Look-Ahead Screening Rules for the Lasso

Larsson, Johan

Published in:
22nd European young statisticians meeting - proceedings

2021

Document Version:
Publisher's PDF, also known as Version of record

Look-ahead screening rules for the Lasso

Johan Larsson[1*]

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

Abstract: The lasso is a popular method to induce shrinkage and sparsity in the solution vector (coefficients) of regression problems, particularly when there are many predictors relative to the number of observations. Solving the lasso in this high-dimensional setting can, however, be computationally demanding. Fortunately, this demand can be alleviated via the use of *screening rules* that discard predictors prior to fitting the model, leading to a reduced problem to be solved. In this paper, we present a new screening strategy: *look-ahead screening*. 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. In experiments we show that these look-ahead screening rules outperform the active warm-start version of the Gap Safe rules.

Keywords: lasso, sparse regression, screening rules, safe screening rules
AMS subject classification: 62J07

1 Introduction

The lasso [6] is a staple among regression models for high-dimensional data. It induces shrinkage and sparsity in the solution vector (regression coefficients) through penalization by the ℓ_1-norm. The optimal level of penalization is, however, usually unknown, which means we typically need to estimate it through model tuning across a grid of candidate values: the regularization path. This leads to a heavy computational load.

Thankfully, the advent of so-called *screening rules* have lead to remarkable advances in tackling this problem. Screening rules discard a subset of the predictors *before* fitting the model, leading to, often considerable, reductions in problem size. There are two types of screening rules: heuristic and safe rules. The latter kind provides a certificate that discarded predictors cannot be active at the optimum—that is, have a non-zero corresponding coefficients—whereas heuristic rules do not. In this paper, we will focus entirely on safe rules.

A prominent type of safe rules are the Gap Safe rules [5, 1], which use the duality gap in a problem to provide effective screening rules. There currently exists sequential versions of the Gap Safe rules, that discard predictors for the next step

*Corresponding author: johan.larsson@stat.lu.se

on the regularization path, as well as dynamic rules, which discard predictors during optimization at the current penalization value.

The objective of this paper is to introduce a new screening strategy based on Gap Safe screening: *look-ahead screening*, which screens predictors for a range of penalization parameters. We show that this method can be used to screen predictors for the entire stretch of the regularization path, leading to substantial improvements in the time to fit the entire lasso path.

2 Look-Ahead Screening

Let $X \in \mathbb{R}^{n \times p}$ be the design matrix with n observations and p predictors and $y \in \mathbb{R}^n$ the response vector. The lasso is represented by the following convex optimization problem:

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \left\{ P(\beta; \lambda) = \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \tag{1}$$

where $P(\beta; \lambda)$ is the *primal* objective. We let $\hat{\beta}_\lambda$ be the solution to (1) for a given λ. Moreover, the dual problem of (1) is

$$\underset{\theta \in \mathbb{R}^n}{\text{maximize}} \left\{ D(\theta; \lambda) = \frac{1}{2} y^T y - \frac{\lambda^2}{2} \left\| \theta - \frac{y}{\lambda} \right\|_2^2 \right\} \tag{2}$$

where $D(\theta; \lambda)$ 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*, defined as

$$G(\beta, \theta; \lambda) = P(\beta; \lambda) - D(\theta; \lambda) = \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 - \lambda \theta^T y + \frac{\lambda^2}{2} \theta^T \theta. \tag{3}$$

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

Suppose, now, that we have solved the lasso for λ; then for any given $\lambda^* \leq \lambda$, the Gap Safe rule [5] discards the jth predictor if

$$|X^T \theta_\lambda|_j + \|x_j\|_2 \sqrt{\frac{1}{\lambda_*^2} G(\beta_\lambda, \theta_\lambda; \lambda^*)} < 1 \tag{4}$$

where

$$\theta_\lambda = \frac{y - X\beta_\lambda}{\max\big(|X^T(y - X\beta_\lambda)|, \lambda\big)}$$

is a dual-feasible point [5] obtained through dual scaling.

Observe that (4) is a quadratic inequality with respect to λ_*, which means that it is trivial to discover the boundary points via the quadratic formula:

$$\lambda_* = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad \text{where} \quad \begin{aligned} a &= \big(1 - |x_j^T \theta_\lambda|\big)^2 - \frac{1}{2} \theta_\lambda^T \theta_\lambda \|x_j\|_2^2, \\ b &= \big(\theta_\lambda^T y - \|\beta_\lambda\|_1\big) \|x_j\|_2^2, \\ c &= -\frac{1}{2} \|y - X\beta_\lambda\|_2^2 \|x_j\|_2^2. \end{aligned}$$

By restricting ourselves to an index j corresponding to a predictor that is inactive at λ and recalling that we have $\lambda_* \leq \lambda$ by construction, we can inspect the signs of a, b, and c and find a range λ values for which predictor j must be inactive. Using this idea for the lasso path—a grid of λ values starting from the null (intercept-only) model, which corresponds to $\lambda_{\max} = \max_i |x_i^T y|$, and finishing at fraction of this (see section 3 for specifics)—we can screen predictor j for all upcoming λs, possibly discarding it for multiple steps on the path rather than just the next step. We call this idea *look-ahead screening.*

To illustrate the effectiveness of this screening method, we consider an instance of employing look-ahead screening for fitting a full lasso path to the *leukemia* data set [3]. At the first step of the path, the screening method discards 99.6% of the predictors for the steps up to and including step 5. The respective figures for steps 10 and 15 are 99.3% and 57%. At step 20, however, the rule does not discard a single predictor. In Figure 1, we have visualized the screening performance of look-ahead screening for a random sample of 25 predictors from this data set.

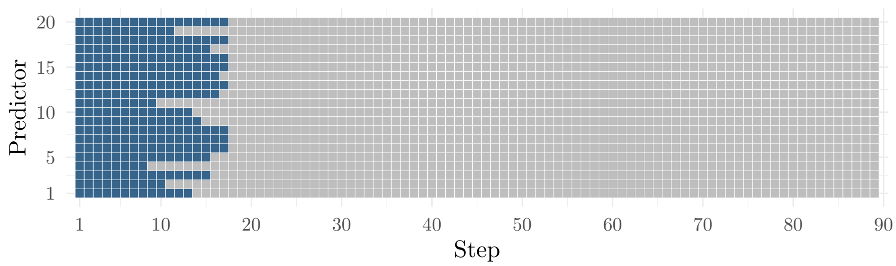

Figure 1: This figure shows the predictors screened at the first step of the lasso path via look-ahead screening for a random sample of 20 predictors from the *leukemia* data set. A blue square indicates that the corresponding predictor can be discarded at the respective step.

As is typical for all screening methods, the effectiveness of look-ahead screening is greatest at the start of the path and diminishes as the strength of penalization decreases later on in the path. Note, however, that all of the quantities involved in the rule are available as a by-product of solving the problem at the previous step, which means that the costs of look-ahead screening are diminutive.

3 Simulations

In this section, we study the effectiveness of the look-ahead screening rules by comparing them against the active warm start version of the Gap Safe rules [1, 5]. We follow the recommendations in [5] and run the screening procedure every tenth pass of the solver. Throughout the experiments, we center the response vector by its mean, as well as center and scale the predictors by their means and uncorrected sample standard deviations respectively.

To construct the regularization path, we employ the standard settings from `glmnet`, using a log-spaced path of 100 λ values from $\lambda_{\max}$ to $\varepsilon\lambda_{\max}$, where $\varepsilon = 10^{-2}$

if $p > n$ and 10^{-4} otherwise. We also use the default path stopping criteria from glmnet, that is, stop the path whenever the deviance ratio, $1-\text{dev}/\text{dev}_{\text{null}}$, is greater than or equal to 0.999, the fractional increase in deviance explained is lower than 10^{-5}, or, if $p \geq n$, when the number of active predictors exceeds or is equal to n.

To fit the lasso, we use cyclical coordinate descent [2]. We consider the solver to have converged whenever the duality gap as a fraction of the primal value for the null model is less than or equal to 10^{-6} and the amount of *infeasibility*, which we define as $\max_j \left(|x_j^T(y - X\beta_\lambda)| - \lambda\right)$, as a fraction of $\lambda_{\max}$ is lower than or equal to 10^{-5}.

Source code for the experiments, including a container to facilitate reproducibility, can be found at https://github.com/jolars/LookAheadScreening/. An HPC cluster node with two Intel Xeon E5-2650 v3 processors (Haswell, 20 compute cores per node) and 64 GB of RAM was used to run the experiments.

We run experiments on a design with $n = 100$ and $p = 50\,000$, drawing the rows of X i.i.d. from $\mathcal{N}(0, \Sigma)$ and y from $\mathcal{N}(X\beta, \sigma^2 I)$ with $\sigma^2 = \beta^T\Sigma\beta/\text{SNR}$, where SNR is the signal-to-noise ratio. We set 5 coefficients, equally spaced throughout the coefficient vector, to 1 and the rest to zero. Taking inspiration from **(author?)** [4], we consider SNR values of 0.1, 1, and 6.

Judging by the results (Figure 2), the addition of look-ahead screening results in sizable reductions in the solving time of the lasso path, particularly in the high signal-to-noise context.

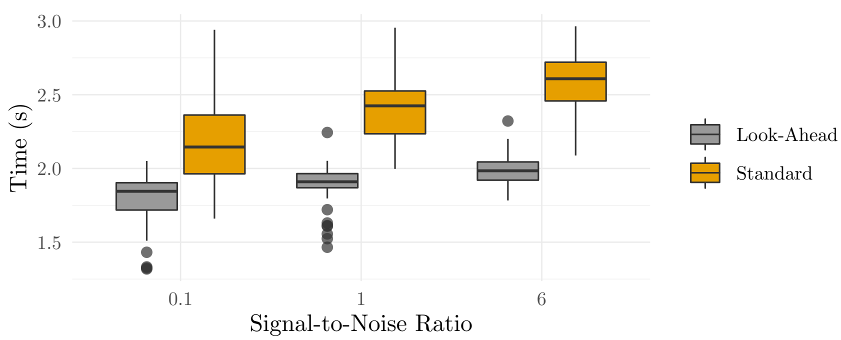

Figure 2: Standard box plots of timings to fit a full lasso path to a simulated data set with $n = 100$, $p = 50\,000$, and five true signals.

4 Discussion

In this paper, we have presented *look-ahead screening*, which is a novel method to screen predictors for a range of penalization values along the lasso regularization path using Gap Safe screening. Our results show that this type of screening can yield considerable improvements in performance for the standard lasso. For other loss functions, (4) may no longer reduce to a quadratic inequality and will hence require more computation. Nevertheless, we believe that applying these rules in these cases is feasible and likely to result in comparable results.

Moreover, the idea is general and can therefore be extended to any type of safe screening rule and also used in tandem with heuristic screening rules in order to avoid expensive KKT computations. Finally, although we only cover one type of cyclical coordinate descent in our experiments, note that our screening method is agnostic to the solver used and that we expect the results hold for any solver that benefits from predictor screening.

Acknowledgements: I would like to thank my supervisor, Jonas Wallin, for valuable feedback on this work. The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at LUNARC partially funded by the Swedish Research Council through grant agreement no. 2017-05973.

Bibliography

[1] O. Fercoq, A. Gramfort, and J. Salmon. Mind the duality gap: Safer rules for the lasso. In F. Bach and D. Blei, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 333–342, Lille, France, July 2015. PMLR.

[2] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. doi: 10.18637/jss.v033.i01.

[3] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, Oct. 1999. ISSN 0036-8075. doi: 10.1126/science.286.5439.531.

[4] T. Hastie, R. Tibshirani, and R. Tibshirani. Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons. *Statistical Science*, 35(4):579–592, Nov. 2020. ISSN 0883-4237. doi: 10. 1214/19-STS733.

[5] 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.

[6] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. ISSN 0035-9246.