De l'usage du saut de dualité pour la pré-sélection dynamique des variables pour le Lasso

Olivier Fercoq [1] & Alexandre Gramfort [2] & Joseph Salmon [3]

[1] *Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI,*
olivier.fercoq@telecom-paristech.fr
[2] *Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI,*
alexandre.gramfort@telecom-paristech.fr
[3] *Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI,*
joseph.salmon@telecom-paristech.fr

Résumé. À l'aide de certificats d'optimalité vérifiés par les solutions du Lasso il est possible d'écarter, avant optimisation, certaines des variables non pertinentes. Ce faisant on peut accélérer drastiquement les algorithmes résolvant le problème du lasso. Nous proposons de nouvelles règles de pré-sélection qui reposent sur le saut de dualité. Elles s'appuient sur la création de régions dites de sécurité, dont le diamètre tend vers zéro, sous l'hypothèse que l'on dispose d'un algorithme convergeant pour résoudre le Lasso. Cette propriété permet à la fois de dépister plus de variables non pertinentes, et de considérer de plus grandes plages pour le paramètre de régularisation. Même si notre cadre englobe tout algorithme résolvant le Lasso, nous démontrons la pertinence de notre approche pour la méthode de descente par coordonnées, particulièrement bien adaptée pour des problèmes de grande dimension. Des gains de temps de calcul importants sont ainsi obtenus par rapport aux précédentes règles de pré-sélection.

Mots-clés. Lasso, règle de sélection, saut de dualité, descente par coordonnées

Abstract. Screening rules allow to early discard irrelevant variables from the optimization in Lasso problems, making solvers faster. 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 helps screening out more variables, often for a wider range of regularization parameter values. 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.

Keywords. Lasso, safe rule, duality gap, coordinate descent

1 Introduction and motivation

Since the mid 1990's, high dimensional statistics has attracted considerable attention, particularly 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 ℓ_1 regularization (called Lasso in statistics [16], or Basis Pursuit in signal processing [4]) has been one of the most popular tools. It enjoys theoretical guarantees [1], 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 [13], bio-statistics [11] and medical imaging [10] to name a few.

Many algorithms exist to approximate Lasso solutions, but it is still an issue to accelerate solvers in high dimensions. Indeed, although some other variable selection and prediction methods exist [8], the best performing methods usually rely on the Lasso. For non-convex approaches such as SCAD [7] or MCP [19], solving the Lasso is often a required preliminary step.

Among possible algorithmic candidates for solving the Lasso, one can mention homotopy methods [14] or LARS [5] that provide the solutions for the full Lasso path, *i.e.,* for all possible choices of tuning parameter λ. More recently, particularly when $p > n$, coordinate descent approaches [9] have proved to be among the best methods to tackle large scale problems.

Following the seminal work by [6], 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 (*cf.* [18] for a nice survey). Zeroing those coefficients allows to focus more precisely on the non-zero ones (likely to represent signal) and helps reducing the computational burden. 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* [17].

The original safe rules operate as follows: for a fixed tuning parameter λ, and before launching any solver, test whether a coordinate can be zeroed or not (equivalently if the corresponding variable can be disregarded or not). Note that the test is performed according to a safe region, *i.e.,* a region containing a dual optimal solution of the Lasso problem. Here, the screening is performed only once prior any optimization iteration.

We aim at improving the screening by interlacing it throughout the optimization algorithm itself: though screening might be useless at the beginning, it might become more and more efficient as the algorithm proceeds towards the optimal solution. We call these strategies *dynamic safe rules* following the terminology introduced in [3, 2].

Based on convex optimization arguments, we leverage duality gap computations to propose a simple dynamic safe rule. We call it GAP SAFE rule.

2 Lasso: model and notation

Our observation vector is $y \in \mathbb{R}^n$ and the design matrix $X = [x_1, \cdots, x_p] \in \mathbb{R}^{n \times p}$ has p explanatory variables 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.

For such a task, we consider the Lasso whose definition is as follows. For a tuning parameter $\lambda > 0$, which controls the trade-off between data-fit versus the sparsity of the solutions, a Lasso estimator $\hat{\beta}^{(\lambda)}$ is any solution of the primal optimization problem

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

Denoting $\Delta_X = \left\{\theta \in \mathbb{R}^n \ : \ \left|x_j^\top \theta\right| \leqslant 1, \forall j \in [p]\right\}$ the dual feasible set, a dual formulation of the Lasso reads (see for instance [12] or [18]):

$$\hat{\theta}^{(\lambda)} = \underset{\theta \in \Delta_X \subset \mathbb{R}^n}{\arg\max} \underbrace{\frac{1}{2} \|y\|_2^2 - \frac{\lambda^2}{2} \left\|\theta - \frac{y}{\lambda}\right\|_2^2}_{=D_\lambda(\theta)}. \tag{2}$$

In particular, note that the dual solution $\hat{\theta}^{(\lambda)}$ is unique, contrarily to the primal $\hat{\beta}^{(\lambda)}$.

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 $\hat{\theta}^{(\lambda)}$ with center c and radius r, *i.e.,* $\mathcal{C} = B(c, r)$. The corresponding safe test is defined as follows:

$$\text{If } |x_j^\top c| + r\|x_j\| < 1, \text{ then } \hat{\beta}_j^{(\lambda)} = 0. \tag{3}$$

Note that for a fixed center, the smaller the radius, the better the safe screening strategy. So the main goal of safe rules is to find sphere with a small radius to eliminate as many variables x_j as possible.

2.2 Dynamic safe rules

For approximating a solution $\hat{\beta}^{(\lambda)}$ of the Lasso primal problem P_λ, iterative algorithms are commonly used. We denote $\beta_k \in \mathbb{R}^p$ the current estimate after k iterations of any iterative algorithm. Dynamic safe rules aim at discovering safe regions that become narrower as k increases. One first needs a dual feasible points: $\theta_k \in \Delta_X$. Following [6] (see also [3]), this can be achieved by a simple transformation of the current residuals $\rho_k = y - X\beta_k$, defining θ_k as

$$\theta_k = \alpha_k \rho_k, \text{ where } \alpha_k = \min\Big[\max\left(\frac{y^\top \rho_k}{\lambda \|\rho_k\|^2}, \frac{-1}{\|X^\top \rho_k\|_\infty}\right), \frac{1}{\|X^\top \rho_k\|_\infty}\Big]. \tag{4}$$

Such dual feasible θ_k is proportional to ρ_k, and is the closest point (for the norm $\|\cdot\|$) to y/λ in Δ_X with such a property. A reason for choosing this dual point is that the dual optimal solution $\hat{\theta}^{(\lambda)}$ is the projection of y/λ on the dual feasible set Δ_X, and the optimal $\hat{\theta}^{(\lambda)}$ is proportional to $y - X\hat{\beta}^{(\lambda)}$.

Our dynamic safe rule consists in choosing as center $c = \theta_k$ in (3). One can prove that a radius equals to $r = \frac{1}{\lambda}\sqrt{2\left(P_\lambda(\beta_k) - D_\lambda\left(\theta_k\right)\right)}$ leads to a safe rule. Note that $P_\lambda(\beta_k) - D_\lambda\left(\theta_k\right)$ is simply the duality gap obtained for primal β_k and dual θ_k.

Remark 1. One can refine the safe sphere rule to a safe dome rule. Unfortunately details are too cumbersome to be given here, *cf.* [18] for more details.

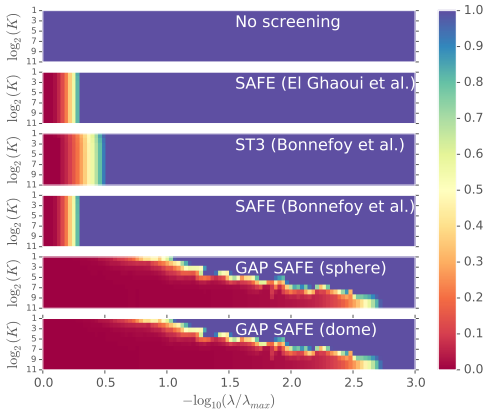
Remark 2. Note that if $\lim_{k\to+\infty} \beta_k = \hat{\beta}^{(\lambda)}$ (convergence of the primal) then we can show that $\lim_{k\to+\infty} \theta_k = \hat{\theta}^{(\lambda)}$ (convergence of the dual), and that the convergence of the primal is unaltered by any safe rule. Screening out unnecessary coefficients can only decrease the distance to a primal solution.

3 Experiments

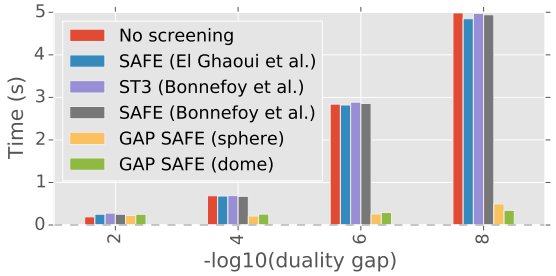
We implemented standard screening rules as well as ours based on the coordinate descent in Scikit-learn [15]. The code is written in Python and Cython to generate low level C code, offering high performance. A low level language is necessary for this algorithm to scale. In practice, we perform the dynamic screening tests every 10 passes through the entire (active) variables. Iterations are stopped when the duality gap is smaller than the target accuracy.

Figure 1,(a) 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 in the coordinate descent implementation of Scikit-Learn [15]. As the SAFE screening rule of [6] is not dynamic, for a given λ the proportion of screened variables does not depend on K. The rules of [3] 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 λ gets small, which is particularly relevant in practice. Moreover, even for very small λ'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.

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 given by the screening. Hence, we compared the time needed to compute the whole Lasso path to a prescribed accuracy for different safe rules. Figures 1(b) present results on the dense, small scale, Leukemia dataset.



(a) Screening proportion as a function of λ and the number of iterations K.



(b) Time to reach convergence using various screening rules.

Figure 1: Leukemia dataset (dense data: $n = 72, p = 7129$).

References

[1] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

[2] A. Bonnefoy, V. Emiya, L. Ralaivola, and R. Gribonval. Dynamic Screening: Accelerating First-Order Algorithms for the Lasso and Group-Lasso. *ArXiv e-prints*, 2014.

[3] A. Bonnefoy, V. Emiya, L. Ralaivola, and R. Gribonval. A dynamic screening principle for the lasso. In *EUSIPCO*, 2014.

[4] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61 (electronic), 1998.

[5] B. Efron, T. Hastie, I. M. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004. With discussion, and a rejoinder by the authors.

[6] L. El Ghaoui, V. Viallon, and T. Rabbani. Safe feature elimination in sparse supervised learning. *J. Pacific Optim.*, 8(4):667–698, 2012.

[7] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001.

[8] J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Statist. Soc. Ser. B*, 70(5):849–911, 2008.

[9] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, 2007.

[10] A. Gramfort, M. Kowalski, and M. Hämäläinen. Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods. *Physics in Medicine and Biology*, 57(7):1937–1961, 2012.

[11] A.-C. Haury, F. Mordelet, P. Vera-Licona, and J.-P. Vert. TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC systems biology*, 6(1):145, 2012.

[12] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. An interior-point method for large-scale l_1-regularized least squares. *IEEE J. Sel. Topics Signal Process.*, 1(4):606–617, 2007.

[13] J. Mairal. *Sparse coding for machine learning, image processing and computer vision*. PhD thesis, École normale supérieure de Cachan, 2010.

[14] M. R. Osborne, B. Presnell, and B. A. Turlach. A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.*, 20(3):389–403, 2000.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

[16] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[17] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *J. Roy. Statist. Soc. Ser. B*, 74(2):245–266, 2012.

[18] Z. J. Xiang, Y. Wang, and P. J. Ramadge. Screening tests for lasso problems. *arXiv preprint arXiv:1405.4897*, 2014.

[19] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942, 2010.