GAP safe screening rules for sparsity enforcing penalties

Joseph Salmon
http://josephsalmon.eu
LTCI, CNRS, Télécom Paristech, Université Paris-Saclay

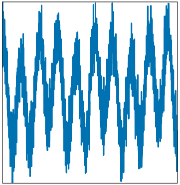
Joint work with:
Olivier Fercoq (Télécom ParisTech)
Alexandre Gramfort (Télécom ParisTech)
Eugene Ndiaye (Télécom ParisTech)

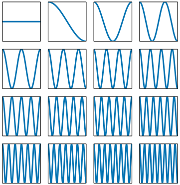
Table of Contents

Sparsity of signals is all around

Signals can often be represented through a combination of a few **atoms** / **features** :

- Fourier decomposition for sounds

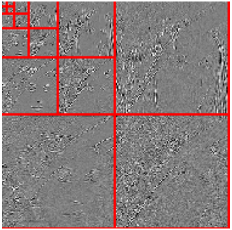




Sparsity of signals is all around

Signals can often be represented through a combination of a few **atoms** / **features** :

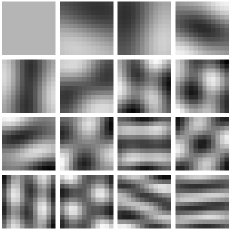
- Fourier decomposition for sounds
- Wavelet for images (1990's)



Sparsity of signals is all around

Signals can often be represented through a combination of a few **atoms** / **features** :

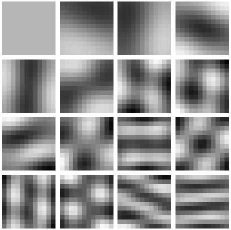
- Fourier decomposition for sounds
- Wavelet for images (1990's)
- Dictionary learning for images (late 2000's)

Sparsity of signals is all around

Signals can often be represented through a combination of a few **atoms** / **features** :

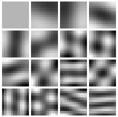
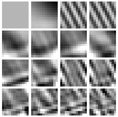
- Fourier decomposition for sounds
- Wavelet for images (1990's)
- Dictionary learning for images (late 2000's)
- etc.



Sparse linear model

Let $y \in \mathbb{R}^n$ be a signal


Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_p] \in \mathbb{R}^{n \times p}$ be a collection of atoms/features: corresponds to a **dictionary**

X well suited if one can approximate the signal $y \approx X\beta$ with a **sparse** vector $\beta \in \mathbb{R}^p$

$$\underbrace{\begin{pmatrix} y \end{pmatrix}}_{y \in \mathbb{R}^n} \approx \underbrace{\begin{pmatrix} \mathbf{x}_1 & \vert & \ldots & \vert & \mathbf{x}_p \end{pmatrix}}_{X \in \mathbb{R}^{n \times p}} \cdot \underbrace{\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}}_{\beta \in \mathbb{R}^p}$$

Objectives:

- Estimation β
- Prediction $X\beta$

Constraints: large p, n, sparse β

The Lasso and variations

Vocabulary: the "Modern least square" Candès *et al.* (2008)

- Statistics: **Lasso** Tibshirani (1996)
- Signal processing variant: **Basis Pursuit** Chen *et al.* (1998)

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

- Uniqueness not automatic, see discussion in Tibshirani (2013)
- Solutions are sparse (for well chosen λ's)
- Need to tune/choose λ (standard is Cross-Validation)
- Theoretical guaranties Bickel, Ritov and Tsybakov (2009)
- Refinements: Adaptive Lasso Zou (2006), $\sqrt{\text{Lasso}}$ Belloni *et al.* (2011), Scaled Lasso Zhang and Zhang (2012)...

The Lasso: algorithmic point of view

Commonly used 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 (*i.e.*, compute solution for "all" λ's). For limits see Mairal and Yu (2012)
- ISTA, Forward - Backward, proximal algorithm:
  useful in signal processing where $r \to X^\top r$ is cheap to compute (*e.g.*, FFT, Fast Wavelet Transform, etc.) Beck and Teboulle (2009)
- Coordinate descent:
  useful for large p and (unstructured) sparse matrix X, *e.g.*, for text encoding Friedman *et al.* (2007)

Objective of this work: speed-up Lasso solvers

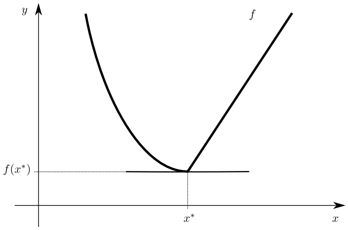
$$\hat{\beta}^{(\lambda)} \in \mathop{\arg\min}_{\beta \in \mathbb{R}^p} \Big(\underbrace{\frac{1}{2}\|y - X\beta\|_2^2}_{\text{data fitting term}} + \underbrace{\lambda\|\beta\|_1}_{\text{sparsity-inducing penalty}} \Big)$$

- Compute $\hat{\beta}^{(\lambda)}$ for **many** λ's: *e.g.*, T values from $\lambda_{\max} := \|X^\top y\|_\infty$ to $\lambda_{\min} = \epsilon\lambda_{\max}$ on log-scale
  Default value in R-glmnet : $T = 100, \epsilon = 0.001$
- **Flexible**: can be adapted to any iterative solver (but not to LARS!), here focus on Coordinate Descent
- **Easy to code** contrarily to **Strong Rule** Tibshirani *et al.* (2012)

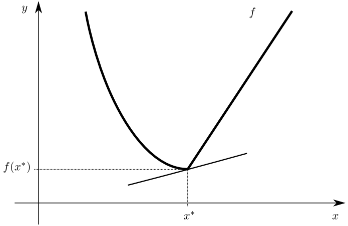
Rem: Starting is clear pick $\lambda = \lambda_{\max}$ but ending is not : $\lambda_{\min}$?

Table of Contents

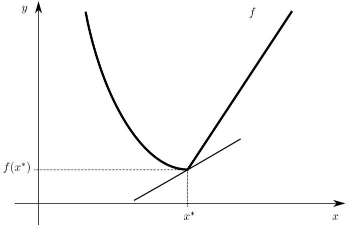
Sub-gradients / sub-differential



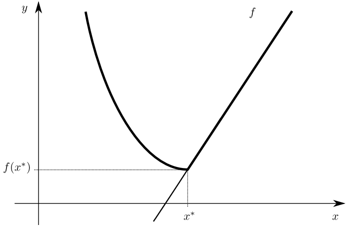
Sub-gradients / sub-differential



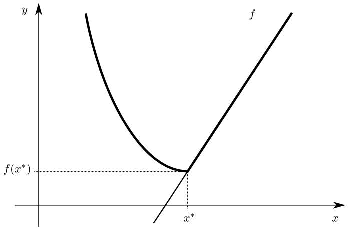
Sub-gradients / sub-differential



Sub-gradients / sub-differential



Sub-gradients / sub-differential



Definition: sub-gradient / sub-differential

For $f : \mathbb{R}^d \to \mathbb{R}$ a convex function, $u \in \mathbb{R}^d$ is a **sub-gradient** of f at x^*, if for all $x \in \mathbb{R}^d$ one has

$$f(x) \geqslant f(x^*) + \langle u, x - x^* \rangle$$

The **sub-differential** is the the <u>set</u>
$\partial f(x^*) = \{u \in \mathbb{R}^d : \forall x \in \mathbb{R}^d, f(x) \geqslant f(x^*) + \langle u, x - x^* \rangle\}$.

<u>Rem</u>: if the sub-gradient is unique, you recover the gradient

Fermat's rule: first order condition

Theorem

A point x^* is a minimum of a convex function $f : \mathbb{R}^d \to \mathbb{R}$ if and only if $0 \in \partial f(x^*)$

Proof: use the definition of sub-gradients:

- 0 is a sub-gradient of f at x^* if and only if $\forall x \in \mathbb{R}^d, f(x) \geqslant f(x^*) + \langle 0, x - x^* \rangle$

Fermat's rule: first order condition

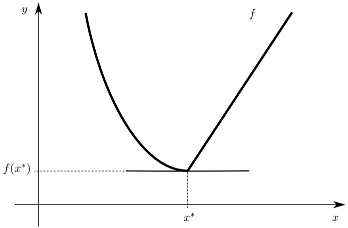
Theorem

A point x^* is a minimum of a convex function $f : \mathbb{R}^d \to \mathbb{R}$ if and only if $0 \in \partial f(x^*)$

Proof: use the definition of sub-gradients:

- 0 is a sub-gradient of f at x^* if and only if $\forall x \in \mathbb{R}^d, f(x) \geqslant f(x^*) + \langle 0, x - x^* \rangle$

Rem: Visually it corresponds to a horizontal tangent



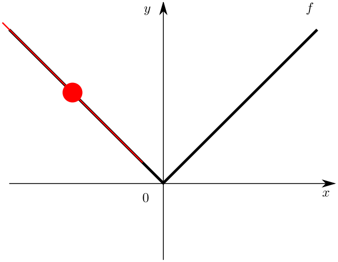
Sub-differential of the absolute value

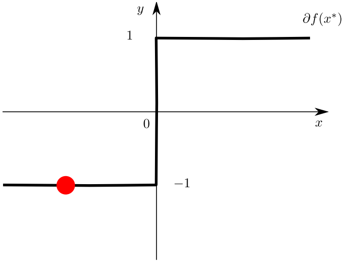
Function (abs):

$$f : \begin{cases} \mathbb{R} & \to \mathbb{R} \\ x & \mapsto |x| \end{cases}$$

Sub-differential (sign)

$$\partial f(x^*) = \begin{cases} \{-1\} & \text{if } x^* \in]-\infty, 0[\\ \{1\} & \text{if } x^* \in]0, \infty[\\ [-1, 1] & \text{if } x^* = 0 \end{cases}$$

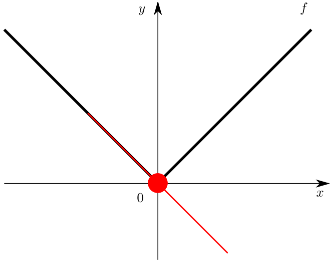




Sub-differential of the absolute value

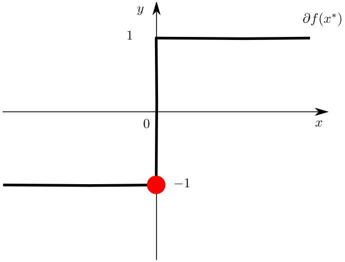
Function (abs):

$$f : \begin{cases} \mathbb{R} & \rightarrow \mathbb{R} \\ x & \mapsto |x| \end{cases}$$



Sub-differential (sign)

$$\partial f(x^*) = \begin{cases} \{-1\} & \text{if } x^* \in]-\infty, 0[\\ \{1\} & \text{if } x^* \in]0, \infty[\\ [-1, 1] & \text{if } x^* = 0 \end{cases}$$



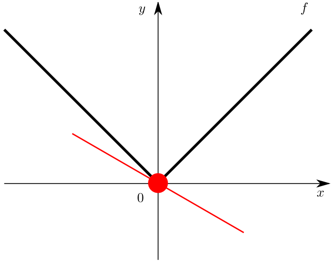
Sub-differential of the absolute value

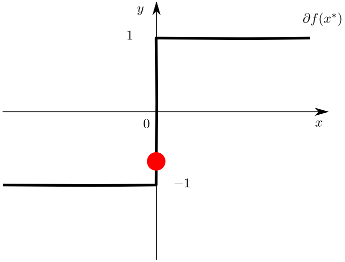
Function (abs):

$$f : \begin{cases} \mathbb{R} & \to \mathbb{R} \\ x & \mapsto |x| \end{cases}$$

Sub-differential (sign)

$$\partial f(x^*) = \begin{cases} \{-1\} & \text{if } x^* \in]-\infty, 0[\\ \{1\} & \text{if } x^* \in]0, \infty[\\ [-1, 1] & \text{if } x^* = 0 \end{cases}$$

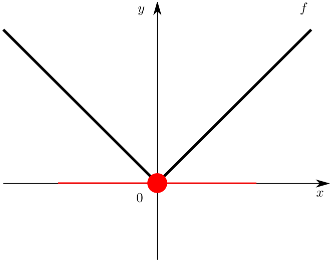




Sub-differential of the absolute value

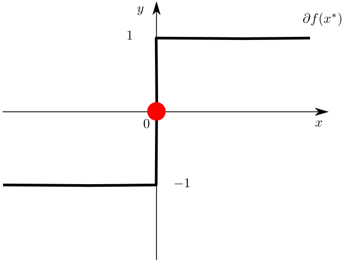
Function (abs):

$$f : \begin{cases} \mathbb{R} & \to \mathbb{R} \\ x & \mapsto |x| \end{cases}$$



Sub-differential (sign)

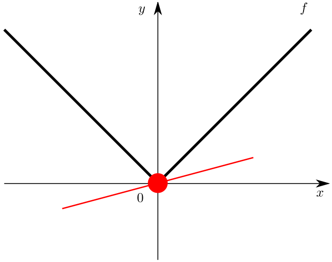
$$\partial f(x^*) = \begin{cases} \{-1\} & \text{if } x^* \in]-\infty, 0[\\ \{1\} & \text{if } x^* \in]0, \infty[\\ [-1, 1] & \text{if } x^* = 0 \end{cases}$$



Sub-differential of the absolute value

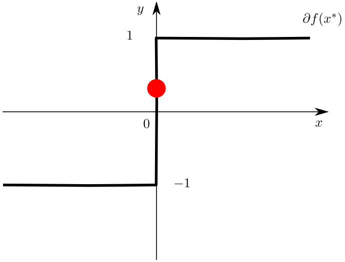
Function (abs):

$$f : \begin{cases} \mathbb{R} & \to \mathbb{R} \\ x & \mapsto |x| \end{cases}$$



Sub-differential (sign)

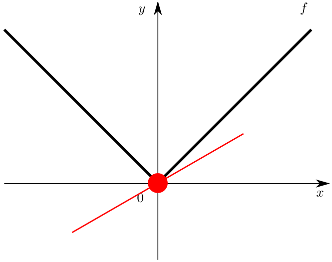
$$\partial f(x^*) = \begin{cases} \{-1\} & \text{if } x^* \in]-\infty, 0[\\ \{1\} & \text{if } x^* \in]0, \infty[\\ [-1, 1] & \text{if } x^* = 0 \end{cases}$$



Sub-differential of the absolute value

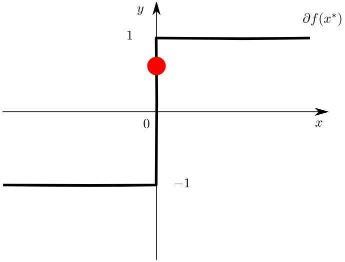
Function (abs):

$$f : \begin{cases} \mathbb{R} & \to \mathbb{R} \\ x & \mapsto |x| \end{cases}$$



Sub-differential (sign)

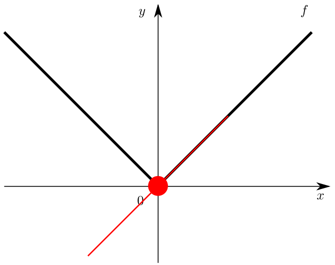
$$\partial f(x^*) = \begin{cases} \{-1\} & \text{if } x^* \in]-\infty, 0[\\ \{1\} & \text{if } x^* \in]0, \infty[\\ [-1, 1] & \text{if } x^* = 0 \end{cases}$$



Sub-differential of the absolute value

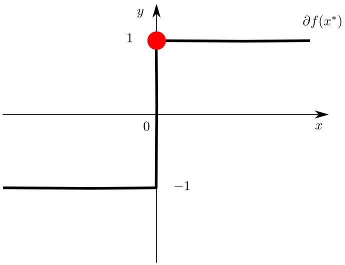
Function (abs):

$$f : \begin{cases} \mathbb{R} & \to \mathbb{R} \\ x & \mapsto |x| \end{cases}$$



Sub-differential (sign)

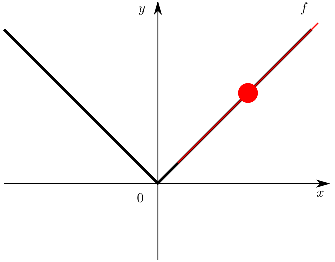
$$\partial f(x^*) = \begin{cases} \{-1\} & \text{if } x^* \in]-\infty, 0[\\ \{1\} & \text{if } x^* \in]0, \infty[\\ [-1, 1] & \text{if } x^* = 0 \end{cases}$$



Sub-differential of the absolute value

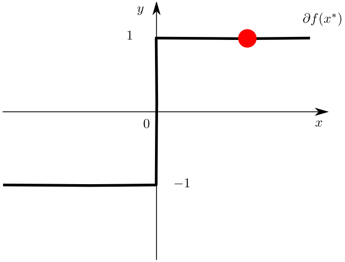
Function (abs):

$$f : \begin{cases} \mathbb{R} & \longrightarrow \mathbb{R} \\ x & \longmapsto |x| \end{cases}$$



Sub-differential (sign)

$$\partial f(x^*) = \begin{cases} \{-1\} & \text{if } x^* \in]-\infty, 0[\\ \{1\} & \text{if } x^* \in]0, \infty[\\ [-1, 1] & \text{if } x^* = 0 \end{cases}$$



The denoising case: $X = \mathrm{Id}_n$

Simple design: $n = p$ and $X = \mathrm{Id}_n$, meaning the atoms are canonical elements: $\mathbf{x}_j = (0, \cdots, 0, \underset{\overset{\uparrow}{j}}{1}, 0, \cdots, 1)^\top$, then

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

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

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

Rem: This is called the **proximal** operator of $\lambda \| \cdot \|_1$

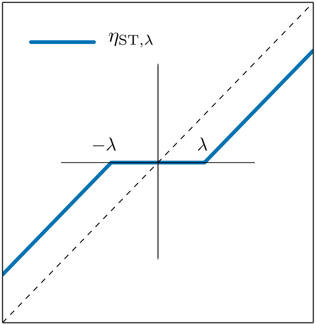
Soft-Thresholding

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

$$\eta_{\mathrm{ST},\lambda}(y) = \underset{\beta \in \mathbb{R}}{\arg\min} \left(\frac{(y-\beta)^2}{2} + \lambda|\beta| \right)$$
$$= \mathrm{sign}(y) \cdot (|y| - \lambda)_+$$

where $(\cdot)_+ = \max(0, \cdot)$

Proof: use sub-gradients of $|\cdot|$ and Fermat condition



Rem: systemetic underestimation / contraction bias; coefficients (greater than λ) are shrinked toward zero by a factor λ

Table of Contents

Dual problem Kim *et al.* (2007)

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

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

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

- $\Delta_X = \left\{\theta \in \mathbb{R}^n : \|X^\top \theta\|_\infty \leqslant 1\right\}$ is a polyhedron
- The dual solution is the **projection** of y/λ over this polyhedron:

$$\hat{\theta}^{(\lambda)} = \underset{\theta \in \Delta_X}{\arg\min} \left\|\frac{y}{\lambda} - \theta\right\|^2 := \Pi_{\Delta_X}\left(\frac{y}{\lambda}\right)$$

Proof in the next slide

Proof of the dual formulation

$$\min_{\beta \in \mathbb{R}^p} \underbrace{\frac{1}{2}\|y - X\beta\|^2}_{f(y-X\beta)} + \lambda \underbrace{\|\beta\|_1}_{\Omega(\beta)} \Leftrightarrow \min_{\beta \in \mathbb{R}^p, z \in \mathbb{R}^n} \begin{cases} f(z) + \lambda\Omega(\beta) \\ \text{s.t.} \quad z = y - X\beta \end{cases}$$

Lagrangian : $\quad \mathcal{L}(z, \beta, \theta) := \frac{1}{2}\|z\|^2 + \lambda\Omega(\beta) + \lambda\theta^\top(y - X\beta - z).$

Find a Lagrangian saddle point $(z^\star, \hat{\beta}^{(\lambda)}, \hat{\theta}^{(\lambda)})$ (Strong duality):

$$\min_{\beta \in \mathbb{R}^p, z \in \mathbb{R}^n} \max_{\theta \in \mathbb{R}^n} \mathcal{L}(z, \beta, \theta) = \max_{\theta \in \mathbb{R}^n} \min_{\beta \in \mathbb{R}^p, z \in \mathbb{R}^n} \mathcal{L}(z, \beta, \theta) =$$

$$\max_{\theta \in \mathbb{R}^n} \left\{ \min_{z \in \mathbb{R}^n} [f(z) - \lambda\theta^\top z] + \min_{\beta \in \mathbb{R}^p} [\lambda\Omega(\beta) - \lambda\theta^\top X\beta] + \lambda\theta^\top y \right\} =$$

$$\max_{\theta \in \mathbb{R}^n} \left\{ -f^*(\lambda\theta) - \lambda\Omega^*(X^\top\theta) + \lambda\theta^\top y \right\}$$

which is the formulation asserted (with conjugacy properties)

Conjugation

For any $f : \mathbb{R}^n \to \mathbb{R}$, the (Fenchel) conjugate f^* is defined as

$$\boxed{f^*(z) = \sup_{x \in \mathbb{R}^n} x^\top z - f(x)}$$

- If $f(\cdot) = \|\cdot\|^2/2$ then $f^*(\cdot) = f(\cdot)$
- If $f(\cdot) = \Omega(\cdot)$ is a norm, then $f^*(\cdot) = \iota_{\mathcal{B}_*(0,1)}(\cdot)$, *i.e.*, it is the indicator function of the dual norm unit ball, where the **dual norm** Ω^* is defined by:

$$\Omega^*(z) = \sup_{x : \Omega(x) \leqslant 1} x^\top z = \iota^*_{\mathcal{B}(0,1)}$$

and

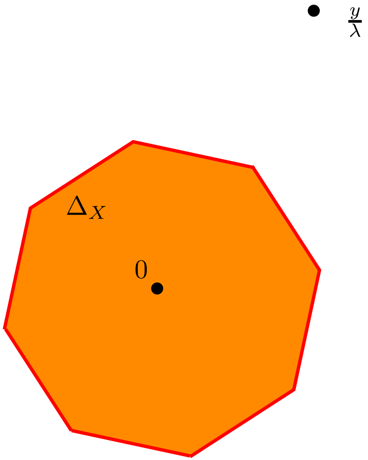
$$\iota_{\mathcal{B}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{B} \\ +\infty & \text{otherwise} \end{cases}, \text{ where } \mathcal{B} = \{x \in \mathbb{R}^n : \Omega(x) \leqslant 1\}$$

Geometric interpretation

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

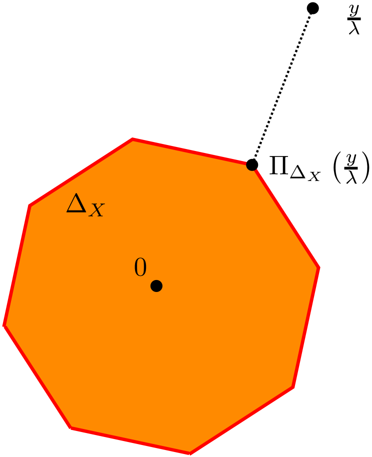
Geometric interpretation

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



Geometric interpretation

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



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

Mother of safe rules: Fermat's rule implies that
if $\lambda \geqslant \lambda_{\max} = \|X^\top y\|_\infty = \max_{j \in [p]} |\mathbf{x}_j^\top \hat{\theta}^{(\lambda)}|$, then $0 \in \mathbb{R}^p$ is the (unique here) primal solution

Proof in next slide

Proof Fermat/KKT + primal/dual link

Lagrangian : $\mathcal{L}(z, \beta, \theta) := \underbrace{\frac{1}{2}\|z\|^2}_{f(z)} + \lambda \underbrace{\|\beta\|_1}_{\Omega(\beta)} + \lambda\theta^\top (y - X\beta - z).$

A saddle point $(z^\star, \hat{\beta}^{(\lambda)}, \hat{\theta}^{(\lambda)})$ of the Lagrangian satisfies:

$$\begin{cases} 0 &= \frac{\partial \mathcal{L}}{\partial z}(z^\star, \hat{\beta}^{(\lambda)}, \hat{\theta}^{(\lambda)}) = \nabla f(z^\star) = z^\star - \lambda\hat{\theta}^{(\lambda)}, \\ 0 &\in \partial \mathcal{L}(z^\star, \cdot, \hat{\theta}^{(\lambda)})(\hat{\beta}^{(\lambda)}) = -\lambda X^\top \hat{\theta}^{(\lambda)} + \lambda \partial\Omega(\hat{\beta}^{(\lambda)}) \\ 0 &= \frac{\partial \mathcal{L}}{\partial \theta}(z^\star, \hat{\beta}^{(\lambda)}, \hat{\theta}^{(\lambda)}) = y - X\hat{\beta}^{(\lambda)} - z^\star. \end{cases}$$

Hence, $y - X\hat{\beta}^{(\lambda)} = z^\star = \lambda\hat{\theta}^{(\lambda)}$ and $X^\top\hat{\theta}^{(\lambda)} \in \partial\Omega(\hat{\beta}^{(\lambda)})$ so

$$\forall j \in \{1, \dots, p\}, \quad \mathbf{x}_j^\top \hat{\theta}^{(\lambda)} \in \partial\|\cdot\|_1(\hat{\beta}^{(\lambda)})$$

Geometric interpretation (II)

A simple dual point is: $y/\lambda_{\max} \in \Delta_X$ where $\lambda_{\max} = \|X^\top y\|_\infty$

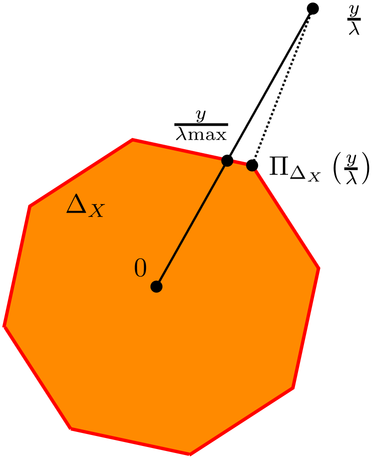


Table of Contents

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

Screening thanks to Fermat's Rule:
$$\boxed{\text{If } |\mathbf{x}_j^\top \hat{\theta}^{(\lambda)}| < 1 \text{ then, } \hat{\beta}_j^{(\lambda)} = 0}$$

Beware: $\hat{\theta}^{(\lambda)}$ is **unknown**, but one can consider a **safe region** $\mathcal{C} \subset \mathbb{R}^n$ containing $\hat{\theta}^{(\lambda)}$, *i.e.*, $\hat{\theta}^{(\lambda)} \in \mathcal{C}$, leading to :

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

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

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

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

so the safe rule becomes

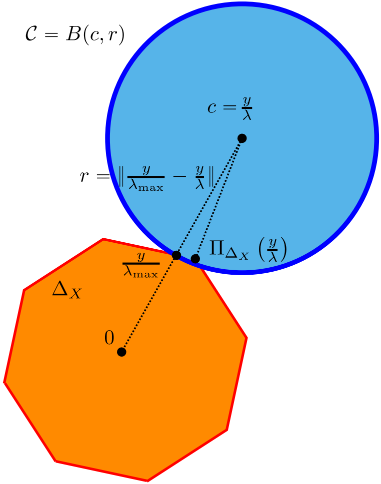
$$\boxed{\text{If } |\mathbf{x}_j^\top c| + r\|\mathbf{x}_j\| < 1 \text{ then } \hat{\beta}_j^{(\lambda)} = 0} \tag{1}$$

Screen-out the all variables $\mathbf{x}_j$ satisfying (1), and remove them from the optimization problem.

New objective:

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

Static safe rules: El Ghaoui *et al.* (2012)



Properties of static safe rules

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

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

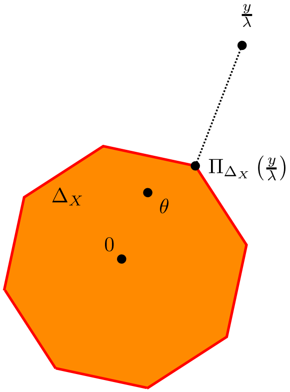
$$\boxed{\text{If } |\mathbf{x}_j^\top y| < \lambda(1 - \|y/\lambda_{\max} - y/\lambda\| \|\mathbf{x}_j\|) \text{ then } \hat{\beta}_j^{(\lambda)} = 0}$$

- Reinterprets screening methods for **variable selection**: "If $|\mathbf{x}_j^\top y|$ is small, discard $\mathbf{x}_j$" as a safe rule for the Lasso
- The corresponding safe test is **useless** as soon as:

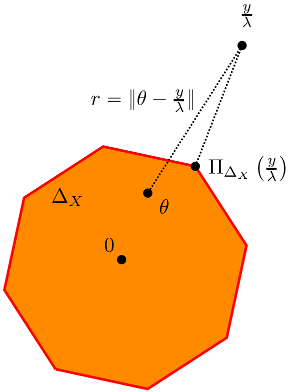
$$\frac{\lambda}{\lambda_{\max}} \leqslant \min_{j \in [p]} \left(\frac{1 + |\mathbf{x}_j^\top y| / (\|\mathbf{x}_j\| \|y\|)}{1 + \lambda_{\max} / (\|\mathbf{x}_j\| \|y\|)} \right)$$

meaning that no variable would be screened-out for such λ's

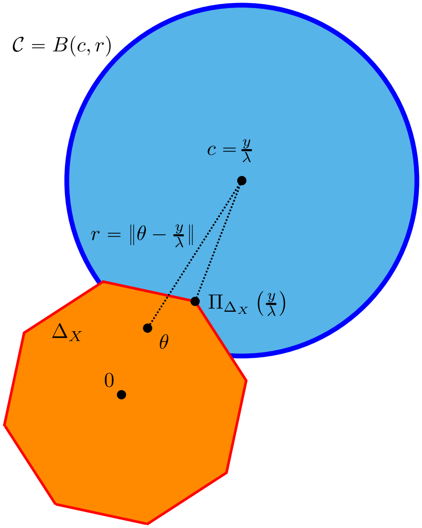
Dynamic safe rules Bonnefoy *et al.* (2014)



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



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



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)

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

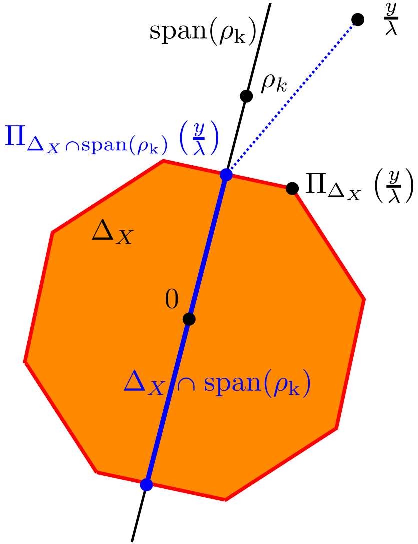
$$\text{Current } \textbf{residual} \text{ for primal point } \beta_k\text{:} \quad \rho_k = y - X\beta_k$$

Dual candidate: choose θ_k proportional to the residual

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

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

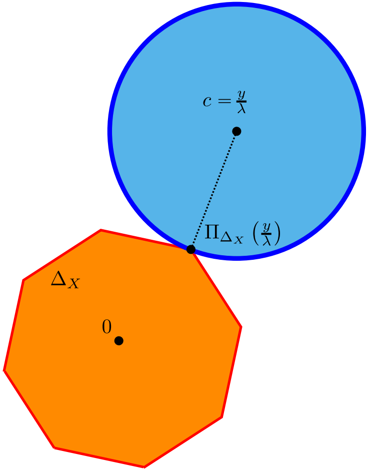
Creating dual points: project on a segment



Limits of previous dynamic rules

For $B(c, r) = B(\theta_k, r_k)$ with $r_k = \|\theta_k - y/\lambda\|$, the radius does not converge to zero, even when $\beta_k \to \hat{\beta}^{(\lambda)}$ and $\theta_k \to \hat{\theta}^{(\lambda)}$ (converging solver). The limiting safe sphere is

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



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

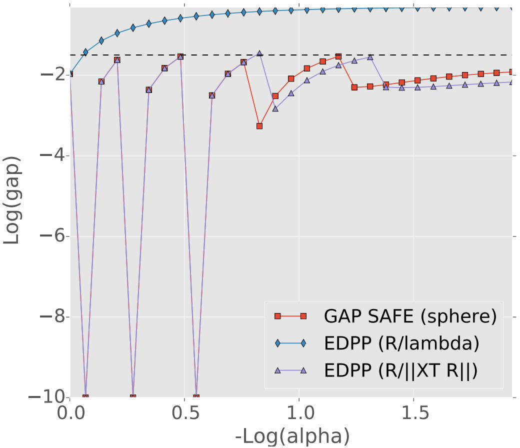
Warm start main idea: to compute the Lasso for T different λ's, say $\lambda_0, \cdots, \lambda_{T-1}$, reuse computation done at λ_{t-1} to get $\hat{\beta}^{(\lambda_t)}$:

- **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**!

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

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



Duality Gap properties

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

Duality gap: for any $\beta \in \mathbb{R}^p, \theta \in \Delta_X$, $G_\lambda(\beta, \theta) = P_\lambda(\beta) - D_\lambda(\theta)$

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

Strong duality: for any $\beta \in \mathbb{R}^p, \theta \in \Delta_X$,

$$D_\lambda(\theta) \leqslant D_\lambda(\hat{\theta}^{(\lambda)}) = P_\lambda(\hat{\beta}^{(\lambda)}) \leqslant P_\lambda(\beta)$$

Consequences:

- $G_\lambda(\beta, \theta) \geqslant 0$, for any $\beta \in \mathbb{R}^p, \theta \in \Delta_X$ (**weak duality**)
- $G_\lambda(\beta, \theta) \leqslant \epsilon \Rightarrow P_\lambda(\beta) - P_\lambda(\hat{\beta}^{(\lambda)}) \leqslant \epsilon$ (stopping criterion!)

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

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 a converging safe rule!

Proof in next slide (if any interest)

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 $\theta_1, \theta_2 \in \mathbb{R}^n$,

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

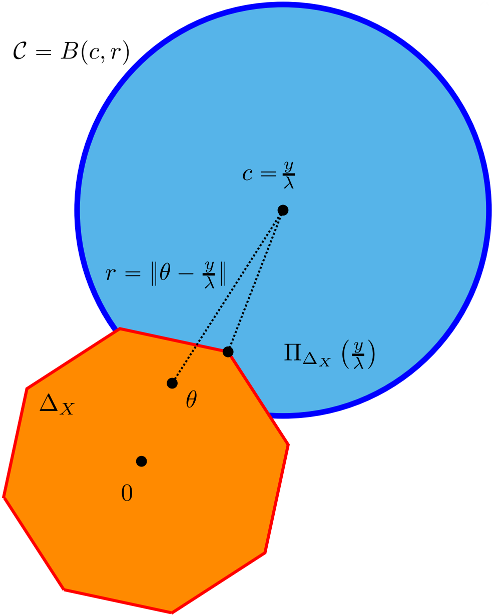
- $\hat{\theta}^{(\lambda)}$ maximizes D_λ over Δ_X, so Fermat's rule yields

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

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

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

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



Dynamic safe sphere Fercoq *et al.* (2015)

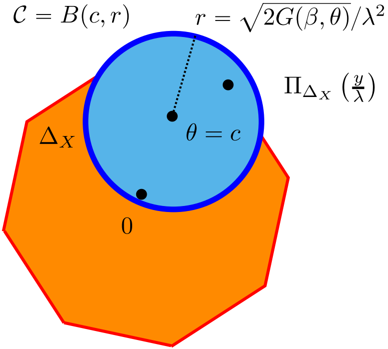


Table of Contents

Recap for safe spheres

$\mathcal{C}_k = B(\theta_k, r_\lambda(\beta_k, \theta_k))$ where β_k and θ_k are the current approximation of the primal and dual solution β_k and θ_k

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

where $\mu_{\mathcal{C}_k}(\mathbf{x}) := \sup_{\theta \in \mathcal{C}_k} |\mathbf{x}^\top \theta| = |\mathbf{x}^\top \theta_k| + r_\lambda(\beta_k, \theta_k)\|\mathbf{x}\|$

Rem: The active set is guaranteed to contain the variable that are in the support of an optimal solution

Algorithm 1 Coordinate descent (Lasso)

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 $k \mod F = 0$ **then** ▷ Screen every F epoch
            Construct $\theta \in \Delta_X$
            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 $j \in [p]$ **do** ▷ Soft-Threshold coordinates
            $\beta_j \leftarrow \mathrm{ST}\big(\frac{\lambda_t}{\|\mathbf{x}_j\|^2}, \beta_j - \frac{\mathbf{x}_j^\top (X\beta - y)}{\|\mathbf{x}_j\|^2}\big)$
        end for
    end for
end for

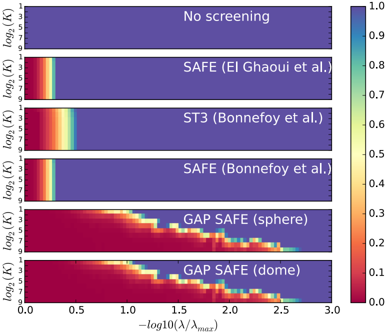
Algorithm 2 Coordinate descent (Lasso) with GAP Safe screening

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 $k \mod F = 0$ **then** ▷ Screen every F epoch
            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**



Proportion of active variables as a function of λ and the number of iterations K on the Leukemia dataset $(n = 72, p = 7129)$

Computing time

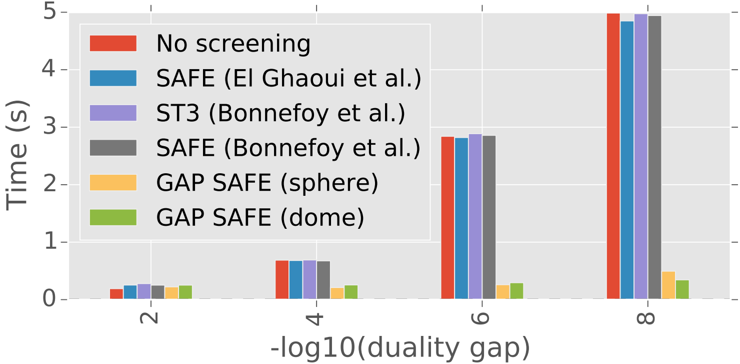


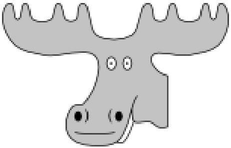
Figure: Time to reach convergence using various screening rules on the Leukemia dataset (dense data: $n = 72, p = 7129$).

Conclusion and future work

- New safe screening rule based on duality gap for the Lasso
- Convergent safe regions (support identification in finite time)
- Improved computational efficiency for Coordinate Descent
- Other regularization can be simply handled: Elastic Net, Group-Lasso
- Other data fitting term: logistic regression for classification (f smooth: gradient Lipschitz)
- On going work: Sparse Group-Lasso ($\ell_1 + \ell_1/\ell_2$) more intricate

More info

- "Mind the duality gap: safer rules for the Lasso"
  Fercoq, Gramfort and S., ICML 2015
- "GAP Safe screening rules for sparse multi-task and multi-class models"
  Ndiaye, Fercoq, Gramfort and S., NIPS 2015
- Python Code on demand (soon available in **scikit-learn** Pedregosa *et al.* (2011))



Powered with **MooseTeX**

Références I

- A. Belloni, V. Chernozhukov, and L. Wang.
  Square-root Lasso: Pivotal recovery of sparse signals via conic programming.
  Biometrika, 98(4):791–806, 2011.
- A. Bonnefoy, V. Emiya, L. Ralaivola, and R. Gribonval.
  Dynamic Screening: Accelerating First-Order Algorithms for the Lasso and Group-Lasso.
  ArXiv e-prints, 2014.
- A. Bonnefoy, V. Emiya, L. Ralaivola, and R. Gribonval.
  A dynamic screening principle for the lasso.
  In *EUSIPCO*, 2014.
- P. J. Bickel, Y. Ritov, and A. B. Tsybakov.
  Simultaneous analysis of Lasso and Dantzig selector.
  Ann. Statist., 37(4):1705–1732, 2009.
- A. Beck and M. Teboulle.
  A fast iterative shrinkage-thresholding algorithm for linear inverse problems.
  SIAM J. Imaging Sci., 2(1):183–202, 2009.

Références II

- S. S. Chen, D. L. Donoho, and M. A. Saunders.
  Atomic decomposition by basis pursuit.
  SIAM J. Sci. Comput., 20(1):33–61 (electronic), 1998.

- E. J. Candès, M. B. Wakin, and S. P. Boyd.
  Enhancing sparsity by reweighted l_1 minimization.
  J. Fourier Anal. Applicat., 14(5-6):877–905, 2008.

- B. Efron, T. Hastie, I. M. Johnstone, and R. Tibshirani.
  Least angle regression.
  Ann. Statist., 32(2):407–499, 2004.
  With discussion, and a rejoinder by the authors.

- L. El Ghaoui, V. Viallon, and T. Rabbani.
  Safe feature elimination in sparse supervised learning.
  J. Pacific Optim., 8(4):667–698, 2012.

- O. Fercoq, A. Gramfort, and J. Salmon.
  Mind the duality gap: safer rules for the lasso.
  In *ICML*, 2015.

Références III

- J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani.
  Pathwise coordinate optimization.
  Ann. Appl. Stat., 1(2):302–332, 2007.

- S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky.
  An interior-point method for large-scale l_1-regularized least squares.
  IEEE J. Sel. Topics Signal Process., 1(4):606–617, 2007.

- J. Mairal and B. Yu.
  Complexity analysis of the lasso regularization path.
  In *ICML*, 2012.

- M. R. Osborne, B. Presnell, and B. A. Turlach.
  A new approach to variable selection in least squares problems.
  IMA J. Numer. Anal., 20(3):389–403, 2000.

- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay.
  Scikit-learn: Machine learning in Python.
  J. Mach. Learn. Res., 12:2825–2830, 2011.

Références IV

- R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani.
  Strong rules for discarding predictors in lasso-type problems.
  J. Roy. Statist. Soc. Ser. B, 74(2):245–266, 2012.

- R. Tibshirani.
  Regression shrinkage and selection via the lasso.
  J. Roy. Statist. Soc. Ser. B, 58(1):267–288, 1996.

- R. J. Tibshirani.
  The lasso problem and uniqueness.
  Electron. J. Stat., 7:1456–1490, 2013.

- J. Wang, J. Zhou, P. Wonka, and J. Ye.
  Lasso screening rules via dual polytope projection.
  In *NIPS*, pages 1070–1078, 2013.

- H. Zou.
  The adaptive lasso and its oracle properties.
  J. Am. Statist. Assoc., 101(476):1418–1429, 2006.

Références V

- C.-H. Zhang and T. Zhang.

  A general theory of concave regularization for high-dimensional sparse estimation problems.

  Statistical Science, 27(4):576–593, 2012.