Convexification

Lecture 3

Recap from last lecture

- **Model**: $y_i = \langle \beta^*, x_i \rangle + \varepsilon_i$, $i = 1, \dots, n$

with $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

- **Notation**:

$$Y := \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} ; \quad f^* := \begin{bmatrix} f^*(x_1) \\ \vdots \\ f^*(x_n) \end{bmatrix} ; \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

and $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \in \mathbb{R}^{n \times p}$

$$\Rightarrow \quad Y = X\beta^* + \varepsilon = f^* + \varepsilon.$$

- **Hidden structure**: we assume that

$$\underbrace{|\beta^*|_0 := \mathrm{card}\{j : \beta_j^* \neq 0\} \text{ is small}}_{\text{Coordinate sparse assumption.}}$$

- For $S \subset \{1, \dots, p\}$, we set

$$\bar{S} = \{X\beta : \mathrm{supp}(\beta) \subset S\}$$

and $\hat{f}^{(S)} := X \hat{\beta}^{(S)}$ with

$$\hat{\beta}^{(S)} \in \underset{\mathrm{supp}(\beta) \subset S}{\mathrm{argmin}} \|Y - X\beta\|^2$$

- **Structure learning**:

$$\hat{S} \in \underset{S \subset \{1, \dots, p\}}{\mathrm{argmin}} \{ \|Y - \hat{f}^{(S)}\|^2 + \mathrm{pen}(S)\sigma^2 \} \qquad (\Pi S)$$

with $\mathrm{pen}(S) = \underset{\substack{\uparrow \\ \text{constant } \gtrsim 2}}{K} |S| \log \frac{ep}{|S|}$

fulfills

$$R(\hat{f}^{(\hat{S})}) \lesssim \underbrace{\frac{\sigma^2}{n} |\beta^*|_0 \log \frac{ep}{|\beta^*|_0}}_{\text{minimax optimal}}$$

- **Main issue**:

prohibitive computational complexity.

Solution(s) ?

→ convex proxy for the minimisation problem (ΠS)

⇝ this lecture

→ greedy / iterative approximate minimisation

⇝ next lecture.

Our goal today :

→ explain and discuss the convexification paradigm in the coordinate sparse setting

→ highlights the strengths and weaknesses of this approach.

To avoid normalizing issues, we assume in the following that the columns $X_{:j}$ of X have been normalized $\|X_{:j}\| = 1$.

Lasso estimator

Let us consider the approximate version of (MS)

$$\hat{S} \in \underset{S \subset \{1,\dots,p\}}{\operatorname{argmin}} \left\{ \|Y - \hat{f}^{(S)}\|^2 + \lambda |S| \right\}, \quad \text{with } \lambda = \underset{\text{constant}}{K} \sigma^2 \log p$$

Since $\hat{f}^{(S)} = X \hat{\beta}^{(S)}$, with $\hat{\beta}^{(S)} \in \underset{\beta :\, \mathrm{supp}(\beta) \subset S}{\operatorname{argmin}} \|Y - X\beta\|^2$, we have

$$\hat{S} \in \underset{S \subset \{1..p\}}{\operatorname{argmin}} \ \min_{\beta :\, \mathrm{supp}(\beta) = S} \left\{ \|Y - X\beta\|^2 + \lambda |\beta|_0 \right\}$$

and

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

:)   :(

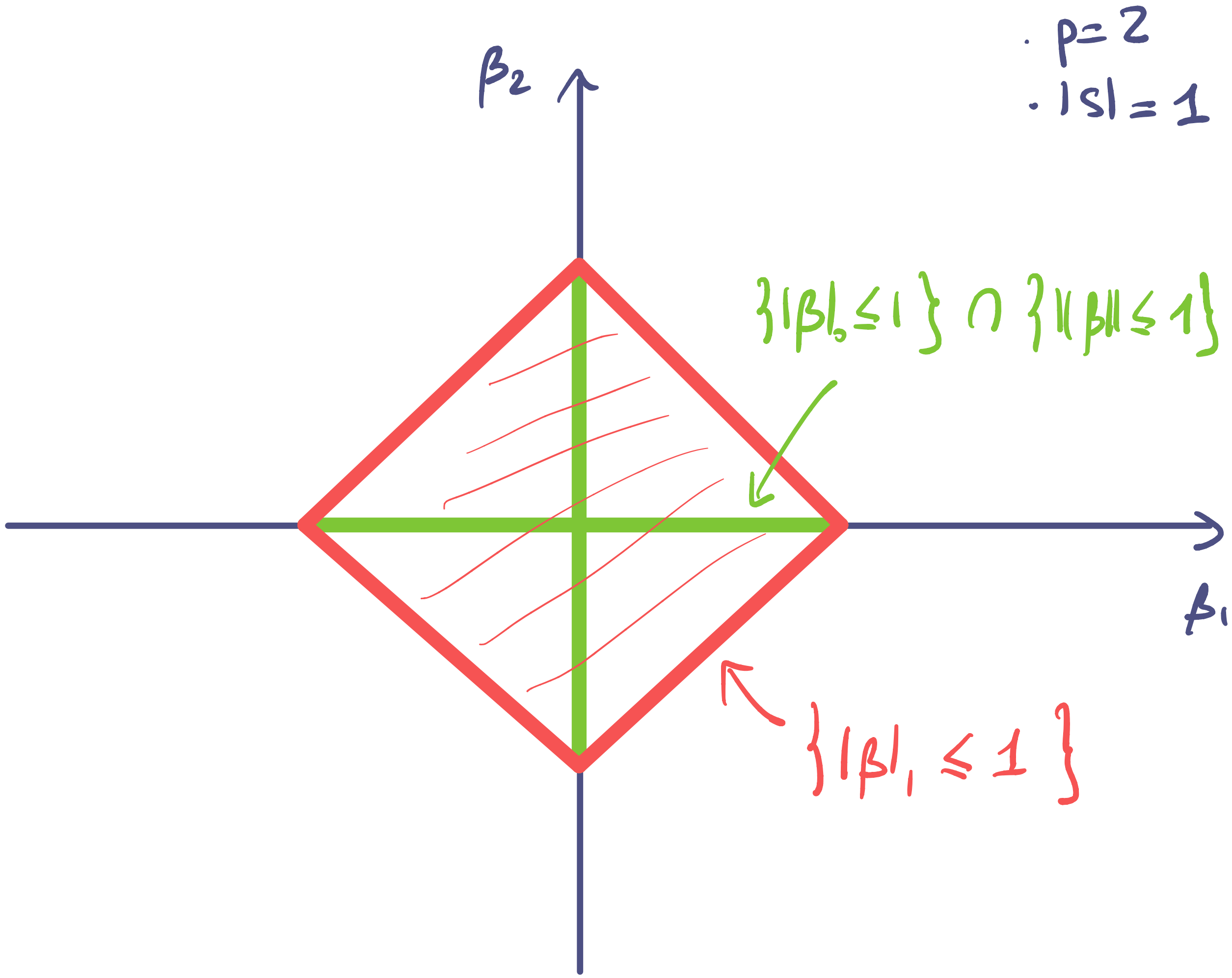
Recipe:

→ constrained version

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

→ convexification

$$|\beta|_0 \leq D \rightsquigarrow |\beta|_1 \leq R$$



Lasso estimator:

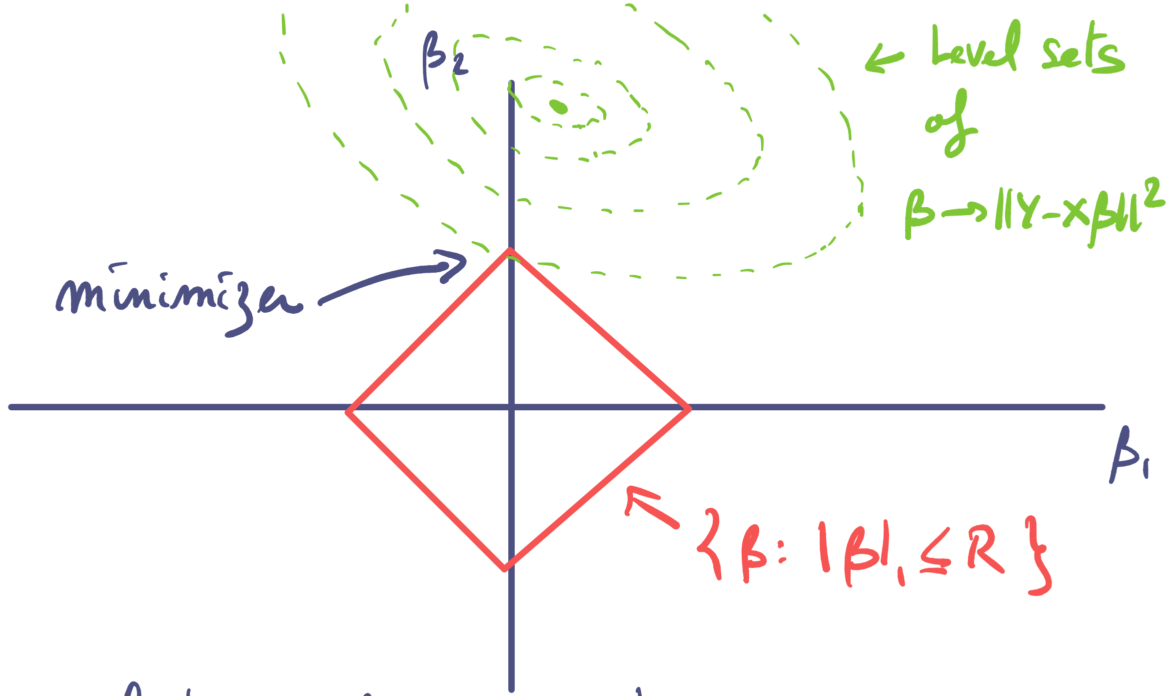
- $\hat{\beta}^{(\lambda)} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \|Y - X\beta\|^2 + \lambda |\beta|_1 \right\}$

- $\hat{f}^{(\lambda)} := X \hat{\beta}^{(\lambda)}$

Geometric interpretation:

constrained version

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



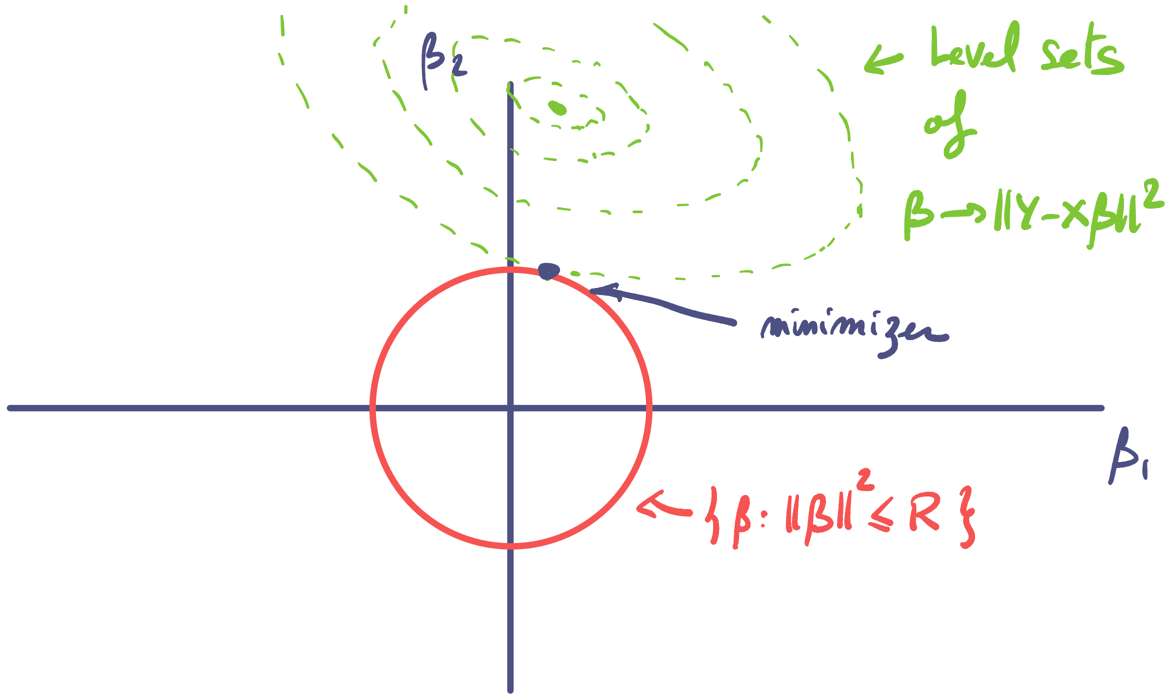
Remark:

singularities of $\{|\beta|_1 \leq R\}$ $\longleftrightarrow$ selection of coordinates

Ridge:

ℓ^1-ball $\rightsquigarrow$ ℓ^2-ball

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



$\rightsquigarrow$ no selection occurs.

Analytic analysis

The objective function

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

is convex but not differentiable where $\beta_j = 0$ for some $j \in \{1, \dots, p\}$

Subdifferential:

$$\partial|\beta|_1 = \left\{ z \in \mathbb{R}^p : \begin{array}{l} \cdot\ z_j = \text{sign}(\beta_j) \text{ if } \beta_j \neq 0 \\ \cdot\ z_j \in [-1, 1] \text{ if } \beta_j = 0 \end{array} \right\}$$

So

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

Since $0 \in \partial \mathcal{L}_\lambda(\hat\beta^{(\lambda)})$, $\exists \hat z \in \partial|\hat\beta^{(\lambda)}|_1$ such that

$$\underbrace{X^T X \hat\beta^{(\lambda)} = X^T Y}_{\Leftrightarrow \text{ least square}} - \underbrace{\frac{\lambda}{2}\hat z}_{\text{selection}}$$

Set $X_{\hat S} := X[\cdot, \hat S]$, where $\hat S = \text{Supp}(\hat\beta^{(\lambda)})$

Then, since $\hat z_{\hat S} = \text{sign}(\hat\beta^{(\lambda)}_{\hat S})$ we have

$$X_{\hat S}^T X_{\hat S} \hat\beta^{(\lambda)}_{\hat S} = X_{\hat S}^T Y - \frac{\lambda}{2}\text{sign}(\hat\beta^{(\lambda)}_{\hat S})$$

so

$$\hat\beta^{(\lambda)}_{\hat S} = \underbrace{(X_{\hat S}^T X_{\hat S})^{-1} X_{\hat S}^T Y}_{= \hat\beta^{(\hat S)} \text{ (least square estimator on } \hat S\text{)}} - \underbrace{\frac{\lambda}{2}(X_{\hat S}^T X_{\hat S})^{-1}\text{sign}(\hat\beta^{(\lambda)}_{\hat S})}_{\text{bias term induced by the } \ell^1\text{-constraint}}$$

Remark: We cannot get an explicit expression for $\hat\beta^{(\lambda)}$, but in the case where X has orthogonal columns:

$$X^T X = I_p.$$

Case $X^T X = I_p$:

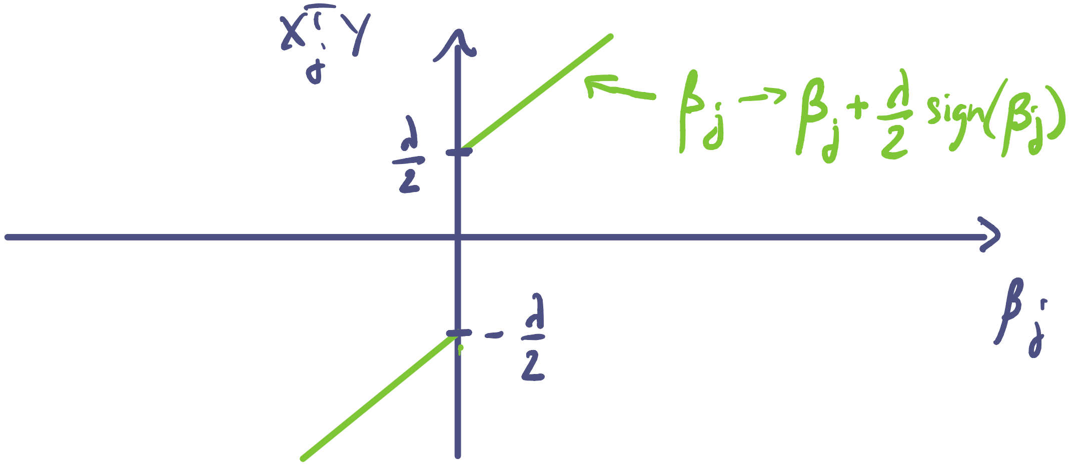
then $\hat{\beta} = X^T Y - \frac{\lambda}{2} \hat{z}$ with $\hat{z} \in \partial |\hat{\beta}|_1$

$\rightarrow$ if $\hat{\beta}_j \neq 0$: then

$$\hat{\beta}_j = x_j^T Y - \frac{\lambda}{2} \operatorname{sign}(\hat{\beta}_j) \quad \text{ie}$$

$$x_j^T Y = \hat{\beta}_j + \frac{\lambda}{2} \operatorname{sign}(\hat{\beta}_j)$$

$\rightsquigarrow$ only possible if $|x_j^T Y| > \lambda/2$



$\rightarrow$ if $\hat{\beta}_j = 0$: then

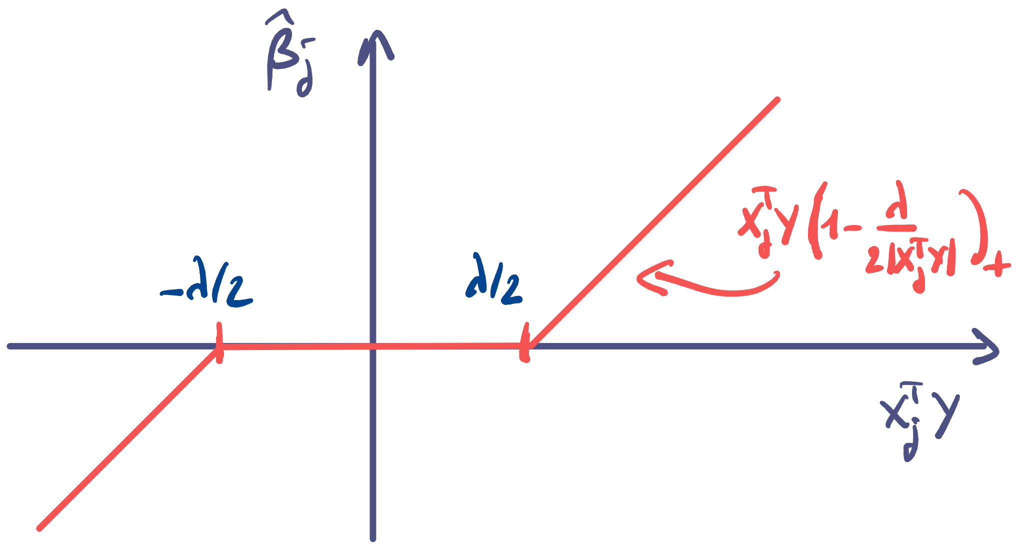
$$\hat{z}_j = \frac{2}{\lambda} x_j^T Y \in [-1, 1]$$

$\rightsquigarrow$ only possible if $|x_j^T Y| \leq \lambda/2$

So

$$\hat{\beta}_j = \begin{cases} 0 & \text{if } |x_j^T Y| \leq \lambda/2 \\ x_j^T Y - \frac{\lambda}{2} \operatorname{sign}(x_j^T Y) & \text{if } |x_j^T Y| > \lambda/2 \end{cases}$$

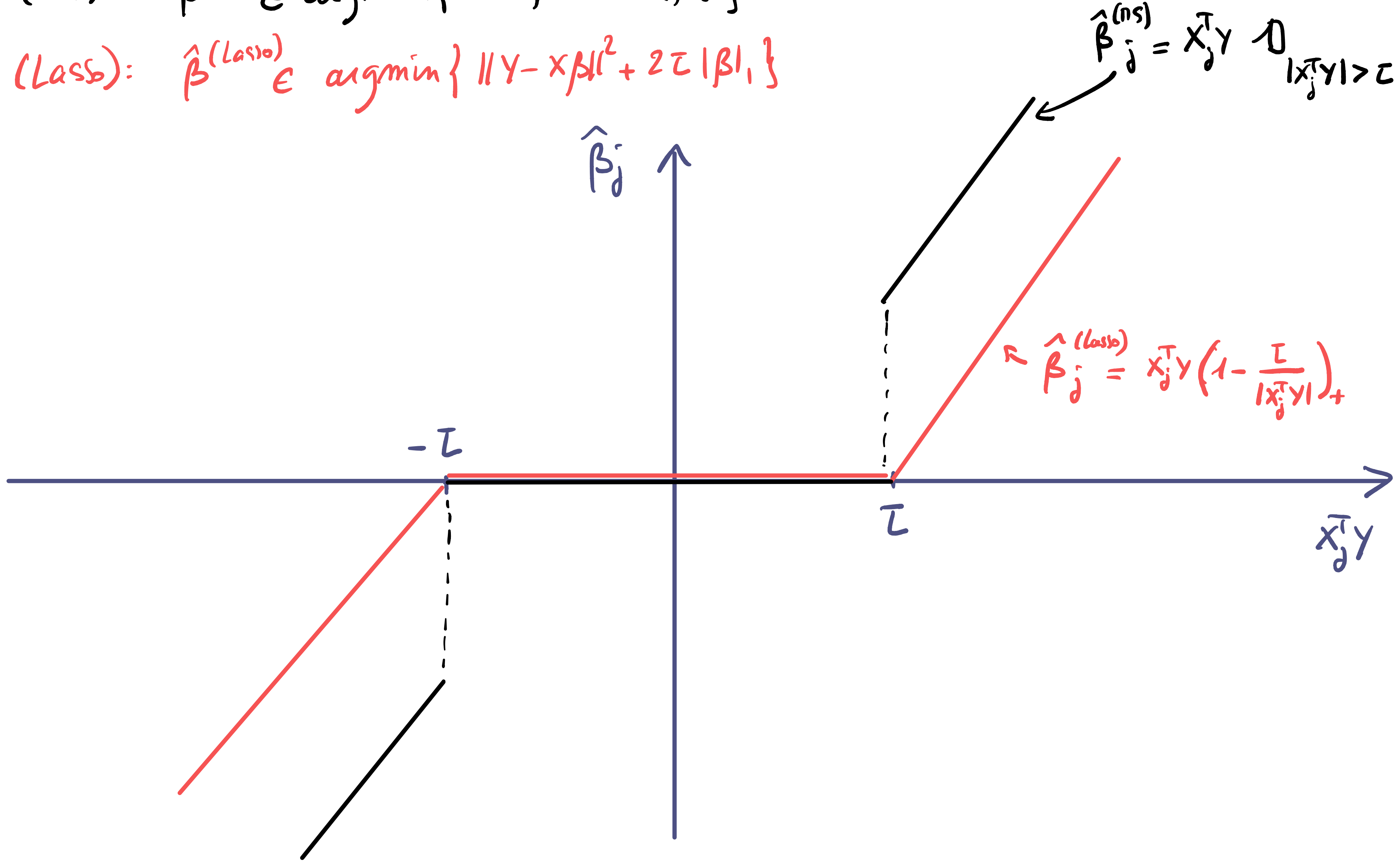
$$= \underbrace{x_j^T Y}_{\text{least square}} \underbrace{\left(1 - \frac{\lambda}{2|x_j^T Y|}\right)_+}_{\text{selects and shrinks}}$$



Comparaison between Lasso and (HS): when $X^TX = I_p$

(HS): $\hat{\beta}^{(HS)} \in \arg\min\{\|Y - X\beta\|^2 + \tau^2 |\beta|_0\}$

(Lasso): $\hat{\beta}^{(Lasso)} \in \arg\min\{\|Y - X\beta\|^2 + 2\tau |\beta|_1\}$



Theoretical garanties

Can we compare the performance of $\hat{\beta}^{(Lasso)}$ to $\hat{\beta}^{(ns)}$?

Compatibility constant:

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

where

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

$\rightsquigarrow$ account for (local) orthogonality

Fact: $\kappa(\beta) \geq \lambda_{min}(X^T X)^{1/2}$

Proof: for any $v \in \mathbb{R}^p$:

$$\|Xv\|^2 \geq \lambda_{min} \|v\|^2 \geq \lambda_{min} \|v_S\|^2$$

$$\underset{C.S.}{\geq} \lambda_{min} \frac{|v_S|_1^2}{|S|^2}$$

$\square$

Theorem: for $\lambda = 3\sigma\sqrt{2\underbrace{K}_{>1}\log p}$,

we have with probability $\geq 1 - \frac{1}{p^{K-1}}$

$$d_n(\hat{f}^{(\lambda)}, f^*)^2 \leq C_K \frac{\sigma^2 |\beta^*|_0 \log p}{n\, \kappa(\beta^*)^2}$$

price to pay for computational tractability

($\frac{1}{\kappa(\beta^*)^2}$ can be huge)

Proof: see Theorem 5.1. $\square$

Bias of Lasso estimators

Example

- We have $n = 60$ noisy observations of $f^*: [0,1) \to \mathbb{R}$

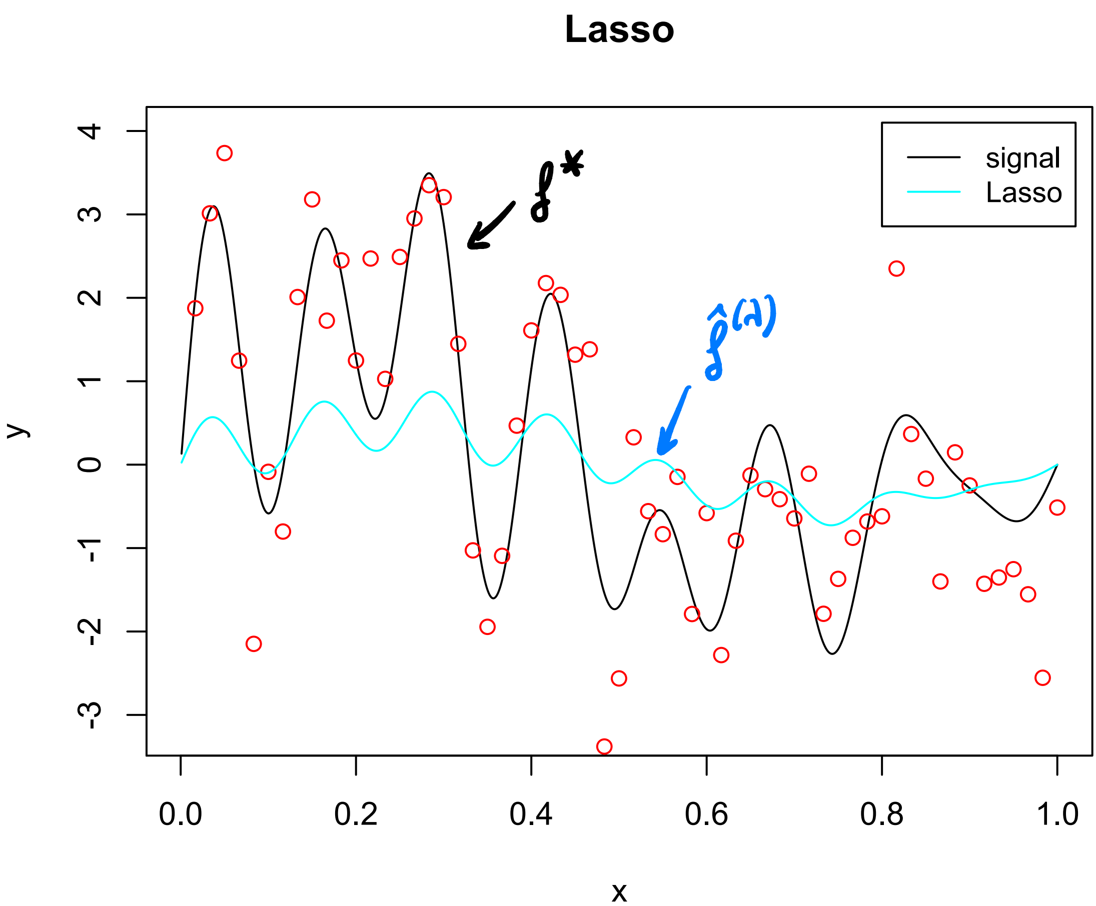
$$y_i = f^*\left(\frac{i}{n}\right) + \varepsilon_i, \qquad i = 1, \dots, n$$

- We expand f^* on the Fourier basis $\{\varphi_j : j \geq 0\}$

$$f^*\left(\frac{i}{n}\right) = \sum_j \beta_j^* \underbrace{\varphi_j(i/n)}_{=: X_{ij}}$$

- We compute $\hat{\beta}^{Lasso}$ and plot the estimator

$$\hat{f}(x) = \sum_j \hat{\beta}_j^{Lasso} \varphi_j(x)$$



Lasso
signal
Lasso
f^*
$\hat{f}(\lambda)$
y
x
4
3
2
1
0
-1
-2
-3
0.0
0.2
0.4
0.6
0.8
1.0

Why?

We have:

$$\hat{\beta}^{(\lambda)} \in \underset{\beta \in \mathbb{R}^p}{\mathrm{argmin}} \; \|Y - X\beta\|^2 + \underbrace{\lambda |\beta|_1}_{\text{promotes small norm solutions}}$$

It can be seen in the formula

$$\hat{\beta}^{(\lambda)}_{\hat{S}} = \underbrace{(X_{\hat{S}}^T X_{\hat{S}})^{-1} X_{\hat{S}}^T Y}_{\text{unbiased Least square estimator on } \hat{S}} - \underbrace{\frac{\lambda}{2} (X_{\hat{S}}^T X_{\hat{S}})^{-1} \mathrm{sign}(\hat{\beta}^{(\lambda)}_{\hat{S}})}_{\text{bias induced by the } \ell^1 \text{ penalty.}}$$

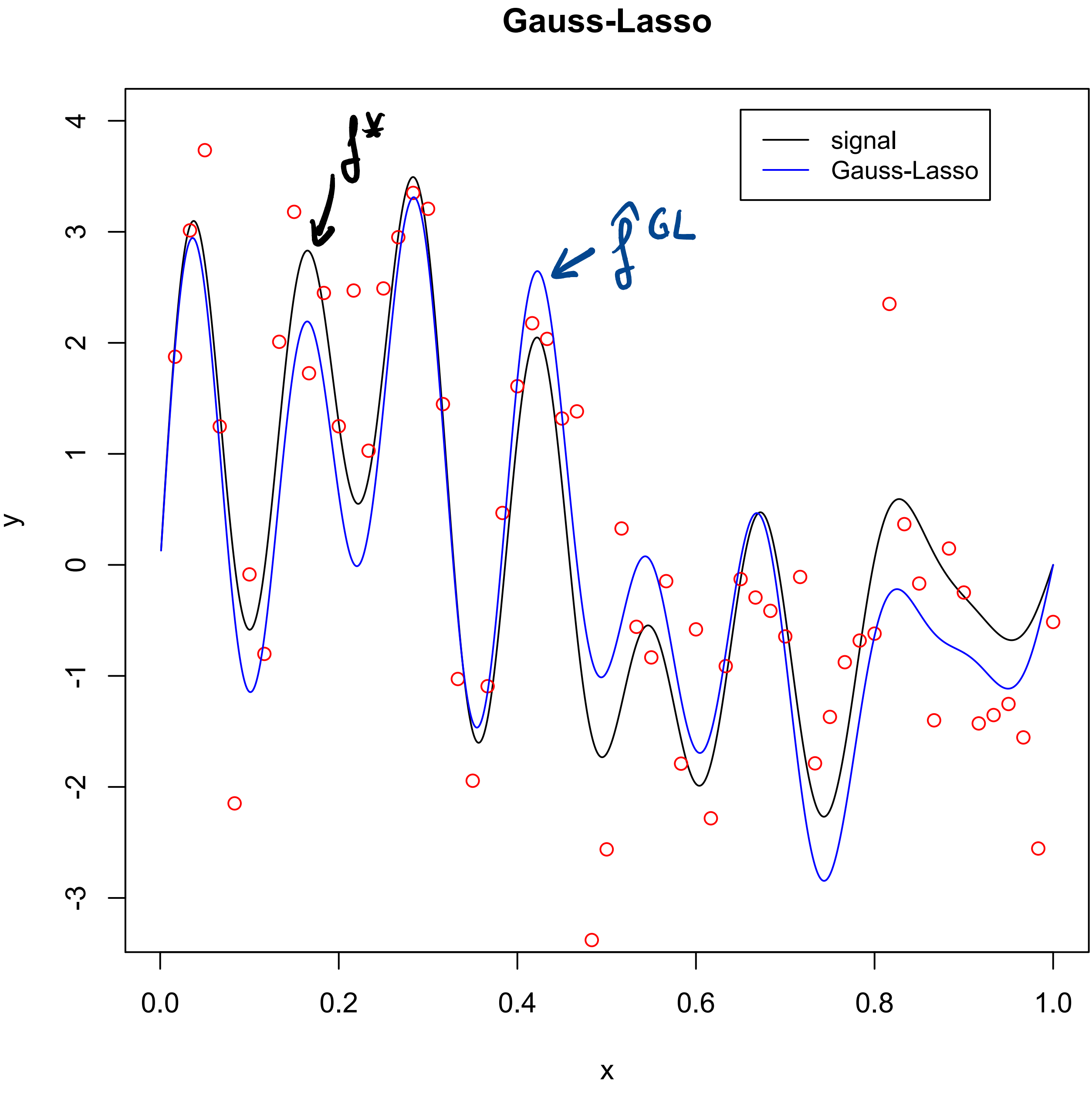
Gauss-Lasso estimator

- Compute $\hat{\beta}^{(\lambda)}$ lasso estimator and set $\hat{S} = \mathrm{supp}(\hat{\beta}^{(\lambda)})$

- Fit the least square estimator on $\hat{S}$:
  - $\hat{\beta}^{GL}_{\hat{S}^c} \equiv 0$
  - $\hat{\beta}^{GL}_{\hat{S}} = (X_{\hat{S}}^T X_{\hat{S}})^{-1} X_{\hat{S}}^T Y$

$\rightsquigarrow$ it removes the shrinkage bias $\frac{\lambda}{2}(X_{\hat{S}}^T X_{\hat{S}})^{-1} \mathrm{sign}(\hat{\beta}^{(\lambda)}_{\hat{S}})$

⚠ For selecting λ, apply cross-validation to $\hat{\beta}^{GL}$, not $\hat{\beta}^{(\lambda)}$.



Gauss-Lasso
signal
Gauss-Lasso
f^*
$\hat{f}^{GL}$
y
x
0.0
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
4

The bias creates a deleterious noise

Ref: W. Su, M. Bogdan, E. Candès
"False discoveries occur early on the Lasso path" (2016)

Setting:

- $X_{:j} \sim \mathcal{N}(0, \frac{1}{n} I_p)$

  (hence $\mathbb{E}[\|X_{:j}\|^2] = 1$)

- $\beta^*_j \overset{iid}{\sim} \alpha\,\delta_0 + (1-\alpha)\,\nu$

  with $\nu(0) = 0$ and $\int x^2 d\nu(x) < +\infty$

  so $S^* := \mathrm{supp}(\beta^*)$ fulfills

  $\mathbb{E}[|S^*|] = \alpha p$

- $n = \delta p$, with $\delta > \alpha$

- $\hat{\beta}^{(\lambda)} \in \underset{\beta}{\mathrm{argmin}} \{\|Y - X\beta\|^2 + \lambda|\beta|_1\}$

  $\hat{S}^{(\lambda)} = \mathrm{support}(\hat{\beta}^{(\lambda)})$

Theorem (informal)

Even if $\sigma = 0$ (no noise),

$$\frac{|\hat{S}^{(\lambda)} \setminus S^*|}{|\hat{S}^{(\lambda)}|} \geq \text{something} > 0$$

with high probability

So, even when there is no noise (!), for any $\lambda > 0$, a positive fraction of the variables selected by the Lasso are not in S^*.

Why? The bias induces a pseudo-noise blurring the residuals

Hand-waving proof: $\sigma^2 = 0$ so that

$$Y = X\beta^* = X_{S^*}\beta^*_{S^*}$$

- We have

$$(X^TX)\hat\beta \overset{(*)}{=} X^TY - \frac{\lambda}{2}\hat z$$

where $\hat z_j = \begin{cases} \mathrm{sign}(\hat\beta_j) & \text{if } \hat\beta_j \neq 0 \\ \in [-1,1] & \text{otherwise} \end{cases}$

From $(*)$, we also have

$$\hat z_j = \frac{2}{\lambda} X_{:j}^T (X\hat\beta - \overbrace{X_{S^*}\beta^*_{S^*}}^{Y})$$

- Let $\tilde\beta_{S^*} = \mathrm{Lasso}(Y, X_{S^*})$

so that

$$X_{S^*}^T X_{S^*}\tilde\beta_{S^*} = X_{S^*}^T X_{S^*}\beta^*_{S^*} - \frac{\lambda}{2}\tilde z_{S^*}$$

i.e

$$\tilde\beta_{S^*} - \beta^*_{S^*} = -\frac{\lambda}{2}(X_{S^*}^T X_{S^*})^{-1}\tilde z_{S^*}$$

hand-waving claim:

if $\left|\frac{2}{\lambda} X_{:j}^T (X_{S^*}\tilde\beta_{S^*} - X_{S^*}\beta^*_{S^*})\right| > 1$

then variable j is selected in $\hat S^{(\lambda)}$.

- for $j \notin S^*$, we have conditionally on X_{S^*}

$$\frac{2}{\lambda} X_{:j}^T X_{S^*}(\tilde\beta_{S^*} - \beta^*_{S^*}) \sim \mathcal{N}(0, v^2)$$

where

$$\begin{aligned} v^2 &= \frac{4}{m\lambda^2}\|X_{S^*}(\tilde\beta_{S^*} - \beta^*_{S^*})\|^2 \\ &\asymp \frac{4}{m\lambda^2}\|\tilde\beta_{S^*} - \beta^*_{S^*}\|^2 \\ &\asymp \frac{1}{m}\|(X_{S^*}^T X_{S^*})^{-1}\tilde z_{S^*}\|^2 \\ &\asymp \frac{1}{m}\|\tilde z_{S^*}\|^2 \\ &\geq \frac{1}{m} \times |S^* \cap \hat S^{(\lambda)}| \\ &= \underbrace{\frac{|S^*|}{m}}_{=\alpha/\delta} \times \frac{|S^* \cap \hat S^{(\lambda)}|}{|S^*|} \end{aligned}$$

$|S^*| = \underbrace{\frac{\alpha}{\delta}}_{<1} m$

Hence $\vartheta^2 \geq$ constant > 0, when $|\hat{S}^{(\lambda)} \cap S^*| \geq$ constant $|S^*|$.

So $\mathbb{P}[\, j$ selected $] \geq$ constant > 0, when $|\hat{S}^{(\lambda)} \cap S^*| \geq$ constant $|S^*|$

and around constant $\times \underbrace{(p - |S^*|)}_{(1-\alpha)p}$ variables $\notin S^*$ are selected,

leading to a non-vanishing False Discovery Proportion.

This mis-selection is due to the non-vanishing bias

$X_{S^*}(\tilde{\beta}_{S^*} - \hat{\beta}_{S^*}) = -\frac{\lambda}{2} X_{S^*}(X_{S^*}^T X_{S^*})^{-1} \tilde{z}_{S^*}$ which is correlated

with the $X_{:j}$, $j \notin S^*$.

Remark: when $|S^*| \leq c \dfrac{n}{\log p}$, the bias is not strong enough

in order to create so many false positives.

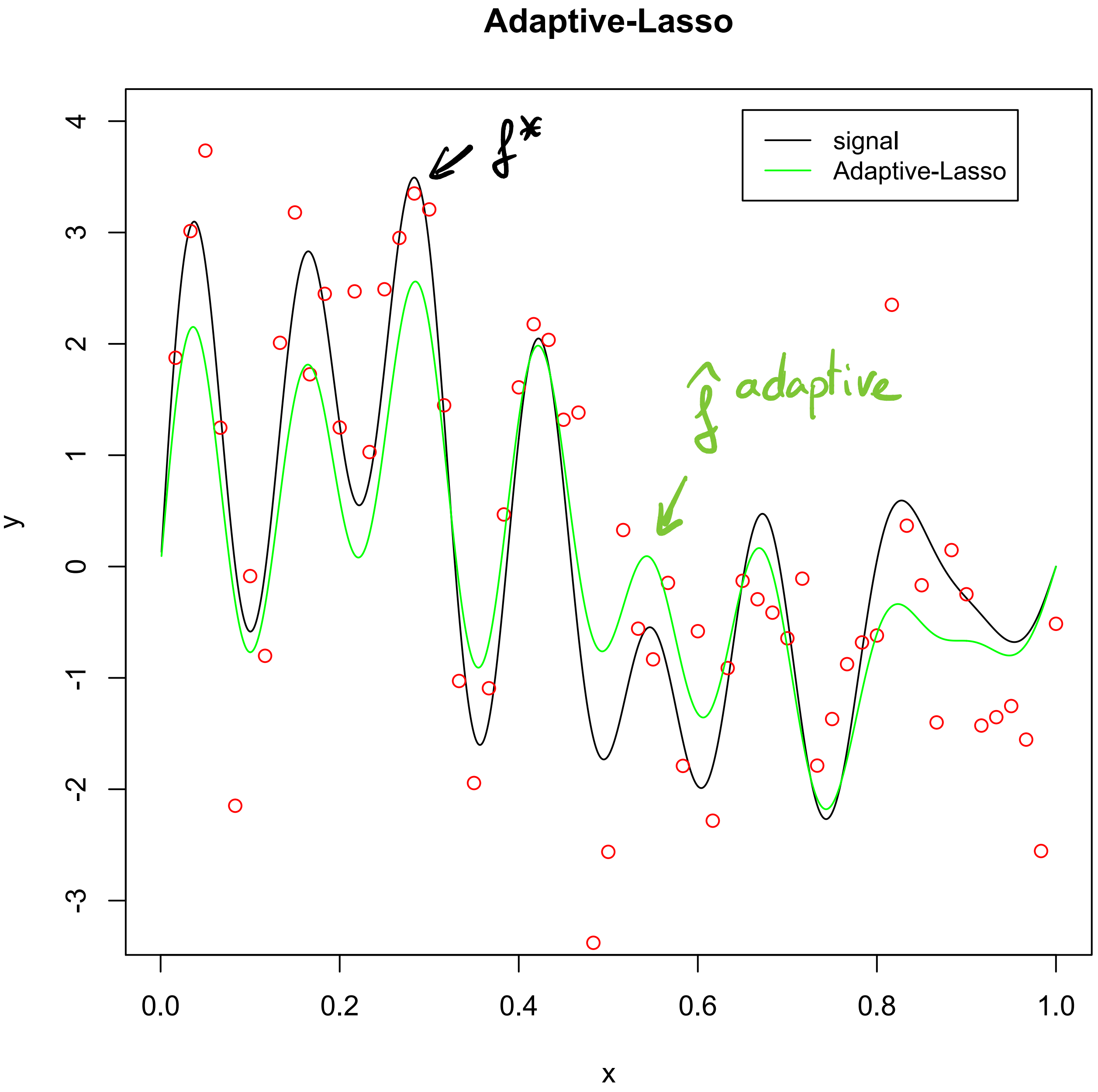
Adaptive-Lasso Estimator

- Take $\hat{\beta}^{init}$ be a first rough estimator of β^*, for example for example a Ridge estimator

for $\beta \approx \hat{\beta}^{init}$, we have $|\beta|_0 \approx \underbrace{\sum_j \frac{|\beta_j|}{|\hat{\beta}_j^{init}|}}_{\text{convex in } \beta}$

$$\hat{\beta}^{adaptive} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \|Y - X\beta\|^2 + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_j^{init}|} \right\}$$

$\rightsquigarrow$ it is still convex and it reduces the bias problems



Adaptive-Lasso
signal
Adaptive-Lasso
f^*
$\hat{f}$ adaptive
y
x
-3
-2
-1
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0