussian distribution. In each case, the dimension is $d = 256$, the size of the true support is $k = 160$, the noise level is $\sigma = 1$ and signal amplitude $\|w\|_\infty = 1$.

References

Argyriou, A., Foygel, R., and Srebro, N. (2012). Sparse prediction with the k-support norm. In *Advances in Neural Information Processing Systems 25*, pages 1466–1474.

Bach, F. (2010). Structured sparsity-inducing norms through submodular functions. In *Adv. NIPS*.

Bach, F. (2013). Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6(2):145–373.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. (2012). Optimization with sparsity-inducing penalties. *Foundation and Trends in Machine Learning*, 1(4):1–106.

Baraniuk, R., Cevher, V., Duarte, M., and Hegde, C. (2010). Model-based compressive sensing. *IEEE Trans. Inf. Theory,*, 56(4):1982–2001.

Barlow, R. and Brunk, H. (1972). The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147.

Bauer, F., Stoer, J., and Witzgall, C. (1961). Absolute and monotonic norms. *Numerische Mathematik*, 3(1):257–264.

Best, M. and Chakravarti, N. (1990). Active set algorithms for isotonic regression; a unifying framework. *Mathematical Programming*, 47(1):425–439.

Bickel, P., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732.

Bien, J., Taylor, J., Tibshirani, R., et al. (2013). A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141.

Bogdan, M., van den Berg, E., Sabatti, C., Su, W., and Candès, E. J. (2015). SLOPE: adaptive variable selection via convex optimization. *Annals of Applied Statistics*, 9(3):1103–1140.

Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, 64(1):115–123.

Boyd, S. P. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Chambolle, A. and Darbon, J. (2009). On total variation minimization and surface evolution using parametric maximum flows. *International Journal of Computer Vision*, 84(3):288–307.

Chandrasekaran, V., Recht, B., Parrilo, P. A., and Willsky, A. S. (2012). The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849.

Dinkelbach, W. (1967). On nonlinear fractional programming. *Management Science*, 13(7):492–498.

Edmonds, J. (2003). Submodular functions, matroids, and certain polyhedra. In *Combinatorial optimization - Eureka, you shrink!*, pages 11–26. Springer.

Figueiredo, M. and Nowak, R. D. (2014). Sparse estimation with strongly correlated variables using ordered weighted ℓ_1 regularization. Technical Report 1409.4005, arXiv.

Fujishige, S. (2005). *Submodular Functions and Optimization*. Elsevier.

Gallo, G., Grigoriadis, M. D., and Tarjan, R. E. (1989). A fast parametric maximum flow algorithm and applications. *SIAM Journal on Computing*, 18(1):30–55.

Groenevelt, H. (1991). Two algorithms for maximizing a separable concave function over a polymatroid feasible region. *Eur. J Oper. Res.*, 54(2):227–236.

He, L. and Carin, L. (2009). Exploiting structure in wavelet-based Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 57:3488–3497.

Hochbaum, D. S. and Hong, S.-P. (1995). About strongly polynomial time algorithms for quadratic optimization over submodular constraints. *Mathematical programming*, 69(1-3):269–309.

Huang, J., Zhang, T., and Metaxas, D. (2011). Learning with structured sparsity. *The JMLR*, 12:3371–3412.

Jacob, L., Obozinski, G., and Vert, J. (2009). Group lasso with overlap and graph lasso. In *ICML*.

Jenatton, R., Audibert, J., and Bach, F. (2011a). Structured variable selection with sparsity-inducing norms. *JMLR*, 12:2777–2824.

Jenatton, R., Mairal, J., Obozinski, G., and Bach, F. (2010). Proximal methods for sparse hierarchical dictionary learning. In *Proc. ICML*.

Jenatton, R., Mairal, J., Obozinski, G., and Bach, F. (2011b). Proximal methods for hierarchical sparse coding. *JMLR*, 12:2297–2334.

Kim, S. and Xing, E. P. (2010). Tree-guided group lasso for multi-task regression with structured sparsity. In *Proc. ICML*.

Lovász, L. (1975). On the ratio of optimal integral and fractional covers. *Discr. Math.*, 13(4):383–390.

Luss, R. and Rosset, S. (2014). Generalized isotonic regression. *Journal of Computational and Graphical Statistics*, 23(1):192–210.

Mairal, J., Jenatton, R., Obozinski, G., and Bach, F. (2011). Convex and network flow optimization for structured sparsity. *JMLR*, 12:2681–2720.

McDonald, A. M., Pontil, M., and Stamos, D. (2015). New perspectives on k-support and cluster norms. *arXiv preprint arXiv:1512.08204*.

Micchelli, C. A., Morales, J. M., and Pontil, M. (2013). Regularizers for structured sparsity. *Advances in Computational Mathematics*, 38(3):455–489.

Negahban, S., Ravikumar, P., Wainwright, M., and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557.

Negahban, S. and Wainwright, M. J. (2008). Joint support recovery under high-dimensional scaling: Benefits and perils of ℓ_1-ℓ_∞-regularization. In *Adv. NIPS*.

Obozinski, G., Jacob, L., and Vert, J.-P. (2011). Group Lasso with overlaps: the Latent Group Lasso approach. *preprint HAL - inria-00628498*.

Pardalos, P. M. and Xue, G. (1999). Algorithms for a class of isotonic regression problems. *Algorithmica*, 23(3):211–222.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5):465–471.

Rockafellar, R. (1970). *Convex Analysis.* Princeton University Press.

Stewart, G. W. and Sun, J. (1990). *Matrix Perturbation Theory.* Academic Press.

Stout, Q. F. (2013). Isotonic regression via partitioning. *Algorithmica*, 66(1):93–112.

van de Geer, S. (2014). Weakly decomposable regularization penalties and structured sparsity. *Scandinavian Journal of Statistics*, 41(1):72–86.

Yan, X. and Bien, J. (2015). Hierarchical sparse modeling: A choice of two regularizers. *arXiv preprint arXiv:1512.01631*.

Yuan, M., Joseph, V. R., and Zou, H. (2009). Structured variable selection and estimation. *The Annals of Applied Statistics*, 3(4):1738–1757.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B*, 68(1):49–67.

Zhao, P., Rocha, G., and Yu, B. (2009a). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, pages 3468–3497.

Zhao, P., Rocha, G., and Yu, B. (2009b). Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *JMLR*, 7:2541–2563.

Zhong, L. W. and Kwok, J. T. (2012). Efficient sparse modeling with automatic feature grouping. *Neural Networks and Learning Systems, IEEE Transactions on*, 23(9):1436–1447.

Zhou, Y., Jin, R., and Hoi, S. C. (2010). Exclusive lasso for multi-task feature selection. In *AISTATS*.

A Form of primal norm

We provide here a proof of lemma 7 which we first recall:

Lemma (7)**.** *Ω_p and Ω_p^* are dual to each other.*

Proof. Let ω_p^A be the function[11] defined by $\omega_p^A(w) = F(A)^{1/q}\,\|w_A\|_p\,\iota_{\{v\,|\,Supp(v)\subset A\}}(w)$ with ι_B the indicator function taking the value 0 on B and ∞ on B^c. Let K_p^A be the set $K_p^A = \{s \mid \|s_A\|_q^q \leq F(A)\}$. By construction, ω_p^A is the *support function* of K_p^A (see Rockafellar, 1970, sec.13), i.e. $\omega_p^A(w) = \max_{s\in K_p^A} w^\top s$. By construction we have $\{s \mid \Omega_p^*(s) \leq 1\} = \cap_{A\subset V} K_p^A$. But this implies that $\iota_{\{s|\Omega_p^*(s)\leq 1\}} = \sum_{A\subset V} \iota_{K_p^A}$. Finally, by definition of Fenchel-Legendre duality,

$$\Omega_p(w) = \max_{w\in\mathbb{R}^d} w^\top s - \sum_{A\subset V} \iota_{K_p^A}(s),$$

or in words Ω_p is the Fenchel-Legendre dual to the sum of the indicator functions $\iota_{K_p^A}$. But since the Fenchel-Legendre dual of a sum of functions is the *infimal convolution* of the duals of these functions (see Rockafellar, 1970, Thm. 16.4 and Corr. 16.4.1, pp. 145-146), and since by definition of a support function $\left(\iota_{K_p^A}\right)^* = \omega_p^A$, then Ω_p is the *infimal convolution* of the functions ω_p^A, i.e.

$$\Omega_p(w) = \inf_{(v^A\in\mathbb{R}^d)_{A\subset V}} \sum_{A\subset V} \omega_p^A(v^A) \quad \text{s.t.} \quad w = \sum_{A\subset V} v^A,$$

which is equivalent to formulation (3). See Obozinski et al. (2011) for a more elementary proof of this result. □

[11] Or gauge function to be more precise.

B Relation between different norms

Proposition 10. *The functions*

$$\Omega : w \mapsto \sum_{B\in\mathcal{G}} d_B^{1/q}\|w_B\| \quad \textit{and} \quad \Omega^\circ : s \mapsto \inf_{z\in\mathcal{V}(s,\mathcal{G})} \max_{B\in\mathcal{G}} \frac{\|z_B\|_q}{d_B^{1/q}}$$

are norms and polar to each other.

Proof. It is clear that Ω is a norm since $\bigcup_{B\in\mathcal{G}} B = \{1,\dots,d\}$. Then Ω° is a convex function because

$$\Omega^\circ(s) = \inf_{z\in\mathcal{V}(s,\mathcal{G})} \psi(s,z) \quad \text{with} \quad \psi : (s,z) \mapsto \max_{B\in\mathcal{G}} d_B^{-1/q}\|z_B\|_q + \iota_{\{s=\sum_{B\in\mathcal{G}} z^B\}},$$

and $\psi(s,z)$ is a proper, l.s.c., jointly convex function in (s,z). Moreover, since Ω° is also symmetric, homogeneous, everywhere finite and satisfies $(\Omega^\circ(s)=0) \Rightarrow (s=0)$, then Ω° is a norm. Finally,

$$\max_{s:\Omega^\circ(s)\le 1} \langle w,s\rangle = \max_z \Big\{ \sum_{B\in\mathcal{G}} \langle w, z^B\rangle \mid \forall B\in\mathcal{G},\ \|z^B\|_q \le d_B^{1/q},\ z^B_{B^c}=0 \Big\} = \sum_{B\in\mathcal{G}} d_B^{1/q}\|w_B\|_p,$$

which shows that Ω is the Fenchel conjugate of $s \mapsto \iota_{\{\Omega^\circ(\cdot)\le 1\}}$. Since Ω and Ω° are norms this establishes that they are polar to each other. □

C Example of the Exclusive Lasso

We showed in Section 4.3 that the ℓ_p exclusive Lasso norm, also called ℓ_p/ℓ_1-norm, defined by the mapping $w \mapsto \Big(\sum_{G\in\mathcal{G}} \|w_G\|_1^p\Big)^{1/p}$, for some partition $\mathcal{G}$, is a norm Ω_p^F providing the ℓ_p tightest convex p.h. relaxation in the sense defined in this paper of a certain combinatorial function F. A computation of the lower combinatorial envelope of that function F yields the function $F_- : A \mapsto \max_{G\in\mathcal{G}} |A\cap G|$.

This last function is also a natural combinatorial function to consider and by the properties of a LCE it has the same convex relaxation. It should be noted that it is however less obvious to show directly that $\Omega_p^{F_-}$ is the ℓ_p/ℓ_1 norm...

We thus show a direct proof of that result since it illustrates how the results on LCE and UCE can be used to analyze norms and derive such results.

Lemma 12. *Let $\mathcal{G} = \{G_1,\dots,G_k\}$ be a partition of V. For $F : A \mapsto \max_{G\in\mathcal{G}} |A\cap G|$, we have $\Omega_\infty^F(w) = \max_{G\in\mathcal{G}} \|w_G\|_1$.*

Proof. Consider the function $f : w \mapsto \max_{G\in\mathcal{G}} \|w_G\|_1$ and the set function $F_0 : A \mapsto f(1_A)$. We have $F_0(A) = \max_{G\in\mathcal{G}} \|1_{A\cap G}\|_1 = F(A)$. But by Lemma 8, this implies that $f(w) \le \Omega_\infty^F(w)$ since $f = f(|\cdot|)$ is convex positively homogeneous and coordinatewise non-decreasing on $\mathbb{R}_+^d$. We could remark first that since $F(A) = f(1_A) \le \Omega_\infty^F(1_A) \le F(A)$, this shows that $F = F_-$ is a lower combinatorial envelope. Now note that

$$(\Omega_\infty^F)^*(s) = \max_{A\subset V, A\neq\varnothing} \min_{G\in\mathcal{G}} \frac{\|s_A\|_1}{|A\cap G|} \ge \max_{A\subset V,\ |A\cap G|=1,\ G\in\mathcal{G}} \|s_A\|_1 = \sum_{G\in\mathcal{G}} \max_{i\in G} |s_i| = \sum_{G\in\mathcal{G}} \|s_G\|_\infty.$$

This shows that $(\Omega_\infty^F)^*(s) \ge \sum_{G\in\mathcal{G}} \|s_G\|_\infty$, which implies for dual norms that $\Omega_\infty^F(w) \le f(w)$. Finally, since we showed above the opposite inequality $\Omega_\infty^F = f$ which shows the result. □

D Properties of the norm Ω_p^F when F is submodular

In this section, we first derive upper bounds and lower bounds for our norms, as well as a local formulation as a sum of ℓ_p-norms on subsets of indices.

D.1 Some important inequalities.

We now derive inequalities which will be useful later in the theoretical analysis. By definition, the dual norm satisfies the following inequalities:

$$\frac{\|s\|_\infty}{M^{\frac{1}{q}}} \le \max_{k\in V} \frac{\|s_{\{k\}}\|_q}{F(\{k\})^{\frac{1}{q}}} \le \Omega_p^*(s) = \max_{A\subset V, A\neq\varnothing} \frac{\|s_A\|_q}{F(A)^{\frac{1}{q}}} \le \frac{\|s\|_q}{\min_{A\subset V, A\neq\varnothing} F(A)^{\frac{1}{q}}} \le \frac{\|s\|_q}{m^{\frac{1}{q}}}, \qquad (21)$$

for $m = \min_{k\in V} F(\{k\})$ and $M = \max_{k\in V} F(\{k\})$. These inequalities imply immediately inequalities for Ω_p (and therefore for f since for $\eta \in \mathbb{R}_+^d,\ f(\eta) = \Omega_\infty(\eta)$):

$$m^{1/q}\|w\|_p \leqslant \Omega_p(w) \leqslant M^{1/q}\|w\|_1.$$

We also have $\Omega_p(w) \leqslant F(V)^{1/q}\|w\|_p$, using the following lower bound for the dual norm: $\Omega_p^*(s) \geqslant \frac{\|s\|_p}{F(V)^{1/q}}$.

Since by submodularity, we in fact have $M = \max_{A,k\notin A} F(A\cup\{k\}) - F(A)$, it makes sense to introduce $\tilde{m} = \min_{A,k,F(A\cup\{k\})>F(A)} F(A\cup\{k\}) - F(A) \le m$. Indeed, we consider in Section 8.2 the norm $\Omega_{p,J}$ (resp. Ω_p^J) associated with *restrictions* of F to J (resp. *contractions* of F on J) and it follows from the previous inequalities that for all $J \subset V$, we have:

$$\tilde{m}^{1/q}\|w\|_p \leqslant m^{1/q}\|w\|_p \leqslant \Omega_{p,J}(w) \leqslant M^{1/q}\|w\|_1 \qquad \text{and} \qquad \tilde{m}^{1/q}\|w\|_p \leqslant \Omega_p^J(w) \leqslant M^{1/q}\|w\|_1.$$

D.2 Some optimality conditions for η.

While exact necessary and sufficient conditions for η to be a solution of Eq. (8) would be tedious to formulate precisely, we provide three necessary and two sufficient conditions, which together characterize a non-trivial subset of the solutions, which will be useful in the subsequent analysis.

Proposition 11 (Optimality conditions for η)**.** *Let F be a non-increasing submodular function. Let $p > 1$ and $w \in \mathbb{R}^d$, $K = Supp(w)$ and J the smallest stable set containing K. Let $H(w)$ the set of minimizers of Eq. (8). Then,*

(a) the set $\{\eta_K,\ \eta \in H(w)\}$ is a singleton with strictly positive components, which we denote $\{\eta_K(w)\}$, i.e., Eq. (8) uniquely determines η_K.

(b) For all $\eta \in H(w)$, then $\eta_{J^c} = 0$.

(c) If $A_1 \cup \cdots \cup A_m$ are the ordered level sets of η_K, i.e., η is constant on each A_j and the values on A_j form a strictly decreasing sequence, then $F(A_1 \cup \cdots \cup A_j) - F(A_1 \cup \cdots \cup A_{j-1}) > 0$ and the value on A_j is equal to $\eta^{A_j}(w) = \frac{\|w_{A_j}\|_p}{[F(A_1\cup\cdots\cup A_j) - F(A_1\cup\cdots\cup A_{j-1})]^{1/p}}$.

(d) If η_K is equal to $\eta_K(w)$, $\max_{k\in J\backslash K}\eta_k \leqslant \min_{k\in K}\eta_k(w)$, and $\eta_{J^c} = 0$, then $\eta \in H(w)$.

(e) There exists $\eta \in H(w)$ such that $\frac{\min_{i\in K}|w_i|}{M^{1/p}} \leqslant \min_{j\in J}\eta_j \leqslant \max_{j\in J}\eta_j \leqslant \frac{\|w\|_p}{m^{1/p}}$.

Proof. (a) Since f is non-decreasing with respect to each of its argument, for any $\eta \in H(w)$, we have $\eta' \in H(w)$ for η' defined through $\eta'_K = \eta_K$ and $\eta_{K^c} = 0$. The set of values of η_K for $\eta \in H(w)$ is therefore the set of solutions problem (8) restricted to K. The latter problem has a unique solution as a consequence of the strict convexity on $\mathbb{R}_+^*$ of $\eta_j \mapsto \frac{|w_j|^p}{\eta_j^{p-1}}$.

(b) If there is $j \in J^c$ such that $\eta \in H(w)$ and $\eta_j \neq 0$, then (since $w_j = 0$) because f is non-decreasing with respect to each of its arguments, we may take η_j infinitesimally small and all other η_k for $k \in K^c$ equal to zero, and we have $f(\eta) = f_K(\eta_K(w)) + \eta_j[F(K \cup \{j\}) - F(K)]$. Since $F(K \cup \{j\}) - F(K) \geqslant F(J \cup \{j\}) - F(J) > 0$ (because J is stable), we have $f(\eta) > f_K(\eta_K(w))$, which is a contradiction.

(c) Given the ordered level sets, we have $f(\eta) = \sum_{j=1}^m \eta^{A_j}[F(A_1 \cup \dots \cup A_j) - F(A_1 \cup \dots \cup A_{j-1})]$, which leads to a closed-form expression $\eta^{A_j}(w) = \frac{\|w_{A_j}\|_p}{[F(A_1 \cup \dots \cup A_j) - F(A_1 \cup \dots \cup A_{j-1})]^{1/p}}$. If $F(A_1 \cup \dots \cup A_j) - F(A_1 \cup \dots \cup A_{j-1}) = 0$, since $\|w_{A_j}\|_p > 0$, we have η^{A_j} as large as possible, i.e., it has to be equal to $\eta^{A_{j-1}}$, thus it is not a possible ordered partition.

(d) With our particular choice for η, we have $\sum_{i \in V} \frac{1}{p} \frac{|w_i|^p}{\eta_i^{p-1}} + \frac{1}{q} f(\eta) = \Omega_K(w_K)$. Since we always have $\Omega(w) \geqslant \Omega_K(w_K)$, then η is optimal in Eq. (8).

(e) We take the largest elements from (d) and bounds the components of η_K using (c). □

Note that from property (c), we can explicit the value of the norm as:

$$\Omega_p(w) = \sum_{j=1}^{k} (F(A_1 \cup \ldots \cup A_j) - F(A_1 \cup \ldots \cup A_{j-1}))^{\frac{1}{q}} \|w_{A_j \setminus A_{j-1}}\|_p \tag{22}$$

$$= \Omega_{p,A_1}(w_{A_1}) + \sum_{j=2}^{k} \Omega_{p,A_j}^{A_{j-1}}(w_{A_j \setminus A_{j-1}}) \tag{23}$$

where $\Omega_{p,B}^A$ is the norm associated with the contraction on A of F restricted to B.

E Proof of Proposition 6 (Decomposability)

Concretely, let $c = \frac{\tilde{m}}{M}$ with $M = \max_{k \in V} F(\{k\})$ and

$$\tilde{m} = \min_{A,k} F(A \cup \{k\}) - F(A) \ \text{ s.t. } \ F(A \cup \{k\}) > F(A)$$

Proposition (6. Weak and local Decomposability)**.** *(a) For any set J and any $w \in \mathbb{R}^d$, we have*

$$\Omega(w) \geq \Omega_J(w_J) + \Omega^J(w_{J^c}).$$

(b) Assume that J is stable, and $\|w_{J^c}\|_p \leq c^{1/p} \min_{i \in J} |w_i|$, then $\Omega(w) = \Omega_J(w_J) + \Omega^J(w_{J^c})$.
(c) Assume that K is non stable and J is the smallest stable set containing K, and that $\|w_{J^c}\|_p \leq c^{1/p} \min_{i \in K} |w_i|$, then $\Omega(w) = \Omega_J(w_J) + \Omega^J(w_{J^c})$.

Proof. We first prove the first statement (a): If $\|s_{A \cap J}\|_p^p \leq F(A \cap J)$ and $\|s_{A \cap J^c}\|_p^p \leq F(A \cup J) - F(J)$ then by submodularity we have $\|s_A\|_p^p \leq F(A \cap J) + F(A \cup J) - F(J) \leq F(A)$. The canonical polyhedra associated with F_J and F^J are respectively defined by

$$\mathcal{P}(F_J) = \{s \in \mathbb{R}_+^d,\ \mathrm{Supp}(s) \subset J,\ s(A) \leq F(A),\ A \subset J\} \quad \text{and}$$
$$\mathcal{P}(F^J) = \{s \in \mathbb{R}_+^d,\ \mathrm{Supp}(s) \subset J^c,\ s(A) \leq F(A \cup J) - F(J)\}$$

Denoting $s^{\circ p} := (s_1^p, \ldots, s_d^p)$, we therefore have

$$\Omega(w) = \max_{\{s^{\circ p} \in \mathcal{P}_F\}} s^\top |w| \geq \max_{\{s_J^{\circ p} \in \mathcal{P}(F_J),\, s_{J^c}^{\circ p} \in \mathcal{P}(F^J)\}} s^\top |w| = \Omega_J(w_J) + \Omega^J(w_{J^c}).$$

In order to prove (b), we consider an optimal η_J for w_J and Ω_J and an optimal η_{J^c} for Ω^J. Because of our inequalities, and because we have assume that J is stable (so that the value m for Ω^J is indeed lower bounded by $\tilde{m}$), we have $\|\eta_{J^c}\|_\infty \leqslant \frac{\|w_{J^c}\|_p}{\tilde{m}^{1/p}}$. Moreover, we have $\min_{j \in J} \eta_j \geqslant \frac{\min_{i \in J} |w_i|}{M^{1/p}}$ (inequality proved in the main paper). Thus when concatenating η_J and η_{J^c} we obtain an optimal η for w (since then the Lovász extension decomposes as a sum of two terms), hence the desired result.

In order to prove (c), we simply notice that since $F(J) = F(K)$, the value of $\eta_{J \backslash K}$ is irrelevant (the variational formulation does not depend on it), and we may take it equal to the largest known possible value, i.e., one which is largest than $\frac{\min_{i \in J} |w_i|}{M^{1/p}}$, and the same reasoning than for (b) applies. □

Note that when $p = \infty$, the condition in (b) becomes $\min_{i \in J} |w_i| \geqslant \max_{i \in J^c} |w_i|$, and we recover exactly the corresponding result from Bach (2010).

F Algorithmic results

F.1 Proximal operator of the norm Ω_p^F and proof of Algorithm 2.

We provide in this section the decomposition algorithm presented as Algorithm 4 to compute the proximal operator of a norm Ω_p^F, for any value of $p \in (1, \infty]$, when F is submodular. Algorithm 2 is the particular instance of that algorithm where all steps are closed form and other simplifications can be made. Algorithm 4 is a particular instance of the decomposition algorithm for the optimization of a convex function over the canonical polyhedron[12] (see e.g. section 8.4 and 9.1 of Bach (2013)). Indeed, denoting $\psi_i(\kappa_i) = \min_{x_i \in \mathbb{R}} \frac{1}{2}(x_i - z_i)^2 + \lambda \kappa_i^{1/q} |x_i|$, the computation of the proximal operator amounts to solving in κ the problem

$$\max_{\kappa \in \mathcal{P}_F} \sum_{i \in V} \psi_i(\kappa_i).$$

Following the decomposition algorithm, one has to solve first

$$\begin{aligned}
& \max_{\kappa \in \mathbb{R}_+^d} \sum_{i \in V} \psi_i(\kappa_i) \quad \text{s.t.} \quad \sum_{i \in V} \kappa_i \leq F(V) \\
= \; & \min_{x \in \mathbb{R}^d} \max_{\kappa \in \mathbb{R}_+^d} \frac{1}{2}\|x - z\|_2^2 + \lambda \sum_{i \in V} \kappa_i^{1/q} |x_i| \quad \text{s.t.} \quad \sum_{i \in V} \kappa_i \leq F(V) \\
= \; & \min_{x \in \mathbb{R}^d} \frac{1}{2}\|x - z\|_2^2 + \lambda F(V)^{1/q} \|x\|_p,
\end{aligned}$$

where the last equation is obtained by solving the maximization problem in κ. Let x^* denote the solution to the above problem.

We consider first the case $p < \infty$.

If $x^* \neq 0$, then $\sum_i \kappa_i^{1/q} |x_i^*| = F(V) \|x^*\|_p$ so that we must have $\kappa_i = \frac{|x_i^*|^p}{\|x^*\|_p^p} F(V)$. If $x^* = 0$, then, given that $x_i^* = (z_i - \lambda \kappa_i^{1/q})_+$, we must also have $z_i - \lambda \kappa_i^{1/q} \leq 0$, which entails $\kappa_i \geq \big(\frac{z_i}{\lambda}\big)^q$. But if

[12] There are different variants of the decomposition algorithm using different constraint sets: the canonical polyhedron or the base polyhedron defined as $\mathcal{B}_F = \mathcal{P}_F \cap \{s \in \mathbb{R}^d \mid s(V) = F(V)\}$ (see Bach, 2013, Sec. 8.4 for details.)

$x^* = 0$, then we must have $\|z\|_q \leq \lambda F(V)^{1/q}$. So that if we set $\kappa_i = \frac{|z_i|^q}{\|z\|_q^q} F(V)$, then $\kappa(V) = F(V)$ and $\kappa_i \geq \left(\frac{z_i}{\lambda}\right)^q$.

In particular, when $p = 2$, we have $x^* = (\|z\|_2 - \lambda\sqrt{F(V)})_+ \frac{z}{\|z\|_2}$ if $z \neq 0$ and $x^* = 0$ else. And since $x^* \propto z$, then $\kappa_i = F(V)\frac{z_i^2}{\|z\|_2^2}$ is always a solution, which explains the simplification made in Algorithm 2.

When $p = \infty$, if $x^* \neq 0$, we have $\sum_i \kappa_i |x_i^*| = F(V)\|x^*\|_\infty$ so that with the constraint $\kappa(V) = F(V)$ we must have $\kappa_i = 1_{\{|x_i^*| = \|x^*\|_\infty\}} F(V)$. The case $x^* = 0$ is the same as for $p < \infty$.

Following the decomposition algorithm, one then has to find the minimizer of the submodular function $A \mapsto F(A) - \kappa(A)$. Then one needs to solve

$$\max_{\kappa_A \in \mathbb{R}_+^{|A|} \cap \mathcal{P}(F_A)} \sum_{i \in A} \psi_i(\kappa_i) \qquad \text{and} \qquad \max_{\kappa_{V\setminus A} \in \mathbb{R}_+^{|V\setminus A|} \cap \mathcal{P}(F^A)} \sum_{i \in V\setminus A} \psi_i(\kappa_i).$$

Using the expression of ψ_i and exchanging as above the minimization in w and the maximization in κ, one obtains directly that these two problems correspond respectively to the computation of the proximal operators of Ω^{F_A} on z_A and of the proximal operator of Ω^{F^A} on $z_{V\setminus A}$.

The decomposition algorithm used here is proved to be correct in section 8.4 of Bach (2013) under the assumption that $\kappa_i \mapsto \psi(\kappa_i)$ is a strictly convex function. The functions we consider here are not strongly convex, and in particular, as mentioned above the solution in κ is not unique in case $w^* = 0$. The proof of Bach (2013) however goes through using any solution of the maximization problem in κ.

Algorithm 4 Computation $x = \text{Prox}_{\lambda\Omega_p^F}(z)$

```
Require: $z \in \mathbb{R}^d$, $\lambda > 0$
Let $A = \{j \mid z_j \neq 0\}$
if $A \neq V$ then
    Set $x_A = \text{Prox}_{\lambda\Omega_p^{F_A}}(z_A)$
    Set $x_{A^c} = 0$
    return $x$ by concatenating $x_A$ and $x_{A^c}$
end if
Let $x = \text{argmin}_y \frac{1}{2}\|y - z\|_2^2 + \lambda F(V)^{\frac{1}{q}}\|y\|_p$
if $x \neq 0$ then
    Let $\kappa \in \mathbb{R}^d$ with $\kappa_i = \frac{|x_i|^p}{\|x\|_p^p} F(V)$ if $p < \infty$ and $\kappa_i = 1_{\{|x_i| = \|x\|_\infty\}} F(V)$ for $p = \infty$
else
    Let $\kappa \in \mathbb{R}^d$ with $\kappa_i = \frac{|z_i|^q}{\|z\|_q^q} F(V)$
end if
Find $A$ minimizing the submodular function $F - \kappa$
if $A = V$ then
    return $x$
end if
Let $x_A = \text{Prox}_{\lambda\Omega_p^{F_A}}(z_A)$
Let $x_{A^c} = \text{Prox}_{\lambda\Omega_p^{F^A}}(z_{A^c})$
return $x$ by concatenating $x_A$ and $x_{A^c}$
```

F.2 Decomposition algorithm to compute the norm

By Equation (6), for $p \in [1, \infty)$, the computation of the norm $\Omega_p^F(z)$ can be formulated as well as the maximization of a separable concave function over the canonical polytope: $\max_{\kappa \in \mathcal{P}_F} \sum_i \psi_i(\kappa_i)$ with $\psi_i(\kappa_i) = \kappa_i^{\frac{1}{q}} |z_i|$. We can therefore apply the same decomposition algorithm of Bach (2013, Sec. 9.1). This yields Algorithm 5.

Algorithm 5 Computation of $\Omega_p^F(z)$

```
Require: $z \in \mathbb{R}^d$.
 1: Let $A = \{j \mid z_j \neq 0\}$.
 2: if $A \neq V$ then
 3:     return $\Omega_p^{F_A}(z_A)$
 4: end if
 5: Let $\kappa \in \mathbb{R}^d$ with $\kappa_i = \frac{|z_i|^p}{\|z\|_p^p} F(V)$
 6: Find $A$ minimizing the submodular function $F - \kappa$
 7: if $A = V$ then
 8:     return $F(V)^{1/q} \|x\|_p$
 9: else
10:     return $\Omega_p^{F_A}(z_A) + \Omega_p^{F^A}(z_{A^c})$
11: end if
```

The derivation of this algorithm is essentially identical to the derivation of Algorithm 2: in the first step of the divide-and-conquer algorithm described, solving $\max_{\kappa(V) \leq F(V)} \sum_i \psi_i(\kappa_i)$ leads to $\kappa_i = \frac{|z_i|^p}{\|z\|_p^p} F(V)$. Finally, then either κ is optimal and the objective equals $F(V)^{\frac{1}{q}} \sum_i \frac{|z_i|^{\frac{p}{q}+1}}{\|z\|_p^{\frac{p}{q}}} = F(V)^{\frac{1}{q}} \|z\|_p$, or solving the two subproblems of steps (4) and (5) corresponds to computing the norms $\Omega_p^{F_A}$ and $\Omega_p^{F^A}$ on the two subvectors z_A and z_{A^c} and summing them.

F.3 Proof of Algorithm 3

Proof. The algorithm is a recursive algorithm, whose principle is to remove a leaf of the tree, to then compute the set A minimizing a new objective G' defined on the resulting reduced tree and to construct the minimizer of $G(B) = \lambda F(B) - s(B)$ from A by possibly adding the removed leaf. To define G', assume that the nodes are indexed in topological order, that the algorithm first removes the node n, and let $s' \in \mathbb{R}^{n-1}$ be defined as

$$\begin{cases} s'_{\pi_n} &= s_{\pi_n} + (s_n - \lambda)_+, \\ s'_j &= s_j, \quad \forall j \in V \backslash \{\pi_n, n\}. \end{cases}$$

Then let $G' : 2^{V \backslash \{n\}} \to \mathbb{R}$ be defined by $G'(A) = \lambda F(A) - s'(A)$. For any $A \subset V$ we have

$$G'(A) = \begin{cases} \min\big(G(A), G(A \cup \{n\})\big) & \text{if } \pi_n \in A, \\ G(A) & \text{else,} \end{cases}$$

because if $\pi_n \in A$, then $G(A \cup \{n\}) = G(A) - (s_n - \lambda)$. It is therefore clear that A is a minimizer of $G'(A)$ if and only if either $A \not\ni \pi_n$ and A is a minimizer of $G(A)$, or $s_n \leq \lambda$ and A is a minimizer of $G(A)$, or $s_n \geq \lambda$ and $A \cup \{n\}$ is a minimizer of G. If $V = \varnothing$, the algorithm returns $A = \varnothing$, which is indeed the unique minimizer, so that the algorithm is correct for a tree with $n = 0$ nodes. Then by the argument above, if the algorithm is correct for a tree with $n - 1$ nodes it is also correct for a tree with n nodes. By induction, this proves the correctness of the algorithm. □

F.4 Proof of lemma 11

We first state a full version of the lemma that covers explicitly the case where $I^c \neq \varnothing$.

Lemma. *For $c \in \mathbb{R}_+^d$, $z \in \mathbb{R}_+^d$ with $z_1 > 0$ and ψ a nonnegative differentiable, decreasing and strictly convex function, consider the optimization problem:*

$$\min_{\eta \in \mathbb{R}^d} \sum_{i=1}^d c_i \psi(\eta_i) + z_i \eta_i \quad s.t. \quad b' \leq \eta_d \leq \ldots \leq \eta_1. \tag{GIRC(c,z,b')}$$

Let $I := \{i \mid c_i \neq 0\}$, $I_\sim := \{i \mid z_i = c_i = 0\}$.

- *If $I^c = \varnothing$ then if x^* is the solution of IRC(ω, y, b) with $\omega_i = c_i$, $y_i = \frac{z_i}{c_i}$ and $b = -\lim_{\eta \to b'} \psi'(\eta)$, then the vector η^* with components $\eta_i^* = (\psi')^{-1}(-x_i^*)$ is the unique solution to GIRC(c, z, b').*
- *If $I_\sim = \varnothing$, then the solution is unique and obtained as follows: Write $I = \{i_1, \ldots, i_K\}$ with $i_1 < \ldots < i_K$, define for all k the set $J_{i_k} := \{j \mid i_{k-1} < j \leq i_k\}$ with $i_0 := 0$ and define $J_+ = \{j \mid j > i_K\}$. Then at the optimum η^*, (a) $\forall i \in I, \forall j \in J_i$, $\eta_j^* = \eta_i^*$, (b) $\forall j \in J_+$, $\eta_j^* = b'$ and (c) if, for any $i \in I$, we let $\bar{z}_i = \sum_{j \in J_i} z_j$, then $(\eta_i^*)_{i \in I}$ is the unique solution of the problem GIRC$\big((c_i)_{i\in I}, (\bar{z}_i)_{i \in I}, b'\big)$, for which the previous case applies.*
- *If $I_\sim \neq \varnothing$, then for any $j \in V$, let $j_+ = \min\{i \geq j \mid i \notin I_\sim\}$ and $j_- = \max\{i \leq j \mid i \notin I_\sim\}$ if $(\eta_i^*)_{i \in I_\sim^c}$ is the unique solution of GIRC$(\tilde{c}, \tilde{z}, b')$ where $\tilde{c}$ and $\tilde{z}$ are respectively the restrictions of c and z on $I_\sim^c$, then the set of solutions is $\{\eta \in \mathbb{R}^d \mid \forall j,\ \eta_{j_+}^* \geq \eta_j \geq \eta_{j_-}^*\}$.*

Proof. First, note that the limit $\lim_{\eta \to b'} \psi'(\eta)$ exists in $\bar{\mathbb{R}}$ since ψ is strictly convex which entails that ψ' is an increasing function.

- When $I^c = \varnothing$, a similar result was shown by Barlow and Brunk (1972) but under essentially more restrictive conditions. With the assumptions of the stated lemma, we have $\min(\eta_1, \ldots, \eta_d) \geq b' > -\infty$, and the value of the objective is lower bounded by $z_1 \eta_1$, so that at the optimum $\eta_i^* \leq \eta_1^* \leq \frac{\psi(0)}{z_1} \sum_i c_i$. This shows that the infimum is attained on a compact set. Given that $c_i > 0$ for all i, the objective is strictly convex, which shows that the minimum exists and is unique. It is characterized by the KKT conditions. The KKT conditions for problem GIRC(c, z, b') are that any primal-dual optimal pair (η, λ) must satisfy the primal feasibility condition $b' \leq \eta_d \leq \ldots \leq \eta_1$ as well as Lagrangian stationarity, dual feasibility and complementary slackness as follows:

  $$\forall i, \qquad c_i \psi'(\eta_i) + z_i - \lambda_i + \lambda_{i-1} = 0, \qquad \lambda_i \geq 0, \qquad \lambda_i(\eta_i - \eta_{i+1}) = 0,$$

  with $\lambda_0 := 0$ and $\eta_{d+1} := b'$. Similarly, the KKT conditions for problem IRC(ω, y, b) are that any primal-dual optimal pair (x, μ) must satisfy $x_1 \leq \ldots \leq x_d \leq b$ and

  $$\forall i, \qquad -(\omega_i x_i - \omega_i y_i + \mu_i - \mu_{i-1}) = 0, \qquad \mu_i \geq 0, \qquad \mu_i(x_i - x_{i+1}) = 0,$$

  with $\mu_0 := 0$ and $x_{d+1} := b$.

  Consider (x^*, μ^*) the unique pair of primal dual solutions to the KKT equations for IRC(ω, y, b) and (η^*, λ^*) the unique pair of primal dual solutions to the KKT equations for GIRC(c, z, b'). The pairs are unique because both primal and dual problems are strongly convex (the objectives are differentiable and strictly convex). Now it is easily seen that, if one sets $\tilde{x}_i := -\psi'(\eta_i^*)$ and $\tilde{\mu}_i = \lambda_i$ for all i, then the pair $(\tilde{x}, \tilde{\mu})$ satisfies the KKT conditions for IRC(ω, y, b), which proves by uniqueness that $(\tilde{x}, \tilde{\mu}) = (x^*, \mu^*)$. So in particular, we have $\eta_i^* = (\psi')^{-1}(-x_i^*)$.

- When $I_\sim = \varnothing$, then the partial minimization with respect to the variables η_j indexed by $j \in J_i \backslash \{i\}$ is obtained in closed form: since $I_\sim = \varnothing$, all coefficients $(z_j)_{j \in J_i \backslash \{i\}}$ are strictly positive which shows that the η_j are equal to their lower bound η_i. The argument for J_+ is the same. Eliminating the variables indexed by $I_\sim$ from the problem, yields a problem which satisfies the assumption that $I^c = \varnothing$ and the result follows.
- When $I_\sim \neq \varnothing$, the variables x_j for $j \in I_\sim$ do not appear in the objective. At the optimum they must therefore simply satisfy the primal inequality constraints, and eliminating them from the objective yields a problem $\mathrm{GIRC}(\tilde{c}, \tilde{z}, b)$ satisfying the previous assumptions.

□

G Theoretical Results

In this section, we prove the propositions on consistency, support recovery and the concentration result of Section 8.2. As there, we consider a fixed design matrix $X \in \mathbb{R}^{n \times p}$ and $y \in \mathbb{R}^n$ a vector of random responses. Given $\lambda > 0$, we define $\hat{w}$ as a minimizer of the regularized least-squares cost:

$$\min_{w \in \mathbb{R}^d} \tfrac{1}{2n} \|y - Xw\|_2^2 + \lambda \Omega(w). \tag{24}$$

G.1 Proof of Proposition 7 (Support recovery)

Proof. We follow the proof of the case $p = \infty$ from Bach (2010). Let $r = \frac{1}{n} X^\top \varepsilon \in \mathbb{R}^d$, which is normal with mean zero and covariance matrix $\sigma^2 Q/n$. We have for any $w \in \mathbb{R}^p$,

$$\Omega(w) \geqslant \Omega_J(w_J) + \Omega^J(w_{J^c}) \geqslant \Omega_J(w_J) + \rho\, \Omega_{J^c}(w_{J^c}) \geqslant \rho\, \Omega(w).$$

This implies that $\Omega^*(r) \geqslant \rho \max\{\Omega_J^*(r_J), (\Omega^J)^*(r_{J^c})\}$.

Moreover, $r_{J^c} - Q_{J^c J} Q_{JJ}^{-1} r_J$ is normal with covariance matrix

$$\frac{\sigma^2}{n}(Q_{J^c J^c} - Q_{J^c J} Q_{JJ}^{-1} Q_{J J^c}) \preccurlyeq \sigma^2/n Q_{J^c J^c}.$$

This implies that with probability larger than $1 - 3P(\Omega^*(r) > \lambda \rho \eta/2)$, we have

$$\Omega_J^*(r_J) \leqslant \lambda/2 \qquad \text{and} \qquad (\Omega^J)^*(r_{J^c} - Q_{J^c J} Q_{JJ}^{-1} r_J) \leqslant \lambda \eta/2.$$

We denote by $\tilde{w}$ the unique (because Q_{JJ} is invertible) minimum of $\frac{1}{2n}\|y - Xw\|_2^2 + \lambda \Omega(w)$, subject to $w_{J^c} = 0$. $\tilde{w}_J$ is defined through $Q_{JJ}(\tilde{w}_J - w_J{}^*) - r_J = -\lambda s_J$ where $s_J \in \partial \Omega_J(\tilde{w}_J)$ (which implies that $\Omega_J^*(s_J) \leqslant 1$) , i.e., $\tilde{w}_J - w_J^* = Q_{JJ}^{-1}(r_J - \lambda s_J)$. We have:

$$\begin{aligned}
\|\tilde{w}_J - w_J^*\|_\infty &\leqslant \max_{j \in J} |\delta_j^\top Q_{JJ}^{-1}(r_J - \lambda s_J)| \\
&\leqslant \max_{j \in J} \Omega_J(Q_{JJ}^{-1}\delta_j) \Omega_J^*(r_J - \lambda s_J)| \\
&\leqslant \max_{j \in J} \|Q_{JJ}^{-1}\delta_j\|_p F(J)^{1-1/p} [\Omega_J^*(r_J) + \lambda \Omega_J^*(s_J)] \\
&\leqslant \max_{j \in J} \kappa^{-1} |J|^{1/p} F(J)^{1-1/p} [\Omega_J^*(r_J) + \lambda \Omega_J^*(s_J)] \leqslant \frac{3}{2} \lambda |J|^{1/p} F(J)^{1-1/p} \kappa^{-1}.
\end{aligned}$$

Thus if $2\lambda |J|^{1/p} F(J)^{1-1/p} \kappa^{-1} \leqslant \nu$, then $\|\tilde{w} - w^*\|_\infty \leqslant \frac{3\nu}{4}$, which implies $\mathrm{Supp}(\tilde{w}) \supset \mathrm{Supp}(w^*)$.

In the neighborhood of $\tilde{w}$, we have an exact decomposition of the norm, hence, to show that $\tilde{w}$ is the unique global minimum, we simply need to show that since we have $(\Omega^J)^*(r_{J^c} - Q_{J^cJ}Q_{JJ}^{-1}r_J) \leqslant \lambda\eta/2$, $\tilde{w}$ is the unique minimizer of Eq. (19). For that it suffices to show that $(\Omega^J)^*(Q_{J^cJ}(\tilde{w}_J - w_J^*) - r_{J^c}) < \lambda$. We have:

$$\begin{aligned}
(\Omega^J)^*(Q_{J^cJ}(\tilde{w}_J - w_J^*) - r_{J^c}) &= (\Omega^J)^*(Q_{J^cJ}Q_{JJ}^{-1}(r_J - \lambda s_J) - r_{J^c}) \\
&\leqslant (\Omega^J)^*(Q_{J^cJ}Q_{JJ}^{-1}r_J - r_{J^c}) + \lambda(\Omega^J)^*(Q_{J^cJ}Q_{JJ}^{-1}s_J) \\
&\leqslant (\Omega^J)^*(Q_{J^cJ}Q_{JJ}^{-1}r_J - r_{J^c}) + \lambda(\Omega^J)^*[(\Omega_J(Q_{JJ}^{-1}Q_{Jj}))_{j\in J^c}] \\
&\leqslant \lambda\eta/2 + \lambda(1-\eta) < \lambda,
\end{aligned}$$

which leads to the desired result. □

G.2 Proof of proposition 8 (Consistency)

Proof. Like for the proof of Proposition 7, we have

$$\Omega(x) \geqslant \Omega_J(x_J) + \Omega^J(x_{J^c}) \geqslant \Omega_J(x_J) + \rho\,\Omega_{J^c}(x_{J^c}) \geqslant \rho\,\Omega(x).$$

Thus, if we assume $\Omega^*(q) \leqslant \lambda\rho/2$, then $\Omega_J^*(q_J) \leqslant \lambda/2$ and $(\Omega^J)^*(q_{J^c}) \leqslant \lambda/2$. Let $\Delta = \hat{w} - w^*$.

We follow the proof from Bickel et al. (2009) by using the decomposition property of the norm Ω. We have, by optimality of $\hat{w}$:

$$\frac{1}{2}\Delta^\top Q\Delta + \lambda\Omega(w^* + \Delta) + q^\top\Delta \leqslant \lambda\Omega(w^* + \Delta) + q^\top\Delta \leqslant \lambda\Omega(w^*)$$

Using the decomposition property,

$$\lambda\Omega_J((w^* + \Delta)_J) + \lambda\Omega^J((w^* + \Delta)_{J^c}) + q_J^\top\Delta_J + q_{J^c}^\top\Delta_{J^c} \leqslant \lambda\Omega_J(w_J^*),$$

$$\lambda\Omega^J(\Delta_{J^c}) \leqslant \lambda\Omega_J(w_J^*) - \lambda\Omega_J(w_J^* + \Delta_J) + \Omega_J^*(q_J)\Omega_J(\Delta_J) + (\Omega^J)^*(q_{J^c})\Omega^J(\Delta_{J^c}), \quad \text{and}$$

$$(\lambda - (\Omega^J)^*(q_{J^c}))\Omega^J(\Delta_{J^c}) \leqslant (\lambda + \Omega_J^*(q_J))\Omega_J(\Delta_J).$$

Thus $\Omega^J(\Delta_{J^c}) \leqslant 3\Omega_J(\Delta_J)$, which implies $\Delta^\top Q\Delta \geqslant \kappa\|\Delta_J\|_2^2$ (by our assumption which generalizes the usual ℓ_1-restricted eigenvalue condition). Moreover, we have:

$$\begin{aligned}
\Delta^\top Q\Delta &= \Delta^\top(Q\Delta) \leqslant \Omega(\Delta)\Omega^*(Q\Delta) \\
&\leqslant \Omega(\Delta)(\Omega^*(q) + \lambda) \leqslant \frac{3\lambda}{2}\Omega(\Delta) \text{ by optimality of } \hat{w} \\
\Omega(\Delta) &\leqslant \Omega_J(\Delta_J) + \rho^{-1}\Omega^J(\Delta_{J^c}) \\
&\leqslant \Omega_J(\Delta_J)(3 + \frac{1}{\rho}) \leqslant \frac{4}{\rho}\Omega_J(\Delta_J).
\end{aligned}$$

This implies that $\kappa\Omega_J(\Delta_J)^2 \leqslant \Delta^\top Q\Delta \leqslant \frac{6\lambda}{\rho}\Omega_J(\Delta_J)$, and thus $\Omega_J(\Delta_J) \leqslant \frac{6\lambda}{\kappa\rho}$, which leads to the desired result, given the previous inequalities.

□

G.3 Proof of proposition 9

Proof. We have $\Omega^*(z) = \max_{A\in\mathcal{D}_F} \frac{\|z_A\|_q}{F(A)^{1/q}}$. Thus, from the union bound, we get

$$\mathbb{P}(\Omega^*(z) > t) \quad \leqslant \quad \sum_{A\in\mathcal{D}_F} \mathbb{P}(\|z_A\|_q^q > t^q F(A)).$$

We can then derive concentration inequalities. We have $\mathbb{E}\|z_A\|_q \leqslant (\mathbb{E}\|z_A\|_q^q)^{1/q} = (|A|\mathbb{E}|\varepsilon|^q)^{1/q} \leqslant 2|A|^{1/q}q^{1/2}$, where ε is a standard normal random variable. Moreover, $\|z_A\|_q \leqslant \|z_A\|_2$ for $q \geqslant 2$, and $\|z_A\|_q \leqslant |A|^{1/q-1/2}\|z_A\|_2$ for $q \leqslant 2$. We can thus use the concentration of Lipschitz-continuous functions of Gaussian variables, to get for $p \geqslant 2$ and $u \geqslant 0$,

$$\mathbb{P}\big(\|z_A\|_q \geqslant 2|A|^{1/q}\sqrt{q} + u\big) \leqslant e^{-u^2/2}.$$

For $p < 2$ (i.e., $q > 2$), we obtain

$$\mathbb{P}\big(\|z_A\|_q \geqslant 2|A|^{1/q}\sqrt{q} + u\big) \leqslant e^{-u^2|A|^{1-2/q}/2}.$$

We can also bound the expected norm $\mathbb{E}[\Omega^*(z)]$, as

$$\mathbb{E}[\Omega^*(z)] \leqslant 4\sqrt{q\log(2|\mathcal{D}_F|)}\max_{A\in\mathcal{D}_F}\frac{|A|^{1/q}}{F(A)^{1/q}}.$$

Together with $\Omega^*(z) \leqslant \|z\|_2 \max_{A\in\mathcal{D}_F}\frac{|A|^{(1/q-1/2)_+}}{F(A)^{1/q}}$, we get

$$\mathbb{P}\bigg(\Omega^*(z) \geqslant 4\sqrt{q\log(2|\mathcal{D}_F|)}\max_{A\in\mathcal{D}_F}\frac{|A|^{1/q}}{F(A)^{1/q}} + u\max_{A\in\mathcal{D}_F}\frac{|A|^{(1/q-1/2)_+}}{F(A)^{1/q}}\bigg) \leqslant e^{-u^2/2}.$$

□