Structured Variable Selection with Sparsity-Inducing Norms

Rodolphe Jenatton, Jean-Yves Audibert, Francis Bach

Structured Variable Selection with Sparsity-Inducing Norms

Rodolphe Jenatton[1,2]
rodolphe.jenatton@inria.fr

Jean-Yves Audibert [2,3]
audibert@certis.enpc.fr

Francis Bach [1,2]
francis.bach@inria.fr

[1] INRIA
[2] WILLOW project-team
[3] Imagine, Université Paris-Est

April 22, 2009

Abstract

We consider the empirical risk minimization problem for linear supervised learning, with regularization by structured sparsity-inducing norms. These are defined as sums of Euclidean norms on certain subsets of variables, extending the usual ℓ_1-norm and the group ℓ_1-norm by allowing the subsets to overlap. This leads to a specific set of allowed nonzero patterns for the solutions of such problems. We first explore the relationship between the groups defining the norm and the resulting nonzero patterns, providing both forward and backward algorithms to go back and forth from groups to patterns. This allows the design of norms adapted to specific prior knowledge expressed in terms of nonzero patterns. We also present an efficient active set algorithm, and analyze the consistency of variable selection for least-squares linear regression in low and high-dimensional settings.

1 Introduction

Regularization by the ℓ_1-norm is now a widespread tool in machine learning, statistics and signal processing: it allows linear variable selection in potentially high dimensions, with both efficient algorithms [Efron et al., 2004, Lee et al., 2007] and well-developed theory for generalization properties and variable selection consistency [Zhao and Yu, 2006, Wainwright, 2006].

However, the ℓ_1-norm cannot easily encode prior knowledge about the patterns of nonzero coefficients ("nonzero patterns") induced in the solution, since they are all theoretically possible. Group ℓ_1-norms [Yuan and Lin, 2006, Roth and Fischer, 2008] consider a partition of all variables into a certain number of subsets and penalize the sum of the Euclidean norms of each one,

[2]WILLOW project-team, Laboratoire d'Informatique de l'Ecole Normale Supérieure, ENS/INRIA/CNRS UMR 8548, 45 rue d'Ulm 75230 Paris Cedex, France.
[3]Imagine, Université Paris-Est, 6 avenue Blaise Pascal, 77455 Marne-la-Vallée, France.

leading to selection of groups rather than individual variables. Moreover, recent works have considered overlapping but nested groups in constrained situations such as trees and directed acyclic graphs [Zhao et al., 2009, Bach, 2008c].

In this paper, we consider all possible sets of groups and characterize exactly what type of prior knowledge can be encoded by considering sums of norms of overlapping groups of variables. We first describe how to go from groups to nonzero patterns (or equivalently zero patterns), then show that is possible to "reverse-engineer" a given set of nonzero patterns, i.e., to build the unique minimal set of groups that will generate these patterns. This allows the automatic design of sparsity-inducing norms, adapted to target sparsity patterns. We give in Section 3 some interesting examples of such designs on two-dimensional grids.

As will be shown in Section 3, for each set of groups, a notion of hull of a nonzero pattern may be naturally defined. In the particular case of the two-dimensional planar grid considered in this paper, this hull is exactly the axis-aligned bounding box or the regular convex hull. We show that, in our framework, the allowed nonzero patterns are exactly those equal to their hull,

and that the hull of the relevant variables is consistently estimated under certain conditions, both in low and high-dimensional settings. Moreover, we present in Section 4 an efficient active set algorithm that scales up to high dimensions. Finally, we illustrate in Section 6 the behavior of our norms with synthetic examples on two-dimensional grids.

Notation. For $x \in \mathbb{R}^p$ and $q \in [1, \infty]$, we denote by $\|x\|_q$ its ℓ_q-norm. Given $w \in \mathbb{R}^p$ and a subset J of $\{1, \dots, p\}$ with cardinality $|J|$, w_J denotes the vector in $\mathbb{R}^{|J|}$ of elements of w indexed by J. Similarly, for a matrix $A \in \mathbb{R}^{p \times p}$, A_{IJ} denotes the (I, J)-block of A. For two vectors u and v in $\mathbb{R}^p$, we denote by $u \circ v = (u_1 v_1, \dots, u_p v_p)^\top \in \mathbb{R}^p$ the elementwise product between u and v.

2 Regularized Risk Minimization

We consider the problem of predicting a random variable $Y \in \mathcal{Y}$ from a (potentially non random) vector $X \in \mathbb{R}^p$, where $\mathcal{Y}$ is the set of responses, typically a subset of $\mathbb{R}$. We assume that we are given n observations $(x_i, y_i) \in \mathbb{R}^p \times \mathcal{Y}$, $i = 1, \dots, n$. We define the *empirical risk* of a loading vector $w \in \mathbb{R}^p$ as $L(w) = \frac{1}{n} \sum_{i=1}^n \ell\left(y_i, w^\top x_i\right)$, where $\ell : \mathcal{Y} \times \mathbb{R} \mapsto \mathbb{R}^+$ is a *loss function*. We assume that ℓ is *convex and continuously differentiable* with respect to the second parameter. Typical examples of loss functions are the square loss for least squares regression, i.e., $\ell(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$ with $y \in \mathbb{R}$, and the logistic loss $\ell(y, \hat{y}) = \log(1 + e^{-y\hat{y}})$ for logistic regression, with $y \in \{-1, 1\}$.

We focus on a general family of sparsity-inducing norms that allow the penalization of subsets of variables grouped together. Let us denote by $\mathcal{G}$ a subset of the power set of $\{1, \dots, p\}$ such that $\bigcup_{G \in \mathcal{G}} G = \{1, \dots, p\}$, i.e., a spanning set of subsets of $\{1, \dots, p\}$. Note that $\mathcal{G}$ does not necessarily define a partition of $\{1, \dots, p\}$, and therefore, *it is possible for elements of $\mathcal{G}$ to overlap*. We consider the norm Ω defined by

$$\Omega(w) = \sum_{G \in \mathcal{G}} \left(\sum_{j \in G} (d_j^G)^2 |w_j|^2 \right)^{\frac{1}{2}} = \sum_{G \in \mathcal{G}} \left\| d^G \circ w \right\|_2 ,$$

where $(d^G)_{G \in \mathcal{G}}$ is a collection of p-dimensional vectors (whose impact will be analyzed in Section 6) such that $d_j^G > 0$ if $j \in G$ and $d_j^G = 0$ otherwise.

This general formulation has several important subcases that we present below, the goal of this paper being to go beyond these, and to consider norms capable to incorporate richer prior knowledge.

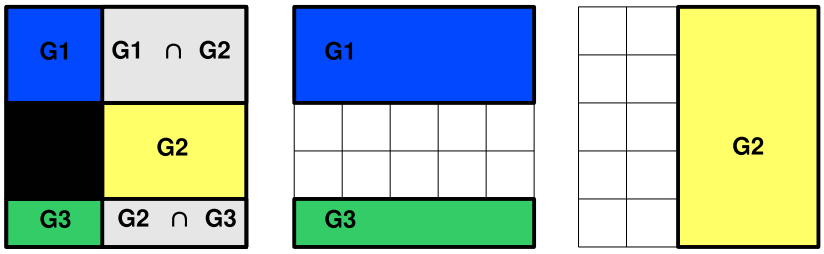

Figure 1: Examples of groups and patterns of zeros/nonzeros: three sparsity-inducing groups (middle and right) with the associated nonzero pattern which is the union of the complements of the groups (left, in black)

- ℓ_2**-norm**: $\mathcal{G}$ is composed of one element, the full set $\{1, \dots, p\}$.
- ℓ_1**-norm**: $\mathcal{G}$ is the set of all singletons, leading to the Lasso [Tibshirani, 1996] for the square loss.
- ℓ_2**-norm and** ℓ_1**-norm**: $\mathcal{G}$ is the set of all singletons and the full set $\{1, \dots, p\}$, leading (up to the squaring of the ℓ_2-norm) to the Elastic net [Zou and Hastie, 2005] for the square loss.
- **Group** ℓ_1**-norm**: $\mathcal{G}$ is any partition of $\{1, \dots, p\}$, leading to the group-Lasso for the square loss [Yuan and Lin, 2006].
- **Hierarchical norms**: when the set $\{1, \dots, p\}$ is embedded into a tree [Zhao et al., 2009] or more generally into a directed acyclic graph [Bach, 2008c], then a set of p groups, each of them composed of descendants of a given variable, is considered.

We study the following regularized problem:

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell \left(y_i, w^\top x_i \right) + \mu \Omega(w), \tag{2.1}$$

where $\mu \geqslant 0$ is a regularization parameter. Note that a non-regularized intersect could be included in this formulation. We denote by $\hat{w}$ any solution of Eq. (2.1). Regularizing by linear combinations of (non-squared) ℓ_2-norms is known to induce sparsity in $\hat{w}$ [Zhao et al., 2009]; our grouping leads to specific patterns that we describe in the next section.

3 Groups and Sparsity Patterns

We now study the relationship between the norm Ω and the nonzero patterns the estimated vector $\hat{w}$ is allowed to have. We first characterize the set of nonzero patterns, then we provide forward and backward procedures to go back and forth from groups to patterns.

3.1 Stable Patterns Generated by $\mathcal{G}$

The regularization term $\Omega(w) = \sum_{G \in \mathcal{G}} \|d^G \circ w\|_2$ is a mixed (ℓ_1, ℓ_2)-norm [Zhao et al., 2009]. At the group level, it behaves like an ℓ_1-norm and, therefore, Ω induces group sparsity (each w_G is encouraged to go to zero); on the other hand, within the groups $G \in \mathcal{G}$, the ℓ_2-norm does not promote sparsity. Intuitively, some of the vectors w_G associated with certain groups G will be

exactly equal to zero, leading to a set of zeros which is the union of these groups G in $\mathcal{G}$. Thus, the set of allowed zero patterns should be the *union-closure* of $\mathcal{G}$, i.e. (see Figure 1 for an example):

$$\mathcal{Z} = \left\{ \bigcup_{G \in \mathcal{G}'} G;\ \mathcal{G}' \subseteq \mathcal{G} \right\}. \tag{3.1}$$

The situation is however slightly more subtle as some zeros can be created by chance (just as regularizing by the ℓ_2-norm may lead, though it is unlikely, to some zeros). Nevertheless, Theorem 3.1 (see proof in Appendix A) ensures us that, under mild conditions, the previous intuition about the set of (non)zero patterns is correct. Before stating the result more precisely, we need to introduce the concept of *$\mathcal{G}$-adapted hull*, or simply hull, that represents the granularity associated to the set of groups $\mathcal{G}$. For any subset $I \subseteq \{1, \dots, p\}$, we define

$$\mathrm{Hull}(I) = \left\{ \bigcup_{G \in (\mathcal{G}_I)^c} G \right\}^c,$$

with $\mathcal{G}_I = \{G \in \mathcal{G};\ G \cap I \neq \varnothing\}$. Note that we always have $I \subseteq \mathrm{Hull}(I)$. As we shall see later, the hull has a clear geometrical interpretation for specific sets $\mathcal{G}$.

Theorem 3.1. *Assume that $Y = (y_1, \dots, y_n)^\top$ is a realization of an absolutely continuous probability distribution. Let us consider the following optimization problem*

$$\min_{w \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^{n} (y_i - w^\top x_i)^2 + \mu \Omega(w), \tag{3.2}$$

and let denote by Q the Gram matrix $\frac{1}{n}\sum_{i=1}^n x_i x_i^\top$.

If for all solution $\hat{w}$ of (3.2) with nonzero pattern $\hat{I} = \{j \in \{1, \dots, p\};\ \hat{w}_j \neq 0\}$, the matrix $Q_{\mathrm{Hull}(\hat{I})\mathrm{Hull}(\hat{I})}$ is invertible, then the problem (3.2) has a unique solution whose set of zeros is in $\mathcal{Z} = \left\{\bigcup_{G \in \mathcal{G}'} G;\ \mathcal{G}' \subseteq \mathcal{G}\right\}$ almost surely.

It is important to note that Theorem 3.1 does not require the invertibility of the full matrix Q. The result can therefore hold in high-dimensional settings where the number of observations n is smaller than the number of variables p (in this case, Q is always singular).

Instead of considering the set of zero patterns $\mathcal{Z}$, it is also convenient to manipulate nonzero patterns, and we define

$$\mathcal{P} = \left\{ \bigcap_{G \in \mathcal{G}'} G^c;\ \mathcal{G}' \subseteq \mathcal{G} \right\} = \{Z^c;\ Z \in \mathcal{Z}\}. \tag{3.3}$$

We can equivalently use $\mathcal{P}$ or $\mathcal{Z}$ by taking the complement of each element of these sets. We have the following usual special cases from Section 2 (we give more examples in Section 3.5):

- **ℓ_2-norm**: the set of allowed nonzero patterns is composed of the empty set and the full set $\{1, \dots, p\}$.
- **ℓ_1-norm**: $\mathcal{P}$ is the set of all possible subsets.
- **ℓ_2-norm and ℓ_1-norm**:$\mathcal{P}$ is also the set of all possible subsets.

- **Grouped ℓ_1-norm**: $\mathcal{P} = \mathcal{Z}$ is the set of all possible unions of the elements of the partition defining $\mathcal{G}$.
- **Hierarchical norms**: the set of patterns $\mathcal{P}$ is then all sets J for which all ancestors of elements in J are included in J [Bach, 2008c].

Two natural questions arise: (1) starting from the groups $\mathcal{G}$, is there an efficient way to generate the set of nonzero patterns $\mathcal{P}$; (2) conversely, and more importantly, given $\mathcal{P}$, how can the groups $\mathcal{G}$—and hence the norm $\Omega(w)$—be designed?

3.2 General Properties of $\mathcal{G}$, $\mathcal{Z}$ and $\mathcal{P}$

Closedness. The set $\mathcal{Z}$ (resp. $\mathcal{P}$) is closed under union (resp. intersection), that is, for all $K \in \mathbb{N}$ and all $z_1, \ldots, z_K \in \mathcal{Z}$, $\bigcup_{k=1}^K z_k \in \mathcal{Z}$ (resp. $p_1, \ldots, p_K \in \mathcal{P}$, $\bigcap_{k=1}^K p_k \in \mathcal{P}$). This implies that when "reverse-engineering" the set of nonzero patterns, we have to assume it is closed under intersection. Otherwise, the best we can do is to deal with its intersection-closure.

Minimality. If a group in $\mathcal{G}$ is the union of other groups, it may be removed from $\mathcal{G}$ without changing the sets $\mathcal{Z}$ or $\mathcal{P}$. This is the main argument behind the pruning backward algorithm in Section 3.4. Moreover, this leads to the notion of a *minimal* set $\mathcal{G}$ of groups, which is such that for all $\mathcal{G}' \subseteq \mathcal{Z}$ whose union-closure spans $\mathcal{Z}$, we have $\mathcal{G} \subseteq \mathcal{G}'$. The existence and unicity of a minimal set is a consequence of classical results in set theory [Doignon and Falmagne, 1998]. The elements of this minimal set are usually referred to as the *atoms* of $\mathcal{Z}$.

Minimal sets of groups are attractive in our setting because they lead to a smaller number of groups and lower computational complexity—for example, for 2 dimensional-grids with rectangular patterns, we have a quadratic possible number of rectangles, i.e., $|\mathcal{Z}| = O(p^2)$, that can be generated by a minimal set $\mathcal{G}$ whose size is $|\mathcal{G}| = O(\sqrt{p})$.

Hull. We recall the definition of the *$\mathcal{G}$-adapted hull*, namely, for any subset $I \subseteq \{1, \ldots, p\}$,

$$\mathrm{Hull}(I) = \left\{ \bigcup_{G \in (\mathcal{G}_I)^c} G \right\}^c,$$

with $\mathcal{G}_I = \{G \in \mathcal{G};\ G \cap I \neq \varnothing\}$. It basically represents the granularity associated to the set of groups $\mathcal{G}$.

If the set $\mathcal{G}$ is formed by all vertical and horizontal half-spaces when the variables are organized in a 2 dimensional-grid (see Figure 4), the hull of a subset $I \subset \{1, \ldots, p\}$ is simply the axis-aligned bounding box of I. Similarly, when $\mathcal{G}$ is the set of all half-spaces with all orientations (e.g., orientations $\pm\pi/4$ are shown in Figure 5), the hull is the regular convex hull. Note that those interpretations of the hull are possible and valid only when we have geometrical information at hand about the set of variables.

Graphs of patterns. We consider the directed acyclic graph (DAG) stemming from the *Hasse diagram* of the partially ordered set (poset) $(\mathcal{G}, \supset)$. The nodes of this graph are the elements G of $\mathcal{G}$ and there is a directed edge from G_1 to G_2 if and only if $G_1 \supset G_2$ and there exists no $G \in \mathcal{G}$ such that $G_1 \supset G \supset G_2$ [Cameron, 1994]. We can also build the corresponding DAG for the set of zero patterns $\mathcal{Z} \supset \mathcal{G}$, which is a super-DAG of the DAG of groups (see Figure 2 for examples).

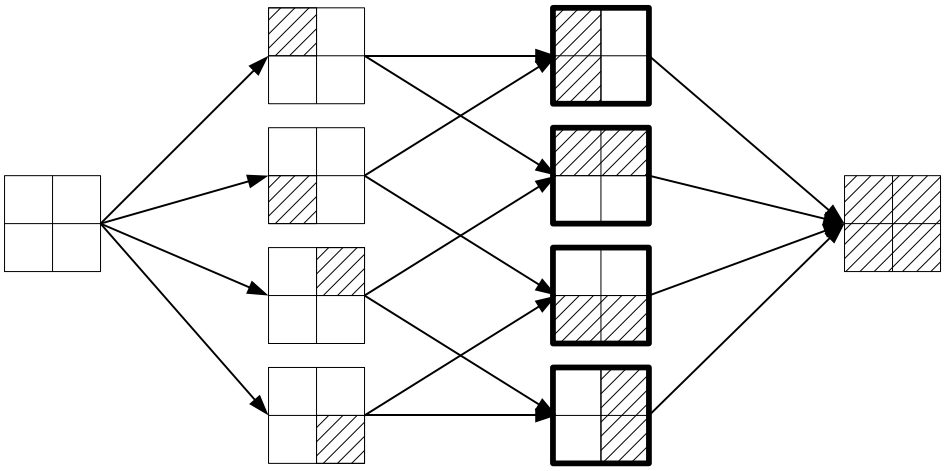

Figure 2: The DAG for the set $\mathcal{Z}$ associated with the 2×2-grid. The members of $\mathcal{Z}$ are the complement of the areas hatched in black. The elements of $\mathcal{G}$ (i.e., the atoms of $\mathcal{Z}$) are highlighted by bold edges.

Note that we obtain also the isomorphic DAG for the nonzero patterns $\mathcal{P}$, although it corresponds to the poset $(\mathcal{P}, \subset)$: this DAG will be used in the active set algorithm presented in Section 4.

Prior works with nested groups [Zhao et al., 2009, Bach, 2008c] have used a similar DAG, which was isomorphic to a DAG on the variables because of the specificity of the hierarchical norm. As opposed to those cases where the DAG was used to give an additional structure to the problem, the DAG we introduce here on the set of groups naturally and always comes up, with no assumption on the variables themselves (for which no DAG is defined in general).

3.3 From Groups to Patterns

The *forward* procedure presented in Algorithm 1, taken from Doignon and Falmagne [1998], allows the construction of $\mathcal{Z}$ from $\mathcal{G}$. It iteratively builds the collection of patterns by taking unions, and has complexity $O(p|\mathcal{Z}||\mathcal{G}|^2)$. The general scheme is straightforward. Namely, by considering increasingly larger subfamilies of $\mathcal{G}$ and the collection of patterns already obtained, all possible unions are formed. However, some attention needs to be paid while checking we are not generating a pattern already encountered. Such a verification is performed by the *if* condition within the inner loop of the algorithm. Indeed, we do not have to scan the whole collection of patterns already obtained (whose size can be exponential in $|\mathcal{G}|$), but we rather use the fact that $\mathcal{G}$ is the base of $\mathcal{Z}$. Note that in general, it is not possible to upper bound the size of $|\mathcal{Z}|$ by a polynomial term in p, even when $\mathcal{G}$ is very small (indeed, $|\mathcal{Z}| = 2^p$ for the ℓ_1-norm).

3.4 From Patterns to Groups

We now assume that we want to impose a priori knowledge on the sparsity structure of $\hat{w}$. This information can be exploited by restricting the patterns allowed by the norm Ω. Namely, from an intersection-closed set of zero patterns $\mathcal{Z}$, we can build back a minimal set of groups $\mathcal{G}$ by iteratively pruning away in the DAG corresponding to $\mathcal{Z}$, all sets which are unions of their parents. See Algorithm 2.

This algorithm can be found under a different form in [Doignon and Falmagne, 1998]—we present it through a pruning algorithm on the DAG, which is natural in our context (the proof of

Algorithm 1 Forward procedure

Input: Set of groups $\mathcal{G} = \{G_1, \ldots, G_M\}$
Output: Collection of zero patterns $\mathcal{Z}$ and nonzero patterns $\mathcal{P}$
Intialization: $\mathcal{Z} = \{\varnothing\}$
for $m = 1$ **to** M **do**
  $C = \{\varnothing\}$
  for each $Z \in \mathcal{Z}$ **do**
    if $(G_m \nsubseteq Z)$ **and**
    $(\forall G \in \{G_1, \ldots, G_{m-1}\},\ G \subseteq Z \cup G_m \Rightarrow G \subseteq Z)$ **then**
      $C \leftarrow C \cup \{Z \cup G_m\}$
    end if
  end for
  $\mathcal{Z} \leftarrow \mathcal{Z} \cup C$
end for
$\mathcal{P} = \{Z^c;\ Z \in \mathcal{Z}\}$.

Algorithm 2 Backward procedure

Input: Intersection-closed family of nonzero patterns $\mathcal{P}$
Output: Set of groups $\mathcal{G}$
Initialization: Compute $\mathcal{Z} = \{P^c;\ P \in \mathcal{P}\}$ and set $\mathcal{G} = \mathcal{Z}$.
Build the Hasse diagram for the poset $(\mathcal{Z}, \supset)$.
for $t = \min_{G \in \mathcal{Z}} |G|$ **to** $\max_{G \in \mathcal{Z}} |G|$ **do**
  for each node with $G \in \mathcal{Z}$ with $|G| = t$ **do**
    if $\left(\bigcup_{C \in \text{Children}(G)} C = G\right)$ **then**
      if $(\text{Parents}(G) \neq \varnothing)$ **then**
        Connect children of G to parents of G
      end if
      Remove G from $\mathcal{G}$
    end if
  end for
end for

the minimality of the procedure can be found in Appendix B). The complexity of Algorithm 2 is $O(p|\mathcal{Z}|^2 + \sum_{z\in\mathcal{Z}} \sum_{c\in\text{Children}(z)} |c|)$. The pruning may reduce significantly the number of groups necessary to generate the whole set of zero patterns, sometimes from exponential in p to polynomial in p (e.g., the ℓ_1-norm). We now give examples where $|\mathcal{G}|$ is also polynomial in p.

3.5 Examples

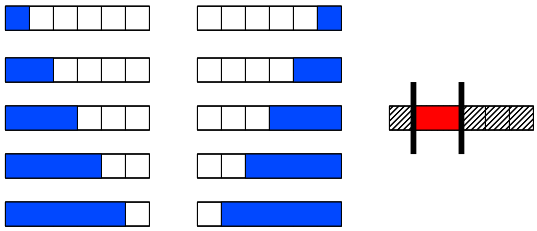

Figure 3: Groups (in blue) that select contiguous patterns in a sequence. On the right, an example of such a pattern (in red).

Sequences Given p variables organized in a sequence, if we want only contiguous nonzero patterns, the backward algorithm will lead to the set of groups which are intervals $[1, k]_{k\in\{1,\dots,p-1\}}$ and $[k, p]_{k\in\{2,\dots,p\}}$, with both $|\mathcal{Z}| = O(p)$ and $|\mathcal{G}| = O(p)$ (see Figure 3).

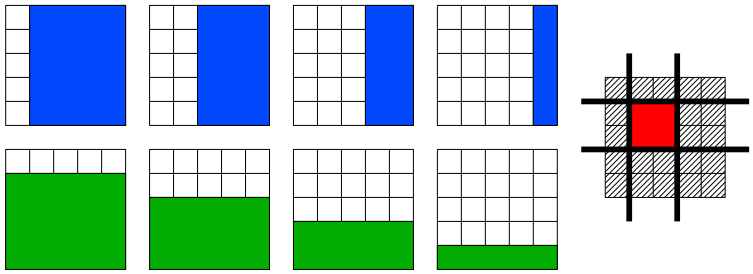

Figure 4: Vertical and horizontal groups (the other half of the groups to form $\mathcal{G}$ are obtained by symmetry) and an example of pattern (in red) that can be recovered in this setting.

Two-dimensional grids In Section 6, we consider for $\mathcal{P}$, the set of all rectangles in two dimensions, leading by the previous algorithm to the set of axis-aligned half-spaces for $\mathcal{G}$ (see Figure 4), with $|\mathcal{Z}| = O(p^2)$ and $|\mathcal{G}| = O(\sqrt{p})$. This type of structure is encountered in object or scene recognition, where the selected rectangle would correspond to a certain box inside an image, that concentrates the predictive power for a given class of object/scene.

By adding more half-planes to $|\mathcal{G}|$ with different angles than 0 and $\pi/2$, the set of nonzero patterns $\mathcal{P}$ tends to the convex sets in the two-dimensional grid [Soille, 2003]. See Figure 5. The number of groups is linear in $\sqrt{p}$ with constant growing linearly with the number of angles, while $|\mathcal{Z}|$ grows more rapidly (typically non-polynomially in the number of angles). This type of structure could be useful in vision as well as in neuroscience, in particular to retrieve brain activity in EEG data, which is usually a small convex-like portion of the scalp.

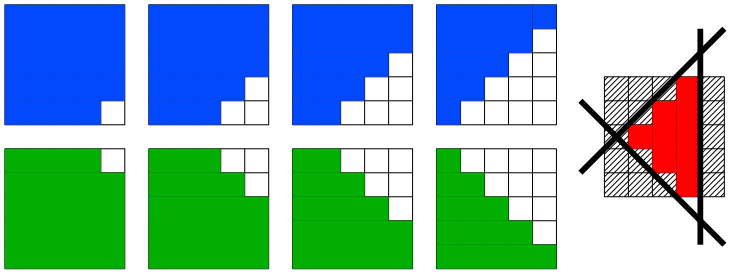

Figure 5: Oblique $\pm\pi/4$ groups (the other half of the groups to form $\mathcal{G}$ are obtained by symmetry) and an example of pattern (in red) that can be recovered in this setting.

4 Active Set Algorithm

For moderate values of p, one may obtain a solution for Eq. (2.1) using generic toolboxes for second-order cone programming—in this paper, we use CVX [Grant and Boyd, 2008], whose time complexity is equal to $O(p^{3.5} + |\mathcal{G}|^{3.5})$, which is not appropriate when p or $|\mathcal{G}|$ are large.

We present in this section an *active set algorithm* (Algorithm 3) that finds a solution for Eq. (2.1) by considering increasingly larger active sets and checking global optimality at each step, with total complexity in $O(p^{1.5})$. Here, the sparsity prior can be used for computational advantages.

It is simpler to derive the algorithm for the following constrained optimization problem—which has the same solution set as the regularized problem of Eq. (2.1) when μ and λ are allowed to vary:

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, w^\top x_i\right) \text{ s. t. } \sum_{G \in \mathcal{G}} \left\|d^G \circ w\right\|_2 \leqslant \lambda. \tag{4.1}$$

In active set methods, the set of nonzero variables, denoted by J, is built incrementally, and the problem is solved only for this reduced set of variables, adding the constraint $w_{J^c} = 0$ to Eq. (4.1). We denote by $L(w) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i, w^\top x_i\right)$ the empirical risk (which is by assumption convex and continuously differentiable) and by L^* its *Fenchel-conjugate*, defined as $L^*(u) = \sup_{w \in \mathbb{R}^p}\{w^\top u - L(w)\}$ [Boyd and Vandenberghe, 2003]. The restriction of L to $\mathbb{R}^{|J|}$ is denoted $L_J(w_J) = L(\tilde{w})$ for $\tilde{w}_J = w_J$ and $\tilde{w}_{J^c} = 0$, with Fenchel-conjugate L_J^*.

For a potential active set $J \subset \{1, \ldots, p\}$ which belongs to the set of allowed nonzero patterns $\mathcal{P}$, we consider the reduced norm $\Omega_J(w_J) = \sum_{G \in \mathcal{G}} \left\|d_J^G \circ w_J\right\|_2$ on $\mathbb{R}^{|J|}$, and its *dual norm* $\Omega_J^*(\kappa_J) = \max_{\Omega_J(w_J) \leqslant 1} w_J^\top \kappa_J$. The next proposition gives the optimization problem dual to the constrained reduced problem (Eq. (4.2) below):

Proposition 4.1. *Let $J \subseteq \{1, \ldots, p\}$. The following two problems:*

$$\min_{\Omega_J(w_J) \leqslant \lambda} L_J(w_J), \tag{4.2}$$

$$\max_{\kappa_J \in \mathbb{R}^{|J|}} -L_J^*(-\kappa_J) - \lambda \Omega_J^*(\kappa_J), \tag{4.3}$$

are dual to each other. In addition, at optimality, we have $-\kappa_J = \nabla L_J(w_J)$.

Proof. The proposition comes from a classic result of Fenchel Duality [Borwein and Lewis, 2006, Theorem 3.3.5 and Exercise 3.3.9] when we consider the convex function

$$h_J : w_J \mapsto \begin{cases} 0 & \text{if } \Omega_J(w_J) \leqslant \lambda, \\ \infty & \text{otherwise.} \end{cases}$$

The Fenchel conjugate of h_J is given by $\kappa_J \mapsto \lambda\Omega_J^*(\kappa_J)$ [Boyd and Vandenberghe, 2003, Exercise 3.26]. Since the set

$$\{w_J \in \mathbb{R}^{|J|};\ h_J(w_J) < \infty\} \cap \{w_J \in \mathbb{R}^{|J|};\ L_J(w_J) < \infty \text{ and } L_J \text{ is continuous at } w_J\}$$

is not empty, we get the first part of the proposition. At optimality, we have moreover the following relationship between the primal-dual variables

$$-\kappa_J \in \partial L_J(w_J) = \{\nabla L_J(w_J)\}$$

where the equality is a consequence of the differentiability of L_J at w_J. We thus have the second part of the proposition.

□

The duality gap of the previous optimization problem is

$$\begin{aligned} &L_J(w_J) + L_J^*(-\kappa_J) + \lambda\Omega_J^*(\kappa_J) \\ =&\Big\{L_J(w_J)+L_J^*(-\kappa_J)+w_J^\top\kappa_J\Big\}+\Big\{\lambda\Omega_J^*(\kappa_J)-w_J^\top\kappa_J\Big\}, \end{aligned}$$

which is a sum of two nonnegative terms (the first nonnegativity is from the Fenchel-Young inequality, while the second one is from the definition of the dual norm). For any feasible w_J, i.e., such that $\Omega_J(w_J) \leq \lambda$, if we choose $\kappa_J = -\nabla L_J(w_J)$, the duality gap then reduces to $\lambda\Omega_J^*(\kappa_J) - w_J^\top\kappa_J$.

In order to check that the reduced solution w_J is optimal for the full problem in Eq. (4.1), we pad w_J with zeros on J^c to define w, compute $\kappa = -\nabla L(w)$, which is such that $\kappa_J = -\nabla L_J(w_J)$, and get a duality gap for the full problem equal to

$$\lambda\Omega^*(\kappa)-w^\top\kappa = \lambda\Omega^*(\kappa)-w_J^\top\kappa_J = \lambda[\Omega^*(\kappa)-\Omega_J^*(\kappa_J)].$$

Computing this gap requires solving an optimization problem which is as hard as the original one, prompting the need for upper and lower bounds. Given a set $J \in \mathcal{P}$, we denote by $\mathcal{G}_J$ the set of active groups, i.e., the set of groups $G \in \mathcal{G}$ such that $G \cap J \neq \varnothing$.

In the light of Theorem 3.1, we can interpret the active set algorithm as a walk through the DAG of nonzero patterns allowed by the norm Ω. The parents $\Pi_{\mathcal{P}}(J)$ of J are exactly the patterns containing the variables that may enter the active set at the next iteration of Algorithm 3. The groups that are exactly at the boundaries of the active set (referred to as the *fringe groups*) are $\mathcal{F}_J = \{G \in (\mathcal{G}_J)^c\ ;\ \nexists G' \in (\mathcal{G}_J)^c,\ G \subseteq G'\}$. Those groups are not contained by any other inactive groups of $(\mathcal{G}_J)^c$. We are interested in a particular subset of $\mathcal{F}_J$ defined as $\bigcup_{K\in\Pi_{\mathcal{P}}(J)} \mathcal{G}_K \backslash \mathcal{G}_J$ (see Figure 6 and Figure 7). This subset can be viewed as the equivalent of $\Pi_{\mathcal{P}}(J)$ for groups.

We have the following optimality conditions (see proofs in Appendix C) that control the progress of Algorithm 3 :

Proposition 4.2 (Necessary condition)**.** *If w is optimal for the full problem in Eq. (4.1), then*

$$\lambda \max_{K\in\Pi_{\mathcal{P}}(J)} \left\{ \sum_{j\in K\backslash J} \frac{|\nabla L(w)_j|^2}{\left(\sum_{H\in\mathcal{G}_K\backslash\mathcal{G}_J, H\ni j} d_j^H\right)^2} \right\}^{\frac{1}{2}} \leqslant -w^\top\nabla L(w). \qquad (N)$$

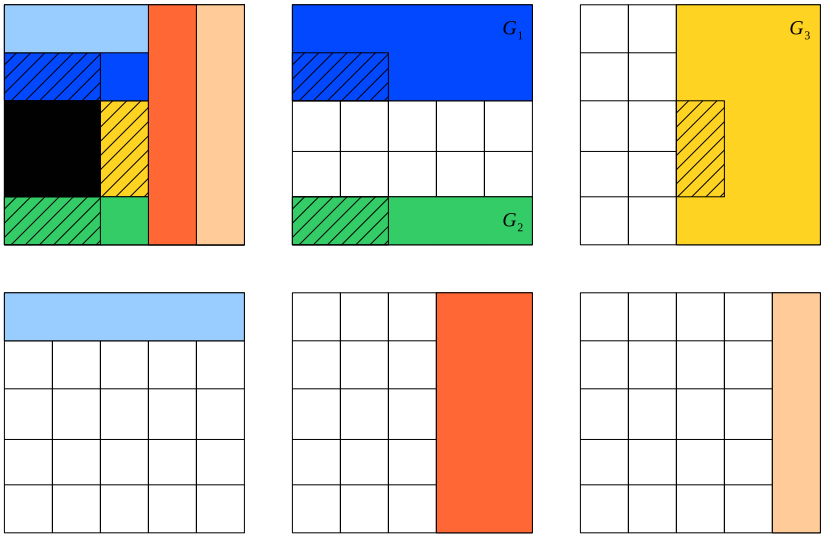

Figure 6: The active set (black) and the candidate patterns of variables, i.e. the variables in $K \backslash J$ (hatched in black) that can become active. The fringe groups are exactly the groups that have the hatched areas (i.e., here we have $\mathcal{F}_J = \bigcup_{K \in \Pi_{\mathcal{P}}(J)} \mathcal{G}_K \backslash \mathcal{G}_J = \{G_1, G_2, G_3\}$).

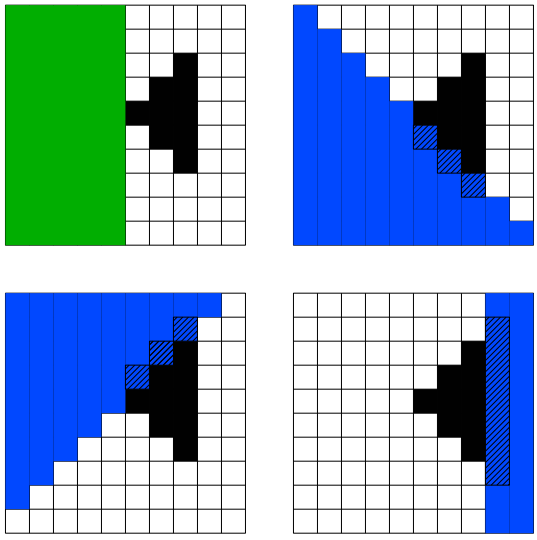

Figure 7: The active set (black) and the candidate patterns of variables, i.e. the variables in $K \backslash J$ (hatched in black) that can become active. The groups in blue are those in $\bigcup_{K \in \Pi_{\mathcal{P}}(J)} \mathcal{G}_K \backslash \mathcal{G}_J$, while the green group is in $\mathcal{F}_J \backslash (\bigcup_{K \in \Pi_{\mathcal{P}}(J)} \mathcal{G}_K \backslash \mathcal{G}_J)$.

Proposition 4.3 (Sufficient condition)**.** *If*

$$\lambda \max_{G \in \mathcal{F}_J} \left\{ \sum_{j \in G} \frac{|\nabla L(w)_j|^2}{\left(\sum_{H \in (\mathcal{G}_J)^c, H \ni j} d_j^H \right)^2} \right\}^{\frac{1}{2}} \leqslant \varepsilon - w^\top \nabla L(w), \qquad (S_\varepsilon)$$

then w is a solution whose duality gap for Eq. (4.1) is less than $\varepsilon \geq 0$.

Note that for the Lasso, the conditions (N) and (S_0) (i.e., the sufficient condition taken with

$\varepsilon = 0$) are both equivalent to the condition $\|\nabla L(w)_{\mathbf{J}^c}\|_\infty \leqslant -w^\top \nabla L(w)$, which is the usual optimality condition [Wainwright, 2006, Tibshirani, 1996]. Moreover, when they are not satisfied, our two conditions provide good heuristics for choosing which $K \in \Pi_{\mathcal{P}}(J)$ to add in the active set.

More precisely, since the necessary condition (N) directly deals with the *variables* (as opposed to groups) that can become active at the next step of Algorithm 3, it suffices to choose the $K \in \Pi_{\mathcal{P}}(J)$ that violates the condition most.

The heuristic for the sufficient condition (S_ε) implies to go from groups to variables. When we have $\mathcal{F}_J = \bigcup_{K \in \Pi_{\mathcal{P}}(J)} \mathcal{G}_K \backslash \mathcal{G}_J$, we simply consider the groups $G \in \mathcal{F}_J$ that violates the sufficient condition most and then take the largest pattern of variables $K \in \Pi_{\mathcal{P}}(J),\ K \cap G \neq \varnothing$, to enter the active set. Note that this situation always happens when we deal with the rectangular groups. In general, we just have the inclusion $(\bigcup_{K \in \Pi_{\mathcal{P}}(J)} \mathcal{G}_K \backslash \mathcal{G}_J) \subseteq \mathcal{F}_J$ and the group that achieves the maximum in (S_ε) might not be a group that is allowed (according to the DAG of nonzero patterns) to become active at this step (see Figure 7). For such a group $H \in \mathcal{F}_J \backslash (\bigcup_{K \in \Pi_{\mathcal{P}}(J)} \mathcal{G}_K \backslash \mathcal{G}_J)$, we look at all the groups $G \in (\bigcup_{K \in \Pi_{\mathcal{P}}(J)} \mathcal{G}_K \backslash \mathcal{G}_J),\ H \cap G \neq \varnothing$ and follow the scheme described before.

A direct consequence of this heuristic is that it is possible for the algorithm to *jump over* the right active set and to consider instead a (slightly) larger active set as optimal. However if the active set is larger than the optimal set, then (it can be proved that) the sufficient condition (S_0) is satisfied, and the reduced problem, which we solve exactly, will still output the correct nonzero pattern.

Moreover, it is worthwile to notice that in Algorithm 3, the active set may sometimes be increased only to make sure that the current solution is optimal (we only check a sufficient condition of optimality).

Algorithm 3 Active set algorithm

Input: Data $\{(x_i, y_i), i = 1, \ldots, n\}$, regularization parameter μ
Duality gap ε, maximum number of variables s
Output: Active set J, loading vector $\hat{w}$
Intialization: $J = \{\varnothing\}$, $\hat{w} = 0$
while (N) is not satisfied **and** $|J| \leqslant s$ **do**
  replace J by violating $K \in \Pi_{\mathcal{P}}(J)$
  solve the reduced problem $\min_{\Omega_J(w_J) \leqslant \lambda} L_J(w_J)$ to get $\hat{w}$
end while
while (S_ε) is not satisfied **and** $|J| \leqslant s$ **do**
  replace J by violating $K \in \Pi_{\mathcal{P}}(J)$
  solve the reduced problem $\min_{\Omega_J(w_J) \leqslant \lambda} L_J(w_J)$ to get $\hat{w}$
end while

Algorithmic complexity. We analyse in detail the time complexity of the active set algorithm when we consider the rectangular groups with $|\mathcal{G}| = O(\sqrt{p})$. The running time necessary to obtain the solution of the reduced problem depends on the number of active groups $|G_J| = |\{G \in \mathcal{G};\ G \cap J \neq \varnothing\}|$, i.e., the groups that contain at least one variable of the active set. Thus, if the number of active variables is upper bounded by $s \ll p$ (which is true if our target is actually sparse), the time complexity of Algorithm 3 is the sum of:

- the computation of the gradient, $O(s\,n\,p)$ for the square loss.

- the cost of the solver, $O(s \max_{J\in\mathcal{P},|J|\leqslant s} |G_J|^{3.5} + s^{4.5})$.

- t_1 times the computation of (N), that is $O(t_1(\sqrt{p}+s) + p^{1.5})$. Indeed, computing $\Pi_{\mathcal{P}}(J)$ costs $O(1)$ with $|\Pi_{\mathcal{P}}(J)| \leq 4$ (i.e., the four edges of the hull for the rectangular groups) and for $K \in \Pi_{\mathcal{P}}(J)$, $|K| = O(s)$ (corresponding to the stripe of variables around the active set) with $|\mathcal{G}_K \backslash \mathcal{G}_J| = 1$. Besides, without elaborated data structures, the cost of getting $\mathcal{G}_K$ is $O(\sqrt{p})$.

  During the initialization (i.e., $J = \varnothing$), we have $|\Pi_{\mathcal{P}}(\varnothing)| = O(p)$ (since we can start with any singletons), and $|\mathcal{G}_K \backslash \mathcal{G}_J| = |\mathcal{G}_K| = O(\sqrt{p})$, which leads to a complexity of $O(p^{1.5})$ for the sum $\sum_{G\in\mathcal{G}_K, j\in G}$. Note however that this sum does not depend on J and can therefore be cached if we need to make several runs with the same set of groups $\mathcal{G}$.

- t_2 times the computation of (S_ε), that is $O(t_2(\sqrt{p}+p^{1.5}))$ with $t_1 + t_2 \leq s$. Computing $\mathcal{F}_J$ requires $O(\sqrt{p})$ with $|\mathcal{F}_J| \leq 4$ (i.e., the four edges of the hull for the rectangular groups) and for all $G \in \mathcal{F}_J$, $|G|$ is upperbounded by $O(p)$. In addition, $|(\mathcal{G}_J)^c| \leq |\mathcal{G}| = O(\sqrt{p})$, so that the computation of $\sum_{G\in(\mathcal{G}_J)^c, G\ni j}$ costs $O(p^{1.5})$.

We finally get complexity in $O(p^{1.5})$, which is much better than $O(p^{3.5})$, without an active set method. Note that the term $s^{4.5}$ could be improved upon by using warm-restart strategies for the sequence of reduced problems.

In our experiments, i.e., with rectangular and $\pm\frac{\pi}{4}$ groups on a 2 dimensional-grid (and more generally, for collections of nested groups along n_δ different directions), we still have a complexity in $O(p^{1.5})$. The main change lies in the computation of the parents $\Pi_{\mathcal{P}}(J)$. It requires to call upon the forward algorithm (see Algorithm 1), from the fringe groups $\mathcal{F}_J$ (whose size is now n_δ) reduced to the set of variables $\left[\bigcup_{G\in(\mathcal{G}_J)^c\backslash\mathcal{F}_J} G\right]^c \backslash J$, whose size is in $O(s)$. Hence, the cost of obtaining the patterns $K \in \Pi_{\mathcal{P}}(J)$ becomes $O(s\, n_\delta^2 \max_{J\in\mathcal{P},|J|\leqslant s} |\Pi_{\mathcal{P}}(J)|)$, as opposed to a constant time before.

Nonzero pattern intersection. We have seen so far how overlapping groups can encore prior information about a desired set of (non)zero patterns. In practice, controlling these overlaps may be delicate and hinges on the choice of the weights $(d^G)_{G\in\mathcal{G}}$ (see the experiments in the Section 6).

Hovewer, it is possible to keep the benefit of overlapping groups whilst limiting their side effects, by taking up the idea of support intersection [Bach, 2008a]. First introduced to stabilize the set of variables recovered by the Lasso, we reuse this technique in a different context, based on the following remark. When dealing with collections of nested groups along multiple directions, the two procedures described below actually lead to the same set of (non)zero patterns:

a) Considering one model with the norm Ω composed of all the groups (i.e., the groups for all directions).

b) First considering one model per direction (the norm Ω of this model being only comprised of the nested groups corresponding to that direction) and then taking the intersection of the nonzero patterns obtained for each of those models. In the example of the sequence (see Figure 3), it boils down to consider one model with the groups starting from the left and one model with the groups starting from the right [1].

Note that, in this second setting, although the training of several models is required (a number of times equals to the number of directions considered, e.g., 2 for the sequence and 4 for the

[1] To be more precise, in order to regularize every variable, we have to add the group $\{1,\dots,p\}$ to each model.

rectangular groups), each of those trainings involves a smaller number of groups. The experiments (see Section 6) will show the superiority of this second approach.

5 Pattern Consistency

In this section, we analyze the model consistency of the solution of Eq. (2.1) for the square loss. Considering the set of nonzero patterns $\mathcal{P}$ derived in Section 3, we can only hope to estimate the correct hull of the generating sparsity pattern, since Theorem 3.1 states that other patterns occur with zero probability. We derive necessary and sufficient conditions for model consistency in a low-dimensional setting, and then consider a high-dimensional result.

We consider the square loss and a fixed-design analysis (i.e., $x_1, \dots, x_n$ are fixed); we assume that for all $i \in \{1, \dots, n\}$, $y_i = \mathbf{w}^\top x_i + \varepsilon_i$ where the vector ε is an i.i.d vector with Gaussian distributions with mean zero and variance $\sigma^2 > 0$, and $\mathbf{w} \in \mathbb{R}^p$ is the population sparse vector; we denote by $\mathbf{J}$ the $\mathcal{G}$-adapted hull of its nonzero pattern. Note that for all $P \in \mathcal{P}$, $\mathrm{Hull}(P) = P$; in other words, if our prior is correctly encoded through $\mathcal{G}$, estimating the hull $\mathbf{J}$ is equivalent to estimating the nonzero pattern of the population vector.

5.1 Consistency Condition

We begin with the low-dimensional setting where n is tending to infinity with p *fixed*. In addition, we also assume that the design is *fixed* and that the Gram matrix $Q = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ is invertible with

$$\lim_{n\to\infty} Q = \mathbf{Q} \succ 0.$$

In this setting, the noise is consequently the only source of randomness. We denote by $\mathbf{r_J}$ the vector defined as

$$\forall j \in \mathbf{J},\ \mathbf{r}_j = \mathbf{w}_j \left(\sum_{G \in \mathcal{G}_\mathbf{J}, G \ni j} (d_j^G)^2 \left\| d^G \circ \mathbf{w} \right\|_2^{-1} \right),$$

or equivalently in the more compact form

$$\mathbf{r} = \sum_{G \in \mathcal{G}_\mathbf{J}} \frac{d^G \circ d^G \circ \mathbf{w}}{\left\| d^G \circ \mathbf{w} \right\|_2}.$$

Besides, we recall that we define $\Omega_\mathbf{J}^c(w_{\mathbf{J}^c}) = \sum_{G \in (\mathcal{G}_\mathbf{J})^c} \left\| d_{\mathbf{J}^c}^G \circ w_{\mathbf{J}^c} \right\|_2$ (which is the norm composed of inactive groups) with its dual norm $(\Omega_\mathbf{J}^c)^*$; note the difference with the norm reduced to $\mathbf{J}^c$, defined as $\Omega_{\mathbf{J}^c}(w_{\mathbf{J}^c}) = \sum_{G \in \mathcal{G}} \left\| d_{\mathbf{J}^c}^G \circ w_{\mathbf{J}^c} \right\|_2$.

In the following Theorem, we give the sufficient and necessary conditions under which the hull of the generating pattern is consistently estimated. Those conditions naturally extend the results of Zhao and Yu [2006] and Bach [2008b] for the Lasso and the group Lasso respectively (see proof in Appendix D).

Theorem 5.1 (Consistency condition)**.** *Assume $\mu \to 0$, $\mu\sqrt{n} \to \infty$ in Eq. (2.1). If the hull is consistently estimated, then $(\Omega_\mathbf{J}^c)^*[\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{JJ}}^{-1}\mathbf{r}_\mathbf{J}] \leqslant 1$. Conversely, if $(\Omega_\mathbf{J}^c)^*[\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{JJ}}^{-1}\mathbf{r}_\mathbf{J}] < 1$, then the hull is consistently estimated, i.e.,*

$$\mathbb{P}\left(\{j \in \{1, \dots, p\}, \hat{w}_j \neq 0\} = \mathbf{J}\right) \underset{\substack{n\to+\infty \\ \mu\to 0,\ \mu\sqrt{n}\to\infty}}{\longrightarrow} 1.$$

The two previous propositions bring into play the dual norm $(\Omega_{\mathbf{J}}^c)^*$ that we cannot compute in closed form, but requires to solve an optimization problem as complex as the inital problem (4.1). However, we can prove the following bounds similar to those obtained in Propositions 4.2 and 4.3 for the necessary and sufficient conditions:

$$(\Omega_{\mathbf{J}}^c)^*[\kappa_{\mathbf{J}^c}] \geqslant \max_{K \in \Pi_{\mathcal{P}}(\mathbf{J})} \left\{ \sum_{j \in K \setminus \mathbf{J}} \frac{|\kappa_j|^2}{\left(\sum_{H \in \mathcal{G}_K \setminus \mathcal{G}_{\mathbf{J}}, H \ni j} d_j^H \right)^2} \right\}^{\frac{1}{2}},$$

$$(\Omega_{\mathbf{J}}^c)^*[\kappa_{\mathbf{J}^c}] \leqslant \max_{G \in \mathcal{F}_{\mathbf{J}}} \left\{ \sum_{j \in G} \frac{|\kappa_j|^2}{\left(\sum_{H \in (\mathcal{G}_{\mathbf{J}})^c, H \ni j} d_j^H \right)^2} \right\}^{\frac{1}{2}}.$$

Comparison with the Lasso. Note that for the ℓ_1-norm, our two bounds lead to the usual consistency conditions for the Lasso, i.e., the quantity $\|\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{JJ}}^{-1}\mathrm{sign}(\mathbf{w}_{\mathbf{J}})\|_\infty$ must be less or strictly less than one.

Let us consider that the Lasso is inconsistent because of variables in $\mathbf{J}^c$ whose hull would form a nonzero pattern which is "far" (in the DAG of nonzero patterns) from the hull of relevant variables. Because of the term $\sum_{G \in (\mathcal{G}_{\mathbf{J}})^c, G \ni j} d_j^G$, the condition for our structured sparsity will tend to be more easily satisfied (see examples in Section 6).

5.2 High-Dimensional Analysis

We prove a high-dimensional variable consistency result (see proof in Appendix E) that extends the corresponding result for the Lasso [Zhao and Yu, 2006, Wainwright, 2006], by assuming that the consistency condition in Theorem 5.1 is satisfied.

Theorem 5.2. *Assume that Q has unit diagonal, $\kappa = \lambda_{\min}(Q_{\mathbf{JJ}}) > 0$ and $(\Omega_{\mathbf{J}}^c)^*[Q_{\mathbf{J}^c\mathbf{J}}Q_{\mathbf{JJ}}^{-1}\mathbf{r}] < 1 - \tau$, with $\tau > 0$. If $\tau\mu\sqrt{n} \geq \sigma C_3(\mathcal{G}, \mathbf{J})$, and $|\mathbf{J}|^{1/2}\mu \leq C_4(\mathcal{G}, \mathbf{J})$, then the probability of incorrect hull selection is upper-bounded by:*

$$\exp\left(-\frac{n\mu^2\tau^2 C_1(\mathcal{G}, \mathbf{J})}{2\sigma^2}\right) + 2|\mathbf{J}| \exp\left(-\frac{nC_2(\mathcal{G}, \mathbf{J})}{2|\mathbf{J}|\sigma^2}\right)$$

where $C_1(\mathcal{G}, \mathbf{J})$, $C_2(\mathcal{G}, \mathbf{J})$, $C_3(\mathcal{G}, \mathbf{J})$ and $C_4(\mathcal{G}, \mathbf{J})$ are constants defined in Appendix E, which essentially depend on the groups, the smallest nonzero coefficient of $\mathbf{w}$ and how close the support $\{j \in \mathbf{J} : \mathbf{w}_j \neq 0\}$ of $\mathbf{w}$ is to its hull $\mathbf{J}$, that is the relevance of the prior information encoded by $\mathcal{G}$.

In the Lasso case, we have $C_1(\mathcal{G}, \mathbf{J}) = O(1)$, $C_2(\mathcal{G}, \mathbf{J}) = O(|\mathbf{J}|^{-2})$, $C_3(\mathcal{G}, \mathbf{J}) = O((\log p)^{1/2})$ and $C_4(\mathcal{G}, \mathbf{J}) = O(|\mathbf{J}|^{-1})$, leading to the usual scaling $n \approx \log p$. In our situation, we may have better scalings, but these are problem-dependent: by reducing the number of allowed zero patterns, we would allow more irrelevant variables (a careful analysis of the group-dependent constants would still be needed in all cases).

6 Experiments

In this section, we carry out several experiments to illustrate the behavior of the sparsity-inducing norm Ω.

Active set algorithm. We first focus on the active set algorithm (see Section 4) and compare its time complexity to the SOCP solver when we are looking for a sparse target. More precisely, for a fixed level of sparsity $|\mathbf{J}|$ and a fixed number of observations n, we analyze the complexity w.r.t. the number of variables p. To this end, we consider a linear model $Y = \mathbf{w}^\top X + \varepsilon$, where the true vector $\mathbf{w} \in \mathbb{R}^p$ has only $|\mathbf{J}| = 9$ nonzero components and p varies in $\{100, 225, 400, 900, 1600, 2500, 3600\}$. In addition, both X and ε are centered and normally distributed, the variance of the noise being set to verify $\left\|\mathbf{w}^\top X\right\|_2 = \|\varepsilon\|_2$ (i.e., the SNR equals 1). The $|\mathbf{J}|$ nonzero components of $\mathbf{w}$ are generated once from a centered Gaussian distribution and are kept fixed for the rest of the procedure.

The linear model is mapped onto a square 2-dimensional grid where the nonzero components of $\mathbf{w}$ form a 3×3 square; we therefore consider the rectangular groups for $\mathcal{G}$. Note that here, we are not especially interested in the statistical behavior of the algorithm w.r.t. different datasets, we just care about the computational aspect, so that it does not matter to focus on a single dataset. We thus consider $n = 5500$ (500 of them dedicated to the test) i.i.d observations $\{(x_i, \varepsilon_i) \in \mathbb{R}^p \times \mathbb{R}, i = 1, \ldots, n\}$. For each value of p and for both the active set algorithm and the SOCP solver, we compute an approximate regularization path for 30 values of λ, then compute the average CPU time over these 30 runs and finally take the best prediction error on this path. We assume that we have a rough idea of the level of sparsity of the true vector and we set the stopping criterion $s = 5|\mathbf{J}|$ (see Algorithm 3), which is a rather conservative choice. We show on the Figure 8 that for the same level of performance, we considerably lower the computational cost. In practice (and as the Figure 8 illustrates it), we have noticed that the active set algorithm scales in $O(\sqrt{p})$, which is in agreement with the analysis Section 4 where the constant before the term in $O(\sqrt{p})$ can be much larger than the one before the term in $O(p^{1.5})$.

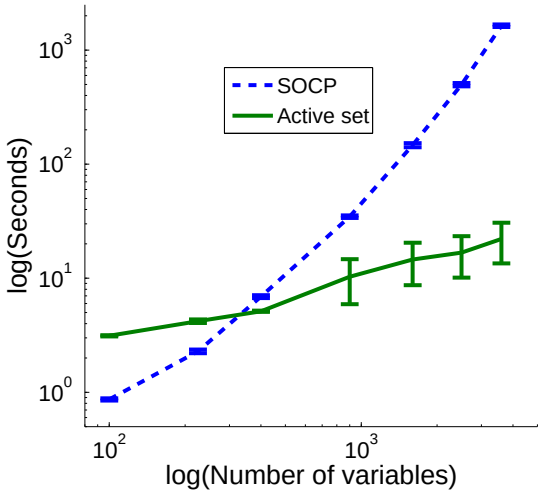

Figure 8: Time complexity comparison: for each value of p, we plot the average CPU time over 30 runs and take the best prediction error. Note we do not display the latter since both algorithms reached exactly the same performance $\{2.89, 3.88, 1.42, 2.31, 3.64, 1.75, 2.71\}$. We can see that the active set algorithm significantly reduces the required computational time.

Not surprisingly, for small values of p, the SOCP solver performs better since it does not have to compute the necessary and sufficient bounds required for the progress of the active set algorithm.

Weight influence and consistency. We illustrate now the influence of the weights $(d^G)_{G\in\mathcal{G}}$. We present in the Figure 9 several settings of $(d^G)_{G\in\mathcal{G}}$ that we will use in the following experiments.

The high-dimensional analysis of the consistency has advocated the choice of weights $(d^G)_{G\in\mathcal{G}}$ that let the sum $\max_{j\in\mathbf{J}}\left\{\sum_{G\in\mathcal{G}_\mathbf{J},G\ni j} d_j^G\right\}$ bounded and independent of p. We will show the importance of this criterion by considering three cases, namely a uniform weighting ($\forall G\in\mathcal{G},\ \forall j\in G,\ d_j^G=1$) and two kinds of weights depending on the size of the groups ($\forall j\in G,\ d_j^G=\frac{1}{|G|}$ and $\forall j\in G,\ d_j^G=\frac{1}{|G|^2}$). Note that only the latter satisfies the previous criterion.

Figure 9: Different choices for the weights $(d^G)_{G\in\mathcal{G}}$: on the left, for all groups, the weights are identical. In the center, we let the groups depend on their sizes, namely for all $G\in\mathcal{G}, \forall j\in G,\ d_j^G=|G|^{-2}$. On the right, we use an exponential weighting that takes into account the geometry of $\mathcal{G}$ (with the factor $\alpha\in(0,1)$; in our experiments, we take $\alpha=0.5$). The latter will be referred to as the exponential weights.

We consider the model $Y=\mathbf{w}^\top X+\varepsilon$, where the vector $\mathbf{w}\in\mathbb{R}^5$ (i.e., $p=5$) has 2 contiguous relevant variables $\{X_1,X_2\}$ and ε is a centered Gaussian noise with the variance fixed to make the SNR equal to 1. We sampled i.i.d. observations (x,y) where $X\sim\mathcal{N}(0,\Sigma)$ is normally distributed and the covariance matrix Σ is designed so that X_5 is highly correlated with $\{X_1,X_2\}$ (the design of Σ follows the first experiment of [Zhao and Yu, 2006]).

In this way, X_5 tends to prevent the consistent estimation of the contiguous pattern $\{X_1,X_2\}$. The norm $\Omega(w)=\sum_{G\in\mathcal{G}}\|d^G\circ w\|_2$ is built to allow only contiguous patterns (as defined in Section 3.5, see Figure 3). For a uniform weighting, the model is inconsistent and all the variables become active together as μ decreases, including X_5, which is an undesirable behavior. By letting the weights depend on group size, we succeed in creating a consistent regime (see Figure 10) and avoiding that all variables come in together.

Convex pattern recovery. We now turn to the ability of the norm Ω to encode a prior that enforces the selection of convex patterns. To this end, we consider a 15×15 grid (i.e., $p=225$) where we are looking for a sparse convex pattern (whose size is $|\mathbf{J}|=40$) in the presence of remote perturbations at the edges of the grid. For the sake of simplicity, we work with the limiting pattern obtained in the asymptotic regime p fixed and $n\to\infty$ (i.e., the pattern of the first order development of the minimizer of (2.1); see the proof of Theorem 5.1 in Appendix D for more

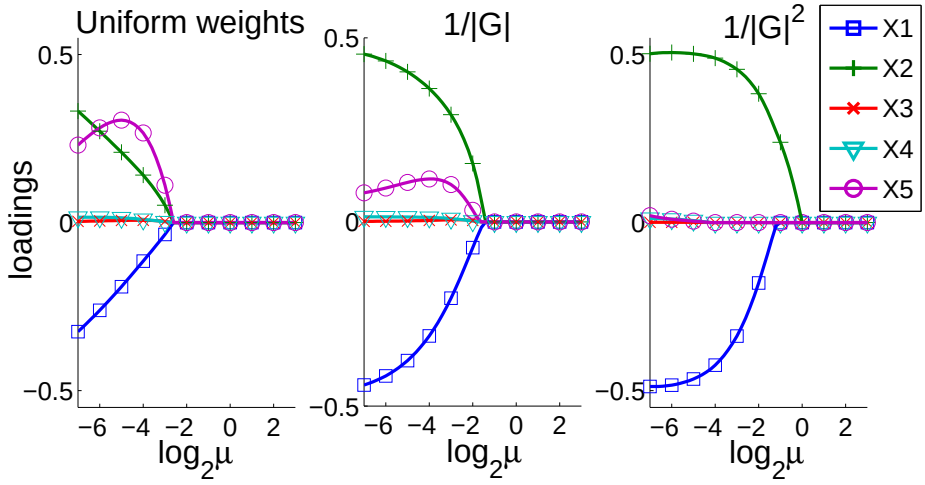

Figure 10: Influence of weights: with uniform weights, all variables enter the active set together, including X_5 that is irrelevant (left); this situation is partially fixed with weights $d_j^G = |G|^{-1}$ (middle) and totally fixed when $d_j^G = |G|^{-2}$ (right): the model consistently selects the correct variables $\{X_1, X_2\}$, and variables enter the model one at a time when μ decreases.

details), so that our experiments depend only on $\mathbf{w}$, $\mathbf{J}$ and the covariance matrices $\mathbf{Q}$ (note that the results obtained in this way could be reproduced for n large enough and the noise variance small enough). The $|\mathbf{J}|$ nonzero components of $\mathbf{w}$ are generated once from a centered Gaussian distribution and are kept fixed for the rest of the procedure. To simulate the remote perturbations, we procceed as follows:

a) We generate a random covariance matrix $\mathbf{Q} \in \mathbb{R}^{p\times p}$ according to a Wishart distribution with p degrees of freedom. Its diagonal is then re-normalized to one. For such matrices, the Lasso consistency condition is known not to hold with high probability [Zhao and Yu, 2006].

b) We compute the vector $z_{\mathbf{J}^c} = \mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{J}\mathbf{J}}^{-1}\text{sign}(\mathbf{w}_{\mathbf{J}}) \in \mathbb{R}^{|\mathbf{J}^c|}$. We randomly permute the variables of $\mathbf{J}^c$, with the constraint that the $4(p^{1/2}-1)$ (i.e., the perimeter of the grid) largest components of $z_{\mathbf{J}^c}$ (in absolute value) must be located at the boundaries of the grid.

We present in the Figure 11 the probabilities of variable selection on the grid for different models, based on an average over 100 covariance matrices generated according to the scheme described above. We compare the average nonzero pattern recovered by the Lasso with our model (referred to as *subsetlasso*) for different choices of the weights $(d^G)_{G\in\mathcal{G}}$.

Although the correct pattern belongs to the active set obtained with the Lasso, the remote variables are also selected, which violates the convex prior. This point may be pivotal in many applications like neuroscience where convex patterns are needed for the sake of interpretation.

Similarly to the previous experiment, the model with the uniform weights fails to recover the correct nonzero pattern, which results in the selection of all the variables. When we consider the weights depending on the size of the groups, we are able to better discriminate the hull of the generating pattern. In addition, we can see the effect of adding the $\pm\frac{\pi}{4}$ groups to the rectangular groups, namely the rectangular hull becomes the $\mathcal{G}$-adapted hull (to the new set of groups $\mathcal{G}$) that approximates the generating pattern more precisely.

Finally, by considering the procedure (introduced at the end of Section 4 and referred to as *intersected subsetlasso*, or simply *I-subsetlasso*) that consists in intersecting the nonzero patterns

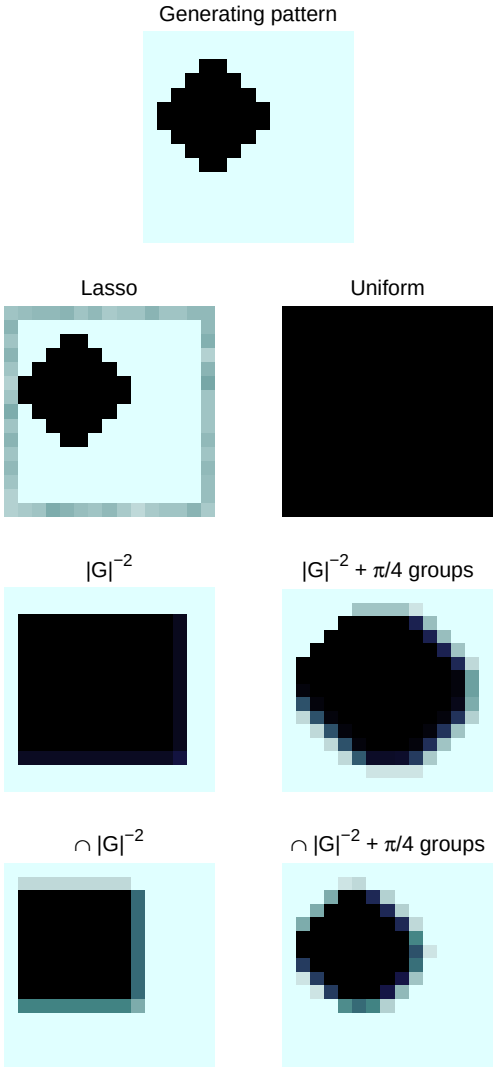

Figure 11: Convex pattern estimation: we plot the probabilities of variable selection for different models, by considering their limiting patterns. The black color indicates a probabilty of one, while the light blue color corresponds to probability of zero. In addition, those probabilities are obtained from an average over 100 covariance matrices generated according to the scheme described before.

obtained for different models, we improve significantly the estimation of the hull by being more discriminative.

Prediction error. In this last experiment, we show that the prior we put through the norm Ω is also a source of improvements for the predictive power. Note that we are currently working on theoretical analyzes to back these experimental results. We are getting to grips with the simplest setting where we can express a prior through Ω, namely the selection of a contiguous pattern on a sequence. More specifically, we consider a sequence of size $p = 100$ with a contiguous generating pattern $\mathbf{J}$ corresponding to the variables 20 to 35, i.e., with $|\mathbf{J}| = 16$. As previously, we are interested in the underlying linear model $Y = \mathbf{w}^\top X + \varepsilon$, where the true vector $\mathbf{w}$ belongs to $\mathbb{R}^{100}$. Again, both X and ε are centered and normally distributed, the variance of the noise being set to

verify $\left\|\mathbf{w}^\top X\right\|_2 = \|\varepsilon\|_2$ (i.e., the SNR equals 1). The $|\mathbf{J}|$ nonzero components of $\mathbf{w}$ are generated once from a centered Gaussian distribution and are kept fixed for the rest of the procedure.

We will compare the Lasso with the subsetlasso and the I-subsetlasso (set for different types of weights $(d^G)_{G\in\mathcal{G}}$). For each of those models and for a given pair $\{X, \varepsilon\}$, we compute an approximate regularization path and pick up the best prediction error on this path. The prediction error is understood here as being the prediction error of the OLS performed on the nonzero pattern obtained by the model considered (note that it is actually a fair way to compare the I-subsetlasso with the other models). For the corresponding value of the regularization parameter (i.e., where the minimum prediction error is reached), we compute the hull estimation error, defined as $\left\|\delta_{\mathbf{J}} - \delta_{\hat{J}}\right\|_2$, where $\delta_I \in \mathbb{R}^p$ represents the indicator vector of the hull of I.

For several values of $n \in \{80, 150, 250, 500, 1000, 1500\}$ (while keeping $p = 100$ fixed), we repeat this procedure 50 times and we present in the Figure 12 the average error for both the prediction and the hull estimation.

Using the norm Ω results in an improvement, both in terms of prediction and hull estimation error. The improvement is sharper for the latter, since it is exactly the task why the norm Ω has been designed for. The experiment underlines again the importance of the choice of the weights $(d^G)_{G\in\mathcal{G}}$. The uniform weights perform poorly compared to the exponential and the size-dependent weights. Moreover, the results show the superiority of the I-subsetlasso procedure.

7 Conclusion

We have shown how to incorporate prior knowledge on the form of nonzero patterns for linear supervised learning. Our solution relies on a regularizing term linearly combining ℓ_2-norms of possibly overlapping groups of variables. We have studied the design of these groups, efficient algorithms and theoretical guarantees of the structured sparsity-inducing method. Natural extensions to this work are to consider bootstrapping since this may improve theoretical guarantees and lead to better variable selection [Bach, 2008a], or to combine this work with the recent approach of Jacob et al. [2009] to handle more general families of (non)zero patterns. Our regularization scheme could also be used for multi-task learning [Argyriou et al., 2008] or multiple kernel learning [Micchelli and Pontil, 2005] when prior knowledge on the structure of the sparse representation is available. Finally, although we have mostly explored in this paper the algorithmic and theoretical issues related to these norms, this type of prior knowledge is of clear interest for the spatially structured data typical in bioinformatics, computer vision and neuroscience applications.

References

R. J. Adler. *An introduction to continuity, extrema, and related topics for general Gaussian processes*, volume 12 of *Lect. Notes*. IMS, Hayward, 1990.

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

F. Bach. Bolasso: model consistent Lasso estimation through the bootstrap. In *Proc. ICML*, 2008a.

F. Bach. Consistency of the group Lasso and multiple kernel learning. *JMLR*, 8:1179–1225, 2008b.

F. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Adv. NIPS*, 2008c.

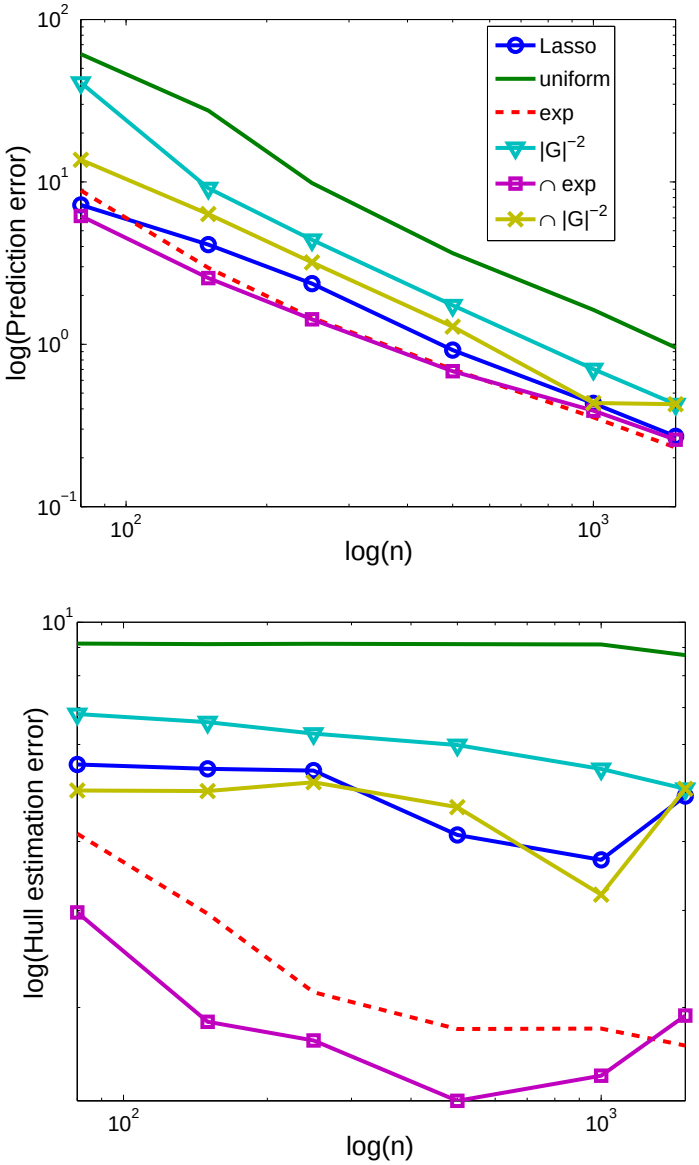

Figure 12: Prediction and hull estimation error versus n: on the top figure, we plot the evolution of the prediction error versus the number of observations n. On the bottom figure, we plot in the same way the hull estimation error. The displayed curves are the average over 50 random pairs $\{X, \varepsilon\}$. The prior we put through the norm Ω improves the predictive power and the ability to estimate the correct contiguous generating pattern.

J. M. Borwein and A. S. Lewis. *Convex Analysis And Nonlinear Optimization: Theory And Examples*. Springer, 2006.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Camb. Univ. P., 2003.

P. J. Cameron. *Combinatorics: Topics, Techniques, Algorithms*. Camb. Univ. P., 1994.

J. P. Doignon and J. C. Falmagne. *Knowledge Spaces*. Springer-Verlag, 1998.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Stat.*, 32:407, 2004.

W. Fu and K. Knight. Asymptotics for lasso-type estimators. *Annals of Statistics*, 28(5):1356–1378, 2000.

M. Grant and S. Boyd. Cvx: Matlab software for disciplined convex programming, 2008.

G. H. Hardy, J. E. Littlewood, and G. Polya. *Inequalities*. Cambridge University Press, 1988.

L. Jacob, G. Obozinski, and J. P. Vert. Group Lasso with Overlap and Graph Lasso. In *Proc. ICML*, 2009. To appear.

H. Lee, A. Battle, R. Raina, and A. Ng. Efficient sparse coding algorithms. In *Adv. NIPS*, 2007.

P. Massart. *Concentration Inequalities and Model Selection: Ecole d'été de Probabilités de Saint-Flour 23*. Springer, 2003.

C. A. Micchelli and M. Pontil. Learning the kernel function via regularization. *JMLR*, 6:1099–1125, 2005.

Y. Nardi and A. Rinaldo. On the asymptotic properties of the group Lasso estimator for linear models. *Electron. J. Statist.*, 2:605–633, 2008.

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

V. Roth and B. Fischer. The group-Lasso for generalized linear models: Uniqueness of solutions and efficient algorithms. In *Proc. ICML*, 2008.

M. Sion, Office of Scientific Research, United States, and A. Force. *General Minimax Theorems*. United States Air Force, Office of Scientific Research, 1957.

P. Soille. *Morphological Image Analysis: Principles and Applications*. Springer, 2003.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. Roy. Stat. Soc. B*, 58:267–288, 1996.

M. J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity. Technical Report 709, Dpt. of Statistics, UC Berkeley, 2006.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. Roy. Stat. Soc. B*, 68(1):49–67, 2006.

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

P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Ann. Stat.*, 2009. To appear.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67(2):301–320, 2005.

A Proof of Theorem 3.1

Proof. We recall that $L(w) = \frac{1}{n}\sum_{i=1}^n \ell\left(y_i, w^\top x_i\right)$. For the square loss, the Hessian of L is Q. Since Q is positive semidefinite, L is convex. In addition, $w \mapsto \Omega(w)$ is convex and goes to infinite when $\|w\|_2$ goes to infinite, so that we can restrict the minimization problem to a compact set of $\mathbb{R}^p$. By Weierstrass' theorem, (3.2) admits a global solution, that we will write w^Y to stress on its dependence on the observed output vector Y. Note that, at this stage of the proof, we do not have the uniqueness of the solution.

Uniqueness: Let us suppose that (3.2) admits more than one solution and let denote by Θ^Y this convex set of solutions. We consider $w^{Y,1} = \text{argmax}_{w\in\Theta^Y} |I^Y(w)|$, the solution having the largest nonzero pattern. We need to discuss two possible cases

a) $I^Y(w^{Y,1}) = \bigcup_{w\in\Theta^Y} I^Y(w)$

b) $I^Y(w^{Y,1}) \neq \bigcup_{w\in\Theta^Y} I^Y(w)$.

If we are in the situation a), we can directly use the assumption on the invertibility of $Q_{\mathrm{Hull}(I^Y(w^{Y,1}))\mathrm{Hull}(I^Y(w^{Y,1}))}$ with any other solutions $w^{Y,2}$ in Θ^Y. The strong convexity of the problem reduced to $\mathrm{Hull}(I^Y(w^{Y,1}))$ leads to the desired conclusion.

The previous argument cannot be reused immedialty in the scenario b). We consider instead $w^{Y,2} \in \Theta^Y$ with $|I^Y(w^{Y,1}) \cup I^Y(w^{Y,2})| > |I^Y(w^{Y,1})|$. By convexity of Θ^Y, we can consider in turn the solution $w^{Y,3} = \beta\, w^{Y,2} + (1-\beta)\, w^{Y,1}$. For $\beta > 0$ sufficiently small, we have

$$I^Y(w^{Y,3}) = I^Y(w^{Y,1}) \cup I^Y(w^{Y,2}),$$

which contradicts the definition of $w^{Y,1}$. Thus, the scenario b) is impossible and we have the uniqueness of the solution.

Stability of the zero patterns: We now prove by contradiction that the zero pattern $Z(w^Y)$ of w^Y almost surely satisfies $Z(w^Y) \in \mathcal{Z}$. Let us assume that

$$\mathbb{P}(Z(w^Y) \notin \mathcal{Z}) = \sum_{K\notin\mathcal{Z}} \mathbb{P}(K = Z(w^Y)) > 0,$$

so that there exists $I \subset \{1,\dots,p\}$ such that $I^c \notin \mathcal{Z}$ with $\mathbb{P}(Z(w^Y) = I^c) > 0$. So, for a large enough compact $\mathcal{S}$, we have $\mathbb{P}(A) > 0$ with $A = \{Y \in \mathcal{S}; Z(w^Y) = I^c\}$.

We now show that this cannot be true by studying the behavior of w^Y around points in A. Let $\mathcal{G}_I = \{G \in \mathcal{G} : G \cap I \neq \emptyset\}$ be the set of active groups and we refer to $\mathrm{Hull}(I)$ as J. We recall that the restriction L_J of L is given by $L_J(w) = L(\tilde{w})$ where $\tilde{w}_J = w$ and $\tilde{w}_{J^c} = 0$ for all $w \in \mathbb{R}^{|J|}$.

The optimality of w^Y when $Z(w^Y) = I^c \supseteq J^c$ implies

$$\nabla L_J(w_J^Y) + r_J(w_J^Y) = 0,$$

where we define the vector $r_J(w_J^Y) \in \mathbb{R}^{|J|}$ as

$$r_j(w_J^Y) = w_j^Y \left(\sum_{G\in\mathcal{G}_I, G\ni j} (d_j^G)^2 \left\| d^G \circ w^Y \right\|_2^{-1} \right),\ \forall j \in J.$$

Let $v^Y \in \mathbb{R}^{|J|}$ be the solution of $f(v, Y) = 0$, with

$$f(v, Y) = \nabla L_J(v) + r_J(v).$$

Let $\tilde{y} \in A$ and $f_1, \dots, f_{|J|}$ be the components of f.

On a small enough ball around $(w_J^{\tilde{y}}, \tilde{y})$, f is continuously differentiable since none of the norms vanishes at $w_J^{\tilde{y}}$. Let H_{JJ} be the matrix whose j-th row is $(\nabla_v f_j)^\top$. The matrix H_{JJ} is actually the sum of

a) the hessian of L_J, i.e., Q_{JJ} that we assumed positive definite, and

b) the hessian of the norm Ω around $(w_J^{\tilde{y}}, \tilde{y})$ that is positive semidefinite on this small ball according to the hessian characterization of convexity [Borwein and Lewis, 2006, Theorem 3.1.11].

Consequently, H_{JJ} is invertible. We can now apply the implicit function theorem to obtain that for Y in a neighborhood of $\tilde{y}$,

$$v^Y = \psi(Y),$$

with $\psi = (\psi_1, \dots, \psi_{|J|})^\top$ a continuously differentiable function satisfying the matricial relation

$$(\dots, \nabla\psi_j, \dots) H_{JJ} + (\dots, \nabla_y f_j, \dots) = 0.$$

Since we supposed that $I^c \notin \mathcal{Z}$, we can consider a fixed $\alpha \in I^c \cap J$.

Let C_α denote the α-th column of H_{JJ}^{-1} and $X^J \in \mathbb{R}^{n \times |J|}$ be the matrix whose $(i,j)-$element is the j-th component of x_i. Since $n(\dots, \nabla_y f_j, \dots) = -X^J$, we have

$$n\nabla\psi_\alpha = X^J C_\alpha.$$

Now, since X^J has full rank and $C_\alpha \neq 0$, we have $\nabla\psi_\alpha \neq 0$.

Without loss of generality, we may assume that $\partial\psi_\alpha/\partial y_1 \neq 0$ on a neighborhood of $\tilde{y}$. We can apply again the implicit function theorem to show that on a neighborhood of $\tilde{y}$ the solution to $\psi_\alpha(Y) = 0$ can be written $y_1 = \varphi(y_2, \dots, y_n)$ with φ a continuously differentiable function.

By Fubini's theorem and by using the fact that the Lebesgue measure of a singleton in $\mathbb{R}^n$ equals zero, we have shown that there exists $\delta_{\tilde{y}} > 0$ such that $\mathbb{P}(Y \in \mathcal{B}(\tilde{y}, \delta_{\tilde{y}}) \cap A) = 0$, where $\mathcal{B}(u, \rho)$ is the open ball in $\mathbb{R}^n$ centered at u and of radius ρ.

Now we have $\mathbb{P}(A) = \sup\{\mathbb{P}(F); F \text{ closed }, F \subset A\}$. For $F \subset A$ closed, we have F closed and in the compact $\mathcal{S}$, hence F is compact. Besides it can be written as $F = \cup_{\tilde{y} \in F}\{\mathcal{B}(\tilde{y}, \delta_{\tilde{y}}) \cap F\}$. By compacity of F, there exists a sequence $(u_m)_{m\in\mathbb{N}}$ of elements in F such that $F = \cup_{m\in\mathbb{N}}\{\mathcal{B}(u_m, \delta_{u_m}) \cap F\}$. So we have $\mathbb{P}(F) \leq \sum_{m\in\mathbb{N}} \mathbb{P}\{\mathcal{B}(u_m, \delta_{u_m}) \cap F\} = 0$, hence $\mathbb{P}(A) = 0$. This concludes the proof by contradiction. □

B Proof of the minimality of the Backward procedure (see Algorithm 2)

Proof. There are 2 points to show:

- $\mathcal{G}$ spans $\mathcal{Z}$.
- $\mathcal{G}$ is minimal.

The first point can be shown by a proof by recurrence on the depth of the DAG. At step t, the base $\mathcal{G}^{(t)}$ verifies $\{\bigcup_{G\in\mathcal{G}'} G,\ \forall\mathcal{G}' \subseteq \mathcal{G}^{(t)}\} = \{G \in \mathcal{Z}, |G| \leqslant t\}$ because an element $G \in \mathcal{Z}$ is either the union of itself or the union of elements strictly smaller. The initialization $t = \min_{G\in\mathcal{Z}} |G|$ is easily verified, the leafs of the DAG being necessarily in $\mathcal{G}$.

As for the second point, we proceed by contradiction. If there exists another base $\mathcal{G}^*$ that spans $\mathcal{Z}$ such that $\mathcal{G}^* \subset \mathcal{G}$, then

$$\exists\, e \in \mathcal{G},\ e \notin \mathcal{G}^*.$$

By definition of the set $\mathcal{Z}$, there exists in turn $\mathcal{G}' \subseteq \mathcal{G}^*,\ \mathcal{G}' \neq \{e\}$ (otherwise, e would belong to $\mathcal{G}^*$), verifying $e = \bigcup_{G \in \mathcal{G}'} G$, which is impossible by construction of $\mathcal{G}$ whose members cannot be the union of elements of $\mathcal{Z}$. □

C Proof of Propositions 4.2 and 4.3

In order to check that the reduced solution w_J is optimal for the full problem in Eq. (4.1), we complete with zeros on J^c to define w, compute $\kappa = -\nabla L(w)$, which is such that $\kappa_J = -\nabla L_J(w_J)$, and get a duality gap for the full problem equal to

$$\lambda\Omega^*(\kappa) - w^\top \kappa = \lambda\Omega^*(\kappa) - w_J^\top \kappa_J = \lambda[\Omega^*(\kappa) - \Omega_J^*(\kappa_J)].$$

By designing upper and lower-bounds for $\Omega^*(\kappa)$, we get sufficient and necessary conditions.

C.1 Proof of Proposition 4.2

Proof. For each $K \in \Pi_{\mathcal{P}}(J)$, we simply need to lower bound $\Omega^*(\kappa) = \max_{\Omega(v) \leqslant 1} v^\top \kappa$. Starting from the Lemma F.7, we have

$$\Omega^*(\kappa) = \min_{\gamma \in \Gamma} \max_{G \in \mathcal{G}} \left\{ \sum_{j \in G} \gamma_{Gj}^2 (d_j^G)^{-2} |\kappa_j|^2 \right\}^{\frac{1}{2}} \tag{C.1}$$

$$= \max_{G \in \mathcal{G}} \left\{ \sum_{j \in G} (\gamma_{Gj}^*)^2 (d_j^G)^{-2} |\kappa_j|^2 \right\}^{\frac{1}{2}} \tag{C.2}$$

where γ^* is a solution of the previous minization. According to Appendix G, all the γ^* corresponding to active groups ($w_G^* \neq 0$) and inactive variables ($w_j^* = 0$) are equal to zero.

We derive a lower-bound on $\Omega^*(\kappa)$ by *restricting* the maximum over the groups $G \in \mathcal{G}$ to those in $\mathcal{G}_K \backslash \mathcal{G}_J$. We also restrict the sum $\sum_{j \in G}$ to the sum $\sum_{j \in G \cap (K \backslash J)}$. Note that for all groups G in $\mathcal{G}_K \backslash \mathcal{G}_J$, we have $(K \backslash J) \subseteq G$ (see Lemma F.9), so that $\sum_{j \in G \cap (K \backslash J)} = \sum_{j \in K \backslash J}$.

To sum up, for all $\gamma \in \Gamma$ with the added constraint that $\gamma_{Gj} = 0$ for $G \in \mathcal{G}_J$ and $j \in J^c$, we have

$$\max_{G \in \mathcal{G}} \left\{ \sum_{j \in G} \gamma_{Gj}^2 (d_j^G)^{-2} |\kappa_j|^2 \right\}^{\frac{1}{2}} \geq \max_{G \in \mathcal{G}_K \backslash \mathcal{G}_J} \left\{ \sum_{j \in K \backslash J} \gamma_{Gj}^2 (d_j^G)^{-2} |\kappa_j|^2 \right\}^{\frac{1}{2}}.$$

Moreover, for all $j \in K \backslash J$,

$$\sum_{\substack{G \in \mathcal{G} \\ G \ni j}} \gamma_{Gj} = \sum_{\substack{G \in \mathcal{G}_{K \backslash J} \\ G \ni j}} \gamma_{Gj} = \sum_{\substack{G \in \mathcal{G}_K \backslash \mathcal{G}_J \\ G \ni j}} \gamma_{Gj} = 1,$$

where we use the added constraint on the active groups and inactive variables. Hence, the quantity

$$\max_{G \in \mathcal{G}_K \backslash \mathcal{G}_J} \left\{ \sum_{j \in K \backslash J} \gamma_{Gj}^2 (d_j^G)^{-2} |\kappa_j|^2 \right\}^{\frac{1}{2}}$$

is minimized for

$$\hat{\gamma}_{Gj} = \frac{d_j^G}{\sum_{H \in \mathcal{G}_K \setminus \mathcal{G}_J,\, H \ni j} d_j^H}.$$

By plugging back $\hat{\gamma}$ into (C.1), we obtain

$$\Omega^*(\nabla L(w)) \geq \left\{ \sum_{j \in K \setminus J} \frac{|\nabla L(w)_j|^2}{\left(\sum_{H \in \mathcal{G}_K \setminus \mathcal{G}_J, H \ni j} d_j^H \right)^2} \right\}^{\frac{1}{2}}.$$

The result follows then from optimality condition for w_J, i.e., $-w^\top \nabla L(w) = -w_J^\top \nabla L(w)_J = \lambda \Omega_J^*(\kappa_J)$. Thus, the duality gap is greater than

$$\lambda \left\{ \sum_{j \in K \setminus J} \frac{|\nabla L(w)_j|^2}{\left(\sum_{H \in \mathcal{G}_K \setminus \mathcal{G}_J, H \ni j} d_j^H \right)^2} \right\}^{\frac{1}{2}} + w^\top \nabla L(w).$$

If $\hat{w}$ is optimal, this quantity must be nonpositive, hence the condition (N). □

C.2 Proof of Proposition 4.3

Proof. We reuse techniques from [Bach, 2008c] to get an upper-bound on the dual norm $\Omega^*(\kappa)$. We have that $(\Omega_J^c)^*(\kappa_{J^c})$ is equal to (see Lemma F.7)

$$\min_{\gamma \in \Gamma_{J^c}} \max_{G \in (\mathcal{G}_J)^c} \left\{ \sum_{j \in G \cap J^c} \gamma_{Gj}^2 (d_j^G)^{-2} |\kappa_j|^2 \right\}^{\frac{1}{2}}.$$

As opposed to the necessary condition, we cannot resort to the optimality condition to derive good candidates for the dual variables γ. In addition, we wish to find an upper-bound on $\Omega_J^c(w_{J^c})$ that takes into account the effect of all inactive variables, while in the necessary condition, we only scan the direct parents of the current nonzero pattern J. To this end, we will consider the minimizer when we replace $\sum_{j \in G}$ by $\max_{j \in G}$ (see Lemma F.8), i.e.,

$$\gamma_{Gj} = \frac{d_j^G}{\sum_{H \in (\mathcal{G}_J)^c, H \ni j} d_j^H},$$

from which we get

$$(\Omega_J^c)^*(\kappa_{J^c}) \leqslant \max_{G \in (\mathcal{G}_J)^c} \left\{ \sum_{j \in G} (d_j^G)^{-2} \frac{|\kappa_i|^2 (d_j^G)^2}{\left(\sum_{H \in (\mathcal{G}_J)^c, H \ni j} d_j^H \right)^2} \right\}^{\frac{1}{2}}.$$

Among all groups $G \in (\mathcal{G}_J)^c$, the ones with the maximum values are the ones in the fringe groups $\mathcal{F}_J = \{G \in (\mathcal{G}_J)^c \,;\, \nexists G' \in (\mathcal{G}_J)^c, G \subseteq G'\}$. We can now upper-bound:

$$\begin{aligned}
\Omega^*(\kappa) &= \max_{\sum_{G \in \mathcal{G}_J} \|d^G \circ v\|_2 + \sum_{G \in (\mathcal{G}_J)^c} \|d^G \circ v\|_2 \leqslant 1} v^\top \kappa \\
&\leqslant \max_{\sum_{G \in \mathcal{G}_J} \left\|d_J^G \circ v_J\right\|_2 + \sum_{G \in (\mathcal{G}_J)^c} \|d^G \circ v\|_2 \leqslant 1} v^\top \kappa \\
&= \max_{\Omega_J(v_J) + (\Omega_J^c)(v_{J^c}) \leqslant 1} v^\top \kappa \\
&= \max \left\{ \Omega_J^*(\kappa_J), (\Omega_J^c)^*[\kappa_{J^c}] \right\}.
\end{aligned}$$

where in the last line, we use the Lemma F.11. Thus the duality gap is less than

$$\lambda[\Omega^*(\kappa) - \Omega_J^*(\kappa_J)] \quad \leqslant \quad \lambda \max\{0, (\Omega_J^c)^*[\kappa_{J^c}] - \Omega_J^*(\kappa_J)\}.$$

Using $-w^\top \nabla L(w) = -w_J^\top \nabla L(w)_J = \lambda \Omega_J^*(\kappa_J)$, we get the desired result. □

D Proof of Theorem 5.1

Proof. Necessary condition: We mostly follow the proof of Bach [2008b], Zou [2006]. Let $\hat{w} \in \mathbb{R}^p$ be the unique solution of

$$\min_{w \in \mathbb{R}^p} L(w) + \mu\, \Omega(w) = \min_{w \in \mathbb{R}^p} F(w).$$

The quantity $\hat{\Delta} = (\hat{w} - \mathbf{w})/\mu$ is the minimizer of $\tilde{F}$ defined as

$$\tilde{F}(\Delta) = \frac{1}{2}\Delta^\top Q \Delta - \mu^{-1} q^\top \Delta + \mu^{-1} \left[\Omega(\mathbf{w} + \mu\Delta) - \Omega(\mathbf{w})\right],$$

where $q = \frac{1}{n}\sum_{i=1}^n \varepsilon_i x_i$. The random variable $\mu^{-1} q^\top \Delta$ is a centered Gaussian with variance $\sqrt{\Delta^\top Q \Delta}/(n\mu^2)$. Since $Q \to \mathbf{Q}$, we obtain that for all $\Delta \in \mathbb{R}^p$,

$$\mu^{-1} q^\top \Delta = o_p(1).$$

Since $\mu \to 0$, we also have by taking the directional derivative of Ω at $\mathbf{w}$ in the direction of Δ

$$\mu^{-1} \left[\Omega(\mathbf{w} + \mu\Delta) - \Omega(\mathbf{w})\right] = \mathbf{r}_\mathbf{J}^\top \Delta_\mathbf{J} + \Omega_\mathbf{J}^c(\Delta_{\mathbf{J}^c}) + o(1),$$

so that for all $\Delta \in \mathbb{R}^p$

$$\tilde{F}(\Delta) = \Delta^\top \mathbf{Q} \Delta + \mathbf{r}_\mathbf{J}^\top \Delta_\mathbf{J} + \Omega_\mathbf{J}^c(\Delta_{\mathbf{J}^c}) + o_p(1) = \tilde{F}_{\text{lin}}(\Delta) + o_p(1).$$

The limiting function $\tilde{F}_{\text{lin}}$ being stricly convex (because $\mathbf{Q} \succ 0$) and $\tilde{F}$ being convex, we have that the minimizer $\hat{\Delta}$ of $\tilde{F}$ tends in probability to the unique minimizer of $\tilde{F}_{\text{lin}}$ [Fu and Knight, 2000] referred to as Δ^*.

By assumption, with probability tending to one, we have $\mathbf{J} = \{j \in \{1, \dots, p\}, \hat{w}_j \neq 0\}$, hence for any $j \in \mathbf{J}^c$ $\mu\hat{\Delta}_j = (\mathbf{w} + \mu\hat{\Delta})_j = 0$. This implies that the nonrandom vector Δ^* verifies $\Delta^*_{\mathbf{J}^c} = 0$.

As a consequence, $\Delta^*_\mathbf{J}$ minimizes $\Delta_\mathbf{J}^\top \mathbf{Q}_{\mathbf{JJ}} \Delta_\mathbf{J} + \mathbf{r}_\mathbf{J}^\top \Delta_\mathbf{J}$, hence $\mathbf{r}_\mathbf{J} = -\mathbf{Q}_{\mathbf{JJ}} \Delta^*_\mathbf{J}$. Besides, since Δ^* is the minimizer of $\tilde{F}_{\text{lin}}$, by taking the directional derivatives as in the proof of Lemma F.10, we have

$$(\Omega_\mathbf{J}^c)^*[\mathbf{Q}_{\mathbf{J}^c\mathbf{J}} \Delta^*_\mathbf{J}] \leq 1.$$

This gives the necessary condition.

Sufficient condition: We turn to the sufficient condition. We first consider the problem reduced to the hull $\mathbf{J}$,

$$\min_{w \in \mathbb{R}^{|\mathbf{J}|}} L_\mathbf{J}(w_\mathbf{J}) + \mu \Omega_\mathbf{J}(w_\mathbf{J}).$$

that is strongly convex since $Q_{\mathbf{JJ}}$ is positive definite and thus admits a unique solution $\hat{w}_\mathbf{J}$. With similar arguments as the ones used in the necessary condition, we can show that $\hat{w}_\mathbf{J}$ tends in probability to the true vector $\mathbf{w}_\mathbf{J}$. We now consider the vector $\hat{w} \in \mathbb{R}^p$ which is the vector $\hat{w}_\mathbf{J}$ padded

with zeros on $\mathbf{J}^c$. Since, from Theorem 3.1, we almost surely have $\mathrm{Hull}(\{j \in \{1,\dots,p\}, \hat{w}_j \neq 0\}) = \{j \in \{1,\dots,p\}, \hat{w}_j \neq 0\}$, we have already that the vector $\hat{w}$ consistently estimates the hull of $\mathbf{w}$ and we have that $\hat{w}$ tends in probability to $\mathbf{w}$. From now on, we thus consider that $\hat{w}$ sufficiently close to $\mathbf{w}$, so that for any $G \in \mathcal{G}_\mathbf{J}$, $\|d^G \circ \hat{w}\|_2 \neq 0$. We may thus introduce

$$\hat{r} = \sum_{G \in \mathcal{G}_\mathbf{J}} \frac{d^G \circ d^G \circ \hat{w}}{\|d^G \circ \hat{w}\|_2}.$$

It remains to show that $\hat{w}$ is indeed optimal for the full problem (that admits a unique solution due to the positiveness of Q). By construction, the optimality condition (see Lemma F.10) relative to the active variables $\mathbf{J}$ is already verified. More precisely, we have

$$\nabla L(\hat{w})_\mathbf{J} + \mu\, \hat{r}_\mathbf{J} = Q_{\mathbf{J}\mathbf{J}}(\hat{w}_\mathbf{J} - \mathbf{w}_\mathbf{J}) - q_\mathbf{J} + \mu\, \hat{r}_\mathbf{J} = 0.$$

Moreover, for all $u_{\mathbf{J}^c} \in \mathbb{R}^{|\mathbf{J}^c|}$, by using the previous expression and the invertibily of Q, we have

$$u_{\mathbf{J}^c}^\top \nabla L(\hat{w})_{\mathbf{J}^c} = u_{\mathbf{J}^c}^\top \left\{ -\mu\, Q_{\mathbf{J}^c\mathbf{J}} Q_{\mathbf{J}\mathbf{J}}^{-1} \hat{r}_\mathbf{J} + Q_{\mathbf{J}^c\mathbf{J}} Q_{\mathbf{J}\mathbf{J}}^{-1} q_\mathbf{J} - q_{\mathbf{J}^c} \right\}$$

The terms related to the noise vanish, having actually $q = o_p(1)$. Since $Q \to \mathbf{Q}$ and $\hat{r}_\mathbf{J} \to \mathbf{r}_\mathbf{J}$, we get for all $u_{\mathbf{J}^c} \in \mathbb{R}^{|\mathbf{J}^c|}$

$$u_{\mathbf{J}^c}^\top \nabla L(\hat{w})_{\mathbf{J}^c} = -\mu\, u_{\mathbf{J}^c}^\top \left\{ \mathbf{Q}_{\mathbf{J}^c\mathbf{J}} \mathbf{Q}_{\mathbf{J}\mathbf{J}}^{-1} \mathbf{r}_\mathbf{J} \right\} + o_p(\mu).$$

Since we assume $(\Omega_\mathbf{J}^c)^*[\mathbf{Q}_{\mathbf{J}^c\mathbf{J}} \mathbf{Q}_{\mathbf{J}\mathbf{J}}^{-1} \mathbf{r}_\mathbf{J}] < 1$, we obtain

$$-u_{\mathbf{J}^c}^\top \nabla L(\hat{w})_{\mathbf{J}^c} < \mu(\Omega_\mathbf{J}^c)[u_{\mathbf{J}^c}] + o_p(\mu),$$

which proves the optimality condition of Lemma F.10 relative to the inactive variables: $\hat{w}$ is therefore optimal for the full problem. □

E Proof of Theorem 5.2

Throughout the following proof, we will have to find lower and upper-bounds for the dual norms $(\Omega_\mathbf{J}^c)^*$ and $(\Omega_\mathbf{J})^*$. Since our analysis takes place in a finite-dimensional space, all the norms defined on this space are equivalent. Therefore, for any norm $\|.\|$ on $\mathbb{R}^{|\mathbf{J}|}$ (e.g., $\|.\|_1$, $\|.\|_2$ or $\|.\|_\infty$), we introduce some equivalence parameters $\mathrm{c}, \mathrm{C} > 0$ such that

$$\forall u \in \mathbb{R}^{|\mathbf{J}|},\ \mathrm{c}_{(\Omega_\mathbf{J}, \|.\|)} \|u\| \leq (\Omega_\mathbf{J})[u] \leq \mathrm{C}_{(\Omega_\mathbf{J}, \|.\|)} \|u\|.$$

We similarly define such $\mathrm{c}, \mathrm{C} > 0$ for the norm $(\Omega_\mathbf{J}^c)$ on $\mathbb{R}^{|\mathbf{J}^c|}$. In addition, we immediatly get by order-reversing

$$\forall u \in \mathbb{R}^{|\mathbf{J}|},\ \mathrm{C}^{-1}_{(\Omega_\mathbf{J}, \|.\|)} \|u\|^* \leq (\Omega_\mathbf{J})^*[u] \leq \mathrm{c}^{-1}_{(\Omega_\mathbf{J}, \|.\|)} \|u\|^*.$$

Note that those parameters *hide* a dependance on the dimension of the space, $|\mathbf{J}|$ or $|\mathbf{J}^c|$, and the weights $(d^G)_{G \in \mathcal{G}}$ of the norms. Since we will intensively use such $\mathrm{c}_{(\Omega_\mathbf{J}, \|.\|)}, \mathrm{C}_{(\Omega_\mathbf{J}, \|.\|)}$ in the proof, it will be crucial to control precisely their scaling w.r.t. $|\mathbf{J}|$ (or $|\mathbf{J}^c|$) and the weights $(d^G)_{G \in \mathcal{G}}$.

Moreover, our proof will rely on the control of the *expected dual norm for isonormal vectors*: $\mathbb{E}\left[(\Omega_\mathbf{J}^c)^*(W)\right]$ with W a centered Gaussian random variable with unique covariance matrix. In the case of the Lasso, it is of order $(\log p)^{1/2}$, but could be much less in our settings, showing that restricting the set of allowed patterns leads to better sampling complexities.

Following [Bach, 2008b] and [Nardi and Rinaldo, 2008], we consider the reduced problem on $\mathbf{J}$,

$$\min_{w \in \mathbb{R}^p} L_{\mathbf{J}}(w_{\mathbf{J}}) + \mu \Omega_{\mathbf{J}}(w_{\mathbf{J}})$$

with solution $\hat{w}_{\mathbf{J}}$, which can be extended to $\mathbf{J}^c$ with zeros. From optimality conditions (see Lemma F.10), we know that

$$\Omega^*_{\mathbf{J}}[Q_{\mathbf{JJ}}(\hat{w}_{\mathbf{J}} - \mathbf{w}_{\mathbf{J}}) - q_{\mathbf{J}}] \leqslant \mu. \tag{E.1}$$

We denote by $\nu = \min\{|\mathbf{w}_j|;\ \mathbf{w}_j \neq 0\}$ the smallest nonzero components of $\mathbf{w}$. We first prove that we must have with high-probability $\|\hat{w}_G\|_2 > 0$ for all $G \in \mathcal{G}_{\mathbf{J}}$, proving that the hull of the active set of $\hat{w}_{\mathbf{J}}$ is exactly $\mathbf{J}$ (i.e., no active groups are missing).

We have

$$\begin{aligned}\|\hat{w}_{\mathbf{J}} - \mathbf{w}_{\mathbf{J}}\|_2 &\leqslant \left\|Q^{-1}_{\mathbf{JJ}}\right\|_2 \|Q_{\mathbf{JJ}}(\hat{w}_{\mathbf{J}} - \mathbf{w}_{\mathbf{J}})\|_2 \\ &\leqslant \kappa^{-1}|\mathbf{J}|^{1/2} \left(\|Q_{\mathbf{JJ}}(\hat{w}_{\mathbf{J}} - \mathbf{w}_{\mathbf{J}}) - q_{\mathbf{J}}\|_\infty + \|q_{\mathbf{J}}\|_\infty\right),\end{aligned}$$

hence from (E.1) and the definition of $\mathrm{C}_{(\Omega_{\mathbf{J}}, \|\cdot\|_1)}$,

$$\|\hat{w}_{\mathbf{J}} - \mathbf{w}_{\mathbf{J}}\|_2 \leq \kappa^{-1}|\mathbf{J}|^{1/2} \left(\mu \mathrm{C}_{(\Omega_{\mathbf{J}}, \|\cdot\|_1)} + \|q_{\mathbf{J}}\|_\infty\right). \tag{E.2}$$

Thus, if we assume $\mu \leqslant \frac{\kappa\nu}{3|\mathbf{J}|^{1/2}\mathrm{C}_{(\Omega_{\mathbf{J}}, \|\cdot\|_1)}}$ and

$$\|q_{\mathbf{J}}\|_\infty \leqslant \frac{\kappa\nu}{3|\mathbf{J}|^{1/2}}, \tag{E.3}$$

we get

$$\|\hat{w}_{\mathbf{J}} - \mathbf{w}_{\mathbf{J}}\|_\infty \leqslant \|\hat{w}_{\mathbf{J}} - \mathbf{w}_{\mathbf{J}}\|_2 \leqslant 2\nu/3, \tag{E.4}$$

so that for all $G \in \mathcal{G}_{\mathbf{J}}$, $\|\hat{w}_G\|_2 \geqslant \frac{\nu}{3}$, hence the hull is indeed selected.

This also ensures that $\hat{w}_{\mathbf{J}}$ satisfies the equation (see Lemma F.10)

$$Q_{\mathbf{JJ}}\left(\hat{w}_{\mathbf{J}} - \mathbf{w}_{\mathbf{J}}\right) - q_{\mathbf{J}} + \mu \hat{r}_{\mathbf{J}} = 0, \tag{E.5}$$

where

$$\hat{r} = \sum_{G \in \mathcal{G}_{\mathbf{J}}} \frac{d^G \circ d^G \circ \hat{w}}{\|d^G \circ \hat{w}\|_2}.$$

We now prove that the $\hat{w}$ padded with zeros on $\mathbf{J}^c$ is indeed optimal for the full problem with high probability. According to Lemma F.10, since we have already proved (E.5), it suffices to show that

$$(\Omega^c_{\mathbf{J}})^*[\nabla L(\hat{w})_{\mathbf{J}^c}] \leq \mu.$$

Defining $q_{\mathbf{J}^c|\mathbf{J}} = q_{\mathbf{J}^c} - Q_{\mathbf{J}^c\mathbf{J}}Q^{-1}_{\mathbf{JJ}}q_{\mathbf{J}}$, we can write the gradient of L on $\mathbf{J}^c$ as

$$\nabla L(\hat{w})_{\mathbf{J}^c} = -q_{\mathbf{J}^c|\mathbf{J}} - \mu Q_{\mathbf{J}^c\mathbf{J}}Q^{-1}_{\mathbf{JJ}}\hat{r}_{\mathbf{J}} = -q_{\mathbf{J}^c|\mathbf{J}} - \mu Q_{\mathbf{J}^c\mathbf{J}}Q^{-1}_{\mathbf{JJ}}(\hat{r}_{\mathbf{J}} - \mathbf{r}_{\mathbf{J}}) - \mu Q_{\mathbf{J}^c\mathbf{J}}Q^{-1}_{\mathbf{JJ}}\mathbf{r}_{\mathbf{J}},$$

which leads us to control the difference $\hat{r}_{\mathbf{J}} - \mathbf{r}_{\mathbf{J}}$. Consider a fixed $j \in \mathbf{J}$. We have $\hat{r}_j - \mathbf{r}_j = R(\hat{w}) - R(\mathbf{w})$ with $R(w) = \sum_{G \in \mathcal{G}_{\mathbf{J}}} (d^G_j)^2 w_j \|d^G \circ w\|_2^{-1}$. Since for any $k \in \mathbf{J}$

$$\frac{\partial R}{\partial w_k}(w) = \sum_{G \in \mathcal{G}_{\mathbf{J}}} \frac{(d^G_j)^2}{\|d^G \circ w\|_2} \mathbb{I}_{j=k} - \sum_{G \in \mathcal{G}_{\mathbf{J}}} \frac{(d^G_j)^2 w_j}{\|d^G \circ w\|_2^3} (d^G_k)^2 w_k,$$

with $\mathbb{I}_{j=k} = 1$ if $j = k$ and 0 otherwise, the mean value theorem gives that for some $w \in [\hat{w}, \mathbf{w}] = \{u \in \mathbb{R}^p;\ u_j \in\ [\hat{w}_j, \mathbf{w}_j]\}$, we have

$$\begin{aligned}|\hat{r}_j - \mathbf{r}_j| &\leqslant \sum_{k\in\mathbf{J}} \left|\frac{\partial R}{\partial w_k}(w)\right| |\hat{w}_k - \mathbf{w}_k| \\ &\leqslant \|\hat{w}_\mathbf{J} - \mathbf{w}_\mathbf{J}\|_\infty \left(\sum_{G\in\mathcal{G}_\mathbf{J}} \frac{(d_j^G)^2}{\|d^G \circ w\|_2} + \sum_{k\in\mathbf{J}}\sum_{G\in\mathcal{G}_\mathbf{J}} \frac{(d_j^G)^2|w_j|}{\|d^G \circ w\|_2^3}(d_k^G)^2|w_k|\right),\end{aligned}$$

hence

$$\|\hat{r}_\mathbf{J} - \mathbf{r}_\mathbf{J}\|_1 \leqslant \|\hat{w}_\mathbf{J} - \mathbf{w}_\mathbf{J}\|_\infty \left(\sum_{G\in\mathcal{G}_\mathbf{J}} \frac{\|d_\mathbf{J}^G\|_2^2}{\|d^G \circ w\|_2} + \sum_{G\in\mathcal{G}_\mathbf{J}} \frac{\|d^G \circ d^G \circ w\|_1^2}{\|d^G \circ w\|_2^3}\right).$$

Let $\mathbf{K} = \{k \in \mathbf{J} : \mathbf{w}_k \neq 0\}$ and

$$\varphi = \sup_{\substack{u\in\mathbb{R}^p:\mathbf{K}\subset\{k\in\mathbf{J}:u_k\neq 0\}\subset\mathbf{J}\\ G\in\mathcal{G}_\mathbf{J}}} \frac{\|d^G \circ d^G \circ u\|_1}{\|d_\mathbf{K}^G \circ d_\mathbf{K}^G \circ u_\mathbf{K}\|_1} \geqslant 1.$$

By using (E.4), we have

$$\|d^G \circ w\|_2^2 \geqslant \|d_\mathbf{K}^G \circ w_\mathbf{K}\|_2^2 \geqslant \|d_\mathbf{K}^G \circ d_\mathbf{K}^G \circ w_\mathbf{K}\|_1 \frac{\nu}{3} \geqslant \|d^G \circ d^G \circ w\|_1 \frac{\nu}{3\varphi},$$

$$\|d^G \circ w\|_2 \geqslant \|d_\mathbf{K}^G \circ w_\mathbf{K}\|_2 \geqslant \|d_\mathbf{K}^G\|_2 \frac{\nu}{3} \geqslant \|d_\mathbf{J}^G\|_2 \frac{\nu}{3\sqrt{\varphi}}$$

and

$$\|w\|_\infty \leqslant \frac{5}{3}\|\mathbf{w}\|_\infty.$$

Therefore we have

$$\begin{aligned}\|\hat{r}_\mathbf{J} - \mathbf{r}_\mathbf{J}\|_1 &\leqslant \|\hat{w}_\mathbf{J} - \mathbf{w}_\mathbf{J}\|_\infty \sum_{G\in\mathcal{G}_\mathbf{J}} \left(\frac{\|d_\mathbf{J}^G\|_2^2}{\|d^G \circ w\|_2} + \frac{5\varphi}{\nu}\frac{\|\mathbf{w}\|_\infty \|d_\mathbf{J}^G \circ d_\mathbf{J}^G\|_1}{\|d^G \circ w\|_2}\right) \\ &\leqslant \frac{3\sqrt{\varphi}\,\|\hat{w}_\mathbf{J} - \mathbf{w}_\mathbf{J}\|_\infty}{\nu}\left(1 + \frac{5\varphi\|\mathbf{w}\|_\infty}{\nu}\right)\sum_{G\in\mathcal{G}_\mathbf{J}} \|d_\mathbf{J}^G\|_2.\end{aligned}$$

Introducing $\mathcal{L} = \frac{18\varphi^{3/2}\|\mathbf{w}\|_\infty}{\nu^2}\sum_{G\in\mathcal{G}_\mathbf{J}} \|d_\mathbf{J}^G\|_2$, we thus have proved

$$\|\hat{r}_\mathbf{J} - \mathbf{r}_\mathbf{J}\|_2 \leq \mathcal{L}\,\|\hat{w}_\mathbf{J} - \mathbf{w}_\mathbf{J}\|_2. \tag{E.6}$$

By writing the Schur complement of Q on the block matrices $Q_{\mathbf{J}^c\mathbf{J}^c}$ and $Q_{\mathbf{JJ}}$, the positiveness of Q implies that the diagonal terms $\mathrm{diag}(Q_{\mathbf{J}^c\mathbf{J}}Q_{\mathbf{JJ}}^{-1}Q_{\mathbf{J}\mathbf{J}^c})$ are less than one, which implies that $\left\|Q_{\mathbf{J}^c\mathbf{J}}Q_{\mathbf{JJ}}^{-1/2}\right\|_\infty \leqslant 1$. We then have

$$\begin{aligned}\left\|Q_{\mathbf{J}^c\mathbf{J}}Q_{\mathbf{JJ}}^{-1}(\hat{r}_\mathbf{J} - \mathbf{r}_\mathbf{J})\right\|_\infty &= \left\|Q_{\mathbf{J}^c\mathbf{J}}Q_{\mathbf{JJ}}^{-1/2}Q_{\mathbf{JJ}}^{-1/2}(\hat{r}_\mathbf{J} - \mathbf{r}_\mathbf{J})\right\|_\infty &\text{(E.7)}\\ &\leqslant \left\|Q_{\mathbf{J}^c\mathbf{J}}Q_{\mathbf{JJ}}^{-1/2}\right\|_\infty \left\|Q_{\mathbf{JJ}}^{-1/2}\right\|_2 \|\hat{r}_\mathbf{J} - \mathbf{r}_\mathbf{J}\|_2 &\text{(E.8)}\\ &\leqslant \kappa^{-1/2}\,\|\hat{r}_\mathbf{J} - \mathbf{r}_\mathbf{J}\|_2 &\text{(E.9)}\\ &\leqslant \kappa^{-3/2}\mathcal{L}|\mathbf{J}|^{1/2}\left(\mu\mathrm{C}_{(\Omega_\mathbf{J},\|.\|_1)} + \|q_\mathbf{J}\|_\infty\right), &\text{(E.10)}\end{aligned}$$

where the last line comes from Eq. (E.2) and (E.6). We get

$$(\Omega^c_{\mathbf{J}})^*[Q_{\mathbf{J}^c\mathbf{J}}Q^{-1}_{\mathbf{J}\mathbf{J}}(\hat{r}_{\mathbf{J}} - \mathbf{r}_{\mathbf{J}})] \leqslant \mathrm{c}^{-1}_{(\Omega^c_{\mathbf{J}},\|.\|_1)} \frac{\mathcal{L}|\mathbf{J}|^{1/2}}{\kappa^{3/2}} \left(\mu \mathrm{C}_{(\Omega_{\mathbf{J}},\|.\|_1)} + \|q_{\mathbf{J}}\|_\infty\right).$$

Thus, if the following inequalities are verified

$$\frac{\mathcal{L}|\mathbf{J}|^{1/2}}{\kappa^{3/2}\mathrm{c}_{(\Omega^c_{\mathbf{J}},\|.\|_1)}}\mu \mathrm{C}_{(\Omega_{\mathbf{J}},\|.\|_1)} \leqslant \frac{\tau}{4}, \tag{E.11}$$

$$\frac{\mathcal{L}|\mathbf{J}|^{1/2}}{\kappa^{3/2}\mathrm{c}_{(\Omega^c_{\mathbf{J}},\|.\|_1)}} \|q_{\mathbf{J}}\|_\infty \leqslant \frac{\tau}{4}, \tag{E.12}$$

$$(\Omega^c_{\mathbf{J}})^*[q_{\mathbf{J}^c|\mathbf{J}}] \leqslant \frac{\mu\tau}{2}, \tag{E.13}$$

we obtain

$$\begin{aligned}(\Omega^c_{\mathbf{J}})^*[\nabla L(\hat{w})_{\mathbf{J}^c}] &\leqslant (\Omega^c_{\mathbf{J}})^*[-q_{\mathbf{J}^c|\mathbf{J}} - \mu Q_{\mathbf{J}^c\mathbf{J}}Q^{-1}_{\mathbf{J}\mathbf{J}}\mathbf{r}_{\mathbf{J}}] \\ &\leqslant (\Omega^c_{\mathbf{J}})^*[-q_{\mathbf{J}^c|\mathbf{J}}] + \mu(1-\tau) + \mu\tau/2 \leqslant \mu,\end{aligned}$$

i.e., $\mathbf{J}$ is exactly selected.

Combined with earlier constraints, this leads to the first part of the desired proposition.

We now need to make sure that the conditions (E.3), (E.12) and (E.13) hold with high-probability. To this end, we upperbound, using Gaussian concentration inequalities, two tail-probabilities. First, $q_{\mathbf{J}^c|\mathbf{J}}$ is a centered Gaussian random vector with covariance matrix

$$\begin{aligned}\mathbb{E}q_{\mathbf{J}^c|\mathbf{J}}q^\top_{\mathbf{J}^c|\mathbf{J}} &= \mathbb{E}\Big(q_{\mathbf{J}^c}q^\top_{\mathbf{J}^c} - q_{\mathbf{J}^c}q^\top_{\mathbf{J}}Q^{-1}_{\mathbf{J}\mathbf{J}}Q_{\mathbf{J}\mathbf{J}^c} - Q_{\mathbf{J}^c\mathbf{J}}Q^{-1}_{\mathbf{J}\mathbf{J}}q_{\mathbf{J}}q^\top_{\mathbf{J}^c} + Q_{\mathbf{J}^c\mathbf{J}}Q^{-1}_{\mathbf{J}\mathbf{J}}q_{\mathbf{J}}q^\top_{\mathbf{J}}Q^{-1}_{\mathbf{J}\mathbf{J}}Q_{\mathbf{J}\mathbf{J}^c}\Big) \\ &= \frac{\sigma^2}{n}Q_{\mathbf{J}^c\mathbf{J}^c|\mathbf{J}},\end{aligned}$$

where $Q_{\mathbf{J}^c\mathbf{J}^c|\mathbf{J}} = Q_{\mathbf{J}^c\mathbf{J}^c} - Q_{\mathbf{J}^c\mathbf{J}}Q^{-1}_{\mathbf{J}\mathbf{J}}Q_{\mathbf{J}\mathbf{J}^c}$. In particular, $(\Omega^c_{\mathbf{J}})^*[q_{\mathbf{J}^c|\mathbf{J}}]$ has the same distribution as $\zeta(W)$, with $\zeta : u \mapsto (\Omega^c_{\mathbf{J}})^*(\sigma n^{-1/2}Q^{1/2}_{\mathbf{J}^c\mathbf{J}^c|\mathbf{J}}u)$ and W a centered Gaussian random variable with unique covariance matrix.

Since for any u we have $u^\top Q_{\mathbf{J}^c\mathbf{J}^c|\mathbf{J}}u \leqslant u^\top Q_{\mathbf{J}^c\mathbf{J}^c}u \leqslant \left\|Q^{1/2}\right\|_2^2 \|u\|_2^2$, by using Sudakov-Fernique inequality [Adler, 1990, Theorem 2.9], we get:

$$\begin{aligned}\mathbb{E}[(\Omega^c_{\mathbf{J}})^*[q_{\mathbf{J}^c|\mathbf{J}}] = \mathbb{E}\sup_{\Omega^c_{\mathbf{J}}(u)\leqslant 1} u^\top q_{\mathbf{J}^c|\mathbf{J}} &\leqslant \sigma n^{-1/2}\|Q\|_2^{1/2}\mathbb{E}\sup_{\Omega^c_{\mathbf{J}}(u)\leqslant 1} u^\top W \\ &\leqslant \sigma n^{-1/2}\|Q\|_2^{1/2}\mathbb{E}[(\Omega^c_{\mathbf{J}})^*(W)]\end{aligned}$$

We have $|\zeta(u)| \leqslant \sigma n^{-1/2}\mathrm{c}^{-1}_{(\Omega^c_{\mathbf{J}},\|.\|_2)}\|Q^{1/2}_{\mathbf{J}^c\mathbf{J}^c|\mathbf{J}}u\|_2$ and $\|Q^{1/2}_{\mathbf{J}^c\mathbf{J}^c|\mathbf{J}}u\|_2 \leqslant \|Q^{1/2}\|_2\|u\|_2$, hence ζ is a Lipschitz function with Lipschitz constant upper bounded by $\sigma n^{-1/2}\mathrm{c}^{-1}_{(\Omega^c_{\mathbf{J}},\|.\|_2)}\|Q\|_2^{1/2}$. Thus by concentration of Lipschitz functions of multivariate standard random variables [Massart, 2003, Theorem 3.4], we have:

$$\mathbb{P}\Big[(\Omega^c_{\mathbf{J}})^*[q_{\mathbf{J}^c|\mathbf{J}}] \geqslant t + \sigma n^{-1/2}\|Q\|_2^{1/2}\mathbb{E}\left[(\Omega^c_{\mathbf{J}})^*(W)\right]\Big] \leqslant \exp\left(-\frac{nt^2}{2\|Q\|_2\sigma^2\mathrm{c}^{-2}_{(\Omega^c_{\mathbf{J}},\|.\|_2)}}\right).$$

Applied for $t = \mu\tau/2 \geqslant 2\sigma n^{-1/2}\|Q\|_2^{1/2}\mathbb{E}\left[(\Omega_{\mathbf{J}}^c)^*(W)\right]$, we get (because $(u-1)^2 \geqslant u^2/4$ for $u \geqslant 2$):

$$\mathbb{P}\left[(\Omega_{\mathbf{J}}^c)^*[q_{\mathbf{J}^c|\mathbf{J}}] \geqslant t\right] \leqslant \exp\left(-\frac{n\mu^2\tau^2}{32\|Q\|_2\sigma^2 \mathrm{c}_{(\Omega_{\mathbf{J}}^c,\|\cdot\|_2)}^{-2}}\right) \leqslant \exp\left(-\frac{n\mu^2\tau^2\mathrm{c}_{(\Omega_{\mathbf{J}}^c,\|\cdot\|_1)}^2}{32\|Q\|_2\sigma^2}\right).$$

It finally remains to control the term $\mathbb{P}(\|q_{\mathbf{J}}\|_\infty \geqslant \xi)$, with

$$\xi = \frac{\kappa\nu}{3|\mathbf{J}|^{1/2}} \min\left\{1, \frac{3\tau\kappa^{1/2}\mathrm{c}_{(\Omega_{\mathbf{J}}^c,\|\cdot\|_1)}}{4\mathcal{L}\nu}\right\}.$$

We can apply classical inequalities for standard random variables [Massart, 2003, Theorem 3.4] that directly lead to

$$\mathbb{P}(\|q_{\mathbf{J}}\|_\infty \geqslant \xi) \leqslant 2|\mathbf{J}| \exp\left(-\frac{n\xi^2}{2\sigma^2}\right).$$

To conclude, Theorem 5.2 holds with

$$C_1(\mathcal{G},\mathbf{J}) = \frac{\mathrm{c}_{(\Omega_{\mathbf{J}}^c,\|\cdot\|_1)}^2}{16\|Q\|_2}, \tag{E.14}$$

$$C_2(\mathcal{G},\mathbf{J}) = \left(\frac{\kappa\nu}{3}\min\left\{1, \frac{\tau\kappa^{1/2}\nu\mathrm{c}_{(\Omega_{\mathbf{J}}^c,\|\cdot\|_1)}}{24\varphi^{3/2}\|\mathbf{w}\|_\infty \sum_{G\in\mathcal{G}_{\mathbf{J}}}\left\|d_{\mathbf{J}}^G\right\|_2}\right\}\right)^2, \tag{E.15}$$

$$C_3(\mathcal{G},\mathbf{J}) = 4\|Q\|_2^{1/2}\mathbb{E}\left[(\Omega_{\mathbf{J}}^c)^*(W)\right], \tag{E.16}$$

and

$$C_4(\mathcal{G},\mathbf{J}) = \frac{\kappa\nu}{3\mathrm{C}_{(\Omega_{\mathbf{J}},\|\cdot\|_1)}}\min\left\{1, \frac{\tau\kappa^{1/2}\mathrm{c}_{(\Omega_{\mathbf{J}}^c,\|\cdot\|_1)}}{24\varphi^{3/2}\sum_{G\in\mathcal{G}_{\mathbf{J}}}\left\|d_{\mathbf{J}}^G\right\|_2}\frac{\nu}{\|\mathbf{w}\|_\infty}\right\},$$

where we recall the definitions: W a centered Gaussian random variable with unit covariance matrix, $\mathbf{K} = \{j \in \mathbf{J} : \mathbf{w}_j \neq 0\}$, $\nu = \min\{|\mathbf{w}_j|;\ j \in \mathbf{K}\}$,

$$\varphi = \sup_{\substack{u\in\mathbb{R}^p:\mathbf{K}\subset\{k\in\mathbf{J}:u_k\neq 0\}\subset\mathbf{J}\\ G\in\mathcal{G}_{\mathbf{J}}}} \frac{\|d^G\circ d^G\circ u\|_1}{\|d_{\mathbf{K}}^G\circ d_{\mathbf{K}}^G\circ u_{\mathbf{K}}\|_1},$$

$\kappa = \lambda_{\min}(Q_{\mathbf{JJ}}) > 0$ and $\tau > 0$ such that $(\Omega_{\mathbf{J}}^c)^*[Q_{\mathbf{J}^c\mathbf{J}}Q_{\mathbf{JJ}}^{-1}\mathbf{r}] < 1-\tau$.

F Technical lemmas

In this last section of the appendix, we give several technical lemmas, mostly relative to the optimization results. In addition, we consider $I \subseteq \{1,\dots,p\}$ and $\mathcal{G}_I = \{G \in \mathcal{G};\ G \cap I \neq \varnothing\} \subseteq \mathcal{G}$, i.e., the set of active groups when the variables I are selected. We will need to deal with two specific convex sets, E and Γ, defined as

$$\mathrm{E}_I = \{\eta \in \mathbb{R}^{|\mathcal{G}_I|};\ \eta_G \geq 0 \text{ and } \sum_{G\in\mathcal{G}_I}\eta_G \leq 1\},$$

and

$$\Gamma_I = \{\gamma \in \mathbb{R}^{|\mathcal{G}_I|\times|I|};\ \gamma_{Gj} \geq 0,\ \sum_{\substack{G\in\mathcal{G}_I\\ G\ni j}}\gamma_{Gj} = 1 \text{ and } \gamma_{Gj} = 0 \text{ if } j \notin G\}.$$

Furthemore, we assume that $\frac{u}{v}$ is defined by continuation at zero by $\frac{u}{0} = \infty$ if $u \neq 0$ and 0 otherwise.

We first start with two lemmas based on the Cauchy-Schwartz inequality [Hardy et al., 1988].

Lemma F.1. *For all vectors x, y in $\mathbb{R}^m$ such that $\forall j \in \{1, \ldots, m\}$, $x_j \geq 0$ and $y_j \geq 0$, we have the following variational equality*

$$x^\top y = \min_{\substack{z \in \mathbb{R}^m \\ z_j \geq 0 \\ \left(\sum_{j=1}^m z_j^2 y_j^2\right)^{1/2} \leq 1}} \left\{ \sum_{j=1}^m z_j^{-2} x_j^2 \right\}^{\frac{1}{2}}, \tag{F.1}$$

and the minimum is obtained for

$$z_j = \begin{cases} (x^\top y)^{-\frac{1}{2}} \frac{x_j^{1/2}}{y_j^{1/2}} & \text{if } y_j \neq 0, \\ 0 & \text{if } x_j = y_j = 0, \\ \infty & \text{otherwise.} \end{cases} \tag{F.2}$$

Similarly, we have for all vectors x, z in $\mathbb{R}^m$

$$\max_{\substack{y \in \mathbb{R}^m \\ \left(\sum_{j=1}^m z_j^2 y_j^2\right)^{1/2} \leq 1}} x^\top y = \left\{ \sum_{j=1}^m z_j^{-2} x_j^2 \right\}^{\frac{1}{2}}, \tag{F.3}$$

whose maximum is obtained for

$$y_j = \begin{cases} \left\{ \sum_{z_i \neq 0} z_i^{-2} x_i^2 \right\}^{-\frac{1}{2}} z_j^{-2} x_j & \text{if } z_j \neq 0, \\ \infty & \text{otherwise.} \end{cases} \tag{F.4}$$

Proof. We apply the Cauchy-Schwartz inequality on

$$\begin{aligned} x^\top y &= \sum_{x_j, y_j \neq 0} z_j y_j z_j^{-1} x_j \\ &\leq \left\{ \sum_{x_j, y_j \neq 0} z_j^2 y_j^2 \right\}^{\frac{1}{2}} \left\{ \sum_{x_j, y_j \neq 0} z_j^{-2} x_j^2 \right\}^{\frac{1}{2}}, \end{aligned}$$

where the right side of the equality becomes infinite if one of the z_j equals zero. The equality in the Cauchy-Schwartz inequality happens when there exists $k > 0$ such that for all j, $z_j^2 y_j^2 = k z_j^{-2} x_j^2$. Put together with the normalization of the vector z, we get the desired result.

The second part of the Lemma follows along similar lines.

If for all j, $z_j \neq 0$, note that this second result comes down to the computation of the dual norm of a weighted ℓ_2-norm.

Moreover, in the degenerated case where one of the z_j is infinite, the corresponding component of $y_j \propto z_j^{-2}$ is necessarily put to zero (with the added convention that $0 \cdot \infty = 0$) to ensure the feasibility of the problem (i.e., $\sum_{j=1}^m z_j^2 y_j^2 \leq 1$). ∎

Lemma F.2. *For all vector x in $\mathbb{R}^m$, we have the following variational equality*

$$\frac{1}{\|x\|_2^2} = \min_{\substack{z\in\mathbb{R}^m \\ z_j\geq 0 \\ \sum_{j=1}^m z_j=1}} \left\{ \sum_{j=1}^m z_j^2 |x_j|^{-2} \right\}, \tag{F.5}$$

and the minimum is obtained for

$$z_j = \frac{|x_j|^2}{\sum_{i=1}^m |x_i|^2}. \tag{F.6}$$

Proof. We apply the Cauchy-Schwartz inequality, starting from

$$\begin{aligned} 1 = \sum_{j=1}^m z_j &= \sum_{z_j\neq 0} z_j |x_j|^{-1} |x_j| \\ &\leq \left\{ \sum_{z_j\neq 0} z_j^2 |x_j|^{-2} \right\}^{\frac{1}{2}} \|x\|_2 , \end{aligned}$$

where the right side of the equality becomes infinite if one of the x_j equals zero. The equality in (F.5) is trivial for $x = 0$ and obtained for $z_j = |x_j|^2 / \|x\|_2^2$ otherwise. □

Throughout this section, we will use a variational representation of $u_I \mapsto \sum_{G\in\mathcal{G}_I} \|d_I^G \circ u_I\|_2$ in terms of the vectors $\eta \in \mathrm{E}_I$:

Lemma F.3. *For all $u_I \in \mathbb{R}^{|I|}$,*

$$\sum_{G\in\mathcal{G}_I} \|d_I^G \circ u_I\|_2 = \min_{\eta\in\mathrm{E}_I} \left\{ \sum_{G\in\mathcal{G}_I} \frac{\|d_I^G \circ u_I\|_2^2}{\eta_G} \right\}^{\frac{1}{2}}. \tag{F.7}$$

and the minimum is obtained for

$$\eta_G = \frac{\|d_I^G \circ u_I\|_2}{\sum_{H\in\mathcal{G}_I} \|d_I^H \circ u_I\|_2}. \tag{F.8}$$

Proof. The result comes from the Lemma F.1 when we take the following vectors

$$x = [\|d_I^G \circ u_I\|_2]_{G\in\mathcal{G}_I}, \; y = [1]_{G\in\mathcal{G}_I}$$

and

$$z = \left[\eta_G^{\frac{1}{2}}\right]_{G\in\mathcal{G}_I}.$$

It suffices to notice that for all vectors $\eta \in \mathrm{E}_I$, the vector $z = \left[\eta_G^{\frac{1}{2}}\right]_{G\in\mathcal{G}_I}$ is feasible. □

For all $\eta \in \mathrm{E}_I$, we introduce the vector $\zeta(\eta) \in \mathbb{R}^{|I|}$ defined for all $j \in I$ as

$$\zeta_j(\eta) = \left[\sum_{\substack{G \in \mathcal{G}_I \\ G \ni j}} \eta_G^{-1} (d_j^G)^2 \right]^{-1} .$$

We can then rewrite the term $\sum_{G \in \mathcal{G}_I} \|d_I^G \circ u_I\|_2$ in terms of $\zeta(\eta)$ and we have

$$\sum_{G \in \mathcal{G}_I} \|d_I^G \circ u_I\|_2 = \min_{\eta \in \mathrm{E}_I} \left\{ \sum_{j \in I} \zeta_j(\eta)^{-1} |u_j|^2 \right\}^{\frac{1}{2}} . \tag{F.9}$$

As we shall see later, the η and ζ will be useful to characterize respectively the (in)active groups and variables.

In the following lemma, we show that $\eta \mapsto \zeta_j(\eta)$ is a concave function of η:

Lemma F.4. *For all $j \in I$ and all $\eta \in \mathrm{E}_I$, we have*

$$\zeta_j(\eta) = \min_{\gamma \in \Gamma_I} \left\{ \sum_{\substack{G \in \mathcal{G}_I \\ G \ni j}} \gamma_{Gj}^2 (d_j^G)^{-2} \eta_G \right\},$$

and $\eta \mapsto \zeta_j(\eta)$ is a concave function of η. In addition, the optimal γ is given by

$$\gamma_{Gj} = \frac{(d_j^G)^2 \, \eta_G^{-1}}{\sum_{\substack{H \in \mathcal{G}_I \\ H \ni j}} (d_j^H)^2 \, \eta_H^{-1}} = \frac{(d_j^G)^2 \, \eta_G^{-1}}{\zeta_j(\eta)^{-1}}. \tag{F.10}$$

Proof. The result is a direct application of the Lemma F.2 when we consider the vectors

$$x = \left[d_j^G \, \eta_G^{-\frac{1}{2}} \right]_{G \in \mathcal{G}_I, G \ni j},$$

and

$$z = \left[\gamma_{Gj} \right]_{G \in \mathcal{G}_I, G \ni j} .$$

The concavity comes from the fact that $\eta \mapsto \zeta_j(\eta)$ is defined as the pointwise infimum of concave (linear in fact) functions of η. □

Lemma F.5. *We have, for all $\eta \in \mathrm{E}_I$ and all $\kappa_I \in \mathbb{R}^{|I|}$,*

$$\max_{\left(\sum_{j \in I} \zeta_j(\eta)^{-1} |u_j|^2\right)^{1/2} \leq 1} u_I^\top \kappa_I = \left\{ \sum_{j \in I} \zeta_j(\eta) |\kappa_j|^2 \right\}^{\frac{1}{2}} . \tag{F.11}$$

Moreover, the maximum is obtained for

$$|u_j| = \begin{cases} \left\{ \sum_{i \in I} \zeta_i(\eta) |\kappa_i|^2 \right\}^{-\frac{1}{2}} \zeta_j(\eta) |\kappa_j| & \text{if } \zeta_j(\eta) < \infty, \\ \infty & \text{otherwise.} \end{cases} \tag{F.12}$$

Proof. We just need to apply the Lemma F.1 with the vectors

$$x = \kappa_I$$

and

$$z = \left[\zeta_j(\eta)^{-\frac{1}{2}}\right]_{j\in I}.$$

□

We now establish a link between the dual norm of $u_I \mapsto \sum_{G\in\mathcal{G}_I} \|d_I^G \circ u_I\|_2$ and the vector ζ.

Lemma F.6. *For any $\kappa_I \in \mathbb{R}^{|I|}$, we have the following relationship*

$$\max_{\sum_{G\in\mathcal{G}_I}\left\|d_I^G\circ u_I\right\|_2\leq 1} u_I^\top \kappa_I = \max_{\eta\in \mathrm{E}_I} \left\{\sum_{j\in I} \zeta_j(\eta)|\kappa_j|^2\right\}^{\frac{1}{2}}.$$

Proof. Thanks to the Lemma F.5, we can write

$$\max_{\eta\in\mathrm{E}_I} \max_{\left(\sum_{j\in I}\zeta_j(\eta)^{-1}|u_j|^2\right)^{1/2}\leq 1} u_I^\top \kappa_I = \max_{\eta\in \mathrm{E}_I} \left\{\sum_{j\in I} \zeta_j(\eta)|\kappa_j|^2\right\}^{\frac{1}{2}}.$$

The maximization problem on the left side of the equality can be first solved w.r.t. $\eta \in \mathrm{E}_I$, letting u fixed [Boyd and Vandenberghe, 2003, page 133]. The relationship (F.9) then provides

$$\max_{\substack{\eta\in\mathrm{E}_I \\ \left(\sum_{j\in I}\zeta_j(\eta)^{-1}|u_j|^2\right)^{1/2}\leq 1}} u_I^\top \kappa_I = \max_{\sum_{G\in\mathcal{G}_I}\left\|d_I^G\circ u_I\right\|_2\leq 1} u_I^\top \kappa_I,$$

which leads to the desired result.

□

We are now in position to express the dual norm of $u_I \mapsto \sum_{G\in\mathcal{G}_I} \|d_I^G \circ u_I\|_2$ through the variables $\gamma \in \Gamma_I$.

Lemma F.7. *We have the following relationship*

$$\max_{\sum_{G\in\mathcal{G}_I}\left\|d_I^G\circ u_I\right\|_2\leq 1} u_I^\top \kappa_I = \min_{\gamma\in\Gamma_I} \max_{G\in\mathcal{G}_J} \left\{\sum_{j\in G\cap I} \gamma_{Gj}^2\,(d_j^G)^{-2}|\kappa_j|^2\right\}^{\frac{1}{2}}.$$

Proof. Starting from the Lemma F.6,

$$\max_{\sum_{G\in\mathcal{G}_I}\left\|d_I^G\circ u_I\right\|_2\leq 1} u_I^\top \kappa_I = \max_{\eta\in \mathrm{E}_I} \left\{\sum_{j\in I} \zeta_j(\eta)|\kappa_j|^2\right\}^{\frac{1}{2}},$$

we use the variational characterization of $\zeta_j(\eta)$ given by

$$\zeta_j(\eta) = \min_{\gamma\in\Gamma_I} \left\{\sum_{\substack{G\in\mathcal{G}_I \\ G\ni j}} \gamma_{Gj}^2 (d_j^G)^{-2}\eta_G\right\}.$$

We minimize through a sum each column of γ independently over the set Γ_I (that also constraints each column of γ independently). We thus get the following max-min problem [Sion et al., 1957, Theorem 4.2]

$$\max_{\eta\in \mathrm{E}_I}\min_{\gamma\in\Gamma_I}\left\{\sum_{j\in I}\sum_{\substack{G\in\mathcal{G}_I\\ G\ni j}}\eta_G\,\gamma_{Gj}^2\,(d_j^G)^{-2}|\kappa_j|^2\right\}^{\frac{1}{2}} = \max_{\eta\in \mathrm{E}_I}\min_{\gamma\in\Gamma_I}\psi(\eta,\gamma).$$

The function

$$\eta\mapsto\psi(\eta,\gamma),$$

is concave and continuous for all $\gamma\in\Gamma_I$. Similarly, the function

$$\gamma\mapsto\psi(\eta,\gamma),$$

is convex for all $\eta\in\mathrm{E}_I$. In addition, the convex sets E_I and Γ_I are compact.

One can invert the max and min, so that we obtain

$$\max_{\eta\in \mathrm{E}_I}\min_{\gamma\in\Gamma_I}\psi(\eta,\gamma) = \min_{\gamma\in\Gamma_I}\max_{\eta\in \mathrm{E}_I}\psi(\eta,\gamma),$$

and the right-hand side can be rewritten as

$$\min_{\gamma\in\Gamma_I}\max_{\eta\in \mathrm{E}_I}\left\{\sum_{G\in\mathcal{G}_I}\eta_G\sum_{j\in G\cap I}\gamma_{Gj}^2\,(d_j^G)^{-2}|\kappa_j|^2\right\}^{\frac{1}{2}}.$$

The optimization over E_I can be performed in closed form and leads to the desired result. □

We provide a lemma to derive a lower-bound on $\min_{\gamma\in\Gamma_I}\max_{G\in\mathcal{G}_I}\left\{\sum_{j\in G\cap I}\gamma_{Gj}^2(d_j^G)^{-2}|\kappa_j|^2\right\}^{\frac{1}{2}}$. The underlying idea of this lemma is to replace the sum $\sum_{j\in G\cap I}$ by $\max_{j\in G\cap I}$.

Lemma F.8. *The quantity*

$$\min_{\gamma\in\Gamma_I}\max_{G\in\mathcal{G}_I}\left\{\sum_{j\in G\cap I}\gamma_{Gj}^2\,(d_j^G)^{-2}|\kappa_j|^2\right\}^{\frac{1}{2}}$$

is lowerbounded by

$$\min_{\gamma\in\Gamma_I}\max_{G\in\mathcal{G}_I}\max_{j\in G\cap I}\left\{\gamma_{Gj}\,(d_j^G)^{-1}|\kappa_j|\right\},$$

and this minimum is obtained for

$$\gamma_{Gj} = \frac{d_j^G}{\sum_{\substack{H\in\mathcal{G}_I\\ H\ni j}} d_j^H}.$$

Proof. The first part of the Lemma is straightforward since we have for all $\gamma\in\Gamma_I$,

$$\max_{G\in\mathcal{G}_I}\left\{\sum_{j\in G\cap I}\gamma_{Gj}^2\,(d_j^G)^{-2}|\kappa_j|^2\right\}^{\frac{1}{2}} \geq \max_{G\in\mathcal{G}_I}\left\{\max_{j\in G\cap I}\left(\gamma_{Gj}^2\,(d_j^G)^{-2}|\kappa_j|^2\right)\right\}^{\frac{1}{2}}.$$

Let us notice that for all $\gamma \in \Gamma_I$, since $t \mapsto t^{1/2}$ is stricly increasing,

$$\max_{G\in\mathcal{G}_I}\left\{\max_{j\in G\cap I}\left(\gamma_{Gj}^2\,(d_j^G)^{-2}|\kappa_j|^2\right)\right\}^{\frac{1}{2}} = \max_{G\in\mathcal{G}_I}\max_{j\in G\cap I}\left\{\gamma_{Gj}\,(d_j^G)^{-1}|\kappa_j|\right\}.$$

We now need to show that $\gamma_{Gj} = \frac{d_j^G}{\sum_{\substack{H\in\mathcal{G}_I\\ H\ni j}} d_j^H}$ is indeed optimal, i.e., for all $\gamma \in \Gamma_I$,

$$\begin{aligned}\max_{G\in\mathcal{G}_I}\max_{j\in G\cap I}\left\{\gamma_{Gj}(d_j^G)^{-1}|\kappa_j|\right\} &\geq \max_{G\in\mathcal{G}_I}\max_{j\in G\cap I}\left\{(d_j^G)^{-1}\frac{d_j^G}{\sum_{\substack{H\in\mathcal{G}_I\\ H\ni j}} d_j^H}|\kappa_j|\right\}\\ &= \max_{j\in I}\left\{\frac{|\kappa_j|}{\sum_{\substack{H\in\mathcal{G}_I\\ H\ni j}} d_j^H}\right\}.\end{aligned}$$

We denote by $j_0 \in I$ one of the indices that achieves the latter maximization. By contradiction, if there exists $\tilde{\gamma} \in \Gamma_I$ such that

$$\max_{G\in\mathcal{G}_I}\max_{j\in G\cap I}\left\{\tilde{\gamma}_{Gj}(d_j^G)^{-1}|\kappa_j|\right\} < \frac{|\kappa_{j_0}|}{\sum_{\substack{H\in\mathcal{G}_I\\ H\ni j_0}} d_{j_0}^H},$$

then, we notably have for all $G \ni j_0$,

$$\tilde{\gamma}_{Gj_0}(d_{j_0}^G)^{-1}|\kappa_{j_0}| < \frac{|\kappa_{j_0}|}{\sum_{\substack{H\in\mathcal{G}_I\\ H\ni j_0}} d_{j_0}^H},$$

which implies in turn that for all $G \ni j_0$,

$$\tilde{\gamma}_{Gj_0} < \frac{d_{j_0}^G}{\sum_{\substack{H\in\mathcal{G}_I\\ H\ni j_0}} d_{j_0}^H}.$$

A summation over $G \ni j_0$ leads to the contradiction $1 < 1$.

□

Given an active set $J \subseteq \{1,\dots,p\}$ and a direct parent $K \in \Pi_{\mathcal{P}}(J)$ of J in the DAG of nonzero patterns, we have the following result:

Lemma F.9. *For all $G \in \mathcal{G}_K\backslash\mathcal{G}_J$, we have*

$$K\backslash J \subseteq G$$

Proof. We proceed by contradiction. We assume there exists $G_0 \in \mathcal{G}_K\backslash\mathcal{G}_J$ such that $K\backslash J \nsubseteq G_0$. Given that $K \in \mathcal{P}$, there exists $\mathcal{G}' \subseteq \mathcal{G}$ verifying $K = \bigcap_{G\in\mathcal{G}'} G^c$. Note that $G_0 \notin \mathcal{G}'$ since by definition $G_0 \cap K \neq \varnothing$.

We can now build the pattern $\tilde{K} = \bigcap_{G\in\mathcal{G}'\cup\{G_0\}} G^c = K \cap G_0^c$ that belongs to $\mathcal{P}$. Moreover, $\tilde{K} = K \cap G_0^c \subset K$ since we assumed $G_0^c \cap K \neq \varnothing$. In addition, we have that $J \subset K$ and $J \subset G_0^c$ because $K \in \Pi_{\mathcal{P}}(J)$ and $G_0 \in \mathcal{G}_K\backslash\mathcal{G}_J$. This results in

$$J \subset \tilde{K} \subset K,$$

which is impossible by definition of K. □

We give below an important Lemma to characterize the solutions of (2.1).

Lemma F.10. *The vector $\hat{w} \in \mathbb{R}^p$ is a solution of*

$$\min_{w \in \mathbb{R}^p} L(w) + \mu\, \Omega(w)$$

if and only if

$$\begin{cases} \nabla L(\hat{w})_{\hat{J}} + \mu\, \hat{r}_{\hat{J}} = 0 \\ (\Omega^c_{\hat{J}})^*[\nabla L(\hat{w})_{\hat{J}^c}] \leq \mu, \end{cases}$$

with $\hat{J}$ the hull of $\{j \in \{1, \dots, p\}, \hat{w}_j \neq 0\}$ and the vector $\hat{r} \in \mathbb{R}^p$ defined as

$$\hat{r} = \sum_{G \in \mathcal{G}_{\hat{J}}} \frac{d^G \circ d^G \circ \hat{w}}{\|d^G \circ \hat{w}\|_2}.$$

In addition, the solution $\hat{w}$ satisfies

$$\Omega^*[\nabla L(\hat{w})] \leq \mu.$$

Proof. The problem

$$\min_{w \in \mathbb{R}^p} L(w) + \mu\, \Omega(w) = \min_{w \in \mathbb{R}^p} F(w)$$

being convex, the directional derivative optimality condition are necessary and sufficient [Borwein and Lewis, 2006, Propositions 2.1.1-2.1.2]. Therefore, the vector $\hat{w}$ is a solution of the previous problem if and only if for all directions $u \in \mathbb{R}^p$, we have

$$\lim_{\substack{\varepsilon \to 0 \\ \varepsilon > 0}} \frac{F(\hat{w} + \varepsilon u) - F(\hat{w})}{\varepsilon} \geq 0.$$

Some algebra leads to the following equivalent formulation

$$\forall u \in \mathbb{R}^p,\ u^\top \nabla L(\hat{w}) + \mu\, u^\top_{\hat{J}} \hat{r}_{\hat{J}} + \mu\, (\Omega^c_{\hat{J}})[u_{\hat{J}^c}] \geq 0. \tag{F.13}$$

The first part of the lemma then comes from the projections on $\hat{J}$ and $\hat{J}^c$.

An application of the Cauchy-Schwartz inequality on $u^\top_{\hat{J}} \hat{r}_{\hat{J}}$ gives for all $u \in \mathbb{R}^p$

$$u^\top_{\hat{J}} \hat{r}_{\hat{J}} \leq (\Omega_{\hat{J}})[u_{\hat{J}}].$$

Combined with the equation (F.13), we get

$$\forall u \in \mathbb{R}^p,\ u^\top \nabla L(\hat{w}) + \mu\, \Omega(u) \geq 0,$$

hence the second part of the lemma. □

We end up with a lemma regarding the dual norm of the sum of two *disjoint* norms (See [Rockafellar, 1970]):

Lemma F.11. *Let A and B be a partition of $\{1, \dots, p\}$, i.e., $A \cap B = \varnothing$ and $A \cup B = \{1, \dots, p\}$. We consider two norms $u_A \in \mathbb{R}^{|A|} \mapsto \|u_A\|_A$ and $u_B \in \mathbb{R}^{|B|} \mapsto \|u_B\|_B$, with dual norms $\|v_A\|^*_A$ and $\|v_B\|^*_B$. We have*

$$\max_{\|u_A\|_A + \|u_B\|_B \leq 1} u^\top v = \max\left\{\|v_A\|^*_A,\ \|v_B\|^*_B\right\}.$$

G Interpretation of the dual variables γ at optimality

The necessary and sufficient conditions we derive for our active set algorithm mostly rely on lower and upper-bounds for $\Omega^*(\nabla L(w))$, where w is a primal solution. Such bounds can be obtained via the dual parameters γ.

More precisely, when we have an optimal solution w^* whose active set is J, we will show that the corresponding γ^*_{Gj} are necessarily equal to zero for $G \in \mathcal{G}_J$ and $j \in J^c$.

When the duality gap for the full problem equals zero, we know that

$$\lambda \Omega^*(\nabla L(w^*)) = -\nabla L(w^*)^\top w^*$$

and

$$\sum_{G \in \mathcal{G}} \|d^G \circ w^*\|_2 \leq \lambda.$$

Therefore, w^* is a solution of the problem

$$\max_{\Omega(v) \leq 1} -v^\top \nabla L(w^*).$$

Following the Lemmas F.4 and F.6 , we consider the $(\eta^*, \zeta(\eta^*), \gamma^*)$ associated with the w^*. At optimality and according to the equations (F.8) and (F.12), we have $\zeta_j(\eta^*) = 0$ for the inactive variables $j \in J^c$ and $\eta^*_G > 0$ for the active groups $G \in \mathcal{G}_J$ (such that $G \cap J \neq \varnothing$).

The relationship (F.10) eventually shows that γ^*_{Gj} is necessarily equal to zero for $G \in \mathcal{G}_J$ and $j \in J^c$.