Journal club: Speeding up LASSO solvers

Marine Le Morvan

CBIO - Mines Paristech, INSERM U900 - Curie institute, Paris, France

March 13^{th}, 2018

Two recent papers

Mind the duality gap: safer rules for the Lasso

Olivier Fercoq OLIVIER.FERCOQ@TELECOM-PARISTECH.FR
Alexandre Gramfort ALEXANDRE.GRAMFORT@TELECOM-PARISTECH.FR
Joseph Salmon JOSEPH.SALMON@TELECOM-PARISTECH.FR
Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI
46 rue Barrault, 75013, Paris, France

Dual Extrapolation for Faster Lasso Solvers

Mathurin Massias, Alexandre Gramfort, Joseph Salmon

(Submitted on 21 Feb 2018 (v1), last revised 22 Feb 2018 (this version, v2))

- Some background
- The GAP safe rules
- Dual extrapolation for faster Lasso solvers

- Some background
- The GAP safe rules
- Dual extrapolation for faster Lasso solvers

Some reminders about the LASSO

Design matrix: $X = [X_1, \ldots, X_p] \in \mathbb{R}^{n \times p}$
Response vector: $y \in \mathbb{R}^n$

Primal formulation of the LASSO:

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

Dual formulation of the LASSO:

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

where $\Delta_X = \left\{ \theta \in \mathbb{R}^n : \left| X_j^T \theta \right| \leq 1, \forall j \in [\![p]\!] \right\}.$

Some reminders about the LASSO

The Karush-Khun-Tucker conditions state that, at optimality:

$$\lambda\hat{\theta}^{(\lambda)} = y - X\hat{\beta}^{(\lambda)} \tag{3}$$

$$\forall j \in [\![p]\!]\,,\ X_j^T\hat{\theta}^{(\lambda)} \in \begin{cases} \{-1, 1\} & \text{if} \quad \hat{\beta}_j^{(\lambda)} \neq 0\,, \\ [-1, 1] & \text{if} \quad \hat{\beta}_j^{(\lambda)} = 0\,. \end{cases} \tag{4}$$

Safe rules exploit KKT condition (4) according to which:

$$\forall j \in [\![p]\!]\,,\ \left| X_j^T\hat{\theta}^{(\lambda)} \right| < 1 \implies \hat{\beta}_j^{(\lambda)} = 0\ .$$

The main idea behind safe rules

According to the KKT conditions we have:

$$\forall j \in [\![p]\!] \, , \, \left| X_j^T \hat{\theta}^{(\lambda)} \right| < 1 \implies \hat{\beta}_j^{(\lambda)} = 0 \, .$$

The idea behind safe rules is to construct a *safe region* $\mathcal{C}$ which is guaranteed to contain $\hat{\theta}^{(\lambda)}$, so that:

$$\forall j \in [\![p]\!] \, , \, \max_{\theta \in \mathcal{C}} \left| X_j^T \theta \right| < 1 \implies \hat{\beta}_j^{(\lambda)} = 0 \, .$$

$\mathcal{C}$ is often chosen as a sphere or a dome so that the quantity $\max_{\theta \in \mathcal{C}} \left| X_j^T \theta \right|$ can be computed easily.

Sphere test

Let $\mathcal{C}$ be a ball $B(c, r)$ with center c and radius r constructed so that it contains $\hat{\theta}^{(\lambda)}$. With such a choice of safe region, the safe test:

$$\forall j \in [\![p]\!]\,,\ \max_{\theta \in \mathcal{C}} \left| X_j^T \theta \right| < 1 \implies \hat{\beta}_j^{(\lambda)} = 0$$

can be written as:

$$\forall j \in [\![p]\!]\,,\ \left| X_j^T c \right| + r \,\|\, X_j \,\| < 1 \implies \hat{\beta}_j^{(\lambda)} = 0$$

Rule	Center	Radius	Ingredients
Static Safe (El Ghaoui et al., 2012)	y/λ	$\check{R}_\lambda(\frac{y}{\lambda_{\max}})$	$\lambda_{\max} = \|X^\top y\|_\infty = \|x_{j^\star}^\top y\|$
Dynamic ST3 (Xiang et al., 2011)	$y/\lambda - \delta x_{j^\star}$	$(\check{R}_\lambda(\theta_k)^2 - \delta^2)^{\frac{1}{2}}$	$\delta = \left(\frac{\lambda_{\max}}{\lambda} - 1\right) / \|x_{j^\star}\|$
Dynamic Safe (Bonnefoy et al., 2014a)	y/λ	$\check{R}_\lambda(\theta_k)$	$\theta_k \in \Delta_X$ (*e.g.*, as in (11))
Sequential (Wang et al., 2013)	$\hat{\theta}^{(\lambda_{t-1})}$	$\left\| \frac{1}{\lambda_{t-1}} - \frac{1}{\lambda_t} \right\| \|y\|$	exact $\hat{\theta}^{(\lambda_{t-1})}$ required
GAP SAFE sphere (proposed)	θ_k	$r_{\lambda_t}(\beta_k, \theta_k) = \frac{1}{\lambda_t}\sqrt{2G_{\lambda_t}(\beta_k, \theta_k)}$	dual gap for β_k, θ_k

Table 1. Review of some common safe sphere tests.

- Some background
- The GAP safe rules
- Dual extrapolation for faster Lasso solvers

- Some background
- The GAP safe rules
- Dual extrapolation for faster Lasso solvers

Construction of the safe sphere

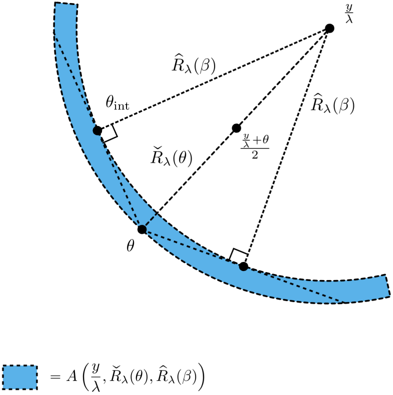
Elements of proof:

- $\hat{\theta}^{(\lambda)}$ is the closest feasible point to $\frac{y}{\lambda}$.

- By the weak duality theorem, for any $\theta \in \Delta_X$ and any $\beta \in \mathbb{R}^p$,

$$\underbrace{\frac{1}{2}\parallel y \parallel_2^2 - \frac{\lambda^2}{2}\left\| \theta - \frac{y}{\lambda} \right\|_2^2}_{D_\lambda(\theta)} \leq \underbrace{\frac{1}{2}\parallel y - X\beta \parallel_2^2 + \lambda \parallel \beta \parallel_1}_{P_\lambda(\beta)}$$

- By convexity of the feasible set Δ_X, the farthest away that $\hat{\theta}^{(\lambda)}$ can be from the dual feasible point θ is if $\hat{\theta}^{(\lambda)}$ is equal to θ_{int}.



Construction of the safe sphere

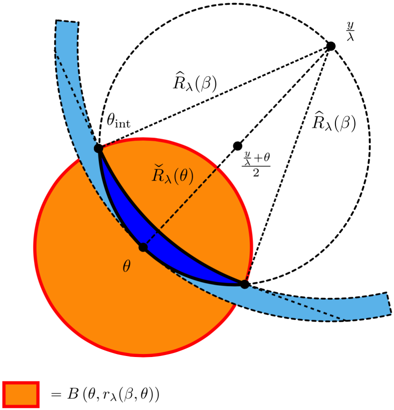
Elements of proof:

- $\hat{\theta}^{(\lambda)}$ is the closest feasible point to $\frac{y}{\lambda}$.

- By the weak duality theorem, for any $\theta \in \Delta_X$ and any $\beta \in \mathbb{R}^p$,

$$\underbrace{\frac{1}{2}\,\|\,y\,\|_2^2 - \frac{\lambda^2}{2}\left\|\,\theta - \frac{y}{\lambda}\,\right\|_2^2}_{D_\lambda(\theta)} \leq \underbrace{\frac{1}{2}\,\|\,y - X\beta\,\|_2^2 + \lambda\,\|\,\beta\,\|_1}_{P_\lambda(\beta)}$$

- By convexity of the feasible set Δ_X, the farthest away that $\hat{\theta}^{(\lambda)}$ can be from the dual feasible point θ is if $\hat{\theta}^{(\lambda)}$ is equal to θ_{int}.



Let $(\beta, \theta) \in \mathbb{R}^p \times \Delta_X$ be any primal-dual feasible pair and let $r_\lambda(\beta, \theta) = \frac{2}{\lambda^2}\left(P_\lambda(\beta) - D_\lambda(\theta)\right)$. The GAP safe sphere test reads:

$$\forall j \in [\![p]\!], \left| X_j^\top \theta \right| + r_\lambda(\beta, \theta) \| X_j \|_2 < 1 \implies \hat{\beta}_j^{(\lambda)} = 0$$

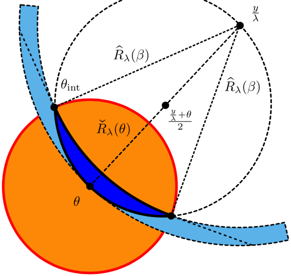
The dual feasible dual point is obtained by dual scaling, i.e,

$$\begin{cases} \theta_k = \alpha_k \rho_k \\ \alpha_k = \min \left[\max \left(\frac{y^\top \rho_k}{\lambda \| \rho_k \|^2}, \frac{-1}{\| X^\top \rho_k \|_\infty} \right), \frac{1}{\| X^\top \rho_k \|_\infty} \right] \end{cases} \tag{5}$$

with $\rho_k = y - W\beta_k$ the current residual.

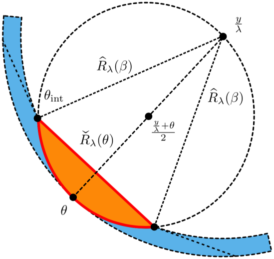
The GAP safe dome

The GAP safe sphere



$= B\left(\theta, r_\lambda(\beta,\theta)\right)$

The GAP safe dome



$= D\left(\frac{\frac{y}{\lambda}+\theta}{2}, \frac{\check{R}_\lambda(\theta)}{2}, 2\left(\frac{\widehat{R}_\lambda(\beta)}{\check{R}_\lambda(\theta)}\right)^2 - 1, \frac{\frac{y}{\lambda}-\theta}{\check{R}_\lambda(\theta)}\right)$

Sequential and dynamic

The GAP safe rules are sequential and dynamic:

- **Sequential**: Suppose $(\beta, \theta) \in \mathbb{R}^p \times \Delta_X$ are the approximate primal and dual solutions of the LASSO for λ_t. Then the ball with centre θ and radius $r_{\lambda+1}(\beta, \theta)$ can be used for screening at λ_{t+1}. In other words, sequential screening rules can be easily warm started.
- **Dynamic**: The safe region can be narrowed down while the optimisation proceeds.

Performance on real data

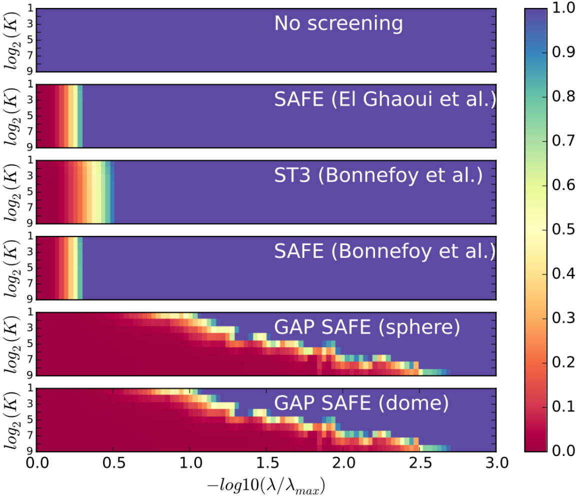


Figure 3. Proportion of active variables as a function of λ and the number of iterations K on the Leukemia dataset. Better strategies have longer range of λ with (red) small active sets.

- Some background
- The GAP safe rules
- Dual extrapolation for faster Lasso solvers

- Some background
- The GAP safe rules
- Dual extrapolation for faster Lasso solvers

Overview of the paper

Contribution:

The authors propose a method to construct an improved dual feasible point.

Consequences:

- ✓ The screening performance of the GAP safe rules is improved.
- ✓ A tighter control of optimality (through the stopping criterion) is obtained.
- ✓ A state-of-the-art LASSO solver is proposed based on an aggressive used of the improved GAP safe rules.

Classical construction of a dual feasible dual point

Let $\rho^t = y - W\beta^t$ be the residual at the t^{th} step of the optimisation. Classically, a dual feasible point is constructed via residuals scaling.

(version used in the GAP paper):

$$\begin{cases} \theta^t = \alpha^t \rho^t \\ \alpha^t = \min\left[\max\left(\frac{y^\top \rho^t}{\lambda \| \rho^t \|^2}, \frac{-1}{\| X^\top \rho^t \|_\infty}\right), \frac{1}{\| X^\top \rho^t \|_\infty}\right] \end{cases}$$

α^t solves:

$$\min_{\alpha \in \mathbb{R}} \left\| \alpha\rho^t - \frac{y}{\lambda} \right\|_2 \text{ s.t. } \left| X^\top \alpha \rho^t \right| \leq 1$$

(simpler version):

$$\begin{cases} \theta^t = \alpha^t \rho^t \\ \alpha^t = \min\left(\frac{1}{\lambda}, \frac{1}{\| X^\top \rho^t \|_\infty}\right) \end{cases}$$

α^t solves:

$$\min_{\alpha \in [0, \frac{1}{\lambda}]} \text{ s.t. } \left| X^\top \alpha \rho^t \right| \leq 1$$

Construction of an improved dual point

Let $K \in \mathbb{N}$ (default K=5).
Let $U^t = \left[\rho^{t+1-K} - \rho^{t-K}, \dots, \rho^t - \rho^{t-1}\right] \in \mathbb{R}^{n \times K}$.
Let $z \in \mathbb{R}^k$ be the solution to the linear system: $\left(U^t\right)^\top U^t z = \mathbf{1}_K$.
Let $c = \frac{1}{z^\top \mathbf{1}_K} z \in \mathbb{R}^K$.

Define:

$$\rho^t_{accel} = \begin{cases} \rho^t, & \text{if} t \leq K \\ \sum_{k=1}^{K} c_k \rho^{t+1-K}, & \text{if} t > K. \end{cases} \tag{6}$$

Then the extrapolated dual point is:

$$\theta^t_{accel} = \alpha^t \rho^t_{accel} \quad \text{where} \quad \alpha^t = \min\left(\frac{1}{\lambda}, \frac{1}{\| X^\top \rho^t \|_\infty}\right). \tag{7}$$

Construction of an improved dual point

The definition of θ^t_{accel} is based on the Minimal Polynomial Extrapolation method (MPE) (Cabay and Jackson, 1976). Let $\{x_n\}_{n\in\mathbb{N}}$ be a sequence of vectors:

- generated by a fixed point iterative method: $x_{n+1} = F(x_n)$,
- which admits a limit $s = \lim_{n\to\infty} x_n$

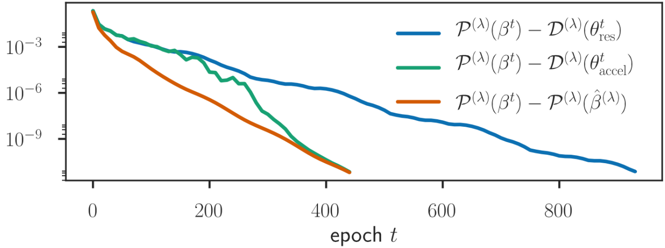
MPE provides an estimate of s that only depends on the k last iterates: $s_n^k = f(x_n, \ldots, x_{n-k})$.

Improved control of the suboptimality gap

Several stopping rules exist to decide when an iterative solver has converged. One of them is the duality gap.

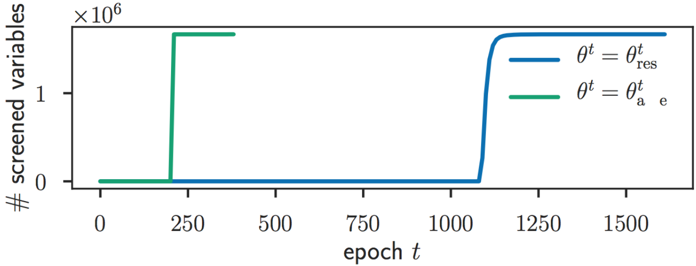
Weak duality implies that for any pair $(\beta, \theta) \in \mathbb{R} \times \Delta_X$, the suboptimality gap is upper bounded by the duality gap, i.e,

$$P^{(\lambda)}(\beta) - P^{(\lambda)}(\hat{\beta}^{(\lambda)}) \leq G^{(\lambda)}(\beta, \theta) \; .$$



Leukemia dataset ($n = 72, p = 7,129$), $\lambda = \frac{\lambda_{max}}{20}$.

Screening performance with the extrapolated dual point



Finance dataset ($n = 16,087$ and $p = 1,668,738$), $\lambda = \frac{\lambda_{max}}{5}$, $\epsilon = 10^{-6}$.

A working set algorithm with aggressive GAP screening

Screening techniques discard irrelevant variables while working set techniques prioritise important variables, and iteratively solve subproblems on the set of prioritised variables until a convergence criterion is met.

For any primal dual feasible pair $(\beta, \theta) \in \mathbb{R}^n \times \Delta_X$, the GAP safe sphere test:

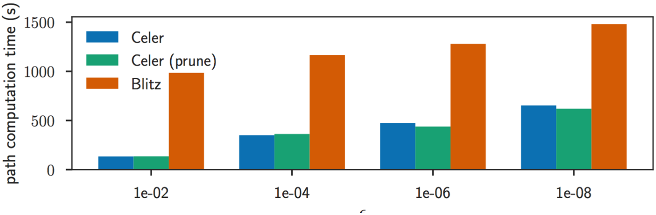
$$\forall j \in [\![p]\!] \,, \left| X_j^\top \theta \right| + r_\lambda(\beta, \theta) \, \| X_j \|_2 < 1 \implies \hat{\beta}_j^{(\lambda)} = 0$$

can be rewritten as

$$\forall j \in [\![p]\!] \,, d_j(\theta) > r_\lambda(\beta, \theta) \implies \hat{\beta}_j^{(\lambda)} = 0 \text{ where } d_j(\theta) = \frac{1 - | X_j^\top \theta |}{\| X_j \|_2}$$

The WS algorithm proposed by the authors, Celer,

1. Use θ^t_{accel} as dual feasible point.
2. Include in the working set the feature with lowest $d_j(\theta^t_{accel})$.



Finance dataset ($n = 16,087$ and $p = 1,668,738$), 100 values of λ from λ_{max} to $\frac{\lambda_{max}}{100}$.

ϵ	10^{-2}	10^{-3}	10^{-4}	10^{-6}
CELER	5	7	8	10
BLITZ	25	26	27	30
scikit-learn	470	1350	2390	-

Finance dataset, $\lambda = \frac{\lambda_{max}}{20}$.

Stan Cabay and LW Jackson. “A polynomial extrapolation method for finding limits and antilimits of vector sequences”. In: *SIAM Journal on Numerical Analysis* 13.5 (1976), pp. 734–752.

T. Johnson and C. Guestrin. “BLITZ: A principled meta-algorithm for scaling sparse optimization”. In: *Proc. 32nd Int. Conf. Mach. Learn. - ICML ’15*. 2015, pp. 1171–1179.

E. Ndiaye et al. “Gap Safe screening rules for sparsity enforcing penalties”. In: *J. Mach. Learn. Res.* 18.128 (2017), pp. 1–33.