Structured sparsity-inducing norms through submodular functions

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

Abstract

Sparse methods for supervised learning aim at finding good linear predictors from as few variables as possible, i.e., with small cardinality of their supports. This combinatorial selection problem is often turned into a convex optimization problem by replacing the cardinality function by its convex envelope (tightest convex lower bound), in this case the ℓ_1-norm. In this paper, we investigate more general set-functions than the cardinality, that may incorporate prior knowledge or structural constraints which are common in many applications: namely, we show that for nondecreasing submodular set-functions, the corresponding convex envelope can be obtained from its Lovász extension, a common tool in submodular analysis. This defines a family of polyhedral norms, for which we provide generic algorithmic tools (subgradients and proximal operators) and theoretical results (conditions for support recovery or high-dimensional inference). By selecting specific submodular functions, we can give a new interpretation to known norms, such as those based on rank-statistics or grouped norms with potentially overlapping groups; we also define new norms, in particular ones that can be used as non-factorial priors for supervised learning.

1 Introduction

The concept of parsimony is central in many scientific domains. In the context of statistics, signal processing or machine learning, it takes the form of variable or feature selection problems, and is commonly used in two situations: First, to make the model or the prediction more interpretable or cheaper to use, i.e., even if the underlying problem does not admit sparse solutions, one looks for the best sparse approximation. Second, sparsity can also be used given prior knowledge that the model should be sparse. In these two situations, reducing parsimony to finding models with low cardinality turns out to be limiting, and structured parsimony has emerged as a fruitful practical extension, with applications to image processing, text processing or bioinformatics (see, e.g., [1, 2, 3, 4, 5, 6, 7] and Section 4). For example, in [4], structured sparsity is used to encode prior knowledge regarding network relationship between genes, while in [6], it is used as an alternative to structured non-parametric Bayesian process based priors for topic models.

Most of the work based on convex optimization and the design of dedicated sparsity-inducing norms has focused mainly on the specific allowed set of sparsity patterns [1, 2, 4, 6]: if $w \in \mathbb{R}^p$ denotes the predictor we aim to estimate, and $\mathrm{Supp}(w)$ denotes its support, then these norms are designed so that penalizing with these norms only leads to supports from a given family of allowed patterns. In this paper, we instead follow the approach of [8, 3] and consider specific penalty functions $F(\mathrm{Supp}(w))$ of the support set, which go beyond the cardinality function, but are not limited or designed to only forbid certain sparsity patterns. As shown in Section 6.2, these may also lead to restricted sets of supports but their interpretation in terms of an *explicit* penalty on the support leads to additional insights into the behavior of structured sparsity-inducing norms (see, e.g., Section 4.1). While direct greedy approaches (i.e., forward selection) to the problem are considered in [8, 3], we provide convex relaxations to the function $w \mapsto F(\mathrm{Supp}(w))$, which extend the traditional link between the ℓ_1-norm and the cardinality function.

This is done for a particular ensemble of set-functions F, namely *nondecreasing submodular functions*. Submodular functions may be seen as the set-function equivalent of convex functions, and exhibit many interesting properties that we review in Section 2—see [9] for a tutorial on submodular analysis and [10, 11] for other applications to machine learning. This paper makes the following contributions:

– We make explicit links between submodularity and sparsity by showing that the convex envelope of the function $w \mapsto F(\mathrm{Supp}(w))$ on the ℓ_∞-ball may be readily obtained from the Lovász extension of the submodular function (Section 3).

– We provide generic algorithmic tools, i.e., subgradients and proximal operators (Section 5), as well as theoretical guarantees, i.e., conditions for support recovery or high-dimensional inference (Section 6), that extend classical results for the ℓ_1-norm and show that many norms may be tackled by the exact same analysis and algorithms.

– By selecting specific submodular functions in Section 4, we recover and give a new interpretation to known norms, such as those based on rank-statistics or grouped norms with potentially overlapping groups [1, 2, 7], and we define new norms, in particular ones that can be used as non-factorial priors for supervised learning (Section 4). These are illustrated on simulation experiments in Section 7, where they outperform related greedy approaches [3].

Notation. For $w \in \mathbb{R}^p$, $\mathrm{Supp}(w) \subset V = \{1, \dots, p\}$ denotes the support of w, defined as $\mathrm{Supp}(w) = \{j \in V,\ w_j \neq 0\}$. For $w \in \mathbb{R}^p$ and $q \in [1, \infty]$, we denote by $\|w\|_q$ the ℓ_q-norm of w. We denote by $|w| \in \mathbb{R}^p$ the vector of absolute values of the components of w. Moreover, given a vector w and a matrix Q, w_A and Q_{AA} are the corresponding subvector and submatrix of w and Q. Finally, for $w \in \mathbb{R}^p$ and $A \subset V$, $w(A) = \sum_{k \in A} w_k$ (this defines a modular set-function).

2 Review of submodular function theory

Throughout this paper, we consider a *nondecreasing submodular* function F defined on the power set 2^V of $V = \{1, \dots, p\}$, i.e., such that:

$$\forall A, B \subset V, \qquad F(A) + F(B) \geqslant F(A \cup B) + F(A \cap B), \quad \text{(submodularity)}$$
$$\forall A, B \subset V, \qquad A \subset B \Rightarrow F(A) \leqslant F(B). \quad \text{(monotonicity)}$$

Moreover, we assume that $F(\varnothing) = 0$. These set-functions are often referred to as *polymatroid set-functions* [12, 13]. Also, without loss of generality, we may 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 (which will lead to the ℓ_1-norm) and, given a partition of V into $B_1 \cup \cdots \cup B_k = V$, the set-function $A \mapsto F(A)$ which is equal to the number of groups $B_1, \dots, B_k$ with non empty intersection with A (which will lead to the grouped ℓ_1/ℓ_∞-norm [1, 14]).

Lovász extension. Given any set-function F, one can define its *Lovász extension* $f : \mathbb{R}_+^p \to \mathbb{R}$, as follows; given $w \in \mathbb{R}_+^p$, 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:

$$f(w) = \sum_{k=1}^p w_{j_k} [F(\{j_1, \dots, j_k\}) - F(\{j_1, \dots, j_{k-1}\})]. \tag{1}$$

The Lovász extension f is always piecewise-linear, and when F is submodular, it is also convex (see, e.g., [12, 9]). Moreover, for all $\delta \in \{0,1\}^p$, $f(\delta) = F(\mathrm{Supp}(\delta))$: f is indeed an extension from vectors in $\{0,1\}^p$ (which can be identified with indicator vectors of sets) to all vectors in $\mathbb{R}_+^p$. Moreover, it turns out that minimizing F over subsets, i.e., minimizing f over $\{0,1\}^p$ is equivalent to minimizing f over $[0,1]^p$ [13].

Submodular polyhedron and greedy algorithm. We denote by $\mathcal{P}$ the *submodular polyhedron* [12], defined as the set of $s \in \mathbb{R}_+^p$ such that for all $A \subset V$, $s(A) \leqslant F(A)$, i.e., $\mathcal{P} = \{s \in \mathbb{R}_+^p,\ \forall A \subset V,\ s(A) \leqslant F(A)\}$, where we use the notation $s(A) = \sum_{k \in A} s_k$. One

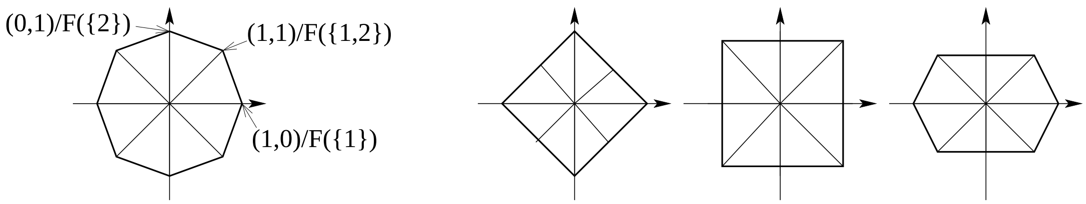

Figure 1: Polyhedral unit ball, for 4 different submodular functions (two variables), with different stable inseparable sets leading to different sets of extreme points; changing values of F may make some of the extreme points disappear. From left to right: $F(A) = |A|^{1/2}$ (all possible extreme points), $F(A) = |A|$ (leading to the ℓ_1-norm), $F(A) = \min\{|A|, 1\}$ (leading to the ℓ_∞-norm), $F(A) = \frac{1}{2} 1_{\{A \cap \{2\} \neq \varnothing\}} + 1_{\{A \neq \varnothing\}}$ (leading to the structured norm $\Omega(w) = \frac{1}{2}|w_2| + \|w\|_\infty$).

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 [12, 9], i.e., for all $w \in \mathbb{R}^p_+$,

$$f(w) = \max_{s \in \mathcal{P}} \; w^\top s. \tag{2}$$

Instead of solving a linear program with $p + 2^p$ contraints, a solution s may then be obtained by the following "greedy algorithm": order the components of w in decreasing order $w_{j_1} \geqslant \cdots \geqslant w_{j_p}$, and then take for all $k \in \{1, \dots, p\}$, $s_{j_k} = F(\{j_1, \dots, j_k\}) - F(\{j_1, \dots, j_{k-1}\})$.

Stable sets. 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 [13], and will correspond to the set of allowed sparsity patterns (see Section 6.2).

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 in [13], the submodular polytope $\mathcal{P}$ has full dimension p as soon as F is strictly positive on all singletons, and its faces are exactly the sets $\{s_k = 0\}$ for $k \in V$ and $\{s(A) = F(A)\}$ for stable *and* inseparable sets A. We denote by $\mathcal{T}$ the set of such sets. This implies that $\mathcal{P} = \{s \in \mathbb{R}^p_+, \; \forall A \in \mathcal{T}, s(A) \leqslant F(A)\}$. These stable inseparable sets will play a role when describing extreme points of unit balls of our new norms (Section 3) and for deriving concentration inequalities in Section 6.3. For the cardinality function, stable and inseparable sets are singletons.

3 Definition and properties of structured norms

We define the function $\Omega(w) = f(|w|)$, where $|w|$ is the vector in $\mathbb{R}^p$ composed of absolute values of w and f the Lovász extension of F. We have the following properties (see proof in [15]), which show that we indeed define a norm and that it is the desired convex envelope:

Proposition 1 (Convex envelope, dual norm) *Assume that the set-function F is submodular, non-decreasing, and strictly positive for all singletons. Define $\Omega : w \mapsto f(|w|)$. Then:*

(i) Ω is a norm on $\mathbb{R}^p$,

(ii) Ω is the convex envelope of the function $g : w \mapsto F(\mathrm{Supp}(w))$ on the unit ℓ_∞-ball,

(iii) the dual norm (see, e.g., [16]) of Ω is equal to $\Omega^*(s) = \max_{A \subset V} \frac{\|s_A\|_1}{F(A)} = \max_{A \in \mathcal{T}} \frac{\|s_A\|_1}{F(A)}$.

We provide examples of submodular set-functions and norms in Section 4, where we go from set-functions to norms, and vice-versa. From the definition of the Lovász extension in Eq. (1), we see that Ω is a polyhedral norm (i.e., its unit ball is a polyhedron). The following proposition gives the set of extreme points of the unit ball (see proof in [15] and examples in Figure 1):

Proposition 2 (Extreme points of unit ball) *The extreme points of the unit ball of Ω are the vectors $\frac{1}{F(A)} s$, with $s \in \{-1, 0, 1\}^p$, $\mathrm{Supp}(s) = A$ and A a stable inseparable set.*

This proposition shows, that depending on the number and cardinality of the inseparable stable sets, we can go from $2p$ (only singletons) to $3^p - 1$ extreme points (all possible sign vectors). We show in Figure 1 examples of balls for $p = 2$, as well as sets of extreme points. These extreme points will play a role in concentration inequalities derived in Section 6.

Figure 2: Sequence and groups: (left) groups for contiguous patterns, (right) groups for penalizing the number of jumps in the indicator vector sequence.

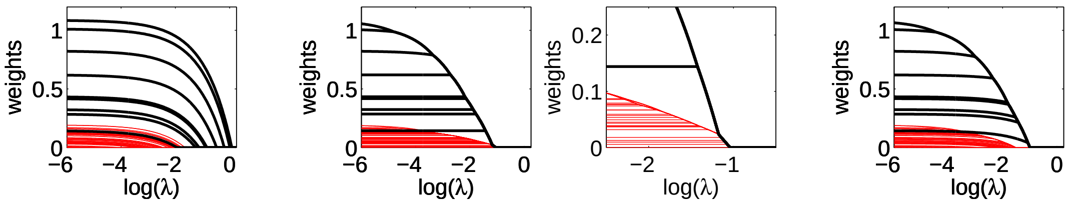

Figure 3: Regularization path for a penalized least-squares problem (black: variables that should be active, red: variables that should be left out). From left to right: ℓ_1-norm penalization (a wrong variable is included with the correct ones), polyhedral norm for rectangles in 2D, with zoom (all variables come in together), mix of the two norms (correct behavior).

4 Examples of nondecreasing submodular functions

We consider three main types of submodular functions with potential applications to regularization for supervised learning. Some existing norms are shown to be examples of our frameworks (Section 4.1, Section 4.3), while other novel norms are designed from specific submodular functions (Section 4.2). Other examples of submodular functions, in particular in terms of matroids and entropies, may be found in [12, 10, 11] and could also lead to interesting new norms. Note that set covers, which are common examples of submodular functions are subcases of set-functions defined in Section 4.1 (see, e.g., [9]).

4.1 Norms defined with non-overlapping or overlapping groups

We consider grouped norms defined with potentially overlapping groups [1, 2], i.e., $\Omega(w) = \sum_{G \subset V} d(G) \|w_G\|_\infty$ where d is a nonnegative set-function (with potentially $d(G) = 0$ when G should not be considered in the norm). It is a norm as soon as $\cup_{G, d(G)>0} G = V$ and it corresponds to the nondecreasing submodular function $F(A) = \sum_{G \cap A \neq \varnothing} d(G)$. In the case where ℓ_∞-norms are replaced by ℓ_2-norms, [2] has shown that the set of allowed sparsity patterns are intersections of complements of groups G with strictly positive weights. These sets happen to be the set of stable sets for the corresponding submodular function; thus the analysis provided in Section 6.2 extends the result of [2] to the new case of ℓ_∞-norms. However, in our situation, we can give a reinterpretation through a submodular function that counts the number of times the support A intersects groups G with non zero weights. This goes beyond restricting the set of allowed sparsity patterns to stable sets. We show later in this section some insights gained by this reinterpretation. We now give some examples of norms, with various topologies of groups.

Hierarchical norms. Hierarchical norms defined on directed acyclic graphs [1, 5, 6] correspond to the set-function $F(A)$ which is the cardinality of the union of ancestors of elements in A. These have been applied to bioinformatics [5], computer vision and topic models [6].

Norms defined on grids. If we assume that the p variables are organized in a 1D, 2D or 3D grid, [2] considers norms based on overlapping groups leading to stable sets equal to rectangular or convex shapes, with applications in computer vision [17]. For example, for the groups defined in the left side of Figure 2 (with unit weights), we have $F(A) = p - 2 + \mathrm{range}(A)$ if $A \neq \varnothing$ and $F(\varnothing) = 0$ (the range of A is equal to $\max(A) - \min(A) + 1$). From empty sets to non-empty sets, there is a gap of $p - 1$, which is larger than differences among non-empty sets. This leads to the undesired result, which has been already observed by [2], of adding all variables in one step, rather than gradually, when the regularization parameter decreases in a regularized optimization problem. In order to counterbalance this effect, adding a constant times the cardinality function has the effect of making the first gap relatively smaller. This corresponds to adding a constant times the ℓ_1-norm and, as shown in Figure 3, solves the problem of having all variables coming together. All patterns are then allowed, but contiguous ones are *encouraged rather than forced*.

Another interesting new norm may be defined from the groups in the right side of Figure 2. Indeed, it corresponds to the function $F(A)$ equal to $|A|$ plus the number of intervals of A. Note that this also favors contiguous patterns but is not limited to selecting a single interval (like the norm obtained from groups in the left side of Figure 2). Note that it is to be contrasted with the total variation (a.k.a. fused Lasso penalty [18]), which is a relaxation of the number of jumps in a vector w rather than in its support. In 2D or 3D, this extends to the notion of perimeter and area, but we do not pursue such extensions here.

4.2 Spectral functions of submatrices

Given a positive semidefinite matrix $Q \in \mathbb{R}^{p\times p}$ and a real-valued function h from $\mathbb{R}_+ \to \mathbb{R}$, one may define $\mathrm{tr}[h(Q)]$ as $\sum_{i=1}^p h(\lambda_i)$ where $\lambda_1, \dots, \lambda_p$ are the (nonnegative) eigenvalues of Q [19]. We can thus define the set-function $F(A) = \mathrm{tr}\, h(Q_{AA})$ for $A \subset V$. The functions $h(\lambda) = \log(\lambda + t)$ for $t \geqslant 0$ lead to submodular functions, as they correspond to entropies of Gaussian random variables (see, e.g., [12, 9]). Thus, since for $q \in (0,1)$, $\lambda^q = \frac{q \sin q\pi}{\pi} \int_0^\infty \log(1 + \lambda/t) t^{q-1} dt$ (see, e.g., [20]), $h(\lambda) = \lambda^q$ for $q \in (0,1]$ are positive linear combinations of functions that lead to nondecreasing submodular functions. Thus, they are also nondecreasing submodular functions, and, to the best of our knowledge, provide novel examples of such functions.

In the context of supervised learning from a design matrix $X \in \mathbb{R}^{n\times p}$, we naturally use $Q = X^\top X$. If h is linear, then $F(A) = \mathrm{tr}\, X_A^\top X_A = \sum_{k\in A} X_k^\top X_k$ (where X_A denotes the submatrix of X with columns in A) and we obtain a weighted cardinality function and hence and a weighted ℓ_1-norm, which is a *factorial prior*, i.e., it is a sum of terms depending on each variable independently.

In a frequentist setting, the Mallows C_L penalty [21] depends on the degrees of freedom, of the form $\mathrm{tr}\, X_A^\top X_A (X_A^\top X_A + \lambda I)^{-1}$. This is a non-factorial prior but unfortunately it does not lead to a submodular function. In a Bayesian context however, it is shown by [22] that penalties of the form $\log\det(X_A^\top X_A + \lambda I)$ (which lead to submodular functions) correspond to marginal likelihoods associated to the set A and have good behavior when used within a non-convex framework. This highlights the need for non-factorial priors which are sub-linear functions of the eigenvalues of $X_A^\top X_A$, which is exactly what nondecreasing submodular function of submatrices are. We do not pursue the extensive evaluation of non-factorial convex priors in this paper but provide in simulations examples with $F(A) = \mathrm{tr}(X_A^\top X_A)^{1/2}$ (which is equal to the trace norm of X_A [16]).

4.3 Functions of cardinality

For $F(A) = h(|A|)$ where h is nondecreasing, such that $h(0) = 0$ and concave, then, from Eq. (1), $\Omega(w)$ is defined from the rank statistics of $|w| \in \mathbb{R}_+^p$, i.e., if $|w_{(1)}| \geqslant |w_{(2)}| \geqslant \cdots \geqslant |w_{(p)}|$, then $\Omega(w) = \sum_{k=1}^p [h(k) - h(k-1)] |w_{(k)}|$. This includes the sum of the q largest elements, and might lead to interesting new norms for unstructured variable selection but this is not pursued here. However, the algorithms and analysis presented in Section 5 and Section 6 apply to this case.

5 Convex analysis and optimization

In this section we provide algorithmic tools related to optimization problems based on the regularization by our novel sparsity-inducing norms. Note that since these norms are polyhedral norms with unit balls having potentially an exponential number of vertices or faces, regular linear programming toolboxes may not be used.

Subgradient. From $\Omega(w) = \max_{s\in\mathcal{P}} s^\top |w|$ and the greedy algorithm[1] presented in Section 2, one can easily get in *polynomial time* one subgradient as one of the maximizers s. This allows to use subgradient descent, with, as shown in Figure 4, slow convergence compared to proximal methods.

Proximal operator. Given regularized problems of the form $\min_{w\in\mathbb{R}^p} L(w) + \lambda\Omega(w)$, where L is differentiable with Lipschitz-continuous gradient, *proximal methods* have been shown to be particularly efficient first-order methods (see, e.g., [23]). In this paper, we consider the methods "ISTA" and its accelerated variants "FISTA" [23], which are compared in Figure 4.

[1]The greedy algorithm to find extreme points of the submodular polyhedron should not be confused with the greedy algorithm (e.g., forward selection) that we consider in Section 7.

To apply these methods, it suffices to be able to solve efficiently problems of the form: $\min_{w \in \mathbb{R}^p} \frac{1}{2}\|w - z\|_2^2 + \lambda\Omega(w)$. In the case of the ℓ_1-norm, this reduces to soft thresholding of z, the following proposition (see proof in [15]) shows that this is equivalent to a particular algorithm for submodular function minimization, namely the minimum-norm-point algorithm, which has no complexity bound but is empirically faster than algorithms with such bounds [12]:

Proposition 3 (Proximal operator) *Let $z \in \mathbb{R}^p$ and $\lambda > 0$, minimizing $\frac{1}{2}\|w - z\|_2^2 + \lambda\Omega(w)$ is equivalent to finding the minimum of the submodular function $A \mapsto \lambda F(A) - |z|(A)$ with the minimum-norm-point algorithm.*

In [15], it is shown how a solution for one problem may be obtained from a solution to the other problem. Moreover, any algorithm for minimizing submodular functions allows to get directly the support of the unique solution of the proximal problem and that with a sequence of submodular function minimizations, the full solution may also be obtained. Similar links between convex optimization and minimization of submodular functions have been considered (see, e.g., [24]). However, these are dedicated to *symmetric* submodular functions (such as the ones obtained from graph cuts) and are thus not directly applicable to our situation of *non-increasing* submodular functions.

Finally, note that using the minimum-norm-point algorithm leads to a *generic* algorithm that can be applied to *any* submodular functions F, and that it may be rather inefficient for simpler subcases (e.g., the ℓ_1/ℓ_∞-norm, tree-structured groups [6], or general overlapping groups [7]).

6 Sparsity-inducing properties

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:

$$\min_{w \in \mathbb{R}^p} \frac{1}{2n}\|y - Xw\|_2^2 + \lambda\Omega(w). \tag{3}$$

We study the sparsity-inducing properties of solutions of Eq. (3), i.e., we determine in Section 6.2 which patterns are allowed and in Section 6.3 which sufficient conditions lead to correct estimation. Like recent analysis of sparsity-inducing norms [25], the analysis provided in this section relies heavily on decomposability properties of our norm Ω.

6.1 Decomposability

For a subset J of V, we denote by $F_J : 2^J \to \mathbb{R}$ the *restriction* of F to J, defined for $A \subset J$ by $F_J(A) = F(A)$, and by $F^J : 2^{J^c} \to \mathbb{R}$ the *contraction* of F by J, defined for $A \subset J^c$ by $F^J(A) = F(A \cup J) - F(A)$. These two functions are submodular and nondecreasing as soon as F is (see, e.g., [12]).

We denote by Ω_J the norm on $\mathbb{R}^J$ defined through the submodular function F_J, and Ω^J the pseudo-norm defined on $\mathbb{R}^{J^c}$ defined through F^J (as shown in Proposition 4, it is a norm only when J is a stable set). Note that Ω_{J^c} (a norm on J^c) is in general different from Ω^J. Moreover, $\Omega_J(w_J)$ is actually equal to $\Omega(\tilde{w})$ where $\tilde{w}_J = w_J$ and $\tilde{w}_{J^c} = 0$, i.e., it is the restriction of Ω to J.

We can now prove the following decomposition properties, which show that under certain circumstances, we can decompose the norm Ω on subsets J and their complements:

Proposition 4 (Decomposition) *Given $J \subset V$ and Ω_J and Ω^J defined as above, we have:*
(i) $\forall w \in \mathbb{R}^p$, $\Omega(w) \geqslant \Omega_J(w_J) + \Omega^J(w_{J^c})$,
(ii) $\forall w \in \mathbb{R}^p$, if $\min_{j \in J} |w_j| \geqslant \max_{j \in J^c} |w_j|$, then $\Omega(w) = \Omega_J(w_J) + \Omega^J(w_{J^c})$,
(iii) Ω^J is a norm on $\mathbb{R}^{J^c}$ if and only if J is a stable set.

6.2 Sparsity patterns

In this section, we do not make any assumptions regarding the correct specification of the linear model. We show that with probability one, only stable support sets may be obtained (see proof in [15]). For simplicity, we assume invertibility of $X^\top X$, which forbids the high-dimensional situation $p \geqslant n$ we consider in Section 6.3, but we could consider assumptions similar to the ones used in [2].

Proposition 5 (Stable sparsity patterns) *Assume $y \in \mathbb{R}^n$ has an absolutely continuous density with respect to the Lebesgue measure and that $X^\top X$ is invertible. Then the minimizer $\hat{w}$ of Eq. (3) is unique and, with probability one, its support $\mathrm{Supp}(\hat{w})$ is a stable set.*

6.3 High-dimensional inference

We now assume that the linear model is well-specified and extend results from [26] for sufficient support recovery conditions and from [25] for estimation consistency. As seen in Proposition 4, the norm Ω is decomposable and we use this property extensively in this section. We denote by $\rho(J) = \min_{B \subset J^c} \frac{F(B \cup J) - F(J)}{F(B)}$; by submodularity and monotonicity of F, $\rho(J)$ is always between zero and one, and, as soon as J is stable it is strictly positive (for the ℓ_1-norm, $\rho(J) = 1$). Moreover, we denote by $c(J) = \sup_{w \in \mathbb{R}^p} \Omega_J(w_J)/\|w_J\|_2$, the equivalence constant between the norm Ω_J and the ℓ_2-norm. We always have $c(J) \leqslant |J|^{1/2} \max_{k \in V} F(\{k\})$ (with equality for the ℓ_1-norm).

The following propositions allow us to get back and extend well-known results for the ℓ_1-norm, i.e., Propositions 6 and 8 extend results based on support recovery conditions [26]; while Propositions 7 and 8 extend results based on restricted eigenvalue conditions (see, e.g., [25]). We can also get back results for the ℓ_1/ℓ_∞-norm [14]. As shown in [15], proof techniques are similar and are adapted through the decomposition properties from Proposition 4.

Proposition 6 (Support recovery) *Assume that $y = Xw^* + \sigma\varepsilon$, where ε is a standard multivariate normal vector. Let $Q = \frac{1}{n} X^\top X \in \mathbb{R}^{p \times p}$. Denote by J the smallest stable set containing the support $\mathrm{Supp}(w^*)$ of w^*. Define $\nu = \min_{j, w_j^* \neq 0} |w_j^*| > 0$, assume $\kappa = \lambda_{\min}(Q_{JJ}) > 0$ and that for $\eta > 0$, $(\Omega^J)^*[(\Omega_J(Q_{JJ}^{-1} Q_{Jj}))_{j \in J^c}] \leqslant 1 - \eta$. Then, if $\lambda \leqslant \frac{\kappa\nu}{2c(J)}$, the minimizer $\hat{w}$ is unique and has support equal to J, with probability larger than $1 - 3P\big(\Omega^*(z) > \frac{\lambda\eta\rho(J)\sqrt{n}}{2\sigma}\big)$, where z is a multivariate normal with covariance matrix Q.*

Proposition 7 (Consistency) *Assume that $y = Xw^* + \sigma\varepsilon$, where ε is a standard multivariate normal vector. Let $Q = \frac{1}{n} X^\top X \in \mathbb{R}^{p \times p}$. Denote by J the smallest stable set containing the support $\mathrm{Supp}(w^*)$ of w^*. Assume that for all Δ such that $\Omega^J(\Delta_{J^c}) \leqslant 3\Omega_J(\Delta_J)$, $\Delta^\top Q \Delta \geqslant \kappa \|\Delta_J\|_2^2$. Then we have $\Omega(\hat{w} - w^*) \leqslant \frac{24 c(J)^2 \lambda}{\kappa \rho(J)^2}$ and $\frac{1}{n}\|X\hat{w} - Xw^*\|_2^2 \leqslant \frac{36 c(J)^2 \lambda^2}{\kappa \rho(J)^2}$, with probability larger than $1 - P\big(\Omega^*(z) > \frac{\lambda\rho(J)\sqrt{n}}{2\sigma}\big)$ where z is a multivariate normal with covariance matrix Q.*

Proposition 8 (Concentration inequalities) *Let z be a normal variable with covariance matrix Q. Let $\mathcal{T}$ be the set of stable inseparable sets. Then $P(\Omega^*(z) > t) \leqslant \sum_{A \in \mathcal{T}} 2^{|A|} \exp\big(-\frac{t^2 F(A)^2/2}{1^\top Q_{AA} 1}\big)$.*

7 Experiments

We provide illustrations on toy examples of some of the results presented in the paper. We consider the regularized least-squares problem of Eq. (3), with data generated as follows: given p, n, k, the design matrix $X \in \mathbb{R}^{n \times p}$ is a matrix of i.i.d. Gaussian components, normalized to have unit ℓ_2-norm columns. A set J of cardinality k is chosen at random and the weights w_J^* are sampled from a standard multivariate Gaussian distribution and $w_{J^c}^* = 0$. We then take $y = Xw^* + n^{-1/2}\|Xw^*\|_2\, \varepsilon$ where ε is a standard Gaussian vector (this corresponds to a unit signal-to-noise ratio).

Proximal methods vs. subgradient descent. For the submodular function $F(A) = |A|^{1/2}$ (a simple submodular function beyond the cardinality) we compare three optimization algorithms described in Section 5, subgradient descent and two proximal methods, ISTA and its accelerated version FISTA [23], for $p = n = 1000$, $k = 100$ and $\lambda = 0.1$. Other settings and other set-functions would lead to similar results than the ones presented in Figure 4: FISTA is faster than ISTA, and much faster than subgradient descent.

Relaxation of combinatorial optimization problem. We compare three strategies for solving the combinatorial optimization problem $\min_{w \in \mathbb{R}^p} \frac{1}{2n}\|y - Xw\|_2^2 + \lambda F(\mathrm{Supp}(w))$ with $F(A) = \mathrm{tr}(X_A^\top X_A)^{1/2}$, the approach based on our sparsity-inducing norms, the simpler greedy (forward selection) approach proposed in [8, 3], and by thresholding the ordinary least-squares estimate. For all methods, we try all possible regularization parameters. We see in the right plots of Figure 4 that

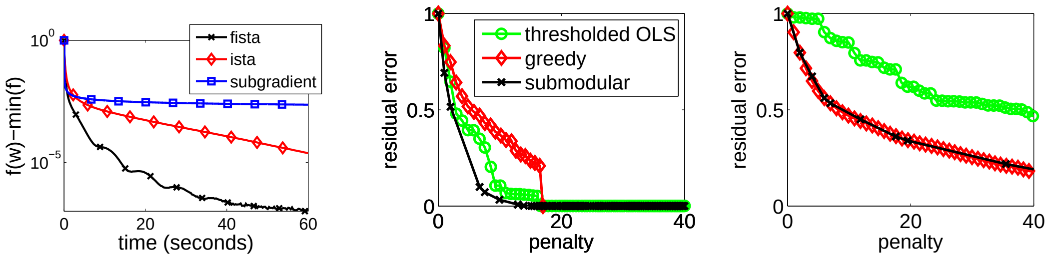

Figure 4: (Left) Comparison of iterative optimization algorithms (value of objective function vs. running time). (Middle/Right) Relaxation of combinatorial optimization problem, showing residual error $\frac{1}{n}\|y - X\hat{w}\|_2^2$ vs. penalty $F(\mathrm{Supp}(\hat{w}))$: (middle) high-dimensional case ($p = 120$, $n = 20$, $k = 40$), (right) lower-dimensional case ($p = 120$, $n = 120$, $k = 40$).

p	n	k	submodular	ℓ_2 vs. submod.	ℓ_1 vs. submod.	greedy vs. submod.
120	120	80	40.8 ± 0.8	-2.6 $\pm$ 0.5	**0.6 $\pm$ 0.0**	**21.8 $\pm$ 0.9**
120	120	40	35.9 ± 0.8	**2.4 $\pm$ 0.4**	**0.3 $\pm$ 0.0**	**15.8 $\pm$ 1.0**
120	120	20	29.0 ± 1.0	**9.4 $\pm$ 0.5**	-0.1 $\pm$ 0.0	**6.7 $\pm$ 0.9**
120	120	10	20.4 ± 1.0	**17.5 $\pm$ 0.5**	-0.2 $\pm$ 0.0	-2.8 $\pm$ 0.8
120	120	6	15.4 ± 0.9	**22.7 $\pm$ 0.5**	-0.2 $\pm$ 0.0	-5.3 $\pm$ 0.8
120	120	4	11.7 ± 0.9	**26.3 $\pm$ 0.5**	-0.1 $\pm$ 0.0	-6.0 $\pm$ 0.8
120	20	80	46.8 ± 2.1	-0.6 $\pm$ 0.5	**3.0 $\pm$ 0.9**	**22.9 $\pm$ 2.3**
120	20	40	47.9 ± 1.9	-0.3 $\pm$ 0.5	**3.5 $\pm$ 0.9**	**23.7 $\pm$ 2.0**
120	20	20	49.4 ± 2.0	0.4 $\pm$ 0.5	**2.2 $\pm$ 0.8**	**23.5 $\pm$ 2.1**
120	20	10	49.2 ± 2.0	0.0 $\pm$ 0.6	1.0 $\pm$ 0.8	**20.3 $\pm$ 2.6**
120	20	6	43.5 ± 2.0	**3.5 $\pm$ 0.8**	**0.9 $\pm$ 0.6**	**24.4 $\pm$ 3.0**
120	20	4	41.0 ± 2.1	**4.8 $\pm$ 0.7**	-1.3 $\pm$ 0.5	**25.1 $\pm$ 3.5**

Table 1: Normalized mean-square prediction errors $\|X\hat{w} - Xw^*\|_2^2/n$ (multiplied by 100) with optimal regularization parameters (averaged over 50 replications, with standard deviations divided by $\sqrt{50}$). The performance of the submodular method is shown, then differences from all methods to this particular one are computed, and shown in bold when they are significantly greater than zero, as measured by a paired t-test with level 5% (i.e., when the submodular method is significantly better).

for hard cases (middle plot) convex optimization techniques perform better than other approaches, while for easier cases with more observations (right plot), it does as well as greedy approaches.

Non factorial priors for variable selection. We now focus on the predictive performance and compare our new norm with $F(A) = \mathrm{tr}(X_A^\top X_A)^{1/2}$, with greedy approaches [3] and to regularization by ℓ_1 or ℓ_2 norms. As shown in Table 1, the new norm based on non-factorial priors is more robust than the ℓ_1-norm to lower number of observations n and to larger cardinality of support k.

8 Conclusions

We have presented a family of sparsity-inducing norms dedicated to incorporating prior knowledge or structural constraints on the support of linear predictors. We have provided a set of common algorithms and theoretical results, as well as simulations on synthetic examples illustrating the good behavior of these norms. Several avenues are worth investigating: first, we could follow current practice in sparse methods, e.g., by considering related adapted concave penalties to enhance sparsity-inducing norms, or by extending some of the concepts for norms of matrices, with potential applications in matrix factorization or multi-task learning (see, e.g., [27] for application of submodular functions to dictionary learning). Second, links between submodularity and sparsity could be studied further, in particular by considering submodular relaxations of other combinatorial functions, or studying links with other polyhedral norms such as the total variation, which are known to be similarly associated with symmetric submodular set-functions such as graph cuts [24].

Acknowledgements. This paper was partially supported by the Agence Nationale de la Recherche (MGA Project) and the European Research Council (SIERRA Project). The author would like to thank Edouard Grave, Rodolphe Jenatton, Armand Joulin, Julien Mairal and Guillaume Obozinski for discussions related to this work.

References

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

[2] R. Jenatton, J.Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. Technical report, arXiv:0904.3523, 2009.

[3] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. In *Proc. ICML*, 2009.

[4] L. Jacob, G. Obozinski, and J.-P. Vert. Group Lasso with overlaps and graph Lasso. In *Proc. ICML*, 2009.

[5] S. Kim and E. Xing. Tree-guided group Lasso for multi-task regression with structured sparsity. In *Proc. ICML*, 2010.

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

[7] J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Network flow algorithms for structured sparsity. In *Adv. NIPS*, 2010.

[8] J. Haupt and R. Nowak. Signal reconstruction from noisy random projections. *IEEE Transactions on Information Theory*, 52(9):4036–4048, 2006.

[9] F Bach. Convex analysis and optimization with submodular functions: a tutorial. Technical Report 00527714, HAL, 2010.

[10] A. Krause and C. Guestrin. Near-optimal nonmyopic value of information in graphical models. In *Proc. UAI*, 2005.

[11] Y. Kawahara, K. Nagano, K. Tsuda, and J.A. Bilmes. Submodularity cuts and applications. In *Adv. NIPS*, 2009.

[12] S. Fujishige. *Submodular Functions and Optimization*. Elsevier, 2005.

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

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

[15] F. Bach. Structured sparsity-inducing norms through submodular functions. Technical Report 00511310, HAL, 2010.

[16] S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[17] R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. In *Proc. AISTATS*, 2009.

[18] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused Lasso. *J. Roy. Stat. Soc. B*, 67(1):91–108, 2005.

[19] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge Univ. Press, 1990.

[20] T. Ando. Concavity of certain maps on positive definite matrices and applications to hadamard products. *Linear Algebra and its Applications*, 26:203–241, 1979.

[21] C. L. Mallows. Some comments on C_p. *Technometrics*, 15(4):661–675, 1973.

[22] D. Wipf and S. Nagarajan. Sparse estimation using general likelihoods and non-factorial priors. In *Adv. NIPS*, 2009.

[23] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

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

[25] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *Adv. NIPS*, 2009.

[26] P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

[27] A. Krause and V. Cevher. Submodular dictionary selection for sparse representation. In *Proc. ICML*, 2010.