Statistical Learning II

Lecture 9 - BSS & LASSO

Bruno Loureiro
@ CSD, DI-ENS & CNRS

brloureiro@gmail.com

DL3 IASO, Université Paris Dauphine-PSL
20.11.2024

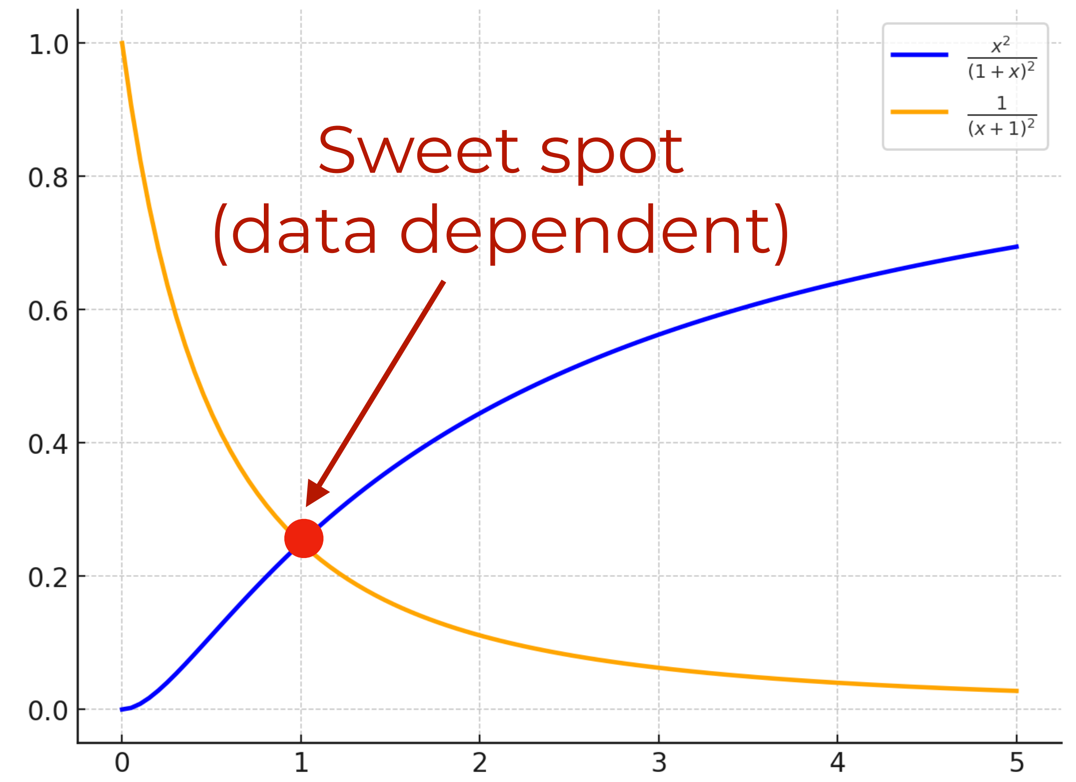
Risk of ridge

Considering the SVD of $\boldsymbol{X} = \sum_{k=1}^{\mathrm{rank}(X)} \lambda_k \boldsymbol{u}_k \boldsymbol{v}_k^\top$, we can also write:

$$\mathscr{B} = \sum_{k=1}^{\mathrm{rank}(X)} \frac{(n\lambda)^2 \lambda_k \langle \boldsymbol{v}_k, \boldsymbol{\theta}_\star \rangle^2}{(\lambda_k + n\lambda)^2} \qquad \mathscr{V} = \sum_{k=1}^{\mathrm{rank}(X)} \frac{\sigma^2 \lambda_k^2}{(\lambda_k + n\lambda)^2}$$

Remarks:

- For $\lambda \to 0^+$, we get the OLS excess risk
- $\mathscr{B}(\lambda)$ is an increasing function of λ
- $\mathscr{V}(\lambda)$ is a decreasing function of λ



Interpretation of variance

Let $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ be a positive definite matrix with decreasing eigenvalues $\mathrm{spec}(\boldsymbol{A}) = \{\lambda_k : k = 1, \cdots, d\}$. Define the cumulative:

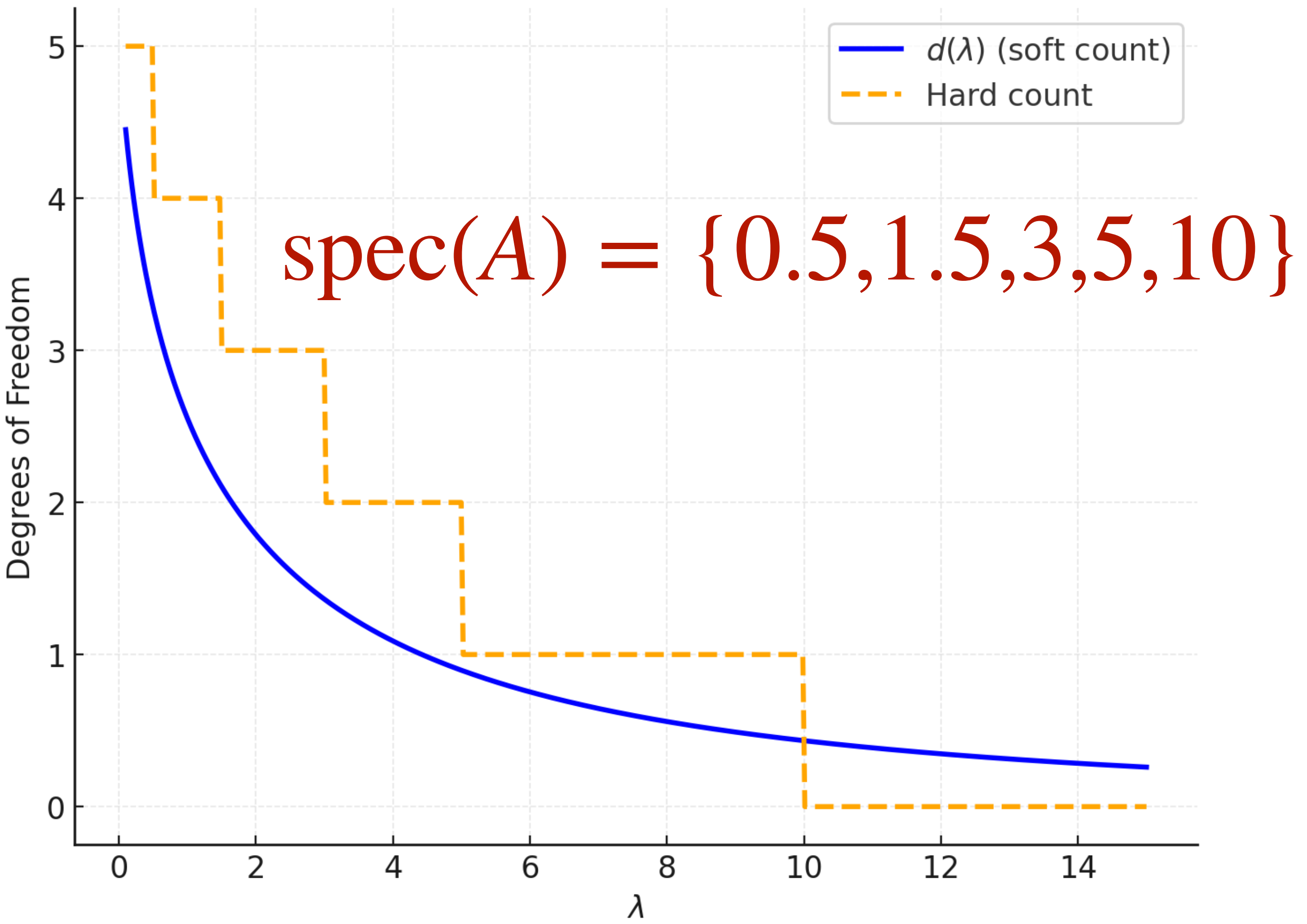
$$\phi(\lambda) = \#\{k : \lambda_k > \lambda\}$$

"Count eigenvalues bigger than λ"

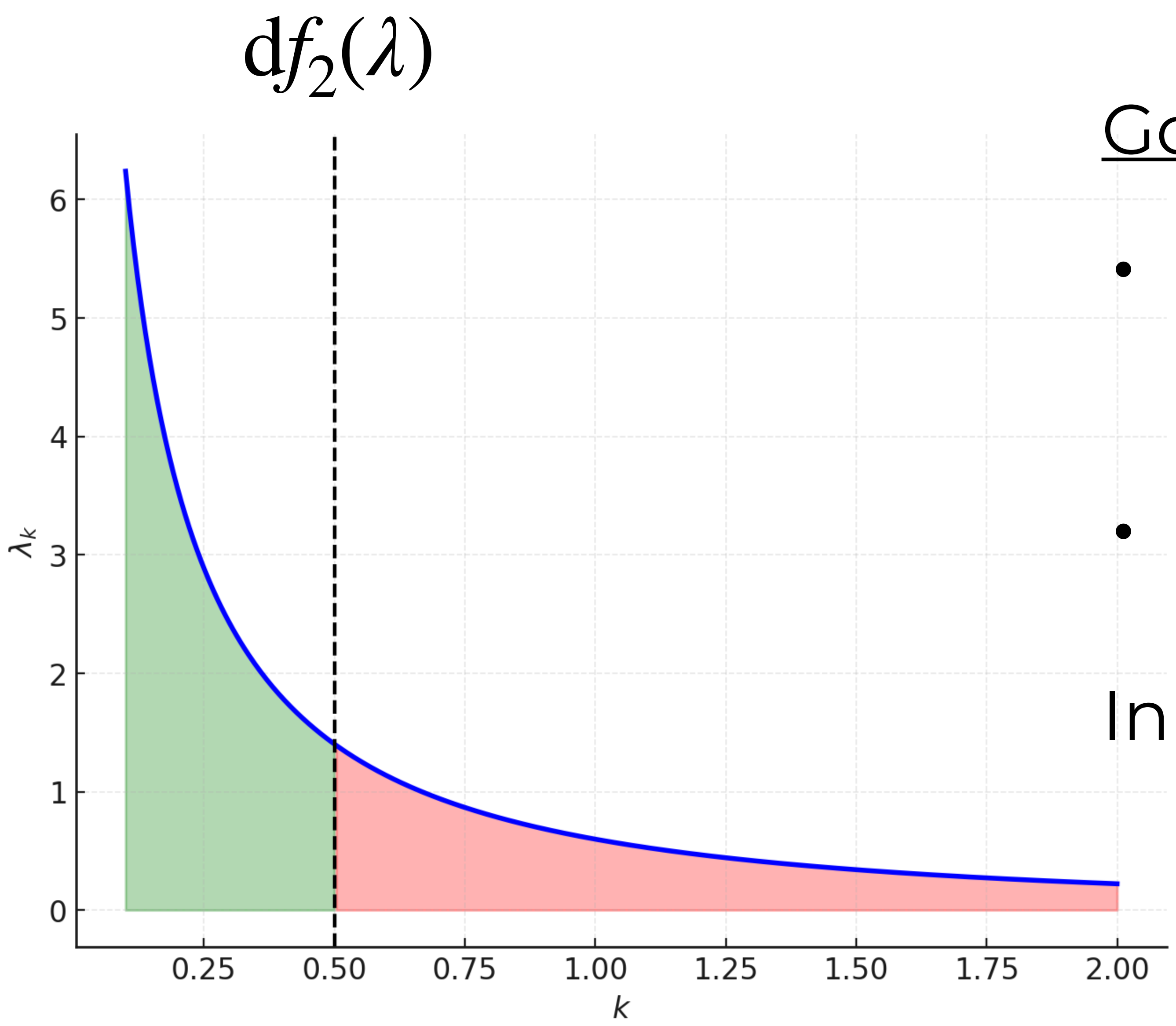
The variance of the ridge risk can be seen as a soft version:

$$\mathrm{d}f_2(\lambda) = \sum_{k=1}^{d} \frac{\lambda_k^2}{(\lambda_k + \lambda)^2}$$

- Fast decay: small λ
- Slow decay: large λ



Choosing regularisation



Low-frequency High-frequency

Goal: pick λ such that:

- directions in $\boldsymbol{X}$ that better correlate with $\boldsymbol{\theta}_\star$ are retained
- Shrink remaining directions

In practice, cross-validation...

Best subset selection & the LASSO

Pitfalls of ridge

The ridge estimation performs uniform shrinkage.

$$\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n}\left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

In other words: ℓ_2 regularisation will control the overall norm $||\hat{\boldsymbol{\theta}}_\lambda||_2^2$ by reducing each entry equally

Pitfalls of ridge

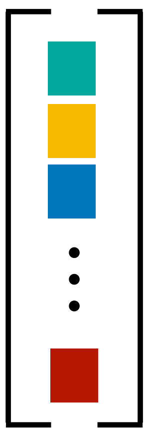
The ridge estimation performs uniform shrinkage.

$$\hat{\boldsymbol{\theta}}_{\lambda}(\boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n}\left(\frac{1}{n}\boldsymbol{X}^{\top}\boldsymbol{X} + \lambda \boldsymbol{I}_d\right)^{-1}\boldsymbol{X}^{\top}\boldsymbol{y}$$

In other words: ℓ_2 regularisation will control the overall norm $||\hat{\boldsymbol{\theta}}_{\lambda}||_2^2$ by reducing each entry equally

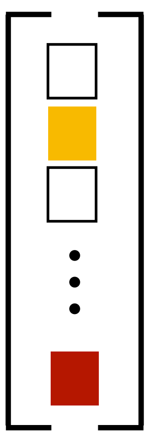
- Good if $\boldsymbol{\theta}_{\star}$ is a dense vector

$$\boldsymbol{\theta}_{\star,j} \neq 0 \qquad i = 1, \cdots, d$$

$\boldsymbol{\theta}_{\star} =$ 

- Bad if $\boldsymbol{\theta}_{\star}$ is a sparse vector

$$\boldsymbol{\theta}_{\star,j} = \begin{cases} 0 & j \in S \subset \{1, \ldots, d\} \\ \neq 0 & j \in \{1, \ldots, d\} \backslash S \end{cases}$$

$\boldsymbol{\theta}_{\star} =$ 

Sparsity is everywhere

Many signals of interest admit a sparse representation in a particular basis.

$$f(\boldsymbol{x}) = \sum_{k \geq 0} f_k \psi_k(\boldsymbol{x})$$

basis

coefficients

Sparsity is everywhere

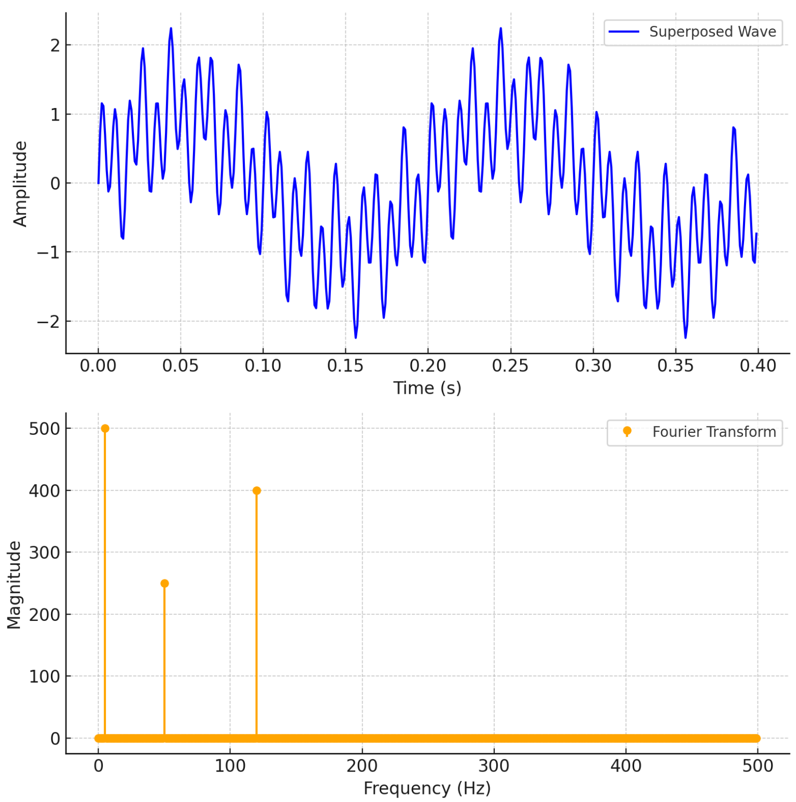
Many signals of interest admit a sparse representation in a particular basis.

$$f(\boldsymbol{x}) = \sum_{k \geq 0} f_k \psi_k(\boldsymbol{x})$$

basis

coefficients

Example: superposition of sine waves



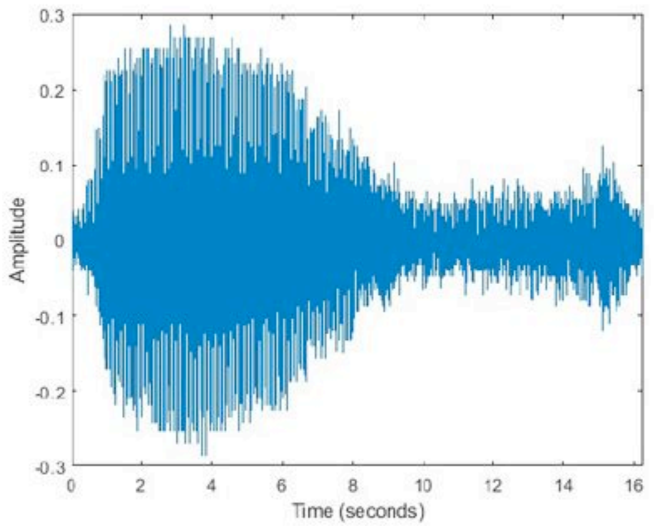
$$f(t) = \sin(10\pi t) + 0.5\sin(100\pi t) + 0.8\sin(240\pi t)$$

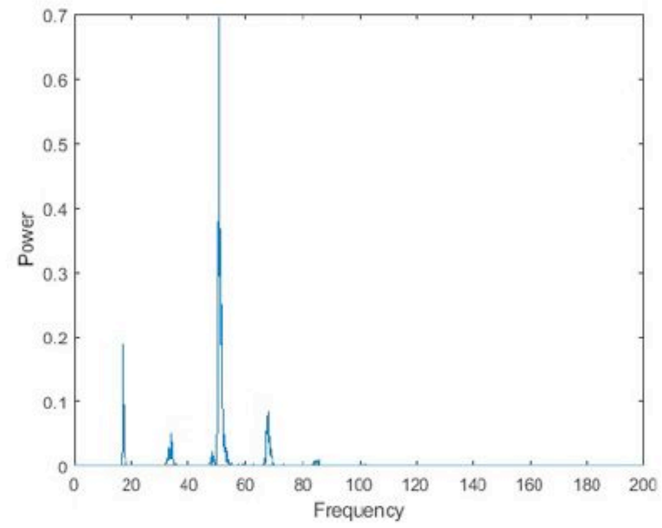
$$\hat{f}(\omega) = \delta_5 + 0.5\ \delta_{50} + 0.8\ \delta_{120}$$

Sparsity is everywhere

Examples:

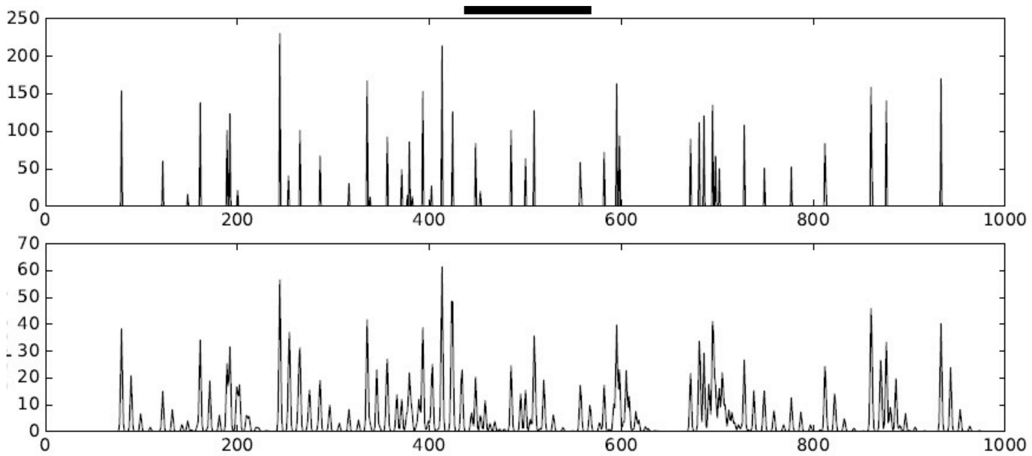
Sound





Images

Scientific signals (mass spectrography)



And many more…

- Portfolio selection (finance)
- Networks (power grids)
- electroencephalogram
- Etc…

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

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

Best subset selection

Idea: encourage solutions which are sparse.

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

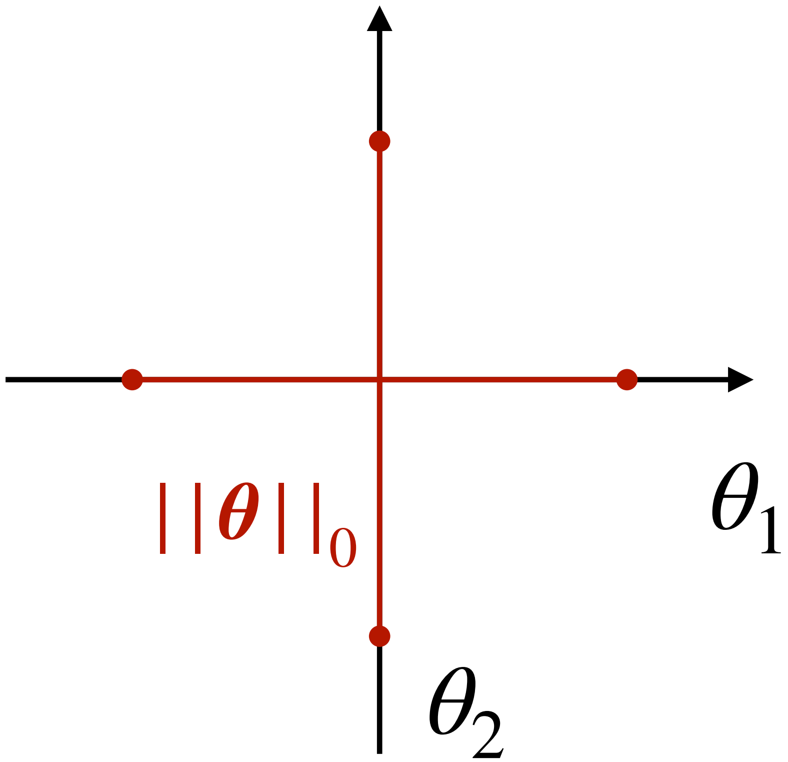
where $||\cdot||_0 : \mathbb{R}^d \to \{0,1,\ldots,d\}$ is the ℓ_0-"norm": Strictly not a norm

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

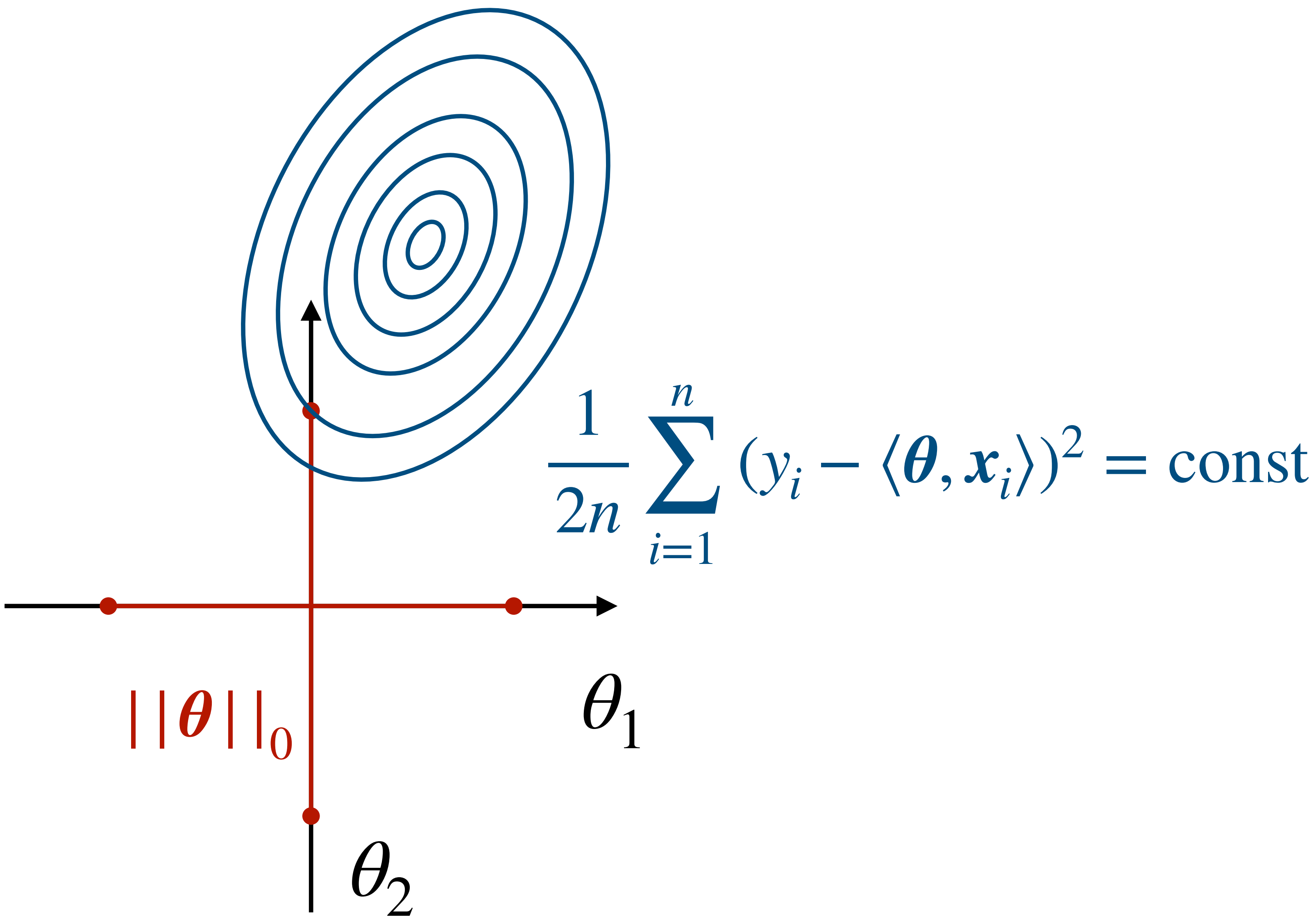
Hence, $\lambda \geq 0$ controls the desired sparsity level

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

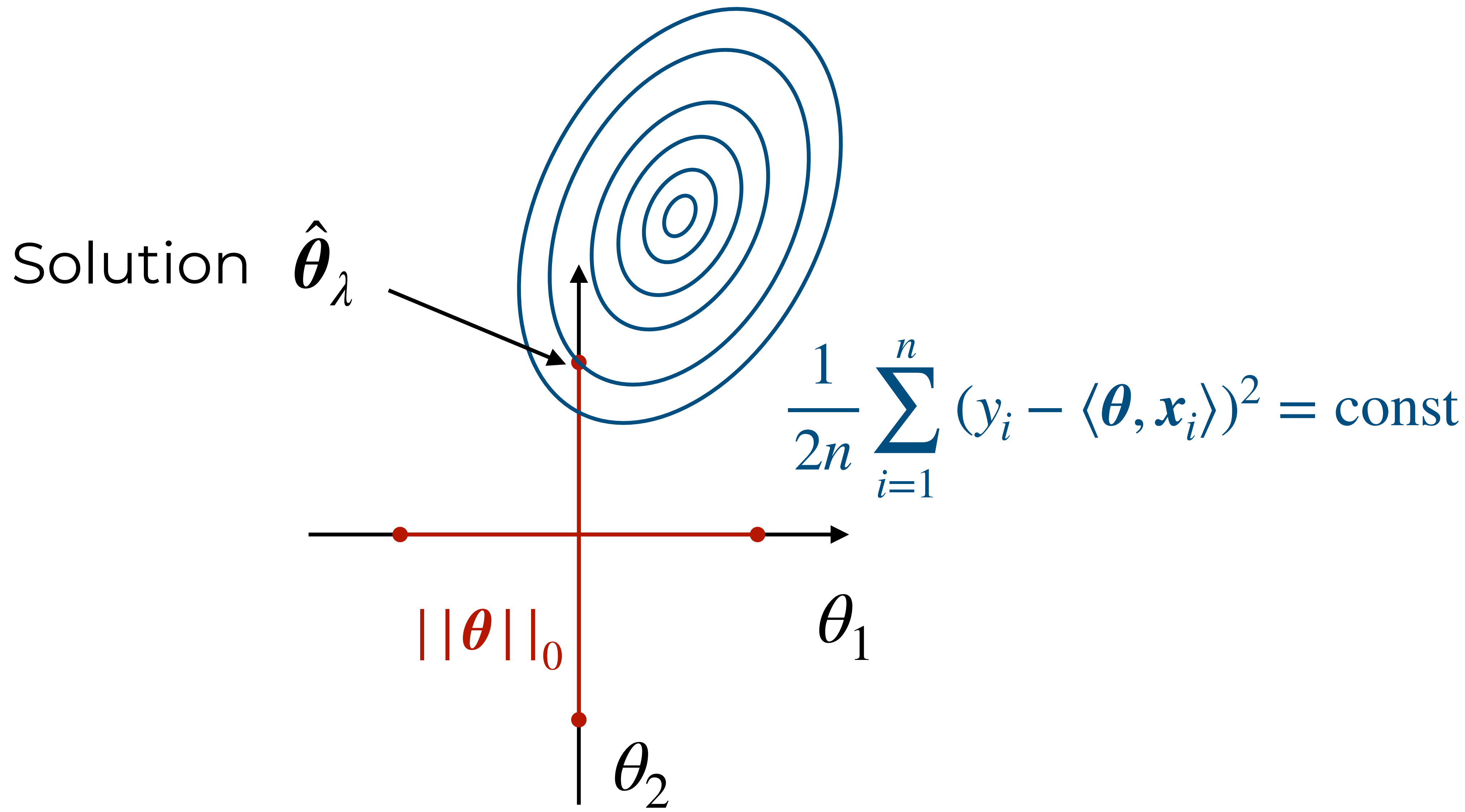
BSS: visualisation



BSS: visualisation



BSS: visualisation



BSS: orthogonal covariates

To get some intuition about this problem, let's consider a simplified setting: assume the covariates are orthogonal

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

BSS: orthogonal covariates

To get some intuition about this problem, let's consider a simplified setting: assume the covariates are orthogonal

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

Then, we can rewrite:

$$||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||_2^2 = ||\boldsymbol{y}||_2^2 + \boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\theta} - 2\boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{y}$$

BSS: orthogonal covariates

To get some intuition about this problem, let's consider a simplified setting: assume the covariates are orthogonal

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

Then, we can rewrite:

$$||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||_2^2 = ||\boldsymbol{y}||_2^2 + \boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\theta} - 2\boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{y}$$

$$= ||\boldsymbol{y}||_2^2 + ||\boldsymbol{\theta}||_2^2 - 2\boldsymbol{\theta}^\top \boldsymbol{z} \qquad (\boldsymbol{z} = \boldsymbol{X}^\top \boldsymbol{y} \in \mathbb{R}^d)$$

BSS: orthogonal covariates

To get some intuition about this problem, let's consider a simplified setting: assume the covariates are orthogonal

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

Then, we can rewrite:

$$\begin{aligned} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||_2^2 &= ||\boldsymbol{y}||_2^2 + \boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\theta} - 2\boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{y} \\ &= ||\boldsymbol{y}||_2^2 + ||\boldsymbol{\theta}||_2^2 - 2\boldsymbol{\theta}^\top \boldsymbol{z} \qquad (\boldsymbol{z} = \boldsymbol{X}^\top \boldsymbol{y} \in \mathbb{R}^d) \\ &= ||\boldsymbol{y}||_2^2 + ||\boldsymbol{z}||_2^2 - ||\boldsymbol{z} - \boldsymbol{\theta}||_2^2 \end{aligned}$$

BSS: orthogonal covariates

To get some intuition about this problem, let's consider a simplified setting: assume the covariates are orthogonal

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

Therefore, under the above:

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

Is equivalent to:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{1}{2n} || \boldsymbol{z} - \boldsymbol{\theta} ||_2^2 + \lambda \, || \boldsymbol{\theta} ||_0$$

Which is a simpler problem since it factorises coordinate-wise.

BSS: orthogonal covariates

Coordinate-wise, we need to solve

$$\min_{\theta_j \in \mathbb{R}} L(\theta_j) := \left\{ \frac{1}{2n}(z_j - \theta_j)^2 + \lambda \mathbb{I}(\theta_j \neq 0) \right\}$$

BSS: orthogonal covariates

Coordinate-wise, we need to solve

$$\min_{\theta_j \in \mathbb{R}} L(\theta_j) := \left\{ \frac{1}{2n}(z_j - \theta_j)^2 + \lambda \mathbb{I}(\theta_j \neq 0) \right\}$$

Note that:

$$L(\theta_j) = \frac{1}{2n}(z_j - \theta_j)^2 + \lambda \mathbb{I}(\theta_j \neq 0) = \begin{cases} \frac{1}{2n} z_j^2 & \text{if } \theta_j = 0 \quad \text{(a)} \\ \frac{1}{2n}(z_j - \theta_j)^2 + \lambda & \text{if } \theta_j \neq 0 \quad \text{(b)} \end{cases}$$

BSS: orthogonal covariates

Coordinate-wise, we need to solve

$$\min_{\theta_j \in \mathbb{R}} L(\theta_j) := \left\{ \frac{1}{2n}(z_j - \theta_j)^2 + \lambda \mathbb{I}(\theta_j \neq 0) \right\}$$

Note that:

$$L(\theta_j) = \frac{1}{2n}(z_j - \theta_j)^2 + \lambda \mathbb{I}(\theta_j \neq 0) = \begin{cases} \frac{1}{2n} z_j^2 & \text{if } \theta_j = 0 \text{ (a)} \\ \frac{1}{2n}(z_j - \theta_j)^2 + \lambda & \text{if } \theta_j \neq 0 \text{ (b)} \end{cases}$$

Note the solution of the problem is not unique:

- In case (a), solution is $\hat{\theta}^{(1)}_{\lambda,j} = 0$
- In case (b), solution is $\hat{\theta}^{(2)}_{\lambda,j} = z_j$

BSS: orthogonal covariates

Coordinate-wise, we need to solve

$$\min_{\theta_j \in \mathbb{R}} L(\theta_j) := \left\{ \frac{1}{2n}(z_j - \theta_j)^2 + \lambda \mathbb{I}(\theta_j \neq 0) \right\}$$

Note that:

$$L(\theta_j) = \frac{1}{2n}(z_j - \theta_j)^2 + \lambda \mathbb{I}(\theta_j \neq 0) = \begin{cases} \frac{1}{2n} z_j^2 & \text{if } \theta_j = 0 \text{ (a)} \\ \frac{1}{2n}(z_j - \theta_j)^2 + \lambda & \text{if } \theta_j \neq 0 \text{ (b)} \end{cases}$$

Note the solution of the problem is not unique:

- In case (a), solution is $\hat{\theta}^{(1)}_{\lambda,j} = 0$
- In case (b), solution is $\hat{\theta}^{(2)}_{\lambda,j} = z_j$

Which one to pick? The one with minimal loss.

BSS: orthogonal covariates

Note the solution of the problem is not unique:

- In case (a), solution is $\hat{\theta}^{(1)}_{\lambda,j} = 0$
- In case (b), solution is $\hat{\theta}^{(2)}_{\lambda,j} = z_j$

Which one to pick? The one with minimal loss.

$$L\left(\hat{\theta}^{(2)}_{\lambda,j}\right) - L\left(\hat{\theta}^{(1)}_{\lambda,j}\right) = -\frac{z_j^2}{2n} + \lambda \underset{?}{\geq} 0$$

BSS: orthogonal covariates

Note the solution of the problem is not unique:

- In case (a), solution is $\hat{\theta}^{(1)}_{\lambda,j} = 0$
- In case (b), solution is $\hat{\theta}^{(2)}_{\lambda,j} = z_j$

Which one to pick? The one with minimal loss.

$$L\left(\hat{\theta}^{(2)}_{\lambda,j}\right) - L\left(\hat{\theta}^{(1)}_{\lambda,j}\right) = -\frac{z_j^2}{2n} + \lambda \underset{?}{\geq} 0 \quad \Leftrightarrow \quad 2n\lambda \geq z_j^2$$

BSS: orthogonal covariates

Note the solution of the problem is not unique:

- In case (a), solution is $\hat{\theta}^{(1)}_{\lambda,j} = 0$
- In case (b), solution is $\hat{\theta}^{(2)}_{\lambda,j} = z_j$

Which one to pick? The one with minimal loss.

$$L\left(\hat{\theta}^{(2)}_{\lambda,j}\right) - L\left(\hat{\theta}^{(1)}_{\lambda,j}\right) = -\frac{z_j^2}{2n} + \lambda \underset{?}{\geq} 0 \quad \Leftrightarrow \quad 2n\lambda \geq z_j^2$$

Hence, the solution is given by:

$$\hat{\theta}_{\lambda,j} = \begin{cases} 0 & \text{if } z_j^2 < 2n\lambda \\ z_j & \text{if } z_j^2 \geq 2n\lambda \end{cases}$$

"Hard threshold" function

BSS: orthogonal covariates

Putting together, the solution of the BSS problem:

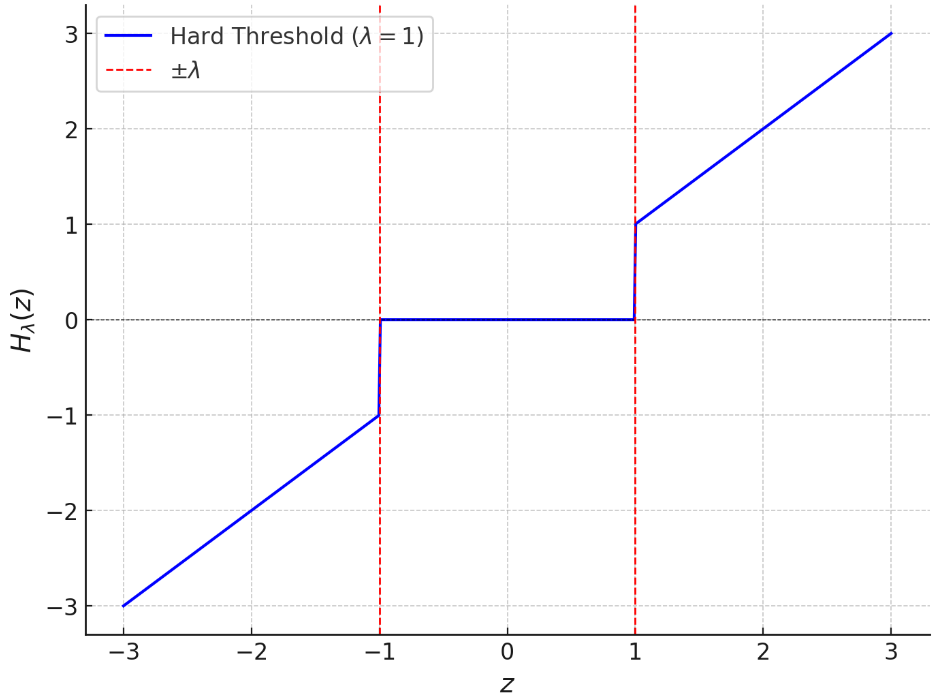
$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i\rangle\right)^2 + \lambda||\boldsymbol{\theta}||_0$$

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

$$\hat{\boldsymbol{\theta}}_\lambda = H_{\sqrt{2n\lambda}}(\boldsymbol{X}^\top\boldsymbol{y})$$

Where:

$$H_\lambda(z) = \begin{cases} 0 & \text{if } |z| < \lambda \\ z & \text{otherwise} \end{cases}$$



BSS: orthogonal covariates

To understand better this solution, consider a linear model for the data:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_\star + \boldsymbol{\varepsilon}$$

With $\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top] = \sigma \boldsymbol{I}_n$, $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$ and $\boldsymbol{\theta}_\star$ a k-sparse vector

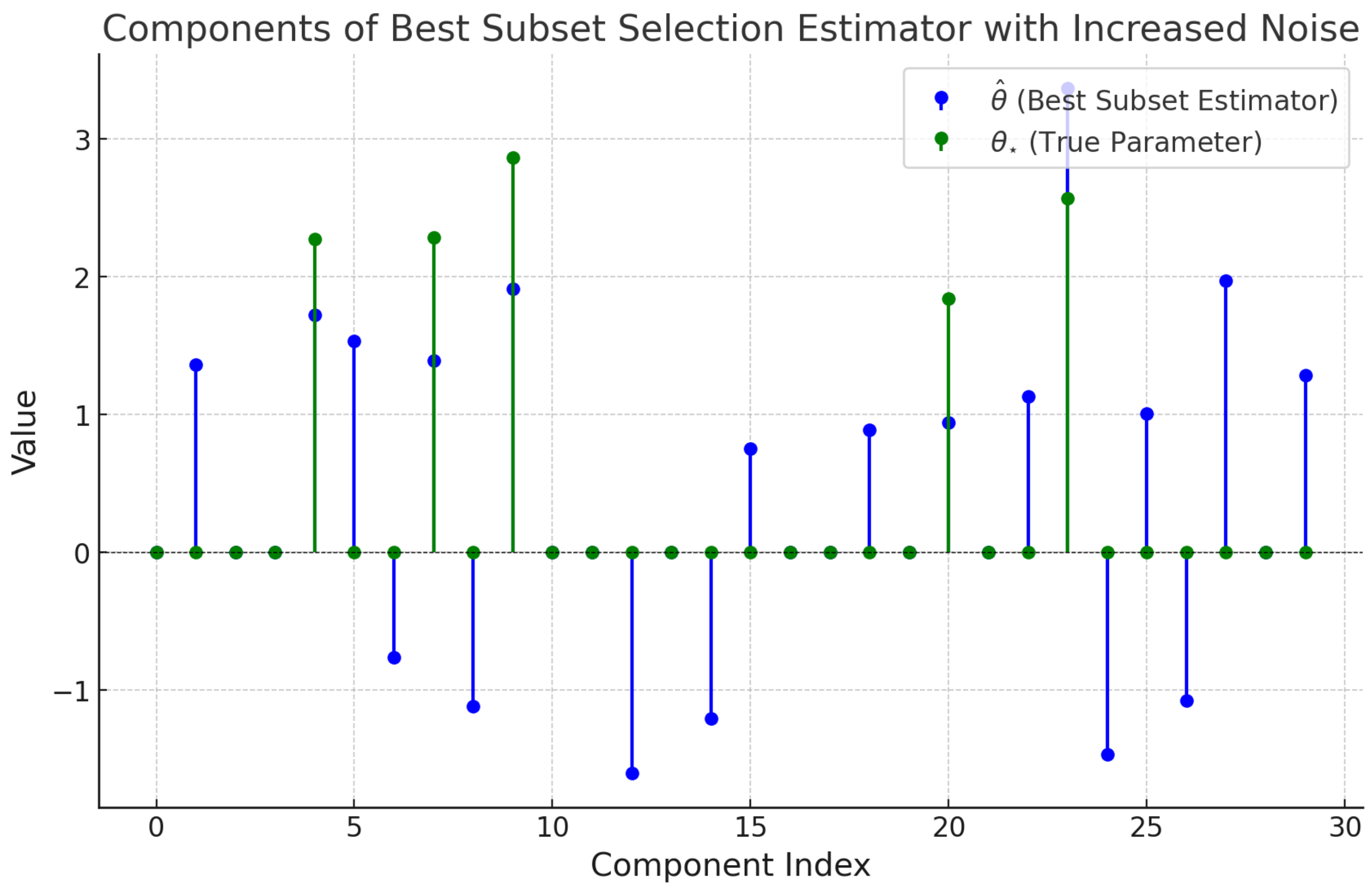
The, the solution is given by:

$$\hat{\boldsymbol{\theta}}_\lambda = H_{\sqrt{2n\lambda}}(\boldsymbol{\theta}_\star + \boldsymbol{X}^\top \boldsymbol{\varepsilon})$$

BSS: orthogonal covariates

Example: $n = 40$ $\lambda = 0.5$ $\boldsymbol{\theta}_\star$ 5-sparse

$d = 30$ $\sigma^2 = 1$ $||\boldsymbol{\theta}_\star||_2^2 = 5.35$



BSS: beyond orthogonal

When the covariates are not orthogonal, an explicit solution is not available. Nevertheless, we can partially characterise it.

BSS: beyond orthogonal

When the covariates are not orthogonal, an explicit solution 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 BSS solution

Denoting:

- $\hat{\boldsymbol{\theta}}_S \in \mathbb{R}^{|S|}$ 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

BSS: beyond orthogonal

When the covariates are not orthogonal, an explicit solution 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 BSS solution

Denoting:

- $\hat{\boldsymbol{\theta}}_S \in \mathbb{R}^{|S|}$ 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

We can write:

$$\hat{\boldsymbol{\theta}}_S = \boldsymbol{X}_S^+ \boldsymbol{y}$$

In other words, BSS = OLS in the support!

The hard part is to find S as a function of $\boldsymbol{X}, \boldsymbol{y}, \lambda$...

Pitfalls of BSS

More generally, BSS is that it is a non-convex problem

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

In particular, for general covariates it is hard to optimise.
(it is actually a NP-hard problem in the worst case)

Pitfalls of BSS

More generally, BSS is that it is a non-convex problem

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

In particular, for general covariates it is hard to optimise.
(it is actually a NP-hard problem in the worst case)

Question: $||\cdot||_0$ is what makes this non-convex. Can we find another regularisation with similar properties but convex?

Pitfalls of BSS

More generally, BSS is that it is a non-convex problem

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

In particular, for general covariates it is hard to optimise.
(it is actually a NP-hard problem in the worst case)

🤔 Question: $||\cdot||_0$ is what makes this non-convex. Can we find another regularisation with similar properties but convex?

💡 That's the key idea of the LASSO.

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:

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

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.

LASSO

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

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

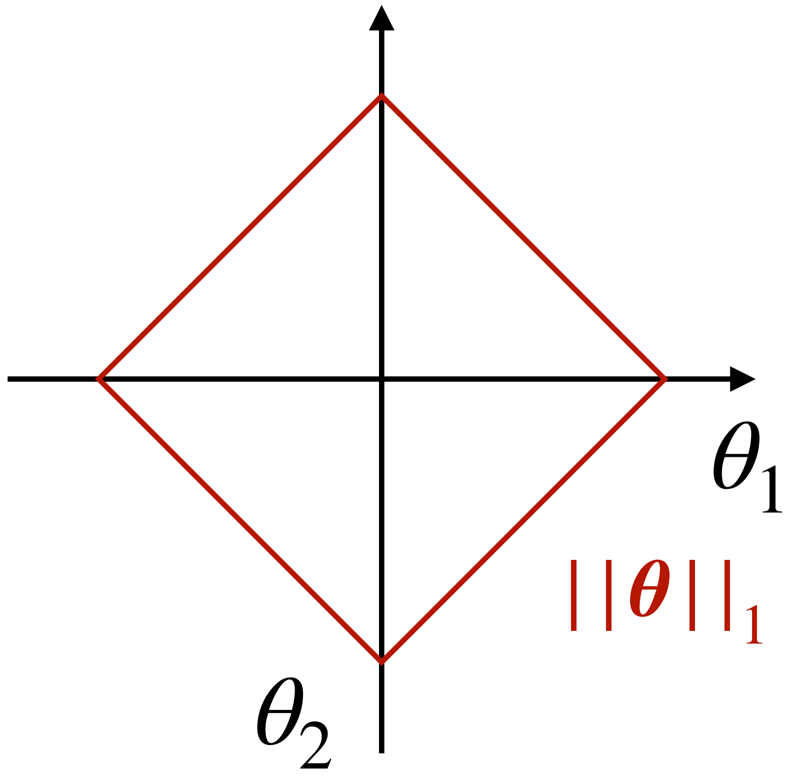
where $|| \cdot ||_1 : \mathbb{R}^d \to \mathbb{R}_+$ is the ℓ_1-norm:

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

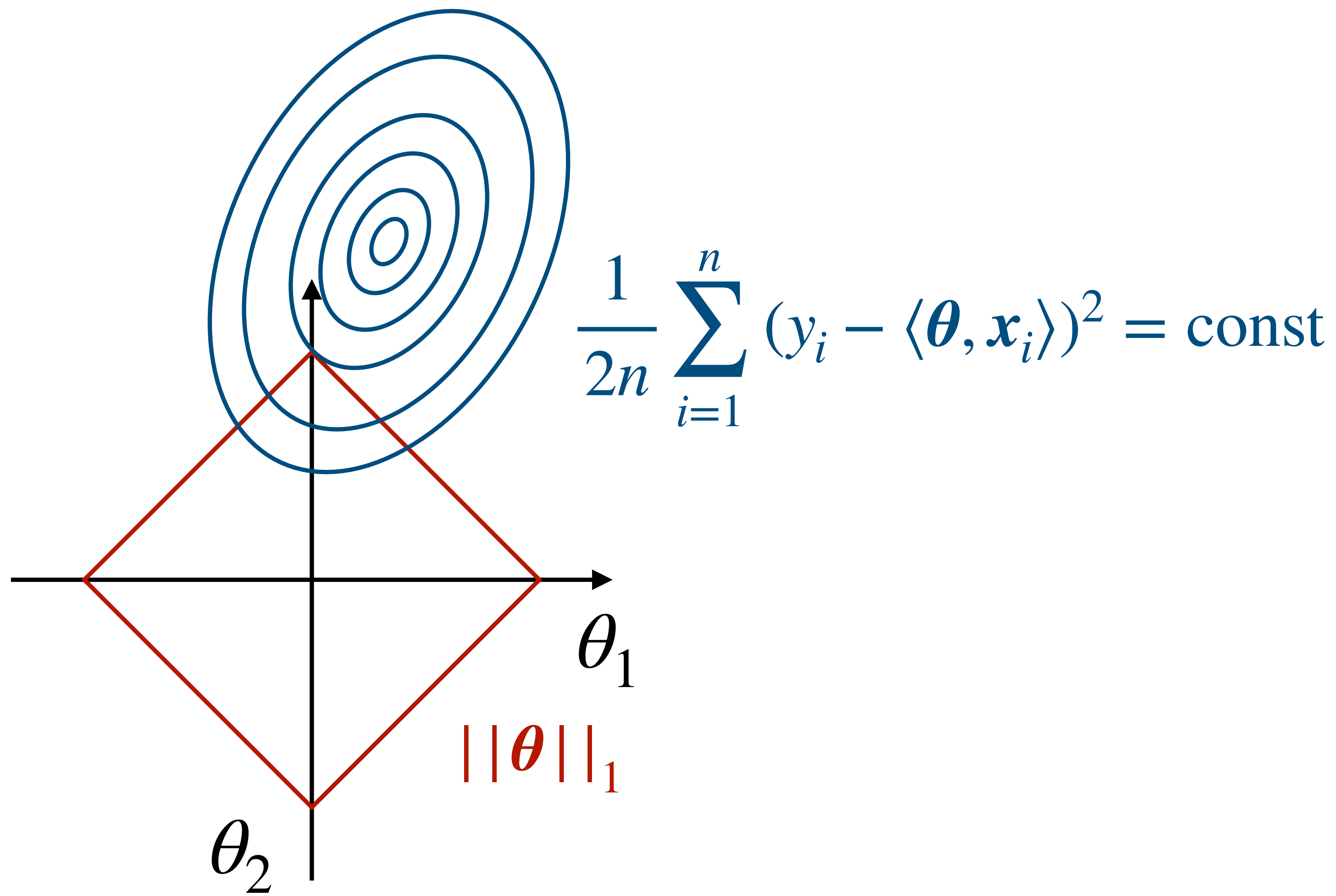
Moreover, this is a convex problem.

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

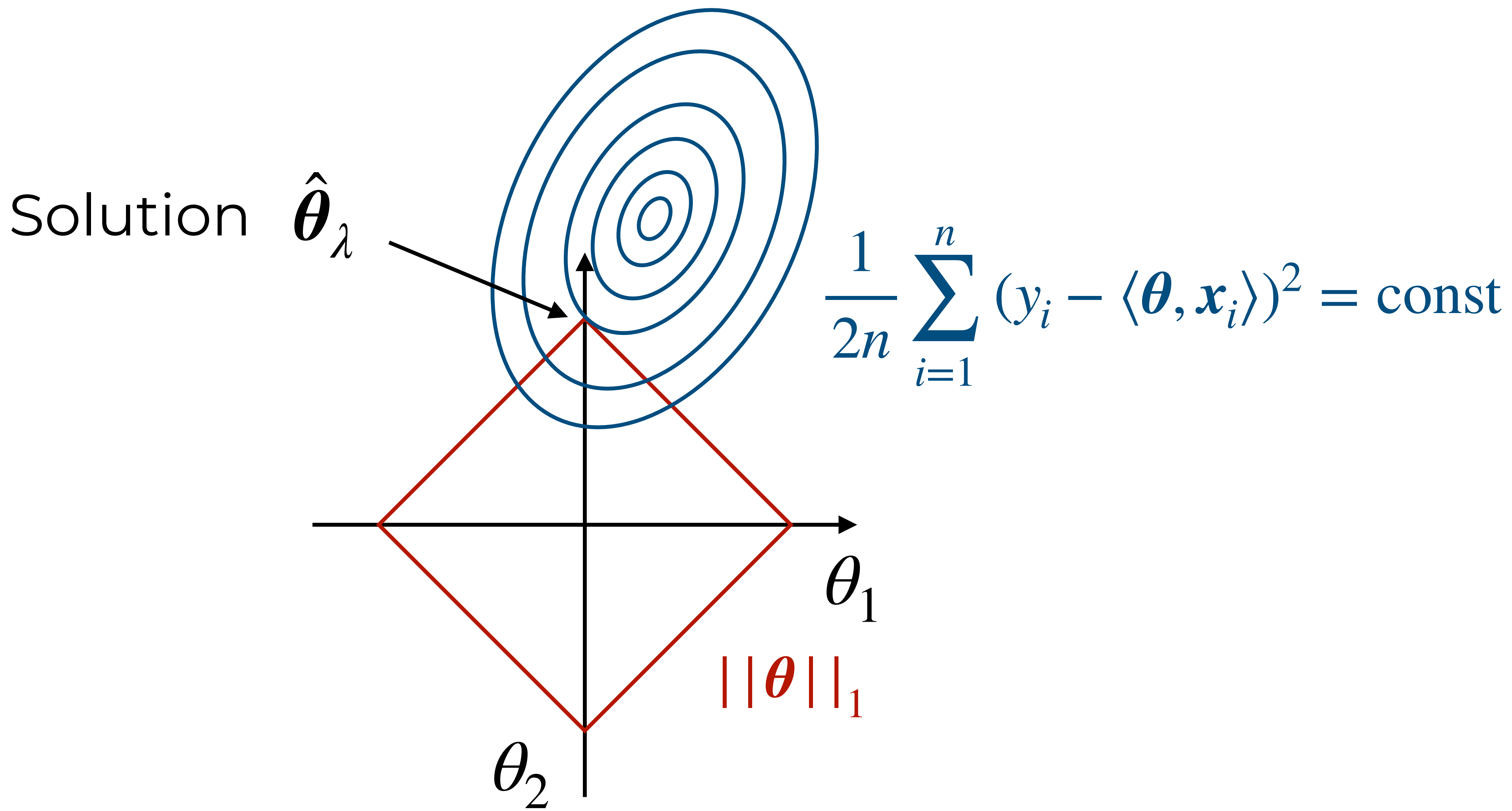
LASSO: visualisation



LASSO: visualisation



LASSO: visualisation



Sharper corners favours sparser solutions!