Look-Ahead Screening Rules for the Lasso

Johan Larsson[*1]

[1] *The Department of Statistics, Lund University*

May 5, 2021

Abstract: The lasso is a popular method for inducing shrinkage and sparsity in the solution vector (coefficients) of regression problems, particularly when the number of predictors far outnumber the number of observations. Solving the lasso for high-dimensional data can, however, be computationally demanding. Fortunately, this computational load can be alleviated via the use of *screening rules*, which screen and discard predictors prior to fitting the model, leading a reduced problem to be solved. Screening rules are particularly effective when fitting a full regularization path: a sequence of models with decreasing penalization. Screening rules can be safe or heuristic. Safe rules certify that discarded predictors are not in the solution; heuristic ones do not. Existing screening rules typically work sequentially or dynamically. Sequential rules screen predictors for the next model along the regularization path, whereas dynamical rules screen during optimization of the current model. There has, however, previously been no attempts to design screening rules that screen further along the path.

In this paper, we present a new screening strategy: *look-ahead* screening rules. Our method uses safe screening rules to find a range of penalty values for which a given predictor cannot enter the model, thereby screening predictors along the remainder of the path. Our screening rules lead to reductions in the time required for screening and also applies to heuristic rules, for which the time required to conduct checks of the optimality conditions to guard against violations of the rules is reduced. In experiments we show that these look-ahead screening rules improve the performance of existing screening strategies and that the additional cost of screening ahead on the path is marginal.

[*] johan.larsson@stat.lu.se

Keywords: lasso, sparse regression, screening rules, safe screening rules

AMS subject classification: 62J07

1 Introduction

2 Preliminaries

Starting with the preliminaries of our problem, we take $X \in \mathbb{R}^{n\times p}$ be the design matrix of n observations and p predictors and $y \in \mathbb{R}^n$ the response vector.

The lasso is represented by the following convex optimization problem:

$$\operatorname*{minimize}_{\beta\in\mathbb{R}^p} P(\beta;\lambda) \tag{1}$$

where

$$P(\beta;\lambda) = \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{2}$$

is the *primal* objective. We let $\hat{\beta}_\lambda$ be the solution to (1) for a given λ. In addition, we also let $\tilde{\beta}_\lambda$ be a numerical approximation to $\hat{\beta}_\lambda$ returned by an algorithm tasked with solving (1).

The Karush–Kuhn–Tucker (KKT) stationarity condition for (1) holds that

$$0 \in X^T(X\beta - y) + \lambda\partial, \tag{3}$$

where ∂ is the subdifferential of the ℓ_1-norm, with its jth element corresponding to

$$\partial_j \in \begin{cases} \{\operatorname{sign}(\beta_j)\} & \text{if } \beta_j \neq 0 \\ [-1, 1] & \text{otherwise.} \end{cases} \tag{4}$$

Moreover, we define the dual problem of (1) as

$$\operatorname*{maximize}_{\theta\in\mathbb{R}^n} D(\theta;\lambda) \tag{5}$$

where

$$D(\theta) = \frac{1}{2}y^T y - \frac{\lambda^2}{2}\left\|\theta - \frac{y}{\lambda}\right\|_2^2 \tag{6}$$

is the *dual* objective. The relationship between the primal and dual problems is given by

$$y = X\hat{\beta}_\lambda + \lambda\hat{\theta}_\lambda.$$

Next, we let G be the so-called *duality gap*, which we define as

$$\begin{aligned} G(\beta, \theta; \lambda) &= P(\beta; \lambda) - D(\theta; \lambda) \\ &= \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 - \frac{1}{2}y^T y + \frac{\lambda^2}{2}\left\|\theta - \frac{y}{\theta}\right\|_2^2 \\ &= \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 - \lambda\theta^T y + \frac{\lambda^2}{2}\theta^T\theta. \end{aligned} \tag{7}$$

In the case of the lasso, strong duality holds, which means that $G(\tilde{\beta}_\lambda, \tilde{\theta}_\lambda; \lambda) = 0$ for any choice of λ.

Suppose, now, that we have solved the lasso for λ; then for any given $\lambda^* \geq \lambda$, the Gap Safe rule [1] for the lasso discards the jth predictor if

$$|X^T \tilde{\theta}_\lambda|_j + \|x_j\|_2 \sqrt{\frac{1}{\lambda_*^2} G(\tilde{\beta}_\lambda, \tilde{\theta}_\lambda; \lambda^*)} < 1 \tag{8}$$

where

$$\tilde{\theta}_\lambda = \frac{y - X\tilde{\beta}_\lambda}{\max\big(\max_j |x_j^T(y - X\tilde{\beta}_\lambda)|, \lambda\big)}$$

is a dual-feasible point [1], which we can always obtain by dual scaling whenever a candidate is infeasible. Using these facts, we now arrive at Theorem 1, which is the main result of this paper.

Theorem 1. *Let* $(\tilde{\beta}_\lambda, \tilde{\theta}_\lambda)$ *be a feasible primal–dual point for the solution to* (1) *and* (5). *Then* $\big(\hat{\beta}_{\lambda^*}\big)_j = 0$ *for any*

$$\lambda_* > \frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

where

$$\begin{aligned} a &= \big(1 - |x_j^T \tilde{\theta}_\lambda|\big)^2 - \frac{1}{2}\tilde{\theta}_\lambda^T \tilde{\theta}_\lambda \|x_j\|_2^2 \\ b &= \big(\tilde{\theta}_\lambda^T y - \|\tilde{\beta}_\lambda\|_1\big)\|x_j\|_2^2 \\ c &= -\frac{1}{2}\|y - X\tilde{\beta}_\lambda\|_2^2 \|x_j\|_2^2. \end{aligned}$$

Proof. □

Note that there must be some λ_*^2, for which it holds that

$$|X^T \tilde{\theta}_\lambda|_j + \|x_j\|_2 \sqrt{\frac{1}{\lambda_*^2} G(\tilde{\beta}_\lambda, \tilde{\theta}_\lambda; \lambda^*)} = 1$$

Acknowledgements: Your acknowledgements.

References

[1] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. Gap safe screening rules for sparsity enforcing penalties. *Journal of Machine Learning Research*, 18(128):1–33, 2017.