Mind the duality gap: safer rules for the Lasso

Alexandre Gramfort
http://alexandre.gramfort.net
Télécom Paristech, CNRS LTCI

Joint work with:
Olivier Fercoq (Télécom ParisTech, CNRS LTCI)
Joseph Salmon (Télécom ParisTech, CNRS LTCI)

Table of Contents

The Lasso

- $y \in \mathbb{R}^n$: target, signal
- $X = [\mathbf{x}_1, \ldots, \mathbf{x}_p] \in \mathbb{R}^{n \times p}$: design, dictionary

Objective: approximate $y \approx X\beta$ with a **sparse** vector $\beta \in \mathbb{R}^p$

The Lasso way:

$$\hat{\beta}^{(\lambda)} \in \mathop{\arg\min}_{\beta \in \mathbb{R}^p} \left(\underbrace{\frac{1}{2}\|y - X\beta\|^2}_{\textbf{data fitting term}} + \underbrace{\lambda\|\beta\|_1}_{\textbf{sparsity-inducing penalty}} \right)$$

- Convex optimization problem
- Need to tune/choose λ (standard is Cross-Validation)

The Lasso

- $y \in \mathbb{R}^n$: target, signal
- $X = [\mathbf{x}_1, \ldots, \mathbf{x}_p] \in \mathbb{R}^{n \times p}$: design, dictionary

Objective: approximate $y \approx X\beta$ with a **sparse** vector $\beta \in \mathbb{R}^p$

The Lasso way:

$$\boxed{\hat{\beta}^{(\lambda)} \in \mathop{\arg\min}_{\beta \in \mathbb{R}^p} \left(\underbrace{\frac{1}{2}\|y - X\beta\|^2}_{\text{data fitting term}} + \underbrace{\lambda\|\beta\|_1}_{\text{sparsity-inducing penalty}} \right)}$$

- Convex optimization problem
- Need to tune/choose λ (standard is Cross-Validation)

The denoising case

Suppose the design is simple: $n = p$ and $X = \mathrm{Id_n}$, meaning the atoms are canonical elements: $\mathbf{x}_j = (0, \cdots, 0, \underset{\substack{\uparrow \\ j}}{1}, 0, \cdots, 1)^\top$

$$\hat{\beta}^{(\lambda)} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \left(\frac{1}{2} \|y - \beta\|^2 + \lambda \|\beta\|_1 \right)$$

$$\hat{\beta}^{(\lambda)} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left(\frac{1}{2} \|y - \beta\|^2 + \lambda \|\beta\|_1 \right) \qquad \text{(strictly convex)}$$

$$\hat{\beta}_j^{(\lambda)} = \underset{\beta_j \in \mathbb{R}}{\arg\min} \left(\frac{1}{2} (y_i - \beta_j)^2 + \lambda |\beta_j| \right), \forall j \in [n] \qquad \text{(separable)}$$

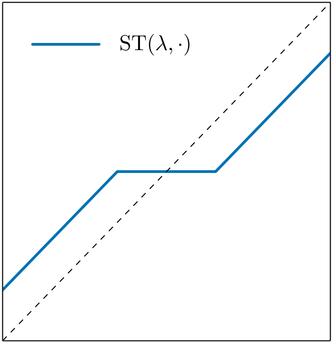
This reduces to a 1D problem.
<u>Rem</u>: The solution is called the **proximal** operator of $\lambda \| \cdot \|_1$

Soft-Thresholding

The 1D problem has a closed form solution: **Soft-Thresholding**:

$$
\begin{aligned}
\mathrm{ST}(\lambda, y) &= \operatorname*{arg\,min}_{\beta \in \mathbb{R}} \left(\frac{1}{2}(y - \beta)^2 + \lambda|\beta| \right) \\
&= \operatorname{sign}(y) \cdot (|y| - \lambda)_+
\end{aligned}
$$

with the notation $(\cdot)_+ = \max(0, \cdot)$



Proof: easy with sub-gradients and Fermat condition

The Lasso: algorithmic point of view

Possible algorithms for solving this **convex** program:

- Homotopy method / LARS : very efficient for small p Osborne *et al.* (2000), Efron *et al.* (2004) and full path
- Forward - Backward / proximal algorithm: useful in signal/image for case where $r \to \mathbf{x}_j^\top r$ is cheap to compute (*e.g.*, with FFT, Fast Wavelet Transform, etc.) Beck and Teboulle (2009)
- Coordinate Descent: very useful for large p and potentially sparse matrix X (*e.g.*, from text encoding) Friedman *et al.* (2007)

Objective of this work: speed-up Lasso solvers

$$\boxed{\hat{\beta}^{(\lambda)} \in \mathop{\arg\min}_{\beta \in \mathbb{R}^p} \left(\underbrace{\frac{1}{2}\|y - X\beta\|_2^2}_{\text{data fitting term}} + \underbrace{\lambda\|\beta\|_1}_{\text{sparsity-inducing penalty}} \right)}$$

- Compute $\hat{\beta}^{(\lambda)}$ for **many** λ's: *e.g.*, T values from $\lambda_{\max} := \|X^\top y\|_\infty$ to $\epsilon\lambda_{\max}$ on log-scale $(T = 100, \epsilon = 0.001)$
- **Flexible**: provide a way that can beneficiate to most solvers (though mainly focused on Coordinate Descent)
- **Easy to code**

Table of Contents

Dual problem

Primal function : $P_\lambda(\beta) = \frac{1}{2}\|y - X\beta\|^2 + \lambda\|\beta\|_1$

Dual feasible set : $\Delta_X = \left\{\theta \in \mathbb{R}^n \ : \ |\mathbf{x}_j^\top \theta| \leqslant 1, \forall j \in [p]\right\}$

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

<u>Rem</u>: The dual feasible set is a polytope

$$\Delta_X = \bigcap_{j=1}^{p} \left\{\theta \in \mathbb{R}^n : |\mathbf{x}_j^\top \theta| \leqslant 1\right\} = \left\{\theta \in \mathbb{R}^n : \|X^\top \theta\|_\infty \leqslant 1\right\}$$

<u>Rem</u>: the dual formulation is obtained using an additional variable $z = (y - X\beta)/\lambda$ and considering the Lagrangian, *cf.* Kim *et al.* (2007)

Multi-task / Multi-class problem

Primal : $$\widehat{\mathrm{B}}^{(\lambda)} \in \underset{\mathrm{B}\in\mathbb{R}^{p\times q}}{\arg\min} \underbrace{\sum_{i=1}^{n} f_i(x_i^\top \mathrm{B}) + \lambda\Omega(\mathrm{B})}_{P_\lambda(\mathrm{B})}$$

Dual feasible set : $$\Delta_X = \left\{\Theta \in \mathbb{R}^{n\times q} \ : \ \|\mathbf{x}_j^\top \Theta\|_2 \leqslant 1, \forall j \in [p]\right\}$$

Dual: $$\widehat{\Theta}^{(\lambda)} = \underset{\Theta\in\Delta_X}{\arg\max} \underbrace{-\sum_{i=1}^{n} f_i^*(-\lambda\Theta_{i,:})}_{D_\lambda(\Theta)}$$

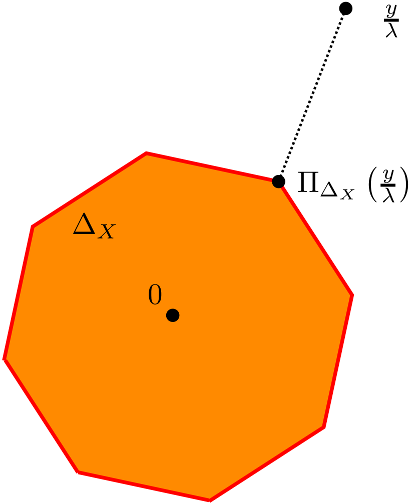
with:

$$\Delta_X = \bigcap_{j=1}^{p} \left\{\Theta \in \mathbb{R}^{n\times q} : |\mathbf{x}_j^\top \Theta|_2 \leqslant 1\right\} = \left\{\Theta \in \mathbb{R}^{n\times q} : \|X^\top \Theta\|_{2\infty} \leqslant 1\right\}$$

Rem: Problem for Gap Safe rules: Compute efficiently Gap and dual feasible points

Geometric interpretation

The dual optimal solution is the projection of y/λ over the dual feasible set $\Delta_X = \left\{\theta \in \mathbb{R}^n : \|X^\top \theta\|_\infty \leqslant 1\right\} : \hat{\theta}^{(\lambda)} = \Pi_{\Delta_X}(y/\lambda)$



Duality Gap properties

- **Primal objective:** P_λ, **Primal solution:** $\hat{\beta}^{(\lambda)} \in \mathbb{R}^p$
- **Dual objective:** D_λ, **Primal solution:** $\hat{\theta}^{(\lambda)} \in \Delta_X \subset \mathbb{R}^n$,

Duality gap: for any $\beta \in \mathbb{R}^p$ and any $\theta \in \Delta_X$,

$$\begin{aligned} G_\lambda(\beta, \theta) =& P_\lambda(\beta) - D_\lambda(\theta) \\ =& \frac{1}{2} \|X\beta - y\|^2 + \lambda \|\beta\|_1 - (\frac{1}{2} \|y\|^2 - \frac{\lambda^2}{2} \left\|\theta - \frac{y}{\lambda}\right\|^2) \end{aligned}$$

Rem: For all $\beta \in \mathbb{R}^p, \theta \in \Delta_X$,

$$D_\lambda(\theta) \leqslant D_\lambda(\hat{\theta}^{(\lambda)}) = P_\lambda(\hat{\beta}^{(\lambda)}) \leqslant P_\lambda(\beta) \quad \textbf{(Strong duality)}$$

Consequences:

- $G_\lambda(\beta, \theta) \geqslant 0$
- $G_\lambda(\beta, \theta) \leqslant \epsilon \implies P_\lambda(\beta) - P_\lambda(\hat{\beta}^{(\lambda)}) \leqslant \epsilon$ (stopping criterion!)

Duality Gap properties

- **Primal objective:** P_λ, **Primal solution:** $\hat{\beta}^{(\lambda)} \in \mathbb{R}^p$
- **Dual objective:** D_λ, **Primal solution:** $\hat{\theta}^{(\lambda)} \in \Delta_X \subset \mathbb{R}^n$,

Duality gap: for any $\beta \in \mathbb{R}^p$ and any $\theta \in \Delta_X$,

$$\begin{aligned} G_\lambda(\beta, \theta) =& P_\lambda(\beta) - D_\lambda(\theta) \\ =& \frac{1}{2} \left\| X\beta - y \right\|^2 + \lambda \left\| \beta \right\|_1 - \left(\frac{1}{2} \left\| y \right\|^2 - \frac{\lambda^2}{2} \left\| \theta - \frac{y}{\lambda} \right\|^2\right) \end{aligned}$$

Rem: For all $\beta \in \mathbb{R}^p, \theta \in \Delta_X$,

$$D_\lambda(\theta) \leqslant D_\lambda(\hat{\theta}^{(\lambda)}) = P_\lambda(\hat{\beta}^{(\lambda)}) \leqslant P_\lambda(\beta) \quad (\textbf{Strong duality})$$

Consequences:

- $G_\lambda(\beta, \theta) \geqslant 0$
- $G_\lambda(\beta, \theta) \leqslant \epsilon \implies P_\lambda(\beta) - P_\lambda(\hat{\beta}^{(\lambda)}) \leqslant \epsilon$ (stopping criterion!)

KKT: Karush-Khun-Tucker (KKT) conditions

- **Primal solution :** $\hat{\beta}^{(\lambda)} \in \mathbb{R}^p$
- **Dual solution :** $\hat{\theta}^{(\lambda)} \in \Delta_X \subset \mathbb{R}^n$

Primal/Dual link: $\boxed{y = X\hat{\beta}^{(\lambda)} + \lambda\hat{\theta}^{(\lambda)}}$

Necessary and sufficient optimality conditions:

KKT/Fermat: $$\boxed{\forall j \in [p],\ x_j^\top \hat{\theta}^{(\lambda)} \in \begin{cases} \{\operatorname{sign}(\hat{\beta}_j^{(\lambda)})\} & \text{if} \quad \hat{\beta}_j^{(\lambda)} \neq 0, \\ [-1, 1] & \text{if} \quad \hat{\beta}_j^{(\lambda)} = 0. \end{cases}}$$

Rem: the KKT implies that $\forall \lambda \geqslant \lambda_{\max} = \|X^\top y\|_\infty$, $0 \in \mathbb{R}^p$ is the (unique here) primal solution for P_λ

Geometric interpretation

A simple dual point is: $y/\lambda_{\max} \in \Delta_X$

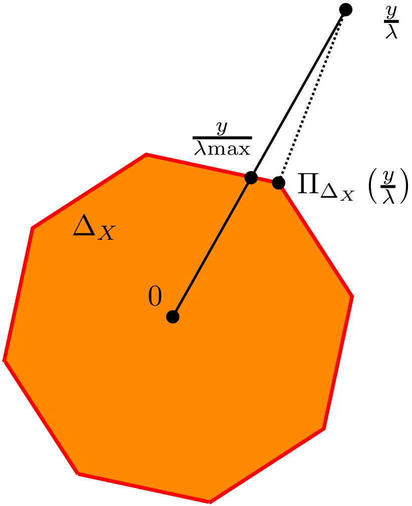


Table of Contents

Safe rules - safe regions
El Ghaoui *et al.* (2012)

Screening thanks to the KKT is possible:

$$\boxed{\text{If } |\mathbf{x}_j^\top \hat{\theta}^{(\lambda)}| < 1 \text{ then, } \hat{\beta}_j^{(\lambda)} = 0}$$

Beware: $\underline{\hat{\theta}^{(\lambda)} \text{ is unknown}}$, so one need to consider a **safe region** $\mathcal{C}$ containing $\hat{\theta}^{(\lambda)}$, *i.e.*, $\hat{\theta}^{(\lambda)} \in \mathcal{C}$, leading to :

$$\textbf{safe rule} : \quad \boxed{\text{If } \sup_{\theta \in \mathcal{C}} |\mathbf{x}_j^\top \theta| < 1 \text{ then } \hat{\beta}_j^{(\lambda)} = 0} \qquad (\star)$$

The new goal is simple, find a region $\mathcal{C}$:

- as narrow as possible containing $\hat{\theta}^{(\lambda)}$
- such that $\mu_{\mathcal{C}} : \begin{cases} \mathbb{R}^n & \mapsto \mathbb{R}^+ \\ \mathbf{x} & \to \sup_{\theta \in \mathcal{C}} |\mathbf{x}^\top \theta| \end{cases}$ is easy to compute

Safe sphere rules

Let $\mathcal{C} = B(c, r)$ be a ball of center $c \in \mathbb{R}^n$ and radius $r > 0$. Then simple computation provide:

$$\mu_{\mathcal{C}}(\mathbf{x}) = |\mathbf{x}^\top c| + r\|\mathbf{x}\|$$

so the safe rule becomes

$$\boxed{\text{If } |\mathbf{x}_j^\top c| + r\|\mathbf{x}_j\| < 1 \text{ then } \hat{\beta}_j^{(\lambda)} = 0} \tag{1}$$

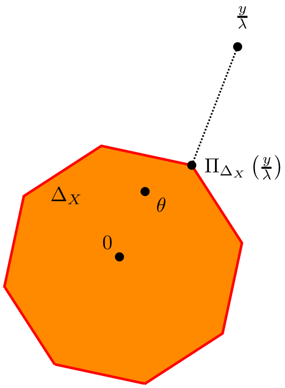
We say we screen-out the variables $\mathbf{x}_j$ satisfying (1)

Active set : $A^{(\lambda)}(\mathcal{C}) = \{j \in [p] : \mu_{\mathcal{C}}(\mathbf{x}_j) \geqslant 1\}$

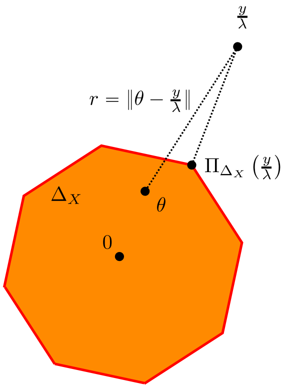
New objective:

- find r as small as possible
- find c as close to $\hat{\theta}^{(\lambda)}$ as possible.

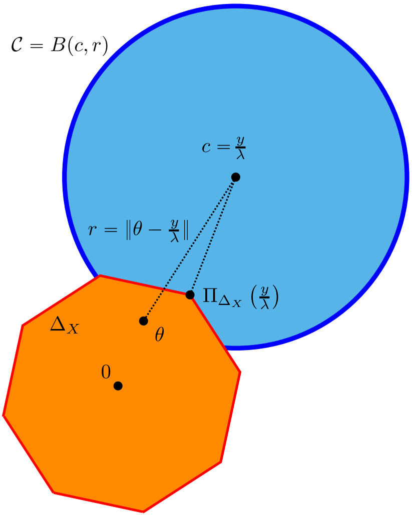
Creating safe sphere



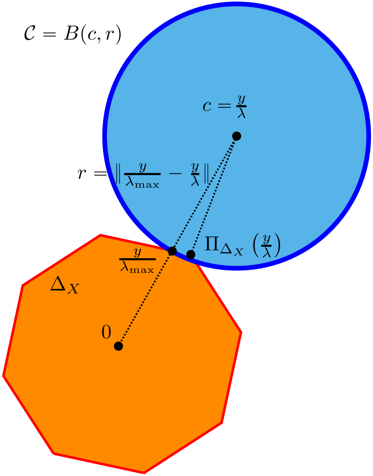
Creating safe sphere



Creating safe sphere



Original safe rule: El Ghaoui *et al.* (2012)



Original static safe rule : El Ghaoui *et al.* (2012)

Static safe region: before any optimization, for a fix λ.

$$\mathcal{C} = B(c, r) = B(y/\lambda, \|y/\lambda_{\max} - y/\lambda\|)$$

$$\boxed{\text{If } |\mathbf{x}_j^\top y| < \lambda(1 - \|y/\lambda_{\max} - y/\lambda\| \|\mathbf{x}_j\|) \text{ then } \hat{\beta}_j^{(\lambda)} = 0} \tag{2}$$

Rem: This reinterprets screening methods for **variable selection**:
"If $|\mathbf{x}_j^\top y|$ is small, remove $\mathbf{x}_j$" as a safe rule for the Lasso

Dynamic safe rule

Dynamic point of view: build $\theta_k \in \Delta_X$, evolving with the solver iterations to get refined safe rules Bonnefoy *et al.* (2014, 2015)

$$\text{Remind link at optimum:} \quad \lambda\hat{\theta}^{(\lambda)} = y - X\hat{\beta}^{(\lambda)}$$

$$\text{Current } \textbf{residual} \text{ for primal point } \beta_k\text{:} \quad \rho_k = y - X\beta_k$$

Dual candidate: choose θ_k proportional to the residual

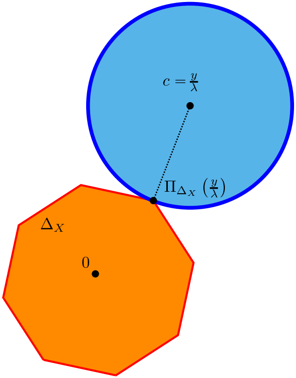
$$\begin{aligned} \theta_k &= \alpha_k \rho_k, \\ \text{where} \quad \alpha_k &= \min\left[\max\left(\frac{y^\top \rho_k}{\lambda \|\rho_k\|^2}, \frac{-1}{\|X^\top \rho_k\|_\infty}\right), \frac{1}{\|X^\top \rho_k\|_\infty}\right]. \end{aligned}$$

Motivation: projecting over the convex set $\Delta_X \cap \operatorname{Span}(\rho_k)$ is cheap

Limits of previous dynamic rules

The radius $r_k = \|\theta_k - y/\lambda\|$ does not converge to zero. The limiting safe sphere is

$$\mathcal{C} = B(y/\lambda, \|\Pi_{\Delta_X}(y/\lambda) - y/\lambda\|)$$



Sequential safe rule Wang *et al.* (2013)

Warm start main idea: to compute the Lasso for T different λ's, say $\lambda_0, \cdots, \lambda_{T-1}$, reuse computation done at λ_{t-1} to get $\hat{\beta}^{(\lambda_t)}$:

- **Warm start** (for the primal) = standard trick to accelerate iterative solvers: Initialize to $\hat{\beta}^{(\lambda_{t-1})}$ to compute $\hat{\beta}^{(\lambda_t)}$
- **Warm start** (for the dual) = sequential safe rule use $\hat{\theta}^{(\lambda_{t-1})}$ to help screening for $\hat{\beta}^{(\lambda_t)}$.

Major issue: in prior works $\hat{\theta}^{(\lambda_{t-1})}$ needs to be **known exactly**!

<u>Rem</u>: Unrealistic except for $\hat{\theta}^{(\lambda_0)} = y/\lambda_{\max} = y/\|X^\top y\|_\infty$

Table of Contents

GAP Safe sphere

For any $\beta \in \mathbb{R}^p, \theta \in \Delta_X$

$$G_\lambda(\beta, \theta) = \frac{1}{2}\left\|X\beta - y\right\|^2 + \lambda\left\|\beta\right\|_1 - \left(\frac{1}{2}\left\|y\right\|^2 - \frac{\lambda^2}{2}\left\|\theta - \frac{y}{\lambda}\right\|^2\right)$$

Gap Safe ball: $\boxed{B(\theta, r_\lambda(\beta, \theta)), \text{ where } r_\lambda(\beta, \theta) = \sqrt{2G_\lambda(\beta, \theta)}/\lambda^2}$

Rem: If $\beta_k \to \hat{\beta}^{(\lambda)}$ and $\theta_k \to \hat{\theta}^{(\lambda)}$ then $G_\lambda(\beta_k, \theta_k) \to 0$: a converging solver leads to converging safe rule!

The GAP SAFE sphere is safe:

- $D_\lambda(\hat{\theta}^{(\lambda)}) \leqslant P_\lambda(\beta_k)$ (weak Duality)
- D_λ is λ^2-strongly concave so for any $\theta_1, \theta_2 \in \mathbb{R}^n$,

$$D_\lambda(\theta_1) \leqslant D_\lambda(\theta_2) + \langle \nabla D_\lambda(\theta_2), \theta_1 - \theta_2 \rangle - \frac{\lambda^2}{2} \|\theta_1 - \theta_2\|_2^2$$

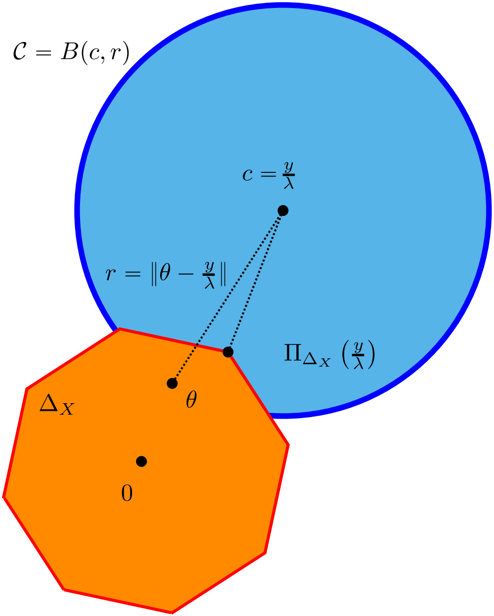
- $\hat{\theta}^{(\lambda)}$ maximizes D_λ over Δ_X, so

$$\forall \theta \in \Delta_X, \qquad \langle \nabla D_\lambda(\hat{\theta}^{(\lambda)}), \theta - \hat{\theta}^{(\lambda)} \rangle \leqslant 0$$

To conclude, for a $\theta \in \Delta_X$:

$$\begin{aligned}\frac{\lambda^2}{2} \left\| \theta - \hat{\theta}^{(\lambda)} \right\|_2^2 &\leqslant D_\lambda(\hat{\theta}^{(\lambda)}) - D_\lambda(\theta) + \langle \nabla D_\lambda(\hat{\theta}^{(\lambda)}), \theta - \hat{\theta}^{(\lambda)} \rangle \\ &\leqslant P_\lambda(\beta_k) - D_\lambda(\theta)\end{aligned}$$

Dynamic safe sphere Bonnefoy *et al.* (2014)



Dynamic GAP safe sphere

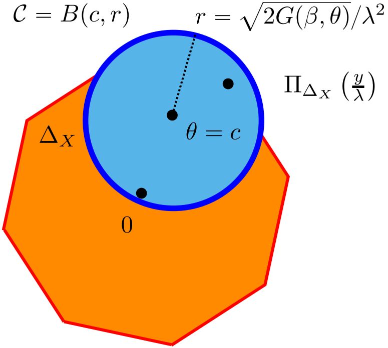


Table of Contents

Algorithm 1 Coordinate descent (Lasso)

```
Input: X, y, ε, K, f, (λ_t)_{t∈[T−1]}
Initialization:   λ_0 = λ_max,   β^{λ_0} = 0
for t ∈ [T − 1] do                                    ▷ Loop over λ's
    β ← β^{λ_{t−1}}                                   ▷ previous ε-solution
    for k ∈ [K] do
        if k mod f = 1 then
            Construct θ ∈ Δ_X
            if G_{λ_t}(β, θ) ⩽ ε then                 ▷ Stop if duality gap small
                β^{λ_t} ← β
                break
            end if
        end if
        for j ∈ [p] do                                ▷ Soft-Threshold coordinates
            β_j ← ST( λ_t / ‖x_j‖², β_j − x_j^⊤(Xβ − y) / ‖x_j‖² )
        end for
    end for
end for
```

Algorithm 2 Coordinate descent (Lasso) with GAP Safe screening

Input: $X, y, \epsilon, K, f, (\lambda_t)_{t \in [T-1]}$

Initialization: $\lambda_0 = \lambda_{\max}, \quad \beta^{\lambda_0} = 0$

for $t \in [T-1]$ **do** ▷ Loop over λ's

  $\beta \leftarrow \beta^{\lambda_{t-1}}$ ▷ previous ϵ-solution

  for $k \in [K]$ **do**

    if $k \mod f = 1$ **then**

      Construct $\theta \in \Delta_X$, $A^{\lambda_t}(\mathcal{C}) = \{j \in [p] : \mu_{\mathcal{C}}(\mathbf{x}_j) \geqslant 1\}$

      if $G_{\lambda_t}(\beta, \theta) \leqslant \epsilon$ **then** ▷ Stop if duality gap small

        $\beta^{\lambda_t} \leftarrow \beta$

        break

      end if

    end if

    for $j \in A^{\lambda_t}(\mathcal{C})$ **do** ▷ Soft-Threshold coordinates

      $\beta_j \leftarrow \mathrm{ST}\left(\frac{\lambda_t}{\|\mathbf{x}_j\|^2}, \beta_j - \frac{\mathbf{x}_j^\top (X\beta - y)}{\|\mathbf{x}_j\|^2}\right)$

    end for

  end for

end for

Gap safe rules: benefits?

- it is a dynamic rule (by construction)
- it is a sequential rule (without any more effort)
- the safe region is converging toward $\{\hat{\theta}^{(\lambda)}\}$
- it works better in practice

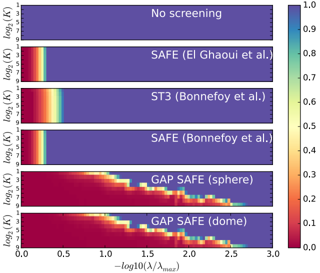


Figure: Proportion of active variables as a function of λ and the number of iterations K on the Leukemia dataset. Better strategies have longer range of λ with (red) small active sets (dense data: $n = 72, p = 7129$).

Computing time

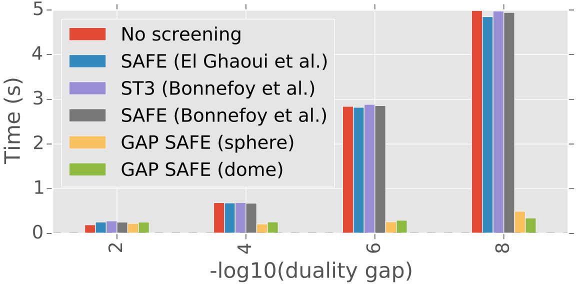


Figure: Time to reach convergence using various screening rules on the Leukemia dataset (dense data: $n = 72, p = 7129$). Full path with 100 values of λ on logarithmic grid from λ_{max} to $\lambda_{max}/1000$.

Computing time

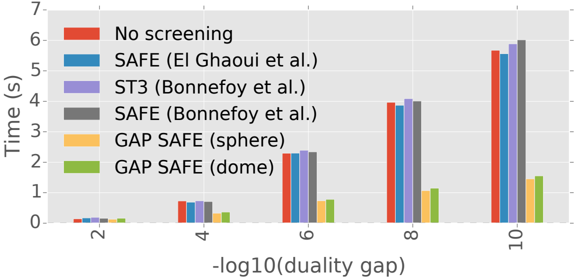


Figure: Time to reach convergence using various screening rules on sparse data (text features from 20 news group, $n = 961, p = 10094$). Full path with 100 values of λ on logarithmic grid from λ_{max} to $\lambda_{max}/1000$.

Conclusion and future work

- New safe screening rule based on duality gap
- Theoretically: convergent safe region
- Improves computational efficiency on Coordinate Descent implementation
- New work: group-Lasso, multitask Lasso, logistic regression with ℓ_1 regularization, multiclass logistic regression with ℓ_1/ℓ_2 regularization to appear in NIPS 2015 conference.
- Python implementation soon in Scikit-Learn (Pedregosa *et al.* JMLR (2011)) `http://scikit-learn.org`

EDDP Wang *et al.* (2013) can remove useful variables