Shaping Level Sets with Submodular Functions

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

Abstract

We consider a class of sparsity-inducing regularization terms based on submodular functions. While previous work has focused on non-decreasing functions, we explore symmetric submodular functions and their Lovász extensions. We show that the Lovász extension may be seen as the convex envelope of a function that depends on level sets (i.e., the set of indices whose corresponding components of the underlying predictor are greater than a given constant): this leads to a class of convex structured regularization terms that impose prior knowledge on the level sets, and not only on the supports of the underlying predictors. We provide unified optimization algorithms, such as proximal operators, and theoretical guarantees (allowed level sets and recovery conditions). By selecting specific submodular functions, we give a new interpretation to known norms, such as the total variation; we also define new norms, in particular ones that are based on order statistics with application to clustering and outlier detection, and on noisy cuts in graphs with application to change point detection in the presence of outliers.

1 Introduction

The concept of parsimony is central in many scientific domains. In the context of statistics, signal processing or machine learning, it may take several forms. Classically, in a variable or feature selection problem, a sparse solution with many zeros is sought so that the model is either more interpretable, cheaper to use, or simply matches available prior knowledge (see, e.g., [1, 2, 3] and references therein). In this paper, we instead consider sparsity-inducing regularization terms that will lead to solutions with *many equal values*. A classical example is the total variation in one or two dimensions, which leads to piecewise constant solutions [4, 5] and can be applied to various image labelling problems [6, 5], or change point detection tasks [7, 8, 9]. Another example is the "Oscar" penalty which induces automatic grouping of the features [10]. In this paper, we follow the approach of [3], who designed sparsity-inducing norms based on *non-decreasing* submodular functions, as a convex approximation to imposing a specific prior on the *supports* of the predictors. Here, we show that a similar parallel holds for some other class of submodular functions, namely non-negative set-functions which are equal to zero for the full and empty set. Our main instance of such functions are *symmetric* submodular functions.

We make the following contributions:

- We provide in Section 3 explicit links between priors on level sets and certain submodular functions: we show that the Lovász extensions (see, e.g., [11] and a short review in Section 2) associated to these submodular functions are the convex envelopes (i.e., tightest convex lower bounds) of specific functions that depend on all level sets of the underlying vector.
- In Section 4, we reinterpret existing norms such as the total variation and design new norms, based on noisy cuts or order statistics. We propose applications to clustering and outlier detection, as well as to change point detection in the presence of outliers.
- We provide unified algorithms in Section 5, such as proximal operators, which are based on a sequence of submodular function minimizations (SFMs), when such SFMs are efficient, or by adapting the generic slower approach of [3] otherwise.
- We derive unified theoretical guarantees for level set recovery in Section 6, showing that even in the absence of correlation between predictors, level set recovery is not always guaranteed, a situation which is to be contrasted with traditional support recovery situations [1, 3].

Notation. For $w \in \mathbb{R}^p$ and $q \in [1, \infty]$, we denote by $\|w\|_q$ the ℓ_q-norm of w. Given a subset A of $V = \{1, \dots, p\}$, $1_A \in \{0,1\}^p$ is the indicator vector of the subset A. Moreover, given a vector w and a matrix Q, w_A and Q_{AA} denote the corresponding subvector and submatrix of w and Q. Finally, for $w \in \mathbb{R}^p$ and $A \subset V$, $w(A) = \sum_{k \in A} w_k = w^\top 1_A$ (this defines a modular set-function). In this paper, for a certain vector $w \in \mathbb{R}^p$, we call *level sets* the sets of indices which are larger (or smaller) or equal to a certain constant α, which we denote $\{w \geqslant \alpha\}$ (or $\{w \leqslant \alpha\}$), while we call *constant sets* the sets of indices which are equal to a constant α, which we denote $\{w = \alpha\}$.

2 Review of Submodular Analysis

In this section, we review relevant results from submodular analysis. For more details, see, e.g., [12], and, for a review with proofs derived from classical convex analysis, see, e.g., [11].

Definition. Throughout this paper, we consider a *submodular* function F defined on the power set 2^V of $V = \{1, \dots, p\}$, i.e., such that $\forall A, B \subset V,\ F(A) + F(B) \geqslant F(A \cup B) + F(A \cap B)$. Unless otherwise stated, we consider functions which are non-negative (i.e., such that $F(A) \geqslant 0$ for all $A \subset V$), and that satisfy $F(\varnothing) = F(V) = 0$. Usual examples are symmetric submodular functions, i.e., such that $\forall A \subset V, F(V \backslash A) = F(A)$, which are known to always have non-negative values. We give several examples in Section 4; for illustrating the concepts introduced in this section and Section 3, we will consider the cut in an undirected chain graph, i.e., $F(A) = \sum_{j=1}^{p-1} |(1_A)_j - (1_A)_{j+1}|$.

Lovász extension. Given any set-function F such that $F(V) = F(\varnothing) = 0$, one can define its *Lovász extension* $f : \mathbb{R}^p \to \mathbb{R}$, as $f(w) = \int_{\mathbb{R}} F(\{w \geqslant \alpha\}) d\alpha$ (see, e.g., [11] for this particular formulation). The Lovász extension is convex if and only if F is submodular. Moreover, f is piecewise-linear and for all $A \subset V$, $f(1_A) = F(A)$, that is, it is indeed an extension from 2^V (which can be identified to $\{0,1\}^p$ through indicator vectors) to $\mathbb{R}^p$. Finally, it is always positively homogeneous. For the chain graph, we obtain the usual total variation $f(w) = \sum_{j=1}^{p-1} |w_j - w_{j+1}|$.

Base polyhedron. We denote by $B(F) = \{s \in \mathbb{R}^p,\ \forall A \subset V,\ s(A) \leqslant F(A),\ s(V) = F(V)\}$ the *base polyhedron* [12], where we use the notation $s(A) = \sum_{k \in A} s_k$. One important result in submodular analysis is that if F is a submodular function, then we have a representation of f as a maximum of linear functions [12, 11], i.e., for all $w \in \mathbb{R}^p$, $f(w) = \max_{s \in B(F)} w^\top s$. Moreover, instead of solving a linear program with $2^p - 1$ contraints, a solution s may 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 $k \in \{1, \dots, p\}$, $s_{j_k} = F(\{j_1, \dots, j_k\}) - F(\{j_1, \dots, j_{k-1}\})$.

Tight and inseparable sets. The polyhedra $\mathcal{U} = \{w \in \mathbb{R}^p, f(w) \leqslant 1\}$ and $B(F)$ are polar to each other (see, e.g., [13] for definitions and properties of polar sets). Therefore, the facial structure of $\mathcal{U}$ may be obtained from the one of $B(F)$. Given $s \in B(F)$, a set $A \subset V$ is said *tight* if $s(A) = F(A)$. It is known that the set of tight sets is a distributive lattice, i.e., if A and B are tight, then so are $A \cup B$ and $A \cap B$ [12, 11]. The faces of $B(F)$ are thus intersections of hyperplanes $\{s(A) = F(A)\}$ for A belonging to certain distributive lattices (see Prop. 3). A set A is said *separable* if there exists a non-trivial partition of $A = B \cup C$ such that $F(A) = F(B) + F(C)$. A set is said inseparable if it is not separable. For the cut in an undirected graph, inseparable sets are exactly connected sets.

3 Properties of the Lovász Extension

In this section, we derive properties of the Lovász extension for submodular functions, which go beyond convexity and homogeneity. Throughout this section, we assume that F is a non-negative submodular set-function that is equal to zero at $\varnothing$ and V. This immediately implies that f is invariant by addition of any constant vector (that is, $f(w + \alpha 1_V) = f(w)$ for all $w \in \mathbb{R}^p$ and $\alpha \in \mathbb{R}$), and that $f(1_V) = F(V) = 0$. Thus, contrary to the non-decreasing case [3], our regularizers are not norms. However, they are norms on the hyperplane $\{w^\top 1_V = 0\}$ as soon as for $A \neq \varnothing$ and $A \neq V$, $F(A) > 0$ (which we assume for the rest of this paper).

We now show that the Lovász extension is the convex envelope of a certain combinatorial function which does depend on all levets sets $\{w \geqslant \alpha\}$ of $w \in \mathbb{R}^p$ (see proof in [14]):

Proposition 1 (Convex envelope) *The Lovász extension $f(w)$ is the convex envelope of the function $w \mapsto \max_{\alpha \in \mathbb{R}} F(\{w \geqslant \alpha\})$ on the set $[0,1]^p + \mathbb{R}1_V = \{w \in \mathbb{R}^p,\ \max_{k \in V} w_k - \min_{k \in V} w_k \leqslant 1\}$.*

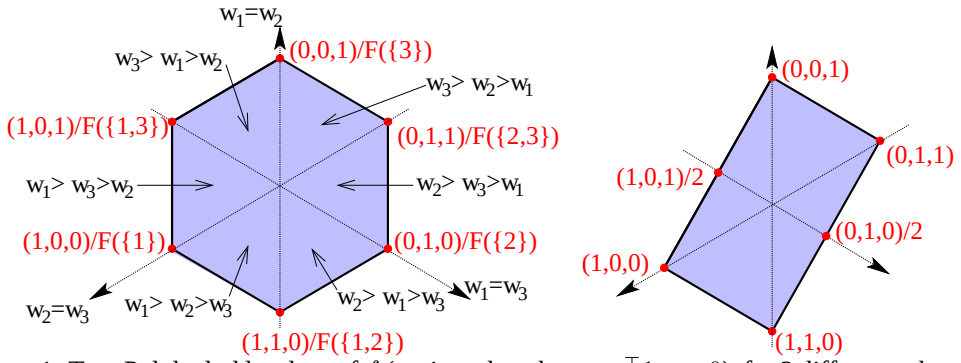


Figure 1: Top: Polyhedral level set of f (projected on the set $w^\top 1_V = 0$), for 2 different submodular symmetric functions of three variables, with different inseparable sets leading to different sets of extreme points; changing values of F may make some of the extreme points disappear. The various extreme points cut the space into polygons where the ordering of the component is fixed. Left: $F(A) = 1_{|A| \in \{1,2\}}$, leading to $f(w) = \max_k w_k - \min_k w_k$ (all possible extreme points); note that the polygon need not be symmetric in general. Right: one-dimensional total variation on three nodes, i.e., $F(A) = |1_{1 \in A} - 1_{2 \in A}| + |1_{2 \in A} - 1_{3 \in A}|$, leading to $f(w) = |w_1 - w_2| + |w_2 - w_3|$, for which the extreme points corresponding to the separable set $\{1, 3\}$ and its complement disappear.

Note the difference with the result of [3]: we consider here a different set on which we compute the convex envelope ($[0, 1]^p + \mathbb{R}1_V$ instead of $[-1, 1]^p$), and not a function of the support of w, but of *all* its level sets.[1] Moreover, the Lovász extension is a convex relaxation of a function of *level sets* (of the form $\{w \geqslant \alpha\}$) and not of *constant sets* (of the form $\{w = \alpha\}$). It would have been perhaps more intuitive to consider for example $\int_{\mathbb{R}} F(\{w = \alpha\}) d\alpha$, since it does not depend on the ordering of the values that w may take; however, to the best of our knowledge, the latter function does not lead to a convex function amenable to polynomial-time algorithms. This definition through level sets will generate some potentially undesired behavior (such as the well-known staircase effect for the one-dimensional total variation), as we show in Section 6.

The next proposition describes the set of extreme points of the "unit ball" $\mathcal{U} = \{w, \ f(w) \leqslant 1\}$, giving a first illustration of sparsity-inducing effects (see example in Figure 1, in particular for the one-dimensional total variation).

Proposition 2 (Extreme points) *The extreme points of the set $\mathcal{U} \cap \{w^\top 1_V = 0\}$ are the projections of the vectors $1_A / F(A)$ on the plane $\{w^\top 1_V = 0\}$, for A such that A is inseparable for F and $V \backslash A$ is inseparable for $B \mapsto F(A \cup B) - F(A)$.*

Partially ordered sets and distributive lattices. A subset $\mathcal{D}$ of 2^V is a (distributive) lattice if it is invariant by intersection and union. We assume in this paper that all lattices contain the empty set $\varnothing$ and the full set V, and we endow the lattice with the inclusion order. Such lattices may be represented as a *partially ordered set (poset)* $\Pi(\mathcal{D}) = \{A_1, \dots, A_m\}$ (with order relationship $\succcurlyeq$), where the sets A_j, $j = 1, \dots, m$, form a *partition* of V (we always assume a topological ordering of the sets, i.e., $A_i \succcurlyeq A_j \Rightarrow i \geqslant j$). As illustrated in Figure 2, we go from $\mathcal{D}$ to $\Pi(\mathcal{D})$, by considering all maximal chains in $\mathcal{D}$ and the differences between consecutive sets. We go from $\Pi(\mathcal{D})$ to $\mathcal{D}$, by constructing all *ideals* of $\Pi(\mathcal{D})$, i.e., sets J such that if an element of $\Pi(\mathcal{D})$ is lower than an element of J, then it has to be in J (see [12] for more details, and an example in Figure 2). Distributive lattices and posets are thus in one-to-one correspondence. Throughout this section, we go back and forth between these two representations. The distributive lattice $\mathcal{D}$ will correspond to all authorized level sets $\{w \geqslant \alpha\}$ for w in a single face of $\mathcal{U}$, while the elements of the poset $\Pi(\mathcal{D})$ are the constant sets (over which w is constant), with the order between the subsets giving *partial* constraints between the values of the corresponding constants.

Faces of $\mathcal{U}$. The faces of $\mathcal{U}$ are characterized by lattices $\mathcal{D}$, with their corresponding posets $\Pi(\mathcal{D}) = \{A_1, \dots, A_m\}$. We denote by $\mathcal{U}_\mathcal{D}^\circ$ (and by $\mathcal{U}_\mathcal{D}$ its closure) the set of $w \in \mathbb{R}^p$ such that (a) $f(w) \leqslant 1$, (b) w is piecewise constant with respect to $\Pi(\mathcal{D})$, with value v_i on A_i, and (c) for all pairs (i, j),

[1] Note that the support $\{w = 0\}$ is a constant set which is the intersection of two level sets.

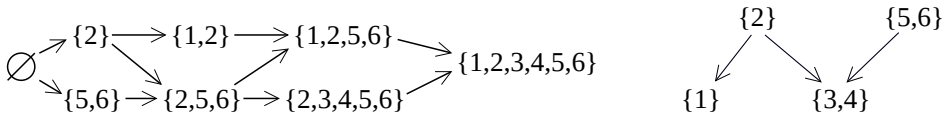


Figure 2: Left: distributive lattice with 7 elements in $2^{\{1,2,3,4,5,6\}}$, represented with the Hasse diagram corresponding to the inclusion order (for a partial order, a Hasse diagram connects A to B if A is smaller than B and there is no C such that A is smaller than C and C is smaller than B). Right: corresponding poset, with 4 elements that form a partition of $\{1, 2, 3, 4, 5, 6\}$, represented with the Hasse diagram corresponding to the order $\succcurlyeq$ (a node points to its immediate smaller node according to $\succcurlyeq$). Note that this corresponds to an "allowed" lattice (see Prop. 3) for the one-dimensional total variation.

$A_i \succcurlyeq A_j \Rightarrow v_i > v_j$. For certain lattices $\mathcal{D}$, these will be exactly the relative interiors of all faces of $\mathcal{U}$ (see proof in [14]):

Proposition 3 (Faces of $\mathcal{U}$) *The (non-empty) relative interiors of all faces of $\mathcal{U}$ are exactly of the form $\mathcal{U}_{\mathcal{D}}^{\circ}$, where $\mathcal{D}$ is a lattice such that:*

(i) the restriction of F to $\mathcal{D}$ is modular, i.e., for all $A, B \in \mathcal{D}$, $F(A)+F(B) = F(A \cup B)+F(A \cap B)$,

(ii) for all $j \in \{1, \dots, m\}$, the set A_j is inseparable for the function $C_j \mapsto F(B_{j-1} \cup C_j) - F(B_{j-1})$, where B_{j-1} is the union of all ancestors of A_j in $\Pi(\mathcal{D})$,

(iii) among all lattices corresponding to the same unordered *partition, $\mathcal{D}$ is a maximal element of the set of lattices satisfying (i) and (ii).*

Among the three conditions, the second one is the easiest to interpret, as it reduces to having constant sets which are inseparable for certain submodular functions, and for cuts in an undirected graph, these will exactly be connected sets. Note also that extreme points from Prop. 2 are recovered with $\mathcal{D} = \{\varnothing, A, V\}$.

Since we are able to characterize *all* faces of $\mathcal{U}$ (of all dimensions) with non-empty relative interior, we have a partition of the space and any $w \in \mathbb{R}^p$ which is not proportional to 1_V, will be, up to the strictly positive constant $f(w)$, in exactly one of these relative interiors of faces; we refer to this lattice as the *lattice associated to* w. Note that from the face w belongs to, we have strong constraints on the constant sets, but we may not be able to determine all level sets of w, because only partial constraints are given by the order on $\Pi(\mathcal{D})$. For example, in Figure 2 for the one-dimensional total variation, w_2 may be larger or smaller than $w_5 = w_6$ (and even potentially equal, but with zero probability, see Section 6).

4 Examples of Submodular Functions

In this section, we provide examples of submodular functions and of their Lovász extensions. Some are well-known (such as cut functions and total variations), some are new in the context of supervised learning (regular functions), while some have interesting effects in terms of clustering or outlier detection (cardinality-based functions).

Symmetrization. From any submodular function G, one may define $F(A) = G(A) + G(V \backslash A) - G(\varnothing) - G(V)$, which is symmetric. Potentially interesting examples which are beyond the scope of this paper are mutual information, or functions of eigenvalues of submatrices [3].

Cut functions. Given a set of *nonnegative* weights $d : V \times V \to \mathbb{R}_+$, define the cut $F(A) = \sum_{k \in A, j \in V \backslash A} d(k, j)$. The Lovász extension is equal to $f(w) = \sum_{k,j \in V} d(k, j)(w_k - w_j)_+$ (which shows submodularity because f is convex), and is often referred to as the total variation. If the weight function d is symmetric, then the submodular function is also symmetric. In this case, it can be shown that inseparable sets for functions $A \mapsto F(A \cup B) - F(B)$ are exactly *connected* sets. Hence, by Props. 3 and 6, constant sets are connected sets, which is the usual justification behind the total variation. Note however that some configurations of connected sets are not allowed due to the other conditions in Prop. 3 (see examples in Section 6). In Figure 5 (right plot), we give an example of the usual chain graph, leading to the one-dimensional total variation [4, 5]. Note that these functions can be extended to cuts in hypergraphs, which may have interesting applications in computer vision [6]. Moreover, directed cuts may be interesting to favor increasing or decreasing jumps along the edges of the graph.

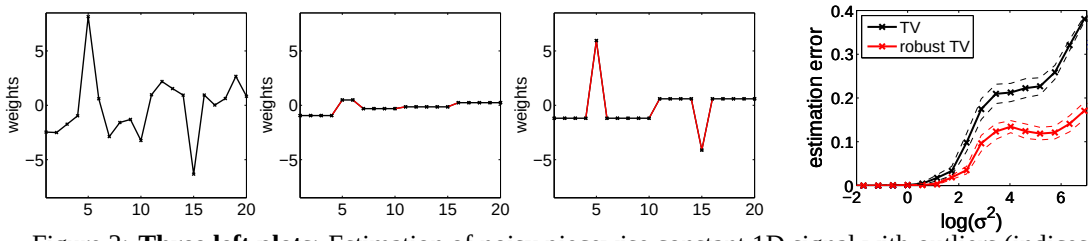


Figure 3: **Three left plots**: Estimation of noisy piecewise constant 1D signal with outliers (indices 5 and 15 in the chain of 20 nodes). Left: original signal. Middle: best estimation with total variation (level sets are not correctly estimated). Right: best estimation with the *robust* total variation based on noisy cut functions (level sets are correctly estimated, with less bias and with detection of outliers). **Right plot**: clustering estimation error vs. noise level, in a sequence of 100 variables, with a single jump, where noise of variance one is added, with 5% of outliers (averaged over 20 replications).

Regular functions and robust total variation. By partial minimization, we obtain so-called *regular functions* [6, 5]. One application is "noisy cut functions": for a given weight function $d : W \times W \to \mathbb{R}_+$, where each node in W is uniquely associated in a node in V, we consider the submodular function obtained as the minimum cut adapted to A in the augmented graph (see an example in the right plot of Figure 5): $F(A) = \min_{B \subset W} \sum_{k \in B,\ j \in W \setminus B} d(k,j) + \lambda |A \Delta B|$. This allows for robust versions of cuts, where some gaps may be tolerated; indeed, compared to having directly a small cut for A, B needs to have a small cut and be close to A, thus allowing some elements to be removed or added to A in order to lower the cut. See examples in Figure 3, illustrating the behavior of the type of graph displayed in the bottom-right plot of Figure 5, where the performance of the robust total variation is significantly more stable in presence of outliers.

Cardinality-based functions. For $F(A) = h(|A|)$ where h is such that $h(0) = h(p) = 0$ and h concave, we obtain a submodular function, and a Lovász extension that depends on the order statistics of w, i.e., if $w_{j_1} \geqslant \cdots \geqslant w_{j_p}$, then $f(w) = \sum_{k=1}^{p-1} h(k)(w_{j_k} - w_{j_{k+1}})$. While these examples do not provide significantly different behaviors for the non-decreasing submodular functions explored by [3] (i.e., in terms of *support*), they lead to interesting behaviors here in terms of *level sets*, i.e., they will make the components w cluster together in specific ways. Indeed, as shown in Section 6, allowed constant sets A are such that A is inseparable for the function $C \mapsto h(|B \cup C|) - h(|B|)$ (where $B \subset V$ is the set of components with higher values than the ones in A), which imposes that the concave function h is not linear on $[|B|, |B|+|A|]$. We consider the following examples:

1. $F(A) = |A| \cdot |V \setminus A|$, leading to $f(w) = \sum_{i,j=1}^{p} |w_i - w_j|$. This function can thus be also seen as the cut in the fully connected graph. All patterns of level sets are allowed as the function h is strongly concave (see left plot of Figure 4). This function has been extended in [15] by considering situations where each w_j is a vector, instead of a scalar, and replacing the absolute value $|w_i - w_j|$ by any norm $\|w_i - w_j\|$, leading to convex formulations for clustering.
2. $F(A) = 1$ if $A \neq \varnothing$ and $A \neq V$, and 0 otherwise, leading to $f(w) = \max_{i,j} |w_i - w_j|$. Two large level sets at the top and bottom, all the rest of the variables are in-between and separated (Figure 4, second plot from the left).
3. $F(A) = \max\{|A|, |V \setminus A|\}$. This function is piecewise affine, with only one kink, thus only one level set of cardinalty greater than one (in the middle) is possible, which is observed in Figure 4 (third plot from the left). This may have applications to multivariate outlier detection by considering extensions similar to [15].

5 Optimization Algorithms

In this section, we present optimization methods for minimizing convex objective functions regularized by the Lovász extension of a submodular function. These lead to convex optimization problems, which we tackle using proximal methods (see, e.g., [16, 17] and references therein). We first start by mentioning that subgradients may easily be derived (but subgradient descent is here rather inefficient as shown in Figure 5). Moreover, note that with the square loss, the regularization paths are piecewise affine, as a direct consequence of regularizing by a polyhedral function.

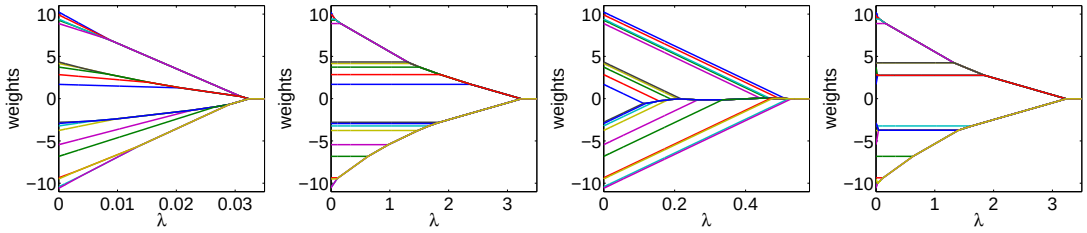


Figure 4: **Left:** Piecewise linear regularization paths of proximal problems (Eq. (1)) for different functions of cardinality. From left to right: quadratic function (all level sets allowed), second example in Section 4 (two large level sets at the top and bottom), piecewise linear with two pieces (a single large level set in the middle). **Right:** Same plot for the one-dimensional total variation. Note that in all these particular cases the regularization paths for orthogonal designs are *agglomerative* (see Section 5), while for general designs, they would still be piecewise affine but not agglomerative.

Subgradient. From $f(w) = \max_{s \in B(F)} s^\top w$ and the greedy algorithm[2] 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 slow convergence compared to proximal methods (see Figure 5).

Proximal problems through sequences of submodular function minimizations (SFMs). Given regularized problems of the form $\min_{w \in \mathbb{R}^p} L(w) + \lambda f(w)$, where L is differentiable with Lipschitz-continuous gradient, *proximal methods* have been shown to be particularly efficient first-order methods (see, e.g., [16]). In this paper, we use the method "ISTA" and its accelerated variant "FISTA" [16]. To apply these methods, it suffices to be able to solve efficiently:

$$\min_{w \in \mathbb{R}^p} \tfrac{1}{2}\|w - z\|_2^2 + \lambda f(w), \tag{1}$$

which we refer to as the *proximal problem*. It is known that solving the proximal problem is related to submodular function minimization (SFM). More precisely, the minimum of $A \mapsto \lambda F(A) - z(A)$ may be obtained by selecting negative components of the solution of a single proximal problem [12, 11]. Alternatively, the solution of the proximal problem may be obtained by a sequence of at most p submodular function minimizations of the form $A \mapsto \lambda F(A) - z(A)$, by a decomposition algorithm adapted from [18], and described in [11].

Thus, computing the proximal operator has polynomial complexity since SFM has polynomial complexity. However, it may be too slow for practical purposes, as the best *generic* algorithm has complexity $O(p^6)$ [19][3]. Nevertheless, this strategy is efficient for families of submodular functions for which dedicated fast algorithms exist:

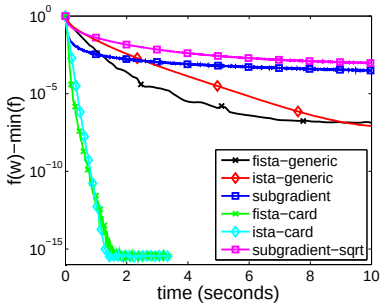
- **Cuts**: Minimizing the cut or the partially minimized cut, plus a modular function, may be done with a min-cut/max-flow algorithm [see, e.g., 6, 5]. For proximal methods, we need in fact to solve an instance of a *parametric max-flow* problem, which may be done using other efficient dedicated algorithms [21, 5] than the decomposition algorithm derived from [18].
- **Functions of cardinality**: minimizing functions of the form $A \mapsto \lambda F(A) - z(A)$ can be done in closed form by sorting the elements of z.

Proximal problems through minimum-norm-point algorithm. In the *generic* case (i.e., beyond cuts and cardinality-based functions), we can follow [12, 3]: since $f(w)$ is expressed as a minimum of linear functions, the problem reduces to the projection on the polytope $B(F)$, for which we happen to be able to easily maximize linear functions (using the greedy algorithm described in Section 2). This can be tackled efficiently by the minimum-norm-point algorithm [12], which iterates between orthogonal projections on affine subspaces and the greedy algorithm for the submodular function[4]. We compare all optimization methods on synthetic examples in Figure 5.

[2]The greedy algorithm to find extreme points of the base polyhedron should not be confused with the greedy algorithm (e.g., forward selection) that is common in supervised learning/statistics.

[3]Note that even in the case of symmetric submodular functions, where more efficient algorithms in $O(p^3)$ for submodular function minimization (SFM) exist [20], the minimization of functions of the form $\lambda F(A) - z(A)$ is provably as hard as general SFM [20].

[4]Interestingly, when used for submodular function minimization (SFM), the minimum-norm-point algorithm has no complexity bound but is empirically faster than algorithms with such bounds [12].



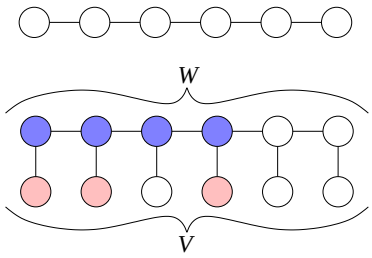


Figure 5: **Left**: Matlab running times of different optimization methods on 20 replications of a least-squares regression problem with $p = 1000$ for a cardinality-based submodular function (best seen in color). Proximal methods with the generic algorithm (using the minimum-norm-point algorithm) are faster than subgradient descent (with two schedules for the learning rate, $1/t$ or $1/\sqrt{t}$). Using the dedicated algorithm (which is not available in all situations) is significantly faster. **Right**: Examples of graphs (top: chain graph, bottom: hidden chain graph, with sets W and V and examples of a set A in light red, and B in blue, see text for details).

Proximal path as agglomerative clustering. When λ varies from zero to $+\infty$, then the unique optimal solution of Eq. (1) goes from z to a constant. We now provide conditions under which the regularization path of the proximal problem may be obtained by agglomerative clustering (see examples in Figure 4):

Proposition 4 (Agglomerative clustering) *Assume that for all sets A, B such that $B \cap A = \varnothing$ and A is inseparable for $D \mapsto F(B \cup D) - F(B)$, we have:*

$$\forall C \subset A, \ \frac{|C|}{|A|}[F(B \cup A) - F(B)] \leqslant F(B \cup C) - F(B). \tag{2}$$

Then the regularization path for Eq. (1) is agglomerative*, that is, if two variables are in the same constant for a certain $\mu \in \mathbb{R}_+$, so are they for all larger $\lambda \geqslant \mu$.*

As shown in [14], the assumptions required for by Prop. 4 are satisfied by (a) all submodular set-functions that only depend on the cardinality, and (b) by the one-dimensional total variation—we thus recover and extend known results from [7, 22, 15].

Adding an ℓ_1-norm. Following [4], we may add the ℓ_1-norm $\|w\|_1$ for additional sparsity of w (on top of shaping its level sets). The following proposition extends the result for the one-dimensional total variation [4, 23] to all submodular functions and their Lovász extensions:

Proposition 5 (Proximal problem for ℓ_1-penalized problems) *The unique minimizer of $\frac{1}{2}\|w - z\|_2^2 + f(w) + \lambda\|w\|_1$ may be obtained by soft-thresholding the minimizers of $\frac{1}{2}\|w - z\|_2^2 + f(w)$. That is, the proximal operator for $f + \lambda\|\cdot\|_1$ is equal to the composition of the proximal operator for f and the one for $\lambda\|\cdot\|_1$.*

6 Sparsity-inducing Properties

Going from the penalization of supports to the penalization of level sets introduces some complexity and for simplicity in this section, we only consider the analysis in the context of orthogonal design matrices, which is often referred to as the denoising problem, and in the context of level set estimation already leads to interesting results. That is, we study the unique global minimum $\hat{w}$ of the proximal problem in Eq. (1) and make some assumption regarding z (typically $z = w^* +$ noise), and provide guarantees related to the recovery of the level sets of w^*. We first start by characterizing the allowed level sets, showing that the partial constraints defined in Section 3 on faces of $\{f(w) \leqslant 1\}$ do not create by chance further groupings of variables (see proof in [14]).

Proposition 6 (Stable constant sets) *Assume $z \in \mathbb{R}^p$ has an absolutely continuous density with respect to the Lebesgue measure. Then, with probability one, the unique minimizer $\hat{w}$ of Eq. (1) has constant sets that define a partition corresponding to a lattice $\mathcal{D}$ defined in Prop. 3.*

We now show that under certain conditions the recovered constant sets are the correct ones:

Theorem 1 (Level set recovery) *Assume that $z = w^* + \sigma\varepsilon$, where $\varepsilon \in \mathbb{R}^p$ is a standard Gaussian random vector, and w^* is consistent with the lattice $\mathcal{D}$ and its associated poset $\Pi(\mathcal{D}) = (A_1, \dots, A_m)$, with values v_j^* on A_j, for $j \in \{1, \dots, m\}$. Denote $B_j = A_1 \cup \cdots \cup A_j$ for $j \in \{1, \dots, m\}$. Assume that there exists some constants $\eta_j > 0$ and $\nu > 0$ such that:*

$$\forall C_j \subset A_j, F(B_{j-1} \cup C_j) - F(B_{j-1}) - \tfrac{|C_j|}{|A_j|}[F(B_{j-1} \cup A_j) - F(B_{j-1})] \geqslant \eta_j \min\big\{\tfrac{|C_j|}{|A_j|}, 1 - \tfrac{|C_j|}{|A_j|}\big\}, \quad (3)$$

$$\forall i, j \in \{1, \dots, m\},\ A_i \not\succcurlyeq A_j \Rightarrow v_i^* - v_j^* \geqslant \nu, \quad (4)$$

$$\forall j \in \{1, \dots, m\},\ \lambda \big| \tfrac{F(B_j) - F(B_{j-1})}{|A_j|} \big| \leqslant \nu/4. \quad (5)$$

Then the unique minimizer $\hat{w}$ of Eq. (1) is associated to the same lattice $\mathcal{D}$ than w^*, with probability greater than $1 - \sum_{j=1}^m \exp\big(-\frac{\nu^2 |A_j|}{32\sigma^2}\big) - 2\sum_{j=1}^m |A_j| \exp\big(-\frac{\lambda^2 \eta_j^2}{2\sigma^2 |A_j|^2}\big)$.

We now discuss the three main assumptions of Theorem 1 as well as the probability estimate:

- Eq. (3) is the equivalent of the support recovery condition for the Lasso [1] or its extensions [3]. The main difference is that for support recovery, this assumption is always met for orthogonal designs, while here it is not always met. Interestingly, the validity of level set recovery implies the agglomerativity of proximal paths (Eq. (2) in Prop. 4).
  Note that if Eq. (3) is satisfied only with $\eta_j \geqslant 0$ (it is then exactly Eq. (2) in Prop. 4), then, even with infinitesimal noise, one can show that in some cases, the wrong level sets may be obtained with non vanishing probability, while if η_j is strictly negative, one can show that in some cases, we *never* get the correct level sets. Eq. (3) is thus essentially sufficient and necessary.
- Eq. (4) corresponds to having distinct values of w^* far enough from each other.
- Eq. (5) is a constraint on λ which controls the bias of the estimator: if it is too large, then there may be a merging of two clusters.
- In the probability estimate, the second term is small if all $\sigma^2 |A_j|^{-1}$ are small enough (i.e., given the noise, there is enough data to correctly estimate the values of the constant sets) and the third term is small if λ is large enough, to avoid that clusters split.

One-dimensional total variation. In this situation, we always get $\eta_j = 0$, but in some cases, it cannot be improved (i.e., the best possible η_j is equal to zero), and as shown in [14], this occurs as soon as there is a "staircase", i.e., a piecewise constant vector, with a sequence of at least two consecutive increases, or two consecutive decreases, showing that in the presence of such staircases, one cannot have consistent support recovery, which is a well-known issue in signal processing (typically, more steps are created). If there is no staircase effect, we have $\eta_j = 1$ and Eq. (5) becomes $\lambda \leqslant \frac{\nu}{8} \min_j |A_j|$. If we take λ equal to the limiting value in Eq. (5), then we obtain a probability less than $1 - 4p \exp(-\frac{\nu^2 \min_j |A_j|^2}{128\sigma^2 \max_j |A_j|^2})$. Note that we could also derive general results when an additional ℓ_1-penalty is used, thus extending results from [24]. Finally, similar (more) negative results may be obtained for the two-dimensional total variation [25, 14].

Clustering with $F(A) = |A| \cdot |V \backslash A|$. In this case, we have $\eta_j = |A_j|/2$, and Eq. (5) becomes $\lambda \leqslant \frac{\nu}{4p}$, leading to the probability of correct support estimation greater than $1 - 4p \exp\big(-\frac{\nu^2}{128p\sigma^2}\big)$. This indicates that the noise variance σ^2 should be small compared to $1/p$, which is not satisfactory and would be corrected with the weighting schemes proposed in [15].

7 Conclusion

We have presented a family of sparsity-inducing norms dedicated to incorporating prior knowledge or structural constraints on the level sets of linear predictors. We have provided a set of common algorithms and theoretical results, as well as simulations on synthetic examples illustrating the 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 capabilities, or by extending some of the concepts for norms of matrices, with potential applications in matrix factorization [26] or multi-task learning [27].

Acknowledgements. This paper was partially supported by the Agence Nationale de la Recherche (MGA Project), the European Research Council (SIERRA Project) and Digiteo (BIOVIZ project).

References

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

[2] 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.

[3] F. Bach. Structured sparsity-inducing norms through submodular functions. In *Adv. NIPS*, 2010.

[4] 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.

[5] 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.

[6] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. PAMI*, 23(11):1222–1239, 2001.

[7] Z. Harchaoui and C. Lévy-Leduc. Catching change-points with Lasso. *Adv. NIPS*, 20, 2008.

[8] J.-P. Vert and K. Bleakley. Fast detection of multiple change-points shared by many signals using group LARS. *Adv. NIPS*, 23, 2010.

[9] M. Kolar, L. Song, and E. Xing. Sparsistent learning of varying-coefficient models with structural changes. *Adv. NIPS*, 22, 2009.

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

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

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

[13] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1997.

[14] F. Bach. Shaping level sets with submodular functions. Technical Report 00542949-v2, HAL, 2011.

[15] T. Hocking, A. Joulin, F. Bach, and J.-P. Vert. Clusterpath: an algorithm for clustering using convex fusion penalties. In *Proc. ICML*, 2011.

[16] 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.

[17] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. Technical Report 00613125, HAL, 2011.

[18] H. Groenevelt. Two algorithms for maximizing a separable concave function over a polymatroid feasible region. *European Journal of Operational Research*, 54(2):227–236, 1991.

[19] J. B. Orlin. A faster strongly polynomial time algorithm for submodular function minimization. *Mathematical Programming*, 118(2):237–251, 2009.

[20] M. Queyranne. Minimizing symmetric submodular functions. *Mathematical Programming*, 82(1):3–12, 1998.

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

[22] H. Hoefling. A path algorithm for the fused Lasso signal approximator. Technical Report 0910.0526v1, arXiv, 2009.

[23] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, 2010.

[24] A. Rinaldo. Properties and refinements of the fused Lasso. *Ann. Stat.*, 37(5):2922–2952, 2009.

[25] V. Duval, J.-F. Aujol, and Y. Gousseau. The TVL1 model: A geometric point of view. *Multiscale Modeling and Simulation*, 8(1):154–189, 2009.

[26] N. Srebro, J. D. M. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In *Adv. NIPS 17*, 2005.

[27] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.