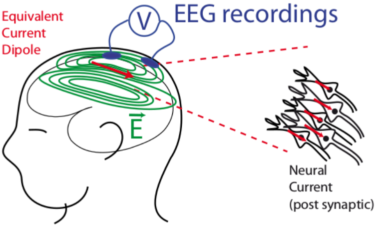
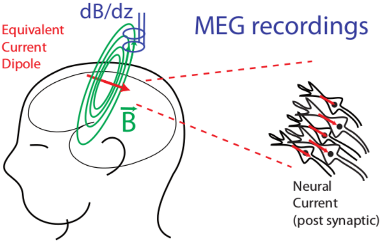
On high dimensional regression: computational and statistical perspectives

Joseph Salmon

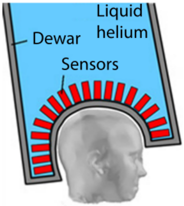
"One" motivation: M/EEG inverse problem

- sensors: magneto- and electro-encephalogram measurements
- sources: brain locations





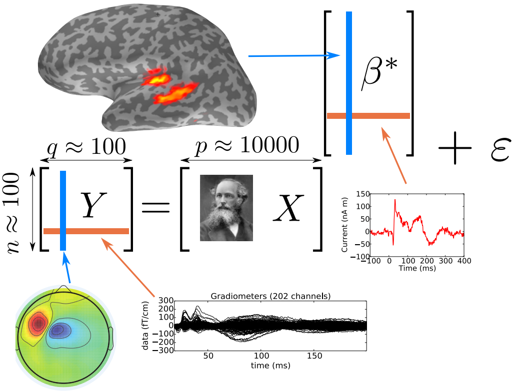
First EEG recordings in 1929 by H. Berger



Hôpital La Timone Marseille, France

(Tribute to **A. Gramfort**)

The M/EEG inverse problem



Dimensions involved:

- n: number of sensors
- q: number of time instants
- p: number of vertices, mesh discretization (features)
- $Y \in \mathbb{R}^{n \times q}$: measurements matrix
- $X = [\mathbf{x}_1, \ldots, \mathbf{x}_p] \in \mathbb{R}^{n \times p}$: matrix describing physical models
- $\beta^* \in \mathbb{R}^{p \times q}$: source activity matrix
- $\varepsilon \in \mathbb{R}^{n \times q}$ additive white noise

The M/EEG inverse problem

Challenging ill-posed problem with particular “constraints”:

- multi-task problem
- regularization must handle specific structures
- heteroscedastic noise (2/3 types of sensors)
- signal might be complex (not only real)
- ...

“La (vraie) vie, c’est pas du gâteau”, let us focus on the simplest sparse linear regression model:

$$\boxed{y = X\beta^* + \varepsilon}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $\|\beta^*\|_0 \ll \min(p, n)$, ($q = 1$ time/task)

The M/EEG inverse problem

Challenging ill-posed problem with particular “constraints”:

- multi-task problem
- regularization must handle specific structures
- heteroscedastic noise (2/3 types of sensors)
- signal might be complex (not only real)
- ...

“La (vraie) vie, c’est pas du gâteau”, let us focus on the simplest sparse linear regression model:

$$\boxed{y = X\beta^* + \varepsilon}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $\|\beta^*\|_0 \ll \min(p, n)$, ($q = 1$ time/task)

The Lasso and variations

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

- Statistics: **Lasso** Tibshirani (1996)
- Signal processing variant: **Basis Pursuit** Chen *et al.* (1998)
- Geophysics: Taylor *et al.* (1979) "Deconvolution with the ℓ_1 norm"

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

- parameter $\lambda > 0$ controlling sparsity/data-fitting trade-off

Rem: usual column normalization $\|\mathbf{x}_j\|^2 = n$ or $\|\mathbf{x}_j\| = 1$

Rem: convex problem, possibly non-uniqueness Tibshirani (2013)

Theory is now (fairly) well understood

Theorem Bickel *et al.* (2009), Dalalyan *et al.* (2017), Giraud (2014)

For Gaussian noise model with X satisfying the "Restricted Eigenvalue" property and $\lambda = 2n\sigma\sqrt{\frac{2\log(p/\delta)}{n}}$, then

$$\frac{1}{n}\|X(\beta^* - \hat{\beta}^{(\lambda)})\|^2 \leq \frac{18}{\kappa_s^2}\frac{\sigma^2 s}{n}\log\left(\frac{p}{\delta}\right)$$

with probability $1-\delta$, where $\hat{\beta}^{(\lambda)}$ is a Lasso solution and $s = \|\beta^*\|_0$

<u>Rem</u>: under the "Restricted Eigenvalue" property, κ_s^2 is a measure of strong convexity of the (quadratic part of the) objective function obtained when extracting s columns of X

Table of Contents

The Lasso: algorithmic point of view

Commonly used algorithms for solving this **convex** program:

- Homotopy method - LARS:
  efficient for small p Osborne *et al.* (2000), Efron *et al.* (2004) and full path (*i.e.*, compute solution for "all" λ's). Specific to Lasso (not flexible) + instabilities Mairal and Yu (2012)

- (F)ISTA, Forward - Backward, proximal(s) algorithm:
  useful in signal processing where $r \to X^\top r$ is cheap to compute (*e.g.*, FFT, Fast Wavelet Transform, etc.) Daubechies *et al.* (2004), Beck and Teboulle (2009), Combettes and Pesquet (2011)

The Lasso: algorithmic point of view

Commonly used algorithms for solving this **convex** program:

- Homotopy method - LARS:
  efficient for small p Osborne *et al.* (2000), Efron *et al.* (2004) and full path (*i.e.*, compute solution for "all" λ's). Specific to Lasso (not flexible) + instabilities Mairal and Yu (2012)

- (F)ISTA, Forward - Backward, proximal(s) algorithm:
  useful in signal processing where $r \to X^\top r$ is cheap to compute (*e.g.*, FFT, Fast Wavelet Transform, etc.) Daubechies *et al.* (2004), Beck and Teboulle (2009), Combettes and Pesquet (2011)

- Coordinate descent (CD):
  useful for large p and unstructured and/or sparse matrix X, *e.g.*, for text encoding Friedman *et al.* (2007)

The Lasso: algorithmic point of view

Commonly used algorithms for solving this **convex** program:

- Homotopy method - LARS:
  efficient for small p Osborne *et al.* (2000), Efron *et al.* (2004) and full path (*i.e.*, compute solution for "all" λ's). Specific to Lasso (not flexible) + instabilities Mairal and Yu (2012)

- (F)ISTA, Forward - Backward, proximal(s) algorithm:
  useful in signal processing where $r \to X^\top r$ is cheap to compute (*e.g.*, FFT, Fast Wavelet Transform, etc.) Daubechies *et al.* (2004), Beck and Teboulle (2009), Combettes and Pesquet (2011)

- Coordinate descent (CD):
  useful for large p and unstructured and/or sparse matrix X, *e.g.*, for text encoding Friedman *et al.* (2007)

Objective: speed-up Lasso solvers with screening

- **Sequential constraint**: compute Lassos for $\lambda_0 > \cdots > \lambda_{T-1}$, possibly for many T's, *i.e.*, get $\hat{\beta}^{(\lambda_0)}, \ldots, \hat{\beta}^{(\lambda_{T-1})}$ (*e.g.*, for choosing the best by Cross-Validation)

  Rem: standard grid is geometric from $\lambda_{\max} := \|X^\top y\|_\infty$ to $\lambda_{\min} = \alpha\lambda_{\max}$ (default in R-`glmnet` / Python-`sklearn`: $T = 100, \alpha = 0.001$)

- **Flexible schemes**: adapt to most iterative solvers and various Lasso-type problems, in particular for (block) coordinate descent (well suited for screening)

Dual problem Kim *et al.* (2007)

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

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

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

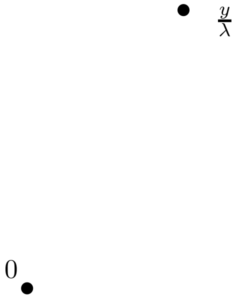
is a closed convex set (finite intersection of half-spaces)

- The (unique) dual solution is the **projection** of y/λ over Δ_X:

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

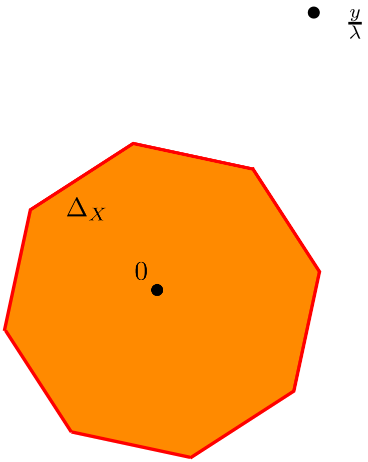
Geometric interpretation

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



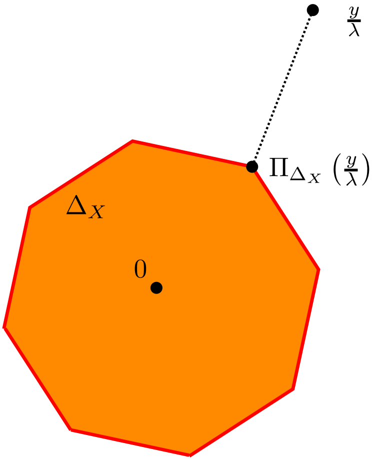
Geometric interpretation

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



Geometric interpretation

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



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

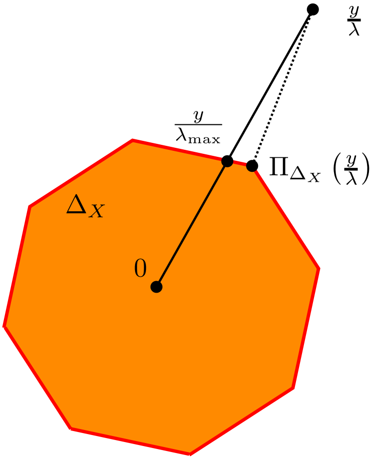
Necessary and sufficient optimality conditions:

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 yields
$\lambda \geq \lambda_{\max} = \|X^\top y\|_\infty = \max_{j \in [p]} |\mathbf{x}_j^\top \hat{\theta}^{(\lambda)}| \Rightarrow \hat{\beta}^{(\lambda)} = 0 \in \mathbb{R}^p$

Geometric interpretation (II)

A simple dual (feasible) point: $\frac{y}{\lambda_{\max}} \in \Delta_X$ where $\lambda_{\max} = \|X^\top y\|_\infty$



Rem: $(y - X \cdot 0)/\lambda \in \Delta_X$ if $\lambda \geq \lambda_{\max}$, hence $\hat{\theta}^{(\lambda)} = y/\lambda, \hat{\beta}^{(\lambda)} = 0$

Table of Contents

Safe screening rules contributions

Joint work with **A. Gramfort**, **O. Fercoq**, **E. Ndiaye**, **V. Leclère**
Fercoq *et al.* (2015), Ndiaye *et al.* (2015), Ndiaye *et al.* (2016), Ndiaye *et al.* (2016), Ndiaye *et al.* (2017)

Safe rules El Ghaoui *et al.* (2012)

Motivation: leverage targeted sparsity to reduce computation

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

BUT: $\hat{\theta}^{(\lambda)}$ is intrinsically **unknown**, not practical

Yet, having a **safe region** $\mathcal{C} \subset \mathbb{R}^n$ containing $\hat{\theta}^{(\lambda)}$, *i.e.*, $\hat{\theta}^{(\lambda)} \in \mathcal{C}$:

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

Remove from the optimization problem the $\mathbf{x}_j$'s satisfying the test!

Goal: find a safe region $\mathcal{C}$

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

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

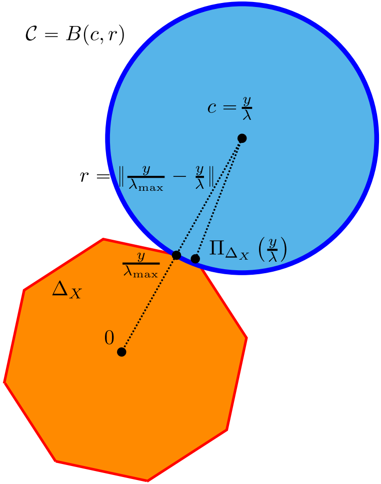
Screening cost for one feature: one dot product (of size n)

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

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)



Static safe rules/variable selection

Static safe region: useful prior any optimization

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

Interpretation: **static rules** = statistical (correlation) "screening":

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

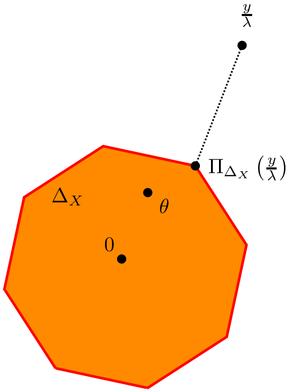
$$\iff (\text{for normalized } \mathbf{x}_j' s)$$

$$\boxed{\text{If } |\mathbf{x}_j^\top y| < C_{X,y} \text{ then } \hat{\beta}_j^{(\lambda)} = 0} \quad (\text{and } \mathbf{x}_j \text{ can be discarded})$$

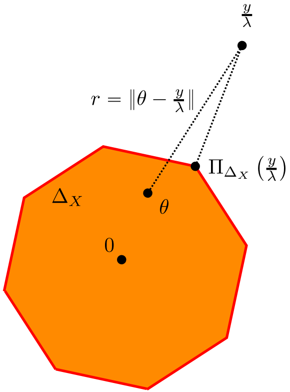
Rem: the corresponding safe test becomes **useless** when

$$\frac{\lambda}{\lambda_{\max}} \leq C'_{X,y} = \min_{j \in [p]} \left(\frac{1 + |\mathbf{x}_j^\top y| / (\|\mathbf{x}_j\| \|y\|)}{1 + \lambda_{\max} / (\|\mathbf{x}_j\| \|y\|)} \right)$$

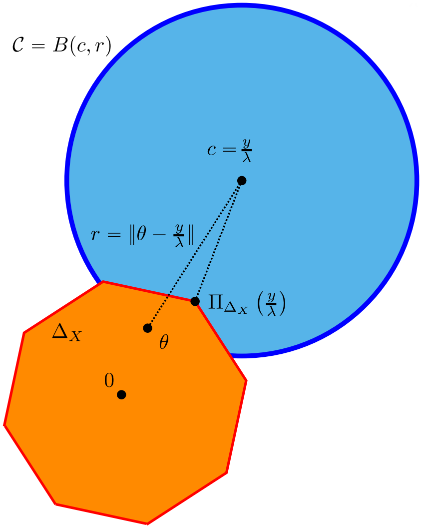
Dynamic safe rules Bonnefoy *et al.* (2014)



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



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



Dynamic safe rule

Dynamic rules: build iteratively $\theta_k \in \Delta_X$, as the solver proceeds to get refined safe regions and improve screening Bonnefoy *et al.* (2014, 2015)

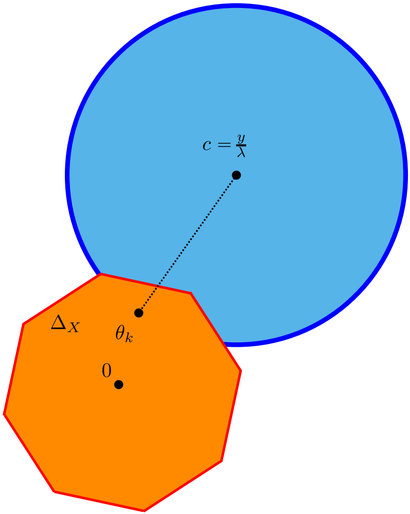
Primal-dual link at optimality : $\lambda\hat{\theta}^{(\lambda)} = y - X\hat{\beta}^{(\lambda)}$
Current **residual** for primal point β_k : $\rho_k = y - X\beta_k$

Dual candidate: choose θ_k as (rescaled) residual

$$\textit{e.g.},\ \theta_k = \frac{\rho_k}{\lambda \vee \|X^\top \rho_k\|_\infty} \in \Delta_X$$

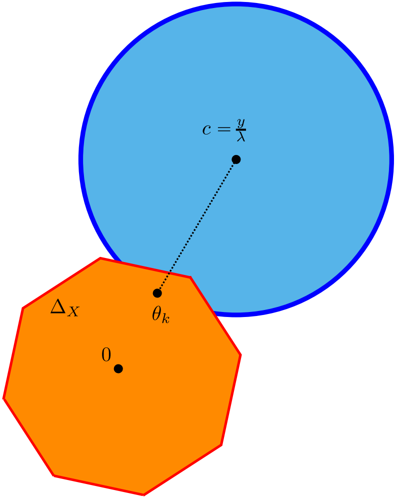
Limits of previous dynamic rules

If $B(c, r) = B(\theta_k, r_k)$ with $r_k = \|\theta_k - y/\lambda\|$, r_k does not converge to zero, even if $\beta_k \to \hat{\beta}^{(\lambda)}$ and $\theta_k \to \hat{\theta}^{(\lambda)}$. Limiting safe sphere:



Limits of previous dynamic rules

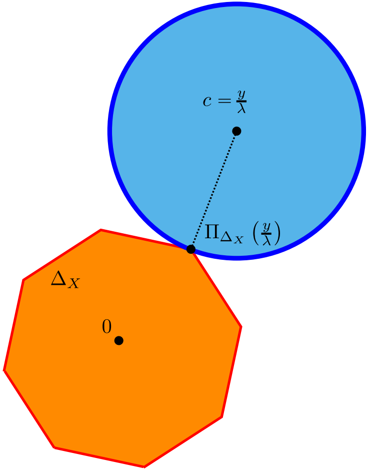
If $B(c, r) = B(\theta_k, r_k)$ with $r_k = \|\theta_k - y/\lambda\|$, r_k does not converge to zero, even if $\beta_k \to \hat{\beta}^{(\lambda)}$ and $\theta_k \to \hat{\theta}^{(\lambda)}$. Limiting safe sphere:



Limits of previous dynamic rules

If $B(c, r) = B(\theta_k, r_k)$ with $r_k = \|\theta_k - y/\lambda\|$, r_k does not converge to zero, even if $\beta_k \to \hat{\beta}^{(\lambda)}$ and $\theta_k \to \hat{\theta}^{(\lambda)}$. Limiting safe sphere:

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



Duality Gap properties

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

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

Consequences:

- $G_\lambda(\beta,\theta) \geq 0$, for any $\beta \in \mathbb{R}^p, \theta \in \Delta_X$
- $G_\lambda(\beta,\theta) \leq \epsilon \Rightarrow P_\lambda(\beta) - P_\lambda(\hat{\beta}^{(\lambda)}) \leq \epsilon$ (stopping criterion)

Gap Safe sphere

For any $\beta \in \mathbb{R}^p, \theta \in \Delta_X, \quad G_\lambda(\beta, \theta) = P_\lambda(\beta) - D_\lambda(\theta)$

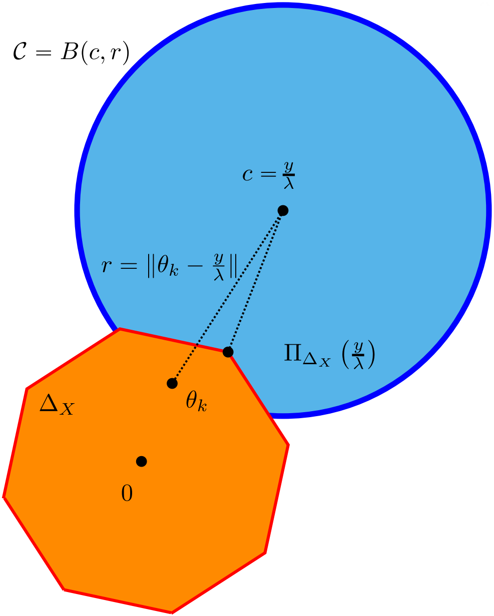
Gap Safe ball: $\boxed{B(\theta, r_\lambda(\beta, \theta)), \text{ where } r_\lambda(\beta, \theta) = \sqrt{2G_\lambda(\beta, \theta)}/\lambda}$

Rem: if $\beta_k \to \hat{\beta}^{(\lambda)}$ and $\theta_k \to \hat{\theta}^{(\lambda)}$ (*e.g.*, by residual scaling) then $G_\lambda(\beta_k, \theta_k) \to 0$: converging solvers lead to a converging safe rules, *i.e.*, the limiting safe spheres converge to $\{\hat{\theta}^{(\lambda)}\}$

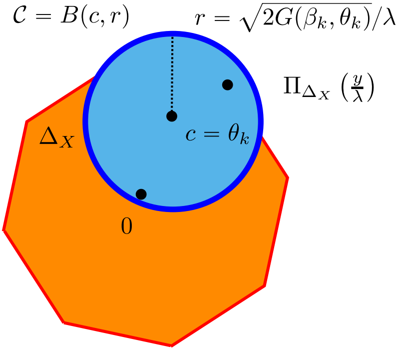
Rem: at optimality, one recovers the **equicorrelation set**

$$E = \left\{ j \in [\![p]\!] : \left| \frac{\mathbf{x}_j^\top (y - X\beta^{(\lambda)})}{\lambda} \right| = 1 \right\} \supseteq \operatorname{supp}(\beta^{(\lambda)})$$

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

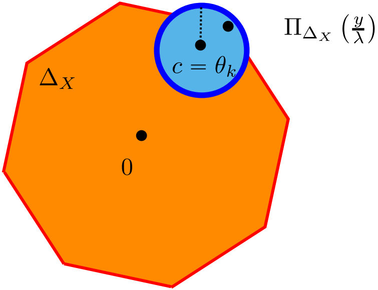


Dynamic safe sphere Fercoq *et al.* (2015)



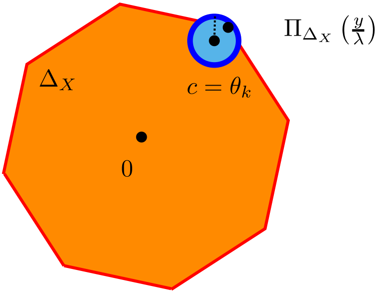
Dynamic safe sphere Fercoq *et al.* (2015)

$\mathcal{C} = B(c, r) \qquad r = \sqrt{2G(\beta_k, \theta_k)}/\lambda$



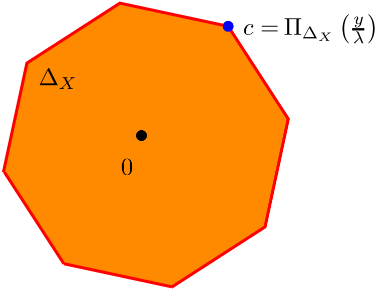
Dynamic safe sphere Fercoq *et al.* (2015)

$\mathcal{C} = B(c, r)$ $\qquad$ $r = \sqrt{2G(\beta_k, \theta_k)}/\lambda$



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

$\mathcal{C} = B(c, r)$ $\quad r = 0$



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

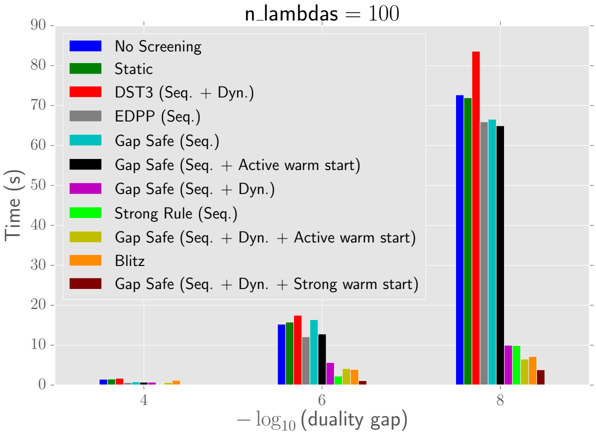
Warm start main idea: to compute the Lasso for T different λ's, say $\lambda_0, \ldots, \lambda_{T-1}$, re-use computation done at λ_{t-1} to get $\hat{\beta}^{(\lambda_t)}$

- Primal warm start: standard trick to accelerate iterative solvers: initialize with $\hat{\beta}^{(\lambda_{t-1})}$ to compute $\hat{\beta}^{(\lambda_t)}$
- Dual warm start: **sequential safe rules** use $\hat{\theta}^{(\lambda_{t-1})}$ to help screening for $\hat{\beta}^{(\lambda_t)}$

Major issue: in prior works $\hat{\theta}^{(\lambda_{t-1})}$ needed to be **known exactly**! Unrealistic except for $\lambda_{t-1} = \lambda_{\max}$, $\hat{\theta}^{(\lambda_0)} = \frac{y}{\lambda_{\max}} = \frac{y}{\|X^\top y\|_\infty}$

Gap safe rules are sequential: use $\tilde{\theta}^{(\lambda_{t-1})} \approx \hat{\theta}^{(\lambda_{t-1})}$ (dual feasible)

Computing time for standard grid with $T = 100$



Time to reach convergence (Leukemia dataset: $n = 72, p = 7129$) for 100 values on a standard grid of λ's

Table of Contents

Active set: aggressive screening

Joint work with **M. Massias**, **A. Gramfort**, Massias *et al.* (2017)

Gap safe rules revisited

Theorem

If $d_j(\theta) := \underbrace{\frac{1 - \left|\mathbf{x}_j^\top \theta\right|}{\|\mathbf{x}_j\|}}_{\text{test statistic}} > \underbrace{\sqrt{\frac{2}{\lambda^2} G_\lambda(\beta, \theta)}}_{\text{threshold}}$ then $\hat{\beta}_j = 0$

for any primal point β and any dual feasible point $\theta \in \Delta_X$

reminder: $G_\lambda(\beta, \theta) = P_\lambda(\beta) - D_\lambda(\theta)$ is the duality gap

Sequential safe rules - strong rules reminder

$$\boxed{d_j(\theta) := \frac{1 - \left|\mathbf{x}_j^\top \theta\right|}{\|\mathbf{x}_j\|}}$$

Assume $\hat{\beta}^{(\lambda')}, \hat{\theta}^{(\lambda')}$ approximated by $\tilde{\beta}^{(\lambda')}, \tilde{\theta}^{(\lambda')}$ and $\lambda' > \lambda$

Sequential safe rules (Wang *et al.* 2013): perform safe screening rule with the safe ball $\mathcal{C} = \mathcal{B}\left(\tilde{\theta}^{(\lambda')}, \left\|\frac{Y}{\lambda} - \frac{Y}{\lambda'}\right\|\right)$:

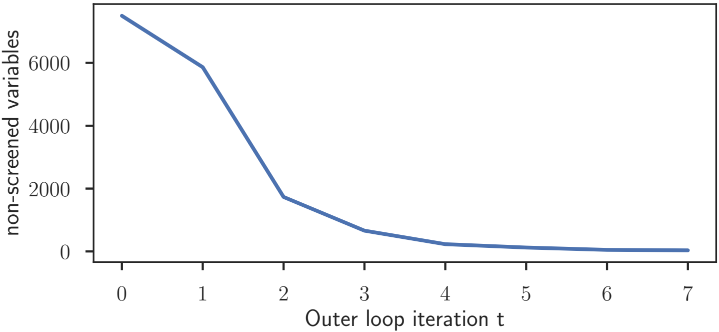
$$\text{If } d_j(\tilde{\theta}^{(\lambda')}) > \|Y\| \left|\frac{1}{\lambda} - \frac{1}{\lambda'}\right| \text{ then } \tilde{\beta}_j^{(\lambda)} = 0$$

Strong rules (Tibshirani *et al.* 2012): relax test to

$$\text{If } d_j(\tilde{\theta}^{(\lambda')}) > \frac{2}{\|\mathbf{x}_j\|} \frac{|\lambda' - \lambda|}{\lambda'} \text{ then } \tilde{\beta}_j^{(\lambda)} = 0$$

Rem: "wrong" screening is possible, post-processing needed

Gap Safe rules: from many to few features



(Leukemia $n = 72, p = 7129$)

<u>Drawback</u>: all features included when starting

Aggressive screening and working sets (WS)

Idea: to reduce the number of features, **drop the safety**

Gap Safe: **exclude** feature $\mathbf{x}_j$ if $d_j(\theta^k) > \sqrt{\frac{2}{\lambda^2} G_\lambda(\beta^k, \theta^k)}$

Aggressive screening and working sets (WS)

Idea: to reduce the number of features, **drop the safety**

Gap Safe: **exclude** feature $\mathbf{x}_j$ if $d_j(\theta^k) > \sqrt{\frac{2}{\lambda^2} G_\lambda(\beta^k, \theta^k)}$

Aggressive Gap: **include** feature $\mathbf{x}_j$ if $d_j(\theta^k) < r\sqrt{\frac{2}{\lambda^2} G_\lambda(\beta^k, \theta^k)}$ for some $r \in [0, 1]$ to be chosen (Jonhson and Guestrin, 2015,16)

- outer loop: only include few features with the smallest $d_j(\theta^k)$

Aggressive screening and working sets (WS)

Idea: to reduce the number of features, **drop the safety**

Gap Safe: **exclude** feature $\mathbf{x}_j$ if $d_j(\theta^k) > \sqrt{\frac{2}{\lambda^2} G_\lambda(\beta^k, \theta^k)}$

Aggressive Gap: **include** feature $\mathbf{x}_j$ if $d_j(\theta^k) < r\sqrt{\frac{2}{\lambda^2} G_\lambda(\beta^k, \theta^k)}$
for some $r \in [0, 1]$ to be chosen (Jonhson and Guestrin, 2015,16)

- outer loop: only include few features with the smallest $d_j(\theta^k)$
- inner loop: solve subproblem keeping only these features (fast)

Aggressive screening and working sets (WS)

Idea: to reduce the number of features, **drop the safety**

Gap Safe: **exclude** feature $\mathbf{x}_j$ if $d_j(\theta^k) > \sqrt{\frac{2}{\lambda^2} G_\lambda(\beta^k, \theta^k)}$

Aggressive Gap: **include** feature $\mathbf{x}_j$ if $d_j(\theta^k) < r\sqrt{\frac{2}{\lambda^2} G_\lambda(\beta^k, \theta^k)}$
for some $r \in [0, 1]$ to be chosen (Jonhson and Guestrin, 2015,16)

- outer loop: only include few features with the smallest $d_j(\theta^k)$
- inner loop: solve subproblem keeping only these features (fast)
- repeat

Aggressive screening and working sets (WS)

Idea: to reduce the number of features, **drop the safety**

Gap Safe: **exclude** feature $\mathbf{x}_j$ if $d_j(\theta^k) > \sqrt{\frac{2}{\lambda^2} G_\lambda(\beta^k, \theta^k)}$

Aggressive Gap: **include** feature $\mathbf{x}_j$ if $d_j(\theta^k) < r\sqrt{\frac{2}{\lambda^2} G_\lambda(\beta^k, \theta^k)}$
for some $r \in [0, 1]$ to be chosen (Jonhson and Guestrin, 2015,16)

- outer loop: only include few features with the smallest $d_j(\theta^k)$
- inner loop: solve subproblem keeping only these features (fast)
- repeat

Aggressive screening and working sets (WS)

Idea: to reduce the number of features, **drop the safety**

Gap Safe: **exclude** feature $\mathbf{x}_j$ if $d_j(\theta^k) > \sqrt{\frac{2}{\lambda^2} G_\lambda(\beta^k, \theta^k)}$

Aggressive Gap: **include** feature $\mathbf{x}_j$ if $d_j(\theta^k) < r\sqrt{\frac{2}{\lambda^2} G_\lambda(\beta^k, \theta^k)}$
for some $r \in [0, 1]$ to be chosen (Jonhson and Guestrin, 2015,16)

- outer loop: only include few features with the smallest $d_j(\theta^k)$
- inner loop: solve subproblem keeping only these features (fast)
- repeat

Working/active set strategies for Lasso-type/SVM problems:
(Joachims 1998), (Roth *et al.* 2008), (Kim & Park, 2010), (Kowalski *et al.* 2011), etc.

AGGressive Gap Greedy with Gram (A5G)

Algorithm: A5G

input : X, y, λ

param: $\beta_0 = 0_{p,q}, \bar{\epsilon} = 10^{-6}, r \in]0, 1[$

```
// Outer loop:
```

for $k = 1, \dots, K$ **do**

  Compute dual point θ^k and dual gap g^k

  if $g^k \leq \bar{\epsilon}$ **then**

    Break

  for $j = 1, \dots, p$ **do**

    Compute $d_j^k = (1 - |\mathbf{x}_j^\top \theta^k|)/\|\mathbf{x}_j\|$

  $\mathcal{W}^k = \{j \in [p] : d_j^k < r\sqrt{2g^k}/\lambda \cup \{j : \beta_j^{k-1} \neq 0\}$

```
  // Inner loop:
```

  Approximately solve problem restricted to $\mathcal{W}^k$ and get β^k

return β^k

AGGressive Gap Greedy with Gram (A5G)

Algorithm: A5G

input : X, y, λ

param: $\beta_0 = 0_{p,q}, \bar{\epsilon} = 10^{-6}, \cancel{r \in]0,1[}, p_0 = 100$ (or other guess)

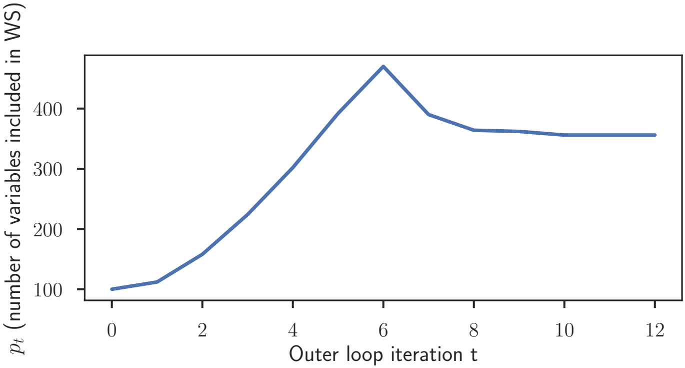
`// Outer loop:`

for $k = 1, \ldots, K$ **do**

- Compute dual point θ^k and dual gap g^k
- **if** $g^k \leq \bar{\epsilon}$ **then**
  - Break
- **for** $j = 1, \ldots, p$ **do**
  - Compute $d_j^k = (1 - |\mathbf{x}_j^\top \theta^k|)/\|\mathbf{x}_j\|$
- $p^k = \max(p_0, \min(2\|\beta^{k-1}\|_0, p))$ `// clipping`
- $\mathcal{W}^k = \{j \in [p] : d_j^k$ among $p^k/2$ smallest ones$\} \cup \{j : \beta_j^{k-1} \neq 0\}$
- `// Inner loop:`
- Approximately solve problem restricted to $\mathcal{W}^k$ and get β^k

return β^k

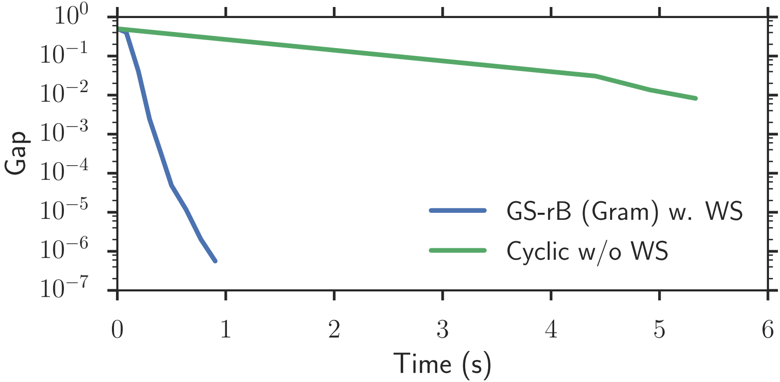
Limiting size of sub-problems solved



(Leukemia $n = 72, p = 7129$)

Smaller subproblems solved $\rightarrow$ Gram matrix $X_{\mathcal{W}^k}^\top X_{\mathcal{W}^k}$ fits in!
Rem: fast inner solver on sub-problems (*e.g.*, Greedy BCD)

Results on MEG data (Multi-task Lasso)



(MEG: $n = 302, p = 7498, q = 181$)

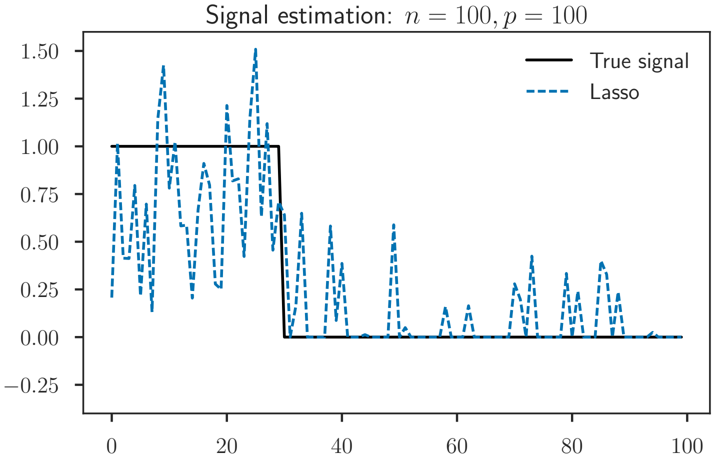
$10\times$ speed-up w.r.t. multi-task Lasso solver from `scikit-learn` (Pedregosa *et al.* 2011)

Table of Contents

Acknowledgments

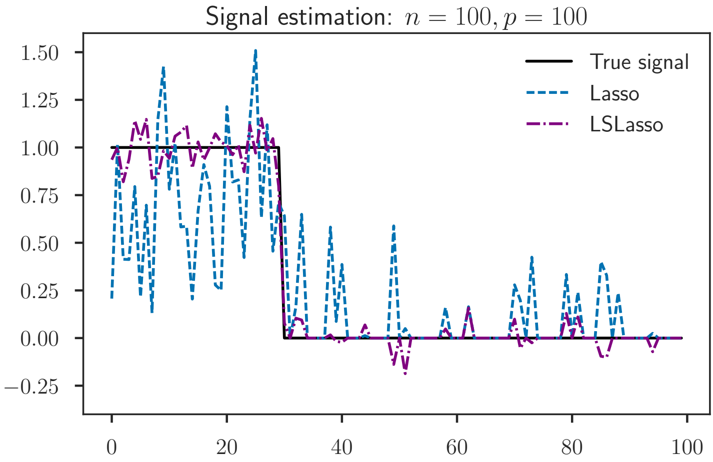
Joint work with **C.-A. Deledalle**, **N. Papadakis** and **S. Vaiter**
Deledalle *et al.* (2015), Deledalle *et al.* (2017)

Lasso and its bias



Gaussian random design (with $\rho = 0.5$)

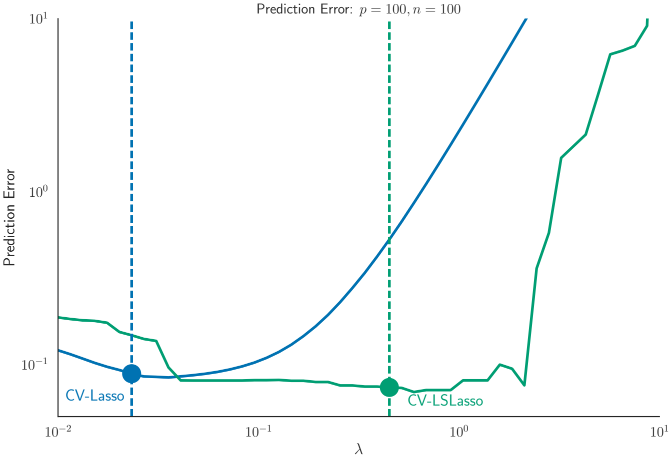
Lasso and its bias



Gaussian random design (with $\rho = 0.5$)
LSLasso: least-squares over estimated support

Interest for CV

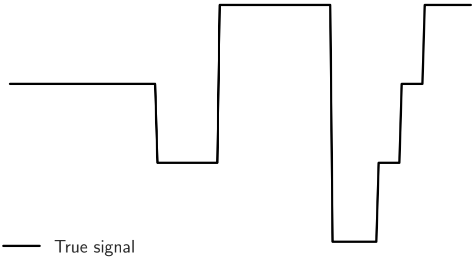
Potentially helps selecting a larger λ (sparser solutions), counter-act the "too small λ issue" of CV-Lasso for selection



(Anisotropic) Total Variation and its bias

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

$D^\top$: discrete gradient (incidence matrix over the path graph)
Rudin *et al.* (1992), Mammen and van de Geer (1997)



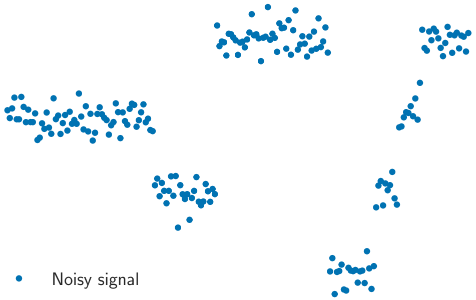
True signal

(Anisotropic) Total Variation and its bias

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

$D^\top$: discrete gradient (incidence matrix over the path graph)

Rudin *et al.* (1992), Mammen and van de Geer (1997)

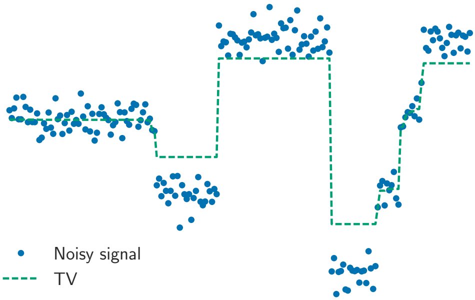


Noisy signal

(Anisotropic) Total Variation and its bias

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

$D^\top$: discrete gradient (incidence matrix over the path graph)
Rudin *et al.* (1992), Mammen and van de Geer (1997)



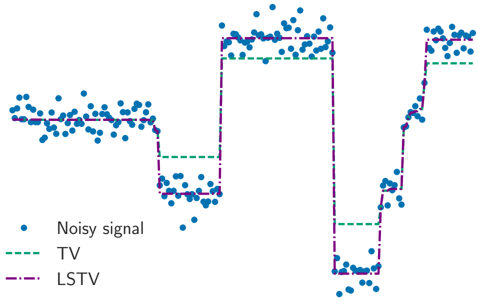
TV

(Anisotropic) Total Variation and its bias

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

$D^\top$: discrete gradient (incidence matrix over the path graph)

Rudin *et al.* (1992), Mammen and van de Geer (1997)

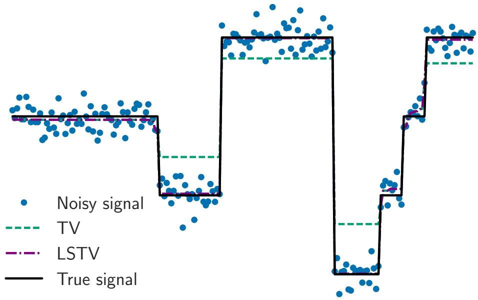


LSTV: Perform least-squares over estimated constant part

(Anisotropic) Total Variation and its bias

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

$D^\top$: discrete gradient (incidence matrix over the path graph)
Rudin *et al.* (1992), Mammen and van de Geer (1997)

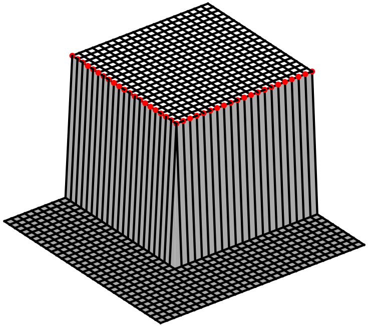


LSTV: Perform least-squares over estimated constant part

(Anisotropic) 2D-Total Variation and its bias

$$\hat{\beta}^{(\lambda)}_{AnisoTV} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \frac{1}{2} \|y - \beta\|^2 + \lambda \left\| D^\top \beta \right\|_1$$

$D^\top$: discrete gradient (incidence matrix over the path graph)

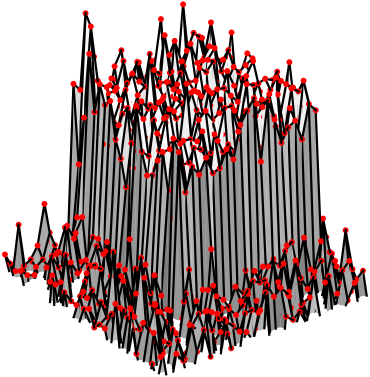


True signal

(Anisotropic) 2D-Total Variation and its bias

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

$D^\top$: discrete gradient (incidence matrix over the path graph)

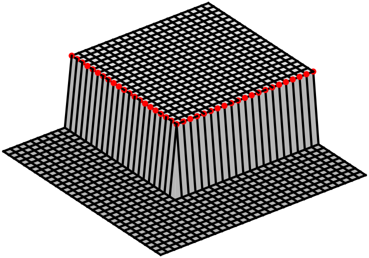


Noisy signal

(Anisotropic) 2D-Total Variation and its bias

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

$D^\top$: discrete gradient (incidence matrix over the path graph)

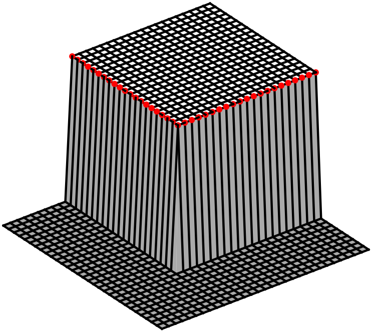


IsoTV

(Anisotropic) 2D-Total Variation and its bias

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

$D^\top$: discrete gradient (incidence matrix over the path graph)



LSIsoTV: Perform least-squares over estimated constant part

Invariant refitting strategy

Definition

The *invariant re-fitting* associated to an a.e. differentiable estimator $y \mapsto \hat{\beta}(y)$ is given for $y \in \mathbb{R}^n$ by

$$\mathcal{R}_{\hat{\beta}}^{\mathsf{inv}}(y) = \hat{\beta}(y) + J(XJ)^{+}(y - X\hat{\beta}(y)) \in \operatorname*{arg\,min}_{\beta \in \mathcal{M}_{\hat{\beta}}(y)} \|X\beta - y\|^2 \quad ,$$

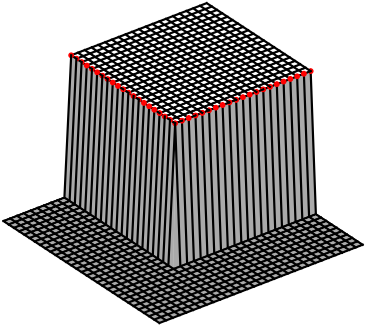
where $J = J_{\hat{\beta}}(y)$ is the Jacobian matrix of $\hat{\beta}$ at y, and the model (affine) space is $\mathcal{M}_{\hat{\beta}}(y) = y + \operatorname{Im}\left[J_{\hat{\beta}}(y)\right]$.

Motivation: extend least-squares refitting from the Lasso case

(Isotropic) 2D-Total Variation and its bias

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

$D^\top$: discrete gradient (incidence matrix over the path graph)

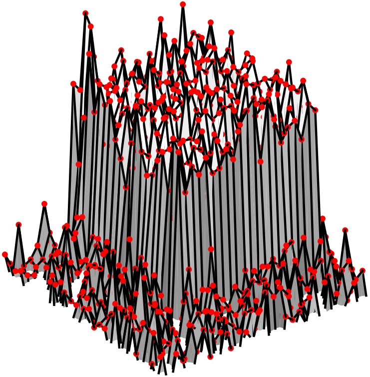


True signal

(Isotropic) 2D-Total Variation and its bias

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

$D^\top$: discrete gradient (incidence matrix over the path graph)

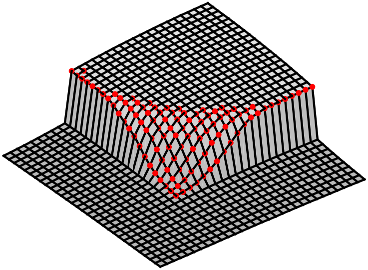


Noisy signal

(Isotropic) 2D-Total Variation and its bias

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

$D^\top$: discrete gradient (incidence matrix over the path graph)

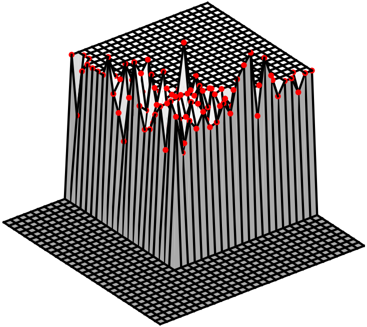


IsoTV

(Isotropic) 2D-Total Variation and its bias

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

$D^\top$: discrete gradient (incidence matrix over the path graph)

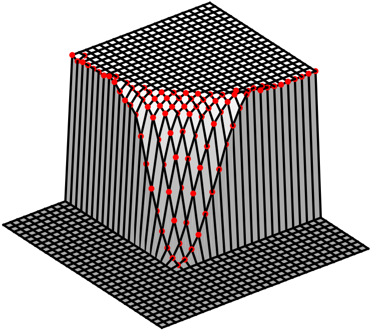


LSIsoTV: Perform least-squares over estimated constant part

(Isotropic) 2D-Total Variation and its bias

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

$D^\top$: discrete gradient (incidence matrix over the path graph)



CLEAR: Perform least-squares covariant refitting

CLEAR Refitting strategy

Definition

The *Covariant LEast-square Re-fitting* associated to an a.e. differentiable estimator $y \mapsto \hat{\beta}(y)$ is, for $y \in \mathbb{R}^n$, given by

$$\mathcal{R}_{\hat{\beta}}(y) = \hat{\beta}(y) + \rho J(y - X\hat{\beta}(y))$$

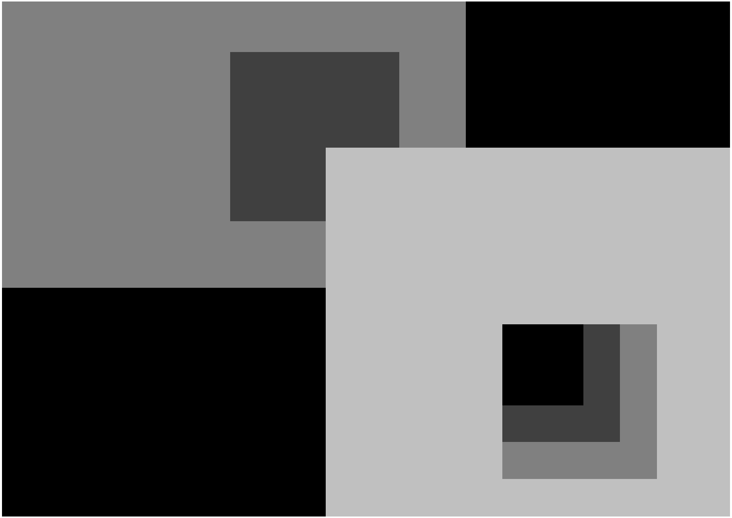
$$\text{with} \quad \rho = \begin{cases} \frac{\langle XJ\delta, \delta \rangle}{\|XJ\delta\|^2} & \text{if } XJ\delta \neq 0 \ , \\ 1 & \text{otherwise} \ , \end{cases}$$

where $\delta = y - X\hat{\beta}(y)$ is the residual and $J = J_{\hat{\beta}}(y)$ is the **Jacobian** matrix of $\hat{\beta}$ at y

Rem: the Jacobian matrix of the original and refitted estimators are preserved, up to a constant (covariant)

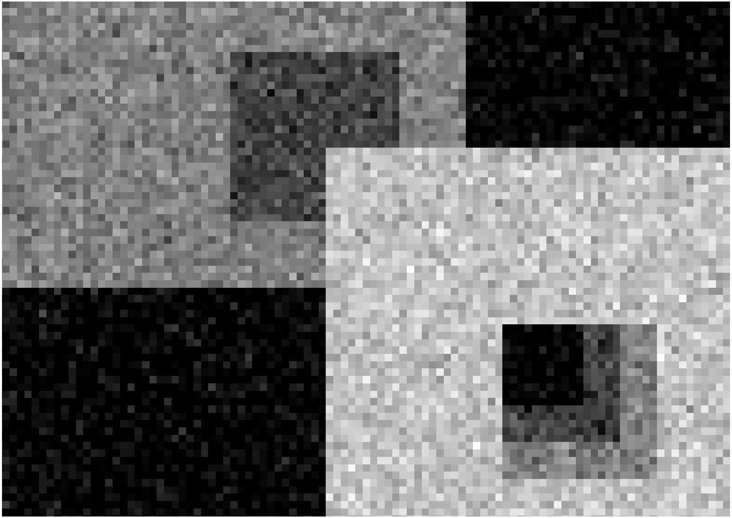
Rem: note that for Lasso and AnisoTV, CLEAR simply reads $\mathcal{R}_{\hat{\beta}}(y) = Jy$ and for Iso-TV, $\mathcal{R}_{\hat{\beta}}(y) = (1 - \rho)\hat{\beta}(y) + \rho Jy$

In practice



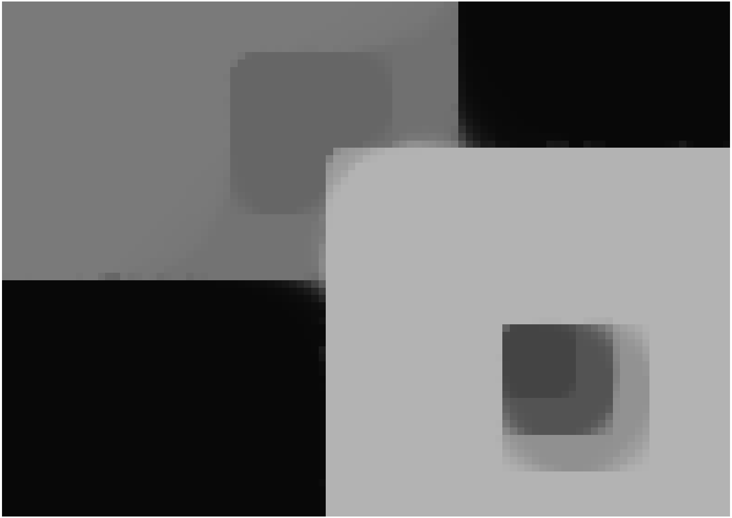
True image

In practice



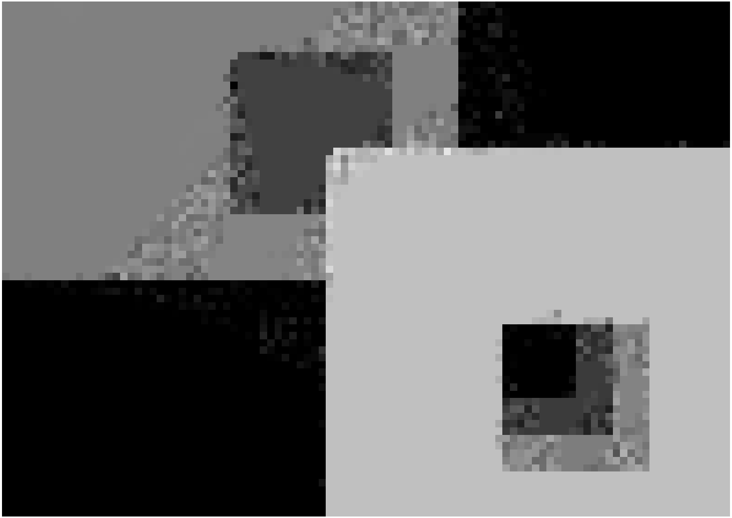
Noisy image

In practice



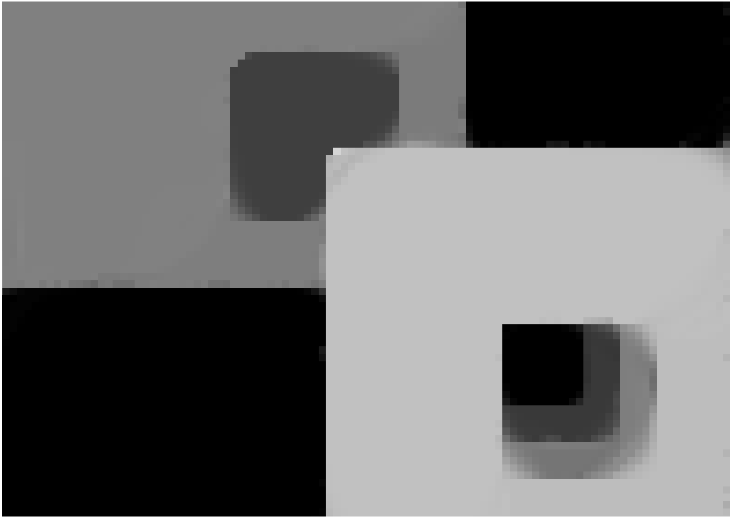
Iso-TV denoising

In practice



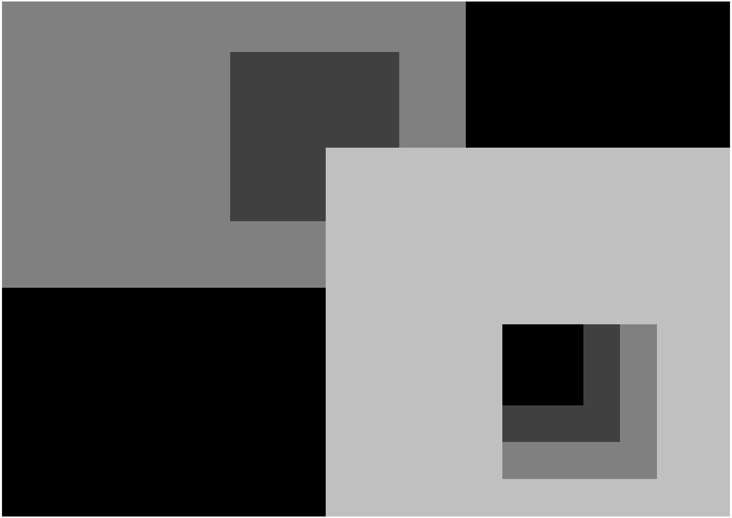
Iso-TV with invariant refitting

In practice



Iso-TV with CLEAR refitting

In practice



True image

Algorithms for CLEAR

Possible numerical schemes:

- Algorithmic differentiation
- Finite difference based differentiation
- Two-step computation for the general case

Rem: it can be applied to other methods, not only variational ones, *e.g.*, NL-Means Buades *et al.* (2005), DDID Knaus and Zwicker (2013), etc.

LSLasso with Coordinate Descent

Algorithm: CD Epoch for CLEAR Lasso (or LSLasso)

input : X, y, λ

param: $\beta = 0_p$, $\widetilde{\beta} = 0_p$, $\forall j \in [\![p]\!], L_j = \|\mathbf{x}_j\|^2$

for $j = 1, \ldots, p$ **do**

$$\widetilde{\beta}_j \leftarrow \left(\widetilde{\beta}_j - \frac{1}{L_j}\mathbf{x}_j^\top (X\widetilde{\beta} - y)\right) \mathbb{1}_{|\beta_j| > \frac{\lambda}{L_j}}$$ `// refitting part`

$$\beta_j \leftarrow \mathrm{ST}\left(\beta_j - \frac{1}{L_j}\mathbf{x}_j^\top (X\beta - y), \frac{\lambda}{L_j}\right)$$

`// soft-thresholding`

return β

<u>Benefits</u>: limit instabilities from two step approach when stopping before support identification

One step further: handling sign constraints

Definition: SLSLasso (Sign Least Squares Lasso)

Constrain the sign not to change after refitting

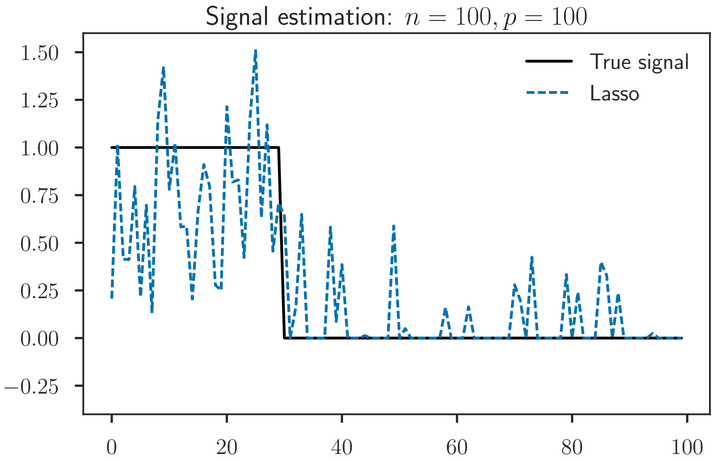
$$\hat{\beta}_E^{\mathrm{SLSLasso}} \in \underset{\beta_E \in \mathbb{R}^{|E|} : \rho_E^{\lambda_1} \odot \beta_E \geq 0}{\arg\min} \quad \|y - X_E \beta_E\|^2 \ ,$$

where the **equicorrelation set** Tibshirani (2013) is

$$E = \{j \in [p] : |\rho_j^\lambda| = 1\} \text{ where } \rho^\lambda = \frac{X^\top (y - X\beta^{(\lambda)})}{\lambda}$$

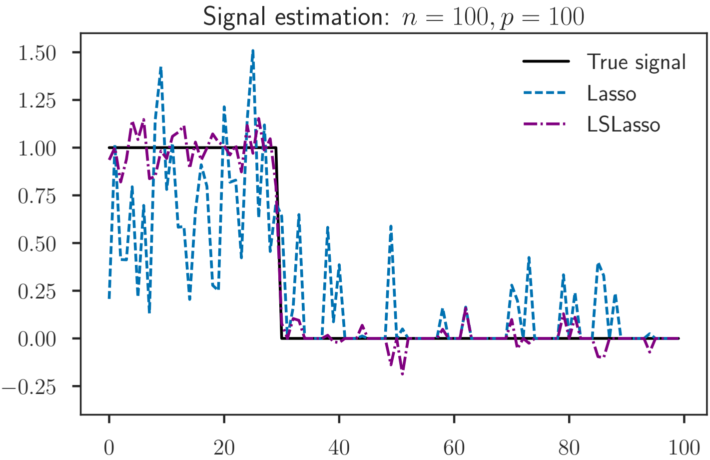
Rem: recover Bregman regularization scheme Brinkmann *et al.* (2016) (used for TV), analyzed for Lasso Chzhen *et al.* (2017)

SLSLasso



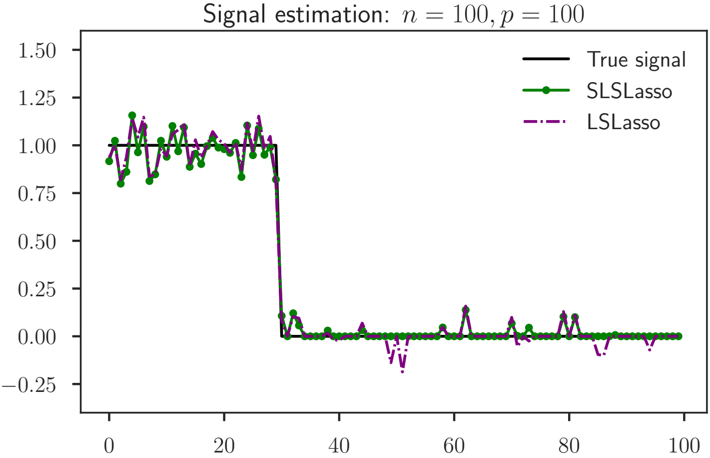
Signal estimation: $n = 100, p = 100$
True signal
Lasso
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
0
20
40
60
80
100

SLSLasso



Signal estimation: $n = 100, p = 100$
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
0
20
40
60
80
100
True signal
Lasso
LSLasso

SLSLasso



Signal estimation: $n = 100, p = 100$
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
0
20
40
60
80
100
True signal
SLSLasso
LSLasso

Table of Contents

Acknowledgments

This part takes its roots in a collaboration with **A. Dalalyan**, **K. Meziani**, **M. Hebiri**, Dalalyan *et al.* (2013)

Recently revisited with:

- **A. Gramfort**, **O. Fercoq**, **V. Leclère**, **E. Ndiaye** (optimization aspects) Ndiaye *et al.* (2017)
- **C. Boyer** and **Y. De Castro** (super-resolution) Boyer *et al.* (2017)
- **M. Massias**, **A. Gramfort** and **O. Fercoq** (heteroscedastic extensions to address problems in neuro-imaging) Massias *et al.* (2017)

The Concomitant Lasso

$$\boxed{(\hat{\beta}^{(\lambda)}, \sigma^{(\lambda)}) \in \mathop{\arg\min}_{\beta \in \mathbb{R}^p, \sigma > 0} \left(\frac{1}{2n\sigma} \|y - X\beta\|^2 + \frac{\sigma}{2} + \lambda \|\beta\|_1 \right)}$$

- Jointly convex method, introduced by Owen (2007)
- Analyzed by Sun and Zhang (2012) as "Scaled-Lasso"
  <u>beware</u>: different from Scaled-Lasso by Städler *et al.* (2010)
- Constraint not closed, might get $\sigma \to 0$: use Fenchel bi-conjugate, so objective accept $\sigma = 0$ for $y = X\beta$
  see also Combettes and Müller (2016), Combettes (2016)
- Roots in Huber (1981)'s work

$$\mathop{\arg\min}_{\beta \in \mathbb{R}^p, \sigma > 0} \frac{1}{n} \sum_{i=1}^{n} \sigma \cdot \ell \left(\frac{y_i - \langle X_{i,:}, \beta \rangle}{\sigma} \right) + \frac{\sigma}{2}$$

Link with the $\sqrt{\text{Lasso}}$ Belloni *et al.* (2011)

- Independently, Belloni *et al.* (2011) analyzed $\sqrt{\text{Lasso}}$ to get "σ free" choice of λ (in theoretical bounds)

$$\boxed{\hat{\beta}^{(\lambda)}_{\sqrt{\text{Lasso}}} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \left(\frac{1}{\sqrt{n}} \|y - X\beta\| + \lambda \|\beta\|_1 \right)}$$

- Connexions with Concomitant Lasso:
$\left(\hat{\beta}^{(\lambda)}_{\sqrt{\text{Lasso}}}, \hat{\sigma}^{(\lambda)}_{\sqrt{\text{Lasso}}}\right)$ is solution of the Concomitant Lasso for

$$\hat{\sigma}^{(\lambda)}_{\sqrt{\text{Lasso}}} = \frac{\left\| y - X\hat{\beta}^{(\lambda)}_{\sqrt{\text{Lasso}}} \right\|}{\sqrt{n}}$$

The Smoothed Concomitant Lasso Ndiaye *et al.* (2016)

To remove issues for small λ (and σ), we have introduced:

$$\boxed{\hat{\beta}^{(\lambda,\sigma_0)}, \hat{\sigma}^{(\lambda,\sigma_0)} \in \underset{\beta\in\mathbb{R}^p, \sigma\geq\sigma_0}{\arg\min} \frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1}$$

- With prior information on the minimal noise level, one can set σ_0 as this bound (and both estimators are the same)
- Setting $\sigma_0 = \epsilon$, smoothing theory asserts that $\frac{\epsilon}{2}$-solutions for the smoothed problem provide ϵ-solutions for the original one Nesterov (2005)

Smoothing aparté Nesterov (2005), Beck and Teboulle (2012)

If f is non-smooth, then make it smooth, using for some $\mu > 0$, f_μ:

$$f_\mu = \mu\omega\left(\frac{\cdot}{\mu}\right)\Box f$$

where $f\Box g(x) = \inf_u f(u) + g(x-u)$ for a predefined function ω

"**Huberization**": $f(\beta) = \frac{\|y-X\beta\|}{\sqrt{n}}$, $\mu = \sigma_0$, $\omega(\beta) = \frac{\|\beta\|^2}{2} + \frac{1}{2}$

$$f_{\sigma_0}(\beta) = \begin{cases} \frac{\|y-X\beta\|^2}{2n\sigma_0} + \frac{\sigma_0}{2} & \text{if } \frac{\|y-X\beta\|}{\sqrt{n}} \leq \sigma_0 \\ \frac{\|y-X\beta\|}{\sqrt{n}} & \text{if } \frac{\|y-X\beta\|}{\sqrt{n}} > \sigma_0 \end{cases}$$

$$= \min_{\sigma \geq \sigma_0} \frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2}$$

Coordinate Descent for Smoothed Concomitant Lasso

Algorithm: CD FOR SMOOTHED CONCOMITANT LASSO

input : X, y, λ, σ_0

param: $\beta = 0_p$, $\forall j \in [\![p]\!], L_j = \|\mathbf{x}_j\|^2$, $\sigma = \sigma_0 \vee \frac{\|y - X\beta\|}{\sqrt{n}}$

for $j = 1, \ldots, p$ **do**

$$\beta_j \leftarrow \mathrm{ST}\left(\beta_j - \frac{1}{L_j}\mathbf{x}_j^\top (X\beta - y), \frac{n\sigma\lambda}{L_j}\right)$$ `// coefficient update`

$$\sigma \leftarrow \sigma_0 \vee \frac{\|y - X\beta\|}{\sqrt{n}}$$ `// standard deviation update`

return β, σ

Rem: previous screening strategies apply straightforwardly;
e.g., **critical value** for $\lambda > \lambda_{\max}$, $\hat{\beta}^{(\lambda)} = 0$

$$\lambda_{\max} = \frac{\left\|X^\top y\right\|_\infty}{\|y\| \sqrt{n}} = \max_{j=1,\ldots,p} \left| \left\langle \mathbf{x}_j, \frac{y}{\|y\| \sqrt{n}} \right\rangle \right|$$

Heteroscedastic ... and more

Motivation: unknown noise level with a piecewise structure *e.g.*, in M/EEG three types of sensors are aggregated Massias *et al.* (2017)

We proposed the convex formulation (works for diagonal Σ):

$$\boxed{\mathop{\arg\min}_{\substack{\beta\in\mathbb{R}^p,\Sigma\in\mathbb{S}^n_{++},\\ \Sigma\succeq\underline{\Sigma}}} \frac{1}{2n}(y-X\beta)^\top\Sigma^{-1}(y-X\beta)+\frac{1}{2n}\operatorname{tr}(\Sigma)+\lambda\left\|\beta\right\|_1}$$

Rem: on going work for handling general correlated model, need more math for covariance update

Rem: alternative to SOCP formulation from Dalalyan *et al.* (2013)

Table of Contents

Other projects involving Phd students

- Matrix completion: **Jean Lafond**, co-supervised with E. Moulines
- Gossip algorithms for decentralized machine learning: **Igor Colin**, co-supervised with S. Clemençon
- Extreme multi-label classification: **Evgenii Chzhen**, co-supervised with M. Hebiri

Rem: **Eugène Ndiaye** (co-supervised with O. Fercoq) and **Mathurin Massias** (co-supervised with A. Gramfort and O. Cappé) already mentioned

Rem: **Simon Amar** and **Jérôme-Alexis Chevalier** (co-supervised with B. Thirion) soon to start!

Other projects involving Phd students

- Matrix completion: **Jean Lafond**, co-supervised with E. Moulines
- Gossip algorithms for decentralized machine learning: **Igor Colin**, co-supervised with S. Clemençon
- Extreme multi-label classification: **Evgenii Chzhen**, co-supervised with M. Hebiri

Rem: **Eugène Ndiaye** (co-supervised with O. Fercoq) and **Mathurin Massias** (co-supervised with A. Gramfort and O. Cappé) already mentioned

Rem: **Simon Amar** and **Jérôme-Alexis Chevalier** (co-supervised with B. Thirion) soon to start!

References I

- E.-M. Brinkmann, M. Burger, J. Rasch, and C. Sutour.
  Bias-Reduction in Variational Regularization.
  ArXiv e-prints, 2016.
- A. Buades, B. Coll, and J.-M. Morel.
  A review of image denoising algorithms, with a new one.
  Multiscale Model. Simul., 4(2):490–530, 2005.
- A. Belloni, V. Chernozhukov, and L. Wang.
  Square-root Lasso: pivotal recovery of sparse signals via conic programming.
  Biometrika, 98(4):791–806, 2011.
- C. Boyer, Y. De Castro, and J. Salmon.
  Adapting to unknown noise level in sparse deconvolution.
  Information and Inference, 6(3):310–348, 2017.
- A. Bonnefoy, V. Emiya, L. Ralaivola, and R. Gribonval.
  A dynamic screening principle for the lasso.
  In *EUSIPCO*, 2014.

References II

- A. Bonnefoy, V. Emiya, L. Ralaivola, and R. Gribonval.
  Dynamic screening: accelerating first-order algorithms for the Lasso and Group-Lasso.
  IEEE Trans. Signal Process., 63(19):20, 2015.
- 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.
- A. Beck and M. Teboulle.
  Smoothing and first order methods: A unified framework.
  SIAM J. Optim., 22(2):557–580, 2012.
- S. S. Chen, D. L. Donoho, and M. A. Saunders.
  Atomic decomposition by basis pursuit.
  SIAM J. Sci. Comput., 20(1):33–61, 1998.

References III

- E. Chzhen, M. Hebiri, and J. Salmon.
  On Lasso refitting strategies.
  ArXiv e-prints, 2017.
- P. L. Combettes and C. L. Müller.
  Perspective functions: Proximal calculus and applications in high-dimensional statistics.
  J. Math. Anal. Appl., 2016.
- P. L. Combettes.
  Perspective functions: Properties, constructions, and examples.
  arXiv preprint arXiv:1610.01552, 2016.
- P. L. Combettes and J.-C. Pesquet.
  Proximal splitting methods in signal processing.
  In *Fixed-point algorithms for inverse problems in science and engineering*, volume 49 of *Springer Optim. Appl.*, pages 185–212. Springer, New York, 2011.

References IV

- 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.
- I. Daubechies, M. Defrise, and C. De Mol.
  An iterative thresholding algorithm for linear inverse problems with a sparsity constraint.
  Comm. Pure Appl. Math., 57(11):1413–1457, 2004.
- A. S. Dalalyan, M. Hebiri, and J. Lederer.
  On the prediction performance of the Lasso.
  Bernoulli, 23(1):552–581, 2017.
- A. S. Dalalyan, M. Hebiri, K. Meziani, and J. Salmon.
  Learning heteroscedastic models by convex programming under group sparsity.
  In *ICML*, 2013.

References V

- C.-A. Deledalle, N. Papadakis, and J. Salmon.
  On debiasing restoration algorithms: applications to total-variation and nonlocal-means.
  In *SSVM*, pages 129–141, 2015.
- C.-A. Deledalle, N. Papadakis, J. Salmon, and S. Vaiter.
  CLEAR: Covariant LEAst-square Re-fitting with applications to image restoration.
  SIAM J. Imaging Sci., 10(1):243–284, 2017.
- B. Efron, T. J. 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.

References VI

- O. Fercoq, A. Gramfort, and J. Salmon.
  Mind the duality gap: safer rules for the lasso.
  In *ICML*, pages 333–342, 2015.
- J. Friedman, T. J. Hastie, H. Höfling, and R. Tibshirani.
  Pathwise coordinate optimization.
  Ann. Appl. Stat., 1(2):302–332, 2007.
- C. Giraud.
  Introduction to high-dimensional statistics, volume 138.
  CRC Press, 2014.
- P. J. Huber.
  Robust Statistics.
  John Wiley & Sons Inc., 1981.

References VII

- T. Joachims.

  Text categorization with support vector machines: Learning with many relevant features.

  In Claire Nédellec and Céline Rouveirol, editors, *Machine Learning: ECML-98*, volume 1398 of *Lecture Notes in Computer Science*, pages 137–142. Springer Berlin Heidelberg, 1998.

- 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. Kim and H. Park.

  Fast active-set-type algorithms for l1-regularized linear regression.

  In *AISTATS*, pages 397–404, 2010.

- M. Kowalski, P. Weiss, A. Gramfort, and S. Anthoine.

  Accelerating ISTA with an active set strategy.

  In *OPT 2011: 4th International Workshop on Optimization for Machine Learning*, page 7, 2011.

References VIII

- C. Knaus and M. Zwicker.
  Dual-domain image denoising.
  In *ICIP*, pages 440–444, 2013.

- M. Massias, O. Fercoq, A. Gramfort, and J. Salmon.
  Heteroscedastic concomitant lasso for sparse multimodal electromagnetic brain imaging.
  In *AISTATS*, 2018.

- M. Massias, A. Gramfort, and J. Salmon.
  From safe screening rules to working sets for faster lasso-type solvers.
  In *NIPS-OPT*, 2017.

- E. Mammen and S. van de Geer.
  Locally adaptive regression splines.
  Ann. Statist., 25(1):387–413, 1997.

- J. Mairal and B. Yu.
  Complexity analysis of the lasso regularization path.
  In *ICML*, pages 353–360, 2012.

References IX

- Y. Nesterov.
  Smooth minimization of non-smooth functions.
  Math. Program., 103(1):127–152, 2005.
- E. Ndiaye, O. Fercoq, A. Gramfort, V. Leclère, and J. Salmon.
  Efficient smoothed concomitant lasso estimation for high dimensional regression.
  Journal of Physics: Conference Series, 904(1):012006, 2017.
- E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon.
  Gap safe screening rules for sparse multi-task and multi-class models.
  In *NIPS*, pages 811–819, 2015.
- E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon.
  GAP safe screening rules for sparse-group-lasso.
  In *NIPS*, 2016.
- E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon.
  Gap safe screening rules for sparsity enforcing penalties.
  J. Mach. Learn. Res., 18(128):1–33, 2017.

References X

- 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.
- A. B. Owen.
  A robust hybrid of lasso and ridge regression.
  Contemporary Mathematics, 443:59–72, 2007.
- 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.
- V. Roth and B. Fischer.
  The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms.
  In *ICML*, pages 848–855, 2008.

References XI

- L. I. Rudin, S. Osher, and E. Fatemi.
  Nonlinear total variation based noise removal algorithms.
  Phys. D, 60(1-4):259–268, 1992.
- N. Städler, P. Bühlmann, and S. van de Geer.
  ℓ_1-penalization for mixture regression models.
  TEST, 19(2):209–256, 2010.
- T. Sun and C.-H. Zhang.
  Scaled sparse linear regression.
  Biometrika, 99(4):879–898, 2012.
- R. Tibshirani, J. Bien, J. Friedman, T. J. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani.
  Strong rules for discarding predictors in lasso-type problems.
  J. R. Stat. Soc. Ser. B Stat. Methodol., 74(2):245–266, 2012.
- H. L. Taylor, S. C. Banks, and J. F. McCoy.
  Deconvolution with the ℓ_1 norm.
  Geophysics, 44(1):39–52, 1979.

References XII

- R. Tibshirani.
  Regression shrinkage and selection via the lasso.
  J. R. Stat. Soc. Ser. B Stat. Methodol., 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.

The Gap safe sphere is safe (proof)

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

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

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

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

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

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