Active Set Algorithm for Structured Sparsity-Inducing Norms

Rodolphe Jenatton[1]   rodolphe.jenatton@inria.fr
Jean-Yves Audibert[1,2]   audibert@certis.enpc.fr
Francis Bach[1]   francis.bach@inria.fr

Abstract

We consider the empirical risk minimization problem for linear supervised learning, with regularization by structured sparsity-inducing norms. These are defined as sums of Euclidean norms on certain subsets of variables, extending the usual ℓ_1-norm and the group ℓ_1-norm by allowing the subsets to overlap. This leads to a specific set of allowed nonzero patterns for the solutions of such problems. We first explore the relationship between the groups defining the norm and the resulting nonzero patterns. In particular, we show how geometrical information about the variables can be encoded by our regularization. We finally present an active set algorithm to efficiently solve the corresponding minimization problem.

1 Introduction

Regularization by the ℓ_1-norm is now a widespread tool in machine learning, statistics and signal processing: it allows linear variable selection in potentially high dimensions, with both efficient algorithms [1] and well-developed theory for generalization properties and variable selection consistency [2, 3, 4].

However, the ℓ_1-norm cannot easily encode prior knowledge about the patterns of nonzero coefficients ("nonzero patterns") induced in the solution, since they are all theoretically possible. Group ℓ_1-norms [5, 6, 7] 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 [8, 9].

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. In particular, when the variables are organized in a 2-dimensional grid, our regularization leads to the selection of convex nonzero patterns. We then present an efficient active set algorithm that scales well to high dimensions, by exploiting the sparsity and the structure of the groups. Finally, the scalability of the algorithm is illustrated on synthetic data.

Notation.   For $x \in \mathbb{R}^p$ and $q \in [1, \infty)$, we denote by $\|x\|_q$ its ℓ_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. Furthermore, for two vectors x and y in $\mathbb{R}^p$, we denote by $x \circ y = (x_1 y_1, \ldots, x_p y_p)^\top \in \mathbb{R}^p$ 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 $L(w) = \frac{1}{n} \sum_{i=1}^n \ell\left(y_i, w^\top x_i\right)$, where $\ell : \mathcal{Y} \times \mathbb{R} \mapsto \mathbb{R}^+$ is a *loss function*. We assume that ℓ is *convex and continuously differentiable* with respect to the second parameter. Typical examples of loss functions are the square loss for least squares regression, i.e., $\ell(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$ with $y \in \mathbb{R}$, and the logistic loss $\ell(y, \hat{y}) = \log(1 + e^{-y\hat{y}})$ for logistic regression, with $y \in \{-1, 1\}$.

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

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

where $(d^G)_{G \in \mathcal{G}}$ is a collection of p-dimensional vectors such that $d_j^G > 0$ if $j \in G$ and $d_j^G = 0$ otherwise. For specific choices of $\mathcal{G}$, Ω leads to standard sparsity-inducing norms. For example, when $\mathcal{G}$ is the set of all singletons, Ω is the usual ℓ_1 norm (assuming that all the weights are equal to 1). We study the following regularized problem:

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

where $\mu \geq 0$ is a regularization parameter. Regularizing by linear combinations of (non-squared) ℓ_2-norms is known to induce sparsity in $\hat{w}$ [8]; our grouping leads to specific patterns that we describe in the next section.

3 Groups and Sparsity Patterns

3.1 Sparsity patterns

The regularization term $\Omega(w) = \sum_{G \in \mathcal{G}} \|d^G \circ w\|_2$ is a mixed (ℓ_1, ℓ_2)-norm [8]. At the group level, it behaves like an ℓ_1-norm and therefore, Ω 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 ℓ_2-norm does not promote sparsity. Intuitively, some of the vectors w_G associated with certain groups G will be exactly equal to zero, leading to a set of zeros which is the union of these groups G in $\mathcal{G}$.

It can actually be proved that the previous intuition is true, in the sense that the solutions to a least-squares regression problem regularized by Ω have a zero pattern which indeed belongs to the *union-closure* of $\mathcal{G}$.

Thus, through the choice of $\mathcal{G}$, we can incorporate prior knowledge on the form of *allowed* zero patterns (or equivalently nonzero patterns).

Due to space limitation, we do not present in details how we can go back and forth, from groups to patterns. We rather focus on examples of sets of groups $\mathcal{G}$ for which the sparsity-inducing effect of Ω has a clear geometrical interpretation.

3.2 Examples

Sequences. Given p variables organized in a sequence, the nonzero patterns allowed by the set of groups which are intervals $[1, k]_{k \in \{1, \ldots, p-1\}}$ and $[k, p]_{k \in \{2, \ldots, p\}}$ are just the contiguous segments of the sequence, with $|\mathcal{G}| = O(p)$ (see Figure 1).
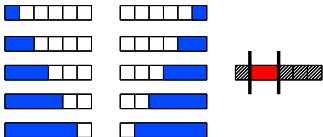
Figure 1: (Left) The set of blue groups to penalize in order to select contiguous patterns in a sequence. (Right) In red, an example of such a pattern.

2

Two-dimensional grids. Similarly, in the case of a 2-dimensional grid where the set of groups $\mathcal{G}$ is the set of all horizontal and vertical half-spaces (see Figure 2), Ω sets to zero some entire horizontal and vertical half-spaces of the grid, and therefore induces rectangular nonzero patterns. Note that a broader set of convex patterns can be obtained by adding in $\mathcal{G}$ half-planes with other orientations. In the experiments, we consider planes with angles that are multiples of $\frac{\pi}{4}$.
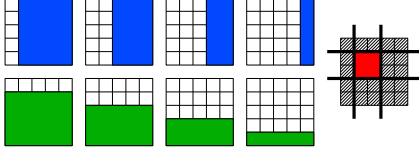

Figure 2: Vertical and horizontal groups: (Left) the set of blue and green groups with their (not displayed) complements to penalize in order to select rectangles. (Right) In red, an example of recovered rectangular pattern in this setting.

4 Optimization and Active Set Algorithm

The minimization problem Eq. (2) is *convex* and *nonsmooth*. For moderate values of p, one may obtain a solution for Eq. (2) using generic toolboxes for second-order cone programming (SOCP) whose time complexity is equal to $O(p^{3.5} + |\mathcal{G}|^{3.5})$ [10], which is not appropriate when p or $|\mathcal{G}|$ are large.

We present in this section an *active set algorithm* (Algorithm 1) that finds a solution for Eq. (2) by considering increasingly larger active sets and checking global optimality at each step, with total complexity in $O(p^{1.75})$. Here, the sparsity prior is exploited for computational advantages.

4.1 Optimality Conditions: from Reduced Problems to Full Problems

It is simpler to derive the algorithm for the following regularized optimization problem which has the same solution set as the regularized problem of Eq. (2) when μ and λ are allowed to vary [11, see Section 3.2]:

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

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 Eq. (3). We denote by $L(w) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, w^\top x_i\right)$ the empirical risk (which is by assumption convex and continuously differentiable) and by L^* its *Fenchel-conjugate*. 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 $L_J^*(\kappa_J) = L^*(\tilde{\kappa})$ for $\tilde{\kappa}_J = \kappa_J$ and $\tilde{\kappa}_{J^c} = 0$.

For a potential active set $J \subset \{1, \ldots, p\}$, we denote by $\mathcal{G}_J$ the set of active groups, i.e., the set of groups $G \in \mathcal{G}$ such that $G \cap J \neq \varnothing$. We consider the reduced norm Ω_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* $\Omega_J^*(\kappa_J) = \max_{\Omega_J(w_J) \leq 1} w_J^\top \kappa_J$. The next proposition gives the optimization problem dual to the reduced problem (Eq. (4) below):

Proposition 4.1 (Dual Problems). *Let $J \subseteq \{1, \ldots, p\}$. The following two problems*

$$\begin{cases} \min_{w_J \in \mathbb{R}^{|J|}} & L_J(w_J) + \frac{\lambda}{2} \left[\Omega_J(w_J)\right]^2, \\ \max_{\kappa_J \in \mathbb{R}^{|J|}} & -L_J^*(-\kappa_J) - \frac{1}{2\lambda} \left[\Omega_J^*(\kappa_J)\right]^2, \end{cases} \tag{4}$$

are dual to each other and the pair of primal-dual variables $\{w_J, \kappa_J\}$ is optimal if and only if we have

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

3

The duality gap of the previous optimization problem is

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

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

which is a sum of two nonnegative terms, the nonnegativity coming from the Fenchel-Young inequality [11, Proposition 3.3.4]. We can think of this duality gap as the sum of two duality gaps, relative to L_J and Ω_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 Eq. (3), we pad w_J with zeros on J^c to define w, compute $\kappa = -\nabla L(w)$, which is such that $\kappa_J = -\nabla L_J(w_J)$, and get a duality gap for the full problem equal to

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

Computing this gap requires solving an optimization problem which is as hard as the original one, prompting the need for upper and lower bounds on Ω^* (see Propositions 4.2 and 4.3 for more details).

4.2 Active set algorithm

The nonzero patterns allowed by the norm Ω are naturally organized in directed acyclic graph (DAG), ordered by inclusion. In light of Section 3.1, we can interpret the active set algorithm as a walk through this DAG. 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 1. The groups that are exactly at the boundaries of the active set (referred to as the *fringe groups*) are $\mathcal{F}_J = \{G \in (\mathcal{G}_J)^c ; \nexists G' \in (\mathcal{G}_J)^c, G \subseteq G'\}$, i.e., the groups that are not contained by any other inactive groups.

In simple settings, e.g., when $\mathcal{G}$ is the set of rectangular groups, the correspondance 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 3).
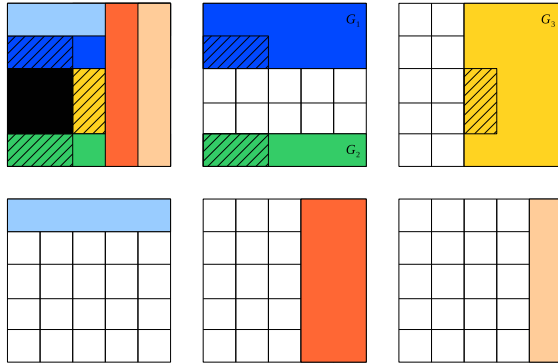

Figure 3: The active set (black) and the candidate patterns of variables, i.e. the variables in $K\backslash J$ (hatched in black) that can become active. The fringe groups are exactly the groups that have the hatched areas (i.e., here we have $\mathcal{F}_J = \bigcup_{K\in\Pi_{\mathcal{P}}(J)} \mathcal{G}_K\backslash\mathcal{G}_J = \{G_1, G_2, G_3\}$).

We now present the optimality conditions that monitor the progress of Algorithm 1 :

Proposition 4.2 (Necessary condition)**.** *If w is optimal for the full problem in Eq. (3), then*

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

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

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

then w is a solution for Eq. (3) whose duality gap is less than $\varepsilon \geq 0$.

4

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 Ω) to the condition $\|\nabla L(w)_{J^c}\|_\infty \leqslant -w^\top \nabla L(w)$, which is the usual optimality condition [12]. 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 1, it suffices to choose the pattern $K \in \Pi_{\mathcal{P}}(J)$ that violates the condition most.

The heuristics for the sufficient condition (S_ε), denoted by (*), implies to go from groups to variables. We simply consider the group $G \in \mathcal{F}_J$ that violates the sufficient condition 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.

Algorithm 1 Active set algorithm

    Input: Data $\{(x_i, y_i),\, i = 1, \dots, n\}$, regularization parameter λ,
             Duality gap precision ε, maximum number of variables s.
    Output: Active set J, loading vector $\hat{w}$.
    Intialization: $J = \{\varnothing\}$, $\hat{w} = 0$.
    while $\big(\,(N)$ is not satisfied $\big)$ **and** $\big(\,|J| \leqslant s\,\big)$ **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** $\big(\,|J| \leqslant s\,\big)$ **do**
         Update J according to the heuristics (*).
         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

Convergence of the active set algorithm. The procedure described in Algorithm 1 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 with a nonzero pattern smaller than the optimal one.

Otherwise, the procedure always converges to an optimal solution, either (1) by validating both the necessary and sufficient conditions (see Propositions 4.2 and 4.3), ending up with fewer than p active variables and a precision of (at least) ε, or (2) by running until the p variables become active, the precision of the solution being given by the underlying solver.

Algorithmic complexity. We analyse 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 3.2. We further assume that $\mathcal{G}$ is sorted by cardinality and by orientation, so that computing $\mathcal{F}_J$ costs $O(1)$.

Thus, if the number of active variables is upper bounded by $s \ll p$ (which is a reasonable assumption if our target is actually sparse), the time complexity of Algorithm 1 has a leading term in $O(sp|\mathcal{G}| + s \max_{J \in \mathcal{P}, |J| \leqslant s} |\mathcal{G}_J|^{3.5})$, which is much better than $O(p^{3.5} + |\mathcal{G}|^{3.5})$, without an active set method. In the example of the 2-dimensional grid (see Section 3.2), we have $|\mathcal{G}| = O(\sqrt{p})$ and a total complexity in $O(sp^{1.75})$.

5 Experiments

We compare the time complexity of 1) the active set algorithm calling upon the SOCP solver with 2) the only 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 $\{100, 225, 400, 900, 1600, 2500\}$. The sparse structured target is a diamond-shaped convex pattern, whose position is randomly selected on a square 2-dimensional grid. Thus, we consider sets of groups $\mathcal{G}$ such a those presented in Section 3.2. We display the median CPU time based on 250 runs (the, not displayed, mean CPU time curve gives a similar result).

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 1), which is a rather conservative choice. We show on

Figure 4 that we considerably lower the computational burden. 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 1).
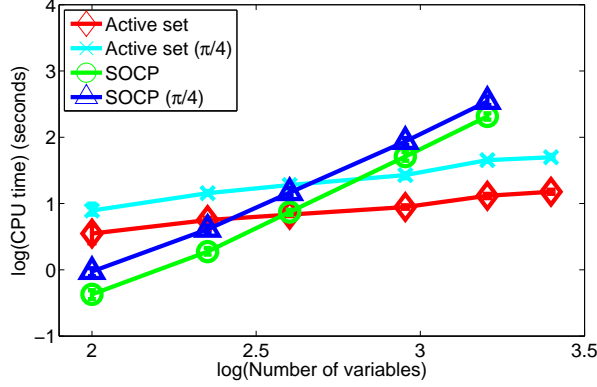

Figure 4: Computational benefit of the active set algorithm: CPU time (in seconds) versus the number of variables p, displayed in $\log$-$\log$ scale. Two sets of groups $\mathcal{G}$ are considered, the rectangular groups with or without the $\pm\pi/4$-groups (denoted by $(\pi/4)$ in the legend). Due to the computational burden, we could not obtain the SOCP's results for $p = 2500$.

6 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 ℓ_2-norms of possibly overlapping groups of variables. We address this nonsmooth convex optimization problem by an active set method that improves upon the scalability with respect to the number of variables. This improvement is illustrated on a synthetic example.

Note that a detailed description of this work with additional experiments can be found in [13].

References

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

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

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

[4] T. Zhang. Some sharp performance bounds for least squares regression with ℓ_1 regularization. *Annals of Statistics*, 2009. To appear.

[5] 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.

[6] V. Roth and B. Fischer. The group-Lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *Proc. of the 25th International Conf. on Machine learning*, 2008.

[7] J. Huang and T. Zhang. The benefit of group sparsity. Technical report, arXiv: 0901.2962, 2009.

[8] P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 2008. To appear.

[9] F. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Adv. in Neural Information Processing Systems*, 2008.

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

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

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

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