Statistical Learning II

Lecture 10 - LASSO (continued)

Bruno Loureiro
@ CSD, DI-ENS & CNRS

brloureiro@gmail.com

DL3 IASO, Université Paris Dauphine-PSL
27.11.2024

Best subset selection

Idea: encourage solutions which are sparse.

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i\rangle\right)^2 + \lambda ||\boldsymbol{\theta}||_0$$

where $||\cdot||_0 : \mathbb{R}^d \to \{0,1,\ldots,d\}$ is the ℓ_0-“norm”:

Strictly not a norm

$$||\boldsymbol{\theta}||_0 = \sum_{j=1}^{d} \mathbb{I}(\theta_j \neq 0) = \text{ \# non-zero entries}$$

Hence, $\lambda \geq 0$ controls the desired sparsity level

- Large $\lambda \gg 1$: encourage more sparsity
- Small $\lambda \ll 1$: encourage less sparsity

LASSO

The Least Absolute Shrinkage and Selection Operator (LASSO) is defined as the solution of the following problem:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle \right)^2 + \lambda ||\boldsymbol{\theta}||_1$$

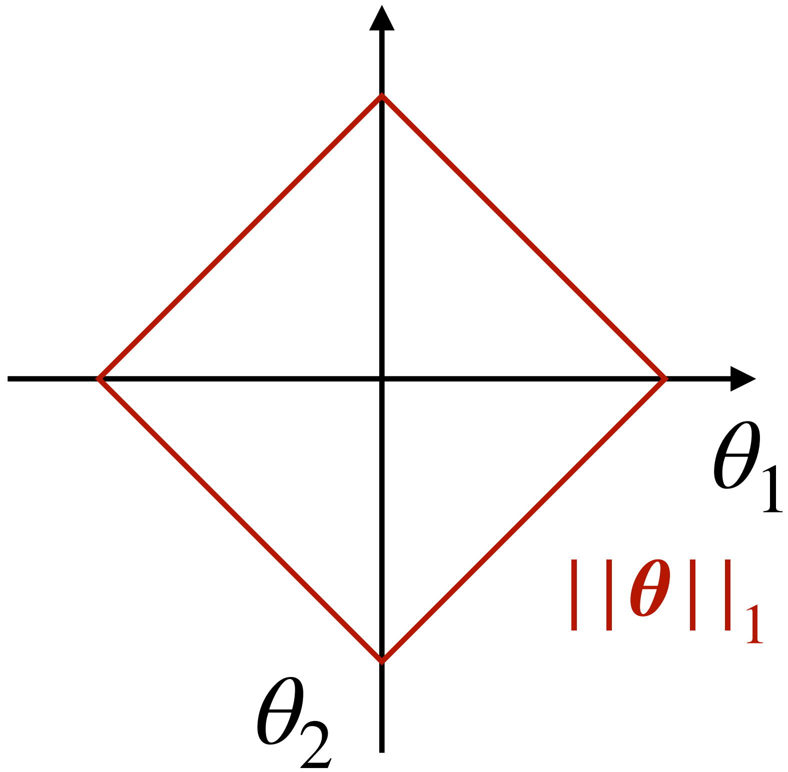
where $||\cdot||_1 : \mathbb{R}^d \to \mathbb{R}_+$ is the ℓ_1-norm:

$$||\boldsymbol{\theta}||_1 = \sum_{j=1}^{d} |\theta_j|$$

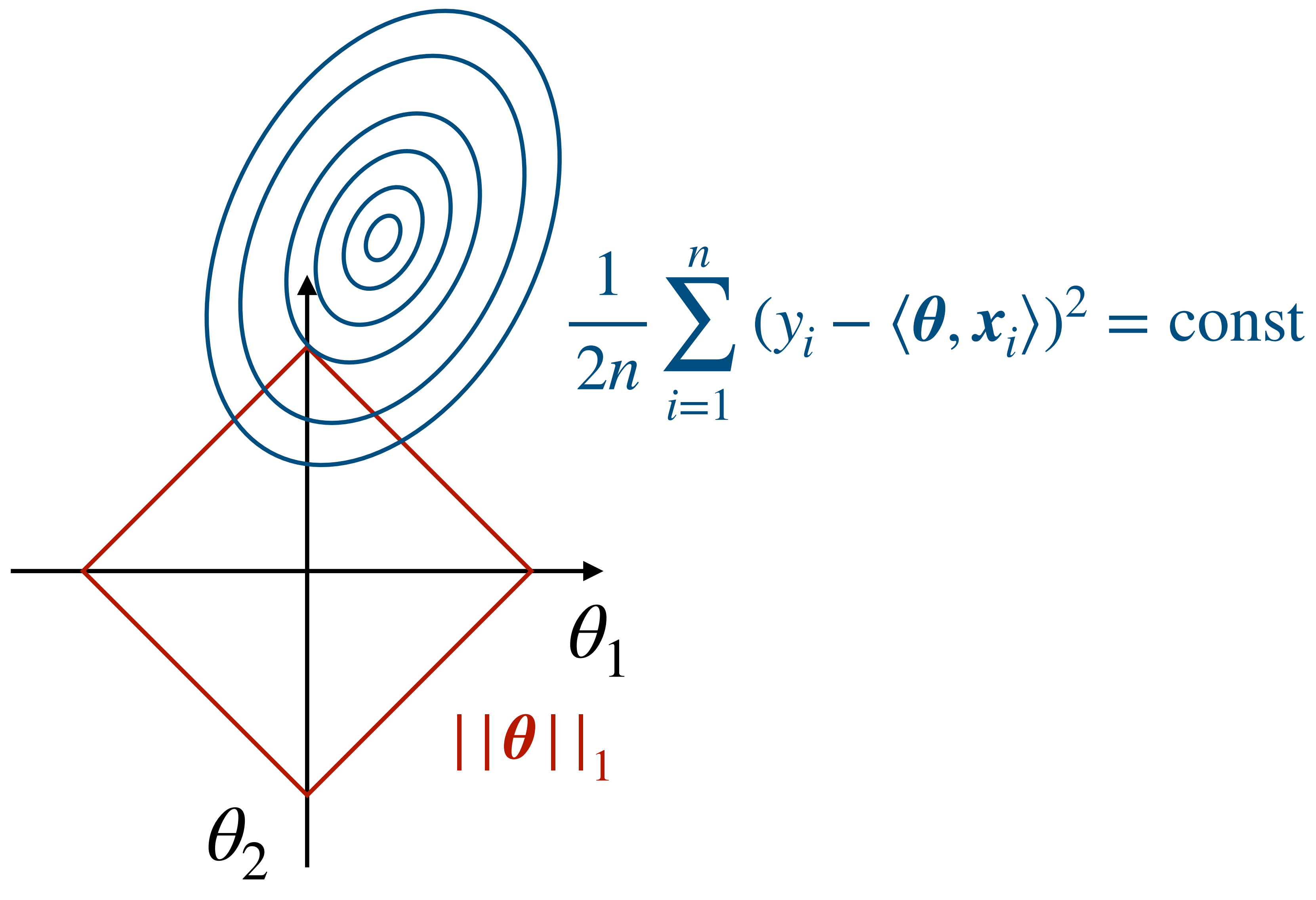
Moreover, this is a convex problem.

Note that both $||\cdot||_1$ and $||\cdot||_2$ are small for sparse vectors… why this is different?

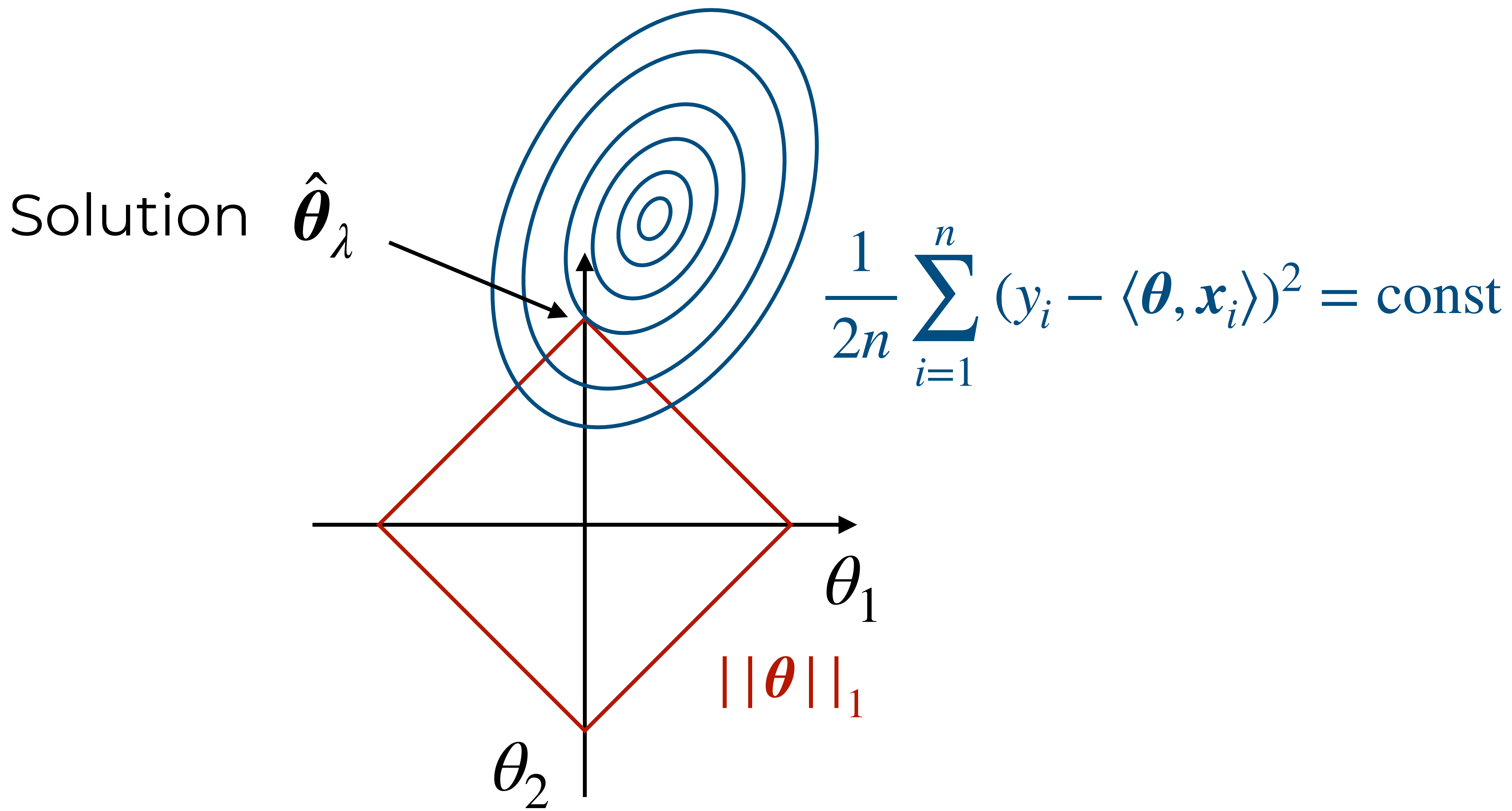
LASSO: visualisation



LASSO: visualisation



LASSO: visualisation



Sharper corners favours sparser solutions!

LASSO: orthogonal covariates

Again, we can get intuition by looking at the orthogonal covariate case:

$$\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}_d \qquad (n \geq d)$$

LASSO: orthogonal covariates

Again, we can get intuition by looking at the orthogonal covariate case:

$$\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}_d \qquad\qquad (n \geq d)$$

Following exactly the same steps from before, in this case we need to solve the following coordinate wise problem:

$$\min_{\theta_j \in \mathbb{R}} L(\theta_j) := \left\{ \frac{1}{2n}(z_j - \theta_j)^2 + \lambda |\theta_j| \right\}$$

LASSO: orthogonal covariates

Again, we can get intuition by looking at the orthogonal covariate case:

$$\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}_d \qquad (n \geq d)$$

Following exactly the same steps from before, in this case we need to solve the following coordinate wise problem:

$$\min_{\theta_j \in \mathbb{R}} L(\theta_j) := \left\{ \frac{1}{2n}(z_j - \theta_j)^2 + \lambda |\theta_j| \right\}$$

As before, we note that: $$L(\theta_j) = \begin{cases} \frac{1}{2n}(z_j - \theta_j)^2 + \lambda\theta_j & \text{for } \theta_\text{j} > 0 \quad \text{(a)} \\ \frac{z_j^2}{2n} & \text{for } \theta_\text{j} = 0 \quad \text{(b)} \\ \frac{1}{2n}(z_j - \theta_j)^2 - \lambda\theta_j & \text{for } \theta_\text{j} < 0 \quad \text{(c)} \end{cases}$$

LASSO: orthogonal covariates

$$L(\theta_j) = \begin{cases} \frac{1}{2n}(z_j - \theta_j)^2 + \lambda\theta_j & \text{for } \theta_j > 0 \quad \text{(a)} \\ \frac{z_j^2}{2n} & \text{for } \theta_j = 0 \quad \text{(b)} \\ \frac{1}{2n}(z_j - \theta_j)^2 - \lambda\theta_j & \text{for } \theta_j < 0 \quad \text{(c)} \end{cases}$$

In case (a), solution is: $\theta_j = z_j - n\lambda$ valid for $z_j > n\lambda$

LASSO: orthogonal covariates

$$L(\theta_j) = \begin{cases} \frac{1}{2n}(z_j - \theta_j)^2 + \lambda\theta_j & \text{for } \theta_\text{j} > 0 \quad \text{(a)} \\ \frac{z_j^2}{2n} & \text{for } \theta_\text{j} = 0 \quad \text{(b)} \\ \frac{1}{2n}(z_j - \theta_j)^2 - \lambda\theta_j & \text{for } \theta_\text{j} < 0 \quad \text{(c)} \end{cases}$$

In case (a), solution is: $\theta_j = z_j - n\lambda$ valid for $z_j > n\lambda$

In case (b), solution is: $\theta_j = 0$

LASSO: orthogonal covariates

$$L(\theta_j) = \begin{cases} \frac{1}{2n}(z_j - \theta_j)^2 + \lambda\theta_j & \text{for } \theta_\text{j} > 0 \quad \text{(a)} \\ \frac{z_j^2}{2n} & \text{for } \theta_\text{j} = 0 \quad \text{(b)} \\ \frac{1}{2n}(z_j - \theta_j)^2 - \lambda\theta_j & \text{for } \theta_\text{j} < 0 \quad \text{(c)} \end{cases}$$

In case (a), solution is: $\theta_j = z_j - n\lambda$ valid for $z_j > n\lambda$

In case (b), solution is: $\theta_j = 0$

In case (c), solution is: $\theta_j = z_j + n\lambda$ valid for $z_j > -n\lambda$

LASSO: orthogonal covariates

$$L(\theta_j) = \begin{cases} \frac{1}{2n}(z_j - \theta_j)^2 + \lambda\theta_j & \text{for } \theta_j > 0 \quad \text{(a)} \\ \frac{z_j^2}{2n} & \text{for } \theta_j = 0 \quad \text{(b)} \\ \frac{1}{2n}(z_j - \theta_j)^2 - \lambda\theta_j & \text{for } \theta_j < 0 \quad \text{(c)} \end{cases}$$

In case (a), solution is: $\theta_j = z_j - n\lambda$ valid for $z_j > n\lambda$

In case (b), solution is: $\theta_j = 0$

In case (c), solution is: $\theta_j = z_j + n\lambda$ valid for $z_j > -n\lambda$

Putting together:

$$\theta_j = \begin{cases} z_j - \text{sign}(z_j)n\lambda & \text{for } |z_j| > \lambda \\ 0 & \text{for } |z_j| \in [-\lambda, \lambda] \end{cases}$$

Soft-thresholding function

LASSO: orthogonal covariates

Putting together, the solution of the LASSO problem:

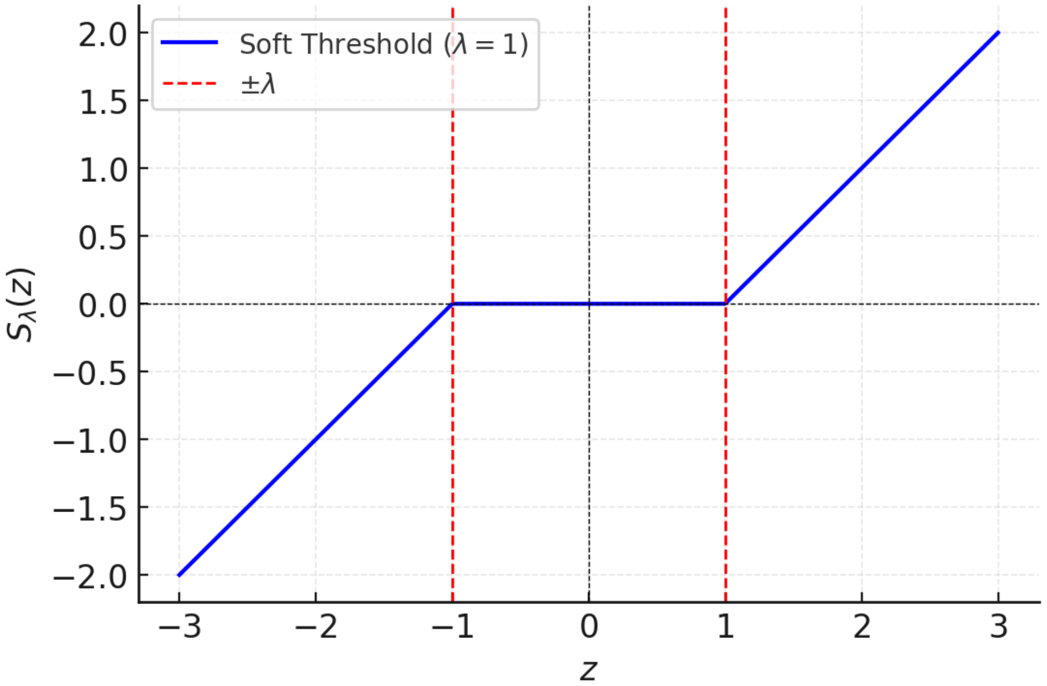
$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i\rangle\right)^2 + \lambda ||\boldsymbol{\theta}||_1$$

Under the assumption of $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}_d$ is given by:

$$\hat{\boldsymbol{\theta}}_\lambda = S_{n\lambda}(\boldsymbol{X}^\top \boldsymbol{y})$$

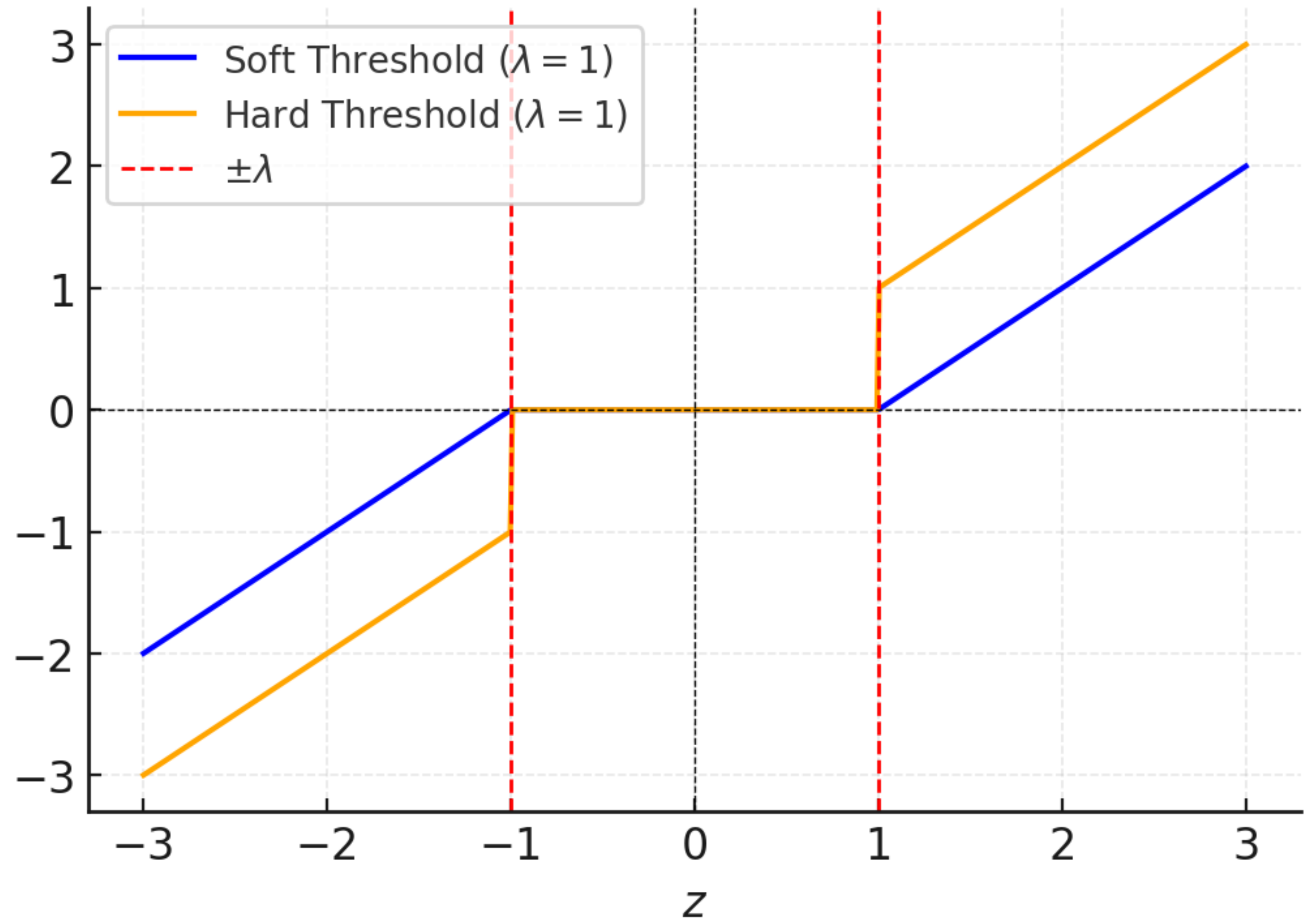
Where:

$$S_\lambda(z) = \begin{cases} z - \text{sign}(z)\lambda & \text{if } |z| > \lambda \\ 0 & \text{if } |z| < \lambda \end{cases}$$



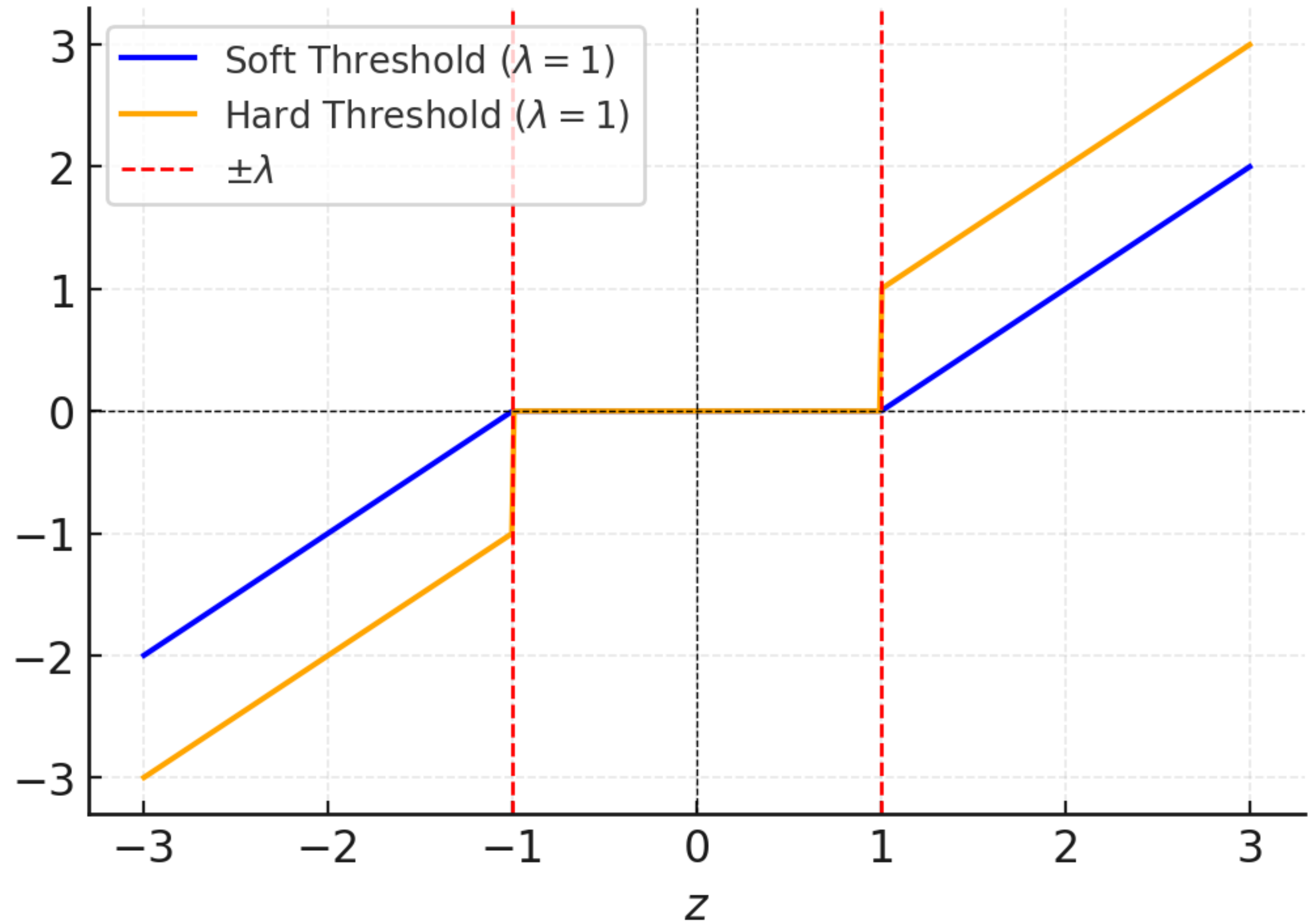
BSS vs. LASSO

It is instructive to compare the BSS and LASSO solutions in the orthogonal covariate case



BSS vs. LASSO

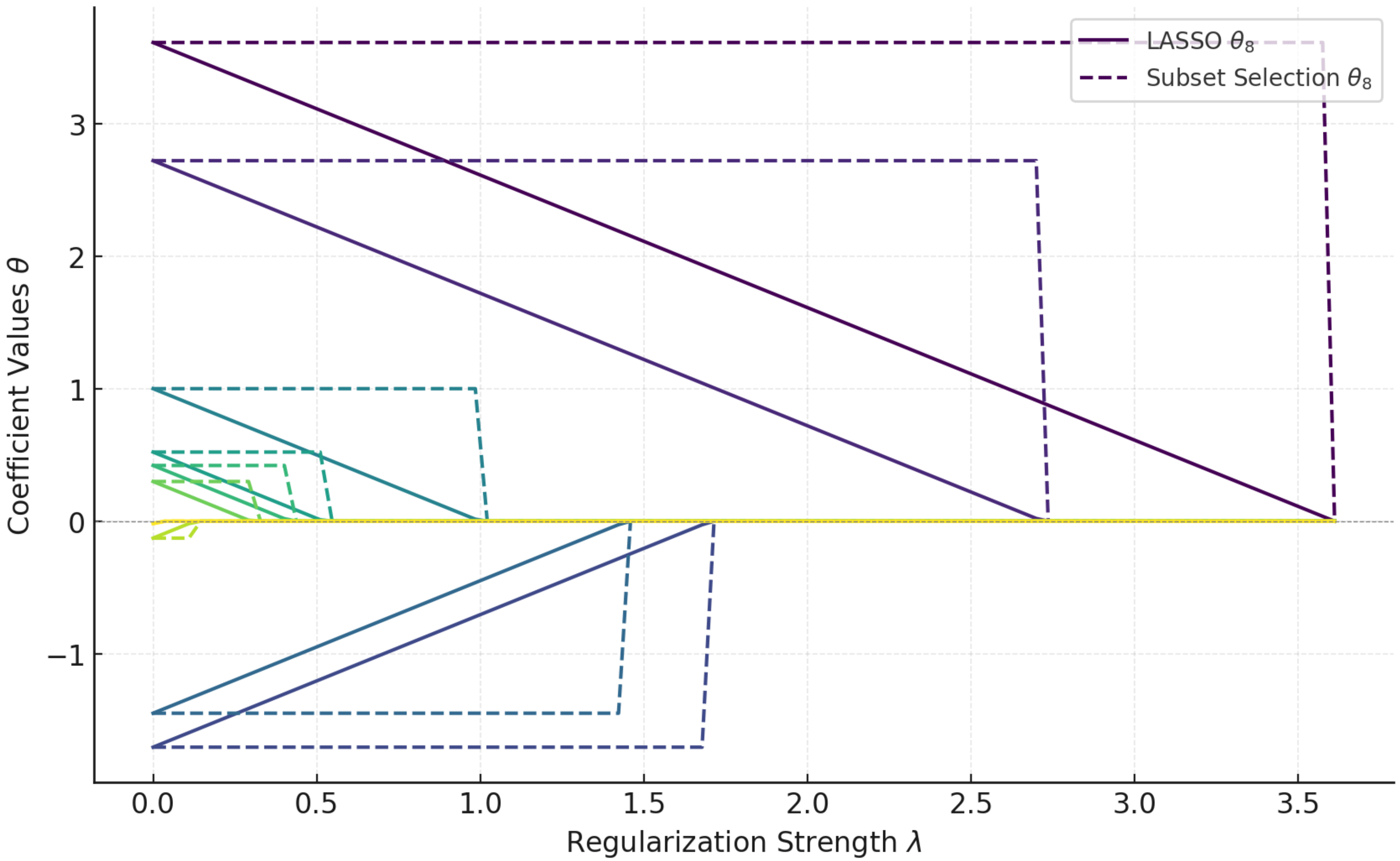
It is instructive to compare the BSS and LASSO solutions in the orthogonal covariate case



- Key similarity: both solutions induce sparsity
- Key differences: LASSO is convex and induce shrinkage (e.g. $z - \lambda$ for $z > \lambda$)

BSS vs. LASSO

$n = 20$ $d = 10$ $y_i = \langle \boldsymbol{\theta}_\star, \boldsymbol{x}_i \rangle + \varepsilon_i$ $\varepsilon_i \sim \mathcal{N}(0,1)$ $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}_{10}$, $\boldsymbol{\theta}_\star$ is 5-sparse



- BSS is discontinuous
- LASSO is piece-wise continuous

For general design, non-zero path not simply a line

LASSO: beyond orthogonal

Again, when the covariates are not orthogonal, an explicit solution for the LASSO is not available. Nevertheless, we can partially characterise it.

LASSO: beyond orthogonal

Again, when the covariates are not orthogonal, an explicit solution for the LASSO is not available. Nevertheless, we can partially characterise it.

Let $S = \{j \in [d] : \hat{\theta}_{\lambda,j} \neq 0\}$ denote the support of the LASSO solution

For this to be well-defined, assume $\hat{\boldsymbol{\theta}}_\lambda$ unique

LASSO: beyond orthogonal

Again, when the covariates are not orthogonal, an explicit solution for the LASSO is not available. Nevertheless, we can partially characterise it.

Let $S = \{j \in [d] : \hat{\theta}_{\lambda,j} \neq 0\}$ denote the support of the LASSO solution

For this to be well-defined, assume $\hat{\boldsymbol{\theta}}_\lambda$ unique

Denote:

- $\hat{\boldsymbol{\theta}}_S \in \mathbb{R}^{|S|}$ the non-zero entries of $\hat{\boldsymbol{\theta}}_\lambda \in \mathbb{R}^d$
- $\boldsymbol{X}_S \in \mathbb{R}^{n \times |S|}$ the corresponding covariates
- $\boldsymbol{s}_S = \mathrm{sign}(\hat{\boldsymbol{\theta}}_\lambda) \in \{-1, +1\}^{|S|}$ the signs.

LASSO: beyond orthogonal

Again, when the covariates are not orthogonal, an explicit solution for the LASSO is not available. Nevertheless, we can partially characterise it.

Let $S = \{j \in [d] : \hat{\theta}_{\lambda,j} \neq 0\}$ denote the support of the LASSO solution

For this to be well-defined, assume $\hat{\boldsymbol{\theta}}_\lambda$ unique

Denote:

- $\hat{\boldsymbol{\theta}}_S \in \mathbb{R}^{|S|}$ the non-zero entries of $\hat{\boldsymbol{\theta}}_\lambda \in \mathbb{R}^d$
- $\boldsymbol{X}_S \in \mathbb{R}^{n \times |S|}$ the corresponding covariates
- $\boldsymbol{s}_S = \mathrm{sign}(\hat{\boldsymbol{\theta}}_\lambda) \in \{-1, +1\}^{|S|}$ the signs.

Then, the by the optimality condition $\hat{\boldsymbol{\theta}}_S \in \mathbb{R}^{|S|}$ satisfies:

$$\boldsymbol{X}_S^\top(\boldsymbol{y} - \boldsymbol{X}_S \hat{\boldsymbol{\theta}}_S) = n\lambda \boldsymbol{s}_S$$

LASSO: beyond orthogonal

Then, the by the optimality condition $\hat{\boldsymbol{\theta}}_S \in \mathbb{R}^{|S|}$ satisfies:

$$\boldsymbol{X}_S^\top(\boldsymbol{y} - \boldsymbol{X}_S\hat{\boldsymbol{\theta}}_S) = n\lambda \boldsymbol{s}_S$$

Therefore, the LASSO solution satisfies:

$$\hat{\boldsymbol{\theta}}_S = \left(\boldsymbol{X}_S^\top \boldsymbol{X}_S\right)^{-1}\left(\boldsymbol{X}_S^\top \boldsymbol{y} - n\lambda \boldsymbol{s}_S\right) \qquad \hat{\boldsymbol{\theta}}_{-S} = \mathbf{0}_{d-|S|}$$

It can be shown uniqueness imply $\boldsymbol{X}_S^\top \boldsymbol{X}_S \succ 0$

LASSO: beyond orthogonal

Then, the by the optimality condition $\hat{\boldsymbol{\theta}}_S \in \mathbb{R}^{|S|}$ satisfies:

$$\boldsymbol{X}_S^\top(\boldsymbol{y} - \boldsymbol{X}_S\hat{\boldsymbol{\theta}}_S) = n\lambda \boldsymbol{s}_S$$

Therefore, the LASSO solution satisfies:

$$\hat{\boldsymbol{\theta}}_S = \left(\boldsymbol{X}_S^\top \boldsymbol{X}_S\right)^{-1}\left(\boldsymbol{X}_S^\top \boldsymbol{y} - n\lambda \boldsymbol{s}_S\right) \qquad \hat{\boldsymbol{\theta}}_{-S} = \boldsymbol{0}_{d-|S|}$$

OLS Shrinkage

⚠ It can be shown uniqueness imply $\boldsymbol{X}_S^\top \boldsymbol{X}_S \succ 0$

LASSO: beyond orthogonal

Then, the by the optimality condition $\hat{\boldsymbol{\theta}}_S \in \mathbb{R}^{|S|}$ satisfies:

$$\boldsymbol{X}_S^\top(\boldsymbol{y} - \boldsymbol{X}_S\hat{\boldsymbol{\theta}}_S) = n\lambda \boldsymbol{s}_S$$

Therefore, the LASSO solution satisfies:

$$\hat{\boldsymbol{\theta}}_S = \left(\boldsymbol{X}_S^\top \boldsymbol{X}_S\right)^{-1}\left(\boldsymbol{X}_S^\top \boldsymbol{y} - n\lambda \boldsymbol{s}_S\right) \qquad \hat{\boldsymbol{\theta}}_{-S} = \boldsymbol{0}_{d-|S|}$$

OLS Shrinkage

It can be shown uniqueness imply $\boldsymbol{X}_S^\top \boldsymbol{X}_S \succ 0$

In particular, note that:

$$||\hat{\boldsymbol{\theta}}_\lambda||_1 = \boldsymbol{s}_S^\top \hat{\boldsymbol{\theta}}_S$$

LASSO: beyond orthogonal

Then, the by the optimality condition $\hat{\boldsymbol{\theta}}_S \in \mathbb{R}^{|S|}$ satisfies:

$$\boldsymbol{X}_S^\top(\boldsymbol{y} - \boldsymbol{X}_S\hat{\boldsymbol{\theta}}_S) = n\lambda \boldsymbol{s}_S$$

Therefore, the LASSO solution satisfies:

$$\hat{\boldsymbol{\theta}}_S = \left(\boldsymbol{X}_S^\top \boldsymbol{X}_S\right)^{-1}\left(\boldsymbol{X}_S^\top \boldsymbol{y} - n\lambda \boldsymbol{s}_S\right) \qquad \hat{\boldsymbol{\theta}}_{-S} = \boldsymbol{0}_{d-|S|}$$

OLS Shrinkage

It can be shown uniqueness imply $\boldsymbol{X}_S^\top \boldsymbol{X}_S \succ 0$

In particular, note that:

$$||\hat{\boldsymbol{\theta}}_\lambda||_1 = \boldsymbol{s}_S^\top \hat{\boldsymbol{\theta}}_S = \boldsymbol{s}_S^\top\left(\boldsymbol{X}_S^\top \boldsymbol{X}_S\right)^{-1}\boldsymbol{X}_S^\top \boldsymbol{y} - n\lambda \boldsymbol{s}_S^\top\left(\boldsymbol{X}_S^\top \boldsymbol{X}_S\right)^{-1}\boldsymbol{s}_S$$

LASSO: beyond orthogonal

Then, the by the optimality condition $\hat{\boldsymbol{\theta}}_S \in \mathbb{R}^{|S|}$ satisfies:

$$\boldsymbol{X}_S^\top(\boldsymbol{y} - \boldsymbol{X}_S\hat{\boldsymbol{\theta}}_S) = n\lambda \boldsymbol{s}_S$$

Therefore, the LASSO solution satisfies:

$$\hat{\boldsymbol{\theta}}_S = \left(\boldsymbol{X}_S^\top \boldsymbol{X}_S\right)^{-1}\left(\boldsymbol{X}_S^\top \boldsymbol{y} - n\lambda \boldsymbol{s}_S\right) \qquad \hat{\boldsymbol{\theta}}_{-S} = \boldsymbol{0}_{d-|S|}$$

OLS Shrinkage

⚠ It can be shown uniqueness imply $\boldsymbol{X}_S^\top \boldsymbol{X}_S \succ 0$

In particular, note that:

$$\begin{aligned} ||\hat{\boldsymbol{\theta}}_\lambda||_1 &= \boldsymbol{s}_S^\top \hat{\boldsymbol{\theta}}_S = \boldsymbol{s}_S^\top \left(\boldsymbol{X}_S^\top \boldsymbol{X}_S\right)^{-1} \boldsymbol{X}_S^\top \boldsymbol{y} - n\lambda \boldsymbol{s}_S^\top \left(\boldsymbol{X}_S^\top \boldsymbol{X}_S\right)^{-1} \boldsymbol{s}_S \\ &< ||\left(\boldsymbol{X}_S^\top \boldsymbol{X}_S\right)^{-1} \boldsymbol{X}_S^\top \boldsymbol{y}||_1 \end{aligned}$$

LASSO: beyond orthogonal

Then, the by the optimality condition $\hat{\boldsymbol{\theta}}_S \in \mathbb{R}^{|S|}$ satisfies:

$$\boldsymbol{X}_S^\top(\boldsymbol{y} - \boldsymbol{X}_S\hat{\boldsymbol{\theta}}_S) = n\lambda \boldsymbol{s}_S$$

Therefore, the LASSO solution satisfies:

$$\hat{\boldsymbol{\theta}}_S = \left(\boldsymbol{X}_S^\top \boldsymbol{X}_S\right)^{-1} \left(\boldsymbol{X}_S^\top \boldsymbol{y} - n\lambda \boldsymbol{s}_S\right) \qquad \hat{\boldsymbol{\theta}}_{-S} = \boldsymbol{0}_{d-|S|}$$

OLS Shrinkage

⚠ It can be shown uniqueness imply $\boldsymbol{X}_S^\top \boldsymbol{X}_S \succ 0$

In particular, note that:

$$||\hat{\boldsymbol{\theta}}_\lambda||_1 = \boldsymbol{s}_S^\top \hat{\boldsymbol{\theta}}_S = \boldsymbol{s}_S^\top \left(\boldsymbol{X}_S^\top \boldsymbol{X}_S\right)^{-1} \boldsymbol{X}_S^\top \boldsymbol{y} - n\lambda \boldsymbol{s}_S^\top \left(\boldsymbol{X}_S^\top \boldsymbol{X}_S\right)^{-1} \boldsymbol{s}_S$$

$$< ||\left(\boldsymbol{X}_S^\top \boldsymbol{X}_S\right)^{-1} \boldsymbol{X}_S^\top \boldsymbol{y}||_1$$

$||\hat{\boldsymbol{\theta}}_{\text{LASSO}}||_1 \leq ||\hat{\boldsymbol{\theta}}_{\text{OLS}}||_1$!!!

LASSO in practice

Beyond the orthogonal case, the LASSO problem:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle\right)^2 + \lambda ||\boldsymbol{\theta}||_1$$

does not admit an explicit solution. How do we do in practice?

LASSO in practice

Beyond the orthogonal case, the LASSO problem:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i\rangle\right)^2 + \lambda ||\boldsymbol{\theta}||_1$$

does not admit an explicit solution. How do we do in practice?

LASSO = OLS + ℓ_1 penalty

Idea: alternate between these two.

LASSO in practice

Beyond the orthogonal case, the LASSO problem:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle \right)^2 + \lambda ||\boldsymbol{\theta}||_1$$

does not admit an explicit solution. How do we do in practice?

LASSO = OLS + ℓ_1 penalty

Idea: alternate between these two.

Iterative Shrinkage-Thresholding Algorithm (ISTA)

$$\boldsymbol{\theta}^{k+1} = S_{\eta\lambda}\left(\boldsymbol{\theta}^k + \frac{\eta}{n} \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}^k) \right)$$

LASSO in practice

$n = 10$

$d = 2$

$y_i = \langle \boldsymbol{\theta}_\star, \boldsymbol{x}_i \rangle + \varepsilon_i$

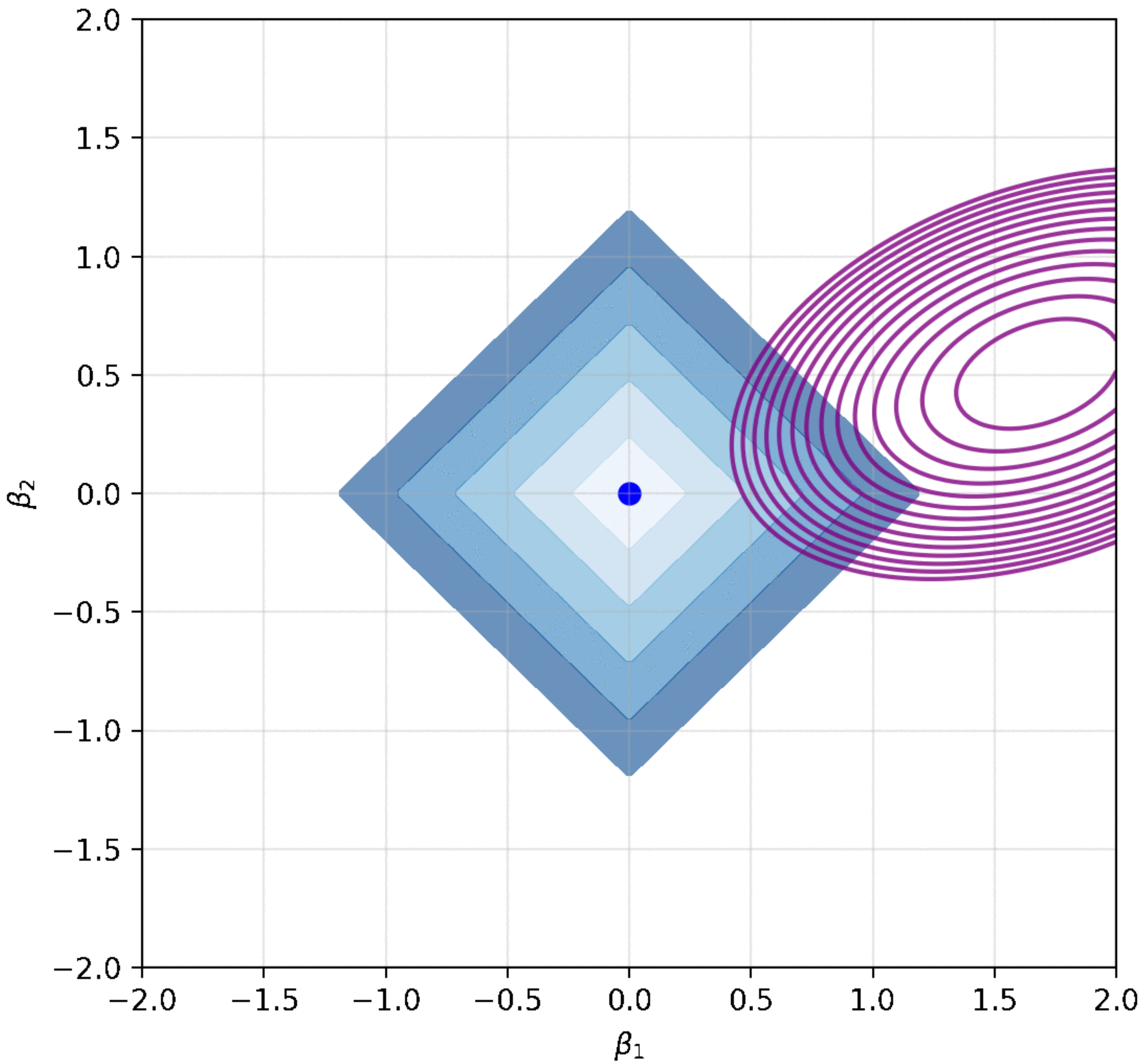
$\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{I}_2)$

$\varepsilon_i \sim \mathcal{N}(0,1)$

$$\boldsymbol{\theta}_\star = \begin{bmatrix} 1.5 \\ 0 \end{bmatrix}$$

$\eta = 0.1$

$\lambda = 0.5$



LASSO in practice

$n = 10$

$d = 2$

$y_i = \langle \boldsymbol{\theta}_\star, \boldsymbol{x}_i \rangle + \varepsilon_i$

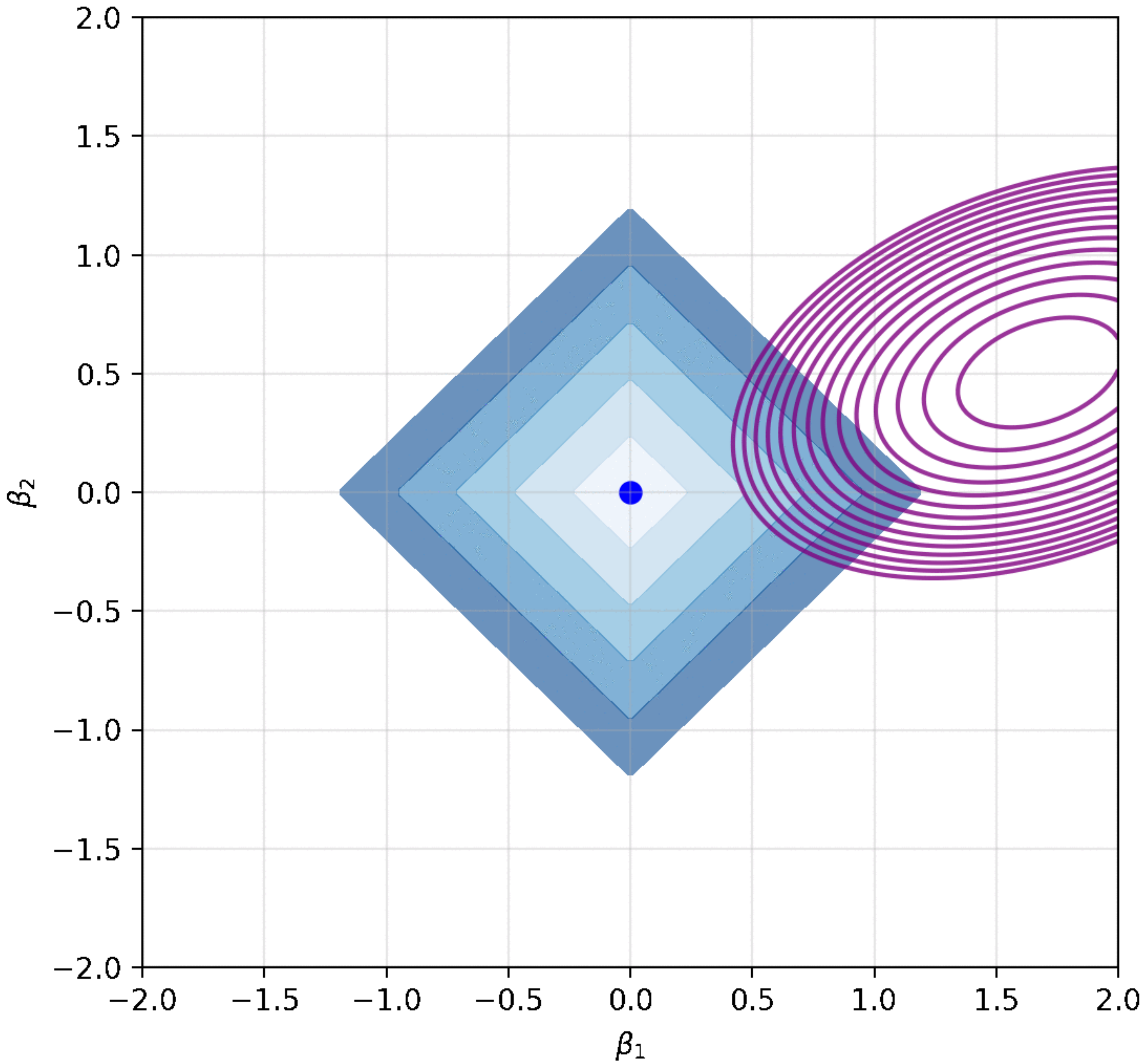
$\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{I}_2)$

$\varepsilon_i \sim \mathcal{N}(0,1)$

$$\boldsymbol{\theta}_\star = \begin{bmatrix} 1.5 \\ 0 \end{bmatrix}$$

$\eta = 0.1$

$\lambda = 0.1$



Elastic Net

The elastic net algorithm combines ridge with LASSO:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i\rangle\right)^2 + \lambda_1 ||\boldsymbol{\theta}||_1 + \frac{\lambda_2}{2}||\boldsymbol{\theta}||_2^2$$

And is particularly suited to the case where the covariate $\boldsymbol{X}$ is badly conditioned.