Shaping Level Sets with Submodular Functions

Francis Bach

Shaping Level Sets with Submodular Functions

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

December 7, 2010

Abstract

We consider a class of sparsity-inducing regularization terms based on submodular functions. While earlier 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: this leads to a class of convex structured regularization terms that impose prior knowledge on the level sets, and not on the supports of the underlying predictors. We provide a unified set of 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, and on noisy cuts in graphs.

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 zeroes is sought so that the model is either more interpretable, cheaper to use, or simply matches available prior knowledge. This can be extended in various ways, e.g., to matrices, where sparsity of singular values is often used as a prior in several problems (see, e.g., Srebro et al., 2005; Negahban & Wainwright, 2009; Rohde & Tsybakov, 2010, and references therein).

In this paper, we 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 (Tibshirani et al., 2005; Chambolle & Darbon, 2009). In this paper, we follow the approach of Bach (2010), who designed sparsity-inducing norms based on *non-decreasing* submodular functions, as a convex approximation to imposing a specific prior on the *support* of the predictor. Here, we show that a similar parallel holds for some submodular functions which are not non-decreasing, namely non-negative set-functions which are equal to zero for the full and empty set. Our main example 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.

– In Section 4, we reinterpret existing norms and design new ones.

– We provide unified algorithms in Section 5.

– We derive unified theoretical guarantees in Section 6.

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 $V = \{1, \dots, p\}$, $1_A \in \mathbb{R}^p$ is the indicator vector of the subset A. 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 $A \subset V$, $w(A) = \sum_{k \in A} w_k = w^\top 1_A$ (this defines a modular set-function). If $w \in \mathbb{R}^p$, and $\alpha \in \mathbb{R}$, then $\{w \geqslant \alpha\}$ (resp. $\{w > \alpha\}$) denotes the subset of $V = \{1, \dots, p\}$ defined as $\{k \in V,\ w_k \geqslant \alpha\}$ (resp. $\{k \in V,\ w_k > \alpha\}$).

2 Review of Submodular Analysis

In this section, we review relevant results from submodular analysis. For more details, see, e.g., Fujishige (2005) and, for a review with proofs derived from classical convex analysis, see Bach (2010).

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:

$$\forall A, B \subset V,\ F(A) + F(B) \geqslant F(A \cup B) + F(A \cap B).$$

Without loss of generality, we assume that $F(\varnothing) = 0$. 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 $F(V) = 0$. Usual examples are symmetric submodular functions, such as cuts in a (directed or undirected) graph with vertex set V.

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 (see, e.g., Bach, 2010):

$$f(w) = \int_{-\infty}^{+\infty} F(\{w \geqslant \alpha\}) d\alpha. \tag{1}$$

The Lovász extension is convex if and only if F is submodular. Moreover, for all $A \subset V$, $f(1_A) = F(A)$ (it is indeed an extension from 2^V to $\mathbb{R}^p$); it is always positively homogeneous, and 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 $\alpha \in \mathbb{R}$).

Base polyhedron. We denote by $\mathcal{B}$ the *base polyhedron* (Fujishige, 2005), defined as the set of $s \in \mathbb{R}^p$ such that for all $A \subset V$, $s(A) \leqslant F(A)$, and $s(V) = F(V)$, i.e.,

$$\mathcal{B} = \{s \in \mathbb{R}^p,\ \forall A \subset V,\ s(A) \leqslant F(A),\ s(V) = F(V)\},$$

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 (Fujishige, 2005; Bach, 2010), i.e., for all $w \in \mathbb{R}^p$,

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

Instead of solving a linear program with 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 $k \in \{1, \ldots, p\}$, $s_{j_k} = F(\{j_1, \ldots, j_k\}) - F(\{j_1, \ldots, j_{k-1}\})$.

Faces of base polyhedron. The polyhedra $\mathcal{U} = \{w \in \mathbb{R}^p, f(w) \leqslant 1\}$ and $\mathcal{B}$ are polar to each other (Rockafellar, 1997). Therefore, the facial structure of $\mathcal{U}$ may be obtained from the one of $\mathcal{B}$. Given $s \in \mathcal{B}$, 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 $A \cap B$ (see, e.g., Bach, 2010).

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. In the case of cuts in an undirected graph, inseparable sets are exactly connected sets.

Minimization of submodular functions. Submodular functions are particularly interesting because they can be minimized in polynomial time. In this paragraph, we consider a submodular function F with potentially negative values (otherwise finding the minimum is trivial). Most algorithms for minimizing submodular functions rely on the following strong duality principle (Edmonds, 2003; Fujishige, 2005):

$$\min_{A \subset V} F(A) = \max_{s \in \mathcal{B}} \sum_{k \in V} \min\{0, s_k\}. \tag{3}$$

Moreover, algorithms for minimizing F will usually output A and s such that $F(A) = \sum_{k \in V} \min\{0, s_k\}$ as a certificate for optimality. The two main types of algorithms are combinatorial algorithms (that explicitly look for A) and ones based on convex optimization (that explicitly look for s). The first type of algorithm leads to strongly polynomial algorithms with best known complexity $O(p^6)$ (Orlin, 2009), while the minimum-norm point algorithm of Fujishige (2005) has no worst-time complexity bounds but is usually much faster in practice (Fujishige, 2005) and is based on the equivalent problem of finding the minimum-norm point in $\mathcal{B}$, i.e., $\min_{s \in \mathcal{B}} \|s\|_2^2$. Note that in this case, the minimum point algorithm also allows the construction of a particular s solution of Eq. (3)—which has several solutions in general.

Minimization of symmetric submodular functions. Such functions can be minimized in time $O(p^3)$ over all *non-trivial* (i.e., different from $\varnothing$ and V) subsets of V (Queyranne, 1998). Moreover, the algorithm is valid for the regular minimization of *posimodular* functions (Nagamochi & Ibaraki, 1998), i.e., functions that satisfy

$$\forall A, B \subset V, \ F(A) + F(B) \geqslant F(A \backslash B) + F(B \backslash A).$$

These include symmetric submodular functions as well as non-decreasing modular functions, and hence the sum of any of those.

General non-negative submodular functions. If F is non-negative but does not satisfy $F(V) = 0$, then F may be decomposed as the sum of $F - s$ and s, for $s \in \mathbb{R}^p$. If $s \in \mathcal{B} \cap \mathbb{R}^p_+$ (which is always possible for non-negative F), then $G = F - s$ is non-negative and such that $G(V) = 0$, and s is submodular non-decreasing. We can thus define a non-negative submodular set-function that is equal to zero at $\varnothing$ and V, from any non-negative submodular function.

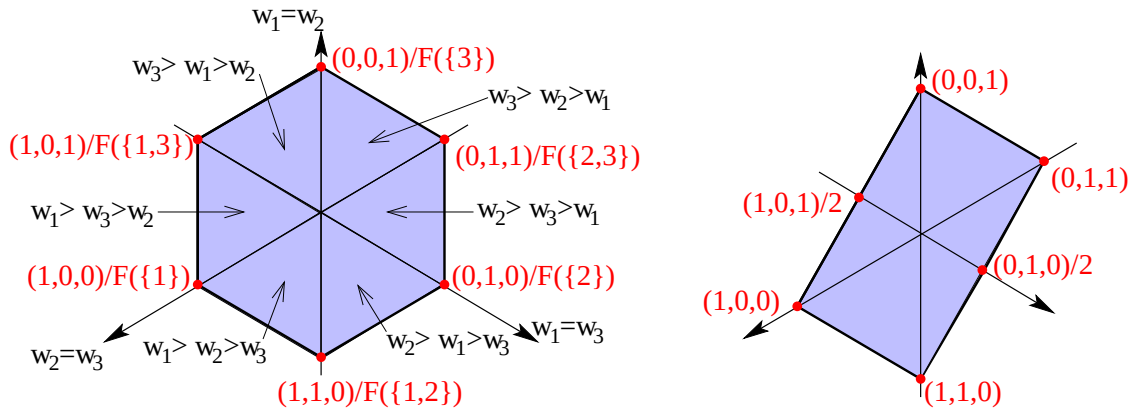

Figure 1: Top: Polyhedral level set of f (projected on the set $w(V) = w^\top 1_V = 0$), for 2 different submodular symmetric functions of three variables, with different inseparable sets leading to different sets of extreme points; changing values of F may make some of the extreme points disappear. The various extreme points cut the space into polygons where the ordering of the component is fixed. Left: $F(A) = 1_{|A|\in\{1,2\}}$ (all possible extreme points), $F(A) = 1_{1\in A,2\notin A} + 1_{2\in A,1\notin A} + 1_{2\in A,3\notin A} + 1_{3\in A,2\notin A}$ (leading to $f(w) = |w_1 - w_2| + |w_2 - w_3|$).

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.

The first proposition shows that the Lovász extension is the convex envelope of a certain combinatorial function which depends on all levets sets $\{w \geqslant \alpha\}$ of $w \in \mathbb{R}^p$. This shows that indeed regularizing by the the Lovász extension leads to imposing a prior on level sets.

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 $[0,1]^p + \mathbb{R}1_V = \{w \in \mathbb{R}^p,\ \max_{k\in V} w_k - \min_{k\in V} w_k \leqslant 1\}$.*

Note the difference with the result of Bach (2010): we consider here a different set on which computing 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 its level sets.

The next proposition gives conditions under which the Lovász extension leads to a norm, once removed the invariance by translation.

Proposition 2 (Norm) *The Lovász extension $f(w)$ is a norm on $\{w^\top 1_V = 0\}$ if and only if, for all $A \neq \varnothing$ and $A \neq V$, $F(A) > 0$.*

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

Proposition 3 (Extreme points) *The extreme points of the set $\mathcal{U} \cap \{w(V) = 0\}$ are the projections of the vectors $1_A/F(A)$ on the plane $\{w(V) = 0\}$, for A such that A is inseparable for F and $V\backslash A$ is inseparable for $B \mapsto F(A \cup B) - F(A)$.*

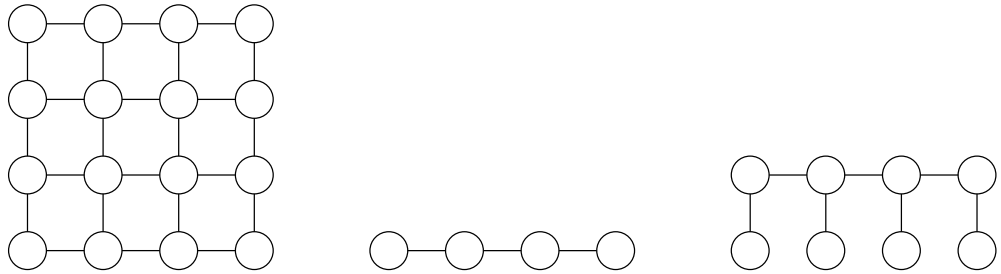

Figure 2: Examples of graphs. Left: Two-dimensional grid. Middle: chain graph. Right: hidden chain graph.

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 (cardinality-based functions).

Note that we could also consider interesting examples of construction from any non-negative function, by decomposing it as a sum of an increasing modular function and a non-negative submodular function which is equal to zero for the full set.

4.1 Cut Functions

Given a set of (non necessarily symmetric) weights $d : V \times V \to \mathbb{R}_+$, define

$$F(A) = \sum_{k \in A, j \in V \setminus A} d(k, j),$$

which we denote $d(A, V \setminus A)$. 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). If the weight function d is symmetric, then the submodular function is also symmetric and the Lovász extension is even.

In Figure 2 (left and middle plots), we give examples of usual graphs, i.e., grids in one or two dimensions, leading to total variations (Tibshirani et al., 2005; Chambolle & Darbon, 2009). Note that these functions can be extended to cuts in hypergraphs, which may have interesting applications in computer vision (Boykov et al., 2001). Moreover, directed cuts may be interesting to favor increasing or decreasing jumps along the edges of the graph.

4.2 Regular Functions

We can also consider partial minimization to obtain "regular functions" (Boykov et al., 2001; Chambolle & Darbon, 2009). Examples lead to $f(w) = \max_{k \in G} w_k - \min_{k \in G} w_k$, which corresponds to $F(A) = 1_{A \cap G \neq \varnothing} + 1_{A^c \cap G = \varnothing} - 1$, for any set $G \subset V$.

It may also lead to "noisy cuts", i.e., for a given weight function $d : V \times V \to \mathbb{R}_+$, we add p nodes (see right plot of Figure 2), each of them associated to the original nodes, and consider the associated

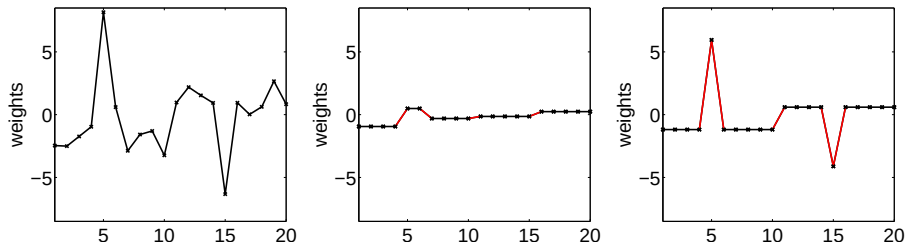

Figure 3: Estimation of noisy piecewise constant 1D signal with outliers (indices 5 and 15 in the chain of 20 nodes). Left: Original signal. Middle: best estimation with total variation (level sets are not correctly estimated). Right: best estimation with *robust* total variation (level sets are correctly estimated, with less bias).

convex and submodular functions:

$$\begin{aligned} f(w) &= \min_{v \in \mathbb{R}^p} \sum_{k,j \in V} d(k,j)(v_k - v_j)_+ + \lambda \|v - w\|_1, \\ F(A) &= \min_{A \subset V} d(B, V \backslash B) + \lambda |A \Delta B|. \end{aligned}$$

This allows for robust versions of cuts, where some gaps may be tolerated. See examples in Figure 3, illustrating the behavior of the graph displayed in the right plot of Figure 2.

4.3 Cardinality-based Functions

For $F(A) = h(|A|)$ where h is nondecreasing, 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. While these examples do not provide significantly different behaviors for the non-decreasing submodular functions explored by Bach (2010) (i.e., in terms of *support*), they lead to interesting behaviors here in terms of *level sets*. See Figure 4 for illustrations. Indeed, as shown in Section 6.1, allowed level sets A are such that A in inseparable for the function $B \mapsto F(B \cup C) - F(B)$ (for some $C \subset V$), which imposes that the interval $[|B|, |B| + |A|]$ is in the linearity space of h.

1. $F(A) = |A| \cdot |V \backslash A| = |A|(p - |A|)$. This function can be also seen as the cut in the fully connected graph. All patterns of level sets are allowed as the function h is strongly convex (see left plots of Figure 4).

2. $F(A) = 1$ if $A \neq \varnothing$ and $A \neq V$, and 0 otherwise. This function is also piecewise affine: two large level sets at the top and bottom, all the rest of the variables are in-between and separated (Figure 4, second plots from the left).

3. $F(A) = \max\{|A|, |V \backslash A|\} = \max\{|A|, p - |A|\}$. This function is piecewise affine, with only one kink, thus only one level set of cardinalty greater than one is possible, which is observed in Figure 4 (third plots from the left).

4. Other piecewise affine functions lead to selecting a given number of large level sets. For example, if $\mathcal{K} \subset \{1, \dots, p\}$, then the function $\min_{k \in \mathcal{K}} k(p - k) + (|A| - k)(p - 2k)$—which is the concave piecewise affine upper approximation of $A \mapsto |A|(p - |A|)$—leads to selecting at most $|\mathcal{K}|$ sets of cardinality strictly greater than one (see right plot of Figure 4).

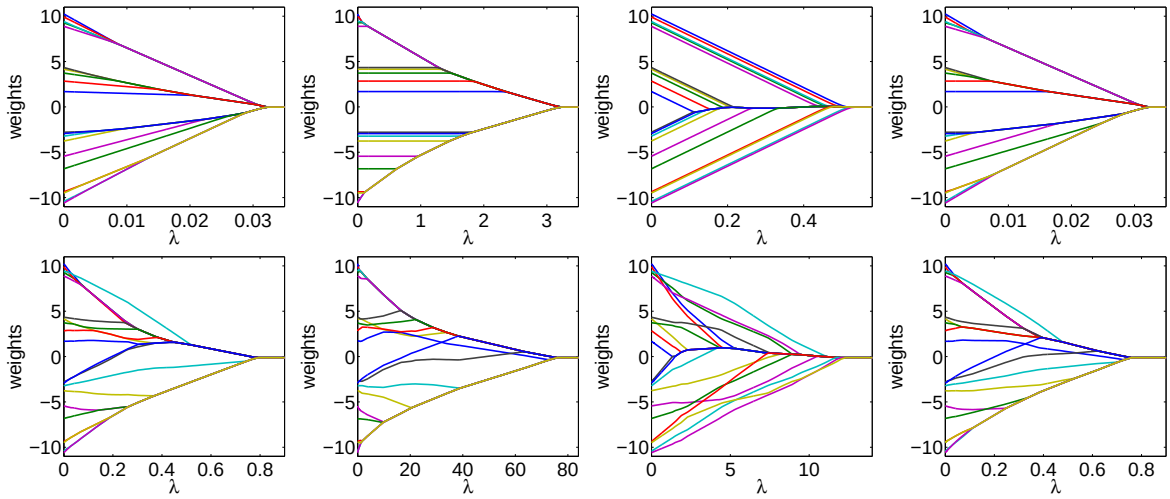

Figure 4: Piecewise linear regularization paths of proximal operators (top) and least-square regression with non-orthogonal design (bottom) for different functions of cardinality. From left to right: quadratic function (all level sets allowed), second example in Section 4.3 (two large level sets in the middle, and single variables at the top and bottom), piecewise linear with two pieces (a single large level set in the middle), piecewise linear with four pieces (four large level sets).

5 Optimization Algorithms

In this section, we present optimization methods for minimizing convex objective functions regularized by the Lovász extension of submodular functions. These lead to convex optimization problems, which we tackle using proximal methods (see, e.g., Beck & Teboulle, 2009; Nesterov, 2007).

We first start by mentioning that subgradients may easily be derived (but subgradient descent is here rather inefficient). Although we have not implemented it, note that with the square loss, the regularization paths are piecewise affine. This section follows closely the corresponding section of Bach (2010) for non-decreasing submodular functions.

Subgradient. From $f(w) = \max_{s \in \mathcal{B}} s^\top w$ 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 usually slow convergence compared to proximal methods.

Proximal operators. 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., Beck & Teboulle (2009)). In this paper, we use the methods "ISTA" and its accelerated variants "FISTA" (Beck & Teboulle, 2009).

To apply these methods, it suffices to be able to solve efficiently problems of the form: $\min_{w \in \mathbb{R}^p} \frac{1}{2}\|w - z\|_2^2 + \lambda f(w)$. In the case of the ℓ_1-norm, this reduces to soft thresholding of z, the following proposition 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 (Fujishige, 2005):

[1]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.

Proposition 4 (Proximal operator) *Let $z \in \mathbb{R}^p$ and $\lambda > 0$, minimizing $\frac{1}{2}\|w - z\|_2^2 + \lambda f(w)$ is equivalent to finding the minimum of the submodular function $A \mapsto \lambda F(A) - z(A)$ with the minimum-norm-point algorithm.*

As shown by Chambolle & Darbon (2009), if any minimization method (and not necessarily the minimum-norm-point algorithm) for the submodular function is used, we obtain that for all $\alpha \in \mathbb{R}$, the minimizers of $\lambda F(A) - z(A) + \alpha|A|$ are exactly the sets A so that $\{w > \alpha\} \subset A \subset \{w \geqslant \alpha\}$, where w is the unique optimum of the proximal problem. The proximal opearator may thus be obtained as a sequence of submodular function minimizations (by any algorithm). This can be done in general by a decomposition algorithm described by Bach (2010, Section 8.6), which can be applied to any submodular functions.

Note that using the minimum-norm-point algorithm or the decomposition algorithm leads to *generic* algorithms that can be applied to *any* submodular functions F, and that it may be rather inefficient for simpler subcases, one of which being cuts, which we now consider.

Fast optimization for cuts. The two examples in Section 4.1 and Section 4.2 are specific, because they lead to a family of submodular functions for which dedicated fast algorithms exist. Indeed, minimizing the cut functions or the partially minimized cut, plus a modular function defined by $z \in \mathbb{R}^p$, may be done with a min-cut/max-flow algorithm (see, e.g., Boykov et al., 2001; Chambolle & Darbon, 2009). For proximal methods, we need to solve an instance of a *parametric max-flow* problem, which may be done using efficient dedicated algorithms (Gallo et al., 1989; Hochbaum, 2001; Chambolle & Darbon, 2009).

Proximal path as agglomerative clustering. We consider the problem $\min_{w \in \mathbb{R}^p} \frac{1}{2}\|w - z\|_2^2 + \lambda f(w)$. When λ is equal to zero, then $w^* = z$, while for λ large enough, w is constant. In this paragraph, we provide conditions under which the regularization path may be obtained by agglomerative clustering (i.e., if two variables are together for a certain λ, they must be so for all greater λ).

Proposition 5 (Agglomerative clustering) *Assume that for all sets A, B, C such that $B \cap A = \varnothing$ and $C \subset A$, A is inseparable for $C \mapsto F(B \cup C) - F(B)$, then $|C|(F(B \cup A) - F(B)) \leqslant |A|(F(B \cup C) - F(B))$. Then the regularization path for the problem $\min_{w \in \mathbb{R}^p} \frac{1}{2}\|w - z\|_2^2 + \lambda f(w)$ is agglomerative, that is, if two variables are in the same group for a certain $\lambda_0 \in \mathbb{R}_+$, so are they for all larger $\lambda \geqslant \lambda_0$.*

The assumptions required for by Prop. 5 are satisfied by (a) all submodular set-functions that only depend on the cardinality, and (b) by one-dimensional total variation–we thus recover results from Harchaoui & Lévy-Leduc (2008); Hoefling (2009).

Adding an ℓ_1-norm. Following Tibshirani et al. (2005), 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 (Tibshirani et al., 2005; Mairal et al., 2010) to all submodular functions and their Lovász extensions.

Proposition 6 (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 $\lambda\|\cdot\|_1$.*

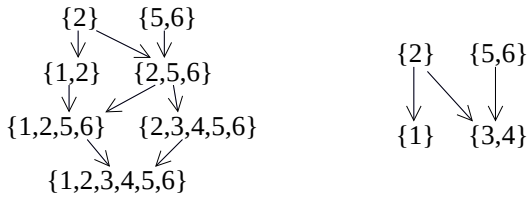

Figure 5: Left: distributive lattice with 7 elements in $2^{\{1,2,3,4,5,6\}}$. Right: corresponding poset, with 4 elements that form a partition of $\{1,2,3,4,5,6\}$ (the graph is a Hasse diagram: $A \succcurlyeq B$ if A is a descendant of B).

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:

$$\min_{w \in \mathbb{R}^p} \frac{1}{2n} \|y - Xw\|_2^2 + \lambda f(w). \tag{4}$$

We study the sparsity-inducing properties of solutions of Eq. (4), i.e., we determine in Section 6.1 which level sets are allowed and in Section 6.2 which sufficient conditions lead to consistent estimation of level sets.

6.1 Sparsity Patterns

Faces of the unit ball $\mathcal{U} = \{w,\ f(w) \leqslant 1\}$ will be defined through lattices and their equivalent poset representations, which now define.

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. Such lattices may be represented as a *partially ordered set (poset)* $\Pi(\mathcal{D}) = \{A_1, \dots, A_m\}$, where the sets A_j, $j = 1, \dots, m$ form a partition of V, which is such that elements of $\mathcal{D}$ may be identified with *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 (Fujishige, 2005). See example in Figure 5. Distributive lattices and posets are thus in one-to-one correspondence. Troughout this section, we go back and forth between these two representations.

Faces of $\mathcal{U}$. The faces of $\mathcal{U}$ are characterized by lattices $\mathcal{D}$ that contain $\varnothing$ and V, with their corresponding poset $\Pi(\mathcal{D}) = \{A_1, \dots, A_m\}$. We denote by $\mathcal{U}_{\mathcal{D}}^{\circ}$ the set of $w \in \mathbb{R}^p$ such that (a) w is piecewise constant with respect to $\Pi(\mathcal{D})$, with value v_i on A_i, and (b) for all pairs (i,j), $A_i \succcurlyeq_{\Pi(\mathcal{D})} A_j \Rightarrow v_i > v_j$. These will be interior of faces of $\mathcal{U}$, as shown in the next proposition:

Proposition 7 (Faces of $\mathcal{U}$) *The 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 maximal chains $\varnothing = S_0 \subset \cdots \subset S_m = V$ of $\mathcal{D}$, the sets S_j, $j \in \{1, \dots, m\}$ are inseparable for the submodular function $U_j \mapsto F(S_{j-1} \cup U_j) - F(S_{j-1})$,

(iii) for all maximal chains $\varnothing = S_0 \subset \cdots \subset S_p = V$ of 2^V compatible with $\mathcal{D}$, then, if s is defined by $s(S_j) = F(S_j)$ for all $j \in \{1, \dots, p\}$, then $\mathcal{D}$ is the set of tight sets for s.

Allowed level sets. In this section, we do not make any assumptions regarding the correct specification of the linear model. We show that with probability one, only certain level sets are allowed. For simplicity, we assume invertibility of $X^\top X$, but we could consider assumptions similar to the ones used by Jenatton et al. (2009).

Proposition 8 (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. (4) is unique and, with probability one, all level sets of $\hat{w}$ correspond to a face of $\mathcal{U}$, i.e., to a set $\mathcal{U}_\mathcal{D}^\circ$ for a certain lattice $\mathcal{D}$ that satisfies the conditions of Prop. 7.*

Note the difference with the analysis of Bach (2010), who considered a specific aspect of the faces of the unit ball, namely the stability of the support.

Cut functions. For cut functions, inseparability for the restricted submodular function is equivalent to inseparability for F, i.e., connectedness in the graph. Thus only connected level sets $\{w = \alpha\}$ are allowed.

6.2 Theoretical Analysis

In this section, we consider a particular interior of a face $\mathcal{U}_\mathcal{D}^\circ$, for a lattice $\mathcal{D}$ and its associated poset $\Pi(\mathcal{D}) = (A_1, \dots, A_m)$. We denote by $M = (1_{A_1}, \dots, 1_{A_m})$ the indicator matrix of the clusters.

Decomposability. We consider an element w of $\mathcal{U}_\mathcal{D}^\circ$; then, for all $t \in \mathbb{R}^p$ small enough so that for all pairs (i,j), $A_i \succcurlyeq_{\Pi(\mathcal{D})} A_j \Rightarrow (w+t)_{A_i} \geqslant (w+t)_{A_j}$, then we have the decomposition $f(w+t) = f(w) + \sum_{j=1}^m f_j(t_{A_j})$, where f_j is the Lovász extension of the submodular function defined on A_j, by $F_j : B \mapsto F(A_1 \cup \cdots \cup A_{j-1} \cup B) - F(A_1 \cup \cdots \cup A_{j-1})$. We have assumed that the sets A_j are ordered along a maximal chain of $\mathcal{D}$ (and by condition (i) of Prop. 7, the function f_j is independent of the choice of the chain). We now use the notation $B_i = A_1 \cup \cdots \cup A_i$.

Support/levels sets recovery. We can now use the decomposability property to derive a general support recovery result. We consider the submodular function defined on 2^{A_j}:

$$G_j : C_j \mapsto F(B_{j-1} \cup C_j) - F(B_{j-1}) - 1_{C_j}^\top QM(M^\top QM)^{-1}t,$$

where $t_j = F(B_{j-1} \cup A_j) - F(B_{j-1})$. It is submodular, such that $G_j(\varnothing) = 0$, and $G_j(A_j) = F(B_{j-1} \cup A_j) - F(B_{j-1}) - \delta_j^\top M^\top QM(M^\top QM)^{-1}t = 0$. We make that the assumption that G_j is *strictly positive* for all nontrivial subsets C_j of A_j (see examples below). We denote by $g_j^*(s)$ the function on $\mathbb{R}^{A_j}$, defined as

$$g_j^*(s) = \max_{C_j \subset A_j, C_j \notin \{\varnothing, A_j\}} \frac{s(C_j)}{G_j(C_j)}.$$

The next theorem provides a general result regarding the recovery of level sets.

Theorem 1 (Level set recovery) *Assume that $y = Xw^* + \sigma\varepsilon$, where ε is a standard multivariate normal vector. Let $Q = \frac{1}{n}X^\top X \in \mathbb{R}^{p \times p}$. Assume w^* belongs to the interior of the face $\mathcal{U}_\mathcal{D}^\circ$, i.e., $w^* = Mv^*$ with $\eta = \min_{A_i \succcurlyeq_{\Pi(\mathcal{D})} A_j} \{v_i^* - v_j^*\} > 0$. Assume that all G_j are strictly positive on non-trivial subsets of A_j.*

Moreover, assume $\lambda\|(M^\top QM)^{-1}t\|_\infty \leqslant \eta/4$ *and* $M^\top QM$ *invertible. Then, with probability greater than*

$$1 - \sum_{j=1}^{m} \mathbb{P}(g_j^*(z_j) > \lambda n^{1/2}\sigma^{-1}) - \sum_{j=1}^{m} \exp\Big(- \frac{n\eta^2}{32\sigma^2((M^\top QM)^{-1})_{jj}}\Big),$$

the estimate $\hat{w}$ *is on the same face than* w^**, with* z_j *normal with zero mean and covariance matrix* $Q_{A_jA_j} - Q_{A_jV}M(M^\top QM)^{-1}M^\top Q_{VA_j}$*.*

Following similar results in support recovery for the ℓ_1-norm or norms associated with non-decreasing submodular functions (Bach, 2010), there are two types of assumptions:

(a) $\eta > 0$ corresponds to having non-zero components which are bounded away from zero (i.e., we are in the interior of a face, and bounded away from its boundary).

(b) Strict positivity of G_j corresponds to the irrepresentable condition. The main difference is that for support recovery, this assumption is always met for the orthogonal design, while here it is not always met, even for the orthogonal design. Interestingly, the validity of level set recovery for the orthogonal design is related to the agglomerativity of proximal paths.

Application to identity design matrix. If we have $Q = I$ and $\tau^2 = \sigma^2/n$. The constraints in the assumptions of Theorem 1 are that, for all $j \in \{1, \dots, m\}$:

$$\lambda\Big|\frac{F(B_j) - F(B_{j-1})}{|A_j|}\Big| \leqslant \eta/4.$$

The function G_j is then equal to:

$$G_j : C_j \mapsto F(B_{j-1} \cup C_j) - F(B_{j-1}) - |C_j|\frac{F(B_{j-1} \cup A_j) - F(B_{j-1})}{|A_j|}.$$

The strict positivity of G_j is not always satisfied, but it is for the 1D total variation and certain functions of the cardinality (as we show below).

Moreover, we have

$$\sum_{j=1}^{m} \exp\Big(- \frac{n\eta^2}{32\sigma^2((M^\top QM)^{-1})_{jj}}\Big) = \sum_{j=1}^{m} \exp\Big(- \frac{n\eta^2|A_j|}{32\sigma^2}\Big).$$

Concentration inequality. For a general submodular function F which is non-negative and satisfies $F(V) = 0$, we can compute an upper bound to $\mathbb{P}(f^*(s) \geqslant u)$ for s normally distributed with zero mean and covariance matrix Σ that has 1_V as a singular vector (which is our case for z_j and G_j in Theorem 1). A simple technique is to use the union bound:

$$\begin{aligned}
f^*(s) &= \max_{A \text{ inseparable}} \frac{s(A)}{F(A)} \\
\mathbb{P}(f^*(s) \geqslant u) &\leqslant \sum_{A \text{ inseparable}} \mathbb{P}(s(A) \geqslant uF(A)) \\
&\leqslant \sum_{A \text{ inseparable}} \exp\Big(- \frac{u^2F(A)^2}{2 \times 1_A^\top \Sigma 1_A}\Big).
\end{aligned}$$

Figure 6: Signal approximation with two-dimensional total variation: For two piecewise constant images with two values, the estimation may (left case) or may not (right case) recover the correct level sets, even with infinitesimal noise. For the two cases, left: original pattern, right: best possible recovered level sets.

One-dimensional total variation. The constraints become $\lambda \leqslant \frac{\eta}{4} \min_j |A_j|$. Moreover, $\frac{G_j(C_j)^2}{|C_j|(1-|C_j|/|A_j|)}$ is lower bounded by $1/|A_j|^2$ for all non trivial inseparable subsets of A_j, and the number of inseparable sets is less than $|A_j|^2$. Thus, we have the probability in Theorem 1 greater than

$$1 - \sum_{j=1}^{m} \exp\Big(- \frac{\eta^2 |A_j|}{32\tau^2} \Big) - \sum_{j=1}^{m} |A_j|^2 \exp\Big(- \frac{\lambda^2}{2\tau^2 |A_j|^2} \Big).$$

We get a probability greater than

$$1 - 2\sum_{j=1}^{m} |A_j|^2 \exp\Big(- \frac{\lambda^2}{2\tau^2 |A_j|^2} \Big).$$

If we choose $\lambda = 4\tau \max_j |A_j| \sqrt{\log p}$, which imposes that $\min_j |A_j| \eta/4 \geqslant 4\tau \max_j |A_j| \sqrt{\log p}$ (i.e., the level sets have approximatively the same number of elements depending on the noise level), we have a probability of correct level set recovery (for no design matrix) which is greater than $1 - 2/p$. Note that we could also derive general results when an additional ℓ_1-penalty is used, thus extending results from Rinaldo (2009).

Two-dimensional total variation. While for the one-dimensional version, the assumption with respect to the positivity of functions G_j is always satisfied for $Q = I$, this is not the case anymore for the two-dimensional version, which has always been noticed in continuous settings (see, e.g., Duval et al., 2009). We illustrate this in Figure 6, where we show that depending on the shape of the level sets (which still have to be connected), we may not recover the correct pattern, even with very small noise.

Clustering with $F(A) = |A| \cdot |A^c|$. If we assume that the design is orthogonal, then $G_j(C_j) = |C_j| \cdot |A_j \backslash C_j|$, and, we can use

$$g_j^*(z_j) \leqslant \frac{2}{|A_j|} \|z_j\|_\infty,$$

to get the following lower bound on the probability of correct support recovery:

$$1 - \sum_{j=1}^{m} \exp\Big(- \frac{\lambda^2 |A_j|^2}{8\tau^2} \Big) - \sum_{j=1}^{m} \exp\Big(- \frac{\eta^2 |A_j|}{32\tau^2} \Big),$$

with the constraint that $\lambda \leqslant \frac{\eta}{4}$. We also get a lower bound on the sizes of the clusters.

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 or multi-task learning (see, e.g., Krause & Cevher, 2010, for application of submodular functions to dictionary learning).

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).

A Proof of Proposition 1

Let $s \in \mathbb{R}^p$, we consider the function $g : w \mapsto \max_{\alpha \in \mathbb{R}} F(\{w \geqslant \alpha\})$, and we compute its Fenchel conjugate (see, e.g. Boyd & Vandenberghe, 2004):

$$\begin{aligned} g^*(s) &= \max_{w \in [0,1]^p + \mathbb{R}1_V} w^\top s - g(w) \\ &= \max_{(A_1,\dots,A_m) \text{ partition}} \quad \max_{t_1 > \cdots > t_m,\ t_1 - t_m \leqslant 1} \sum_{j=1}^{m-1} (t_j - t_{j+1}) s(A_1 \cup \cdots \cup A_j) + t_m s(V) \\ &\qquad\qquad - \max_{j \in \{1,\dots,m\}} F(A_1 \cup \cdots \cup A_j) \\ &= \iota_{s(V)=0}(s) + \max_{(A_1,\dots,A_m) \text{ partition}} \left\{ \max_{j \in \{1,\dots,m\}} s(A_1 \cup \cdots \cup A_j) - \max_{j \in \{1,\dots,m\}} F(A_1 \cup \cdots \cup A_j) \right\}, \end{aligned}$$

where $\iota_{s(V)=0}$ is the indicator function of the set $\{s(V) = 0\}$ (with values 0 or $+\infty$).

Let $h(s) = \iota_{s(V)=0}(s) + \max_{A \subset V} s(A) - F(A)$. We clearly have $g^*(s) \geqslant h(s)$, because we take a maximum over a larger set (consider $m = 2$). Moreover, for all partitions $(A_1, \dots, A_m)$, if $s(V) = 0$, $\max_{j \in \{1,\dots,m\}} s(A_1 \cup \cdots \cup A_j) \leqslant \max_{j \in \{1,\dots,m\}} (h(s) + F(A_1 \cup \cdots \cup A_j)) = h(s) + \max_{j \in \{1,\dots,m\}} F(A_1 \cup \cdots \cup A_j)$, which implies that $g^*(s) \leqslant h(s)$. Thus $g^*(s) = h(s)$.

Moreover, we have, for f symmetric,

$$\begin{aligned} \max_{w \in [0,1]^p + \mathbb{R}1_V} w^\top s - f(w) &= \iota_{s(V)=0}(s) + \max_{w \in [0,1]^p} w^\top s - f(w) \\ &= \iota_{s(V)=0}(s) + \max_{A \subset V} s(A) - F(A) = h(s). \end{aligned}$$

Thus f and g have the same Fenchel conjugates. The result follows from the convexity of f.

B Proof of Proposition 2

What needs to be proved is that $f(w) = 0$ implies that w is a constant vector. This is implied by the strict positivity of F on all non-trivial subsets of V and the definition of the Lovász extension f.

C Proof of Proposition 3

Extreme points of $\mathcal{U}$ correspond to full-dimensional faces of $\mathcal{B}$. From Fujishige (2005, Corollary 3.4.4), these facets are exactly the ones that correspond to sets A with the given conditions. These facets are defined as the intersection of $\{s(A) = F(A)\}$ and $\{s(V) = F(V)\}$, which leads to the desired result.

D Proof of Proposition 5

Optimality conditions for a w in a certain face interior $\mathcal{U}_{\mathcal{D}}^{\circ}$, i.e., which is constant on A_i (with value v_i) and such that $v_i > v_j$ for all $A_i \succcurlyeq_{\Pi(\mathcal{D})} A_j$ are that:

(a) v minimizes
$$\frac{1}{2}\|z - Mv\|_2^2 + \lambda t^\top v,$$
where $t_i = F(A_1 \cup \cdots \cup A_i) - F(A_1 \cup \cdots \cup A_{i-1})$ and $M \in \mathbb{R}^{p \times m}$ is the matrix of indicator vectors of the sets A_i. We thus get
$$v = (M^\top M)^{-1}(M^\top z - \lambda t).$$

(b) v is such that $(z - Mv)/\lambda \in \mathcal{B}$.

Note that
$$z - Mv = (I - M(M^\top M)^{-1}M^\top)z + \lambda M(M^\top M)^{-1}t,$$
and that for all $i \in \{1, \ldots, m\}$,
$$1_{A_i}^\top (I - M(M^\top M)^{-1}M^\top)z = 0,$$
and
$$1_{A_i}^\top M(M^\top M)^{-1}t = t_i = F(A_1 \cup \cdots \cup A_i) - F(A_1 \cup \cdots \cup A_{i-1}),$$
so that, if $B_i = A_1 \cup \cdots \cup A_i$, $[(I - M(M^\top M)^{-1}M^\top)z](B_i) = 0$, $[M(M^\top M)^{-1}t](B_i) = F(B_i)$.

Thus if $(z - Mv)/\lambda \in \mathcal{B}$ for a certain λ, with our assumption $(z - Mv)/\mu \in \mathcal{B}$ for all μ larger than λ. Indeed, we then have
$$\forall C_i \subset A_i,\ [M(M^\top M)^{-1}t](C_i) \leqslant F(B_{i-1} \cup C_i) - F(B_{i-1}),$$
which implies by submodularity that $[M(M^\top M)^{-1}t](C) \leqslant F(C)$ for all $C \subset V$. Indeed, we have:
$$\begin{aligned}
[M(M^\top M)^{-1}t](C) &= \sum_{j=1}^{m} [M(M^\top M)^{-1}t](C \cap A_j) \\
&\leqslant \sum_{i=1}^{m} F(B_{i-1} \cup (C \cap A_j)) - F(B_{i-1}) \\
&\leqslant \sum_{i=1}^{m} F((B_{i-1} \cap C) \cup (C \cap A_j)) - F(B_{i-1} \cap C) \\
&= \sum_{i=1}^{m} F(B_i \cap C) - F(B_{i-1} \cap C) = F(C).
\end{aligned}$$
This shows that when μ increases, the only way to break the optimality conditions is by having some $v_i = v_j$, i.e., clusters A_i and A_j merge.

E Proof of Proposition 6

We denote by w the unique mininum of $\frac{1}{2}\|w-z\|_2^2 + f(w)$ and s the associated dual vector in $\mathcal{B}$. The optimality conditions lead to $w = z - s$. We assume that w takes disctinct values $v_1, \dots, v_m$ on the sets $A_1, \dots, A_m$. We define t as $t_k = \operatorname{sign}(w_k)(|w_k| - \lambda)_+$. The level sets of t are A_j, for j such that $|v_j| > \lambda$ and zero for the unions of A_j such that $|v_j| \leqslant \lambda$. We have $t = z - s + t - w$, with $t - w$ optimal for $\max_{\|u\|_\infty \leqslant 1} u^\top w$. We only need to show that s is optimal for $\max_{s \in \mathcal{B}} s^\top w$, which true since the level sets of w are finer than the ones of t, with no change of ordering (see Bach, 2010).

F Proof of Proposition 7

Given that the polyhedra $\mathcal{U}$ and $\mathcal{B}$ are polar to each other (Rockafellar, 1997), the proposition follows from (Fujishige, 2005, Theorem 3.43).

G Proof of Theorem 1

We denote by $M = (1_{A_1}, \dots, 1_{A_m})$ the indicator matrices of the clusters, and we assume that $w^* = Mv^*$. We first minimize with respect to $w = Mv$ for $v \in \mathbb{R}^m$. The cost function may be written as

$$\frac{1}{2}(w - w^*)^\top Q(w - w^*) - q^\top (w - w^*) + \lambda f(w),$$

where $Q = \frac{1}{n} X^\top X \in \mathbb{R}^{p \times p}$ and $q = \sigma X^\top \varepsilon / n \in \mathbb{R}^p$. If $w = Mv$, we have to minimize

$$\frac{1}{2}(v - v^*)^\top M^\top Q M (v - v^*) - q^\top M (v - v^*) + \lambda v^\top t,$$

with $t_i = F(A_1 \cup \cdots \cup A_i) - F(A_1 \cup \cdots \cup A_{i-1})$, with solution $\hat{v} = v^* + (M^\top Q M)^{-1}(M^\top q - \lambda t)$. Then $\hat{w} = M\hat{v}$ is the optimal solution if (note that it is then the unique solution because of the invertibility of $M^\top QM$):

(a) $\hat{v}_i > \hat{v}_j$ as soon as $A_i \succcurlyeq_{\Pi(\mathcal{D})} A_j$; this is implied by

$$\max\{\|(M^\top QM)^{-1} M^\top q\|_\infty, \|\lambda (M^\top QM)^{-1} t\|_\infty\} \leqslant \eta/4,$$

for $\eta = \min_{A_i \succcurlyeq_{\Pi(\mathcal{D})} A_j} v_i^* - v_j^*$. Note that $(M^\top QM)^{-1} M^\top q$ is normally distributed with zero mean and covariance matrix

$$\frac{\sigma^2}{n}(M^\top QM)^{-1},$$

which leads to the second part of the probability using standard bound on Gaussian tails.

(b) For all Δ such that for all j, $\Delta(A_j) = 0$, we have $\lambda f_j(\Delta_{A_j}) + (Q_{A_j V} M(\hat{v} - v^*) - q_{A_j})^\top \Delta_{A_j} \geqslant 0$, for f_j the Lovász extension corresponding to the submodular function $C_j \mapsto F(C_j \cup B_{j-1}) - F(B_{j-1})$. This is equivalent to

$$\lambda f_j(\Delta_{A_j}) + (Q_{A_j V} M (M^\top QM)^{-1}(M^\top q - \lambda t) - q_{A_j})^\top \Delta_{A_j} \geqslant 0$$

$$\lambda f_j(\Delta_{A_j}) \geqslant \lambda \Delta_{A_j}^\top Q_{A_j V} M (M^\top QM)^{-1} t + \Delta_{A_j}^\top (q_{A_j} - Q_{A_j V} M (M^\top QM)^{-1} M^\top q),$$

$$\lambda f_j(\Delta_{A_j}) \geqslant \lambda \Delta_{A_j}^\top Q_{A_j V} M (M^\top Q M)^{-1} t + \frac{\sigma}{n} \Delta_{A_j}^\top X_{A_j}^\top (I - X^\top M (M^\top X^\top X M)^{-1} M^\top X^\top)\varepsilon,$$

This is equivalent to $Q_{A_j V} M (M^\top Q M)^{-1} t + \frac{\sigma}{\lambda n} X_{A_j}^\top (I - X^\top M (M^\top X^\top X M)^{-1} M^\top X^\top)\varepsilon \in \mathcal{B}_j + \mathbb{R} 1_{A_j}$, where $\mathcal{B}_j$ is the base polyhedron associated with F_j.

Note that for any submodular function, $s \in B(F) + \mathbb{R} 1_V$ if and only if for all $A \subset V$, $s(A) - \frac{s(V)}{|V|}|A| \leqslant F(A) - \frac{F(V)}{|V|}|A|$.

Moreover, $z_{A_j} = \frac{1}{n^{1/2}} X_{A_j}^\top (I - X^\top M (M^\top X^\top X M)^{-1} M^\top X^\top)\varepsilon$ is normally distributed with covariance matrix $Q_{A_j A_j} - Q_{A_j V} M (M^\top Q M)^{-1} M^\top Q_{V A_j}$, and this matrix has 1_{A_j} has a singular vector, since $Q_{A_j V} M (M^\top Q M)^{-1} M^\top Q_{V A_j} 1_{A_j} = Q_{A_j A_j} 1_{A_j}$.

The condition for global optimality is thus exactly that $g_j^*(z_{A_j}) \leqslant \frac{\lambda n^{1/2}}{\sigma}$.

References

Bach, F. Structured sparsity-inducing norms through submodular functions. In *Advances in Neural Information Processing Systems*, 2010.

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

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

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

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

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

Duval, Vincent, Aujol, Jean-Francois, and Gousseau, Yann. The tvl1 model: A geometric point of view. *Multiscale Modeling and Simulation*, 8(1):154–189, 2009.

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

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

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

Harchaoui, Z. and Lévy-Leduc, C. Catching change-points with Lasso. In *Advances in Neural Information Processing Systems*, 2008.

Hochbaum, D.S. An efficient algorithm for image segmentation, Markov random fields and related problems. *Journal of the ACM (JACM)*, 48(4):686–701, 2001.

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

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

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

Mairal, J., Bach, F., Ponce, J., and Sapiro, G. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, 2010. ISSN 1532-4435.

Nagamochi, H. and Ibaraki, T. A note on minimizing submodular functions. *Information Processing Letters*, 67(5):239–244, 1998.

Negahban, S. and Wainwright, M.J. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. Technical Report 0912.5100, Arxiv, 2009.

Nesterov, Y. Gradient methods for minimizing composite objective function. Technical report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, Tech. Rep, 2007.

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

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

Rinaldo, A. Properties and refinements of the fused Lasso. *Ann. Statist*, 37(5B):2922–2952, 2009.

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

Rohde, A. and Tsybakov, A.B. Estimation of high-dimensional low-rank matrices. Technical Report 0912.5338, arXiv, 2010.

Srebro, N., Rennie, J. D. M., and Jaakkola, T. S. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems 17*, 2005.

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