Journal of Machine Learning Research 12 (2011) 2777-2824 Submitted 9/09; Revised 3/10; Published 10/11

# Structured Variable Selection with Sparsity-Inducing Norms

**Rodolphe Jenatton** RODOLPHE.JENATTON@INRIA.FR
*INRIA - SIERRA Project-team,*
*Laboratoire d'Informatique de l'Ecole Normale Supérieure (INRIA/ENS/CNRS UMR 8548)*
*23, avenue d'Italie*
*75214 Paris, France*

**Jean-Yves Audibert** AUDIBERT@CERTIS.ENPC.FR
*Imagine (ENPC/CSTB), Université Paris-Est,*
*Laboratoire d'Informatique de l'Ecole Normale Supérieure (INRIA/ENS/CNRS UMR 8548)*
*6 avenue Blaise Pascal*
*77455 Marne-la-Vallée, France*

**Francis Bach** FRANCIS.BACH@INRIA.FR
*INRIA - SIERRA Project-team,*
*Laboratoire d'Informatique de l'Ecole Normale Supérieure (INRIA/ENS/CNRS UMR 8548),*
*23, avenue d'Italie*
*75214 Paris, France*

**Editor:** Bin Yu

## 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 $\ell_1$-norm and the group $\ell_1$-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.

**Keywords:** sparsity, consistency, variable selection, convex optimization, active set algorithm

## 1. Introduction

Sparse linear models have emerged as a powerful framework to deal with various supervised estimation tasks, in machine learning as well as in statistics and signal processing. These models basically seek to predict an output by linearly combining only a small subset of the features describing the data. To simultaneously address this variable selection and the linear model estimation, $\ell_1$-norm regularization has become a popular tool, that benefits both from efficient algorithms (see, e.g., Efron et al., 2004; Lee et al., 2007; Beck and Teboulle, 2009; Yuan et al., 2010; Bach et al., 2011, and multiple references therein) and well-developed theory for generalization properties and variable selection consistency (Zhao and Yu, 2006; Wainwright, 2009; Bickel et al., 2009; Zhang, 2009; Negahban et al., 2009).

When regularizing by the $\ell_1$-norm, sparsity is yielded by treating each variable individually, regardless of its position in the input feature vector, so that existing relationships and structures between the variables (e.g., spatial, hierarchical or related to the physics of the problem at hand) are merely disregarded. However, many practical situations could benefit from this type of prior knowledge, potentially both for interpretability purposes and for improved predictive performance.

For instance, in neuroimaging, one is interested in localizing areas in functional magnetic resonance imaging (fMRI) or magnetoencephalography (MEG) signals that are discriminative to distinguish between different brain states (Gramfort and Kowalski, 2009; Xiang et al., 2009; Jenatton et al., 2011a, and references therein). More precisely, fMRI responses consist in voxels whose three-dimensional spatial arrangement respects the anatomy of the brain. The discriminative voxels are thus expected to have a specific localized spatial organization (Xiang et al., 2009), which is important for the subsequent identification task performed by neuroscientists. In this case, regularizing by a plain $\ell_1$-norm to deal with the ill-conditionedness of the problem (typically only a few fMRI responses described by tens of thousands of voxels) would ignore this spatial configuration, with a potential loss in interpretability and performance.

Similarly, in face recognition, robustness to occlusions can be increased by considering as features, sets of pixels that form small convex regions on the face images (Jenatton et al., 2010b). Again, a plain $\ell_1$-norm regularization fails to encode this specific spatial locality constraint (Jenatton et al., 2010b). The same rationale supports the use of *structured sparsity* for background subtraction tasks (Cevher et al., 2008; Huang et al., 2009; Mairal et al., 2010b). Still in computer vision, object and scene recognition generally seek to extract bounding boxes in either images (Harzallah et al., 2009) or videos (Dalal et al., 2006). These boxes concentrate the predictive power associated with the considered object/scene class, and have to be found by respecting the spatial arrangement of the pixels over the images. In videos, where series of frames are studied over time, the temporal coherence also has to be taken into account. An unstructured sparsity-inducing penalty that would disregard this spatial and temporal information is therefore not adapted to select such boxes.

Another example of the need for higher-order prior knowledge comes from bioinformatics. Indeed, for the diagnosis of tumors, the profiles of array-based comparative genomic hybridization (arrayCGH) can be used as inputs to feed a classifier (Rapaport et al., 2008). These profiles are characterized by plenty of variables, but only a few samples of such profiles are available, prompting the need for variable selection. Because of the specific spatial organization of bacterial artificial chromosomes along the genome, the set of discriminative features is expected to have specific contiguous patterns. Using this prior knowledge on top of a standard sparsity-inducing method leads to improvement in classification accuracy (Rapaport et al., 2008). In the context of multi-task regression, a genetic problem of interest is to find a mapping between a small subset of single nucleotide polymorphisms (SNP's) that have a phenotypic impact on a given family of genes (Kim and Xing, 2010). This target family of genes has its own structure, where some genes share common genetic characteristics, so that these genes can be embedded into a underlying hierarchy (Kim and Xing, 2010). Exploiting directly this hierarchical information in the regularization term outperforms the unstructured approach with a standard $\ell_1$-norm. Such hierarchical structures have been likewise useful in the context of wavelet regression (Baraniuk et al., 2010; Zhao et al., 2009; Huang et al., 2009; Jenatton et al., 2011b), kernel-based non linear variable selection (Bach, 2008a), for topic modelling (Jenatton et al., 2011b) and for template selection in natural language processing (Martins et al., 2011).

These real world examples motivate the need for the design of sparsity-inducing regularization schemes, capable of encoding more sophisticated prior knowledge about the expected sparsity patterns.

As mentioned above, the $\ell_1$-norm focuses only on *cardinality* and cannot easily specify side information about the patterns of nonzero coefficients ("nonzero patterns") induced in the solution, since they are all theoretically possible. Group $\ell_1$-norms (Yuan and Lin, 2006; Roth and Fischer, 2008; Huang and Zhang, 2010) consider a partition of all variables into a certain number of subsets and penalize the sum of the Euclidean norms of each one, 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, 2008a; Kim and Xing, 2010; Jenatton et al., 2010a, 2011b; Schmidt and Murphy, 2010).

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. Before describing how to go from groups to nonzero patterns (or equivalently zero patterns), we show that it is possible to "reverse-engineer" a given set of nonzero patterns, that is, 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 in specific geometric and structured configurations, which covers the type of prior knowledge available in the real world applications described previously.

As will be shown in Section 3, for each set of groups, a notion of hull of a nonzero pattern may be naturally defined. For example, 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 well to high dimensions. Finally, we illustrate in Section 6 the behavior of our norms with synthetic examples on specific geometric settings, such as lines and two-dimensional grids.

### 1.1 Notation

For $x \in \mathbb{R}^p$ and $q \in [1,\infty)$, we denote by $\|x\|_q$ its $\ell_q$-norm defined as $(\sum_{j=1}^p |x_j|^q)^{1/q}$ and $\|x\|_\infty = \max_{j\in\{1,\ldots,p\}} |x_j|$. Given $w \in \mathbb{R}^p$ and a subset $J$ of $\{1,\ldots,p\}$ with cardinality $|J|$, $w_J$ denotes the vector in $\mathbb{R}^{|J|}$ of elements of $w$ indexed by $J$. Similarly, for a matrix $M \in \mathbb{R}^{p\times m}$, $M_{IJ} \in \mathbb{R}^{|I|\times|J|}$ denotes the sub-matrix of $M$ reduced to the rows indexed by $I$ and the columns indexed by $J$. For any finite set $A$ with cardinality $|A|$, we also define the $|A|$-tuple $(y^a)_{a\in A} \in \mathbb{R}^{p\times|A|}$ as the collection of $p$-dimensional vectors $y^a$ indexed by the elements of $A$. Furthermore, for two vectors $x$ and $y$ in $\mathbb{R}^p$, we denote by $x\circ y = (x_1y_1,\ldots,x_py_p)^\top \in \mathbb{R}^p$ the elementwise product of $x$ and $y$.

## 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 $(x_i,y_i) \in \mathbb{R}^p \times \mathcal{Y}$, $i = 1,\ldots,n$. We define the *empirical risk* of a loading vector $w \in \mathbb{R}^p$ as $L(w) = \frac{1}{n}\sum_{i=1}^n \ell\big(y_i, w^\top x_i\big)$, where $\ell : \mathcal{Y}\times\mathbb{R} \mapsto \mathbb{R}^+$ is a *loss function*. We assume that $\ell$

is *convex and continuously differentiable* with respect to the second parameter. Typical examples of loss functions are the square loss for least squares regression, that is, $\ell(y,\hat{y}) = \frac{1}{2}(y-\hat{y})^2$ 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 $\bigcup_{G\in\mathcal{G}} G = \{1,\dots,p\}$, that is, 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 $\Omega$ defined by

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

where $(d^G)_{G\in\mathcal{G}}$ is a $|\mathcal{G}|$-tuple of $p$-dimensional vectors such that $d_j^G > 0$ if $j \in G$ and $d_j^G = 0$ otherwise. A same variable $w_j$ belonging to two different groups $G_1, G_2 \in \mathcal{G}$ is allowed to be weighted differently in $G_1$ and $G_2$ (by respectively $d_j^{G_1}$ and $d_j^{G_2}$). We do not study the more general setting where each $d^G$ would be a (non-diagonal) positive-definite matrix, which we defer to future work. Note that a larger family of penalties with similar properties may be obtained by replacing the $\ell_2$-norm in Equation (1) by other $\ell_q$-norm, $q > 1$ (Zhao et al., 2009). Moreover, non-convex alternatives to Equation (1) with quasi-norms in place of norms may also be interesting, in order to yield sparsity more aggressively (see, e.g., Jenatton et al., 2010b).

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

- **$\ell_2$-norm**: $\mathcal{G}$ is composed of one element, the full set $\{1,\dots,p\}$.
- **$\ell_1$-norm**: $\mathcal{G}$ is the set of all singletons, leading to the Lasso (Tibshirani, 1996) for the square loss.
- **$\ell_2$-norm and $\ell_1$-norm**: $\mathcal{G}$ is the set of all singletons and the full set $\{1,\dots,p\}$, leading (up to the squaring of the $\ell_2$-norm) to the Elastic net (Zou and Hastie, 2005) for the square loss.
- **Group $\ell_1$-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, 2008a), then a set of $p$ groups, each of them composed of descendants of a given variable, is considered.

We study the following regularized problem:

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

where $\mu \geq 0$ is a regularization parameter. Note that a non-regularized constant term could be included in this formulation, but it is left out for simplicity. We denote by $\hat{w}$ any solution of Problem (2). Regularizing by linear combinations of (non-squared) $\ell_2$-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.

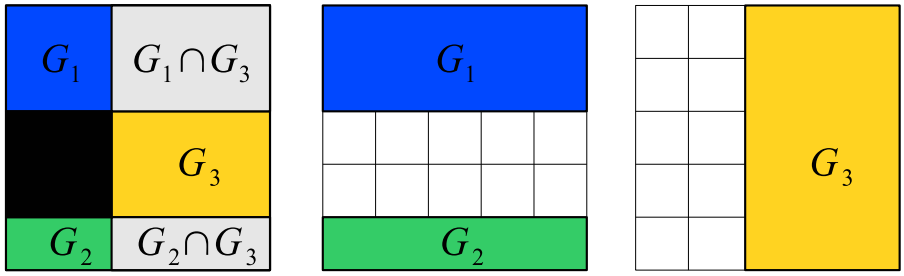



Figure 1: Groups and induced nonzero pattern: three sparsity-inducing groups (middle and right, denoted by $\{G_1, G_2, G_3\}$) with the associated nonzero pattern which is the complement of the union of groups, that is, $(G_1 \cup G_2 \cup G_3)^c$ (left, in black).

## 3. Groups and Sparsity Patterns

We now study the relationship between the norm $\Omega$ defined in Equation (1) 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 $(\ell_1, \ell_2)$-norm (Zhao et al., 2009). At the group level, it behaves like an $\ell_1$-norm and therefore, $\Omega$ induces group sparsity. In other words, each $d^G \circ w$, and equivalently each $w_G$ (since the support of $d^G$ is exactly $G$), is encouraged to go to zero. On the other hand, within the groups $G \in \mathcal{G}$, the $\ell_2$-norm does not promote sparsity. Intuitively, for a certain subset of groups $\mathcal{G}' \subseteq \mathcal{G}$, the vectors $w_G$ associated with the groups $G \in \mathcal{G}'$ will be exactly equal to zero, leading to a set of zeros which is the union of these groups, $\bigcup_{G \in \mathcal{G}'} G$. Thus, the set of allowed zero patterns should be the *union-closure* of $\mathcal{G}$, that is, (see Figure 1 for an example):

$$\mathcal{Z} = \Big\{ \bigcup_{G \in \mathcal{G}'} G;\ \mathcal{G}' \subseteq \mathcal{G} \Big\}.$$

The situation is however slightly more subtle as some zeros can be created by chance (just as regularizing by the $\ell_2$-norm may lead, though it is unlikely, to some zeros). Nevertheless, Theorem 2 shows that, under mild conditions, the previous intuition about the set of zero patterns is correct. Note that instead of considering the set of zero patterns $\mathcal{Z}$, it is also convenient to manipulate nonzero patterns, and we define

$$\mathcal{P} = \Big\{ \bigcap_{G \in \mathcal{G}'} G^c;\ \mathcal{G}' \subseteq \mathcal{G} \Big\} = \big\{ Z^c;\ Z \in \mathcal{Z} \big\}.$$

We can equivalently use $\mathcal{P}$ or $\mathcal{Z}$ by taking the complement of each element of these sets.

The following two results characterize the solutions of Problem (2). We first gives sufficient conditions under which this problem has a unique solution. We then formally prove the aforementioned intuition about the zero patterns of the solutions of (2), namely they should belong to $\mathcal{Z}$. In the following two results (see proofs in Appendix A and Appendix B), we assume that $\ell : (y, y') \mapsto \ell(y, y')$ is nonnegative, twice continuously differentiable with positive second derivative with respect to the

second variable and non-vanishing mixed derivative, that is, for any $y,y'$ in $\mathbb{R}$, $\frac{\partial^2 \ell}{\partial y'^2}(y,y') > 0$ and $\frac{\partial^2 \ell}{\partial y \partial y'}(y,y') \neq 0$.

**Proposition 1** *Let $Q$ denote the Gram matrix $\frac{1}{n}\sum_{i=1}^n x_i x_i^\top$. We consider the optimization problem in Equation (2) with $\mu > 0$. If $Q$ is invertible or if $\{1,\dots,p\}$ belongs to $\mathcal{G}$, then the problem in Equation (2) admits a unique solution.*

Note that the invertibility of the matrix $Q$ requires $p \leq n$. For high-dimensional settings, the uniqueness of the solution will hold when $\{1,\dots,p\}$ belongs to $\mathcal{G}$, or as further discussed at the end of the proof, as soon as for any $j,k \in \{1,\dots,p\}$, there exists a group $G \in \mathcal{G}$ which contains both $j$ and $k$. Adding the group $\{1,\dots,p\}$ to $\mathcal{G}$ will in general not modify $\mathcal{P}$ (and $\mathcal{Z}$), but it will cause $\mathcal{G}$ to lose its minimality (in a sense introduced in the next subsection). Furthermore, adding the full group $\{1,\dots,p\}$ has to be put in parallel with the equivalent (up to the squaring) $\ell_2$-norm term in the elastic-net penalty (Zou and Hastie, 2005), whose effect is to notably ensure strong convexity. For more sophisticated uniqueness conditions that we have not explored here, we refer the readers to Osborne et al. (2000, Theorem 1, 4 and 5), Rosset et al. (2004, Theorem 5) or Dossal (2007, Theorem 3) in the Lasso case, and Roth and Fischer (2008) for the group Lasso setting. We now turn to the result about the zero patterns of the solution of the problem in Equation (2):

**Theorem 2** *Assume that $Y = (y_1,\dots,y_n)^\top$ is a realization of an absolutely continuous probability distribution. Let $k$ be the maximal number such that any $k$ rows of the matrix $(x_1,\dots,x_n) \in \mathbb{R}^{p\times n}$ are linearly independent. For $\mu > 0$, any solution of the problem in Equation (2) with at most $k-1$ nonzero coefficients has a zero pattern in $\mathcal{Z} = \left\{\bigcup_{G\in\mathcal{G}'} G;\ \mathcal{G}' \subseteq \mathcal{G}\right\}$ almost surely.*

In other words, when $Y = (y_1,\dots,y_n)^\top$ is a realization of an absolutely continuous probability distribution, the sparse solutions have a zero pattern in $\mathcal{Z} = \left\{\bigcup_{G\in\mathcal{G}'} G;\ \mathcal{G}' \subseteq \mathcal{G}\right\}$ almost surely. As a corollary of our two results, if the Gram matrix $Q$ is invertible, the problem in Equation (2) has a unique solution, whose zero pattern belongs to $\mathcal{Z}$ almost surely. Note that with the assumption made on $Y$, Theorem 2 is not directly applicable to the classification setting. Based on these previous results, we can look at the following usual special cases from Section 2 (we give more examples in Section 3.5):

- **$\ell_2$-norm**: the set of allowed nonzero patterns is composed of the empty set and the full set $\{1,\dots,p\}$.
- **$\ell_1$-norm**: $\mathcal{P}$ is the set of all possible subsets.
- **$\ell_2$-norm and $\ell_1$-norm**: $\mathcal{P}$ is also the set of all possible subsets.
- **Group $\ell_1$-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, 2008a).

Two natural questions now 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?

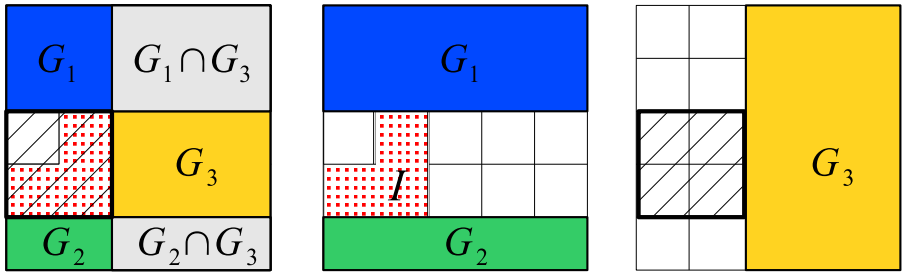



Figure 2: $\mathcal{G}$-adapted hull: the pattern of variables $I$ (left and middle, red dotted surface) and its hull (left and right, hatched square) that is defined by the complement of the union of groups that do not intersect $I$, that is, $(G_1 \cup G_2 \cup G_3)^c$.

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

We now study the different properties of the set of groups $\mathcal{G}$ and its corresponding sets of patterns $\mathcal{Z}$ and $\mathcal{P}$.

#### 3.2.1 CLOSEDNESS

The set of zero patterns $\mathcal{Z}$ (respectively, the set of nonzero patterns $\mathcal{P}$) is closed under union (respectively, intersection), that is, for all $K \in \mathbb{N}$ and all $z_1, \ldots, z_K \in \mathcal{Z}$, $\bigcup_{k=1}^K z_k \in \mathcal{Z}$ (respectively, $p_1, \ldots, p_K \in \mathcal{P}$, $\bigcap_{k=1}^K p_k \in \mathcal{P}$). 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. For instance, if we consider a sequence (see Figure 4), we cannot take $\mathcal{P}$ to be the set of contiguous patterns with length two, since the intersection of such two patterns may result in a singleton (that does not belong to $\mathcal{P}$).

#### 3.2.2 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 3.3. 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 uniqueness 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, that is, $|\mathcal{Z}| = O(p^2)$, that can be generated by a minimal set $\mathcal{G}$ whose size is $|\mathcal{G}| = O(\sqrt{p})$.

#### 3.2.3 HULL

Given a set of groups $\mathcal{G}$, we can define for any subset $I \subseteq \{1, \ldots, p\}$ the *$\mathcal{G}$-adapted hull*, or simply *hull*, as:

$$\text{Hull}(I) = \Big\{ \bigcup_{G \in \mathcal{G},\ G \cap I = \varnothing} G \Big\}^c,$$

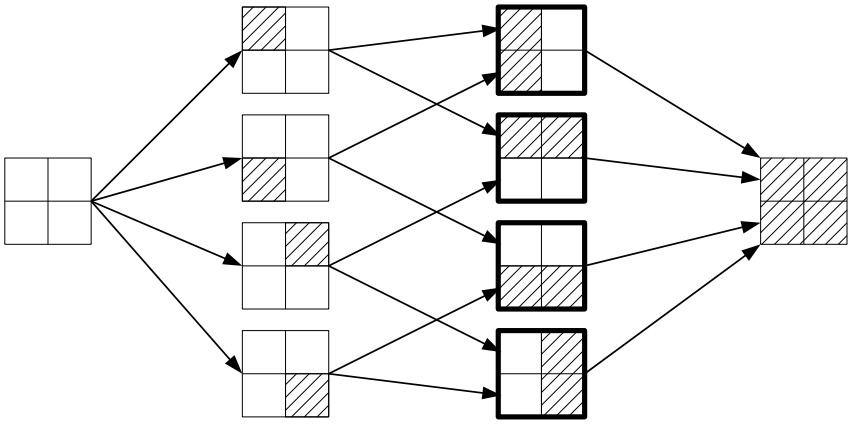


Figure 3: The DAG for the set $\mathcal{Z}$ associated with the $2\times 2$-grid. The members of $\mathcal{Z}$ are the complement of the areas hatched in black. The elements of $\mathcal{G}$ (i.e., the atoms of $\mathcal{Z}$) are highlighted by bold edges.

which is the smallest set in $\mathcal{P}$ containing $I$ (see Figure 2); we always have $I \subseteq \text{Hull}(I)$ with equality if and only if $I \in \mathcal{P}$. The hull has a clear geometrical interpretation for specific sets $\mathcal{G}$ of groups. For instance, 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 5), the hull of a subset $I \subset \{1,\dots,p\}$ is simply the axis-aligned bounding box of $I$. Similarly, when $\mathcal{G}$ is the set of all half-spaces with all possible orientations (e.g., orientations $\pm\pi/4$ are shown in Figure 6), the hull becomes the regular convex hull.[1] Note that those interpretations of the hull are possible and valid only when we have geometrical information at hand about the set of variables.

### 3.2.4 GRAPHS OF PATTERNS

We consider the directed acyclic graph (DAG) stemming from the *Hasse diagram* (Cameron, 1994) of the partially ordered set (poset) $(\mathcal{G},\supset)$. By definition, 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 3 for examples). 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, 2008a; Kim and Xing, 2010; Jenatton et al., 2010a; Schmidt and Murphy, 2010) have also used a similar DAG structure, with the slight difference that in these works, the corresponding hierarchy of variables is built from the prior knowledge about the problem at hand (e.g., the tree of wavelets in Zhao et al., 2009, the decomposition of kernels in Bach, 2008a or the hierarchy of genes in Kim and Xing, 2010). 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).

1. We use the term *convex* informally here. It can however be made precise with the notion of convex subgraphs (Chung, 1997).

### 3.3 From Patterns to Groups

We now assume that we want to impose a priori knowledge on the sparsity structure of a solution $\hat{w}$ of our regularized problem in Equation (2). This information can be exploited by restricting the patterns allowed by the norm $\Omega$. 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 1. 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 the minimality of the procedure can be found in Appendix C). The complexity of Algorithm 1 is $O(p|\mathcal{Z}|^2)$. 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 $\ell_1$-norm). In Section 3.5, we give other examples of interest where $|\mathcal{G}|$ (and $|\mathcal{P}|$) is also polynomial in $p$.

**Algorithm 1** 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 $G \in \mathcal{Z}$ such that $|G| = t$ **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**

### 3.4 From Groups to Patterns

The *forward* procedure presented in Algorithm 2, 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 sub-families 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 $\mathcal{G}$ generates $\mathcal{Z}$. 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, $|\mathcal{Z}| = 2^p$ and $|\mathcal{G}| = p$ for the $\ell_1$-norm).

**Algorithm 2** Forward procedure

**Input:** Set of groups $\mathcal{G} = \{G_1, \ldots, G_M\}$.
**Output:** Collection of zero patterns $\mathcal{Z}$ and nonzero patterns $\mathcal{P}$.
**Initialization:** $\mathcal{Z} = \{\varnothing\}$.
**for** $m = 1$ **to** $M$ **do**
  $C = \{\varnothing\}$
  **for each** $Z \in \mathcal{Z}$ **do**
    **if** $(G_m \not\subseteq Z)$ **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**
$\mathcal{P} = \{Z^c;\ Z \in \mathcal{Z}\}$.

### 3.5 Examples

We now present several examples of sets of groups $\mathcal{G}$, especially suited to encode geometric and temporal prior information.

#### 3.5.1 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 $[1,k]_{k\in\{1,\ldots,p-1\}}$ and $[k,p]_{k\in\{2,\ldots,p\}}$, with both $|\mathcal{Z}| = O(p^2)$ and $|\mathcal{G}| = O(p)$ (see Figure 4). Imposing the contiguity of the nonzero patterns is for instance relevant for the diagnosis of tumors, based on the profiles of arrayCGH (Rapaport et al., 2008).

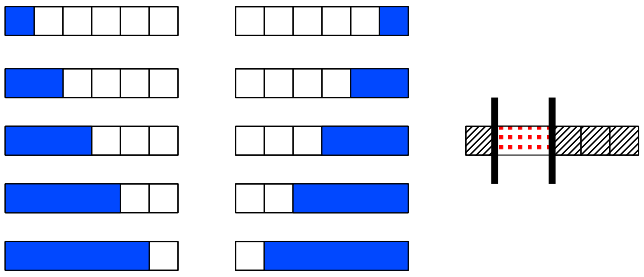


Figure 4: (Left and middle) The set of groups (blue areas) to penalize in order to select contiguous patterns in a sequence. (Right) An example of such a nonzero pattern (red dotted area) with its corresponding zero pattern (hatched area).

#### 3.5.2 TWO-DIMENSIONAL GRIDS

In Section 6, we notably 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 5), 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 (Harzallah et al., 2009).

Larger set of convex patterns can be obtained by adding in $\mathcal{G}$ half-planes with other orientations than vertical and horizontal. For instance, if we use planes with angles that are multiples of $\pi/4$, the nonzero patterns of $\mathcal{P}$ can have polygonal shapes with up to 8 faces. In this sense, if we keep on adding half-planes with finer orientations, the nonzero patterns of $\mathcal{P}$ can be described by polygonal shapes with an increasingly larger number of faces. The standard notion of convexity defined in $\mathbb{R}^2$ would correspond to the situation where an infinite number of orientations is considered (Soille, 2003). See Figure 6. 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). Imposing such convex-like regions turns out to be useful in computer vision. For instance, in face recognition, it enables the design of localized features that improve upon the robustness to occlusions (Jenatton et al., 2010b). In the same vein, regularizations with similar two-dimensional sets of groups have led to good performances in background subtraction tasks (Mairal et al., 2010b), where the pixel spatial information is crucial to avoid scattered results. Another application worth being mentioned is the design of topographic dictionaries in the context of image processing (Kavukcuoglu et al., 2009; Mairal et al., 2011). In this case, dictionaries self-organize and adapt to the underlying geometrical structure encoded by the two-dimensional set of groups.

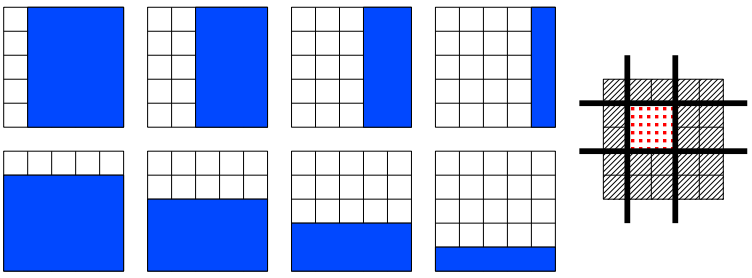


Figure 5: Vertical and horizontal groups: (Left) the set of groups (blue areas) with their (not displayed) complements to penalize in order to select rectangles. (Right) An example of nonzero pattern (red dotted area) recovered in this setting, with its corresponding zero pattern (hatched area).

### 3.5.3 EXTENSIONS

The sets of groups presented above can be straightforwardly extended to more complicated topologies, such as three-dimensional spaces discretized in cubes or spherical volumes discretized in slices. Similar properties hold for such settings. For instance, if all the axis-aligned half-spaces are considered for $\mathcal{G}$ in a three-dimensional space, then $\mathcal{P}$ is the set of all possible rectangular boxes with $|\mathcal{P}| = O(p^2)$ and $|\mathcal{G}| = O(p^{1/3})$. Such three-dimensional structures are interesting to retrieve discriminative and local sets of voxels from fMRI/MEEG responses. In particular, they have recently proven useful for modelling brain resting-state activity (Varoquaux et al., 2010). Moreover, while the two-dimensional rectangular patterns described previously are adapted to find bounding boxes in static images (Harzallah et al., 2009), scene recognition in videos requires to deal with a third temporal dimension (Dalal et al., 2006). This may be achieved by designing appropriate sets of

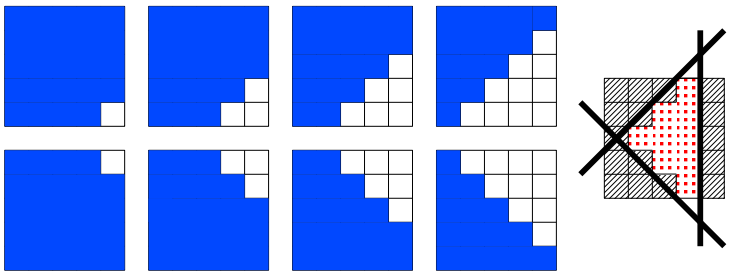


Figure 6: Groups with $\pm\pi/4$ orientations: (Left) the set of groups (blue areas) with their (not displayed) complements to penalize in order to select diamond-shaped patterns. (Right) An example of nonzero pattern (red dotted area) recovered in this setting, with its corresponding zero pattern (hatched area).

groups, embedded in the three-dimensional space obtained by tracking the frames over time. Finally, in the context of matrix-based optimization problems, for example, multi-task learning and dictionary learning, sets of groups $\mathcal{G}$ can also be designed to encode *structural constraints* the solutions must respect. This notably encompasses banded structures (Levina et al., 2008) and *simultaneous* row/column sparsity for CUR matrix factorization (Mairal et al., 2011).

### 3.5.4 REPRESENTATION AND COMPUTATION OF $\mathcal{G}$

The sets of groups described so far can actually be represented in a same form, that lends itself well to the analysis of the next section. When dealing with a discrete sequence of length $l$ (see Figure 4), we have

$$\begin{aligned}\mathcal{G} &= \{g_-^k;\, k\in\{1,\dots,l-1\}\}\cup\{g_+^k;\, k\in\{2,\dots,l\}\},\\ &= \mathcal{G}_{\text{left}}\cup\mathcal{G}_{\text{right}},\end{aligned}$$

with $g_-^k=\{i;\, 1\le i\le k\}$ and $g_+^k=\{i;\, l\ge i\ge k\}$. In other words, the set of groups $\mathcal{G}$ can be rewritten as a partition[2] in two sets of nested groups, $\mathcal{G}_{\text{left}}$ and $\mathcal{G}_{\text{right}}$.

The same goes for a two-dimensional grid, with dimensions $h\times l$ (see Figure 5 and Figure 6). In this case, the nested groups we consider are defined based on the following groups of variables

$$g^{k,\theta}=\{(i,j)\in\{1,\dots,l\}\times\{1,\dots,h\};\ \cos(\theta)i+\sin(\theta)j\le k\},$$

where $k\in\mathbb{Z}$ is taken in an appropriate range. The nested groups we obtain in this way are therefore parameterized by an angle[3] $\theta$, $\theta\in(-\pi;\pi]$. We refer to this angle as an *orientation*, since it defines the normal vector $\binom{\cos(\theta)}{\sin(\theta)}$ to the line $\{(i,j)\in\mathbb{R}^2;\cos(\theta)i+\sin(\theta)j=k\}$. In the example of the rectangular groups (see Figure 5), we have four orientations, with $\theta\in\{0,\pi/2,-\pi/2,\pi\}$. More generally, if we denote by $\Theta$ the set of the orientations, we have

$$\mathcal{G}=\bigcup_{\theta\in\Theta}\mathcal{G}_\theta,$$

2. Note the subtlety: the sets $\mathcal{G}_\theta$ are disjoint, that is $\mathcal{G}_\theta\cap\mathcal{G}_{\theta'}=\varnothing$ for $\theta\neq\theta'$, but groups in $\mathcal{G}_\theta$ and $\mathcal{G}_{\theta'}$ can overlap.
3. Due to the discrete nature of the underlying geometric structure of $\mathcal{G}$, angles $\theta$ that are not multiple of $\pi/4$ (i.e., such that $\tan(\theta)\notin\mathbb{Z}$) are dealt with by rounding operations.

where $\theta \in \Theta$ indexes the partition of $\mathcal{G}$ in sets $\mathcal{G}_\theta$ of nested groups of variables. Although we have not detailed the case of $\mathbb{R}^3$, we likewise end up with a similar partition of $\mathcal{G}$.

## 4. Optimization and Active Set Algorithm

For moderate values of $p$, one may obtain a solution for Problem (2) using generic toolboxes for second-order cone programming (SOCP) whose time complexity is equal to $O(p^{3.5}+|\mathcal{G}|^{3.5})$ (Boyd and Vandenberghe, 2004), which is not appropriate when $p$ or $|\mathcal{G}|$ are large. This time complexity corresponds to the computation of a solution of Problem (2) for a single value of the regularization parameter $\mu$.

We present in this section an *active set algorithm* (Algorithm 3) that finds a solution for Problem (2) by considering increasingly larger active sets and checking global optimality at each step. When the rectangular groups are used, the total complexity of this method is in $O(s\max\{p^{1.75},s^{3.5}\})$, where $s$ is the size of the active set at the end of the optimization. Here, the sparsity prior is exploited for computational advantages. Our active set algorithm needs an underlying *black-box* SOCP solver; in this paper, we consider both a first order approach (see Appendix H) and a SOCP method[4]—in our experiments, we use `SDPT3` (Toh et al., 1999; Tütüncü et al., 2003). Our active set algorithm extends to general overlapping groups the work of Bach (2008a), by further assuming that it is computationally possible to have a time complexity polynomial in the number of variables $p$.

We primarily focus here on finding an efficient active set algorithm; we defer to future work the design of specific SOCP solvers, for example, based on proximal techniques (see, e.g., Nesterov, 2007; Beck and Teboulle, 2009; Combettes and Pesquet, 2010, and numerous references therein), adapted to such non-smooth sparsity-inducing penalties.

### 4.1 Optimality Conditions: From Reduced Problems to Full Problems

It is simpler to derive the algorithm for the following regularized optimization problem[5] which has the same solution set as the regularized problem of Equation (2) when $\mu$ and $\lambda$ are allowed to vary (Borwein and Lewis, 2006, see Section 3.2):

$$\min_{w\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \ell\big(y_i, w^\top x_i\big) + \frac{\lambda}{2}\left[\Omega(w)\right]^2. \tag{3}$$

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 Equation (3). In the subsequent analysis, we will use arguments based on duality to monitor the optimality of our active set algorithm. We denote by $L(w)=\frac{1}{n}\sum_{i=1}^{n}\ell\big(y_i,w^\top x_i\big)$ the empirical risk (which is by assumption convex and continuously differentiable) and by $L^*$ its *Fenchel-conjugate*, defined as (Boyd and Vandenberghe, 2004; Borwein and Lewis, 2006):

$$L^*(u) = \sup_{w\in\mathbb{R}^p}\{w^\top u - L(w)\}.$$

4. The C++/Matlab code used in the experiments may be downloaded from the authors website.
5. It is also possible to derive the active set algorithm for the constrained formulation $\min_{w\in\mathbb{R}^p}\frac{1}{n}\sum_{i=1}^{n}\ell\big(y_i,w^\top x_i\big)$ such that $\Omega(w)\leq\lambda$. However, we empirically found it more difficult to select $\lambda$ in this latter formulation.

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 $L_J^*$. Note that, as opposed to $L$, we do not have in general $L_J^*(\kappa_J) = L^*(\tilde{\kappa})$ for $\tilde{\kappa}_J = \kappa_J$ and $\tilde{\kappa}_{J^c} = 0$.

For a potential active set $J \subset \{1,\dots,p\}$ which belongs to the set of allowed nonzero patterns $\mathcal{P}$, we denote by $\mathcal{G}_J$ the set of active groups, that is, the set of groups $G \in \mathcal{G}$ such that $G \cap J \neq \varnothing$. We consider the reduced norm $\Omega_J$ defined on $\mathbb{R}^{|J|}$ as

$$\Omega_J(w_J) = \sum_{G\in\mathcal{G}} \|d_J^G \circ w_J\|_2 = \sum_{G\in\mathcal{G}_J} \|d_J^G \circ w_J\|_2 ,$$

and its *dual norm* $\Omega_J^*(\kappa_J) = \max_{\Omega_J(w_J)\leq 1} w_J^\top \kappa_J$, also defined on $\mathbb{R}^{|J|}$. The next proposition (see proof in Appendix D) gives the optimization problem dual to the reduced problem (Equation (4) below):

**Proposition 3 (Dual Problems)** *Let $J \subseteq \{1,\dots,p\}$. The following two problems*

$$\min_{w_J\in\mathbb{R}^{|J|}} L_J(w_J) + \frac{\lambda}{2}\left[\Omega_J(w_J)\right]^2, \tag{4}$$

$$\max_{\kappa_J\in\mathbb{R}^{|J|}} -L_J^*(-\kappa_J) - \frac{1}{2\lambda}\left[\Omega_J^*(\kappa_J)\right]^2,$$

*are dual to each other and strong duality holds. The pair of primal-dual variables $\{w_J, \kappa_J\}$ is optimal if and only if we have*

$$\begin{cases} \kappa_J & = -\nabla L_J(w_J), \\ w_J^\top \kappa_J & = \frac{1}{\lambda}\left[\Omega_J^*(\kappa_J)\right]^2 = \lambda\left[\Omega_J(w_J)\right]^2. \end{cases}$$

As a brief reminder, the duality gap of a minimization problem is defined as the difference between the primal and dual objective functions, evaluated for a feasible pair of primal/dual variables (Boyd and Vandenberghe, 2004, see Section 5.5). This gap serves as a certificate of (sub)optimality: if it is equal to zero, then the optimum is reached, and provided that strong duality holds, the converse is true as well (Boyd and Vandenberghe, 2004, see Section 5.5).

The previous proposition enables us to derive the duality gap for the optimization Problem (4), that is reduced to the active set of variables $J$. In practice, this duality gap will always vanish (up to the precision of the underlying SOCP solver), since we will sequentially solve Problem (4) for increasingly larger active sets $J$. We now study how, starting from the optimality of the problem in Equation (4), we can control the optimality, or equivalently the duality gap, for the full problem in Equation (3). More precisely, the duality gap of the optimization problem in Equation (4) is

$$\begin{aligned} & L_J(w_J) + L_J^*(-\kappa_J) + \frac{\lambda}{2}\left[\Omega_J(w_J)\right]^2 + \frac{1}{2\lambda}\left[\Omega_J^*(\kappa_J)\right]^2 \\ = \; & \left\{L_J(w_J) + L_J^*(-\kappa_J) + w_J^\top \kappa_J\right\} + \left\{\frac{\lambda}{2}\left[\Omega_J(w_J)\right]^2 + \frac{1}{2\lambda}\left[\Omega_J^*(\kappa_J)\right]^2 - w_J^\top \kappa_J\right\}, \end{aligned}$$

which is a sum of two nonnegative terms, the nonnegativity coming from the Fenchel-Young inequality (Borwein and Lewis, 2006; Boyd and Vandenberghe, 2004, Proposition 3.3.4 and Section 3.3.2 respectively). We can think of this duality gap as the sum of two duality gaps, respectively

relative to $L_J$ and $\Omega_J$. Thus, if we have a primal candidate $w_J$ and we choose $\kappa_J = -\nabla L_J(w_J)$, the duality gap relative to $L_J$ vanishes and the total duality gap then reduces to

$$\frac{\lambda}{2}\left[\Omega_J(w_J)\right]^2 + \frac{1}{2\lambda}\left[\Omega_J^*(\kappa_J)\right]^2 - w_J^\top \kappa_J.$$

In order to check that the reduced solution $w_J$ is optimal for the full problem in Equation (3), we pad $w_J$ with zeros on $J^c$ to define $w$ and compute $\kappa = -\nabla L(w)$, which is such that $\kappa_J = -\nabla L_J(w_J)$. For our given candidate pair of primal/dual variables $\{w, \kappa\}$, we then get a duality gap for the full problem in Equation (3) equal to

$$\begin{aligned}
& \frac{\lambda}{2}\left[\Omega(w)\right]^2 + \frac{1}{2\lambda}\left[\Omega^*(\kappa)\right]^2 - w^\top \kappa \\
= & \frac{\lambda}{2}\left[\Omega(w)\right]^2 + \frac{1}{2\lambda}\left[\Omega^*(\kappa)\right]^2 - w_J^\top \kappa_J \\
= & \frac{\lambda}{2}\left[\Omega(w)\right]^2 + \frac{1}{2\lambda}\left[\Omega^*(\kappa)\right]^2 - \frac{\lambda}{2}\left[\Omega_J(w_J)\right]^2 - \frac{1}{2\lambda}\left[\Omega_J^*(\kappa_J)\right]^2 \\
= & \frac{1}{2\lambda}\left(\left[\Omega^*(\kappa)\right]^2 - \left[\Omega_J^*(\kappa_J)\right]^2\right) \\
= & \frac{1}{2\lambda}\left(\left[\Omega^*(\kappa)\right]^2 - \lambda w_J^\top \kappa_J\right).
\end{aligned}$$

Computing this gap requires computing the dual norm which itself is as hard as the original problem, prompting the need for upper and lower bounds on $\Omega^*$ (see Propositions 4 and 5 for more details).

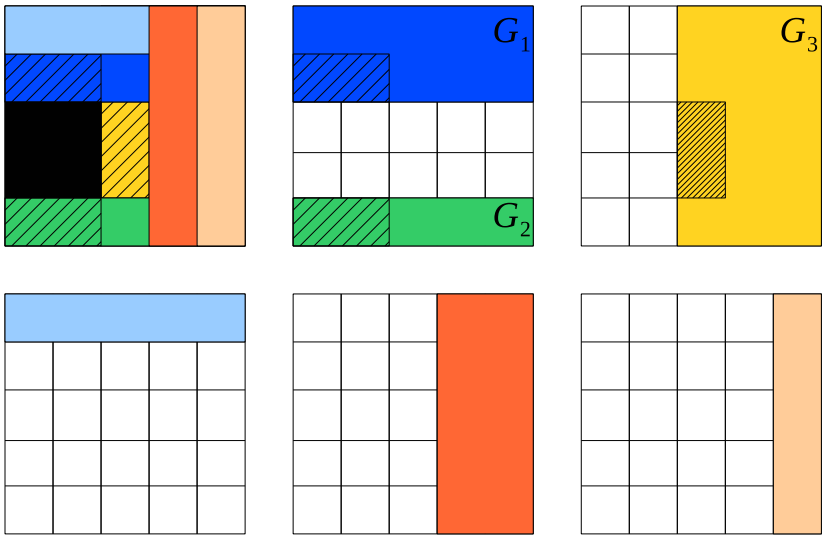



Figure 7: The active set (black) and the candidate patterns of variables, that is, the variables in $K \backslash J$ (hatched in black) that can become active. The fringe groups are exactly the groups that have the hatched areas (i.e., here we have $\mathcal{F}_J = \bigcup_{K \in \Pi_{\mathcal{P}}(J)} \mathcal{G}_K \backslash \mathcal{G}_J = \{G_1, G_2, G_3\}$).

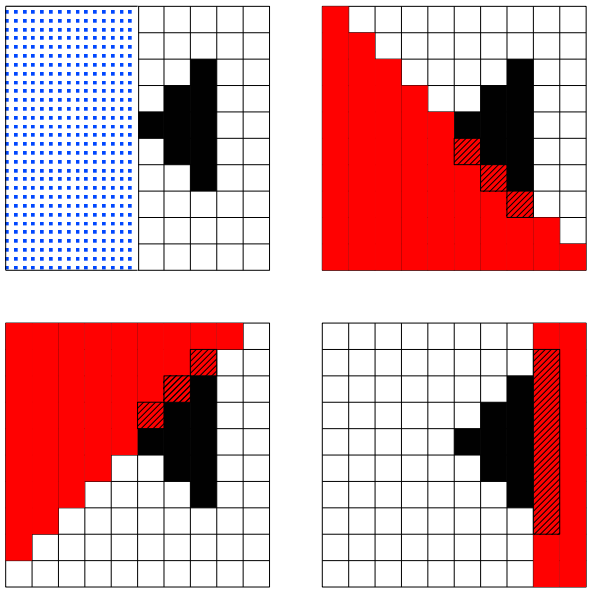


Figure 8: The active set (black) and the candidate patterns of variables, that is, the variables in $K \backslash J$ (hatched in black) that can become active. The groups in red are those in $\bigcup_{K \in \Pi_{\mathcal{P}}(J)} \mathcal{G}_K \backslash \mathcal{G}_J$, while the blue dotted group is in $\mathcal{F}_J \backslash (\bigcup_{K \in \Pi_{\mathcal{P}}(J)} \mathcal{G}_K \backslash \mathcal{G}_J)$. The blue dotted group does not intersect with any patterns in $\Pi_{\mathcal{P}}(J)$.

### 4.2 Active Set Algorithm

We can interpret the active set algorithm as a walk through the DAG of nonzero patterns allowed by the norm $\Omega$. The parents $\Pi_{\mathcal{P}}(J)$ of $J$ in this DAG 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 $\mathcal{F}_J = \{G \in (\mathcal{G}_J)^c \; ; \; \nexists G' \in (\mathcal{G}_J)^c, \; G \subseteq G'\}$, that is, the groups that are not contained by any other inactive groups.

In simple settings, for example, when $\mathcal{G}$ is the set of rectangular groups, the correspondence between groups and variables is straightforward since we have $\mathcal{F}_J = \bigcup_{K \in \Pi_{\mathcal{P}}(J)} \mathcal{G}_K \backslash \mathcal{G}_J$ (see Figure 7). However, 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 some elements of $\mathcal{F}_J$ might not correspond to any patterns of variables in $\Pi_{\mathcal{P}}(J)$ (see Figure 8).

We now present the optimality conditions (see proofs in Appendix E) that monitor the progress of Algorithm 3:

**Proposition 4 (Necessary condition)** *If $w$ is optimal for the full problem in Equation (3), then*

$$\max_{K \in \Pi_{\mathcal{P}}(J)} \frac{\left\| \nabla L(w)_{K \backslash J} \right\|_2}{\sum_{H \in \mathcal{G}_K \backslash \mathcal{G}_J} \left\| d^H_{K \backslash J} \right\|_\infty} \leq \left\{ -\lambda w^\top \nabla L(w) \right\}^{\frac{1}{2}}. \tag{N}$$

**Proposition 5 (Sufficient condition)** *If*

$$\max_{G \in \mathcal{F}_J} \left\{ \sum_{k \in G} \left\{ \frac{\nabla L(w)_k}{\sum_{H \ni k, \, H \in (\mathcal{G}_J)^c} d^H_k} \right\}^2 \right\}^{\frac{1}{2}} \leq \left\{ \lambda (2\varepsilon - w^\top \nabla L(w)) \right\}^{\frac{1}{2}}, \tag{$S_\varepsilon$}$$

*then $w$ is an approximate solution for Equation (3) whose duality gap 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 (up to the squaring of $\Omega$) to the condition $\|\nabla L(w)_{J^c}\|_\infty \leq -w^\top \nabla L(w)$, which is the usual optimality condition (Fuchs, 2005; Tibshirani, 1996; Wainwright, 2009). Moreover, when they are not satisfied, our two conditions provide good heuristics for choosing which $K \in \Pi_{\mathcal{P}}(J)$ should enter 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 pattern $K \in \Pi_{\mathcal{P}}(J)$ that violates most the condition.

The heuristics for the sufficient condition ($S_\varepsilon$) implies that, to go from groups to variables, we simply consider the group $G \in \mathcal{F}_J$ violating the sufficient condition the most and then take all the patterns of variables $K \in \Pi_{\mathcal{P}}(J)$ such that $K \cap G \neq \varnothing$ to enter the active set. If $G \cap (\bigcup_{K \in \Pi_{\mathcal{P}}(J)} K) = \varnothing$, we look at all the groups $H \in \mathcal{F}_J$ such that $H \cap G \neq \varnothing$ and apply the scheme described before (see Algorithm 4).

A direct consequence of this heuristics 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 worthwhile 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 $\lambda$,
Duality gap precision $\varepsilon$, maximum number of variables $s$.
**Output:** Active set $J$, loading vector $\hat{w}$.
**Initialization:** $J = \{\varnothing\}$, $\hat{w} = 0$.
**while** $\big($ $(N)$ is not satisfied $\big)$ **and** $\big($ $|J| \leq s$ $\big)$ **do**
  Replace $J$ by violating $K \in \Pi_{\mathcal{P}}(J)$ in $(N)$.
  Solve the reduced problem $\min_{w_J \in \mathbb{R}^{|J|}} L_J(w_J) + \frac{\lambda}{2}\left[\Omega_J(w_J)\right]^2$ to get $\hat{w}$.
**end while**
**while** $\big($ $(S_\varepsilon)$ is not satisfied $\big)$ **and** $\big($ $|J| \leq s$ $\big)$ **do**
  Update $J$ according to Algorithm 4.
  Solve the reduced problem $\min_{w_J \in \mathbb{R}^{|J|}} L_J(w_J) + \frac{\lambda}{2}\left[\Omega_J(w_J)\right]^2$ to get $\hat{w}$.
**end while**

### 4.2.1 CONVERGENCE OF THE ACTIVE SET ALGORITHM

The procedure described in Algorithm 3 can terminate in two different states. If the procedure stops because of the limit on the number of active variables $s$, the solution might be suboptimal. Note that, in any case, we have at our disposal a upper-bound on the duality gap.

Otherwise, the procedure always converges to an optimal solution, either (1) by validating both the necessary and sufficient conditions (see Propositions 4 and 5), ending up with fewer than $p$ active variables and a precision of (at least) $\varepsilon$, or (2) by running until the $p$ variables become active, the precision of the solution being given by the underlying solver.

**Algorithm 4** Heuristics for the sufficient condition ($S_\varepsilon$)

Let $G \in \mathcal{F}_J$ be the group that violates ($S_\varepsilon$) most.
**if** $(G \cap (\bigcup_{K \in \Pi_{\mathcal{P}}(J)} K) \neq \varnothing)$ **then**
  **for** $K \in \Pi_{\mathcal{P}}(J)$ such that $K \cap G \neq \varnothing$ **do**
    $J \leftarrow J \cup K$.
  **end for**
**else**
  **for** $H \in \mathcal{F}_J$ such that $H \cap G \neq \varnothing$ **do**
    **for** $K \in \Pi_{\mathcal{P}}(J)$ such that $K \cap H \neq \varnothing$ **do**
      $J \leftarrow J \cup K$.
    **end for**
  **end for**
**end if**

#### 4.2.2 ALGORITHMIC COMPLEXITY

We analyze in detail the time complexity of the active set algorithm when we consider sets of groups $\mathcal{G}$ such as those presented in the examples of Section 3.5. We recall that we denote by $\Theta$ the set of orientations in $\mathcal{G}$ (for more details, see Section 3.5). For such choices of $\mathcal{G}$, the fringe groups $\mathcal{F}_J$ reduces to the largest groups of each orientation and therefore $|\mathcal{F}_J| \leq |\Theta|$. We further assume that the groups in $\mathcal{G}_\theta$ are sorted by cardinality, so that computing $\mathcal{F}_J$ costs $O(|\Theta|)$.

Given an active set $J$, both the necessary and sufficient conditions require to have access to the direct parents $\Pi_{\mathcal{P}}(J)$ of $J$ in the DAG of nonzero patterns. In simple settings, for example, when $\mathcal{G}$ is the set of rectangular groups, this operation can be performed in $O(1)$ (it just corresponds to scan the (up to) four patterns at the edges of the current rectangular hull).

However, for more general orientations, computing $\Pi_{\mathcal{P}}(J)$ requires to find the smallest nonzero patterns that we can generate from the groups in $\mathcal{F}_J$, reduced to the stripe of variables around the current hull. This stripe of variables can be computed as $\big[\bigcup_{G \in (\mathcal{G}_J)^c \setminus \mathcal{F}_J} G\big]^c \setminus J$, so that getting $\Pi_{\mathcal{P}}(J)$ costs $O(s2^{|\Theta|} + p|\mathcal{G}|)$ in total.

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, $O(snp)$ for the square loss.
- if the underlying solver called upon by the active set algorithm is a standard SOCP solver, $O(s \max_{J \in \mathcal{P}, |J| \leq s} |\mathcal{G}_J|^{3.5} + s^{4.5})$ (note that the term $s^{4.5}$ could be improved upon by using warm-restart strategies for the sequence of reduced problems).
- $t_1$ times the computation of $(N)$, that is $O(t_1(s2^{|\Theta|} + p|\mathcal{G}| + sn_\theta^2) + p|\mathcal{G}|) = O(t_1 p|\mathcal{G}|)$.

  During the initialization (i.e., $J = \varnothing$), we have $|\Pi_{\mathcal{P}}(\varnothing)| = O(p)$ (since we can start with any singletons), and $|\mathcal{G}_K \setminus \mathcal{G}_J| = |\mathcal{G}_K| = |\mathcal{G}|$, which leads to a complexity of $O(p|\mathcal{G}|)$ for the sum $\sum_{G \in \mathcal{G}_K \setminus \mathcal{G}_J} = \sum_{G \in \mathcal{G}_K}$. 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 $(S_\varepsilon)$, that is $O(t_2(s2^{|\Theta|} + p|\mathcal{G}| + |\Theta|^2 + |\Theta|p + p|\mathcal{G}|)) = O(t_2 p|\mathcal{G}|)$, with $t_1 + t_2 \leq s$.

We finally get complexity with a leading term in $O(sp|\mathcal{G}| + s\max_{J\in\mathcal{P},|J|\leq s}|\mathcal{G}_J|^{3.5} + s^{4.5})$, which is much better than $O(p^{3.5} + |\mathcal{G}|^{3.5})$, without an active set method. In the example of the two-dimensional grid (see Section 3.5), we have $|\mathcal{G}| = O(\sqrt{p})$ and $O(s\max\{p^{1.75}, s^{3.5}\})$ as total complexity. The simulations of Section 6 confirm that the active set strategy is indeed useful when $s$ is much smaller than $p$. Moreover, the two extreme cases where $s \approx p$ or $p \ll 1$ are also shown not to be advantageous for the active set strategy, since either it is cheaper to use the SOCP solver directly on the $p$ variables, or we uselessly pay the additional fixed-cost of the active set machinery (such as computing the optimality conditions). Note that we have derived here the *theoretical* complexity of the active set algorithm when we use an interior point method as underlying solver. With the first order method presented in Appendix H, we would instead get a total complexity in $O(sp^{1.5})$.

### 4.3 Intersecting Nonzero Patterns

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 Section 6). In particular, the weights have to take into account that some variables belonging to several overlapping groups are penalized multiple times.

However, it is possible to keep the benefit of overlapping groups whilst limiting their side effects, by taking up the idea of support intersection (Bach, 2008c; Meinshausen and Bühlmann, 2010). First introduced to stabilize the set of variables recovered by the Lasso, we reuse this technique in a different context, based on the fact that $\mathcal{Z}$ is closed under union.

If we deal with the same sets of groups as those considered in Section 3.5, it is natural to rewrite $\mathcal{G}$ as $\bigcup_{\theta\in\Theta}\mathcal{G}_\theta$, where $\Theta$ is the set of the orientations of the groups in $\mathcal{G}$ (for more details, see Section 3.5). Let us denote by $\hat{w}$ and $\hat{w}^\theta$ the solutions of Problem (3), where the regularization term $\Omega$ is respectively defined by the groups in $\mathcal{G}$ and by the groups[6] in $\mathcal{G}_\theta$.

The main point is that, since $\mathcal{P}$ is closed under intersection, the two procedures described below actually lead to the same set of allowed nonzero patterns:

a) Simply considering the nonzero pattern of $\hat{w}$.

b) Taking the *intersection* of the nonzero patterns obtained for each $\hat{w}^\theta$, $\theta$ in $\Theta$.

With the latter procedure, although the learning of several models $\hat{w}^\theta$ is required (a number of times equals to the number of orientations considered, for example, 2 for the sequence, 4 for the rectangular groups and more generally $|\Theta|$ times), each of those learning tasks involves a smaller number of groups (that is, just the ones belonging to $\mathcal{G}_\theta$). In addition, this procedure is a *variable selection* technique that therefore needs a second step for estimating the loadings (restricted to the selected nonzero pattern). In the experiments, we follow Bach (2008c) and we use an ordinary least squares (OLS). The simulations of Section 6 will show the benefits of this variable selection approach.

6. To be more precise, in order to regularize every variable, we add the full group $\{1,\ldots,p\}$ to $\mathcal{G}_\theta$, which does not modify $\mathcal{P}$.

## 5. Pattern Consistency

In this section, we analyze the model consistency of the solution of the problem in Equation (2) 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 2 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., $x_1, \ldots, x_n$ are fixed). The extension of the following consistency results to other loss functions is beyond the scope of the paper (see for instance Bach, 2009). We assume that for all $i \in \{1, \ldots, n\}$, $y_i = \mathbf{w}^\top x_i + \varepsilon_i$ where the vector $\varepsilon$ 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 estimating the $\mathcal{G}$-adapted hull of $\mathbf{w}$ is equivalent to estimating the nonzero pattern of $\mathbf{w}$ if and only if this nonzero pattern belongs to $\mathcal{P}$. This happens when our prior information has led us to consider an appropriate set of groups $\mathcal{G}$. Conversely, if $\mathcal{G}$ is misspecified, recovering the hull of the nonzero pattern of $\mathbf{w}$ may be irrelevant, which is for instance the case if $\mathbf{w} = \binom{\mathbf{w}_1}{0} \in \mathbb{R}^2$ and $\mathcal{G} = \{\{1\}, \{1,2\}\}$. Finding the appropriate structure of $\mathcal{G}$ *directly from the data* would therefore be interesting future work.

### 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 positive-definite (i.e., invertible) limit: $\lim_{n\to\infty} Q = \mathbf{Q} \succ 0$. In this setting, the noise is the only source of randomness. We denote by $\mathbf{r}_\mathbf{J}$ the vector defined as

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

In the Lasso and group Lasso setting, the vector $\mathbf{r}_\mathbf{J}$ is respectively the sign vector $\mathrm{sign}(\mathbf{w}_\mathbf{J})$ and the vector defined by the blocks $(\frac{\mathbf{w}_G}{\|\mathbf{w}_G\|_2})_{G \in \mathcal{G}_\mathbf{J}}$.

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

The following Theorem gives 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 F).

**Theorem 6 (Consistency condition)** *Assume $\mu \to 0$, $\mu\sqrt{n} \to \infty$ in Equation (2). If the hull is consistently estimated, then $(\Omega^c_\mathbf{J})^*[\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{J}\mathbf{J}}^{-1}\mathbf{r}_\mathbf{J}] \leq 1$. Conversely, if $(\Omega^c_\mathbf{J})^*[\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{J}\mathbf{J}}^{-1}\mathbf{r}_\mathbf{J}] < 1$, then the hull is consistently estimated, that is,*

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

The two previous propositions bring into play the dual norm $(\Omega^c_\mathbf{J})^*$ that we cannot compute in closed form, but requires to solve an optimization problem as complex as the initial problem in

Equation (3). However, we can prove bounds similar to those obtained in Propositions 4 and 5 for the necessary and sufficient conditions.

#### 5.1.1 COMPARISON WITH THE LASSO AND GROUP LASSO

For the $\ell_1$-norm, our two bounds lead to the usual consistency conditions for the Lasso, that is, the quantity $\|\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{J}\mathbf{J}}^{-1}\operatorname{sign}(\mathbf{w}_\mathbf{J})\|_\infty$ must be less or strictly less than one. Similarly, when $\mathcal{G}$ defines a partition of $\{1,\dots,p\}$ and if all the weights equal one, our two bounds lead in turn to the consistency conditions for the group Lasso, that is, the quantity $\max_{G\in(\mathcal{G}_\mathbf{J})^c}\|\mathbf{Q}_{G\,\mathrm{Hull}(\mathbf{J})}\mathbf{Q}^{-1}_{\mathrm{Hull}(\mathbf{J})\mathrm{Hull}(\mathbf{J})}(\frac{\mathbf{w}_G}{\|\mathbf{w}_G\|_2})_{G\in\mathcal{G}_\mathbf{J}}\|_2$ must be less or strictly less than one.

### 5.2 High-Dimensional Analysis

We prove a high-dimensional variable consistency result (see proof in Appendix G) that extends the corresponding result for the Lasso (Zhao and Yu, 2006; Wainwright, 2009), by assuming that the consistency condition in Theorem 6 is satisfied.

**Theorem 7** *Assume that $Q$ has unit diagonal, $\kappa = \lambda_{\min}(Q_{\mathbf{J}\mathbf{J}}) > 0$ and $(\Omega^c_\mathbf{J})^*[Q_{\mathbf{J}^c\mathbf{J}}Q_{\mathbf{J}\mathbf{J}}^{-1}\mathbf{r}_\mathbf{J}] < 1-\tau$, with $\tau > 0$. If $\tau\mu\sqrt{n} \geqslant \sigma C_3(\mathcal{G},\mathbf{J})$, and $\mu|\mathbf{J}|^{1/2} \leqslant C_4(\mathcal{G},\mathbf{J})$, 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 G, 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 $n \approx \log p$ and $\mu \approx \sigma(\log p/n)^{1/2}$.

We can also give the scaling of these constants in simple settings where groups overlap. For instance, let us consider that the variables are organized in a sequence (see Figure 4). Let us further assume that the weights $(d^G)_{G\in\mathcal{G}}$ satisfy the following two properties:

a) The weights take into account the overlaps, that is,

$$d^G_j = \beta(|\{H\in\mathcal{G}\,;\, H\ni j,\ H\subset G \text{ and } H\neq G\}|),$$

with $t\mapsto\beta(t)\in(0,1]$ a non-increasing function such that $\beta(0)=1$,

b) The term

$$\max_{j\in\{1,\dots,p\}}\sum_{G\ni j, G\in\mathcal{G}} d^G_j$$

is upper bounded by a constant $\mathcal{K}$ independent of $p$.

Note that we consider such weights in the experiments (see Section 6). Based on these assumptions, some algebra directly leads to

$$\|u\|_1 \leqslant \Omega(u) \leqslant 2\mathcal{K}\|u\|_1, \text{ for all } u\in\mathbb{R}^p.$$

We thus obtain a scaling similar to the Lasso (with, *in addition*, a control of the allowed nonzero patterns). With stronger assumptions on the possible positions of $\mathbf{J}$, we may have better scalings, but these are problem-dependent (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 $\Omega$. We denote by *Structured-lasso*, or simply *Slasso*, the models regularized by the norm $\Omega$. In addition, the procedure (introduced in Section 4.3) that consists in intersecting the nonzero patterns obtained for different models of Slasso will be referred to as *Intersected Structured-lasso*, or simply *ISlasso*.

Throughout the experiments, we consider noisy linear models $Y = X\mathbf{w} + \varepsilon$, where $\mathbf{w} \in \mathbb{R}^p$ is the generating loading vector and $\varepsilon$ is a standard Gaussian noise vector with its variance set to satisfy $\|X\mathbf{w}\|_2 = 3\|\varepsilon\|_2$. This consequently leads to a fixed signal-to-noise ratio. We assume that the vector $\mathbf{w}$ is sparse, that is, it has only a few nonzero components, that is, $|\mathbf{J}| \ll p$. We further assume that these nonzero components are either organized on a sequence or on a two-dimensional grid (see Figure 9). Moreover, we consider sets of groups $\mathcal{G}$ such as those presented in Section 3.5. We also consider different choices for the weights $(d^G)_{G\in\mathcal{G}}$ that we denote by **(W1)**, **(W2)** and **(W3)** (we will keep this notation throughout the following experiments):

**(W1)**: Uniform weights, $d_j^G = 1$,

**(W2)**: Weights depending on the size of the groups, $d_j^G = |G|^{-2}$,

**(W3)**: Weights for overlapping groups, $d_j^G = \rho^{|\{H\in\mathcal{G}; H\ni j,\ H\subset G \text{ and } H\neq G\}|}$, for some $\rho \in (0,1)$.

For each orientation in $\mathcal{G}$, the third type of weights **(W3)** aims at reducing the unbalance caused by the overlapping groups. Specifically, given a group $G \in \mathcal{G}$ and a variable $j \in G$, the corresponding weight $d_j^G$ is all the more small as the variable $j$ already belongs to other groups with the same orientation. Unless otherwise specified, we use the third type of weights **(W3)** with $\rho = 0.5$. In the following experiments, the loadings $w_{\mathbf{J}}$, as well as the design matrices, are generated from a standard Gaussian distribution with identity covariance matrix. The positions of $\mathbf{J}$ are also random and are uniformly drawn.

### 6.1 Consistent Hull Estimation

We first illustrate Theorem 6 that establishes necessary and sufficient conditions for consistent hull estimation. To this end, we compute the probability of correct hull estimation when we consider diamond-shaped generating patterns of $|\mathbf{J}| = 24$ variables on a $20\times 20$-dimensional grid (see Figure 9h). Specifically, we generate 500 covariance matrices $\mathbf{Q}$ distributed according to a Wishart distribution with $\delta$ degrees of freedom, where $\delta$ is uniformly drawn in $\{1, 2, \ldots, 10p\}$.[7] The diagonal terms of $\mathbf{Q}$ are then re-normalized to one. For each of these covariance matrices, we compute an

7. We have empirically observed that this choice of degrees of freedom enables to cover well the consistency transition regime around zero in Figure 10.

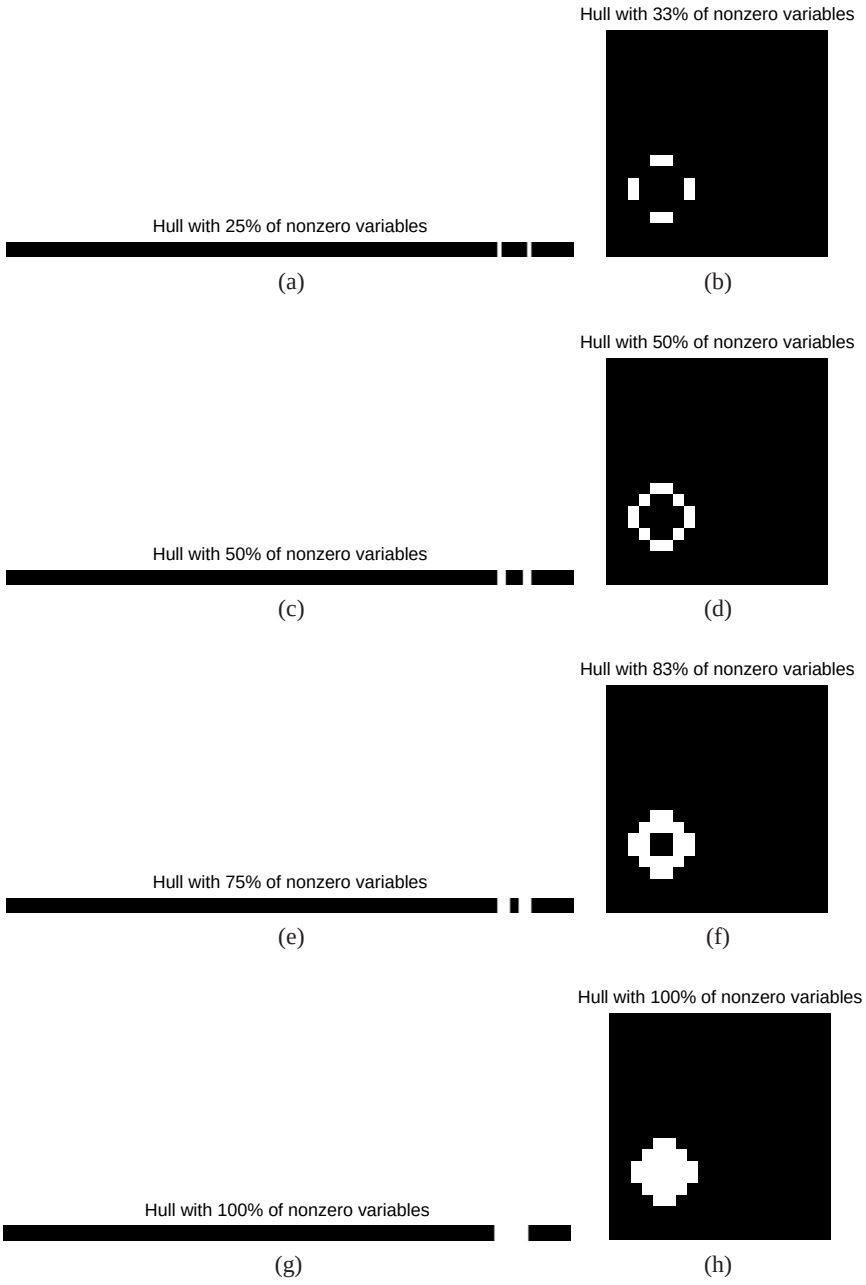



Figure 9: Examples of generating patterns (the zero variables are represented in black, while the nonzero ones are in white): (Left column, in white) generating patterns that are used for the experiments on 400-dimensional sequences; those patterns all form the same hull of 24 variables, that is, the contiguous sequence in (g). (Right column, in white) generating patterns that we use for the $20 \times 20$-dimensional grid experiments; again, those patterns all form the same hull of 24 variables, that is, the diamond-shaped convex in (h). The positions of these generating patterns are randomly selected during the experiments. For the grid setting, the hull is defined based on the set of groups that are half-planes, with orientations that are multiple of $\pi/4$ (see Section 3.5).

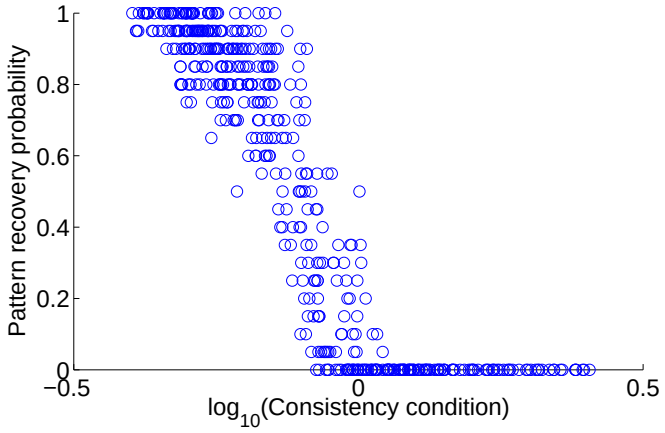



Figure 10: Consistent hull estimation: probability of correct hull estimation versus the consistency condition $(\Omega^c_{\mathbf{J}})^*[\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{J}\mathbf{J}}^{-1}\mathbf{r}_{\mathbf{J}}]$. The transition appears at zero, in good agreement with Theorem 6.

entire regularization path based on one realization of $\{\mathbf{J},\mathbf{w},X,\varepsilon\}$, with $n=3000$ samples. The quantities $\{\mathbf{J},\mathbf{w},\varepsilon\}$ are generated as described previously, while the $n$ rows of $X$ are Gaussian with covariance $\mathbf{Q}$. After repeating 20 times this computation for each $\mathbf{Q}$, we eventually report in Figure 10 the probabilities of correct hull estimation versus the consistency condition $(\Omega^c_{\mathbf{J}})^*[\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{J}\mathbf{J}}^{-1}\mathbf{r}_{\mathbf{J}}]$. In good agreement with Theorem 6, comparing $(\Omega^c_{\mathbf{J}})^*[\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{J}\mathbf{J}}^{-1}\mathbf{r}_{\mathbf{J}}]$ to 1 determines whether the hull is consistently estimated.

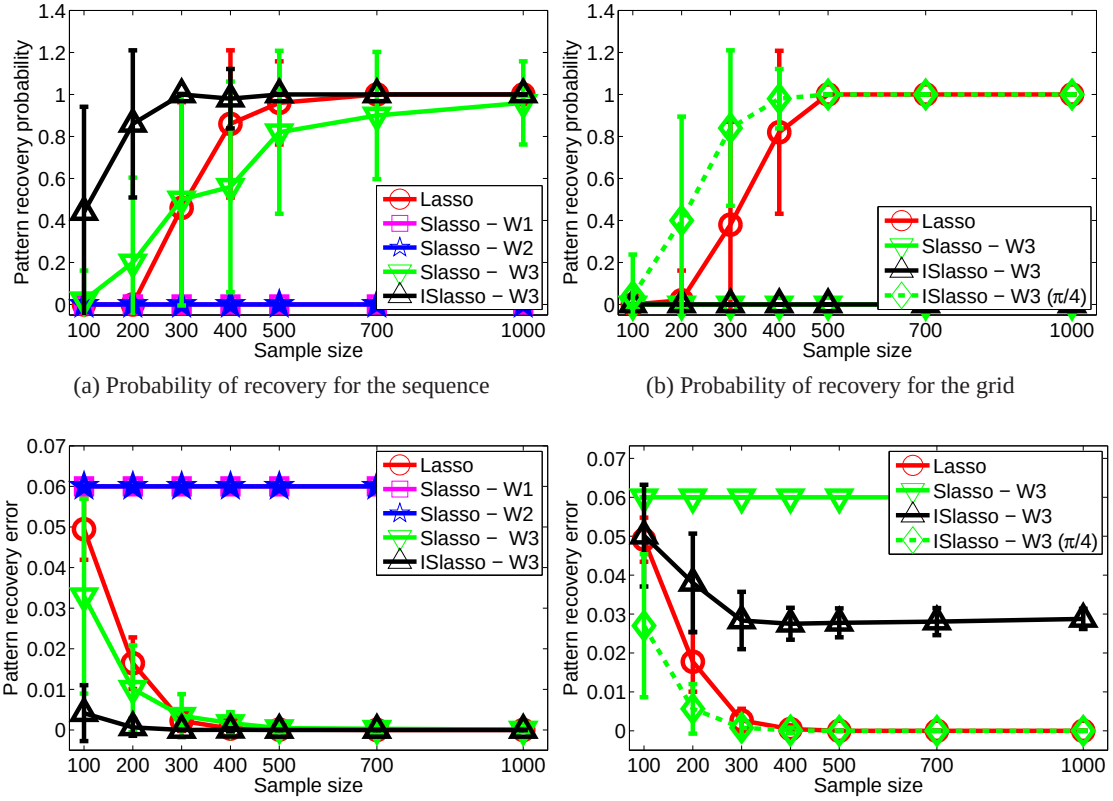
### 6.2 Structured Variable Selection

We show in this experiment that the prior information we put through the norm $\Omega$ improves upon the ability of the model to recover spatially structured nonzero patterns. We are looking at two situations where we can express such a prior through $\Omega$, namely (1) the selection of a contiguous pattern on a sequence (see Figure 9g) and (2) the selection of a convex pattern on a grid (see Figure 9h).

In what follows, we consider $p=400$ variables with generating patterns $\mathbf{w}$ whose hulls are composed of $|\mathbf{J}|=24$ variables. For different sample sizes $n\in\{100,200,300,400,500,700,1000\}$, we consider the probabilities of correct recovery and the (normalized) Hamming distance to the true nonzero patterns. For the realization of a random setting $\{\mathbf{J},\mathbf{w},X,\varepsilon\}$, we compute an entire regularization path over which we collect the closest Hamming distance to the true nonzero pattern and whether it has been exactly recovered for some $\mu$. After repeating 50 times this computation for each sample size $n$, we report the results in Figure 11.

First and foremost, the simulations highlight how important the weights $(d^G)_{G\in\mathcal{G}}$ are. In particular, the uniform **(W1)** and size-dependent weights **(W2)** perform poorly since they do not take into account the overlapping groups. The models learned with such weights do not manage to recover the correct nonzero patterns (in that case, the best model found on the path corresponds to the empty solution, with a normalized Hamming distance of $|\mathbf{J}|/p=0.06$—see Figure 11c).

Although groups that moderately overlap do help (e.g., see Slasso with the weights **(W3)** on Figure 11c), it remains delicate to handle groups with many overlaps, even with an appropriate choice of $(d^G)_{G\in\mathcal{G}}$ (e.g., see Slasso on Figure 11d). The ISlasso procedure does not suffer from this issue since it reduces the number of overlaps whilst keeping the desirable effects of overlapping




(a) Probability of recovery for the sequence

(b) Probability of recovery for the grid

(c) Distance to the true pattern for the sequence

(d) Distance to the true pattern for the grid

Figure 11: For different sample sizes, the probabilities of correct recovery and the (normalized) Hamming distance to the true nonzero patterns are displayed. In the grid case, two sets of groups $\mathcal{G}$ are considered, the rectangular groups with or without the $\pm\pi/4$-groups (denoted by $(\pi/4)$ in the legend). The points and the error bars on the curves respectively represent the mean and the standard deviation, based on 50 random settings $\{\mathbf{J}, \mathbf{w}, X, \varepsilon\}$.

groups. Another way to yield a better level of sparsity, even with many overlaps, would be to consider non-convex alternatives to $\Omega$ (see, e.g., Jenatton et al., 2010b). Moreover, adding the $\pm\pi/4$-groups to the rectangular groups enables to recover a nonzero pattern closer to the generating pattern. This is illustrated on Figure 11d where the error of ISlasso with only rectangular groups (in black) corresponds to the selection of the smallest rectangular box around the generating pattern.

### 6.3 Prediction Error and Relevance of the Structured Prior

In the next simulation, we start from the same setting as Section 6.2 where we additionally evaluate the relevance of the contiguous (or convex) prior by varying the number of zero variables that are contained in the hull (see Figure 9). We then compute the prediction error for different sample sizes $n \in \{250, 500, 1000\}$. The prediction error is understood here as $\|X^{\text{test}}(\mathbf{w} - \hat{w})\|_2^2 / \|X^{\text{test}}\mathbf{w}\|_2^2$, where $\hat{w}$ denotes the OLS estimate, performed on the nonzero pattern found by the model considered (i.e., either Lasso, Slasso or ISlasso). The regularization parameter is chosen by 5-fold cross-validation and the test set consists of 500 samples. For each value of $n$, we display on Figure 12 the median errors over 50 random settings $\{\mathbf{J}, \mathbf{w}, X, \varepsilon\}$, for respectively the sequence and the grid. Note that we have dropped for clarity the models that performed already poorly in Section 6.2.

The experiments show that if the prior about the generating pattern is relevant, then our structured approach performs better that the standard Lasso. Indeed, as soon as the hull of the generating pattern does not contain too many zero variables, Slasso/ISlasso outperform Lasso. In fact, the sample complexity of the Lasso depends on the number of nonzero variables in $\mathbf{w}$ (Wainwright, 2009) as opposed to the size of the hull for Slasso/ISlasso. This also explains why the error for Slasso/ISlasso is almost constant with respect to the number of nonzero variables (since the hull has a constant size).

### 6.4 Active Set Algorithm

We finally 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 structured target. More precisely, for a fixed level of sparsity $|\mathbf{J}| = 24$ and a fixed number of observations $n = 3500$, we analyze the complexity with respect to the number of variables $p$ that varies in $\{100, 225, 400, 900, 1600, 2500\}$. We consider the same experimental protocol as above except that we display the median CPU time based only[8] on 5 random settings $\{\mathbf{J}, \mathbf{w}, X, \varepsilon\}$. We assume that we have a rough idea of the level of sparsity of the true vector and we set the stopping criterion $s = 4|\mathbf{J}|$ (see Algorithm 3), which is a rather conservative choice. We show on Figure 13 that we considerably lower the computational cost for the same level of performance.[9] As predicted by the complexity analysis of the active set algorithm (see the end of Section 4), considering the set of rectangular groups with or without the $\pm\pi/4$-groups results in the same complexity (up to constant terms). We empirically obtain an average complexity of $\approx O(p^{2.13})$ for the SOCP solver and of $\approx O(p^{0.45})$ for the active set algorithm.

Not surprisingly, for small values of $p$, the SOCP solver is faster than the active set algorithm, since the latter has to check its optimality by computing necessary and sufficient conditions (see Algorithm 3 and the discussion in the algorithmic complexity paragraph of Section 4).

8. Note that it already corresponds to several hundreds of runs for both the SOCP and the active set algorithms since we compute a 5-fold cross-validation for each regularization parameter of the (approximate) regularization path.
9. We have not displayed this second figure that is just the superposition of the error curves for the SOCP and the active set algorithms.

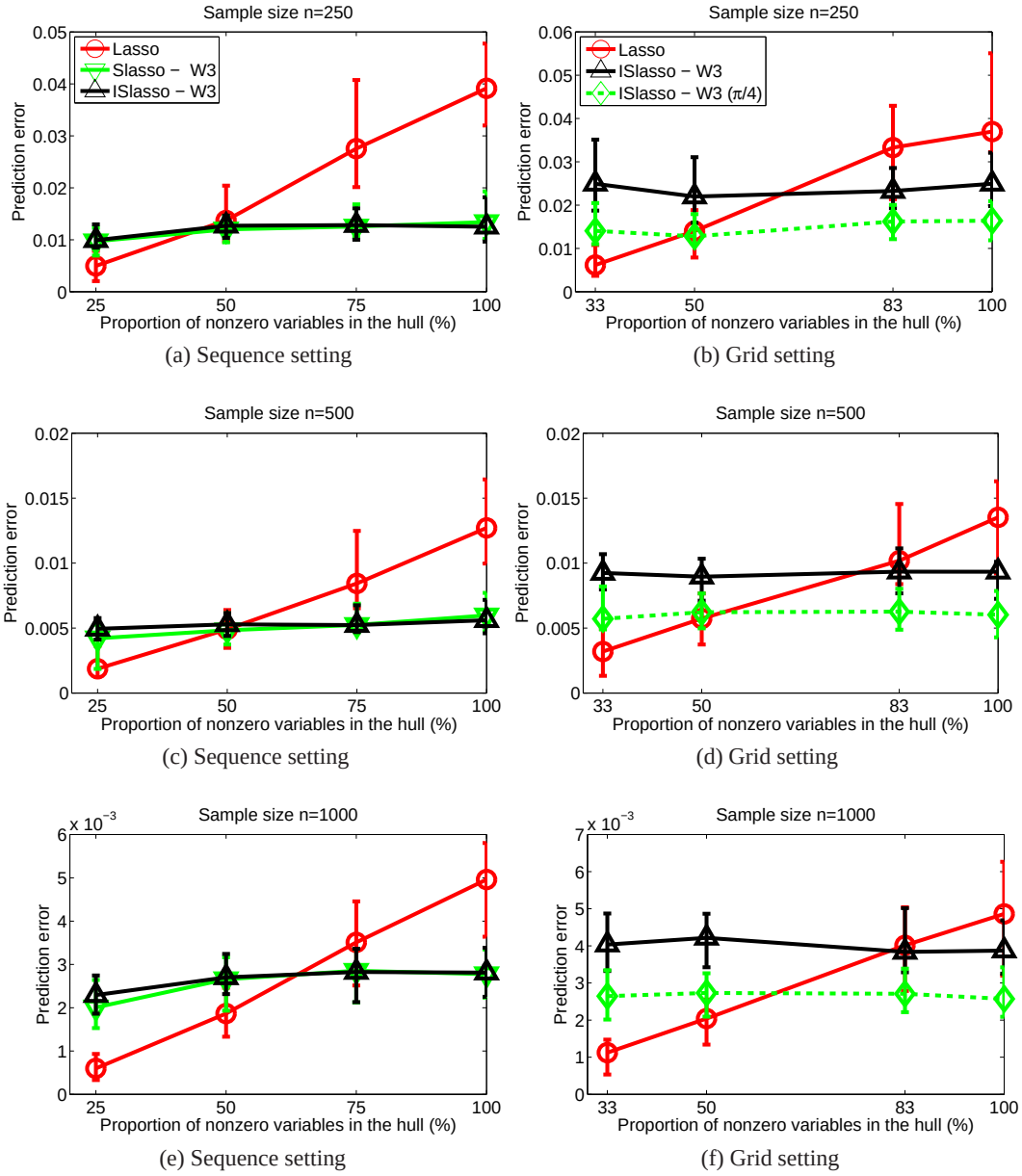



Figure 12: For the sample size $n \in \{250, 500, 1000\}$, we plot the prediction error versus the proportion of nonzero variables in the hull of the generating pattern. In the grid case, two sets of groups $\mathcal{G}$ are considered, the rectangular groups with or without the $\pm\pi/4$-groups (denoted by $(\pi/4)$ in the legend). The points, the lower and upper error bars on the curves respectively represent the median, the first and third quartile, based on 50 random settings $\{\mathbf{J}, \mathbf{w}, X, \varepsilon\}$.

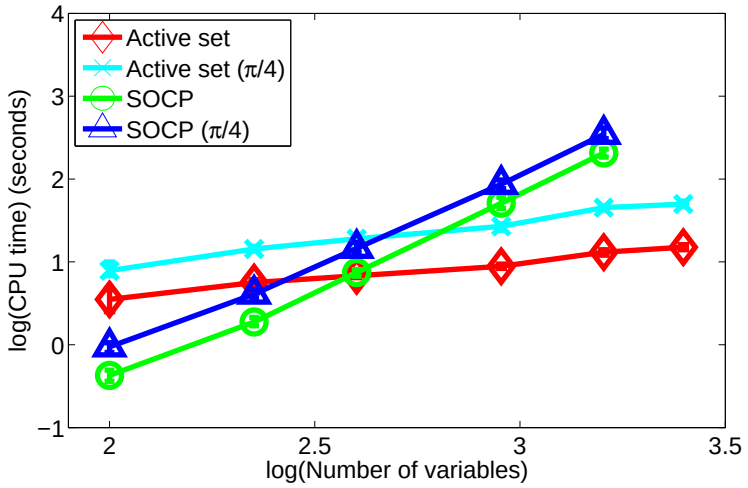



Figure 13: Computational benefit of the active set algorithm: CPU time (in seconds) versus the number of variables $p$, displayed in log-log scale. The points, the lower and upper error bars on the curves respectively represent the median, the first and third quartile. Two sets of groups $\mathcal{G}$ are considered, the rectangular groups with or without the $\pm\pi/4$-groups (denoted by $(\pi/4)$ in the legend). Due to the computational burden, we could not obtain the SOCP's results for $p = 2500$.

## 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 which linearly combines $\ell_2$-norms of possibly overlapping groups of variables. Our framework brings into play intersection-closed families of nonzero patterns, such as all rectangles on a two-dimensional grid. We have studied the design of these groups, efficient algorithms and theoretical guarantees of the structured sparsity-inducing method. Our experiments have shown to which extent our model leads to better prediction, depending on the relevance of the prior information.

A natural extension to this work is to consider bootstrapping since this may improve theoretical guarantees and result in better variable selection (Bach, 2008c; Meinshausen and Bühlmann, 2010). In order to deal with broader families of (non)zero patterns, it would be interesting to combine our approach with recent work on structured sparsity: for instance, Baraniuk et al. (2010) and Jacob et al. (2009) consider union-closed collections of nonzero patterns, He and Carin (2009) exploit structure through a Bayesian prior while Huang et al. (2009) handle non-convex penalties based on information-theoretic criteria.

More generally, our regularization scheme could also be used for various learning tasks, as soon as prior knowledge on the structure of the sparse representation is available, for example, for multiple kernel learning (Micchelli and Pontil, 2006), multi-task learning (Argyriou et al., 2008; Obozinski et al., 2009; Kim and Xing, 2010) and sparse matrix factorization problems (Mairal et al., 2010a; Jenatton et al., 2010b, 2011b).

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 and tem-

porally structured data typical in bioinformatics (Kim and Xing, 2010), computer vision (Jenatton et al., 2010b; Mairal et al., 2010b) and neuroscience applications (see, e.g., Varoquaux et al., 2010).

## Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments that improve the clarity and the overall quality of the manuscript. We also thank Julien Mairal and Guillaume Obozinski for insightful discussions. This work was supported in part by a grant from the Agence Nationale de la Recherche (MGA Project) and a grant from the European Research Council (SIERRA Project).

## Appendix A. Proof of Proposition 1

We recall that $L(w) = \frac{1}{n}\sum_{i=1}^{n} \ell\big(y_i, w^\top x_i\big)$. Since $w \mapsto \Omega(w)$ is convex and goes to infinite when $\|w\|_2$ goes to infinite, and since $L$ is lower bounded, by Weierstrass' theorem, the problem in Equation (2) admits at least one global solution.

•*First case: $Q$ invertible.* The Hessian of $L$ is

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top \frac{\partial^2 \ell}{\partial y'^2}(y_i, w^\top x_i).$$

It is positive definite since $Q$ is positive definite and $\min_{i\in\{1,\ldots,n\}} \frac{\partial^2 \ell}{\partial y'^2}(y_i, w^\top x_i) > 0$. So $L$ is strictly convex. Consequently the objective function $L + \mu\Omega$ is strictly convex, hence the uniqueness of its minimizer.

•*Second case: $\{1,\ldots,p\}$ belongs to $\mathcal{G}$.* We prove the uniqueness by contradiction. Assume that the problem in Equation (2) admits two different solutions $w$ and $\tilde{w}$. Then one of the two solutions is different from 0, say $w \neq 0$.

By convexity, it means that any point of the segment $[w,\tilde{w}] = \big\{aw + (1-a)\tilde{w};\ a \in [0,1]\big\}$ also minimizes the objective function $L + \mu\Omega$. Since both $L$ and $\mu\Omega$ are convex functions, it means that they are both linear when restricted to $[w,\tilde{w}]$.

Now, $\mu\Omega$ is only linear on segments of the form $[v,tv]$ with $v \in \mathbb{R}^p$ and $t > 0$. So we necessarily have $\tilde{w} = tw$ for some positive $t$. We now show that $L$ is strictly convex on $[w,tw]$, which will contradict that it is linear on $[w,\tilde{w}]$. Let $E = \mathrm{Span}(x_1,\ldots,x_n)$ and $E^\perp$ be the orthogonal of $E$ in $\mathbb{R}^p$. The vector $w$ can be decomposed in $w = w' + w''$ with $w' \in E$ and $w'' \in E^\perp$. Note that we have $w' \neq 0$ (since if it was equal to 0, $w''$ would be the minimizer of $\mu\Omega$, which would imply $w'' = 0$ and contradict $w \neq 0$). We thus have $(w^\top x_1,\ldots,w^\top x_n) = (w'^\top x_1,\ldots,w'^\top x_n) \neq 0$.

This implies that the function $s \mapsto \ell(y_i, sw^\top x_i)$ is a polynomial of degree 2. So it is not linear. This contradicts the existence of two different solutions, and concludes the proof of uniqueness.

**Remark 8** *Still by using that a sum of convex functions is constant on a segment if and only if the functions are linear on this segment, the proof can be extended in order to replace the alternative assumption "$\{1,\ldots,p\}$ belongs to $\mathcal{G}$" by the weaker but more involved assumption: for any $(j,k) \in \{1,\ldots,p\}^2$, there exists a group $G \in \mathcal{G}$ which contains both $j$ and $k$.*

## Appendix B. Proof of Theorem 2

For $w \in \mathbb{R}^p$, we denote by $Z(w)$ its zero pattern (i.e., the indices of zero-components of $w$). To prove the result, it suffices to prove that for any set $I \subset \{1,\dots,p\}$ with $I^c \notin \mathcal{Z}$ and $|I| \leq k-1$, the probability of

$$\mathcal{E}_I = \big\{Y \in \mathbb{R}^n\text{: there exists } w \text{ solution of the problem in Equation (2) with } Z(w) = I^c\big\}$$

is equal to 0. We will prove this by contradiction: assume that there exists a set $I \subset \{1,\dots,p\}$ with $I^c \notin \mathcal{Z}$, $|I| \leq k-1$ and $\mathbb{P}(\mathcal{E}_I) > 0$. Since $I^c \notin \mathcal{Z}$, there exists $\alpha \in \mathrm{Hull}(I) \setminus I$. Let $J = I \cup \{\alpha\}$ and $\mathcal{G}_I = \{G \in \mathcal{G} : G \cap I \neq \emptyset\}$ be the set of active groups. Define $\mathbb{R}^J = \{w \in \mathbb{R}^p : w_{J^c} = 0\}$. The restrictions $L_J : \mathbb{R}^J \to \mathbb{R}$ and $\Omega_J : \mathbb{R}^J \to \mathbb{R}$ of $L$ and $\Omega$ are continuously differentiable functions on $\big\{w \in \mathbb{R}^J : w_I \neq 0\big\}$ with respective gradients

$$\nabla L_J(w) = \left(\frac{\partial L_J}{\partial w_j}(w)\right)_{j\in J}^\top \quad \text{and} \quad \nabla \Omega_J(w) = \left(w_j\Big(\sum_{\substack{G\in\mathcal{G}_I,\\ G\ni j}} (d_j^G)^2 \,\|d^G \circ w\|_2^{-1}\Big)\right)_{j\in J}^\top.$$

Let $f(w,Y) = \nabla L_J(w) + \mu \nabla \Omega_J(w)$, where the dependence in $Y$ of $f(w,Y)$ is hidden in $\nabla L_J(w) = \frac{1}{n}\sum_{i=1}^n (x_i)_J \frac{\partial \ell}{\partial y}(y_i, w^\top x_i)$.

For $Y \in \mathcal{E}_I$, there exists $w \in \mathbb{R}^J$ with $Z(w) = I^c$, which minimizes the convex function $L_J + \mu\Omega_J$. The vector $w$ satisfies $f(w,Y) = 0$. So we have proved $\mathcal{E}_I \subset \mathcal{E}_I'$, where

$$\mathcal{E}_I' = \big\{Y \in \mathbb{R}^n : \text{ there exists } w \in \mathbb{R}^J \text{ with } Z(w) = I^c \text{ and } f(w,Y) = 0\big\}.$$

Let $\tilde{y} \in \mathcal{E}_I$. Consider the equation $f(w,\tilde{y}) = 0$ on $\big\{w \in \mathbb{R}^J : w_j \neq 0 \text{ for any } j \in I\big\}$. By construction, we have $|J| \leq k$, and thus, by assumption, the matrix $X^J = \big((x_1)_J, ..., (x_n)_J\big)^\top \in \mathbb{R}^{n\times|J|}$ has rank $|J|$. As in the proof of Proposition 1, this implies that the function $L_J$ is strictly convex, and then, the uniqueness of the minimizer of $L_J + \mu\Omega$, and also the uniqueness of the point at which the gradient of this function vanishes. So the equation $f(w,\tilde{y}) = 0$ on $\big\{w \in \mathbb{R}^J : w_j \neq 0 \text{ for any } j \in I\big\}$ has a unique solution, which we will write $w^{\tilde{y}}$.

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 $(f_j)_{j\in J}$ be the components of $f$ and $H_{JJ} = \big(\frac{\partial f_j}{\partial w_k}\big)_{j\in J, k\in J}$. The matrix $H_{JJ}$ is actually the sum of:

a) the Hessian of $L_J$, which is positive definite (still from the same argument as in the proof of Theorem 1),

b) the Hessian of the norm $\Omega_J$ 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 $\tilde{y}$,

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

with $\psi = (\psi_j)_{j\in J}$ a continuously differentiable function satisfying the matricial relation

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

Let $C_\alpha$ denote the column vector of $H_{JJ}^{-1}$ corresponding to the index $\alpha$, and let $D$ the diagonal matrix whose $(i,i)$-th element is $\frac{\partial^2 \ell}{\partial y \partial y'}(y_i, w^\top x_i)$. Since $n(\ldots, \nabla_y f_j, \ldots) = DX^J$, we have

$$n\nabla\psi_\alpha = -DX^J C_\alpha.$$

Now, since $X^J$ has full rank, $C_\alpha \neq 0$ and none of the diagonal elements of $D$ is null (by assumption on $\ell$), 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 an open ball in $\mathbb{R}^n$ centered at $\tilde{y}$, say $\mathcal{B}_{\tilde{y}}$, the solution to $\psi_\alpha(Y) = 0$ can be written $y_1 = \varphi(y_2, \ldots, y_n)$ with $\varphi$ 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 get that the set $A(\tilde{y}) = \left\{Y \in \mathcal{B}_{\tilde{y}} : \psi_\alpha(Y) = 0\right\}$ has thus zero probability. Let $\mathcal{S} \subset \mathcal{E}_I$ be a compact set. We thus have $\mathcal{S} \subset \mathcal{E}_I'$.

By compacity, the set $\mathcal{S}$ can be covered by a finite number of ball $\mathcal{B}_{\tilde{y}}$. So there exist $\tilde{y}_1, \ldots, \tilde{y}_m$ such that we have $\mathcal{S} \subset A(\tilde{y}_1) \cup \cdots \cup A(\tilde{y}_m)$. Consequently, we have $\mathbb{P}(\mathcal{S}) = 0$.

Since this holds for any compact set in $\mathcal{E}_I$ and since the Lebesgue measure is regular, we have $\mathbb{P}(\mathcal{E}_I) = 0$, which contradicts the definition of $I$, and concludes the proof.

## Appendix C. Proof of the Minimality of the Backward Procedure (See Algorithm 1)

There are essentially two points to show: (1) $\mathcal{G}$ spans $\mathcal{Z}$, and (2) $\mathcal{G}$ 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 $\{\bigcup_{G\in\mathcal{G}'} G,\ \forall \mathcal{G}' \subseteq \mathcal{G}^{(t)}\} = \{G \in \mathcal{Z}, |G| \leq t\}$ 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 $\mathcal{Z}$.

## Appendix D. Proof of Proposition 3

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

$$h_J : w_J \mapsto \frac{\lambda}{2}\left[\Omega_J(w_J)\right]^2,$$

whose Fenchel conjugate $h_J^*$ is given by $\kappa_J \mapsto \frac{1}{2\lambda}\left[\Omega_J^*(\kappa_J)\right]^2$ (Boyd and Vandenberghe, 2004, example 3.27). 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. Moreover, the primal-dual variables $\{w_J, \kappa_J\}$ is optimal if and only if

$$\begin{cases} -\kappa_J & \in \quad \partial L_J(w_J), \\ \kappa_J & \in \quad \partial[\frac{\lambda}{2}[\Omega_J(w_J)]^2] = \lambda\Omega_J(w_J)\partial\Omega_J(w_J), \end{cases}$$

where $\partial\Omega_J(w_J)$ denotes the subdifferential of $\Omega_J$ at $w_J$. The differentiability of $L_J$ at $w_J$ then gives $\partial L_J(w_J) = \{\nabla L_J(w_J)\}$. It now remains to show that

$$\kappa_J \in \lambda\Omega_J(w_J)\partial\Omega_J(w_J) \tag{5}$$

is equivalent to

$$w_J^\top \kappa_J = \frac{1}{\lambda}[\Omega_J^*(\kappa_J)]^2 = \lambda[\Omega_J(w_J)]^2. \tag{6}$$

As a starting point, the Fenchel-Young inequality (Borwein and Lewis, 2006, Proposition 3.3.4) gives the equivalence between Equation (5) and

$$\frac{\lambda}{2}[\Omega_J(w_J)]^2 + \frac{1}{2\lambda}[\Omega_J^*(\kappa_J)]^2 = w_J^\top \kappa_J. \tag{7}$$

In addition, we have (Rockafellar, 1970)

$$\partial\Omega_J(w_J) = \{u_J \in \mathbb{R}^{|J|}; u_J^\top w_J = \Omega_J(w_J) \text{ and } \Omega_J^*(u_J) \leq 1\}. \tag{8}$$

Thus, if $\kappa_J \in \lambda\Omega_J(w_J)\partial\Omega_J(w_J)$ then $w_J^\top \kappa_J = \lambda[\Omega_J(w_J)]^2$. Combined with Equation (7), we obtain $w_J^\top \kappa_J = \frac{1}{\lambda}[\Omega_J^*(\kappa_J)]^2$.

Reciprocally, starting from Equation (6), we notably have

$$w_J^\top \kappa_J = \lambda[\Omega_J(w_J)]^2.$$

In light of Equation (8), it suffices to check that $\Omega_J^*(\kappa_J) \leq \lambda\Omega_J(w_J)$ in order to have Equation (5). Combining Equation (6) with the definition of the dual norm, it comes

$$\frac{1}{\lambda}[\Omega_J^*(\kappa_J)]^2 = w_J^\top \kappa_J \leq \Omega_J^*(\kappa_J)\Omega_J(w_J),$$

which concludes the proof of the equivalence between Equation (5) and Equation (6).

## Appendix E. Proofs of Propositions 4 and 5

In order to check that the reduced solution $w_J$ is optimal for the full problem in Equation (3), 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

$$\frac{1}{2\lambda}\left([\Omega^*(\kappa)]^2 - \lambda w_J^\top \kappa_J\right).$$

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

### E.1 Proof of Proposition 4

Let us suppose that $w^* = \binom{w_J^*}{0_{J^c}}$ is optimal for the full problem in Equation (3). Following the same derivation as in Lemma 14 (up to the squaring of the regularization $\Omega$), we have that $w^*$ is a solution of Equation (3) if and only if for all $u \in \mathbb{R}^p$,

$$u^\top \nabla L(w^*) + \lambda \Omega(w^*)(u_J^\top r_J + (\Omega_J^c)[u_{J^c}]) \geq 0,$$

with

$$r = \sum_{G \in \mathcal{G}_J} \frac{d^G \circ d^G \circ w^*}{\|d^G \circ w^*\|_2}.$$

We project the optimality condition onto the variables that can possibly enter the active set, that is, the variables in $\Pi_{\mathcal{P}}(J)$. Thus, for each $K \in \Pi_{\mathcal{P}}(J)$, we have for all $u_{K\backslash J} \in \mathbb{R}^{|K\backslash J|}$,

$$u_{K\backslash J}^\top \nabla L(w^*)_{K\backslash J} + \lambda \Omega(w^*) \sum_{G \in \mathcal{G}_{K\backslash J} \cap (\mathcal{G}_J)^c} \left\| d^G_{K\backslash J} \circ u_{G \cap K\backslash J} \right\|_2 \geq 0.$$

By combining Lemma 13 and the fact that $\mathcal{G}_{K\backslash J} \cap (\mathcal{G}_J)^c = \mathcal{G}_K \backslash \mathcal{G}_J$, we have for all $G \in \mathcal{G}_K \backslash \mathcal{G}_J$, $K\backslash J \subseteq G$ and therefore $u_{G \cap K\backslash J} = u_{K\backslash J}$. Since we cannot compute the dual norm of $u_{K\backslash J} \mapsto \|d^G_{K\backslash J} \circ u_{K\backslash J}\|_2$ in closed-form, we instead use the following upperbound

$$\left\| d^G_{K\backslash J} \circ u_{K\backslash J} \right\|_2 \leq \|d^G_{K\backslash J}\|_\infty \left\| u_{K\backslash J} \right\|_2,$$

so that we get for all $u_{K\backslash J} \in \mathbb{R}^{|K\backslash J|}$,

$$u_{K\backslash J}^\top \nabla L(w^*)_{K\backslash J} + \lambda \Omega(w^*) \sum_{G \in \mathcal{G}_K \backslash \mathcal{G}_J} \|d^G_{K\backslash J}\|_\infty \left\| u_{K\backslash J} \right\|_2 \geq 0.$$

Finally, Proposition 3 gives $\lambda \Omega(w^*) = \left\{ -\lambda {w^*}^\top \nabla L(w^*) \right\}^{\frac{1}{2}}$, which leads to the desired result.

### E.2 Proof of Proposition 5

The goal of the proof is to upper bound the dual norm $\Omega^*(\kappa)$ by taking advantage of the structure of $\mathcal{G}$; we first show how we can upper bound $\Omega^*(\kappa)$ by $(\Omega_J^c)^*[\kappa_{J^c}]$. We indeed have:

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

where in the last line, we use Lemma 15. Thus the duality gap is less than

$$\frac{1}{2\lambda}\left( [\Omega^*(\kappa)]^2 - [\Omega_J^*(\kappa_J)]^2 \right) \leq \frac{1}{2\lambda} \max\{0, [(\Omega_J^c)^*[\kappa_{J^c}]]^2 - [\Omega_J^*(\kappa_J)]^2\},$$

and a sufficient condition for the duality gap to be smaller than $\varepsilon$ is

$$(\Omega_J^c)^*[\kappa_{J^c}] \leq (2\lambda\varepsilon + [\Omega_J^*(\kappa_J)]^2)^{\frac{1}{2}}.$$

Using Proposition 3, we have $-\lambda w^\top \nabla L(w) = [\Omega_J^*(\kappa_J)]^2$ and we get the right-hand side of Proposition 5. It now remains to upper bound $(\Omega_J^c)^*[\kappa_{J^c}]$. To this end, we call upon Lemma 11 to obtain:

$$(\Omega_J^c)^*[\kappa_{J^c}] \leq \max_{G \in (\mathcal{G}_J)^c} \left\{ \sum_{j \in G} \left\{ \frac{\kappa_j}{\sum_{H \in j, H \in (\mathcal{G}_J)^c} d_j^H} \right\}^2 \right\}^{\frac{1}{2}}.$$

Among all groups $G \in (\mathcal{G}_J)^c$, the ones with the maximum values are the largest ones, that is, those in the fringe groups $\mathcal{F}_J = \{G \in (\mathcal{G}_J)^c \ ; \ \nexists G' \in (\mathcal{G}_J)^c, G \subseteq G'\}$. This argument leads to the result of Proposition 5.

## Appendix F. Proof of Theorem 6

*Necessary condition:* We mostly follow the proof of Zou (2006) and Bach (2008b). 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 $\Omega$ at $\mathbf{w}$ in the direction of $\Delta$

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

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

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

The limiting function $\tilde{F}_{\lim}$ 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 $\tilde{F}_{\lim}$ (Fu and Knight, 2000) referred to as $\Delta^*$.

By assumption, with probability tending to one, we have $\mathbf{J} = \{j \in \{1, \ldots, p\}, \hat{w}_j \neq 0\}$, hence for any $j \in \mathbf{J}^c$ $\mu\hat{\Delta}_j = (\hat{w} - \mathbf{w})_j = 0$. This implies that the nonrandom vector $\Delta^*$ 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 $\Delta^*$ is the minimizer of $\tilde{F}_{\lim}$, by taking the directional derivatives as in the proof of Lemma 14, 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 2, 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 consider that $\hat{w}$ is 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 14) relative to the active variables $\mathbf{J}$ is already verified. More precisely, we have

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

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

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

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 $u_{\mathbf{J}^c}\in\mathbb{R}^{|\mathbf{J}^c|}$

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

Since we assume $(\Omega_{\mathbf{J}}^c)^*[\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{JJ}}^{-1}\mathbf{r}_{\mathbf{J}}] < 1$, 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 14 relative to the inactive variables: $\hat{w}$ is therefore optimal for the full problem.

## Appendix G. Proof of Theorem 7

Since our analysis takes place in a finite-dimensional space, all the norms defined on this space are equivalent. Therefore, we introduce the equivalence parameters $a(\mathbf{J}), A(\mathbf{J}) > 0$ such that

$$\forall u\in\mathbb{R}^{|\mathbf{J}|},\ a(\mathbf{J})\|u\|_1 \le \Omega_{\mathbf{J}}[u] \le A(\mathbf{J})\|u\|_1.$$

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

$$\forall u\in\mathbb{R}^{|\mathbf{J}|},\ A(\mathbf{J})^{-1}\|u\|_\infty \le (\Omega_{\mathbf{J}})^*[u] \le a(\mathbf{J})^{-1}\|u\|_\infty.$$

For any matrix $\Gamma$, we also introduce the operator norm $\|\Gamma\|_{m,s}$ defined as

$$\|\Gamma\|_{m,s} = \sup_{\|u\|_s \leq 1} \|\Gamma u\|_m.$$

Moreover, our proof will rely on the control of the *expected dual norm for isonormal vectors*: $\mathbb{E}\left[(\Omega^c_{\mathbf{J}})^*(W)\right]$ with $W$ a centered Gaussian random variable with unit covariance matrix. In the case of the Lasso, it is of order $(\log p)^{1/2}$.

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 14), we know that

$$\Omega^*_{\mathbf{J}}[Q_{\mathbf{JJ}}(\hat{w}_{\mathbf{J}} - \mathbf{w}_{\mathbf{J}}) - q_{\mathbf{J}}] \leq \mu, \tag{9}$$

where the vector $q \in \mathbb{R}^p$ is defined as $q = \frac{1}{n}\sum_{i=1}^n \varepsilon_i x_i$. 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 $\|\hat{w}_G\|_\infty > 0$ 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 group is missing).

We have

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

hence from (9) and the definition of $A(\mathbf{J})$,

$$\|\hat{w}_{\mathbf{J}} - \mathbf{w}_{\mathbf{J}}\|_\infty \leq |\mathbf{J}|^{1/2}\kappa^{-1}\left(\mu A(\mathbf{J}) + \|q_{\mathbf{J}}\|_\infty\right). \tag{10}$$

Thus, if we assume $\mu \leq \frac{\kappa\nu}{3|\mathbf{J}|^{1/2}A(\mathbf{J})}$ and

$$\|q_{\mathbf{J}}\|_\infty \leq \frac{\kappa\nu}{3|\mathbf{J}|^{1/2}}, \tag{11}$$

we get

$$\|\hat{w}_{\mathbf{J}} - \mathbf{w}_{\mathbf{J}}\|_\infty \leq 2\nu/3, \tag{12}$$

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

This also ensures that $\hat{w}_{\mathbf{J}}$ satisfies the equation (see Lemma 14)

$$Q_{\mathbf{JJ}}(\hat{w}_{\mathbf{J}} - \mathbf{w}_{\mathbf{J}}) - q_{\mathbf{J}} + \mu\hat{r}_{\mathbf{J}} = 0, \tag{13}$$

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 14, since we have already proved (13), it suffices to show that

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

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

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

which leads us to control the difference $\hat{r}_{\mathbf{J}} - \mathbf{r}_{\mathbf{J}}$. Using Lemma 12, we get

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

where $w = t_0\hat{w} + (1-t_0)\mathbf{w}$ for some $t_0 \in (0,1)$.

Let $\bar{\mathbf{J}} = \{k \in \mathbf{J} : \mathbf{w}_k \neq 0\}$ and let $\varphi$ be defined as

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

The term $\varphi$ basically measures how close $\mathbf{J}$ and $\bar{\mathbf{J}}$ are, that is, how relevant the prior encoded by $\mathcal{G}$ about the hull $\mathbf{J}$ is. By using (12), we have

$$\|d^G \circ w\|_2^2 \geq \left\|d^G_{\bar{\mathbf{J}}} \circ w_{\bar{\mathbf{J}}}\right\|_2^2 \geq \|d^G_{\bar{\mathbf{J}}} \circ d^G_{\bar{\mathbf{J}}} \circ w_{\bar{\mathbf{J}}}\|_1 \frac{\nu}{3} \geq \|d^G \circ d^G \circ w\|_1 \frac{\nu}{3\varphi},$$

$$\|d^G \circ w\|_2 \geq \|d^G_{\bar{\mathbf{J}}} \circ w_{\bar{\mathbf{J}}}\|_2 \geq \|d^G_{\bar{\mathbf{J}}}\|_2 \frac{\nu}{3} \geq \left\|d^G_{\mathbf{J}}\right\|_2 \frac{\nu}{3\sqrt{\varphi}}$$

and

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

Therefore we have

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

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

$$\|\hat{r}_{\mathbf{J}} - \mathbf{r}_{\mathbf{J}}\|_1 \leq \alpha \|\hat{w}_{\mathbf{J}} - \mathbf{w}_{\mathbf{J}}\|_\infty. \tag{14}$$

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

$$\begin{aligned}
\left\|Q_{\mathbf{J}^c\mathbf{J}}Q_{\mathbf{J}\mathbf{J}}^{-1}(\hat{r}_{\mathbf{J}} - \mathbf{r}_{\mathbf{J}})\right\|_\infty &= \left\|Q_{\mathbf{J}^c\mathbf{J}}Q_{\mathbf{J}\mathbf{J}}^{-1/2}Q_{\mathbf{J}\mathbf{J}}^{-1/2}(\hat{r}_{\mathbf{J}} - \mathbf{r}_{\mathbf{J}})\right\|_\infty \\
&\leq \|Q_{\mathbf{J}^c\mathbf{J}}Q_{\mathbf{J}\mathbf{J}}^{-1/2}\|_{\infty,2}\|Q_{\mathbf{J}\mathbf{J}}^{-1/2}\|_2 \|\hat{r}_{\mathbf{J}} - \mathbf{r}_{\mathbf{J}}\|_2 \\
&\leq \kappa^{-1/2}\|\hat{r}_{\mathbf{J}} - \mathbf{r}_{\mathbf{J}}\|_1 \\
&\leq \kappa^{-3/2}\alpha|\mathbf{J}|^{1/2}\left(\mu A(\mathbf{J}) + \|q_{\mathbf{J}}\|_\infty\right),
\end{aligned}$$

where the last line comes from Equation (10) and (14). We get

$$(\Omega_{\mathbf{J}}^c)^*[Q_{\mathbf{J}^c\mathbf{J}}Q_{\mathbf{J}\mathbf{J}}^{-1}(\hat{r}_{\mathbf{J}}-\mathbf{r}_{\mathbf{J}})] \leq \frac{\alpha|\mathbf{J}|^{1/2}}{\kappa^{3/2}a(\mathbf{J}^c)}\left(\mu A(\mathbf{J})+\|q_{\mathbf{J}}\|_\infty\right).$$

Thus, if the following inequalities are verified

$$\frac{\alpha|\mathbf{J}|^{1/2}A(\mathbf{J})}{\kappa^{3/2}a(\mathbf{J}^c)}\mu \leq \frac{\tau}{4},$$

$$\frac{\alpha|\mathbf{J}|^{1/2}}{\kappa^{3/2}a(\mathbf{J}^c)}\|q_{\mathbf{J}}\|_\infty \leq \frac{\tau}{4}, \tag{15}$$

$$(\Omega_{\mathbf{J}}^c)^*[q_{\mathbf{J}^c|\mathbf{J}}] \leq \frac{\mu\tau}{2}, \tag{16}$$

we obtain

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

that is, $\mathbf{J}$ 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 (11), (15) and (16) 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

$$\begin{aligned}\mathbb{E}\left[q_{\mathbf{J}^c|\mathbf{J}}q_{\mathbf{J}^c|\mathbf{J}}^\top\right] &= \mathbb{E}\left[q_{\mathbf{J}^c}q_{\mathbf{J}^c}^\top - q_{\mathbf{J}^c}q_{\mathbf{J}}^\top Q_{\mathbf{J}\mathbf{J}}^{-1}Q_{\mathbf{J}\mathbf{J}^c} - Q_{\mathbf{J}^c\mathbf{J}}Q_{\mathbf{J}\mathbf{J}}^{-1}q_{\mathbf{J}}q_{\mathbf{J}^c}^\top + Q_{\mathbf{J}^c\mathbf{J}}Q_{\mathbf{J}\mathbf{J}}^{-1}q_{\mathbf{J}}q_{\mathbf{J}}^\top Q_{\mathbf{J}\mathbf{J}}^{-1}Q_{\mathbf{J}\mathbf{J}^c}\right]\\ &= \frac{\sigma^2}{n}Q_{\mathbf{J}^c\mathbf{J}^c|\mathbf{J}},\end{aligned}$$

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

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

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

In addition, we have

$$|\psi(u)-\psi(v)| \leq \psi(u-v) \leq \sigma n^{-1/2}a(\mathbf{J}^c)^{-1}\left\|Q_{\mathbf{J}^c\mathbf{J}^c|\mathbf{J}}^{1/2}(u-v)\right\|_\infty.$$

On the other hand, since $Q$ has unit diagonal and $Q_{\mathbf{J}^c\mathbf{J}}Q_{\mathbf{J}\mathbf{J}}^{-1}Q_{\mathbf{J}\mathbf{J}^c}$ has diagonal terms less than one, $Q_{\mathbf{J}^c\mathbf{J}^c|\mathbf{J}}$ also has diagonal terms less than one, which implies that $\|Q_{\mathbf{J}^c\mathbf{J}^c|\mathbf{J}}^{1/2}\|_{\infty,2} \leq 1$. Hence $\psi$ is a Lipschitz function with Lipschitz constant upper bounded by $\sigma n^{-1/2}a(\mathbf{J}^c)^{-1}$. Thus by concentration

of Lipschitz functions of multivariate standard random variables (Massart, 2003, Theorem 3.4), we have for $t > 0$:

$$\mathbb{P}\left[(\Omega_{\mathbf{J}}^c)^*[q_{\mathbf{J}^c|\mathbf{J}}] \geq t + \sigma n^{-1/2}\|Q\|_2^{1/2}\mathbb{E}\left[(\Omega_{\mathbf{J}}^c)^*(W)\right]\right] \leq \exp\left(-\frac{nt^2 a(\mathbf{J}^c)^2}{2\sigma^2}\right).$$

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

$$\mathbb{P}\left[(\Omega_{\mathbf{J}}^c)^*[q_{\mathbf{J}^c|\mathbf{J}}] \geq t\right] \leq \exp\left(-\frac{n\mu^2\tau^2 a(\mathbf{J}^c)^2}{32\sigma^2}\right).$$

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

$$\xi = \frac{\kappa\nu}{3}\min\left\{1, \frac{3\tau\kappa^{1/2}a(\mathbf{J}^c)}{4\alpha\nu}\right\}.$$

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

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

To conclude, Theorem 7 holds with

$$\begin{aligned}
C_1(\mathcal{G},\mathbf{J}) &= \frac{a(\mathbf{J}^c)^2}{16}, \\
C_2(\mathcal{G},\mathbf{J}) &= \left(\frac{\kappa\nu}{3}\min\left\{1, \frac{\tau\kappa^{1/2}a(\mathbf{J}^c)\nu}{24\varphi^{3/2}\|\mathbf{w}\|_\infty \sum_{G\in\mathcal{G}_{\mathbf{J}}}\left\|d_{\mathbf{J}}^G\right\|_2}\right\}\right)^2, \\
C_3(\mathcal{G},\mathbf{J}) &= 4\|Q\|_2^{1/2}\mathbb{E}\left[(\Omega_{\mathbf{J}}^c)^*(W)\right],
\end{aligned}$$

and

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

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

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

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

## Appendix H. A First Order Approach to Solve Problems (2) and (3)

Both regularized minimization problems in Equation (2) and Equation (3) (that just differ in the squaring of $\Omega$) can be solved by using generic toolboxes for second-order cone programming (SOCP) (Boyd and Vandenberghe, 2004). We propose here a first order approach that takes up

ideas from Micchelli and Pontil (2006) and Rakotomamonjy et al. (2008) and that is based on the following variational equalities: for $x \in \mathbb{R}^p$, we have

$$\|x\|_1^2 = \min_{\substack{z \in \mathbb{R}_+^p, \\ \sum_{j=1}^p z_j \leq 1}} \sum_{j=1}^p \frac{x_j^2}{z_j},$$

whose minimum is uniquely attained for $z_j = |x_j| / \|x\|_1$. Similarly, we have

$$2\|x\|_1 = \min_{z \in \mathbb{R}_+^p} \sum_{j=1}^p \frac{x_j^2}{z_j} + \|z\|_1,$$

whose minimum is uniquely obtained for $z_j = |x_j|$. Thus, we can equivalently rewrite Equation (2) as

$$\min_{\substack{w \in \mathbb{R}^p, \\ (\eta^G)_{G \in \mathcal{G}} \in \mathbb{R}_+^{|\mathcal{G}|}}} \frac{1}{n} \sum_{i=1}^n \ell\big(y_i, w^\top x_i\big) + \frac{\mu}{2} \sum_{j=1}^p w_j^2 \zeta_j^{-1} + \frac{\mu}{2} \big\|(\eta^G)_{G \in \mathcal{G}}\big\|_1, \tag{17}$$

with $\zeta_j = (\sum_{G \ni j} (d_j^G)^2 (\eta^G)^{-1})^{-1}$. In the same vein, Equation (3) is equivalent to

$$\min_{\substack{w \in \mathbb{R}^p, \\ (\eta^G)_{G \in \mathcal{G}} \in \mathbb{R}_+^{|\mathcal{G}|}, \\ \sum_{G \in \mathcal{G}} \eta^G \leq 1}} \frac{1}{n} \sum_{i=1}^n \ell\big(y_i, w^\top x_i\big) + \frac{\lambda}{2} \sum_{j=1}^p w_j^2 \zeta_j^{-1}, \tag{18}$$

where $\zeta_j$ is defined as above. The reformulations Equation (17) and Equation (18) are *jointly* convex in $\{w, (\eta^G)_{G \in \mathcal{G}}\}$ and lend themselves well to a simple alternating optimization scheme between $w$ (for instance, $w$ can be computed in closed-form when the square loss is used) and $(\eta^G)_{G \in \mathcal{G}}$ (whose optimal value is always a closed-form solution). If the variables $(\eta^G)_{G \in \mathcal{G}} \in \mathbb{R}_+^{|\mathcal{G}|}$ are bounded away from zero by a smoothing parameter, the convergence of this scheme is guaranteed by standard results about block coordinate descent procedures (Bertsekas, 1999).

This first order approach is computationally appealing since it allows *warm-restart*, which can dramatically speed up the computation over regularization paths. Moreover, it does not make any assumptions on the nature of the family of groups $\mathcal{G}$.

## Appendix I. Technical Lemmas

In this last section of the appendix, we give several technical lemmas. We consider $I \subseteq \{1, \ldots, p\}$ and $\mathcal{G}_I = \{G \in \mathcal{G};\ G \cap I \neq \varnothing\} \subseteq \mathcal{G}$, that is, the set of active groups when the variables $I$ are selected.

We begin with a dual formulation of $\Omega^*$ obtained by conic duality (Boyd and Vandenberghe, 2004):

**Lemma 9** *Let $u_I \in \mathbb{R}^{|I|}$. We have*

$$\begin{aligned} (\Omega_I)^*[u_I] \quad &= \quad \min_{(\xi_I^G)_{G \in \mathcal{G}_I}} \max_{G \in \mathcal{G}_I} \|\xi_I^G\|_2 \\ &\quad s.t. \quad u_j + \sum_{G \in \mathcal{G}_I, G \ni j} d_j^G \xi_j^G = 0 \text{ and } \xi_j^G = 0 \text{ if } j \notin G. \end{aligned}$$

**Proof** By definition of $(\Omega_I)^*[u_I]$, we have

$$(\Omega_I)^*[u_I] = \max_{\Omega_I(v_I)\leq 1} u_I^\top v_I.$$

By introducing the primal variables $(\alpha_G)_{G\in\mathcal{G}_I} \in \mathbb{R}^{|\mathcal{G}_I|}$, we can rewrite the previous maximization problem as

$$(\Omega_I)^*[u_I] = \max_{\sum_{G\in\mathcal{G}_I}\alpha_G\leq 1} u_I^\top v_I, \quad \text{s.t.} \quad \forall\, G\in\mathcal{G}_I,\ \|d_I^G \circ u_{G\cap I}\|_2 \leq \alpha_G,$$

which is a second-order cone program (SOCP) with $|\mathcal{G}_I|$ second-order cone constraints. This primal problem is convex and satisfies Slater's conditions for generalized conic inequalities, which implies that strong duality holds (Boyd and Vandenberghe, 2004). We now consider the Lagrangian $\mathcal{L}$ defined as

$$\mathcal{L}(v_I,\alpha_G,\gamma,\tau_G,\xi_I^G) = u_I^\top v_I + \gamma(1-\sum_{G\in\mathcal{G}_I}\alpha_G) + \sum_{G\in\mathcal{G}_I}\begin{pmatrix}\alpha_G\\ d_I^G\circ u_{G\cap I}\end{pmatrix}^\top\begin{pmatrix}\tau_G\\ \xi_I^G\end{pmatrix},$$

with the dual variables $\{\gamma,(\tau_G)_{G\in\mathcal{G}_I},(\xi_I^G)_{G\in\mathcal{G}_I}\} \in \mathbb{R}_+\times\mathbb{R}^{|\mathcal{G}_I|}\times\mathbb{R}^{|I|\times|\mathcal{G}_I|}$ such that for all $G\in\mathcal{G}_I$, $\xi_j^G = 0$ if $j\notin G$ and $\|\xi_I^G\|_2\leq\tau_G$. The dual function is obtained by taking the derivatives of $\mathcal{L}$ with respect to the primal variables $v_I$ and $(\alpha_G)_{G\in\mathcal{G}_I}$ and equating them to zero, which leads to

$$\begin{array}{lll}\forall j\in I, & u_j + \sum_{G\in\mathcal{G}_I, G\ni j} d_j^G\xi_j^G & = 0\\ \forall G\in\mathcal{G}_I, & \gamma-\tau_G & = 0.\end{array}$$

After simplifying the Lagrangian, the dual problem then reduces to

$$\min_{\gamma,(\xi_I^G)_{G\in\mathcal{G}_I}} \gamma \quad \text{s.t.} \begin{cases}\forall j\in I, u_j + \sum_{G\in\mathcal{G}_I,G\ni j} d_j^G\xi_j^G = 0 \text{ and } \xi_j^G = 0 \text{ if } j\notin G,\\ \forall G\in\mathcal{G}_I, \|\xi_I^G\|_2\leq\gamma,\end{cases}$$

which is equivalent to the displayed result. ■

Since we cannot compute in closed-form the solution of the previous optimization problem, we focus on a different *but closely related* problem, that is, when we replace the objective $\max_{G\in\mathcal{G}_I}\|\xi_I^G\|_2$ by $\max_{G\in\mathcal{G}_I}\|\xi_I^G\|_\infty$, to obtain a *meaningful* feasible point:

**Lemma 10** *Let $u_I\in\mathbb{R}^{|I|}$. The following problem*

$$\begin{array}{ll}\min_{(\xi_I^G)_{G\in\mathcal{G}_I}} & \max_{G\in\mathcal{G}_I}\|\xi_I^G\|_\infty\\ \textit{s.t.} & u_j + \sum_{G\in\mathcal{G}_I,G\ni j} d_j^G\xi_j^G = 0 \textit{ and } \xi_j^G = 0 \textit{ if } j\notin G,\end{array}$$

*is minimized for* $(\xi_j^G)^* = -\dfrac{u_j}{\sum_{H\in j,H\in\mathcal{G}_I} d_j^H}$.

**Proof** We proceed by contradiction. Let us assume there exists $(\xi_I^G)_{G\in\mathcal{G}_I}$ such that

$$\begin{aligned}\max_{G\in\mathcal{G}_I}\|\xi_I^G\|_\infty &< \max_{G\in\mathcal{G}_I}\|(\xi_I^G)^*\|_\infty\\ &= \max_{G\in\mathcal{G}_I}\max_{j\in G}\frac{|u_j|}{\sum_{H\in j,H\in\mathcal{G}_I} d_j^H}\\ &= \frac{|u_{j_0}|}{\sum_{H\in j_0,H\in\mathcal{G}_I} d_{j_0}^H},\end{aligned}$$

where we denote by $j_0$ an argmax of the latter maximization. We notably have for all $G \ni j_0$:

$$|\xi^G_{j_0}| < \frac{|u_{j_0}|}{\sum_{H \in j_0, H \in \mathcal{G}_I} d^H_{j_0}}.$$

By multiplying both sides by $d^G_{j_0}$ and by summing over $G \ni j_0$, we get

$$|u_{j_0}| = |\sum_{G \in \mathcal{G}_I, G \ni j_0} d^G_{j_0} \xi^G_{j_0}| \leq \sum_{G \ni j_0} d^G_{j_0} |\xi^G_{j_0}| < |u_{j_0}|,$$

which leads to a contradiction. ■

We now give an upperbound on $\Omega^*$ based on Lemma 9 and Lemma 10:

**Lemma 11** *Let $u_I \in \mathbb{R}^{|I|}$. We have*

$$(\Omega_I)^*[u_I] \leq \max_{G \in \mathcal{G}_I} \left\{ \sum_{j \in G} \left\{ \frac{u_j}{\sum_{H \in j, H \in \mathcal{G}_I} d^H_j} \right\}^2 \right\}^{\frac{1}{2}}.$$

**Proof** We simply plug the minimizer obtained in Lemma 10 into the problem of Lemma 9. ■

We now derive a lemma to control the difference of the gradient of $\Omega_J$ evaluated in two points:

**Lemma 12** *Let $u_J, v_J$ be two nonzero vectors in $\mathbb{R}^{|J|}$. Let us consider the mapping $w_J \mapsto r(w_J) = \sum_{G \in \mathcal{G}_J} \frac{d^G_J \circ d^G_J \circ w_J}{\|d^G_J \circ w_J\|_2} \in \mathbb{R}^{|J|}$. There exists $z_J = t_0 u_J + (1-t_0) v_J$ for some $t_0 \in (0,1)$ such that*

$$\|r(u_J) - r(v_J)\|_1 \leq \|u_J - v_J\|_\infty \left( \sum_{G \in \mathcal{G}_J} \frac{\|d^G_J\|_2^2}{\|d^G_J \circ z_J\|_2} + \sum_{G \in \mathcal{G}_J} \frac{\|d^G_J \circ d^G_J \circ z_J\|_1^2}{\|d^G_J \circ z_J\|_2^3} \right).$$

**Proof** For $j,k \in J$, we have

$$\frac{\partial r_j}{\partial w_k}(w_J) = \sum_{G \in \mathcal{G}_J} \frac{(d^G_j)^2}{\|d^G_J \circ w_J\|_2} \mathbb{I}_{j=k} - \sum_{G \in \mathcal{G}_J} \frac{(d^G_j)^2 w_j}{\|d^G_J \circ w_J\|_2^3} (d^G_k)^2 w_k,$$

with $\mathbb{I}_{j=k} = 1$ if $j = k$ and 0 otherwise. We then consider $t \in [0,1] \mapsto h_j(t) = r_j(t u_J + (1-t) v_J)$. The mapping $h_j$ being continuously differentiable, we can apply the mean-value theorem: there exists $t_0 \in (0,1)$ such that

$$h_j(1) - h_j(0) = \frac{\partial h_j(t)}{\partial t}(t_0).$$

We then have

$$\begin{aligned} |r_j(u_J) - r_j(v_J)| &\leq \sum_{k \in J} \left| \frac{\partial r_j}{\partial w_k}(z) \right| |u_k - v_k| \\ &\leq \|u_J - v_J\|_\infty \left( \sum_{G \in \mathcal{G}_J} \frac{(d^G_j)^2}{\|d^G_J \circ z_J\|_2} + \sum_{k \in J} \sum_{G \in \mathcal{G}_J} \frac{(d^G_j)^2 |z_j|}{\|d^G_J \circ z_J\|_2^3} (d^G_k)^2 |z_k| \right), \end{aligned}$$

which leads to

$$\|r(u_J) - r(v_J)\|_1 \leq \|u_J - v_J\|_\infty \left( \sum_{G \in \mathcal{G}_J} \frac{\|d_J^G\|_2^2}{\|d_J^G \circ z_J\|_2} + \sum_{G \in \mathcal{G}_J} \frac{\|d_J^G \circ d_J^G \circ z_J\|_1^2}{\|d_J^G \circ z_J\|_2^3} \right).$$

■

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

**Lemma 13** *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 Problem (2).

**Lemma 14** *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*

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

*with $\hat{J}$ the hull of $\{j \in \{1,\ldots,p\}, \hat{w}_j \neq 0\}$ 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

$$\forall u \in \mathbb{R}^p,\ u^\top \nabla L(\hat{w}) + \mu\, u_{\hat{J}}^\top \hat{r}_{\hat{J}} + \mu(\Omega_{\hat{J}}^c)[u_{\hat{J}^c}] \geq 0. \tag{19}$$

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 $u_{\hat{J}}^\top \hat{r}_{\hat{J}}$ gives for all $u \in \mathbb{R}^p$

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

Combined with Equation (19), 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 15** *Let $A$ and $B$ be a partition of $\{1,\dots,p\}$, that is, $A \cap B = \varnothing$ 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*

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

## References


R. J. Adler. *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes*. IMS, 1990.

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

F. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in Neural Information Processing Systems*, 2008a.

F. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008b.

F. Bach. Bolasso: model consistent Lasso estimation through the bootstrap. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2008c.

F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2009.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Convex optimization with sparsity-inducing norms. In *Optimization for Machine Learning*. MIT press, 2011. To appear.

R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56:1982–2001, 2010.

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

D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

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

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

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

P. J. Cameron. *Combinatorics: Topics, Techniques, Algorithms*. Cambridge University Press, 1994.

V. Cevher, M. F. Duarte, C. Hegde, and R. G. Baraniuk. Sparse signal recovery using markov random fields. In *Advances in Neural Information Processing Systems*, 2008.

F. R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer, 2010.

N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2006.

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

C. Dossal. A necessary and sufficient condition for exact recovery by l1 minimization. Technical report, HAL-00164738:1, 2007.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32 (2):407–451, 2004.

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

J. J. Fuchs. Recovery of exact sparse representations in the presence of bounded noise. *IEEE Transactions on Information Theory*, 51(10):3601–3608, 2005.

A. Gramfort and M. Kowalski. Improving M/EEG source localization with an inter-condition sparse prior. In *IEEE International Symposium on Biomedical Imaging*, 2009.

H. Harzallah, F. Jurie, and C. Schmid. Combining efficient object localization and image classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2009.

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

J. Huang and T. Zhang. The benefit of group sparsity. *Annals of Statistics*, 38(4):1978–2004, 2010.

J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

L. Jacob, G. Obozinski, and J.-P. Vert. Group Lasso with overlaps and graph Lasso. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010a.

R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010b.

R. Jenatton, A. Gramfort, V. Michel, G. Obozinski, F. Bach, and B. Thirion. Multi-scale mining of fMRI data with hierarchical structured sparsity. In *International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, 2011a.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12:2297–2334, 2011b.

K. Kavukcuoglu, M. A. Ranzato, R. Fergus, and Y. Le-Cun. Learning invariant features through topographic filter maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

S. Kim and E. P. Xing. Tree-guided group Lasso for multi-task regression with structured sparsity. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems*, 2007.

E. Levina, A. Rothman, and J. Zhu. Sparse estimation of large covariance matrices via a nested Lasso penalty. *Annals of Applied Statistics*, 2(1):245–263, 2008.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(1):19–60, 2010a.

J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Network flow algorithms for structured sparsity. In *Advances in Neural Information Processing Systems*, 2010b.

J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Convex and network flow optimization for structured sparsity. Technical report, Preprint arXiv:1104.1872, 2011. To appear in Journal Machine Learning Research.

A. F. T. Martins, N. A. Smith, P. M. Q. Aguiar, and M. A. T. Figueiredo. Structured sparsity in structured prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011.

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

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society. Series B*, 72(4):417–473, 2010.

C. A. Micchelli and M. Pontil. Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6(2):1099, 2006.

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

S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, 2009.

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

G. Obozinski, B. Taskar, and M. I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, pages 1–22, 2009.

M. R. Osborne, B. Presnell, and B. A. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9:319–337, 2000.

A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

F. Rapaport, E. Barillot, and J.-P. Vert. Classification of arrayCGH data using fused SVM. *Bioinformatics*, 24(13):i375–i382, Jul 2008.

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

S. Rosset, J. Zhu, and T. Hastie. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5:941–973, 2004.

V. Roth and B. Fischer. The group-Lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

M. Schmidt and K. Murphy. Convex structure learning in log-linear models: Beyond pairwise potentials. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

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

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, pages 267–288, 1996.

K. C. Toh, M. J. Todd, and R. H. Tütüncü. SDPT3–a MATLAB software package for semidefinite programming, version 1.3. *Optimization Methods and Software*, 11(1):545–581, 1999.

R. H. Tütüncü, K. C. Toh, and M. J. Todd. Solving semidefinite-quadratic-linear programs using SDPT3. *Mathematical Programming*, 95(2):189–217, 2003.

G. Varoquaux, R. Jenatton, A. Gramfort, G. Obozinski, B. Thirion, and F. Bach. Sparse structured dictionary learning for brain resting-state activity modeling. In *NIPS Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions*, 2010.

M. J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using $\ell_1$-constrained quadratic programming. *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.

Z. J. Xiang, Y. T. Xi, U. Hasson, and P. J. Ramadge. Boosting with spatial regularization. In *Advances in Neural Information Processing Systems*, 2009.

G. X. Yuan, K. W. Chang, C. J. Hsieh, and C. J. Lin. Comparison of optimization methods and software for large-scale l1-regularized linear classification. *Journal of Machine Learning Research*, 11:3183–3234, 2010.

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

T. Zhang. Some sharp performance bounds for least squares regression with l1 regularization. *Annals of Statistics*, 37(5A):2109–2144, 2009.

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

P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, 37(6A):3468–3497, 2009.

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.