Sparsity by Worst-Case Quadratic Penalties

Yves Grandvalet

Sorbonne universités, Université de technologie de Compiègne, CNRS,
Heudiasyc UMR 7253, CS 60 319, 60 203 Compiègne cedex, France

Julien Chiquet, Christophe Ambroise

Laboratoire de Mathématiques et Modèlisation d'Évry val d'Essonne
CNRS UMR 8071 & Université d'Évry 23, Boulevard de France
91000 Évry, FRANCE

Abstract

This paper proposes a new robust regression interpretation of sparse penalties such as the elastic net and the group-lasso. Beyond providing a new viewpoint on these penalization schemes, our approach results in a unified optimization strategy. Our evaluation experiments demonstrate that this strategy, implemented on the elastic net, is computationally extremely efficient for small to medium size problems. Our accompanying software solves problems at machine precision in the time required to get a rough estimate with competing state-of-the-art algorithms.

Keywords: sparse regression, robust optimization, lasso, elastic net, group-lasso

1. Introduction

Inferential statistics aim at drawing conclusions from data and from some kind of assumptions or prior information about the underlying distribution. It is well known that processes where data guide the choice of assumptions can lead to paradoxes, resulting from overfitting issues, in particular in the large dimensional setup when data are used to select explicative variables:

Email addresses: `yves.grandvalet@utc.fr` (Yves Grandvalet),
`julien.chiquet@genopole.cnrs.fr` (Julien Chiquet),
`christophe.ambroise@genopole.cnrs.fr` (Christophe Ambroise)

the seeming explanatory power of weakly relevant variables can be important when the number of variables is similar to the number of data points (Freedman, 1983; Ambroise and McLachlan, 2002). In this context, we present here an unusual reformulation of variable selection methods based on sparsity-inducing penalties: we show that these methods can be interpreted as adaptive penalties, where adaptivity refers to the choice of the eventual penalty from data. Contrary to the available reformulations so far, ours shows that the adaptation of the sparsity-inducing penalty to data follows a disagreement principle, where the least-favorable penalty is selected from data.

We believe that our new viewpoint can be instrumental in unified analyses of algorithms and their derived estimators, and we show here two such examples: first, we provide a generic algorithm for solving the data fitting problem; then, we derive the general form of an optimality gap that may be used to monitor convergence. Our experimental section illustrates the effectiveness of the generic algorithm motivated by our interpretation, when instanced on the elastic-net estimator: the algorithm, which relies on solving linear systems is accurate, and computationally efficient up to medium scale problems (thousands of variables). As a side experimental result, we show that solving problems with high precision, as with the proposed approach, benefits to the performances, either measured in terms of prediction accuracy or in terms of support error rate.

2. Adaptive Quadratic Penalties

2.1. Background

We consider the linear regression model

$$Y = X\boldsymbol{\beta}^\star + \varepsilon \ , \tag{1}$$

where Y is a continuous response variable, $X = (X_1, \ldots, X_p)$ is a vector of p predictor variables, $\boldsymbol{\beta}^\star$ is the vector of unknown parameters and ε is a zero-mean Gaussian error variable with variance σ^2. We will assume throughout this paper that $\boldsymbol{\beta}^\star$ has few non-zero coefficients.

The estimation and inference of $\boldsymbol{\beta}^\star$ is based on training data, consisting of a vector $\mathbf{y} = (y_1, \ldots, y_n)^\intercal$ for responses and a $n \times p$ design matrix $\mathbf{X}$ whose jth column contains $\mathbf{x}_j = (x_j^1, \ldots, x_j^n)^\intercal$, the n observations for variable X_j. For clarity, we assume that both $\mathbf{y}$ and $\{\mathbf{x}_j\}_{j=1,\ldots,p}$ are centered so as to eliminate the intercept from fitting criteria.

Penalization methods that build on the ℓ_1-norm, referred to as *Lasso* procedures (Least Absolute Shrinkage and Selection Operator), are now widely used to tackle simultaneously variable estimation and selection in sparse problems. They define a shrinkage estimator of the form

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda \|\boldsymbol{\beta}\| \quad , \tag{2}$$

where $\|\cdot\|_2$ is the Euclidean norm and $\|\cdot\|$ is an arbitrary norm, chosen to induce some assumed sparsity pattern (typically ℓ_1 or $\ell_{c,1}$ norms, where $c \in (1, \infty]$).

The existence of computationally efficient optimization procedures plays an important role in the popularity of these methods. Though various general-purpose convex optimization solvers could be used (Boyd and Vandenberghe, 2004), exploiting the structure of the regularization problem, and especially the sparsity of solutions, is essential in terms of computational efficiency. Bach et al. (2012) provided an overview of the families of techniques specifically designed for solving this type of problems: proximal methods, coordinate descent algorithms, reweighted-ℓ_2 algorithms, working-set methods. Stochastic gradient methods (Moulines and Bach, 2011), the Frank-Wolfe algorithm (Lacoste-Julien et al., 2012) or ADMM (Alternating Direction Method of Multipliers, Boyd et al., 2011) have also recently gained in popularity to the resolution of sparse problems.

We present below a new formulation of Problem (2) that motivates an algorithm that may seem reminiscent of reweighted-ℓ_2 algorithms, but which is in fact more closely related to working-set methods. As for reweighted-ℓ_2 algorithms, our proposal is based on the reformulation of the sparsity-inducing penalty in terms of penalties that are simpler to handle (linear or quadratic). However, whereas reweighted-ℓ_2 algorithms rely on a variational formulation of the sparsity-inducing norm that ends up in an augmented minimization problem, our proposal is rooted in the duality principle, eventually leading to a minimax problem that lends itself to a working-set algorithm that will be presented in Section 4.

2.2. Dual Norms

When the sparsity-inducing penalty is a norm, its sublevel sets can always be defined as the intersection of linear or quadratic sublevel sets. In other terms, if the optimization problem is written in the form of a constrained optimization problem with inequality constraints pertaining to the

penalty, then, the feasible region can be defined as the intersection of linear or quadratic regions. This fact, which is illustrated in Figures 1 and 2, stems from the definition of dual norms:

$$\|\boldsymbol{\beta}\| = \max_{\boldsymbol{\gamma} \in \mathcal{B}_*} \boldsymbol{\gamma}^\mathsf{T} \boldsymbol{\beta} \ ,$$

where $\mathcal{B}_*$ is the unit ball centered at the origin defined from the dual norm $\|\cdot\|_*$, $\mathcal{B}_* = \{\boldsymbol{\gamma} \in \mathbb{R}^p : \|\boldsymbol{\gamma}\|_* \leq 1\}$. Using this definition, Problem (2) can be reformulated as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \max_{\boldsymbol{\gamma} \in \mathcal{B}_*} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda \boldsymbol{\gamma}^\mathsf{T} \boldsymbol{\beta} \ . \tag{3}$$

Technically, this formulation is the primal form of the original Problem (2) using the coupling function defined by the dual norm (see e.g. Gilbert, 2016; Bonnans et al., 2006). It is interesting in the sense that the problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda \boldsymbol{\gamma}^\mathsf{T} \boldsymbol{\beta}$$

is simple to solve for any value of $\boldsymbol{\gamma}$, since it only requires solving a linear system. The problem

$$\begin{aligned} \hat{\boldsymbol{\gamma}} &= \arg\max_{\boldsymbol{\gamma} \in \mathcal{B}_*} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda \boldsymbol{\gamma}^\mathsf{T} \boldsymbol{\beta} \\ &= \arg\max_{\boldsymbol{\gamma} \in \mathcal{B}_*} \boldsymbol{\gamma}^\mathsf{T} \boldsymbol{\beta} \end{aligned}$$

is usually straightforward to solve. Besides the sparsity of $\hat{\boldsymbol{\beta}}$, the overall efficiency of our algorithm relies also on the fact that $\hat{\boldsymbol{\gamma}}$ is often invariant with respect to large changes in $\boldsymbol{\beta}$ for the penalties we are interested in, typically taking a finite number of possible values. This number typically increases exponentially in p, but, with the working-set strategy, the number of configuration actually visited usually grows linearly with the number of non-zero coefficients in the solution $\hat{\boldsymbol{\beta}}$.

2.3. Relations with Other Methods

The expansion in dual norm expressed in Problem (3) bears some similarities with the first step of the derivation of very general duality schemes, such as Fenchel duality or Lagrangian duality. It is however dedicated to the category of problems expressed as in (2), thereby offering an interesting

novel view of this category of problems. In particular, it provides geometrical insights on these methods and a generic algorithm for computing solutions. The associated algorithm, that relies on solving linear systems is accurate, and efficient up to medium scale problems (thousands of variables).

3. Assumptions on the Spurious Regression Coefficients

Our framework is amenable to many variations. Here, we simply present two examples following the same pattern: assuming a given regularity on the regression coefficients $\boldsymbol{\beta}^{\star}$, we consider the adversarial dual assumption on the spurious coefficients $\boldsymbol{\gamma}$. When the initial regularity conditions on $\boldsymbol{\beta}^{\star}$ are expressed by ℓ_1 or ℓ_{∞} norms, this process results in uncertainty sets $\mathcal{D}_{\boldsymbol{\gamma}}$ which are convex polytopes that are easy to manage when solving Problem (**??**), since they can be defined as the convex hulls of a finite number possible perturbations.

The two sparsity-inducing penalizers presented below have a grouping effect. The elastic net implements this grouping without predefining the group structure: strongly correlated predictors tend to be in or out of the model together (Zou and Hastie, 2005). The $\ell_{\infty,1}$ group-Lasso that is presented subsequently is based on a prescribed group structure and favors regression coefficients with identical magnitude within activated groups.

3.1. Elastic Net

As an introductory example, let us consider the regularity assumption stating that the ℓ_1-norm of $\boldsymbol{\beta}^{\star}$ should be small:

$$\mathcal{H}_{\boldsymbol{\beta}^*}^{\text{Lasso}} = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_1 \leq \eta_{\beta}\} \quad .$$

The dual assumption is that the ℓ_{∞}-norm of $\boldsymbol{\gamma}$ should be controlled, say:

$$\mathcal{D}_{\boldsymbol{\gamma}}^{\text{Lasso}} = \left\{\boldsymbol{\gamma} \in \mathbb{R}^p : \sup_{\boldsymbol{\beta} \in \mathcal{H}_{\boldsymbol{\beta}^*}^{\text{Lasso}}} \boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{\beta} \leq 1\right\}$$
$$= \{\boldsymbol{\gamma} \in \mathbb{R}^p : \|\boldsymbol{\gamma}\|_{\infty} \leq \eta_{\gamma}\}$$
$$= \mathbf{conv}\left\{\{-\eta_{\gamma}, \eta_{\gamma}\}^p\right\} \quad,$$

where $\eta_{\gamma} = 1/\eta_{\beta}$ and **conv** denotes convex hull, so that Problem (**??**) reads:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \max_{\boldsymbol{\gamma} \in \{-\eta_{\gamma}, \eta_{\gamma}\}^p} \left\{\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2^2\right\}$$
$$\Leftrightarrow \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + 2\lambda\eta_{\gamma}\|\boldsymbol{\beta}\|_1 + \lambda\|\boldsymbol{\beta}\|_2^2 \quad, \tag{4}$$
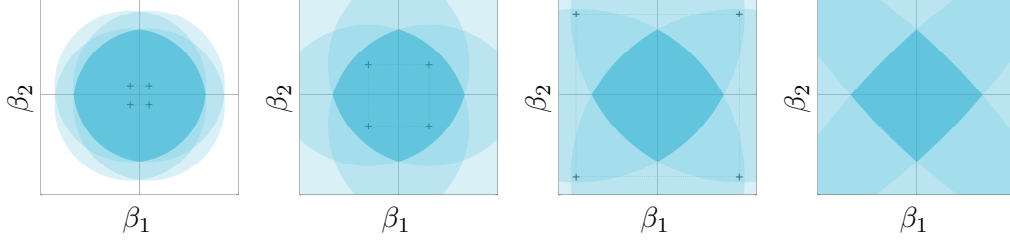
Figure 1: Sublevel sets for elastic net penalties (represented by the darker colored patches). Each set is defined as the intersection of the the Euclidean balls (represented by the lighter color patches) whose centers are represented by crosses.

which is recognized as an elastic net problem. When η_γ is null, we recover ridge regression, and when the magnitude of η_γ grows, the problem approaches a Lasso problem. A 2D pictorial illustration of this evolution is given in Figure 1, where the shape of the uncertainty set $\mathcal{D}_\gamma$ is the convex hull of the points located at $(\pm\eta_\gamma, \pm\eta_\gamma)^\intercal$, which are identified by the cross markers. Then, the sublevel set $\{\boldsymbol{\beta} : \max_{\gamma \in \mathcal{D}_\gamma} \|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2^2 \leq t\}$ is simply defined as the intersection of the four sublevel sets $\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2^2 \leq t\}$ for $\boldsymbol{\gamma} = (\pm\eta_\gamma, \pm\eta_\gamma)^\intercal$, which are Euclidean balls centered at these $\boldsymbol{\gamma}$ values.

3.2. Group-Lasso

We consider here the $\ell_{\infty,1}$ variant of the group-Lasso, which was first proposed by Turlach et al. (2005) to perform variable selection in the multiple response setup, which is first detailed before introducing the general situation. Following the previous example, we now consider the regularity assumption stating that the ℓ_∞-norm of $\boldsymbol{\beta}^\star$ should be small:

$$\mathcal{H}_{\boldsymbol{\beta}^*}^{\mathrm{Max}} = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_\infty \leq \eta_\beta\} \quad .$$

The dual assumption is that the ℓ_1-norm of $\boldsymbol{\gamma}$ should be controlled:

$$\begin{aligned}
\mathcal{D}_{\boldsymbol{\gamma}}^{\mathrm{Max}} &= \left\{\boldsymbol{\gamma} \in \mathbb{R}^p : \sup_{\boldsymbol{\beta} \in \mathcal{H}_{\boldsymbol{\beta}^*}^{\mathrm{Max}}} \boldsymbol{\gamma}^\intercal \boldsymbol{\beta} \leq 1\right\} \\
&= \{\boldsymbol{\gamma} \in \mathbb{R}^p : \|\boldsymbol{\gamma}\|_1 \leq \eta_\gamma\} \\
&= \mathbf{conv}\left\{\eta_\gamma \mathbf{e}_1^p, \ldots, \eta_\gamma \mathbf{e}_p^p, -\eta_\gamma \mathbf{e}_1^p, \ldots, -\eta_\gamma \mathbf{e}_p^p\right\} \quad ,
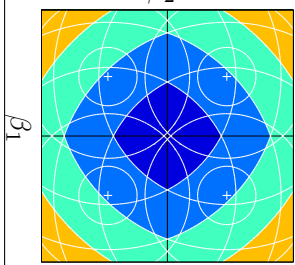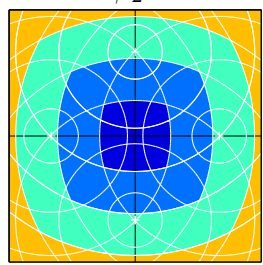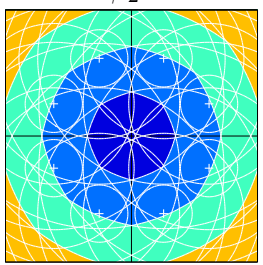\end{aligned}$$

	Oscar	Group ℓ_∞	Elastic Net
Norm	$\{\beta \in \mathbb{R}^p : \|\beta\|_1 + c\sum_{j<k}\max(|\beta_j|,|\beta_k|) \leq \eta_\beta\}$	$\{\beta \in \mathbb{R}^p : \sum_{k=1}^K \|\beta_{\mathcal{G}_k}\|_\infty \leq \eta_\beta\}$	$\{\beta \in \mathbb{R}^p : \|\beta\|_1 \leq \eta_\beta\}$
Dual Norm	$\mathcal{D}_\gamma = \left\{ \gamma \in \mathbb{R}^p \mid \gamma = \begin{pmatrix} \alpha_1 1 \\ \alpha_2(c+1) \\ \alpha_3(2c+1) \\ \vdots \\ \alpha_p(p-1)c+1 \end{pmatrix}, \right.$ $\left. c \in \mathbb{R}^+, \|(\alpha_1, \cdots, \alpha_p)\|_\infty \leq \eta_\alpha \right\}$ and the permutation matrix $P_\beta = \left\{ \mathbb{I}_{(\text{rank}(|\beta|_{(i)})=i)} \right\}_{i=1\cdots p, j=1\cdots p}$	$\mathcal{D}_\gamma^{\ell_\infty,1} = \left\{ \gamma \in \mathbb{R}^p : \sup_{\beta \in \mathcal{H}_{\beta^*}^{\ell_\infty,1}} \gamma^T \beta \leq 1 \right\}$ $= \left\{ \gamma \in \mathbb{R}^p : \sum_{k=1}^K \|\gamma_{\mathcal{G}_k}\|_1 \leq \eta_\gamma \right\}$ $= \mathbf{conv}\{\eta_\gamma e_1^{p_1}, \ldots, \eta_\gamma e_1^{p_1},$ $-\eta_\gamma e_1^{p_1}, \ldots, -\eta_\gamma e_1^{p_1}, \times \{\eta_\gamma e_1^{p_K}, \ldots, \eta_\gamma e_{p_K}^{p_K},$ $-\eta_\gamma e_1^{p_K}, \ldots, -\eta_\gamma e_{p_K}^{p_K}\}\},$ where $e_j^{p_k}$ is the jth element of the canonical basis of $\mathbb{R}^{p_k}$	$\mathcal{D}_\gamma^{\text{Lasso}} = \left\{ \gamma \in \mathbb{R}^p : \sup_{\beta \in \mathcal{H}_{\beta^*}^{\text{Lasso}}} \gamma^T \beta \leq 1 \right\}$ $= \{\gamma \in \mathbb{R}^p : \|\gamma\|_\infty \leq \eta_\gamma\}$ $= \mathbf{conv}\{-\eta_\gamma, \eta_\gamma\}^p\},$ where $\eta_\gamma = 1/\eta_\beta$ and **conv** denotes convex hull
Criterion	$\min_{\beta} \max_{\gamma \in \mathcal{D}_\gamma} \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda\|\beta + P_\beta\gamma\|_2^2$ $\Leftrightarrow \min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|_2^2 +$ $\lambda\sum_{j=1}^p (c(j-1)+1)|\beta|_{(j)} + \lambda\|\beta\|_2^2$	$\min_{\beta \in \mathbb{R}^p} \max_{\gamma \in \mathcal{D}_\gamma} \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + 2\lambda\eta_\gamma \sum_{k=1}^K \|\beta_{\mathcal{G}_k}\|_\infty$ $\Leftrightarrow \min_{\beta \in \mathbb{R}^p} \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda\|\beta - \gamma\|_2^2$	$\min_{\beta \in \mathbb{R}^p} \max_{\gamma \in \mathcal{D}_{\text{Lasso}}} \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda\|\beta - \gamma\|_2^2$ $\Leftrightarrow \min_{\beta \in \mathbb{R}^p} \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + 2\lambda\eta_\gamma\|\beta\|_1 + \lambda\|\beta\|_2^2$
2D Representation	(β_2 vs β_1)	(β_2 vs β_1)	(β_2 vs β_1)

Table 1: Various classical sparse problems may be expressed by means of a quadratic penalty. Each line of the table presents a different problem with its associated worst case quadratic penalty.

where $\eta_\gamma = 1/\eta_\beta$ and $\mathbf{e}_j^p$ is the jth element of the canonical basis of $\mathbb{R}^p$, that is $e_{jj'} = 1$ if $j = j'$ and $e_{jj'} = 0$ otherwise. Then, Problem (**??**) becomes:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \max_{\boldsymbol{\gamma} \in \mathcal{D}_\gamma^{\mathrm{Max}}} \left\{ \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda \|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2^2 \right\}$$

$$\Leftrightarrow \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + 2\lambda\eta_\gamma \|\boldsymbol{\beta}\|_\infty + \lambda \|\boldsymbol{\beta}\|_2^2 \quad ,$$

Now, consider the more general situation where a group structure is defined on the set of variables by setting a partition of the index set $\mathcal{I} = \{1, \ldots, p\}$, that is,

$$\mathcal{I} = \bigcup_{k=1}^K \mathcal{G}_k \ , \ \text{with} \ \mathcal{G}_k \cap \mathcal{G}_\ell = \emptyset \ \text{for} \ k \neq \ell \ .$$

Let p_k denote the cardinality of group k, and $\boldsymbol{\beta}_{\mathcal{G}_k} \in \mathbb{R}^{p_k}$ be the vector $(\beta_j)_{j \in \mathcal{G}_k}$. We now consider the regularity assumption stating that the $\ell_{\infty,1}$ mixed-norm of $\boldsymbol{\beta}^\star$ (that is, its groupwise ℓ_∞-norm) should be small:

$$\mathcal{H}_{\boldsymbol{\beta}^*}^{\ell_{\infty,1}} = \left\{ \boldsymbol{\beta} \in \mathbb{R}^p : \sum_{k=1}^K \|\boldsymbol{\beta}_{\mathcal{G}_k}\|_\infty \leq \eta_\beta \right\} \ .$$

The dual assumption is that the groupwise ℓ_1-norm of $\boldsymbol{\gamma}$ should be controlled:

$$\mathcal{D}_\gamma^{\ell_{\infty,1}} = \left\{ \boldsymbol{\gamma} \in \mathbb{R}^p : \sup_{\boldsymbol{\beta} \in \mathcal{H}_{\boldsymbol{\beta}^*}^{\ell_{\infty,1}}} \boldsymbol{\gamma}^\mathsf{T}\boldsymbol{\beta} \leq 1 \right\}$$

$$= \left\{ \boldsymbol{\gamma} \in \mathbb{R}^p : \max_{k \in \{1,\ldots,K\}} \|\boldsymbol{\gamma}_{\mathcal{G}_k}\|_1 \leq \eta_\gamma \right\}$$

$$= \mathbf{conv}\left\{ \left\{ \eta_\gamma \mathbf{e}_1^{p_1}, \ldots, \eta_\gamma \mathbf{e}_{p_1}^{p_1}, -\eta_\gamma \mathbf{e}_1^{p_1}, \ldots, -\eta_\gamma \mathbf{e}_{p_1}^{p_1} \right\} \times \ldots \right.$$
$$\left. \times \left\{ \eta_\gamma \mathbf{e}_1^{p_K}, \ldots, \eta_\gamma \mathbf{e}_{p_K}^{p_K}, -\eta_\gamma \mathbf{e}_1^{p_K}, \ldots, -\eta_\gamma \mathbf{e}_{p_K}^{p_K} \right\} \right\} \ ,$$

so that Problem (**??**) becomes:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \max_{\boldsymbol{\gamma} \in \mathcal{D}_\gamma^{\ell_{\infty,1}}} \left\{ \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda \|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2^2 \right\}$$

$$\Leftrightarrow \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + 2\lambda\eta_\gamma \sum_{k=1}^K \|\boldsymbol{\beta}_{\mathcal{G}_k}\|_\infty + \lambda \|\boldsymbol{\beta}\|_2^2 \quad ,$$
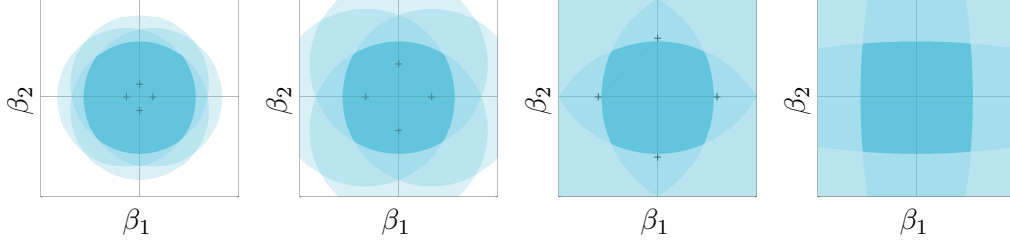
8

Figure 2: Sublevel sets for the $\ell_{\infty,1}$ group-Lasso penalties (represented by the darker colored patches). Each set is defined as the intersection of the the Euclidean balls (represented by the lighter color patches) whose centers are represented by crosses.

Notice that the limiting cases of this penalty are two classical problems: ridge regression and the $\ell_{\infty,1}$ group-Lasso. A 2D pictorial illustration of this evolution is given in Figure 2, where the shape of the uncertainty set $\mathcal{D}_{\boldsymbol{\gamma}}$ is the convex hull of the points located on the axes at $\pm\eta_{\gamma}$, which are identified by the cross markers. Then, the sublevel set $\{\boldsymbol{\beta} : \max_{\boldsymbol{\gamma} \in \mathcal{D}_{\boldsymbol{\gamma}}} \|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2^2 \leq t\}$ is simply defined as the intersection of the four sublevel sets $\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2^2 \leq t\}$ for $\boldsymbol{\gamma} = \pm\eta_{\gamma}\mathbf{e}_1^2$ and $\boldsymbol{\gamma} = \pm\eta_{\gamma}\mathbf{e}_1^2$, which are Euclidean balls centered at these $\boldsymbol{\gamma}$ values.

4. Algorithm

The unified derivation for the problems presented in Section 3 suggests a unified processing based on the iterative resolution of quadratic problems. This general algorithm is summarized in this section. We then show that the new derivation can also be used in the analysis of this algorithm by describing an alternative to Fenchel duality (used for example by Bach et al., 2012) to assess convergence.

4.1. Active Set Approach

The efficient approaches developed for sparse regression take advantage of the sparsity of the solution by solving a series of small linear systems, whose sizes are incrementally increased/decreased. Here, as for the Lasso (Osborne et al., 2000; Efron et al., 2004), this process boils down to an iterative optimization scheme involving the resolution of quadratic problems.
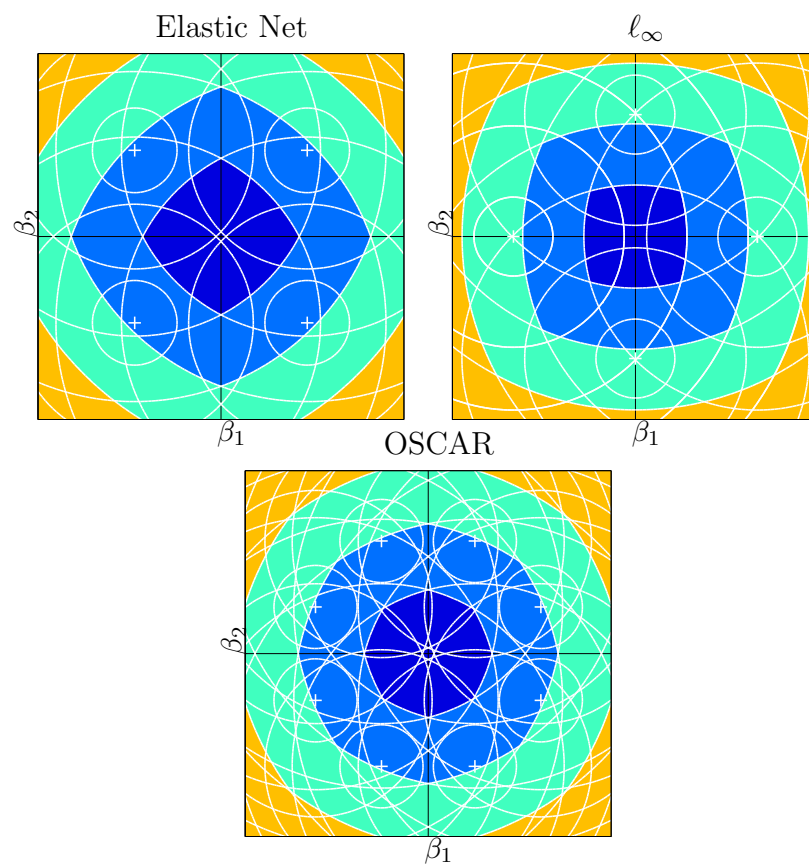
Figure 3: Penalty shapes (patches) built from the quadratic functions whose isocontour are displayed in white.

The algorithm is based on the iterative update of the set of "active" variables, $\mathcal{A}$, indexing the coefficients $\boldsymbol{\beta}_{\mathcal{A}}$ currently identified as being non-zero. It is started from a sparse initial guess, say $\mathcal{A} = \emptyset$ ($\boldsymbol{\beta} = 0$), and iterates the three following steps:

1. the first step solves Problem (**??**) considering that $\mathcal{A}$, the set of "active" variables, is correct; that is, the objective (**??**) is optimized with respect to $\boldsymbol{\beta}_{\mathcal{A}}$. This penalized least squares problem is defined from $\mathbf{X}_{.\mathcal{A}}$, which is the submatrix of $\mathbf{X}$ comprising all rows and the columns indexed by $\mathcal{A}$ and $\boldsymbol{\gamma}_{\mathcal{A}}$, which is set to its current most adversarial value. [1]

2. the second step updates $\boldsymbol{\beta}_{\mathcal{A}}$ if necessary (and possibly $\boldsymbol{\gamma}_{\mathcal{A}}$), so that $\boldsymbol{\gamma}_{\mathcal{A}}$ is indeed (one of) the most adversarial value of the current $\boldsymbol{\beta}_{\mathcal{A}}$. This is easily checked with the problems given in Section 3, where $\mathcal{D}_{\boldsymbol{\gamma}}$ is a convex polytope whose vertices (that is, extreme $\boldsymbol{\gamma}$-values) are associated with a cone of coherent $\boldsymbol{\beta}$-value.

3. the last step updates the active set $\mathcal{A}$. It relies on the "worst-case gradient" with respect to $\boldsymbol{\beta}$, where $\boldsymbol{\gamma}$ is chosen so as to minimize infinitesimal improvements of the current solution. Again picking the right $\boldsymbol{\gamma}$ is easy for the problems given in Section 3. Once this is done, we first check whether some variables should quit the active set, and if this is not the case, we assess the completeness of $\mathcal{A}$, by checking the optimality conditions with respect to inactive variables. We add the variable, or the group of variables that most violates the worst-case optimality conditions. When no such violation exists, the current solution is optimal, since, at this stage, the problem is solved exactly within the active set $\mathcal{A}$.

Algorithm 1 provides a more comprehensive technical description. Note that the structure is essentially identical to the one proposed by Osborne et al. (2000) or Efron et al. (2004) for the Lasso, but that it applies to any penalty that can be decomposed as in Problem (**??**). Our viewpoint is also radically different, as the global non-smooth problem is dealt with subdifferentials by Osborne et al. (2000), whereas we rely on the maximum of smooth functions. This approach suggests a new assessment of convergence, as detailed below.

---

[1] When several $\boldsymbol{\gamma}_{\mathcal{A}}$ are equally unfavorable to $\boldsymbol{\beta}_{\mathcal{A}}$, we use gradient information to pick the worst one among those when $\boldsymbol{\beta}_{\mathcal{A}}$ moves along the steepest descent direction.

Algorithm 1: Worst-Case Quadratic Penalty Active Set Algorithm

---

Init. $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta}^0$

Determine the active set: $\mathcal{A} \leftarrow \{j : \beta_j > 0\}$

Pick a worst admissible $\boldsymbol{\gamma}$, that is, $\boldsymbol{\gamma} \in \arg\max\limits_{\mathbf{g}\in\mathcal{D}_\gamma} \|\boldsymbol{\beta} - \mathbf{g}\|_2^2$

Step 1 Update active variables $\boldsymbol{\beta}_{\mathcal{A}}$ assuming that $\mathcal{A}$ and $\boldsymbol{\gamma}_{\mathcal{A}}$ are optimal

$\boldsymbol{\beta}_{\mathcal{A}}^{\text{old}} \leftarrow \boldsymbol{\beta}_{\mathcal{A}}$

$\boldsymbol{\beta}_{\mathcal{A}} \leftarrow \left(\mathbf{X}_{.\mathcal{A}}^{\mathsf{T}}\mathbf{X}_{.\mathcal{A}} + \lambda\mathbf{I}_{|\mathcal{A}|}\right)^{-1}\left(\mathbf{X}_{.\mathcal{A}}^{\mathsf{T}}\mathbf{y} + \lambda\boldsymbol{\gamma}_{\mathcal{A}}\right)$

Step 2 Verify coherence of $\boldsymbol{\gamma}_{\mathcal{A}}$ with the updated $\boldsymbol{\beta}_{\mathcal{A}}$

if $\|\boldsymbol{\beta}_{\mathcal{A}} - \boldsymbol{\gamma}_{\mathcal{A}}\|_2^2 < \max\limits_{\mathbf{g}\in\mathcal{D}_\gamma} \|\boldsymbol{\beta}_{\mathcal{A}} - \mathbf{g}_{\mathcal{A}}\|_2^2$ **then**   `// if `$\gamma_\mathcal{A}$` is not worst-case`

> `/* Backtrack towards the last `$\gamma_\mathcal{A}$`-coherent solution:        */`
>
> $\boldsymbol{\beta}_{\mathcal{A}} \leftarrow \boldsymbol{\beta}_{\mathcal{A}}^{\text{old}} + \rho(\boldsymbol{\beta}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^{\text{old}})$
>
> $\boldsymbol{\gamma}_{\mathcal{A}}$ is worst-case for $\boldsymbol{\beta}_{\mathcal{A}}$, and there is another worst-case value $\widetilde{\boldsymbol{\gamma}}_{\mathcal{A}}$
>
> `/* Check whether progress can be made with `$\widetilde{\gamma}_\mathcal{A}$`        */`
>
> $\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} \leftarrow \left(\mathbf{X}_{.\mathcal{A}}^{\mathsf{T}}\mathbf{X}_{.\mathcal{A}} + \lambda\mathbf{I}_{|\mathcal{A}|}\right)^{-1}\left(\mathbf{X}_{.\mathcal{A}}^{\mathsf{T}}\mathbf{y} + \lambda\widetilde{\boldsymbol{\gamma}}_{\mathcal{A}}\right)$
>
> **if** $\left\|\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} - \widetilde{\boldsymbol{\gamma}}_{\mathcal{A}}\right\|_2^2 = \max\limits_{\mathbf{g}\in\mathcal{D}_\gamma} \left\|\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} - \mathbf{g}_{\mathcal{A}}\right\|_2^2$ **then** `// if `$\widetilde{\gamma}_\mathcal{A}$` is worst-case...`
>
> > $(\boldsymbol{\beta}_{\mathcal{A}}, \boldsymbol{\gamma}_{\mathcal{A}}) \leftarrow (\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}, \widetilde{\boldsymbol{\gamma}}_{\mathcal{A}})$ `// `$(\widetilde{\beta}_\mathcal{A}, \widetilde{\gamma}_\mathcal{A})$` is better than `$(\beta_\mathcal{A}, \gamma_\mathcal{A})$

`/* The current `$\gamma_\mathcal{A}$` is coherent with `$\beta_\mathcal{A}$`                    */`

Step 3 Update active set $\mathcal{A}$

$g_j \leftarrow \min\limits_{\gamma\in\mathcal{D}_\gamma} \left|\mathbf{x}_j^{\mathsf{T}}(\mathbf{X}_{.\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}} - \mathbf{y}) + \lambda(\beta_j - \gamma_j)\right| \quad j = 1, \ldots, p$ `// worst-case`

`gradient`

if $\exists j \in \mathcal{A} : \beta_j = 0 \ and \ g_j = 0$ **then**

> $\mathcal{A} \leftarrow \mathcal{A}\backslash\{j\}$;   `// Downgrade `j
>
> `/* Go to Step 1                                      */`

else

> **if** $\max_{j\in\mathcal{A}^c} g_j \neq 0$ **then**
>
> > `/* Identify the greatest violation of optimality`
> >
> > `   conditions:                                     */`
> >
> > $j^\star \leftarrow \arg\max\limits_{j\in\mathcal{A}^c} g_j, \quad \mathcal{A} \leftarrow \mathcal{A}\cup\{j^\star\}$;   `// Upgrade `$j^\star$
> >
> > `/* Go to Step 1                                     */`
>
> **else**
>
> > `/* Stop and return `$\boldsymbol{\beta}$`, which is optimal           */`

4.2. Monitoring Convergence

At each iteration of the algorithm, the current $\boldsymbol{\beta}$ is computed assuming that the current active set $\mathcal{A}$ and the current $\boldsymbol{\gamma}_{\mathcal{A}}$-value are optimal. When the current active set is not optimal, the current $\boldsymbol{\beta}$ (where $\boldsymbol{\beta}_{\mathcal{A}}$ is completed by zeros on the complement $\mathcal{A}^c$) is nevertheless optimal for a $\boldsymbol{\gamma}$-value defined in $\mathbb{R}^p$ (where $\boldsymbol{\gamma}_{\mathcal{A}}$ is completed by ad hoc values on the complement $\mathcal{A}^c$). However this $\boldsymbol{\gamma}$ fails to belong to $\mathcal{D}_{\boldsymbol{\gamma}}$ (otherwise, the problem would be solved: $\mathcal{A}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ would indeed be optimal). The following proposition relates the current objective function, associated with an infeasible $\boldsymbol{\gamma}$-value ($\boldsymbol{\gamma} \notin \mathcal{D}_{\boldsymbol{\gamma}}$), to the global optimum of the optimization problem.

Proposition 1. *For any $\eta_{\boldsymbol{\gamma}} > 0$, and for all vectorial norm $\|\cdot\|_*$, when $\mathcal{D}_{\boldsymbol{\gamma}}$ is defined as $\mathcal{D}_{\boldsymbol{\gamma}} = \{\boldsymbol{\gamma} \in \mathbb{R}^p : \|\boldsymbol{\gamma}\|_* \leq \eta_{\boldsymbol{\gamma}}\}$, then, $\forall \boldsymbol{\gamma} \in \mathbb{R}^p : \|\boldsymbol{\gamma}\|_* \geq \eta_{\boldsymbol{\gamma}}$, we have:*

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \max_{\boldsymbol{\gamma}' \in \mathcal{D}_{\boldsymbol{\gamma}}} J_\lambda(\boldsymbol{\beta}, \boldsymbol{\gamma}') \geq \frac{\eta_{\boldsymbol{\gamma}}}{\|\boldsymbol{\gamma}\|_*} J_\lambda\left(\boldsymbol{\beta}^\star(\boldsymbol{\gamma}), \boldsymbol{\gamma}\right) - \frac{\lambda \eta_{\boldsymbol{\gamma}}(\|\boldsymbol{\gamma}\|_* - \eta_{\boldsymbol{\gamma}})}{\|\boldsymbol{\gamma}\|_*^2} \|\boldsymbol{\gamma}\|_2^2 \quad,$$

where

$$J_\lambda(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda \|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2^2 \quad and \quad \boldsymbol{\beta}^\star(\boldsymbol{\gamma}) = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} J_\lambda(\boldsymbol{\beta}, \boldsymbol{\gamma}) \quad.$$

See proof in Appendix Appendix A.3.

This proposition can be used to compute an optimality gap at Step 3 of Algorithm 1, by picking a $\boldsymbol{\gamma}$-value such that the current worst-case gradient $\mathbf{g}$ is null (the current $\boldsymbol{\beta}$-value then being the optimal $\boldsymbol{\beta}^\star(\boldsymbol{\gamma})$). Note that more precise upper bounds could be computed relying on significant extra computation. The generic optimality gap computed from Proposition 1 differs from the Fenchel duality gap (see Bach et al., 2012). For the elastic net expressed in (4), Fenchel inequality (see details in Mairal, 2010) yields the following optimality gap:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \max_{\boldsymbol{\gamma}' \in \mathcal{D}_{\boldsymbol{\gamma}}} J_\lambda(\boldsymbol{\beta}, \boldsymbol{\gamma}') \geq J_\lambda\left(\boldsymbol{\beta}^\star(\boldsymbol{\gamma}), \boldsymbol{\gamma}\right) - \frac{\eta_{\boldsymbol{\gamma}}^2}{\|\boldsymbol{\gamma}\|_*^2} \left(\|\mathbf{X}\boldsymbol{\beta}^\star(\boldsymbol{\gamma}) - \mathbf{y}\|_2^2 + \lambda \|\boldsymbol{\beta}^\star(\boldsymbol{\gamma})\|_2^2 \right)$$
$$- 2 \frac{\eta_{\boldsymbol{\gamma}}}{\|\boldsymbol{\gamma}\|_*} \left(\mathbf{X}\boldsymbol{\beta}^\star(\boldsymbol{\gamma}) - \mathbf{y}\right)^\mathsf{T} \mathbf{y} \quad.$$

The two optimality gaps are empirically compared in Figure 4 for the elastic net, along a short regularization path with five values of the ℓ_1-penalization
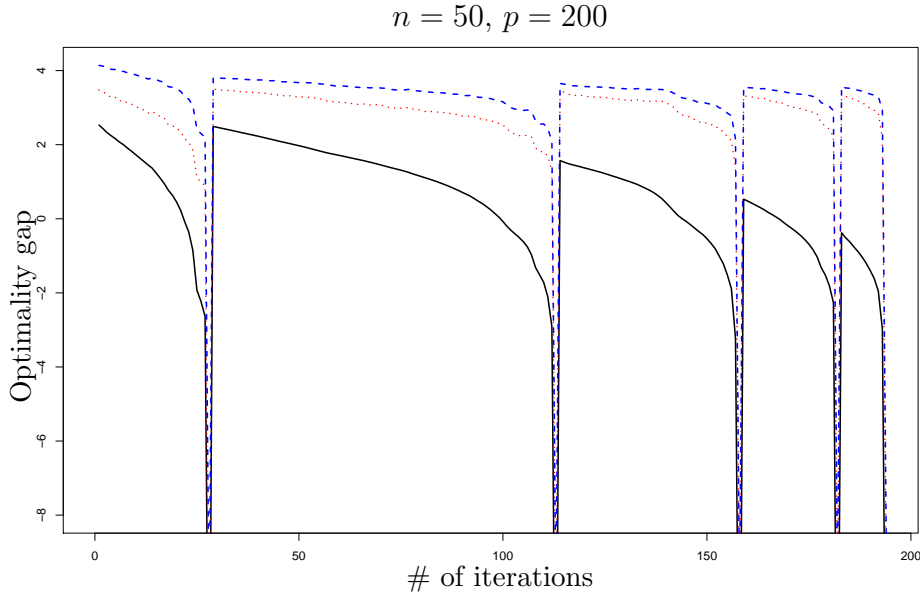
Figure 4: Monitoring convergence: true optimality gap (solid black) versus our generic upper bound (dashed blue) and Fenchel's duality gap for elastic net (dotted red) computed at each iteration of Algorithm 1.

parameter. We see that the two optimality gaps can be used to assess convergence, though Fenchel's duality gap is more accurate for the rougher solutions. Note however that both gaps are fairly coarse until a very accurate solution is reached, which makes them both unsuitable for deriving loose stopping criteria. Proposition 1 is thus of limited scope, but it illustrates that, besides its algorithmic consequences, our original view of sparse penalties opens new ways for analysis. As a final note on this topic, we provide a slightly tighter inequality for computing the optimality gap in Appendix Appendix A.3, and we conjecture that it could be further tightened (see in particular the derivation of inequality (A.7)).

5. Numerical Experiments

This section compares the performances of our algorithm to its state-of-the-art competitors from an optimization viewpoint. Efficiency may then be assessed by accuracy and speed: accuracy is the difference between the optimum of the objective function and its value at the solution returned by the algorithm; speed is the computing time required for returning this solution. Obviously, the timing of two algorithms/packages has to be compared at

14

similar precision requirements, which are rather crude in statistical learning, far from machine precision (Bottou and Bousquet, 2008).

We compare the performance of the proposed quadratic solver with representatives of the most successful competing algorithms. First, we use our own implementations of all competitors, so as to provide comparisons without implementation biases (language, library, etc.). We use `R` with most of the matrix calculus done in `C++` using the `RcppArmadillo` package (Eddelbuettel and François, 2011; Sanderson, 2010) that relies on `BLAS/LAPACK` libraries for linear algebra operations. Second, we compare our code to the leading standalone packages that are available today, so as to provide comparisons avoiding a possible competence bias.

We use simulated data to obtain representative average results. Their generation covers the typical attributes of the real data encountered in post genomic and signal processing. In these domains, the main optimization difficulties result from ill-conditioning, which is either due to the high correlation between predictors, or to underdetermination when the number of variables exceeds the sample size (also known as the high-dimensional or the "large p small n" setup). For the optimization algorithms based on active set strategies, bad conditioning is somehow alleviated when the objective function has a regular behavior when restricted to the subspace containing the solution. All other things being equal, this local conditioning is thus governed by the sparsity of the unknown true parameter (which affects the sparsity of the solution), which also heavily impacts the running times of most optimization algorithms available today.

5.1. Data Generation

The above-mentioned characteristics are explored without difficulty in the framework of linear regression. We generate samples of size n from the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^\star + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \ ,$$

with σ chosen so as to reach a rather strong coefficient of determination ($R^2 \approx 0.8$). The design matrix $\mathbf{X}$ is drawn from a multivariate normal distribution in $\mathbb{R}^p$, and the conditioning of $\mathbf{X}^\intercal \mathbf{X}$ is ruled by the correlation between variables. We use the same correlation coefficient ρ for all pairs of variables. The sparsity of the true regression coefficients is controlled by a

parameter s, with

$$\boldsymbol{\beta}^\star = \big(\underbrace{2, \ldots, 2}_{s/2}, \underbrace{-2, \ldots, -2}_{s/2}, \underbrace{0, \ldots, 0}_{p-s} \big) \ .$$

Finally, the ratio p/n quantifies the well/ill-posedness of the problem.

5.2. Comparing Optimization Strategies

We compare here the performance of three state-of-the-art optimization strategies implemented in our own computational framework: accelerated proximal method (see, e.g., Beck and Teboulle, 2009), coordinate descent (popularized by Friedman et al., 2007), and our algorithm, that will respectively be named hereafter `proximal`, `coordinate` and `quadratic`. Our implementations estimate the solution to the elastic net problem

$$J^{\text{enet}}_{\lambda_1, \lambda_2}(\boldsymbol{\beta}) = \frac{1}{2} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{2} \|\boldsymbol{\beta}\|_2^2 \ , \tag{5}$$

which is strictly convex when $\lambda_2 > 0$ and thus admits a unique solution even if $n < p$.

The three implementations are embedded in the same active set routine, which approximately solves the optimization problem with respect to a limited number of variables as in Algorithm 1. They only differ regarding the inner optimization problem with respect to the current active variables, which is performed by an accelerated proximal gradient method for `proximal`, by coordinate descent for `coordinate`, and by the resolution of the worst-case quadratic problem for `quadratic`. We followed the practical recommendations of Bach et al. (2012) for accelerating the proximal and coordinate descent implementations, and we used the same halting condition for the three implementations, based on the approximate satisfaction of the first-order optimality conditions:

$$\max_{j\{\in 1 \ldots p\}} \left| \mathbf{x}_j^\intercal \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \right) + \lambda_2 \hat{\boldsymbol{\beta}} \right| < \lambda_1 + \tau, \tag{6}$$

where the threshold τ was fixed to $\tau = 10^{-2}$ in our simulations. [2] Finally, the active set algorithm is itself wrapped in a warm-start routine, where the

---

[2]The rather loose threshold is favorable to `coordinate` and `proximal`, which reach the threshold, while `quadratic` ends up with a much smaller value, due to the exact resolution, up to machine precision, of the inner quadratic problem.

approximate solution to $J_{\lambda_1,\lambda_2}^{\text{enet}}$ is used as the starting point for the resolution of $J_{\lambda_1',\lambda_2}^{\text{enet}}$ for $\lambda_1' < \lambda_1$.

Our benchmark considers small-scale problems, with size $p = 100$, and the nine situations stemming from the choice of three following parameters:

- low, medium and high correlation between predictors ($\rho \in \{0.1, 0.4, 0.8\}$),

- low, medium and high-dimensional setting ($p/n \in \{2, 1, 0.5\}$),

- low, medium and high levels of sparsity ($s/p \in \{0.6, 0.3, 0.1\}$).

Each solver computes the elastic net for the tuning parameters λ_1 and λ_2 on a 2D-grid of 50×50 values, and their running times have been averaged over 100 runs.

All results are qualitatively similar regarding the dimension and sparsity settings. Figure 5 displays the high-dimensional case ($p = 2n$) with a medium level of sparsity ($s = 30$) for the three levels of correlation. Each map represents the log-ratio between the timing of either `coordinate` or `proximal` versus `quadratic`, according to (λ_1, λ_2) for a given correlation level. Dark regions with a value of 1 indicate identical running times while lighter regions with a value of 10 indicate that `quadratic` is 10 times faster. Figure 5 illustrates that `quadratic` outperforms both `coordinate` and `proximal`, by running much faster in most cases, even reaching 300-fold speed increases. The largest gains are observed for small (λ_1, λ_2) penalty parameters for which the problem is ill-conditioned, including many active variables, resulting in a huge slowdown of the first-order methods `coordinate` and `proximal`. As the penalty parameters increase, smaller gains are observed, especially when λ_2, attached to the quadratic penalty, reaches high values for which all problems are well-conditioned, and where the elastic net is leaning towards univariate soft thresholding, in which case all algorithms behave similarly.

5.3. Comparing Stand-Alone Implementations

We now proceed to the evaluation of our code with three other stand-alone programs publicly available as `R` packages. We chose three leading state-of-the-art packages, namely `glmnet` (Generalized Linear Models regularized by Lasso and elastic-NET, Friedman et al., 2010), `lars` (Least Angle Regression, lasso and forward Stagewise, Efron et al., 2004) and `SPAMS` (SPArse Modeling Software, Bach et al., 2012), with two options `SPAMS-FISTA`, which implements an accelerated proximal method, and `SPAMS-LARS` which is a `lars` substitute.
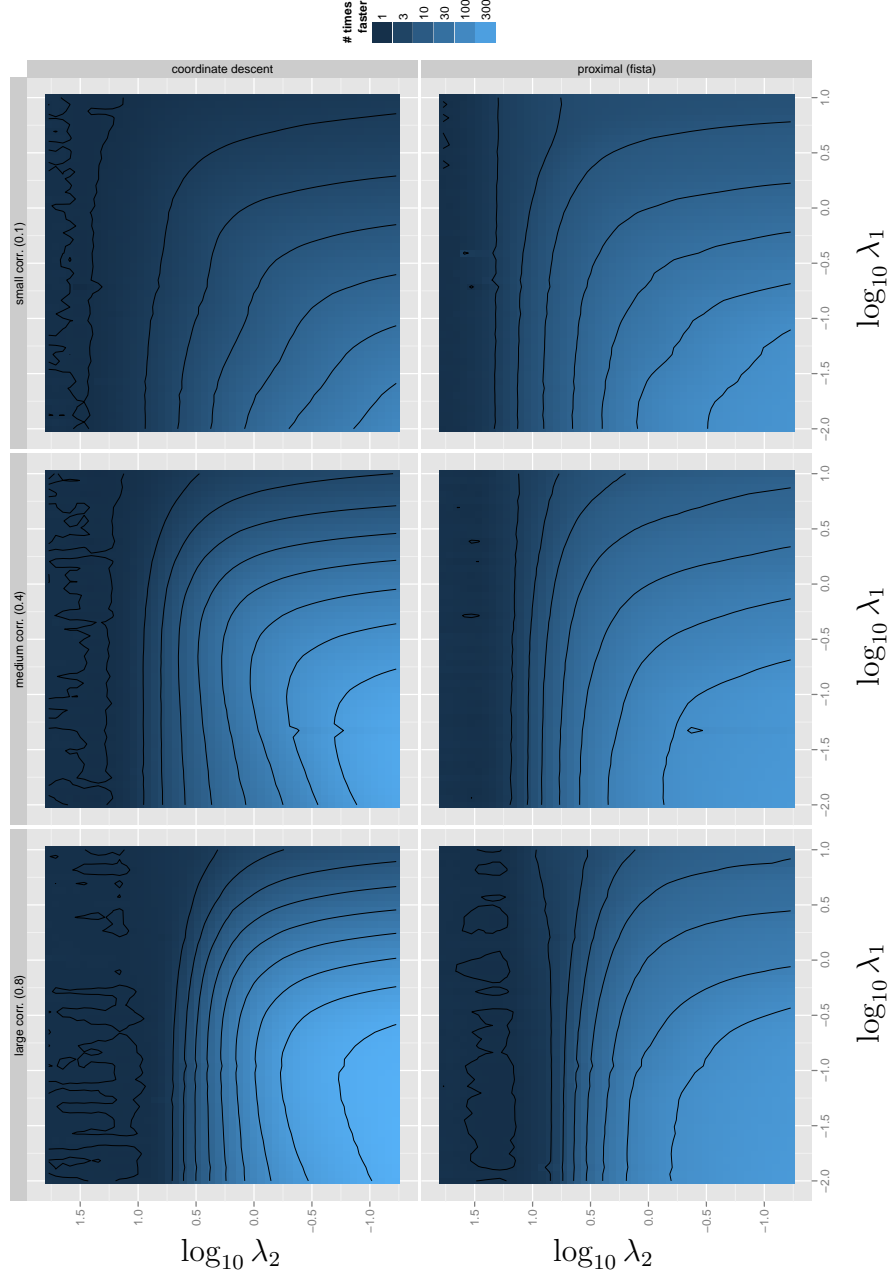
Figure 5: Log-ratio of computation times for `coordinate` (left), and `proximal` (right), versus `quadratic`, for $(p, n) = (100, 50)$, $s = 30$ and correlation $\rho \in \{0.1, 0.4, 0.8\}$ (top, middle and bottom respectively).

Note that `glmnet` does most of its internal computations in `Fortran`, `lars` in `R`, and `SPAMS` in `C++`. Our own implementation, by resolution of the worst-case quadratic problem, is shipped within an `R` package `quadrupen` publicly available at the second author's web page[3].

We benchmark these packages by computing regularization paths for the Lasso[4], that is, the elastic net Problem (5) with $\lambda_2 = 0$. The inaccuracy of the solutions produced is measured by the gap in the objective function compared to a reference solution, considered as being the true optimum. We use the `lars` solution as a reference, since it solves the Lasso problem up to the machine precision, relying on the `LAPACK/BLAS` routines. Furthermore, `lars` provides the solution path for the Lasso, that is, the set of solutions computed for each penalty parameter value for which variable activation or deletion occurs, from the empty model to the least-mean squares model. This set of reference penalty parameters is used here to define a sensible reproducible choice.

In high dimensional setups, the computational cost of returning the solutions for the largest models may be overwhelming compared to the one necessary for exploring the interesting part of the regularization path (Simon et al., 2011; Friedman et al., 2010). This is mostly due to numerical instability problems that may be encountered in these extreme settings, where the Lasso solution is overfitting as it approaches the set of solutions to the underdetermined least squares problem. We avoid a comparison mostly relying on these spurious cases by restricting the set of reference penalty parameters to the first $\min(n, p)$ steps of `lars` (similar settings are used by Friedman et al., 2010).

Henceforth, the distance D of a given `method` to the optimum is evaluated on the whole set of penalties Λ used along the path, by

$$
\mathrm{D}(\texttt{method}) = \left(\frac{1}{|\Lambda|} \sum_{\lambda \in \Lambda} \left(J_\lambda^{\mathrm{lasso}} \left(\hat{\boldsymbol{\beta}}_\lambda^{\texttt{lars}} \right) - J_\lambda^{\mathrm{lasso}} \left(\hat{\boldsymbol{\beta}}_\lambda^{\texttt{method}} \right) \right)^2 \right)^{1/2} ,
$$

where $J_\lambda^{\mathrm{lasso}}(\boldsymbol{\beta}) = J_{\lambda,0}^{\mathrm{enet}}(\boldsymbol{\beta})$ is the objective function of the Lasso evaluated at $\boldsymbol{\beta}$, and $\hat{\boldsymbol{\beta}}_\lambda^{\texttt{method}}$ is the estimated optimal solution provided by the `method`

---

[3]or directly at . This will be made available on the CRAN (Comprehensive R Archive Network)

[4]We benchmark the packages on a Lasso problem since the parametrization of the elastic net problem differs among packages, hindering fair comparisons.

package currently tested.

The data sets are generated according to the linear model described above, in three different high-dimensional settings and small to medium number of variables: $(p, n) = (100, 40)$, $(p, n) = (1\,000, 200)$ and $(p, n) = (10\,000, 400)$. The sparsity of the true underlying $\boldsymbol{\beta}^\star$ is governed by $s = 0.25\min(n, p)$, and the correlation between predictors is set by $\rho \in \{0.1, 0.4, 0.8\}$. For each value of ρ, we averaged the timings over 50 simulations, ensuring that each package computes the solutions at identical λ values, as defined above.

We pool together the runtimes obtained for the three levels of correlation for quadrupen, SPAMS-LARS and lars, which are not sensible to the correlation between features. In each plot of Figure 6, each of these methods is thus represented by a single point marking the average precision and the average distance to the optimum over the 150 runs (50 runs for each $\rho \in \{0.1, 0.4, 0.8\}$). Note that for lars only the abscissa is meaningful since D(lars) is zero by definition. Besides, quadrupen, which solves each quadratic problem up to the machine precision, tends to be within this precision of the lars solution.The SPAMS-LARS is also very precise, up to 10^{-6}, which is the typical precision of the approximate resolution of linear systems. It is the fastest alternative for solving the Lasso when the problem is high-dimensional with a large number of variables (Figure 6, bottom-left).

In contrast, the (precision,timing)-values of glmnet and SPAMS-FISTA are highly affected by the threshold parameters[5] that control their stopping conditions. The computational burden to reach a given precision is also affected by the level of correlation, as illustrated in Figure 6. Obviously, a precise solution is difficult to reach with first-order descent algorithms in a high correlation setup, which corresponds to an ill-conditioned linear system. It may be surprising to observe that SPAMS-FISTA is about ten time slower than glmnet, as proximal and coordinate descent methods were experimentally shown to be roughly equivalent in our preceding analysis and by Bach et al. (2012). However, these two comparisons were carried out with the same active set strategy (that is, *with* active set for ours and *without* active set for Bach et al., 2012). We believe that this difference in the handling of active

---

[5]In glmnet, convergence is monitored by the stability of the objective function, measured between two optimization steps, and optimization is halted when changes fall below the specified threshold (scaled by the null deviance). In SPAMS-FISTA, the stopping condition of the algorithm is based on the relative change of parameters between two iterations.
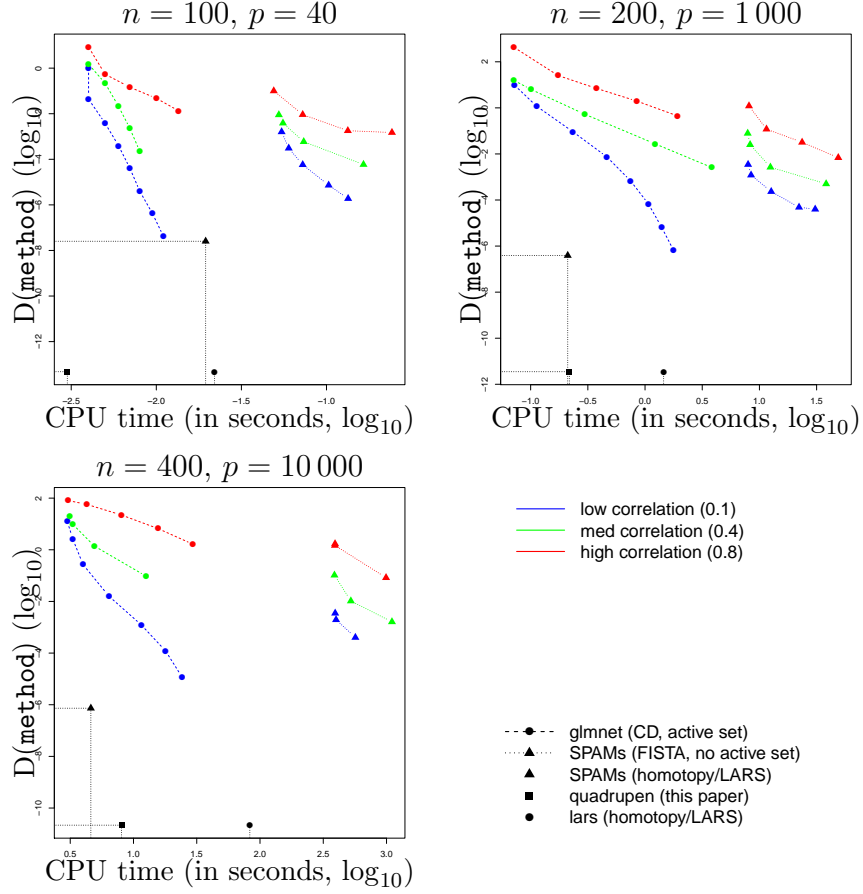
Figure 6: Distance to optimum versus CPU time for three different high-dimensional settings: $(p, n) = (100, 40)$ (top left), $(p, n) = (1\,000, 200)$ (top right) and $p = (10\,000, 400)$ (bottom left).

variables explains the relative bad performance of `SPAMS-FISTA`, which optimizes all variables along the regularization path, while `glmnet` uses a greedy active set strategy.

Overall, our implementation is highly competitive, that is, very accurate, at the `lars` level, and much faster. The speed improvements of `glmnet` are only observed for very rough approximate solutions and `SPAMS-FISTA` is dominated by `glmnet`. Our experiments, in the framework of active set methods, agree with the results of Bach et al. (2012): indeed, they observed that first-order methods are competitive with second-order ones only for low correlation levels and small penalties (which entails a large number of active variables). Conversely, our results may appear to contradict some of the experimental findings of Friedman et al. (2010): first, we observe that `glmnet` is quite sensitive to correlations, and second, the optimized second-order methods are competitive with `glmnet`. These differences in conclusions arise from the differences in experimental protocols: while we compare running times at a given accuracy, they are compared at a given threshold on the stopping criterion by Friedman et al. (2010). Regarding the influence of correlations, the stability-based criterion can be fooled due to the tiny step size that typically occurs for ill-conditioned problems, leading to a sizable early stopping. Regarding the second point, even though the `R` implementation of `lars` may indeed be slow compared to `glmnet`, considerable improvements can be obtained using optimized second-order methods such as `quadrupen` as soon as a sensible accuracy is required, especially when correlation increases.

Finally, among the accurate solvers, `SPAMS-LARS` is insignificantly less accurate than `quadrupen` or `lars` in a statistical context. It is always faster than `lars` and slightly faster than `quadrupen` for the largest problem sizes (Figure 6, bottom-left) and much slower for the smallest problem (Figure 6, top-left).

5.4. Link between accuracy of solutions and prediction performances

When the "irrepresentable condition" (Zhao and Yu, 2006) holds, the Lasso should select the true model consistently. However, even when this rather restrictive condition is fulfilled, perfect support recovery obviously requires numerical accuracy: rough estimates may speed up the procedure, but whatever optimization strategy is used, stopping an algorithm is likely to prevent either the removal of all irrelevant coefficients or the insertion of all relevant ones. The support of the solution may then be far from the optimal one.

We advocate here that our quadratic solver is very competitive in computation time when support recovery matters, that is, when high level of accuracy is needed, in small (few hundreds of variables) and medium sized problems (few thousands). As an illustration, we generate 100 data sets under the linear model described above, with a rather strong coefficient of determination ($R^2 \approx 0.8$ on average), a rather high level of correlation between predictors ($\rho = 0.8$) and a medium level of sparsity ($s/p = 30\%$). The number of variable is kept low ($p = 100$) and the difficulty of the estimation problem is tuned by the n/p ratio. For each data set, we also generate a test set sufficiently large (say, $10n$) to evaluate the quality of the prediction without depending on any sampling fluctuation. We compare the Lasso solutions computed by `quadrupen` to the ones returned by `glmnet` with various level of accuracy[6]. Figure 7 reports performances, as measured by the mean squared test error and the support error rate.

As expected, the curves show that selecting variables and searching for the best prediction are two different problems. The selection problem (bottom of Figure 7) always requires a sparser model than the prediction problem. But despite this obvious difference, the more accurate the solution returned by the algorithm, the better the performances for any levels of penalty and for both performance measures.

methods	quadrupen	glmnet low	glmnet med	glmnet high
timing (ms)	8	7	8	64
accuracy (dist. to opt.)	5.9×10^{-14}	7.2×10^{0}	6.04×10^{0}	1.47×10^{-2}

Table 2: Median timings and solution accuracies

Now focusing on `glmnet` performances, the better the accuracy, the smaller the MSE and the support error rate. But the better the accuracy, the slower the algorithm becomes. Using the default settings allows to have a result very close to our quadratic solver, and the performance differences become negligible between our approach and `glmnet` running with high precision. However, Table 2 illustrates that high accuracy is achieved at a high computational cost: to be at par with `quadrupen` with regards to test performances, `glmnet` is about ten times slower than our solver.

---

[6]This is done via the `thresh` argument of the `glmnet` procedure, whose default value is `1e-7`. In our experiments, `low`, `med` and `high` level of accuracy for `glmnet` respectively correspond to `thresh` set to `1e-1`, `1e-4`, and `1e-9`.
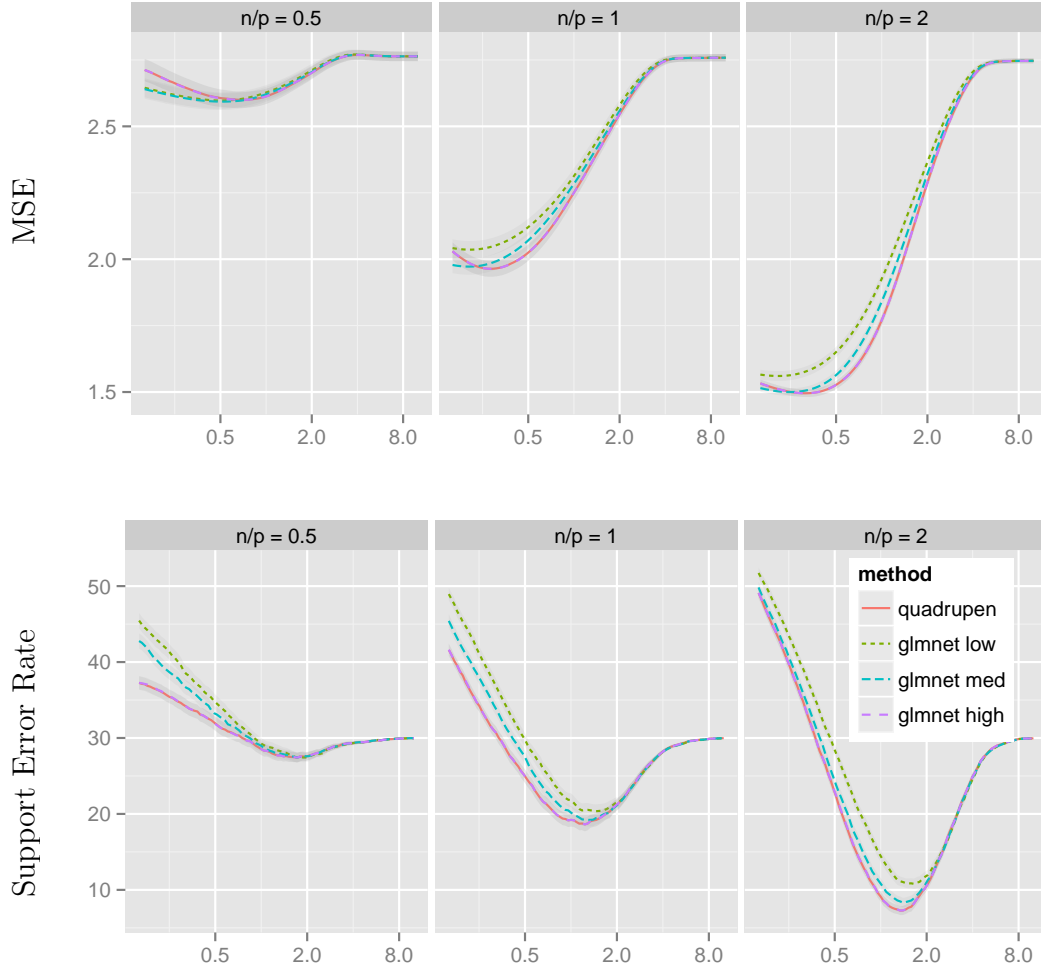
Figure 7: Test performances according to the penalty parameter for the Lasso estimates returned by `quadrupen` and `glmnet` at various level of accuracy. Three high-dimensional setups are illustrated: from left to right $n/p = 1/2$, $n/p = 1$ and $n/p = 2$; top: mean squared test error; bottom: support error rate.

6. Discussion

This paper presents a new viewpoint on sparsity-inducing penalties stemming from a quadratic view of the feasible sets associated with these penalties. This viewpoint can be seen as a dual formulation of the original problem where the dual variables defines a series of quadratic convex set the intersection of which defines the feasible set of the original problem.

This viewpoint enables to cast in the same framework several well-known penalties. In particular, we detailed how the Lasso and the group-Lasso (with the $\ell_{\infty,1}$ mixed norm), possibly applied together with an ℓ_2 ridge penalty (leading to what is known as the elastic net for the Lasso) can be derived.

We derived a general-purpose algorithm that computes the solution to their companion penalized regression problem, and to obtain a lower bound on the minimum of the objective function that provides an assessment of convergence. The proposed algorithm solves a series of quadratic problems defined from the dual variables. It has been thoroughly tested and compared with state-of-the-art implementations for the elastic net and the Lasso, comparing favorably in most cases to all its competitors.

From a practical viewpoint, an important feature of our approach is that it solves the original problem up to machine precision: we shown that when variable selection is involved, optimization with a high level of precision is mandatory to recover the true model.

Moreover, our algorithm is well suited for computing a solution path rather than for just one value of $\lambda_1, (\lambda_2)$ since we are generally interested in cross-validating these paths for model selection purpose (at least in genomics).

Regarding future development, the algorithm can be adapted to non-quadratic loss functions for addressing other learning problems such as classification, but this generalization, which requires solving non-quadratic problems, may not be as efficient compared to the existing alternatives. We are now examining how to address a wider range of penalties by extending the framework in two directions: first, to accommodate additional general ℓ_2 penalties in the form of arbitrary symmetric positive semidefinite matrix instead of the simple ridge, in particular to provide an efficient implementation of the structured elastic net (Slawski et al., 2010) ; second, we plan to derive

similar views on a wider range of sparsity-inducing penalties, such as the fused-Lasso or the OSCAR (Bondell and Reich, 2008).

Ambroise, C., McLachlan, G. J., 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. Proceedings of the National Academy of Sciences 99 (10), 6562–6566.

Bach, F., Jenatton, R., Mairal, J., Obozinski, G., 2012. Optimization with sparsity-inducing penalties. Foundations and Trends in Machine Learning 4 (1), 1–106.

Beck, A., Teboulle, M., 2009. Fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences 2, 183–202.

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

Bonnans, J. F., Gilbert, J. C., Lemaréchal, C., Sagastizábal, C., 2006. Numerical Optimization: Theoretical and Practical Aspects, 2nd Edition. Springer.

Bottou, L., Bousquet, O., 2008. The tradeoffs of large scale learning. In: Platt, J. C., Koller, D., Singer, Y., Roweis, S. (Eds.), Advances in Neural Information Processing Systems, 20 (NIPS 2008). pp. 161–168.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine Learning 3 (1), 1–122.

Boyd, S., Vandenberghe, L., 2004. Convex optimization. Cambridge university press.

Eddelbuettel, D., François, R., 4 2011. Rcpp: Seamless R and C++ integration. Journal of Statistical Software 40 (8), 1–18.
URL http://www.jstatsoft.org/v40/i08

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. The Annals of Statistics 32 (2), 407–499, with discussion, and a rejoinder by the authors.

Freedman, D. A., 1983. A note on screening regression equations. The American Statistician 37 (2), 152–155.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., 2007. Pathwise coordinate optimization. The Annals of Applied Statistics 1 (2), 302–332.

Friedman, J. H., Hastie, T., Tibshirani, R., 2 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33 (1), 1–22.
URL http://www.jstatsoft.org/v33/i01

Gilbert, J. C., 2016. Fragments d'optimisation différentiable – théorie et algorithmes.
URL http://www-rocq.inria.fr/~gilbert/ensta/optim.html

Lacoste-Julien, S., Jaggi, M., Schmidt, M., Pletscher, P., 2012. Block-coordinate Frank-Wolfe optimization for structural svms. arXiv preprint arXiv:1207.4747.

Mairal, J., 2010. Sparse coding for machine learning, image processing and computer vision. Ph.D. thesis, École Normale Supérieure de Cachan.

Moulines, E., Bach, F. R., 2011. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In: Advances in Neural Information Processing Systems. pp. 451–459.

Osborne, M. R., Presnell, B., Turlach, B. A., 2000. On the LASSO and its dual. Journal of Computational and Graphical Statistics 9 (2), 319–337.

Sanderson, C., 2010. Armadillo: An open source C++ linear algebra library for fast prototyping and computationally intensive experiments. Tech. rep., NICTA.
URL http://http://arma.sourceforge.net/

Simon, N., Friedman, J. H., Hastie, T., Tibshirani, R., 3 2011. Regularization paths for cox's proportional hazards model via coordinate descent. Journal of Statistical Software 39 (5), 1–13.
URL http://www.jstatsoft.org/v39/i05

Slawski, M., zu Castell, W., Tutz, G., 2010. Feature selection guided by structural information. The Annals of Applied Statistics 4 (2), 1056–1080.

Turlach, B. A., Venables, W. N., Wright, S. J., 2005. Simultaneous variable selection. Technometrics 47 (3), 349–363.

Zhao, P., Yu, B., 2006. On model selection consistency of lasso. Journal of Machine Learning Research 7, 2541–2567.

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

Appendix A. Proofs

Appendix A.1. Proof of Proposition **??**

We first show that the objective function of Problem (**??**) is, up to an irrelevant constant, upper bounded by the objective function of Problem (**??**):

$$\left\| (\mathbf{X} - \boldsymbol{\Delta}_{\mathbf{X}})\boldsymbol{\beta} + \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon} - \mathbf{y} \right\|_2 \leq \left\| \mathbf{X}(\boldsymbol{\beta} + \boldsymbol{\gamma}) - \mathbf{y} \right\|_2 + \left\| \boldsymbol{\Delta}_{\mathbf{X}}\boldsymbol{\beta} \right\|_2 + \left\| \boldsymbol{\epsilon} \right\|_2 \tag{A.1}$$

$$\leq \left\| \mathbf{X}(\boldsymbol{\beta} + \boldsymbol{\gamma}) - \mathbf{y} \right\|_2 + \eta_X \left\| \boldsymbol{\beta} \right\|_2 + \eta_\epsilon \tag{A.2}$$

where (A.3) stems from the triangular inequality and (A.4) follows from the definition of the uncertainty sets. The right-hand side of (A.4) is, up to an irrelevant constant, identical to the objective function of Problem (**??**) after the change of variable $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} + \boldsymbol{\gamma}$.

We conclude, by showing that, for all $(\boldsymbol{\beta}, \boldsymbol{\gamma})$, there is $(\boldsymbol{\Delta}_{\mathbf{X}}^\star, \boldsymbol{\epsilon}^\star) \in \mathcal{D}_{\mathbf{X}} \times \mathcal{D}_{\boldsymbol{\epsilon}}$ such that the upper bound on the right-hand side of (A.4) is reached:

$$\boldsymbol{\Delta}_{\mathbf{X}}^\star = \eta_X \frac{\mathbf{X}(\boldsymbol{\beta} + \boldsymbol{\gamma}) - \mathbf{y}}{\left\| \mathbf{X}(\boldsymbol{\beta} + \boldsymbol{\gamma}) - \mathbf{y} \right\|_2} \frac{\boldsymbol{\beta}^\mathsf{T}}{\left\| \boldsymbol{\beta} \right\|_2} \qquad \boldsymbol{\epsilon}^\star = \eta_\epsilon \frac{\mathbf{X}(\boldsymbol{\beta} + \boldsymbol{\gamma}) - \mathbf{y}}{\left\| \mathbf{X}(\boldsymbol{\beta} + \boldsymbol{\gamma}) - \mathbf{y} \right\|_2} .$$

With these choices, the vectors $\mathbf{X}(\boldsymbol{\beta} + \boldsymbol{\gamma}) - \mathbf{y}$, $\boldsymbol{\Delta}_{\mathbf{X}}^\star \boldsymbol{\beta}$ and $\boldsymbol{\epsilon}^\star$ are colinear, so that minimizing the left-hand-side of (A.3) is equivalent to minimize the right-hand-side of (A.4).

Appendix A.2. Proof of Proposition **??**

first show that the objective function of Problem (**??**) is, up to an irrelevant constant, upper bounded by the objective function of Problem (**??**):

$$\|(\mathbf{X} - \mathbf{\Delta_X})\boldsymbol{\beta} + \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon} - \mathbf{y}\|_2 \leq \|\mathbf{X}(\boldsymbol{\beta} + \boldsymbol{\gamma}) - \mathbf{y}\|_2 + \|\mathbf{\Delta_X}\boldsymbol{\beta}\|_2 + \|\boldsymbol{\epsilon}\|_2 \tag{A.3}$$

$$\leq \|\mathbf{X}(\boldsymbol{\beta} + \boldsymbol{\gamma}) - \mathbf{y}\|_2 + \eta_X \|\boldsymbol{\beta}\|_2 + \eta_\epsilon \tag{A.4}$$

where (A.3) stems from the triangular inequality and (A.4) follows from the definition of the uncertainty sets. The right-hand side of (A.4) is, up to an irrelevant constant, identical to the objective function of Problem (**??**) after the change of variable $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} + \boldsymbol{\gamma}$.

We conclude, by showing that, for all $(\boldsymbol{\beta}, \boldsymbol{\gamma})$, there is $(\mathbf{\Delta_X^\star}, \boldsymbol{\epsilon}^\star) \in \mathcal{D}_\mathbf{X} \times \mathcal{D}_\epsilon$ such that the upper bound on the right-hand side of (A.4) is reached:

$$\mathbf{\Delta_X^\star} = \eta_X \frac{\mathbf{X}(\boldsymbol{\beta} + \boldsymbol{\gamma}) - \mathbf{y}}{\|\mathbf{X}(\boldsymbol{\beta} + \boldsymbol{\gamma}) - \mathbf{y}\|_2} \frac{\boldsymbol{\beta}^\mathsf{T}}{\|\boldsymbol{\beta}\|_2} \qquad \boldsymbol{\epsilon}^\star = \eta_\epsilon \frac{\mathbf{X}(\boldsymbol{\beta} + \boldsymbol{\gamma}) - \mathbf{y}}{\|\mathbf{X}(\boldsymbol{\beta} + \boldsymbol{\gamma}) - \mathbf{y}\|_2} \; .$$

With these choices, the vectors $\mathbf{X}(\boldsymbol{\beta} + \boldsymbol{\gamma}) - \mathbf{y}$, $\mathbf{\Delta_X^\star}\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}^\star$ are colinear, so that minimizing the left-hand-side of (A.3) is equivalent to minimize the right-hand-side of (A.4).

Appendix A.3. Proof of Proposition 1

We detail here a proof yielding a slightly tighter bound. Proposition 1 is simply a corollary of Proposition 2 stated and proved below.

When $\mathcal{D}_\gamma$ is defined by a norm as in Proposition 1, the following Lemma relates the penalty associated with a infeasible $\boldsymbol{\gamma}$-value ($\|\boldsymbol{\gamma}\|_* > \eta_\gamma$) to the one obtained by shrinking this $\boldsymbol{\gamma}$-value to reach the boundary of $\mathcal{D}_\gamma$.

Lemma 1. *Let* $\mathcal{S} \subseteq \{1, ..., p\}$, $\alpha \in (0, 1)$, $\|\cdot\|_*$ *be a vectorial norm and* $\varphi^*(\cdot, \mathcal{S}, \alpha) : \mathbb{R}^p \to \mathbb{R}^p$ *be defined as follows:*

$$\begin{cases} \varphi_\mathcal{S}^*(\boldsymbol{\gamma}, \mathcal{S}, \alpha) = \boldsymbol{\gamma}_\mathcal{S} \\ \varphi_{\mathcal{S}^c}^*(\boldsymbol{\gamma}, \mathcal{S}, \alpha) = \alpha \boldsymbol{\gamma}_{\mathcal{S}^c} \end{cases} \; .$$

Then,

$$\|\boldsymbol{\beta} - \varphi^*(\boldsymbol{\gamma}, \mathcal{S}, \alpha)\|_2^2 \geq \alpha \|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2^2 - \alpha(1 - \alpha) \|\boldsymbol{\gamma}_{\mathcal{S}^c}\|_2^2 \; . \tag{A.5}$$

Proof.

$$\|\boldsymbol{\beta} - \varphi^*(\boldsymbol{\gamma}, \mathcal{S}, \alpha)\|_2^2 = \|\boldsymbol{\beta}_{\mathcal{S}} - \boldsymbol{\gamma}_{\mathcal{S}}\|_2^2 + \alpha \|\boldsymbol{\beta}_{\mathcal{S}^c} - \boldsymbol{\gamma}_{\mathcal{S}^c}\|_2^2 + (1 - \alpha) \|\boldsymbol{\beta}_{\mathcal{S}^c}\|_2^2 - \alpha(1 - \alpha) \|\boldsymbol{\gamma}_{\mathcal{S}^c}\|_2^2$$
$$\geq \|\boldsymbol{\beta}_{\mathcal{S}} - \boldsymbol{\gamma}_{\mathcal{S}}\|_2^2 + \alpha \|\boldsymbol{\beta}_{\mathcal{S}^c} - \boldsymbol{\gamma}_{\mathcal{S}^c}\|_2^2 - \alpha(1 - \alpha) \|\boldsymbol{\gamma}_{\mathcal{S}^c}\|_2^2$$
$$\geq \alpha \|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2^2 - \alpha(1 - \alpha) \|\boldsymbol{\gamma}_{\mathcal{S}^c}\|_2^2 \quad .$$

$\square$

Proposition 2. *For any $\eta_{\boldsymbol{\gamma}} > 0$, and for all vectorial norm $\|\cdot\|_*$, when $\mathcal{D}_{\boldsymbol{\gamma}}$ is defined as $\mathcal{D}_{\boldsymbol{\gamma}} = \{\boldsymbol{\gamma} \in \mathbb{R}^p : \|\boldsymbol{\gamma}\|_* \leq \eta_{\boldsymbol{\gamma}}\}$, then, $\forall \boldsymbol{\gamma} \in \mathbb{R}^p : \|\boldsymbol{\gamma}\|_* \geq \eta_{\boldsymbol{\gamma}}$, and $\forall(\mathcal{S}, \alpha) \in 2^{\{1,\dots,p\}} \times (0, 1)$ such that $\|(\boldsymbol{\gamma}_{\mathcal{S}}, \alpha \boldsymbol{\gamma}_{\mathcal{S}^c})\|_* \leq \eta_{\boldsymbol{\gamma}}$, we have:*

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \max_{\boldsymbol{\gamma}' \in \mathcal{D}_{\boldsymbol{\gamma}}} J_\lambda(\boldsymbol{\beta}, \boldsymbol{\gamma}') \geq \alpha J_\lambda(\boldsymbol{\beta}^\star(\boldsymbol{\gamma}), \boldsymbol{\gamma}) - \lambda \alpha (1 - \alpha) \|\boldsymbol{\gamma}_{\mathcal{S}^c}\|_2^2 \quad ,$$

where

$$J_\lambda(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda \|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2^2 \quad and \quad \boldsymbol{\beta}^\star(\boldsymbol{\gamma}) = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} J_\lambda(\boldsymbol{\beta}, \boldsymbol{\gamma}) \quad .$$

Proof. For all $\boldsymbol{\beta} \in \mathbb{R}^p$ and for any $(\boldsymbol{\gamma}, \mathcal{S}, \alpha) \in \mathbb{R}^p \times 2^{\{1,\dots,p\}} \times (0, 1)$ such that $\varphi^*(\boldsymbol{\gamma}, \mathcal{S}, \alpha) \in \mathcal{D}_{\boldsymbol{\gamma}}$, with φ^* defined as in Lemma 1 we have:

$$\max_{\boldsymbol{\gamma}' \in \mathcal{D}_{\boldsymbol{\gamma}}} J_\lambda(\boldsymbol{\beta}, \boldsymbol{\gamma}') \geq J_\lambda(\boldsymbol{\beta}, \varphi^*(\boldsymbol{\gamma}, \mathcal{S}, \alpha)) \quad , \tag{A.6}$$

since $\varphi^*(\boldsymbol{\gamma}, \mathcal{S}, \alpha)$ belongs to $\mathcal{D}_{\boldsymbol{\gamma}}$. We now compute a lower bound of the right-hand-side for $\boldsymbol{\gamma}$ such that $\|\boldsymbol{\gamma}\|_* \geq \eta_{\boldsymbol{\gamma}}$:

$$J_\lambda(\boldsymbol{\beta}, \varphi^*(\boldsymbol{\gamma}, \mathcal{S}, \alpha)) = \alpha \left(\frac{1}{\alpha} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \frac{\lambda}{\alpha} \|\boldsymbol{\beta} - \varphi^*(\boldsymbol{\gamma}, \mathcal{S}, \alpha)\|_2^2 \right)$$

$$\geq \alpha \left(\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \frac{\lambda}{\alpha} \|\boldsymbol{\beta} - \varphi^*(\boldsymbol{\gamma}, \mathcal{S}, \alpha)\|_2^2 \right) \tag{A.7}$$

$$\geq \alpha \left(\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda \|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2^2 \right) - \lambda \alpha (1 - \alpha) \|\boldsymbol{\gamma}_{\mathcal{S}^c}\|_2^2 \quad ,$$

where the last inequality stems from Lemma 1. This inequality holds for any given $\boldsymbol{\beta}$-value, in particular for $\boldsymbol{\beta}^\star(\varphi^*(\boldsymbol{\gamma}, \mathcal{S}, \alpha)) = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} J_\lambda(\boldsymbol{\beta}, \varphi^*(\boldsymbol{\gamma}, \mathcal{S}, \alpha))$:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} J_\lambda(\boldsymbol{\beta}, \varphi^*(\boldsymbol{\gamma}, \mathcal{S}, \alpha)) \geq \alpha J_\lambda(\boldsymbol{\beta}^\star(\varphi^*(\boldsymbol{\gamma}, \mathcal{S}, \alpha))) - \lambda \alpha (1 - \alpha) \|\boldsymbol{\gamma}_{\mathcal{S}^c}\|_2^2$$

$$\geq \alpha J_\lambda(\boldsymbol{\beta}^\star(\boldsymbol{\gamma}), \boldsymbol{\gamma}) - \lambda \alpha (1 - \alpha) \|\boldsymbol{\gamma}_{\mathcal{S}^c}\|_2^2 \quad , \tag{A.8}$$

30

where the second inequality follows from the definition of $\boldsymbol{\beta}^\star(\boldsymbol{\gamma})$. Inequality (A.8) can be restated as:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} J_\lambda\left(\boldsymbol{\beta}, \varphi^*(\boldsymbol{\gamma}, \mathcal{S}, \alpha)\right) \geq \alpha \min_{\boldsymbol{\beta}\in\mathbb{R}^p} J_\lambda(\boldsymbol{\beta}, \boldsymbol{\gamma}) - \lambda\alpha(1-\alpha)\|\boldsymbol{\gamma}_{\mathcal{S}^c}\|_2^2 \quad.$$

We finally remark that, since $\varphi^*(\boldsymbol{\gamma}, \mathcal{S}, \alpha) \in \mathcal{D}_{\boldsymbol{\gamma}}$, we trivially have:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \max_{\boldsymbol{\gamma}\in\mathcal{D}_{\boldsymbol{\gamma}}} J_\lambda(\boldsymbol{\beta}, \boldsymbol{\gamma}) \geq \min_{\boldsymbol{\beta}\in\mathbb{R}^p} J_\lambda\left(\boldsymbol{\beta}, \varphi^*(\boldsymbol{\gamma}, \mathcal{S}, \alpha)\right) \quad,$$

which concludes the proof. $\qquad\square$