Convex relaxation

- Problem: solving the model selection minimization criterion

$$\hat{m} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left\{ \|Y - \hat{f}_m\|^2 + \operatorname{pen}(m)\,\sigma^2 \right\}$$

is impossible in practice when $\mathcal{M}$ is very large.

Ex: coordinate sparse regression $\mathcal{M} = \mathcal{P}(\{1, \dots, p\})$, so we must evaluate $|\mathcal{M}| = 2^p$ quantities :-(

- Today's recipe: modify the model selection criterion, in order to obtain a convex criterion, amenable to numerical computations.

① Lasso estimator

- Sparse linear regression:
  - $Y = X\beta^* + \varepsilon$ with $|\beta^*|_0$ small
  - in all this lecture, we assume that the columns of X are normalized: $\|X_j\| = 1$ for $j = 1, \dots, p$

- Model selection estimator:
  - $\mathcal{M} = \mathcal{P}(\{1, \dots, p\})$, $S_m = \operatorname{Span}\{X_j : j \in m\}$, $\hat{f}_m = \operatorname{Proj}_{S_m} Y$
  - $\pi_m = \exp(-|m| \log p)$
  - $\hat{m} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left\{ \|Y - \hat{f}_m\|^2 + \lambda |m| \right\}$, with $\lambda = K\left(1 + \sqrt{2\log p}\right)^2 \sigma^2$

- **Convexification:**

  - $\hat{m} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \left\{ \|Y - \hat{f}_m\|^2 + \lambda |m| \right\}$

  - we have $\hat{f}_m = X \hat{\beta}_m$ where $\hat{\beta}_m \in \underset{\beta : \operatorname{supp}(\beta) = m}{\operatorname{argmin}} \|Y - X\beta\|^2$

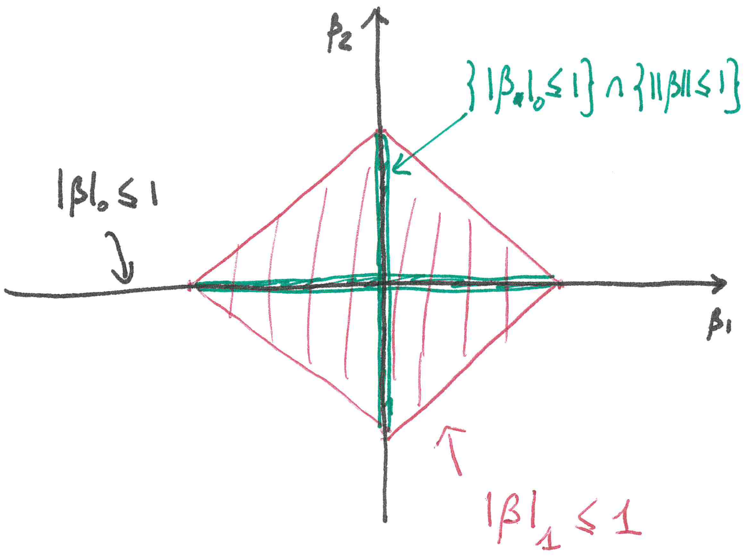
So

$$\hat{m} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \ \min_{\beta : \operatorname{supp}(\beta) = m} \left\{ \|Y - X\beta\|^2 + \lambda |\beta|_0 \right\}$$

and

$$\hat{\beta}_{\hat{m}} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \underbrace{\|Y - X\beta\|^2}_{\text{convex :)}} + \lambda \underbrace{|\beta|_0}_{\text{highly-non convex :(}} \right\}.$$

- recipe



constrained version:

$$\min_{|\beta|_0 \le D} \|Y - X\beta\|^2$$

convexification:

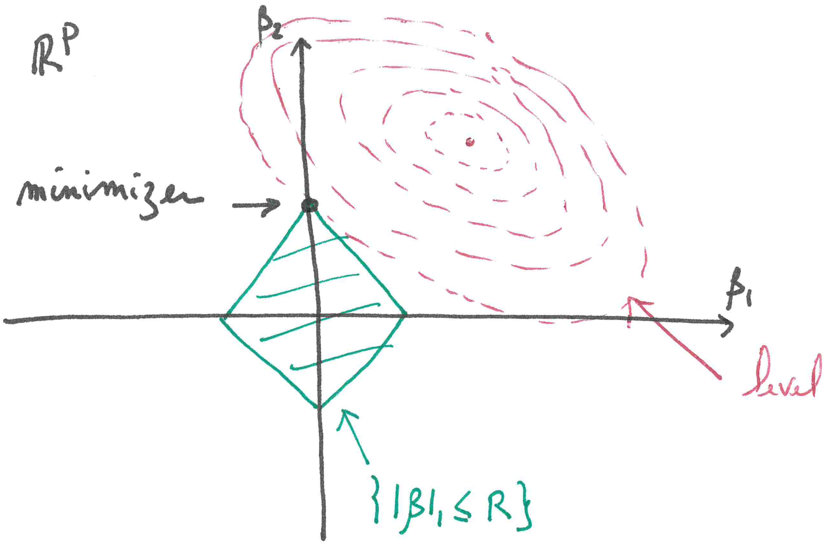
$|\beta|_0 \le D \leadsto |\beta|_1 \le R$

Lasso:

- $\hat{\beta}_\lambda \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \underbrace{\|Y - X\beta\|^2 + \lambda |\beta|_1}_{=: \mathcal{L}_\lambda(\beta) \ \text{convex}} \right\}$ for $\lambda > 0$

- $\hat{f}_\lambda = X \hat{\beta}_\lambda$

- Geometric interpretation

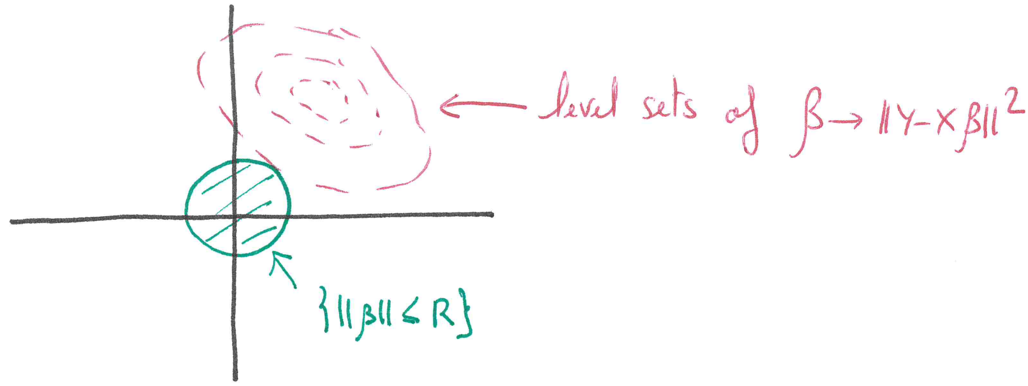


Constrained version:

$$\min_{|\beta|_1 \leq R} \|Y - X\beta\|^2$$

Singularities of $\{|\beta|_1 \leq R\}$ $\longleftrightarrow$ selection of variables

Remark: if we replace $|\beta|_1$ by $\|\beta\|^2$, no selection occurs

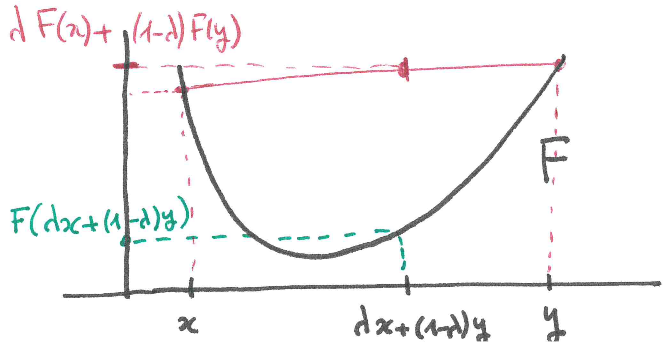


To do: exercise 5.5.7 parts A) and B). (ridge & elastic net)

- Analytic interpretation

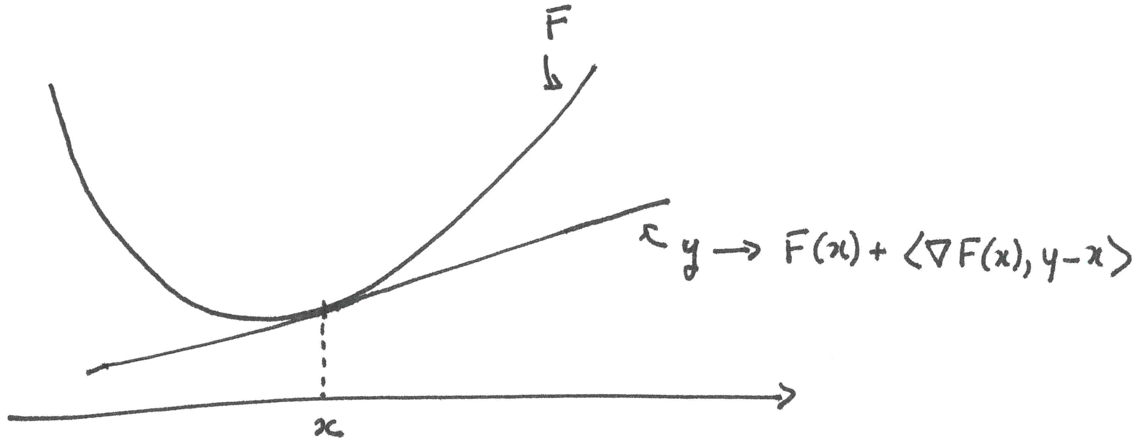
Reminder on convex functions and subdifferentials

- $F: \mathbb{R}^d \to \mathbb{R}$ is convex iff $\begin{cases} F(\lambda x + (1-\lambda)y) \leq \lambda F(x) + (1-\lambda) F(y) \\ \forall \lambda \in [0,1] \text{ and } \forall x, y \in \mathbb{R}^d \end{cases}$



Lemma D1: if F is convex and differentiable in x

$$F(y) \geq F(x) + \langle \nabla F(x), y-x \rangle, \quad \forall y \in \mathbb{R}^d$$



Subdifferential: if F is convex, we define the subdifferential

$$\partial F(x) = \{ w \in \mathbb{R}^d : F(y) \geq F(x) + \langle w, y-x \rangle \ \ \forall y \in \mathbb{R}^d \}$$

$\uparrow$ subgradient

Lemma D2: Let $F : \mathbb{R}^d \to \mathbb{R}$ be a convex function

1. $\partial F(x) \neq \emptyset$
2. if F is diff. in x: $\partial F(x) = \{\nabla F(x)\}$.

Properties: $F : \mathbb{R}^d \to \mathbb{R}$ convex

1. monotonicity: $\forall w_x \in \partial F(x)$ and $\forall w_y \in \partial F(y)$

$$\langle w_x - w_y, x-y \rangle \geq 0$$

2. minimum:

$$x^* \in \operatorname*{argmin}_{x \in \mathbb{R}^d} F(x) \iff 0 \in \partial F(x^*)$$

Proof:

1- from the very definition:

$$F(y) \geq F(x) + \langle w_x, y-x \rangle$$

$$F(x) \geq F(y) + \langle w_y, x-y \rangle$$

$\oplus$

$$0 \geq \langle w_x - w_y, y-x \rangle$$

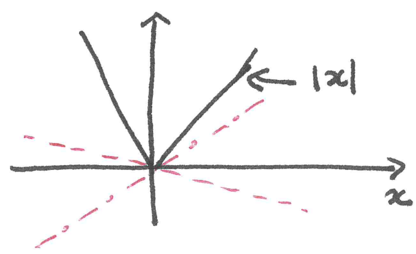
2- both statements are equivalent to

$$F(y) \geq F(x^*) + \langle 0, y-x^* \rangle \quad \forall y \in \mathbb{R}^d$$

$\square$

Ex: $\partial |x|_1$?

- dim = 1: $|x|_1 = |x|$



So $\partial |x| = \begin{cases} \mathrm{sign}(x) & \text{if } x \neq 0 \\ [-1,1] & \text{if } x = 0 \end{cases}$

- dim = d: $|x|_1 = \sum_{j=1}^{d} |x_j|$

So $\partial |x|_1 = \{ z \in \mathbb{R}^d : z_j = \mathrm{sign}(x_j) \text{ if } x_j \neq 0 \text{ and } z_j \in [-1,1] \text{ if } x_j = 0 \}$

. it is insightful to recover this result from a more principled way.

reminder: $|x|_1 = \sup_{|\phi|_\infty \leq 1} \langle \phi, x \rangle$.

We will prove $\partial |x|_1 = \mathcal{D}_x$ where $\mathcal{D}_x = \{ \phi : \langle \phi, x \rangle = |x|_1, |\phi|_\infty \leq 1 \}$

proof:

($\supset$) for $\phi \in \mathcal{D}_x$:

$$|y|_1 \underset{\substack{\uparrow \\ |\phi|_\infty \leq 1}}{\geq} \langle \phi, y \rangle \underset{\substack{\uparrow \\ \langle \phi, x \rangle = |x|_1}}{=} |x|_1 + \langle \phi, y-x \rangle$$

Hence $\phi \in \partial |x|_1$.

(C) for $w \in \partial |x|_1$:

- $y = 2x$: $\quad 2|x|_1 \geqslant |x|_1 + \langle w, x \rangle$
- $y = 0$: $\quad 0 \geqslant |x|_1 + \langle w, -x \rangle$

$$\left.\begin{array}{l} 2|x|_1 \geqslant |x|_1 + \langle w, x \rangle \\ 0 \geqslant |x|_1 + \langle w, -x \rangle \end{array}\right\} \Rightarrow |x|_1 \leqslant \langle w, x \rangle \leqslant |x|_1$$

- we also have $|w|_\infty = \langle w, z \rangle$ with $|z|_1 = 1$

so

$$|x|_1 + \underbrace{|z|_1}_{=1} \overset{\Delta}{\geqslant} |x + z|_1 \overset{w \in \partial|x|_1}{\geqslant} |x|_1 + \underbrace{\langle w, z \rangle}_{= |w|_\infty} \Rightarrow |w|_\infty \leqslant 1$$

and hence $w \in D_x$.

$\square$

Lasso expression:

since $\beta \to \mathcal{L}_\lambda(\beta) = \|Y - X\beta\|^2 + \lambda |\beta|_1$ is convex

$$\partial \mathcal{L}_\lambda(\beta) = \left\{ -2X^T(Y - X\beta) + \lambda z : \text{ with } z \in \partial|\beta|_1 \right\}$$

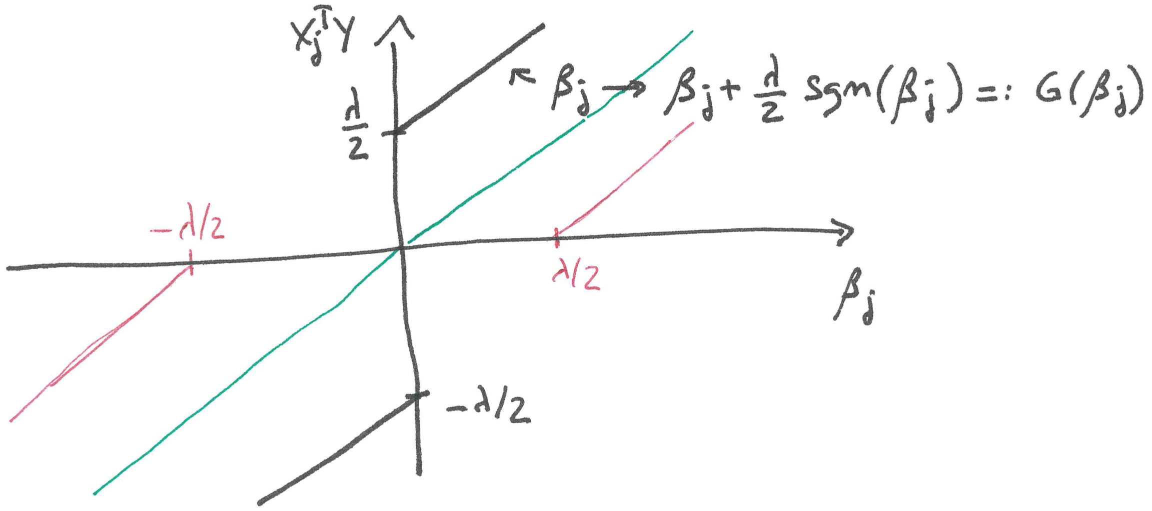
and $\exists\, \hat{z} \in \partial|\hat{\beta}_\lambda|_1$ such that

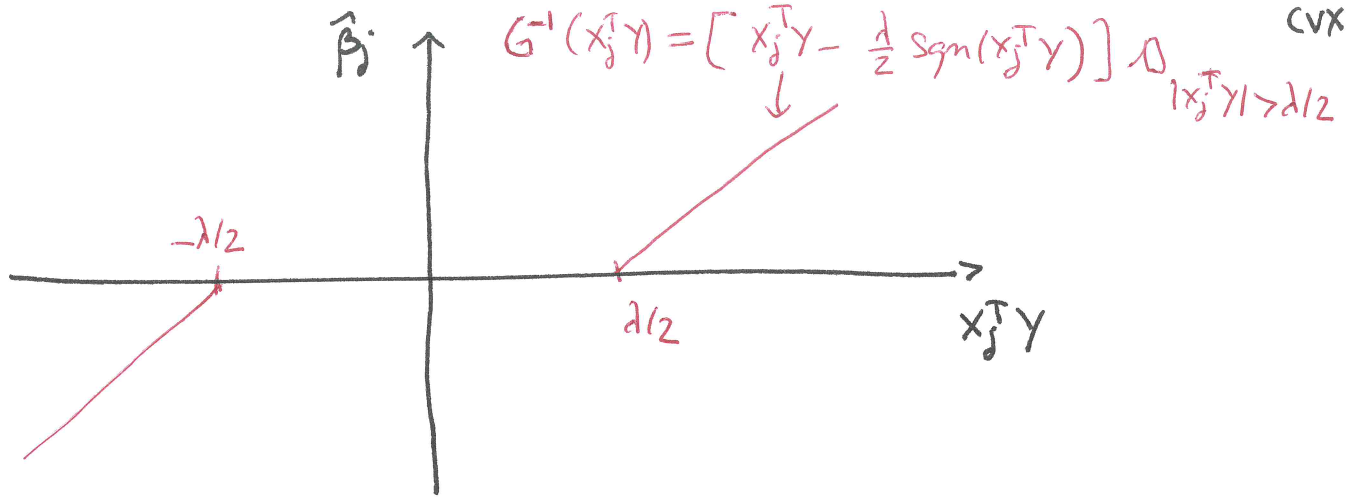
$$X^T X \hat{\beta}_\lambda = X^T Y - \frac{\lambda}{2} \hat{z}$$

no explicit expression, but in the orthogonal case.

- case $X^T X = I_d$: $\quad \hat{\beta} = X^T Y - \frac{\lambda}{2} \hat{z}$.

if $\hat{\beta}_j \neq 0$: $\quad \hat{\beta}_j = X_j^T Y - \frac{\lambda}{2} \operatorname{sgn}(\hat{\beta}_j) \Rightarrow X_j^T Y = \hat{\beta}_j + \frac{\lambda}{2} \operatorname{sign}(\hat{\beta}_j)$





So: • if $|X_j^T Y| > \lambda/2$: $\hat\beta_j = X_j^T Y - \frac{\lambda}{2}\operatorname{sgn}(X_j^T Y)$

• if $|X_j^T Y| \le \lambda/2$: $\hat\beta_j = 0$ and $\hat z_j = \frac{2}{\lambda} X_j^T Y$

compact formula: soft-thresholding

$$\hat\beta_j = X_j^T Y \left(1 - \frac{\lambda}{2|X_j^T Y|}\right)_+ \qquad \text{for } j = 1, \dots, p.$$

② Statistical analysis of Lasso estimator

Compatibility constant: account for (local) orthogonality

$$\kappa(\beta) = \min\left\{ \frac{\sqrt{|\beta|_0}\,\|Xv\|}{|v_S|_1} \; : \; v \in \mathcal{C}(\beta) \right\}$$

where • $S = \operatorname{supp}(\beta)$

• $\mathcal{C}(\beta) = \{ v \in \mathbb{R}^p : 5|v_S|_1 > |v_{S^c}|_1 \}$.

Exercise: check that

1/ if $X^T X = I$ then $\kappa(\beta) \ge 1$

2) we always have $\kappa(\beta) \ge \lambda_{\min}(X^T X)^{1/2}$

Theorem 5.1 Deterministic bound

For any $\lambda > 3|X^T\varepsilon|_\infty$, we have

$$\|X\hat{\beta}_\lambda - X\beta^*\|^2 \leq \inf_{\beta\in\mathbb{R}^p} \left\{ \|X\beta - X\beta^*\|^2 + \frac{\lambda^2}{\kappa(\beta)^2}|\beta|_0 \right\}$$

Corollary 5.3

Assume that $\begin{cases} \cdot\ \|X_j\| = 1, \quad j=1,\dots,p \\ \cdot\ \varepsilon \sim \mathcal{N}(0,\sigma^2 I_n) \\ \cdot\ \lambda = 3\sigma\sqrt{2K\log p} \end{cases}$, with $K>1$

Then, with probability $\geq 1 - \frac{1}{p^{K-1}}$

$$\|X\hat{\beta}_\lambda - X\beta^*\|^2 \leq \inf_{\beta\in\mathbb{R}^p} \left\{ \|X\beta - X\beta^*\|^2 + \frac{18K\sigma^2\log p}{\kappa(\beta)^2}|\beta|_0 \right\}$$

$$\leq \inf_{m\subset\{1,\dots,p\}} \left\{ \|X\hat{\beta}_m - X\beta^*\|^2 + \frac{18K\sigma^2\log p}{\kappa(\hat{\beta}_m)^2}|m| \right\}$$

price to pay for computational tractability

Proof Cor. 5.3:

- $|X^T\varepsilon|_\infty = \max_{j=1,\dots,p} |X_j^T\varepsilon|$, with $X_j^T\varepsilon \sim \mathcal{N}(0,\sigma^2)$

- Hence

$$\mathbb{P}\left[|X^T\varepsilon|_\infty > \sigma\sqrt{2K\log(p)}\right] \leq p\,\mathbb{P}\left[|\mathcal{N}(0,1)| > \sqrt{2K\log p}\right]$$

$$\leq p\exp\left(-\tfrac{1}{2}(2K\log p)\right) = \frac{1}{p^{K-1}}$$

Lemma B.4.

$\square$

Proof of Theorem 5.1:

- optimality condition: $0 \in \partial \mathcal{L}_\lambda(\hat{\beta}_\lambda)$

  So $\exists \hat{z} \in \partial |\hat{\beta}_\lambda|_1$ such that $2X^T(X\hat{\beta}_\lambda - Y) + \lambda \hat{z} = 0$.

- $\forall \beta \in \mathbb{R}^p$:

$$2 \langle X^T(X\hat{\beta}_\lambda - X\beta^* - \varepsilon), \hat{\beta}_\lambda - \beta \rangle + \lambda \langle \hat{z}, \hat{\beta}_\lambda - \beta \rangle = 0$$

i.e. $$2 \langle X\hat{\beta}_\lambda - X\beta^*, X\hat{\beta}_\lambda - X\beta \rangle - 2\langle X^T\varepsilon, \hat{\beta}_\lambda - \beta \rangle + \lambda \langle \hat{z}, \hat{\beta}_\lambda - \beta \rangle = 0$$

- Monotonicity: $\forall z \in \partial |\beta|_1$: $\langle \hat{z}, \hat{\beta}_\lambda - \beta \rangle \geq \langle z, \hat{\beta}_\lambda - \beta \rangle$

  So $\forall \beta \in \mathbb{R}^p$, $\forall z \in \partial |\beta|_1$

$$\underbrace{2 \langle X(\hat{\beta}_\lambda - \beta^*), X(\hat{\beta}_\lambda - \beta) \rangle}_{=: A(\hat{\beta}_\lambda)} \leq 2 \langle X^T\varepsilon, \hat{\beta}_\lambda - \beta \rangle - \lambda \langle z, \hat{\beta}_\lambda - \beta \rangle \quad (1)$$

- Lemma 5.2: $S = \mathrm{supp}(\beta)$

$$\boxed{A(\hat{\beta}_\lambda) \underset{(i)}{\leq} \frac{\lambda}{3} \left(5 |(\hat{\beta}_\lambda - \beta)_S|_1 - |(\hat{\beta}_\lambda - \beta)_{S^c}|_1 \right) \underset{(ii)}{\leq} 2\lambda |(\hat{\beta}_\lambda - \beta)_S|_1}$$

Proof Lemma 5.2: • define z by: • $z_j = \mathrm{sign}(\beta_j)$ for $j \in S$

• $z_j = \mathrm{sign}(\hat{\beta}_j - \beta_j)$ for $j \in S^c$

- by construction $z \in \partial |\beta|_1$:

$$A(\hat{\beta}_\lambda) \overset{(1)}{\leq} 2 \underbrace{|X^T\varepsilon|_\infty}_{\leq \lambda/3} |\hat{\beta}_\lambda - \beta|_1 - \lambda \langle z_S, (\hat{\beta}_\lambda - \beta)_S \rangle - \lambda \langle z_{S^c}, (\hat{\beta}_\lambda - \beta)_{S^c} \rangle$$

$\uparrow$

hypothesis of Theorem 5.1.

From the definition of $\mathfrak{Z}$:

$$A(\hat\beta_\lambda) \leqslant \frac{2\lambda}{3}|\hat\beta_\lambda-\beta|_1 + \lambda|(\hat\beta_\lambda-\beta)_S|_1 - \lambda|(\hat\beta_\lambda-\beta)_{S^c}|_1$$

$$= \frac{\lambda}{3}\left(5|(\hat\beta_\lambda-\beta)_S|_1 - |(\hat\beta-\beta)_{S^c}|_1\right) \quad \leftarrow \text{gives (i)}$$

$$\leqslant 2\lambda|(\hat\beta_\lambda-\beta)_S|_1 \quad \leftarrow \text{gives (ii)}.$$

$\square$

We can conclude the proof of Theorem 5.1.

→ if $A(\hat\beta_\lambda) \overset{\text{Al-Kashi}}{=} \|X(\hat\beta_\lambda-\beta^*)\|^2 + \|X(\hat\beta_\lambda-\beta)\|^2 - \|X(\beta-\beta^*)\|^2 \leqslant 0$

then $\|X(\hat\beta_\lambda-\beta^*)\|^2 \leqslant \|X(\beta-\beta^*)\|^2$:-)

→ if $A(\hat\beta_\lambda) > 0$, then $\hat\beta_\lambda - \beta \in \mathcal{C}(\beta)$ from (i), and from ~~(i)~~ ~~and~~ (ii)

$$A(\hat\beta_\lambda) = \|X(\hat\beta_\lambda-\beta^*)\|^2 + \|X(\hat\beta_\lambda-\beta)\|^2 - \|X(\beta-\beta^*)\|^2$$

$$\overset{(ii)}{\leqslant} 2\lambda|(\hat\beta_\lambda-\beta)_S|_1$$

definition of $\kappa(\beta)$ → $\leqslant 2\lambda\frac{\sqrt{|\beta|_0}}{\kappa(\beta)}\|X(\hat\beta_\lambda-\beta)\| \underset{2ab\leqslant a^2+b^2}{\leqslant} \frac{\lambda^2|\beta|_0}{\kappa(\beta)^2} + \|X(\hat\beta_\lambda-\beta)\|^2$

$$\Longrightarrow \|X(\hat\beta_\lambda-\beta^*)\|^2 \leqslant \|X(\beta-\beta^*)\|^2 + \frac{\lambda^2|\beta|_0}{\kappa(\beta)^2} \quad \text{:-)}$$

$\square$

3 Computing the Lasso

a/ Algebraic computation (LARS algorithm)

- $S_\lambda = \text{supp}(\hat\beta_\lambda)$
- optimality: $\exists\, z_\lambda$ such that $\begin{cases} [z_\lambda]_{S_\lambda} = \text{sign}([\hat\beta_\lambda]_{S_\lambda}) \\ \|[z_\lambda]_{S_\lambda^c}\|_\infty \leq 1 \end{cases}$

and

$$X^T X \hat\beta_\lambda = X^T Y - \frac{\lambda}{2} z_\lambda$$

$$\Rightarrow \begin{cases} \text{on } S_\lambda: & X_{S_\lambda}^T X_{S_\lambda} [\hat\beta_\lambda]_{S_\lambda} = X_{S_\lambda}^T Y - \frac{\lambda}{2}\,\text{sign}([\hat\beta_\lambda]_{S_\lambda}) \\ \text{on } S_\lambda^c: & |X_{S_\lambda^c}^T Y - X_{S_\lambda^c}^T X \hat\beta_\lambda|_\infty \leq \frac{2}{\lambda} \end{cases}$$

- Since $\lambda \to S_\lambda$ is piecewise constant, $\lambda \to \hat\beta_\lambda$ is piecewise linear

  ⇝ see Figure 5.2

- LARS algorithm computes algebraically the break points $\hat\beta_{\lambda_1}, \hat\beta_{\lambda_2}, \dots$
  Then, for $\lambda \in (\lambda_{k+1}, \lambda_k)$, $\hat\beta_\lambda$ is computed by linear interpolation.

⚠
- while the computations are of algebraic nature, we do not have explicit formulas for $\hat\beta_\lambda$
- matrix inversions are done up to some precision level

b/ (Accelerated) proximal method

- Consider $\min_{\beta} \{ F(\beta) + \lambda |\beta|_1 \}$ with F convex and smooth

(in our case $F(\beta) = \|Y - X\beta\|^2$)

$$F(\beta) = F(\beta^t) + \langle \nabla F(\beta^t), \beta - \beta^t \rangle + O(\|\beta - \beta^t\|^2)$$

iterate ⟶ $\beta^{t+1} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ F(\beta^t) + \langle \nabla F(\beta^t), \beta - \beta^t \rangle + \frac{1}{2\eta} \|\beta - \beta^t\|^2 + \lambda |\beta|_1 \right\}$

- Why easier?

1) $\beta^{t+1} \in \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2\eta} \|\beta - \beta^t + \eta \nabla F(\beta^t)\|^2 + \underbrace{F(\beta^t) - \frac{\eta}{2} \|\nabla F(\beta^t)\|^2}_{\text{no } \beta \text{ here}} + \lambda |\beta|_1 \right\}$

$$= \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\beta - (\beta^t - \eta \nabla F(\beta^t))\|^2 + \lambda \eta |\beta|_1 \right\}$$

2) we have seen that $S_\lambda(\alpha) := \begin{bmatrix} \alpha_1 (1 - \lambda / |\alpha_1|)_+ \\ \vdots \\ \alpha_p (1 - \lambda / |\alpha_p|)_+ \end{bmatrix}$ is solution

to $\underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\beta - \alpha\|^2 + \lambda |\beta|_1 \right\}$

Soft thresholding operator

$$\Longrightarrow \boxed{\beta^{t+1} = S_{\lambda\eta}(\beta^t - \eta \nabla F(\beta^t))}$$

(here $\nabla F(\beta^t) = 2X^T(X\beta^t - Y)$)

Remarks:

1) What do you recognize for $\lambda = 0$?

2) the scheme converges to $\hat{\beta}_\lambda$ for η small enough

3) acceleration: we can accelerate the convergence with Nesterov's acceleration scheme → FISTA algorithm (see Section 5.2.4)

c/ Others algorithms

- There exists many other algorithms.
- A simple one: coordinate descent algorithm

  $\rightsquigarrow$ To Do: exercise 5.5.7

④ Bias of the Lasso

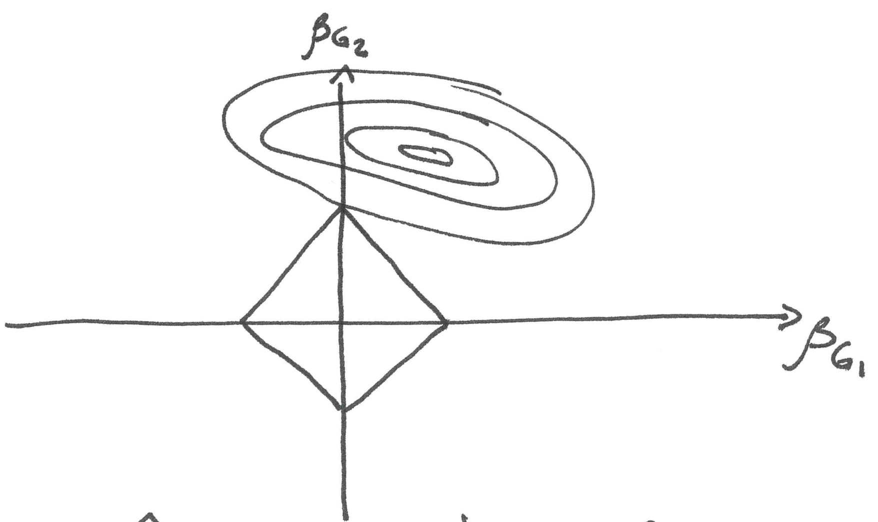
$\longrightarrow$ look at the slides

⑤ Other sparsity structures

Ex: group sparsity

$$\beta^* = \begin{bmatrix} \beta^*_{G_1} \\ \vdots \\ \beta^*_{G_n} \end{bmatrix} \begin{matrix} \}\, G_1 \\ \vdots \\ \}\, G_n \end{matrix} \qquad \text{with a few } \beta^*_{G_k} \neq 0$$

Coordinate sparse	group sparse
$\text{card}\{j : \beta_j \neq 0\}$ small	$\text{card}\{k : \beta_{G_k} \neq 0\}$ small
$\Downarrow$	$\Downarrow$
$\sum_{j=1}^{p} \lvert\beta_j\rvert$ penalisation	$\sum_{k=1}^{M} \lVert\beta_{G_k}\rVert$



Group Lasso: $\hat{\beta}_\lambda \in \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \left\{ \lVert Y - X\beta \rVert^2 + \sum_k \lambda_k \lVert \beta_{G_k} \rVert \right\}$

⇝ again a convex criterion

⇝ usually $\lambda_k = \lambda \sqrt{|G_k|}$

. <u>Exercise</u>: following similar arguments as for $|x|_1$ check that

$$\left\{ \partial\left(\sum_k \lambda_k \|\beta_{G_k}\|\right) = \left\{ z \in \mathbb{R}^p : \begin{array}{ll} \cdot\ z_{G_k} = \lambda_k \dfrac{\beta_{G_k}}{\|\beta_{G_k}\|} & \text{if } \beta_{G_k} \neq 0 \\ \cdot\ \|z_{G_k}\| \leq \lambda_k & \text{if } \beta_{G_k} = 0 \end{array} \right\} \right.$$

6 Take home message

<u>Convex relaxation</u>

→ a principled way to derive practical estimators from model selection

→ good theoretical garantees

→ but suffers from some bias

Model selection is too hard ?

Just relax !