Mind the duality gap: safer rules for the Lasso

Olivier Fercoq
Alexandre Gramfort
Joseph Salmon
Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI
46 rue Barrault, 75013, Paris, France

Abstract
Screening rules allow to early discard irrelevant variables from the optimization in Lasso problems, or its derivatives, making solvers faster. In this paper, we propose new versions of the so-called safe rules for the Lasso. Based on duality gap considerations, our new rules create safe test regions whose diameters converge to zero, provided that one relies on a converging solver. This property helps screening out more variables, for a wider range of regularization parameter values. In addition to faster convergence, we prove that we correctly identify the active sets (supports) of the solutions in finite time. While our proposed strategy can cope with any solver, its performance is demonstrated using a coordinate descent algorithm particularly adapted to machine learning use cases. Significant computing time reductions are obtained with respect to previous safe rules.

1. Introduction
Since the mid 1990’s, high dimensional statistics has attracted considerable attention, especially in the context of linear regression with more explanatory variables than observations: the so-called \( p > n \) case. In such a context, the least squares with \( \ell_1 \) regularization, referred to as the Lasso (Tibshirani, 1996) in statistics, or Basis Pursuit (Chen et al., 1998) in signal processing, has been one of the most popular tools. It enjoys theoretical guarantees (Bickel et al., 2009), as well as practical benefits: it provides sparse solutions and fast convex solvers are available. This has made the Lasso a popular method in modern data-science toolkits. Among successful fields where it has been applied, one can mention dictionary learning (Mairal, 2010), bio-statistics (Haury et al., 2012) and medical imaging (Lustig et al., 2007; Gramfort et al., 2012) to name a few.

Many algorithms exist to approximate Lasso solutions, but it is still a burning issue to accelerate solvers in high dimensions. Indeed, although some other variable selection and prediction methods exist (Fan & Lv, 2008), the best performing methods usually rely on the Lasso. For stability selection methods (Meinshausen & Bühlmann, 2010; Bach, 2008; Varoquaux et al., 2012), hundreds of Lasso problems need to be solved. For non-convex approaches such as SCAD (Fan & Li, 2001) or MCP (Zhang, 2010), solving the Lasso is often a required preliminary step (Zou, 2006; Zhang & Zhang, 2012; Candès et al., 2008).

Among possible algorithmic candidates for solving the Lasso, one can mention homotopy methods (Osborne et al., 2000), LARS (Efron et al., 2004), and approximate homotopy (Mairal & Yu, 2012), that provide solutions for the full Lasso path, i.e., for all possible choices of tuning parameter \( \lambda \). More recently, particularly for \( p > n \), coordinate descent approaches (Friedman et al., 2007) have proved to be among the best methods to tackle large scale problems.

Following the seminal work by El Ghaoui et al. (2012), screening techniques have emerged as a way to exploit the known sparsity of the solution by discarding features prior to starting a Lasso solver. Such techniques are coined safe rules when they screen out coefficients guaranteed to be zero in the targeted optimal solution. Zeroing those coefficients allows to focus more precisely on the non-zero ones (likely to represent signal) and helps reducing the computational burden. We refer to (Xiang et al., 2014) for a concise introduction on safe rules. Other alternatives have tried to screen the Lasso relaxing the “safety”. Potentially, some variables are wrongly disregarded and post-processing is needed to recover them. This is for instance the strategy adopted for the strong rules (Tibshirani et al., 2012).

The original basic safe rules operate as follows: one
chooses a fixed tuning parameter $\lambda$, and before launching any solver, tests whether a coordinate can be zeroed or not (equivalently if the corresponding variable can be disregarded or not). We will refer to such safe rules as *static safe rules*. Note that the test is performed according to a safe region, *i.e.*, a region containing a dual optimal solution of the Lasso problem. In the static case, the screening is performed only once, prior any optimization iteration. Two directions have emerged to improve on static strategies.

- The first direction is oriented towards the resolution of the Lasso for a large number of tuning parameters. Indeed, practitioners commonly compute the Lasso over a grid of parameters and select the best one in a data-driven manner, *e.g.*, by cross-validation. As two consecutive $\lambda$’s in the grid lead to similar solutions, knowing the first solution may help improve screening for the second one. We call *sequential safe rules* such strategies, also referred to as recursive safe rules in (El Ghaoui et al., 2012). This road has been pursued in (Wang et al., 2013; Xu & Ramadge, 2013; Xiang et al., 2014), and can be thought of as a “warm start” of the screening (in addition to the warm start of the solution itself). When performing sequential safe rules, one should keep in mind that generally, only an approximation of the previous dual solution is computed. Though, the safety of the rule is guaranteed only if one uses the exact solution. Neglecting this issue, leads to “unsafe” rules: relevant variables might be wrongly disregarded.

- The second direction aims at improving the screening by interlacing it throughout the optimization algorithm itself: although screening might be useless at the beginning of the algorithm, it might become (more) efficient as the algorithm proceeds towards the optimal solution. We call these strategies *dynamic safe rules* following (Bonnefoy et al., 2014a;b).

Based on convex optimization arguments, we leverage duality gap computations to propose a simple strategy unifying both sequential and dynamic safe rules. We coined GAP SAFE rules such safe rules.

The main contributions of this paper are 1) the introduction of new safe rules which demonstrate a clear practical improvement compared to prior strategies 2) the definition of a theoretical framework for comparing safe rules by looking at the convergence of their associated safe regions.

In Section 2, we present the framework and the basic concepts which guarantee the soundness of static and dynamic screening rules. Then, in Section 3, we introduce the new concept of converging safe rules. Such rules identify in finite time the active variables of the optimal solution (or equivalently the inactive variables), and the tests become more and more precise as the optimization algorithm proceeds. We also show that our new GAP SAFE rules, built on dual gap computations, are converging safe rules since their associated safe regions have a diameter converging to zero. We also explain how our GAP SAFE tests are sequential by nature. Application of our GAP SAFE rules with a coordinate descent solver for the Lasso problem is proposed in Section 4. Using standard data-sets, we report the time improvement compared to prior safe rules.

1.1. Model and notation

We denote by $[d]$ the set $\{1, \ldots, d\}$ for any integer $d \in \mathbb{N}$. Our observation vector is $y \in \mathbb{R}^n$ and the design matrix $X = [x_1, \ldots, x_p] \in \mathbb{R}^{n \times p}$ has $p$ explanatory variables (or features) column-wise. We aim at approximating $y$ as a linear combination of few variables $x_j$’s, hence expressing $y$ as $X \beta$ where $\beta \in \mathbb{R}^p$ is a sparse vector. The standard Euclidean norm is written $\|\cdot\|_2$, the $\ell_1$ norm $\|\cdot\|_1$, the $\ell_\infty$ norm $\|\cdot\|_\infty$, and the matrix transposition of a matrix $Q$ is denoted by $Q^T$. We denote $(t)_+ = \max(0, t)$.

For such a task, the Lasso is often considered (see Bühlmann & van de Geer (2011) for an introduction). For a tuning parameter $\lambda > 0$, controlling the trade-off between data fidelity and sparsity of the solutions, a Lasso estimator $\hat{\beta}(\lambda)$ is any solution of the primal optimization problem

$$\hat{\beta}(\lambda) = \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{2} \|X\beta - y\|^2_2 + \lambda \|\beta\|_1. \quad (1)$$

Denoting $\Delta_X = \{ \theta \in \mathbb{R}^n \mid x_j^T \theta \leq 1, \forall j \in [p]\}$ the dual feasible set, a dual formulation of the Lasso reads (see for instance Kim et al. (2007) or Xiang et al. (2014)):

$$\hat{\theta}(\lambda) = \arg \max_{\theta \in \Delta_X \subset \mathbb{R}^n} \frac{1}{2} \|y - \lambda \theta\|^2_2 - \frac{\lambda^2}{2} \|\theta\|_2^2. \quad (2)$$

We can reinterpret Eq. (2) as $\hat{\theta}(\lambda) = \Pi_{\Delta_X}(y/\lambda)$, where $\Pi_{\Delta}$ refers to the projection onto a closed convex set $\Delta$. In particular, this ensures that the dual solution $\hat{\theta}(\lambda)$ is always unique, contrarily to the primal $\hat{\beta}(\lambda)$.

1.2. A KKT detour

For the Lasso problem, a primal solution $\hat{\beta}(\lambda) \in \mathbb{R}^p$ and the dual solution $\hat{\theta}(\lambda) \in \mathbb{R}^n$ are linked through the relation:

$$y = X \hat{\beta}(\lambda) + \lambda \hat{\theta}(\lambda). \quad (3)$$

The Karush-Kuhn-Tucker (KKT) conditions state:

$$\forall j \in [p], \ x_j^T \hat{\theta}(\lambda) \in \{ \text{sign}(\hat{\beta}_j^{(\lambda)}) \} \quad \text{if} \quad \hat{\beta}_j^{(\lambda)} \neq 0, \quad \text{and} \quad \{ -1, 1 \} \quad \text{if} \quad \hat{\beta}_j^{(\lambda)} = 0. \quad (4)$$
Mind the duality gap: safer rules for the Lasso

<table>
<thead>
<tr>
<th>Rule</th>
<th>Center</th>
<th>Radius</th>
<th>Ingredients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static Safe (El Ghaoui et al., 2012)</td>
<td>( y/\lambda )</td>
<td>( R_{\lambda} \left( \frac{y}{\lambda \max} \right) )</td>
<td>( \lambda_{\max} =</td>
</tr>
<tr>
<td>Dynamic ST3 (Xiang et al., 2011)</td>
<td>( y/\lambda - \delta x_j )</td>
<td>( \left( R_{\lambda}(\theta_k) + \delta^2 \right)^{1/2} )</td>
<td>( \delta = \left( \frac{\lambda_{\max}}{\lambda} - 1 \right) )</td>
</tr>
<tr>
<td>Dynamic Safe (Bonnefoy et al., 2014a)</td>
<td>( y/\lambda )</td>
<td>( R_{\lambda}(\theta_k) )</td>
<td>( \theta_k = \Delta X ) (e.g., as in (11))</td>
</tr>
<tr>
<td>Sequential (Wang et al., 2013)</td>
<td>( \hat{\theta}^{(\lambda-1)} )</td>
<td>( \frac{1}{\lambda_{\max} - \frac{1}{\lambda}} |y| )</td>
<td>exact ( \hat{\theta}^{\lambda-1} ) required</td>
</tr>
<tr>
<td>GAP SAFE sphere (proposed)</td>
<td>( \theta_k )</td>
<td>( x_{\lambda}(\beta_k, \theta_k) = \frac{1}{\lambda_{\max}} \sqrt{2G_{\lambda}(\beta_k, \theta_k)} )</td>
<td>dual gap for ( \beta_k, \theta_k )</td>
</tr>
</tbody>
</table>

Table 1. Review of some common safe sphere tests.

See for instance (Xiang et al., 2014) for more details. The KKT conditions lead to the fact that for \( \lambda \geq \lambda_{\max} = |X^Ty|/x \), 0 \( \in \mathbb{R}^p \) is a primal solution. It can be considered as the mother of all safe screening rules. So from now on, we assume that \( \lambda \leq \lambda_{\max} \) for all the considered \( \lambda \)’s.

2. Safe rules

Safe rules exploit the KKT condition (4). This equation implies that \( \hat{\beta}^{(\lambda)} = 0 \) as soon as \( |x_j^T\hat{\theta}^{(\lambda)}| < 1 \). The main challenge is that the dual optimal solution is unknown. Hence, a safe rule aims at constructing a set \( C \subset \mathbb{R}^n \) containing \( \hat{\beta}^{(\lambda)} \). We call such a set \( C \) a safe region. Safe regions are all the more helpful that for many \( j \)’s, \( \mu_c(x_j) := \sup_{\theta \in \mathcal{C}} |x_j^T\theta| < 1 \), hence for many \( j \)’s, \( \hat{\beta}^{(\lambda)} = 0 \).

Practical benefits are obtained if one can construct a region \( C \) for which it is easy to compute its support function, denoted by \( \sigma_C \) and defined for any \( x \in \mathbb{R}^n \) by:

\[
\sigma_C(x) = \max_{\theta \in \mathcal{C}} x^T \theta. \tag{5}
\]

Cast differently, for any safe region \( C \), any \( j \in [p] \), and any primal optimal solution \( \beta^{(\lambda)} \), the following holds true:

If \( \mu_c(x_j) = \max(\sigma_c(x_j), \sigma_c(-x_j)) < 1 \) then \( \hat{\beta}^{(\lambda)} = 0 \). \tag{6}

We call safe test or safe rule, a test associated to \( C \) and screening out explanatory variables thanks to Eq. (6).

Remark 1. Reminding that the support function of a set is the same as the support function of its closed convex hull (Hiriart-Urruty & Lemaréchal, 1993)[Proposition V.2.2.1], we restrict our search to closed convex safe regions.

Based on a safe region \( C \) one can partition the explanatory variables into a safe active set \( A^{(\lambda)}(C) \) and a safe zero set \( Z^{(\lambda)}(C) \) where:

\[
A^{(\lambda)}(C) = \{ j \in [p] : \mu_c(x_j) \geq 1 \}, \tag{7}
\]

\[
Z^{(\lambda)}(C) = \{ j \in [p] : \mu_c(x_j) < 1 \}. \tag{8}
\]

Note that for nested safe regions \( C_1 \subset C_2 \) then \( A^{(\lambda)}(C_1) \subset A^{(\lambda)}(C_2) \). Consequently, a natural goal is to find safe regions as small as possible: narrowing safe regions can only increase the number of screened out variables.

Remark 2. If \( C = \{ \hat{\beta}(\lambda) \} \), the safe active set is the equicorrelation set \( A^{(\lambda)}(C) = \mathcal{E}_\lambda := \{ j \in [p] : |x_j^T\hat{\theta}(\lambda)| = 1 \} \) (in most cases (Tibshirani, 2013) it is exactly the active set of \( \hat{\beta}(\lambda) \)). If the Lasso has a unique solution, its support is exactly the equicorrelation set. If it is not unique, the equicorrelation set contains all the solutions’ supports and there exists a Lasso solution whose support is exactly this set (Tibshirani, 2013)[Lemma 12]. The other extreme case is when \( C = \Delta X \), and \( A^{(\lambda)}(C) = [p] \). Here, no variable is screened out: \( Z^{(\lambda)}(C) = \emptyset \) and the screening is useless.

We now consider common safe regions whose support functions are easy to obtain in closed form. For simplicity we focus only on balls and domes, though more complicated regions could be investigated (Xiang et al., 2014).

2.1. Sphere tests

Following previous work on safe rules, we call sphere tests, tests relying on balls as safe regions. For a sphere test, one chooses a ball containing \( \hat{\beta}(\lambda) \) with center \( c \) and radius \( r \), i.e., \( C = B(c, r) \). Due to their simplicity, safe spheres have been the most commonly investigated safe regions (see for instance Table 1 for a brief review). The corresponding test is defined as follows:

If \( \mu_{B(c,r)}(x_j) = |x_j^Tc| + r|x_j| < 1 \), then \( \hat{\beta}^{(\lambda)} = 0 \). \tag{9}

Note that for a fixed center, the smaller the radius, the better the safe screening strategy.

Example 1. The first introduced sphere test (El Ghaoui et al., 2012) consists in using the center \( c = y/\lambda \) and radius \( r = |1/\lambda - 1/\lambda_{\max}|\|y\| \). Given that \( \hat{\beta}(\lambda) = \Pi_{\Delta X}(y/\lambda) \), this is a safe region since \( y/\lambda_{\max} \in \Delta X \) and \( \|y/\lambda_{\max} - \Pi_{\Delta X}(y/\lambda)\| \leq \|y\|/1 - 1/\lambda_{\max} \). However, one can check that this static safe rule is useless as soon as

\[
\frac{\lambda}{\lambda_{\max}} \leq \min_{j \in [p]} \left( \frac{1 + |x_j^Ty|/\|x_j\|/\|y\|)}{1 + \lambda_{\max}/\|x_j\|/\|y\|} \right). \tag{10}
\]
Remark 3. The dual feasible point \( \theta \in \text{rel} \) immediately provides a ball that contains \( \tilde{\theta}(\lambda) \) since
\[
\left\| \tilde{\theta}(\lambda) - y \right\| = \min_{\theta' \in \hat{\Delta}_X} \left\| \theta' - y \right\| \leq \left\| \theta - y \right\| := \hat{R}_\lambda(\theta).
\]

The ball \( B\left(y, \hat{R}_\lambda(\theta_k)\right) \) corresponds to the simplest safe region introduced in (Bonneyfo et al., 2014a;b) (cf Figure 2 for more insights). When the algorithm proceeds, one expects that \( \theta_k \) gets closer to \( \tilde{\theta}(\lambda) \), so \( \|\theta_k - y/\lambda\| \) should get closer to \( \|\tilde{\theta}(\lambda) - y/\lambda\| \). Similarly to Example 1, this dynamic rule becomes useless once \( \lambda \) is too small. More precisely, this occurs as soon as
\[
\frac{\lambda}{\lambda_{\max}} \leq \min_{j \in [p]} \left( \frac{1 + |x_j^T y|/\|x_j\|/\|y\|}{\lambda_{\max}/\|\tilde{\theta}(\lambda)\|/\|y\| + \lambda_{\max}/\|x_j\|/\|y\|} \right).
\]

Noticing that \( \|\tilde{\theta}(\lambda)\| \leq \|y/\lambda\| \) (since \( \Pi_{\hat{\Delta}_X} \) is a contraction and \( 0 \in \hat{\Delta}_X \)) and proceeding as for (10), one can show that this dynamic safe rule is inefficient when:
\[
\frac{\lambda}{\lambda_{\max}} \leq \min_{j \in [p]} \left( \frac{|x_j^T y|/\|x_j\|/\|y\|}{\lambda_{\max}} \right).
\]

This is a critical threshold, yet the screening might stop even at a larger \( \lambda \) thanks to Eq. (13). In practice the bound in Eq. (13) cannot be evaluated a priori due to the term \( \|\tilde{\theta}(\lambda)\| \). Note also that the bound in Eq. (14) is close to the one in Eq. (10), explaining the similar behavior observed in our experiments (see Figure 3 for instance).

3. New contributions on safe rules

3.1. Support discovery in finite time

Let us first introduce the notions of converging safe regions and converging safe tests.

Definition 1. Let \( \{C_k\}_{k \in \mathbb{N}} \) be a sequence of closed convex sets in \( \mathbb{R}^n \) containing \( \tilde{\theta}(\lambda) \). It is a converging sequence of safe regions for the Lasso with parameter \( \lambda \) if the diameters of the sets converge to zero. The associated safe screening rules are referred to as converging safe tests.

Not only converging safe regions are crucial to speed up computation, but they are also helpful to reach exact active set identification in a finite number of steps. More precisely, we prove that one recovers the equicorrelation set of the Lasso (cf. Remark 2) in finite time with any converging strategy: after a finite number of steps, the equicorrelation set \( \mathcal{E}_\lambda \) is exactly identified. Such a property is
Theorem 1. Let \((C_k)_{k \in \mathbb{N}}\) be a sequence of converging safe regions. The estimated support provided by \(C_k\),
\[ A^{(\lambda)}(C_k) = \{ j \in [p]: \max_{x \in C_k} \{ x^\top x \} \geq 1 \}, \]
satisfies \(\lim_{k \to \infty} A^{(\lambda)}(C_k) = \mathcal{E}_\lambda\), and there exists \(k_0 \in \mathbb{N}\) such that for all \(k \geq k_0\), one gets \(A^{(\lambda)}(C_k) = \mathcal{E}_\lambda\).

Proof. The main idea of the proof is to use that \(\lim_{k \to \infty} C_k = \{ \hat{\theta}^{(\lambda)} \}, \lim_{k \to \infty} \mu_{C_k}(x) = \mu_{\{ \hat{\theta}^{(\lambda)} \}}(x) = |x^\top \hat{\theta}^{(\lambda)}|\) and that the set \(A^{(\lambda)}(C_k)\) is discrete. Details are delayed to the Appendix.

Remark 5. A more general result is proved for a specific algorithm (Forward-Backward) in Liang et al. (2014). Interestingly, our scheme is independent of the algorithm considered (e.g., Forward-Backward (Beck & Teboulle, 2009), Primal Dual (Chambolle & Pock, 2011), coordinate-descent (Tseng, 2001; Friedman et al., 2007)) and relies on the convergence of a sequence of safe regions.

3.2. GAP SAFE regions: leveraging the duality gap

In this section, we provide new dynamic safe rules built on converging safe regions.

Theorem 2. Let us take any \((\beta, \theta) \in \mathbb{R}^p \times \Delta_X\). Denote \(\tilde{R}_\lambda(\beta) := \frac{1}{\lambda}(\|y\|^2 - \|X\beta - y\|^2 - 2\lambda \|\beta\|_1^{1/2}), \tilde{R}_\lambda(\theta) := \|\theta - y/\lambda\|, \tilde{\theta}^{(\lambda)}\) the dual optimal Lasso solution and \(\tilde{\theta}_\lambda(\beta, \theta) := \sqrt{\tilde{R}_\lambda(\theta)^2 - \tilde{R}_\lambda(\beta)^2}\), then
\[ \hat{\theta}^{(\lambda)} \in B\left(\theta, \tilde{\theta}_\lambda(\beta, \theta)\right). \]

Proof. The construction of the ball \(B(\theta, \tilde{\theta}_\lambda(\beta, \theta))\) is based on the weak duality theorem (cf. (Rockafellar & Wets, 1998) for a reminder on weak and strong duality). Fix \(\theta \in \Delta_X\) and \(\beta \in \mathbb{R}^p\), then it holds that
\[ \frac{1}{2} \|y\|^2 - \frac{\lambda}{2} \|\theta - y/\lambda\|^2 \leq \frac{1}{2} \|X\beta - y\|^2 + \lambda \|\beta\|_1. \]
Hence,
\[ \|\theta - y/\lambda\| \geq \sqrt{\left(\|y\|^2 - \|X\beta - y\|^2 - 2\lambda \|\beta\|_1\right)/\lambda}. \]

In particular, this provides \(\|\hat{\theta}^{(\lambda)} - y/\lambda\| \geq \tilde{R}_\lambda(\beta)\). Combining (12) and (16), asserts that \(\hat{\theta}^{(\lambda)}\) belongs to the annulus \(A(y/\lambda, \tilde{R}_\lambda(\theta), \tilde{R}_\lambda(\beta))\) := \(\{ z \in \mathbb{R}^n: \tilde{R}_\lambda(\beta) \leq \|z - y/\lambda\| \leq \tilde{R}_\lambda(\theta) \}\) (the light blue zone in Figure 2).

Remind that the dual feasible set \(\Delta_X\) is convex, hence \(\Delta_X \cap B(y/\lambda, \tilde{R}_\lambda(\theta))\) is also convex. Thanks to (16), \(\Delta_X \cap B(y/\lambda, \tilde{R}_\lambda(\theta)) = \Delta_X \cap A(y/\lambda, \tilde{R}_\lambda(\theta), \tilde{R}_\lambda(\beta))\), and then \(\Delta_X \cap A(y/\lambda, \tilde{R}_\lambda(\theta), \tilde{R}_\lambda(\beta))\) is convex too. Hence, \(\hat{\theta}^{(\lambda)}\) is inside the annulus \(A(y/\lambda, \tilde{R}_\lambda(\theta), \tilde{R}_\lambda(\beta))\) and so is \([\theta, \tilde{\theta}_\lambda(\beta)] \subseteq A(y/\lambda, \tilde{R}_\lambda(\theta), \tilde{R}_\lambda(\beta))\) by convexity (see Figure 2.(a) and Figure 2.(b)). Moreover, \(\hat{\theta}^{(\lambda)}\) is the point of \([\theta, \tilde{\theta}_\lambda(\beta)]\) which is closest to \(y/\lambda\). The farthest where \(\tilde{\theta}_\lambda(\beta)\) can be according to this information would be if \([\theta, \tilde{\theta}_\lambda(\beta)]\) were tangent to the inner ball \(B(y/\lambda, \tilde{R}_\lambda(\beta))\) and \(\|\hat{\theta}^{(\lambda)} - y/\lambda\| = \tilde{R}_\lambda(\beta)\). Let us denote \(\theta_{\text{int}}\) such a point. The tangency property reads \(\|\theta_{\text{int}} - y/\lambda\| = \tilde{R}_\lambda(\beta)\) and \((\theta - \theta_{\text{int}})^\top (y/\lambda - \theta_{\text{int}}) = 0\). Hence, with the later and the definition of \(\tilde{R}_\lambda(\theta)\), \(\|\theta - y/\lambda\|^2 = \|\theta - \theta_{\text{int}}\|^2 + \|\theta_{\text{int}} - y/\lambda\|^2\) and \(\|\theta - \theta_{\text{int}}\|^2 = (\tilde{R}_\lambda(\theta)^2 - \tilde{R}_\lambda(\beta)^2)/2\).

Since by construction \(\hat{\theta}^{(\lambda)}\) cannot be further away from \(\theta\) than \(\theta_{\text{int}}\) (again, insights can be gleaned from Figure 2), we conclude that \(\hat{\theta}^{(\lambda)} \in B(\theta, (\tilde{R}_\lambda(\theta)^2 - \tilde{R}_\lambda(\beta)^2)^{1/2})\).

Remark 6. Choosing \(\beta = 0\) and \(\theta = y/\lambda_{\text{max}}\), then one recovers the static safe rule given in Example 1.
With the definition of the primal (resp. dual) objective for $P_{\lambda}(\beta)$, (resp. $D_{\lambda}(\theta)$), the duality gap reads as $G_{\lambda}(\beta, \theta) = P_{\lambda}(\beta) - D_{\lambda}(\theta)$. Remind that if $G_{\lambda}(\beta, \theta) \leq \epsilon$, then one has $P_{\lambda}(\beta) - P_{\lambda}(\hat{\beta}(\lambda)) \leq \epsilon$, which is a standard stopping criterion for Lasso solvers. The next proposition establishes a connection between the radius $r_{\lambda}(\beta, \theta)$ and the duality gap $G_{\lambda}(\beta, \theta)$.

**Proposition 1.** For any $(\beta, \theta) \in \mathbb{R}^p \times \Delta_X$, the following holds

$$\hat{r}_{\lambda}(\beta, \theta)^2 \leq r_{\lambda}(\beta, \theta)^2 := \frac{2}{\lambda^2} G_{\lambda}(\beta, \theta).$$  \hspace{1cm} (17)

**Proof.** Use the fact that $\hat{R}_{\lambda}(\theta)^2 = \|\theta - y/\lambda\|^2$ and $R_{\lambda}(\beta)^2 \geq \|y\|^2 - \|X\beta - y\|^2 - 2\lambda\|\beta\|_1/\lambda^2$. \hfill $\square$

If we could choose the “oracle” $\theta = \hat{\theta}(\lambda)$ and $\beta = \hat{\beta}(\lambda)$ in (15) then we would obtain a zero radius. Since those quantities are unknown, we rather pick dynamically the current available estimates given by an optimization algorithm: $\beta = \beta_k$ and $\theta = \theta_k$ as in Eq. (11). Introducing GAP SAFE spheres and domes as below, Proposition 2 ensures that they are converging safe regions.

**GAP SAFE sphere:**

$$C_k = B(\theta_k, r_{\lambda}(\beta, \theta)).$$ \hspace{1cm} (18)

**GAP SAFE dome:**

$$C_k = D\left(\frac{\theta_k + \theta_k}{2}, \frac{\theta_k}{2}, \frac{\sqrt{2}}{2} \left(\frac{R_{\lambda}(\beta_k)}{R_{\lambda}(\theta_k)}\right)^2 - 1, \frac{\theta_k}{2}, \frac{\theta_k}{2} \right).$$ \hspace{1cm} (19)

**Proposition 2.** For any converging primal sequence $(\beta_k)_{k \in \mathbb{N}_0}$ and dual sequence $(\theta_k)_{k \in \mathbb{N}_0}$ defined as in Eq. (11), then the GAP SAFE sphere and the GAP SAFE dome are converging safe regions.

**Proof.** For the GAP SAFE sphere the result follows from strong duality, Remark 4 and Proposition 1 yield $\lim_{k \to \infty} r_{\lambda}(\beta_k, \theta_k) = 0$, since $\lim_{k \to \infty} \theta_k = \hat{\theta}(\lambda)$ and $\lim_{k \to \infty} \beta_k = \hat{\beta}(\lambda)$. For the GAP SAFE dome, one can check that it is included in the GAP SAFE sphere, therefore inherits the convergence (see also Figure 2(c) and (d)). \hfill $\square$

**Remark 7.** The radius $r_{\lambda}(\beta_k, \theta_k)$ can be compared with the radius considered for the Dynamic Safe rule and Dynamic ST3 (Bonnefoy et al., 2014a) respectively: $\hat{R}_{\lambda}(\theta) = \|\theta - y/\lambda\|^2$ and $(\hat{R}_{\lambda}(\theta_k)^2 - \delta^2)^{1/2}$, where $\delta = (\lambda_{\max}/\lambda - 1)/\|x_j\|$. We have proved that $\lim_{k \to \infty} r_{\lambda}(\beta_k, \theta_k) = 0$, but a weaker property is satisfied by the two other radius: $\lim_{k \to \infty} \hat{R}_{\lambda}(\theta_k) = \hat{R}_{\lambda}(\hat{\theta}(\lambda)) = \|\hat{R}_{\lambda}(\hat{\theta}(\lambda)) - y/\lambda\|^2$ and $\lim_{k \to \infty}(\hat{R}_{\lambda}(\theta_k)^2 - \delta^2)^{1/2} = (\hat{R}_{\lambda}(\hat{\theta}(\lambda))^2 - \delta^2)^{1/2} > 0$.

**3.3. GAP SAFE rules : sequential for free**

As a byproduct, our dynamic screening tests provide a warm start strategy for the safe regions, making our GAP SAFE rules inherently sequential. The next proposition shows their efficiency when attacking a new tuning parameter, after having solved the Lasso for a previous $\lambda$, even only approximately. Handling approximate solutions is a critical issue to produce safe sequential strategies: without taking into account the approximation error, the screening might disregard relevant variables, especially the one near the safe regions boundaries. Except for $\lambda_{\max}$, it is unrealistic to assume that one can dispose of exact solutions.

Consider $\lambda_0 = \lambda_{\max}$ and a non-increasing sequence of $T - 1$ tuning parameters $\lambda_t \in [0, \lambda_{\max})$. In practice, we choose the common grid (Bühlmann & van de Geer, 2011)(2.12.11): $\lambda_t = \lambda_0 10^{-t(\delta/(T-1))}$ (for instance in Figure 3, we considered $\delta = 3$). The next result controls how the duality gap, or equivalently, the diameter of our GAP SAFE regions, evolves from $\lambda_{t-1}$ to $\lambda_t$.

**Proposition 3.** Suppose that $t \geq 1$ and $(\beta, \theta) \in \mathbb{R}^p \times \Delta_X$. Reminding $r_{\lambda_t}^2(\beta, \theta) = 2G_{\lambda_t}(\beta, \theta)/\lambda_t^2$, the following holds

$$r_{\lambda_t}^2(\beta, \theta) = \left(\frac{\lambda_{t-1}}{\lambda_t} - 1\right) \frac{\lambda_{t-1}}{\lambda_t} \|X \beta - y\|^2 - \left(\frac{\lambda_{t-1}}{\lambda_t} - 1\right) \|\theta\|^2.$$ \hspace{1cm} (20)

**Proof.** Details are given in the Appendix. \hfill $\square$

This proposition motivates to screen sequentially as follows: having $(\beta, \theta) \in \mathbb{R}^p \times \Delta_X$ such that $G_{\lambda_{t-1}}(\beta, \theta) \leq \epsilon$, then, we can screen using the GAP SAFE sphere with center $\theta$ and radius $r_{\lambda_t}(\beta, \theta)$. The adaptation to the GAP SAFE dome is straightforward and consists in replacing $\theta_k, \beta_k, \lambda$ by $\beta, \theta, \lambda_t$ in the GAP SAFE dome definition.

**Remark 8.** The basic sphere test of (Wang et al., 2013) requires the exact dual solution $\theta = \hat{\theta}(\lambda_{t-1})$ for center, and has radius $1/\lambda_t - 1/\lambda_{t-1}$ $\|y\|$, which is strictly larger than ours. Indeed, if one has access to dual and primal optimal solutions at $\lambda_{t-1}$, i.e., $(\beta, \theta) = (\hat{\theta}(\lambda_{t-1}), \hat{\theta}(\lambda_{t-1}))$, then $r_{\lambda_{t-1}}^2(\beta, \theta) = 0$ and

$$r_{\lambda_t}^2(\beta, \theta) = \left(\frac{\lambda_{t-1}}{\lambda_t} - 1\right) \left(\frac{\lambda_{t-1}}{\lambda_t} - 1\right) \|\theta\|^2,$$

$$\leq \left(\frac{1}{\lambda_t} - 1\right)^2 \|y\|^2,$$

since $\|\theta\| \leq \|y\|/\lambda_{t-1}$ for $\theta = \hat{\theta}(\lambda_{t-1})$.

Note that contrarily to former sequential rules (Wang et al., 2013), our introduced GAP SAFE rules still work when one has only access to approximations of $\hat{\theta}(\lambda_{t-1})$. 
4. Experiments

4.1. Coordinate Descent

Screening procedures can be used with any optimization algorithm. We chose coordinate descent because it is well suited for machine learning tasks, especially with sparse and/or unstructured design matrix $X$. Coordinate descent requires to extract efficiently columns of $X$ which is typically not easy in signal processing applications where $X$ is commonly an implicit operator (e.g. Fourier or wavelets).

Algorithm 1 Coordinate descent with GAP SAFE rules

<table>
<thead>
<tr>
<th>Input</th>
<th>$X, y, \epsilon, K, f, (\lambda_t)_{t=1}^{T-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initialization:</td>
<td>$\lambda_0 = \lambda_{\text{max}}$</td>
</tr>
<tr>
<td>$\beta^{0}_t = 0$</td>
<td></td>
</tr>
<tr>
<td>for $t \in [T-1]$ do</td>
<td>$\beta_t \leftarrow \beta^{t-1}_t$ (previous $\epsilon$-solution)</td>
</tr>
<tr>
<td>for $k \in [K]$ do</td>
<td>$\beta_k \leftarrow \beta$</td>
</tr>
<tr>
<td>if $k \mod f = 1$ then</td>
<td>Get $A^\lambda_t(C) = { j \in [p] : \mu_C(x_j) \geq 1 }$ as in (7)</td>
</tr>
<tr>
<td>if $G_{\lambda_t}(\beta, \theta) \leq \epsilon$ then</td>
<td>$\beta^{t}_k \leftarrow \beta$</td>
</tr>
<tr>
<td>break</td>
<td>$\beta^t_j \leftarrow \text{ST}(\frac{\lambda}{|x_j|^2}, \beta^t_j - \frac{x^T_j (X \beta^t - y)}{|x_j|^2})$</td>
</tr>
<tr>
<td>$# \text{ST}(u, x) = \text{sign}(x)(</td>
<td>x</td>
</tr>
</tbody>
</table>

We implemented the screening rules of Table 1 based on the coordinate descent in Scikit-learn (Pedregosa et al., 2011). This code is written in Python and Cython to generate low level C code, offering high performance. A low level language is necessary for this algorithm to scale. Two implementations were written to work efficiently with both dense data stored as Fortran ordered arrays and sparse data stored in the compressed sparse column (CSC) format. Our pseudo-code is presented in Algorithm 1. In practice, we perform the dynamic screening tests every $f = 10$ passes through the entire (active) variables. Iterations are stopped when the duality gap is smaller than the target accuracy.

The naive computation of $\theta_k$ in (11) involves the computation of $\|X^\top \rho_k\|_p$ ($\rho_k$ being the current residual), which costs $O(np)$ operations. This can be avoided as one knows when using a safe rule that the index achieving the maximum for this norm is in $A^\lambda_t(C)$. Indeed, by construction $\arg \max_{j \in A^\lambda_t(C)} |x_j^\top \theta_k| = \arg \max_{j \in [p]} |x_j^\top \theta_k| = \arg \max_{j \in [p]} |x_j^\top \rho_k|$. In practice the evaluation of the dual gap is therefore not a $O(np)$ but $O(nq)$ where $q$ is the size of $A^\lambda_t(C)$. In other words, using screening also speeds up the evaluation of the stopping criterion.

We did not compare our method against the strong rules of Tibshirani et al. (2012) because they are not safe and therefore need complex post-processing with parameters to tune. Also we did not compare against the sequential rule of Wang et al. (2013) (e.g., EDDP) because it requires the exact dual optimal solution of the previous Lasso problem, which is not available in practice and can prevent the solver from actually converging: this is a phenomenon we always observed on our experiments.

4.2. Number of screened variables

Figure 3 presents the proportion of variables screened by several safe rules on the standard Leukemia dataset. The screening proportion is presented as a function of the number of iterations $K$. As the SAFE screening rule of El Ghaoui et al. (2012) is sequential but not dynamic, for a given $\lambda$ the proportion of screened variables does not depend on $K$. The rules of Bonnefoy et al. (2014a) are more efficient on this dataset but they do not benefit much from the dynamic framework. Our proposed GAP SAFE tests screen much more variables, especially when the tuning parameter $\lambda$ gets small, which is particularly relevant in practice. Moreover, even for very small $\lambda$’s (notice the logarithmic scale) where no variable is screened at the beginning of the optimization procedure, the GAP SAFE rules manage to screen more variables, especially when $K$ increases. Finally, the figure demonstrates that the GAP SAFE dome test only brings marginal improvement over the sphere.
Figure 4. Time to reach convergence using various screening rules on the Leukemia dataset (dense data: \(n = 72, p = 7129\)).

4.3. Gains in the computation of Lasso paths

The main interest of variable screening is to reduce computation costs. Indeed, the time to compute the screening itself should not be larger than the gains it provides. Hence, we compared the time needed to compute Lasso paths to prescribed accuracy for different safe rules. Figures 4, 5 and 6 illustrate results on three datasets. Figure 4 presents results on the dense, small scale, Leukemia dataset. Figure 5 presents results on a medium scale sparse dataset obtained with bag of words features extracted from the 20newsgroup dataset (comp.graphics vs. talk.religion.misc with TF-IDF removing English stop words and words occurring only once or more than 95% of the time). Text feature extraction was done using Scikit-Learn. Figure 6 focuses on the large scale sparse RCV1 (Reuters Corpus Volume 1) dataset, cf. (Schmidt et al., 2013).

In all cases, Lasso paths are computed as required to estimate optimal regularization parameters in practice (when using cross-validation one path is computed for each fold). For each Lasso path, solutions are obtained for \(T = 100\) values of \(\lambda\)’s, as detailed in Section 3.3. Remark that the grid used is the default one in both Scikit-Learn and the glmnet R package. With our proposed GAP SAFE screening we obtain on all datasets substantial gains in computational time. We can already get an up to 3x speedup when we require a duality gap smaller than \(10^{-4}\). The interest of the screening is even clearer for higher accuracies: GAP SAFE sphere is 11x faster than its competitors on the Leukemia dataset, at accuracy \(10^{-8}\). One can observe that with the parameter grid used here, the larger is \(p\) compared to \(n\), the higher is the gain in computation time.

In our experiments, the other safe screening rules did not show much speed-up. As one can see on Figure 3, those screening rules keep all the active variables for a wide range of \(\lambda\)’s. The algorithm is thus faster for large \(\lambda\)'s but slower afterwards, since we still compute the screening tests. Even if one can avoid some of these useless computations thanks to formulas like (14) or (10), the corresponding speed-up would not be significant.

5. Conclusion

We have presented new results on safe rules for accelerating algorithms solving the Lasso problem (see Appendix for extension to the Elastic Net). First, we have introduced the framework of converging safe rules, a key concept independent of the implementation chosen. Our second contribution was to leverage duality gap computations to create two safer rules satisfying the aforementioned convergence properties. Finally, we demonstrated the important practical benefits of those new rules by applying them to standard dense and sparse datasets using a coordinate descent solver. Future works will extend our framework to generalized linear model and group-Lasso.

Acknowledgment

We acknowledge the support from Chair Machine Learning for Big Data at Télécom ParisTech and from the Orange/Télécom ParisTech think tank phi-TAB. This work benefited from the support of the “FMJH Program Gaspard Monge in optimization and operation research”, and from the support to this program from EDF.
Mind the duality gap: safer rules for the Lasso

References


Mind the duality gap: safer rules for the Lasso


