Screening Tests for the LASSO

Hanxiao Liu
hanxiaol@cs.cmu.edu

November 3, 2015

Outline

- LASSO
  - Primal & Dual form
  - Primal-dual correspondence
- Safe Test for LASSO
  - Static case (example: sphere test)
  - Dynamic case
- Better safe test based on duality gap
  - Geometric illustration
  - Convergence
- Empirical Results

$X \in \mathbb{R}^{n \times p}$ where $p \gg n$

$$\min_{\beta \in \mathbb{R}^p} \ \frac{1}{2}\|X\beta - y\|_2^2 + \lambda\|\beta\|_1 \tag{1}$$

- Commonly used for high-dimensional feature selection

Today's topic—screening tests

- Rules to early discard irrelevant features prior to starting a LASSO solver without affecting the final opt solution.
- The "chicken-and-egg problem"?

LASSO Dual

$$\text{Primal:} \quad \min_{\beta\in\mathbb{R}^p, z\in\mathbb{R}^n} \frac{1}{2}\|y-z\|_2^2 + \lambda\|\beta\|_1 \quad \text{s.t.} \quad z = X\beta \tag{2}$$

$$\text{Dual:} \quad \underbrace{\max_{u\in\mathbb{R}^n} \min_{\beta,z} \underbrace{\frac{1}{2}\|y-z\|_2^2 + \lambda\|\beta\|_1 + u^\top(z - X\beta)}_{\text{Lagrangian } \mathcal{L}(\beta,z,u)}}_{\text{Dual Objective } g(u)} \tag{3}$$

$$\implies \max_u \left[\min_z \left(\frac{1}{2}\|y-z\|_2^2 + u^\top z\right) - \lambda \max_\beta \left(\frac{u^\top X}{\lambda}\beta - \|\beta\|_1\right)\right] \tag{4}$$

$$\implies \max_u \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|u-y\|_2^2 - \lambda \mathbb{I}_{\left\{\left\|\frac{X^\top u}{\lambda}\right\|_\infty \le 1\right\}} \tag{5}$$

$$\xrightarrow{\theta=\frac{u}{\lambda}} \max_{\theta\in\mathbb{R}^n} \frac{1}{2}\|y\|_2^2 - \frac{\lambda^2}{2}\left\|\theta - \frac{y}{\lambda}\right\|_2^2 \quad \text{s.t.} \quad \left\|X^\top\theta\right\|_\infty \le 1 \tag{6}$$

Karush-Khun-Tucker Condition

Primal-dual correspondence

$$\hat{\beta}, \hat{z} \in \operatorname{argmin}_{\beta,z} \ \mathcal{L}\left(\beta, z, \hat{\theta}\right) \tag{7}$$

$$\Downarrow$$

$$\begin{aligned} 0_n &\in \partial_\beta \mathcal{L}\left(\hat{\beta}, \hat{z}, \hat{\theta}\right) && (8) \\ &= \partial_\beta \left[\frac{1}{2}\|y - \hat{z}\|_2^2 + \lambda\|\hat{\beta}\|_1 + \lambda\hat{\theta}^\top\left(\hat{z} - X\hat{\beta}\right)\right] && (9) \\ &= \lambda\partial_\beta\|\hat{\beta}\|_1 - \lambda X^\top\hat{\theta} && (10) \end{aligned}$$

$$\Downarrow$$

$$x_j^\top\hat{\theta} \in \partial_{\beta_j}|\hat{\beta}_j| = \begin{cases} \operatorname{sign}(\hat{\beta}_j) & \hat{\beta}_j \neq 0 \\ [-1, 1] & \hat{\beta}_j = 0 \end{cases} \quad \forall j \in [p] \tag{11}$$

Key observation: $|x_j^\top\hat{\theta}| < 1 \implies \hat{\beta}_j = 0$

Safe Test

$|x_j^\top \hat{\theta}| < 1 \implies \hat{\beta}_j = 0$

- Challenge: dual solution $\hat{\theta}$ is unknown
- Workaround: relaxation

Let $\mathcal{C}$ be a set containing $\hat{\theta}$ and define $\mu_{\mathcal{C}}(x_j) := \sup_{\theta \in \mathcal{C}} |x_j^\top \theta|$. Obviously $|x_j^\top \hat{\theta}| \leq \mu_{\mathcal{C}}(x_j)$.

Safe Test

$$\mu_{\mathcal{C}}(x_j) < 1 \implies \hat{\beta}_j = 0 \tag{12}$$

The test is useful when

1. $\mu_{\mathcal{C}}(x_j)$ can be evaluated efficiently.
2. $\mathcal{C}$ is small—thus leading to small $\mu_{\mathcal{C}}(x_j)$.

Goal: Find $\mathcal{C}$ satisfying both conditions above.

Parametrize $\mathcal{C}$ as a closed ℓ_2-ball $B(c, r) = \{\theta : \|\theta - c\|_2 \le r\}$.

$$\mu_{\mathcal{C}}(x_j) = \mu_{B(c,r)}(x_j) = \sup_{\theta \in B(c,r)} |x_j^\top \theta| = |c^\top x_j| + r\|x_j\|_2 \quad (13)$$

- Q: How to find $B(c, r)$ that contains $\hat{\theta}$ without knowing $\hat{\theta}$?
- A: Any dual-feasible θ' defines a ball over $\hat{\theta}$

$$\Big\|\hat{\theta} - \underbrace{\frac{y}{\lambda}}_{c}\Big\|_2 \le \underbrace{\Big\|\theta' - \frac{y}{\lambda}\Big\|_2}_{r} \quad (14)$$

  - Recall: $\max_{\theta \in \mathbb{R}^n} \ \frac{1}{2}\|y\|_2^2 - \frac{\lambda^2}{2}\|\theta - \frac{y}{\lambda}\|_2^2$ s.t. $\underbrace{\|X^\top \theta\|_\infty \le 1}_{\theta \in \Delta_X}$

- A trivial feasible point: $\theta' = \frac{y}{\lambda_{\max}}$ where $\lambda_{\max} := \|X^\top y\|_\infty$ [1].

$$\implies \quad c = \frac{y}{\lambda}, \quad r = \|y\| \left| \frac{1}{\lambda} - \frac{1}{\lambda_{\max}} \right| \quad (15)$$

[1] In fact, θ' obtained in this manner is the dual solution when $\lambda = \lambda_{\max}$, corresponding to all-zero primal solution.

Dynamic Safe Tests

To iteratively apply safe tests as the algorithm proceeds

- Recall $\forall \theta' \in \Delta_X$ defines an ℓ_2-ball containing $\hat{\theta}$
- Let $\theta_k \in \Delta_X$ be a dual-feasible point at iteration k, $\{\theta_k\}_{k\in\mathbb{N}}$ defines a sequence of balls $\left\{B\left(\frac{y}{\lambda}, \left\|\theta_k - \frac{y}{\lambda}\right\|_2\right)\right\}_{k\in\mathbb{N}}$
- Each ball defines a safe test

How θ_k is defined via β_k?

- $\theta_k := \Pi_{\Delta_X \cap \mathrm{span}(\rho_k)}\left(\frac{y}{\lambda}\right)$ where $\rho_k := y - X\beta_k$
- Intuition
    1. Dual opt: $\hat{\theta} = \Pi_{\Delta_X}\left(\frac{y}{\lambda}\right)$
    2. Primal-dual correspondence: $\hat{\theta} \in \mathrm{span}\left(\hat{\rho}\right)$ where $\hat{\rho} := y - X\hat{\beta}$

Can we better bound $\hat{\theta}$ by also leveraging primal info?

$\forall\,(\beta, \theta) \in \mathbb{R}^p \times \Delta_X$, we claim

$$\hat{R}\,(\beta) \leq \|\hat{\theta} - \frac{y}{\lambda}\| \leq \check{R}\,(\theta) \tag{16}$$

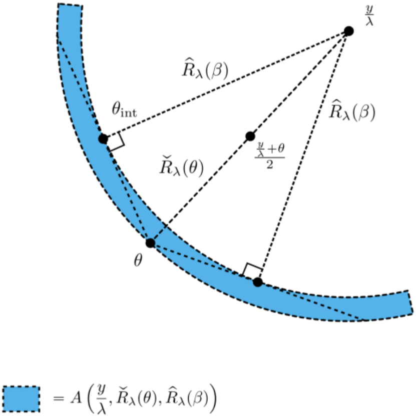
where

- $\check{R}\,(\theta) = \|\theta - \frac{y}{\lambda}\|$ (dual optimality)
- $\hat{R}\,(\beta) = \frac{1}{\lambda}\sqrt{(\|y\|^2 - \|y - X\beta\|^2 - 2\lambda\|\beta\|_2)_+}$ (duality gap)

weak duality

$$\frac{1}{2}\|y\|^2 - \frac{\lambda^2}{2}\left\|\theta - \frac{y}{\lambda}\right\|^2 \leq \frac{1}{2}\|y - X\beta\|^2 + \lambda\|\beta\|_1 \tag{17}$$

Therefore, $\hat{\theta}$ lies in an annulus, i.e. $A\left(\frac{y}{\lambda}, \check{R}\,(\theta)\,, \hat{R}\,(\beta)\right)$



Recall $\hat{R}(\beta) \leq \left\|\hat{\theta} - \frac{y}{\lambda}\right\| \leq \check{R}(\theta)$

[1] [Fercoq et al., 2015]

Fine-grained Analysis

Two geometrical observations

1. $\left[\theta, \hat{\theta}\right] \subseteq A\left(\frac{y}{\lambda}, \check{R}(\theta), \hat{R}(\theta)\right)$

Proof.

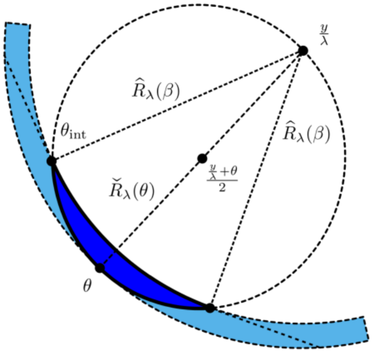
Convexity of polyhedron $\Delta_X \implies$ convexity of $\Delta_X \cap B\left(\frac{y}{\lambda}, \check{R}(\theta)\right)$ $\implies$ convexity of $\Delta_X \cap A\left(\frac{y}{\lambda}, \check{R}(\theta), \hat{R}(\beta)\right)$ containing θ, $\hat{\theta}$ □

2. $vecAngle\left(\theta - \hat{\theta}, \frac{y}{\lambda} - \hat{\theta}\right) \geq 90°$

Proof.

$\forall \theta' \in \left[\theta, \hat{\theta}\right]$ we have $\|\frac{y}{\lambda} - \hat{\theta}\| \leq \|\frac{y}{\lambda} - \theta'\|$ as $\hat{\theta}, \theta' \in \Delta_X$ and $\hat{\theta} = \Pi_{\Delta_X}\left(\frac{y}{\lambda}\right)$. However, suppose $vecAngle\left(\theta - \hat{\theta}, \frac{y}{\lambda} - \hat{\theta}\right) < 90°$, contradiction occurs by setting $\theta' := \Pi_{[\theta, \hat{\theta}]}\left(\frac{y}{\lambda}\right)$. □

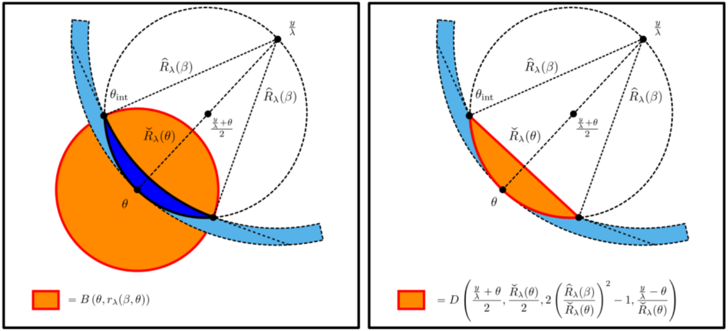
Geometric Illustration



$$= B\left(\frac{y}{\lambda}, \widehat{R}_\lambda(\beta)\right)^c \cap B\left(\frac{\frac{y}{\lambda}+\theta}{2}, \frac{\check{R}_\lambda(\theta)}{2}\right)$$

Recall $vecAngle\left(\theta - \hat{\theta}, \frac{y}{\lambda} - \hat{\theta}\right) \geq 90^\circ$

[1][Fercoq et al., 2015]



Two convex relaxation schemes

$$\text{Sphere} \quad \mathcal{C}_{relaxed} = B\left(\theta, \underbrace{\sqrt{\check{R}\left(\theta\right)^2 - \hat{R}\left(\beta\right)^2}}_{\tilde{r}(\theta,\beta)}\right) \tag{18}$$

$$\text{Dome} \quad \mathcal{C}_{relaxed} = conv\left(darkBlueRegion\right) \tag{19}$$

[1] [Fercoq et al., 2015]

Convergence

Proposition

Let $G(\beta, \theta)$ *be the LASSO duality gap,* $\forall (\beta, \theta) \in \mathbb{R}^p \times \Delta_X$ *we have* $\tilde{r}(\beta, \theta)^2 \leq \frac{2}{\lambda^2} G(\beta, \theta)$

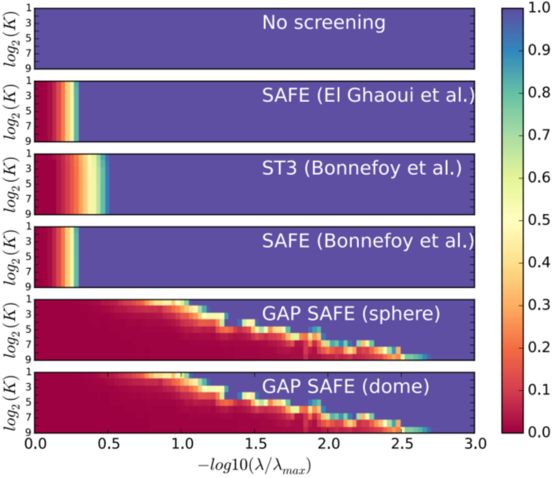
$$\underbrace{\frac{1}{2}\|y\|^2 - \frac{\lambda^2}{2}\left\|\theta - \frac{y}{\lambda}\right\|^2}_{\text{dual obj}} + G(\beta, \theta) = \underbrace{\frac{1}{2}\|y - X\beta\|^2 + \lambda\|\beta\|_1}_{\text{primal obj}} \quad (20)$$

$$\frac{2}{\lambda^2} G(\beta, \theta) = \underbrace{\left\|\theta - \frac{y}{\lambda}\right\|^2}_{=\check{R}(\theta)^2} - \underbrace{\frac{1}{\lambda^2}\left(\|y\|^2 - \|y - X\beta\|^2 - 2\lambda\|\beta\|_1\right)}_{\leq \hat{R}(\beta)^2} \quad (21)$$

$$\geq \tilde{r}(\beta, \theta)^2 \quad (22)$$

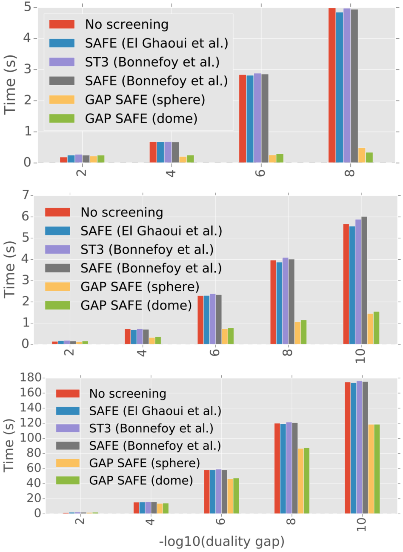
$\implies \lim_{k\to\infty} \tilde{r}(\beta_k, \theta_k) = 0$. Convergence of domes is also implied.

Proportion of active variables v.s. (1) num of iterations (2) λ



[1] [Fercoq et al., 2015]

Experiment Results (b)



Leukemia

$\frac{p}{n} = \frac{7129}{72} \approx 99.0$

20NewsGroup

$\frac{p}{n} = \frac{10094}{961} \approx 10.5$

RCV1

$\frac{p}{n} = \frac{47236}{20242} \approx 2.3$

[1] [Fercoq et al., 2015]

Conclusion

Summary

- LASSO - primal & dual
- Safe rules - discard irrelevant variables prior to optimization
- Refined safe rules based on duality gap

Additional Note

- Safe rules can be applied to other models as well, e.g. support vector machines [Ogawa et al., 2013]

Reference I

El Ghaoui, L., Viallon, V., and Rabbani, T. (2010).
Safe feature elimination in sparse supervised learning technical report no.
Technical report, UCB/EECS-2010-126, EECS Department, University of California, Berkeley.

Fercoq, O., Gramfort, A., and Salmon, J. (2015).
Mind the duality gap: safer rules for the lasso.
arXiv preprint arXiv:1505.03410.

Ogawa, K., Suzuki, Y., and Takeuchi, I. (2013).
Safe screening of non-support vectors in pathwise svm computation.
In *Proceedings of the 30th International Conference on Machine Learning*, pages 1382–1390.

Xiang, Z. J., Wang, Y., and Ramadge, P. J. (2014).
Screening tests for lasso problems.
arXiv preprint arXiv:1405.4897.