Statistical Learning II

Lecture 11 - LASSO & Feature maps

Bruno Loureiro
@ CSD, DI-ENS & CNRS

brloureiro@gmail.com

DL3 IASO, Université Paris Dauphine-PSL
03.12.2025

BSS: orthogonal covariates

Putting together, the solution of the BSS problem:

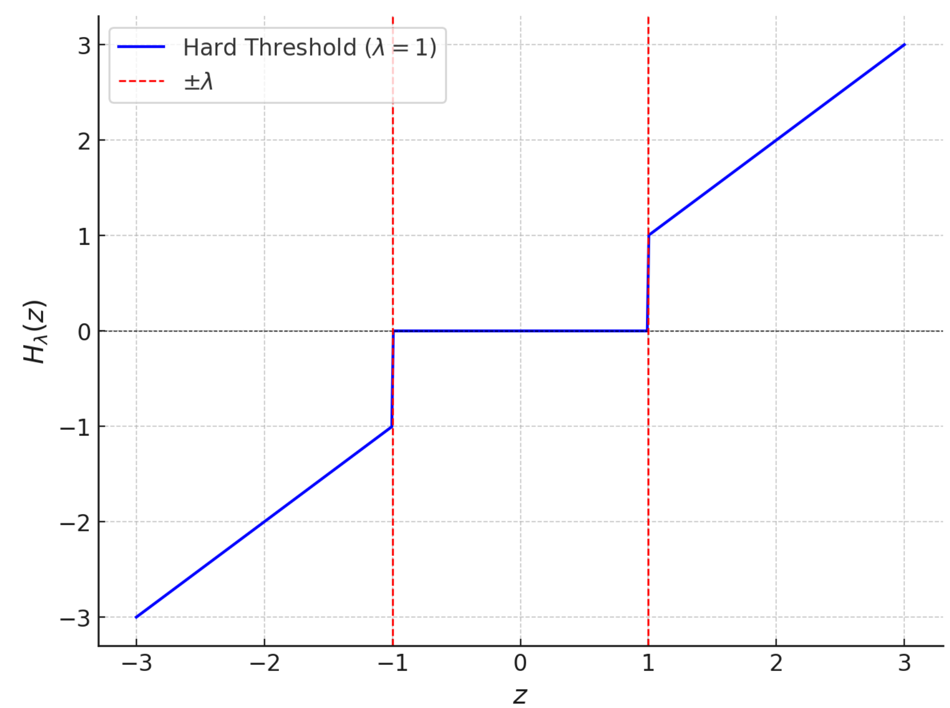
$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle\right)^2 + \lambda ||\boldsymbol{\theta}||_0$$

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

$\mathbb{E}[\boldsymbol{\varepsilon}] = 0$

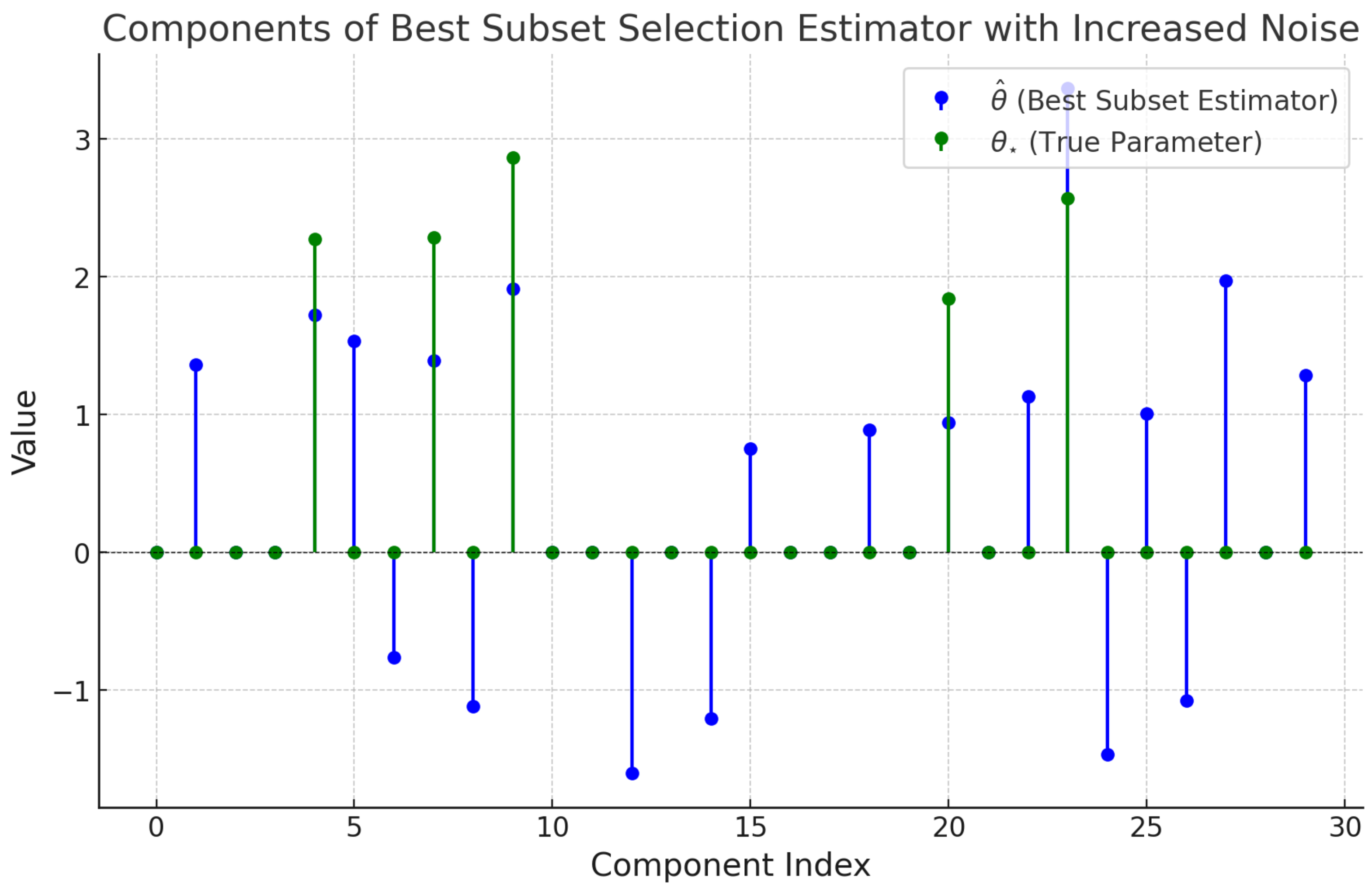
The, the solution is given by:

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

BSS: orthogonal covariates

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

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



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:

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

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:

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

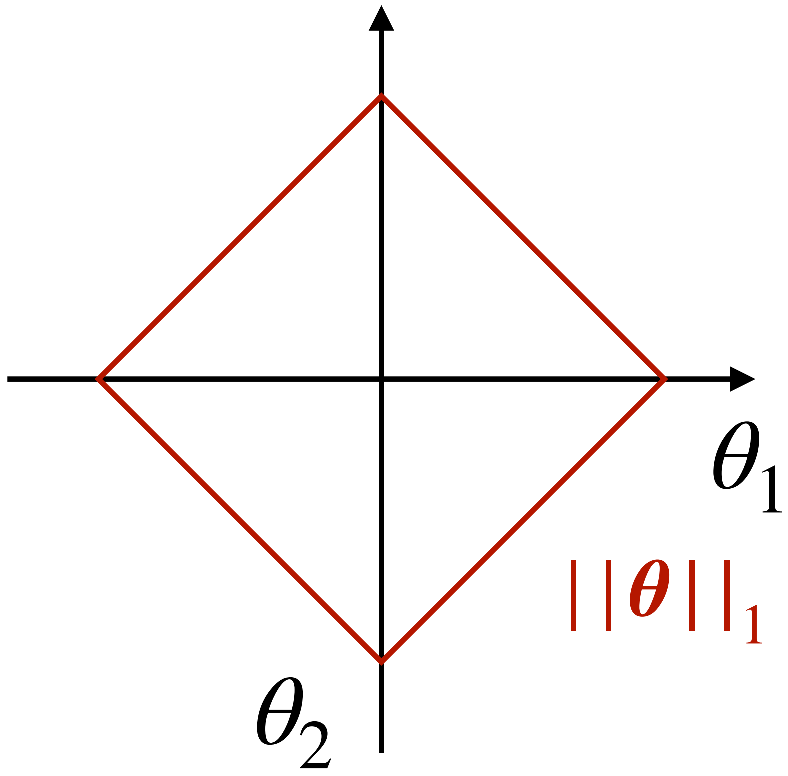
where $||\cdot||_1 : \mathbb{R}^d \to \mathbb{R}_+$ is the ℓ_1-norm:

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

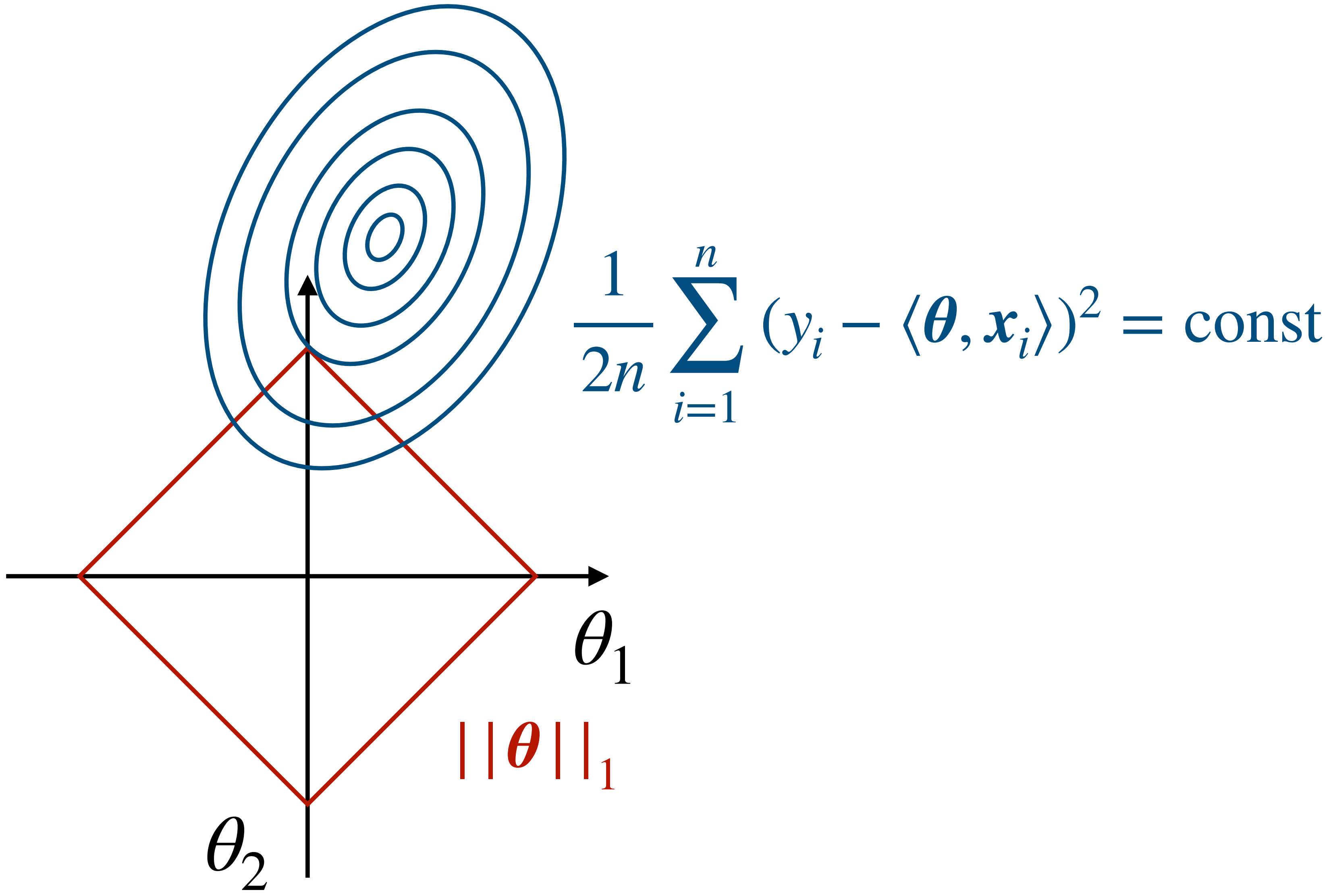
Moreover, this is a convex problem.

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

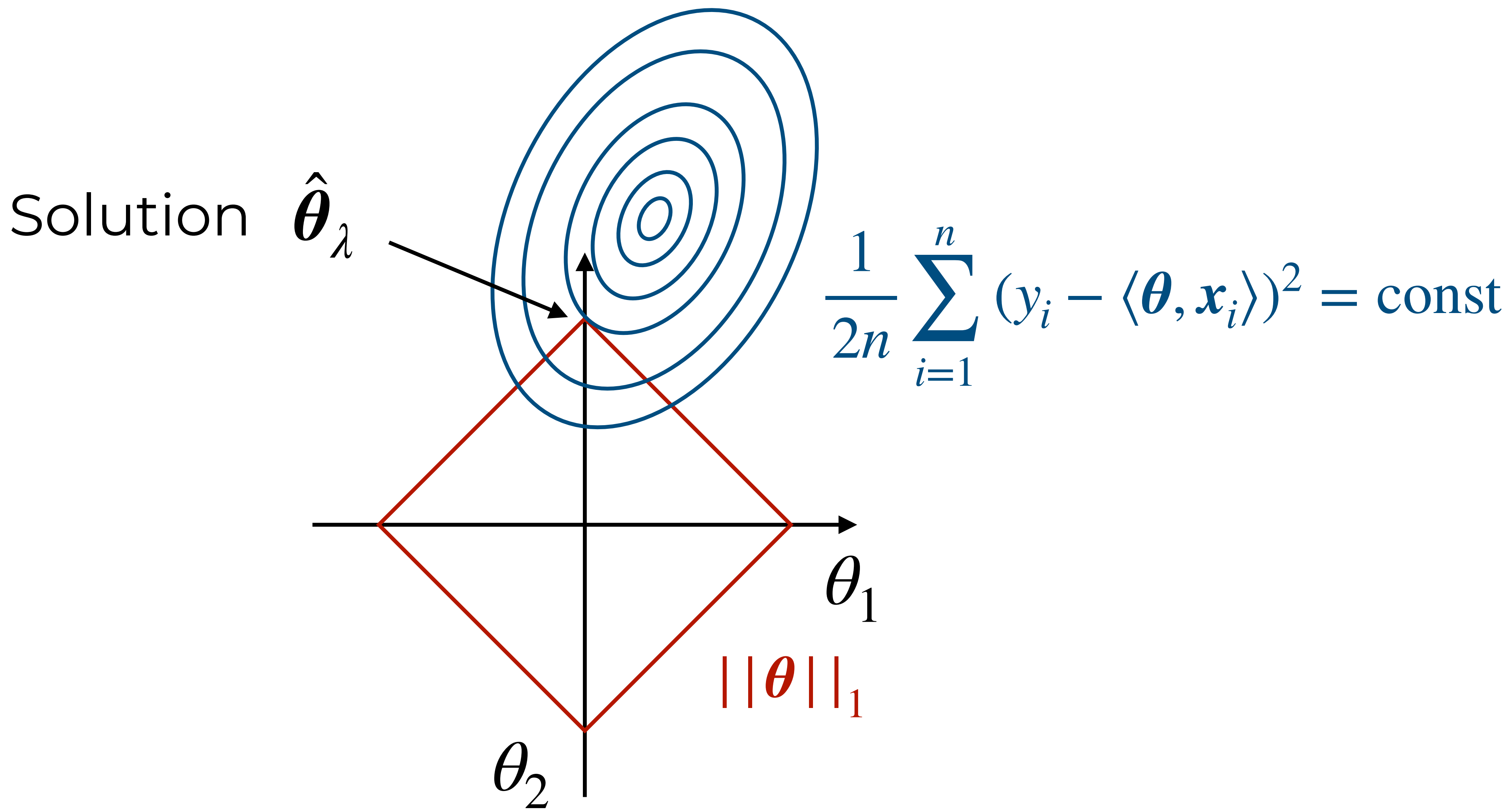
LASSO: visualisation



LASSO: visualisation



LASSO: visualisation



Sharper corners favours sparser solutions!

LASSO: orthogonal covariates

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

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

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:

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

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

LASSO: orthogonal covariates

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

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

LASSO: orthogonal covariates

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

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

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

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 $z_j > -n\lambda$

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 $z_j > -n\lambda$

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:

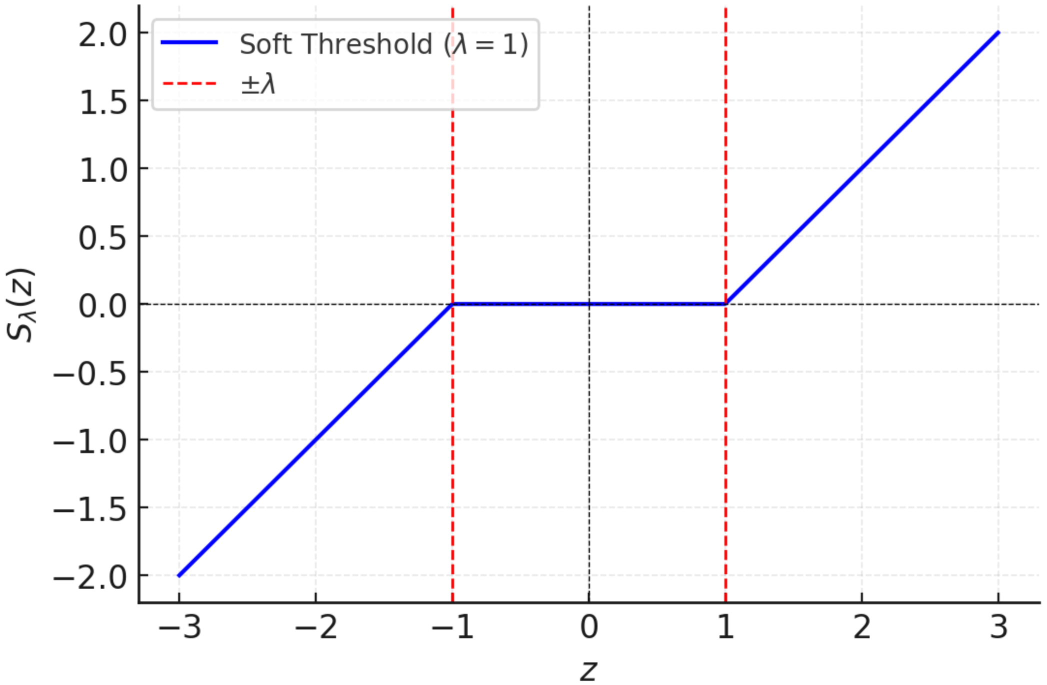
$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle\right)^2 + \lambda ||\boldsymbol{\theta}||_1$$

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

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

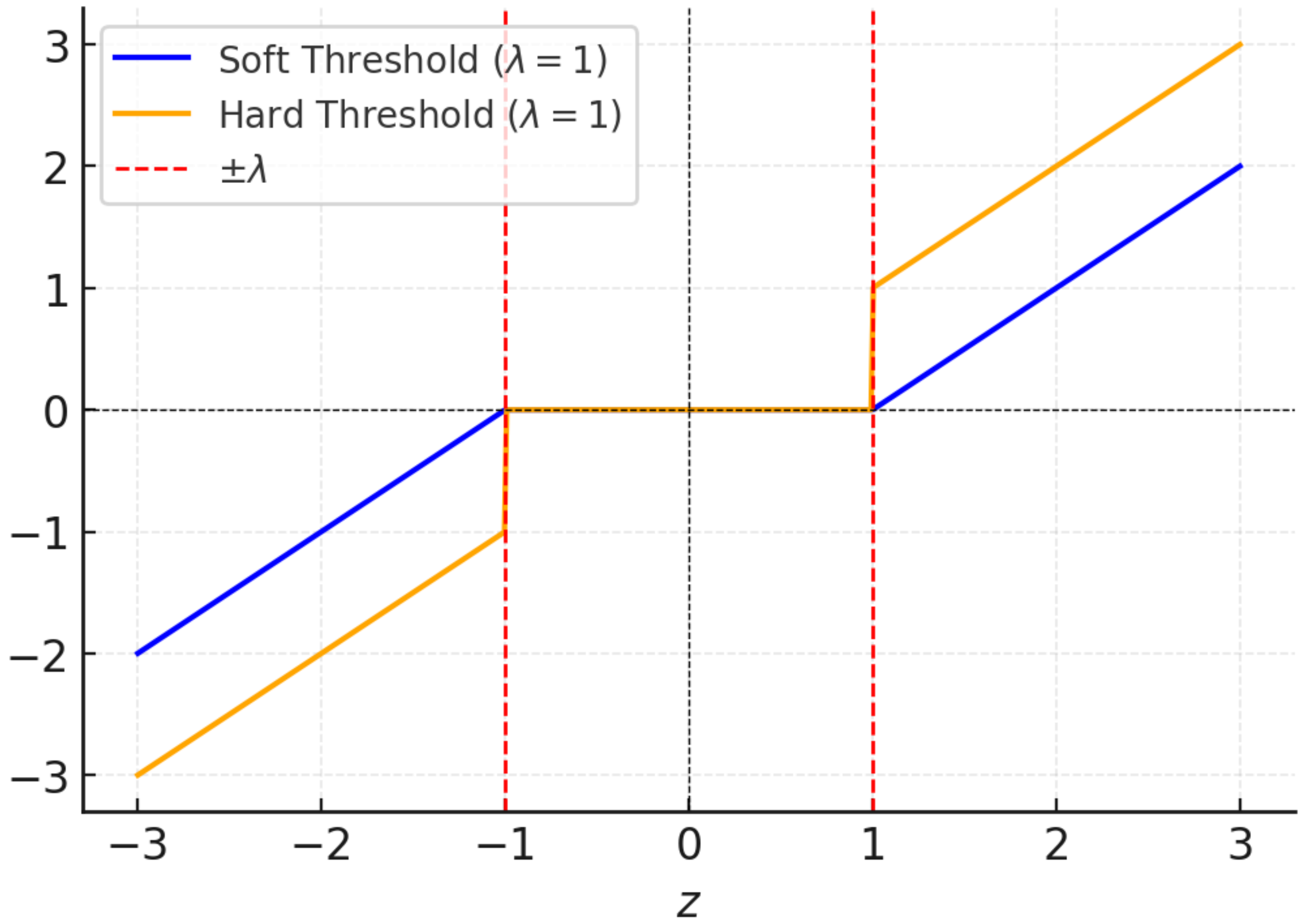
Where:

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



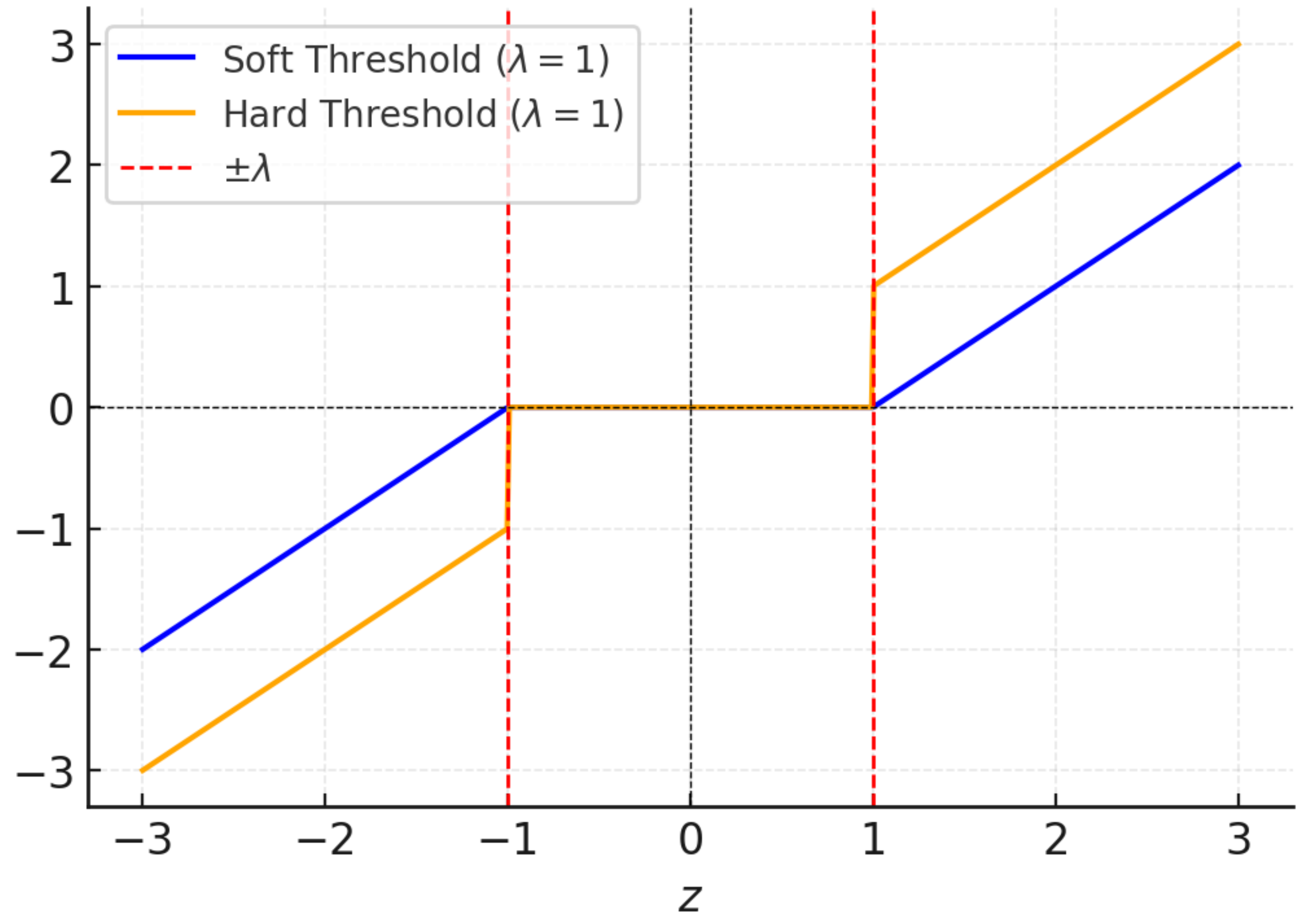
BSS vs. LASSO

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



BSS vs. LASSO

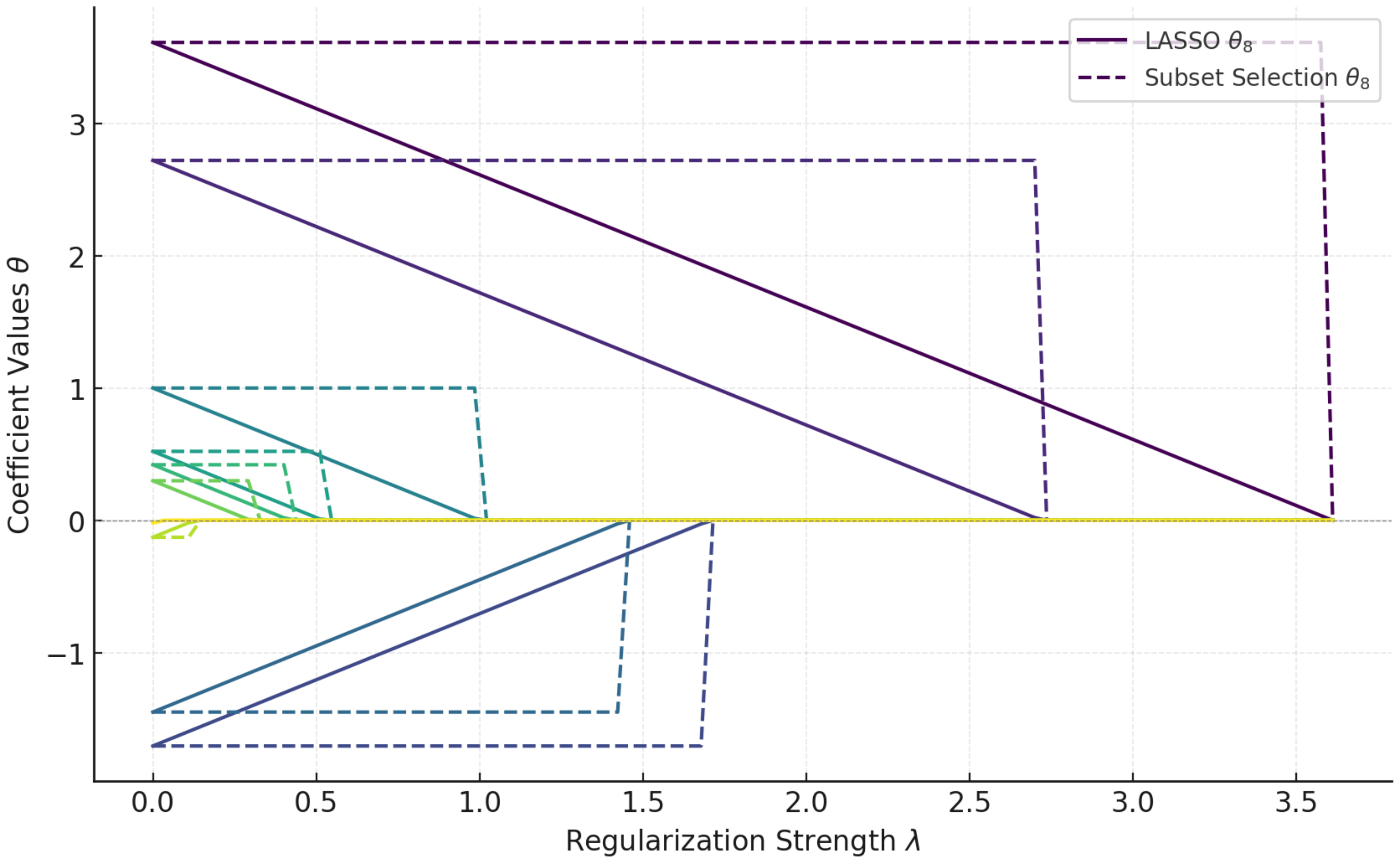
It is instructive to compare the BSS and LASSO solutions in the orthogonal covariate case



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

BSS vs. LASSO

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

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

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:

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

⚠ 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:

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

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:

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

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:

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

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:

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

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

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

$n = 10$

$d = 2$

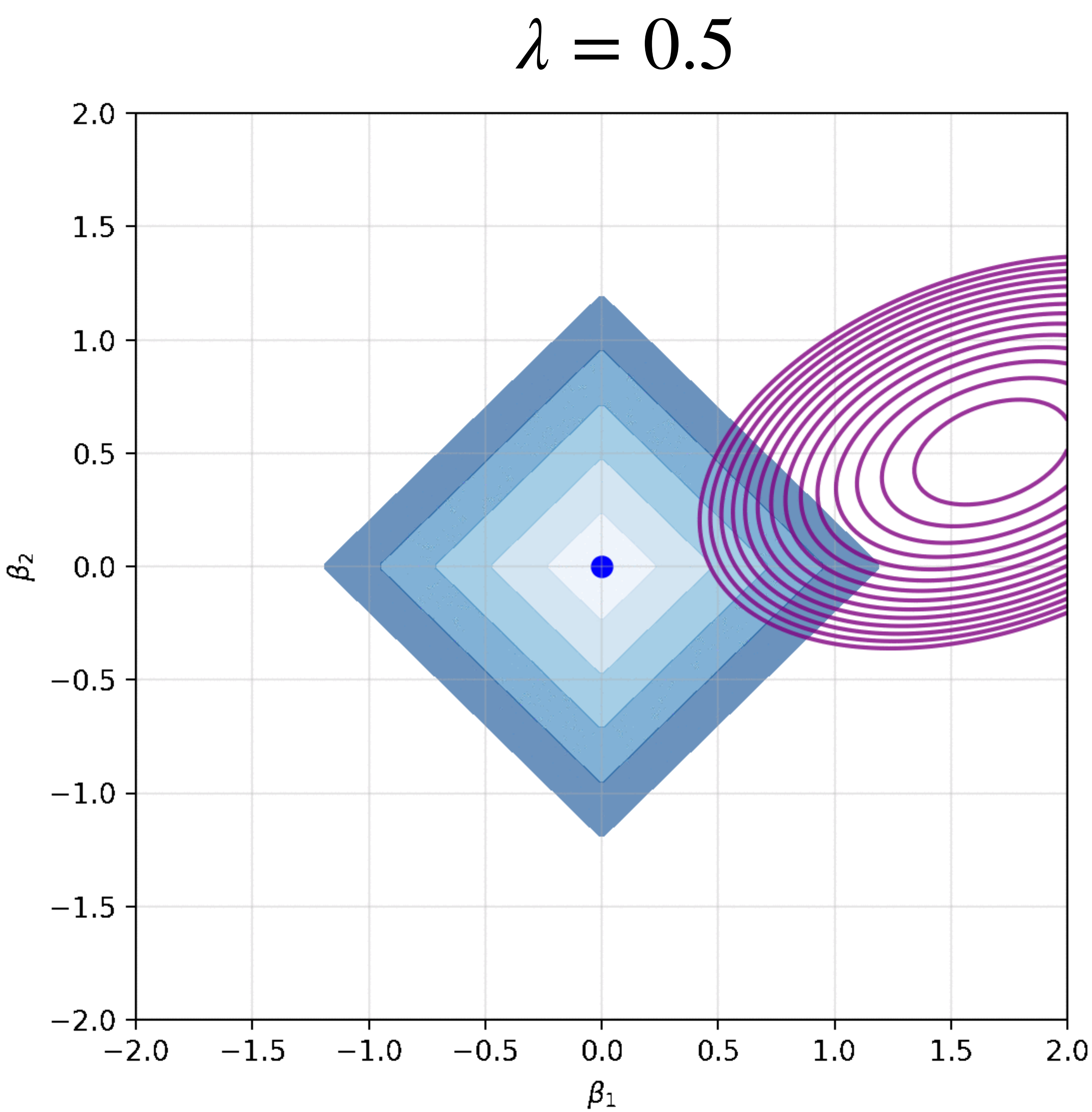
$y_i = \langle \boldsymbol{\theta}_\star, \boldsymbol{x}_i \rangle + \varepsilon_i$

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

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

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

$\eta = 0.1$



LASSO in practice

$n = 10$

$d = 2$

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

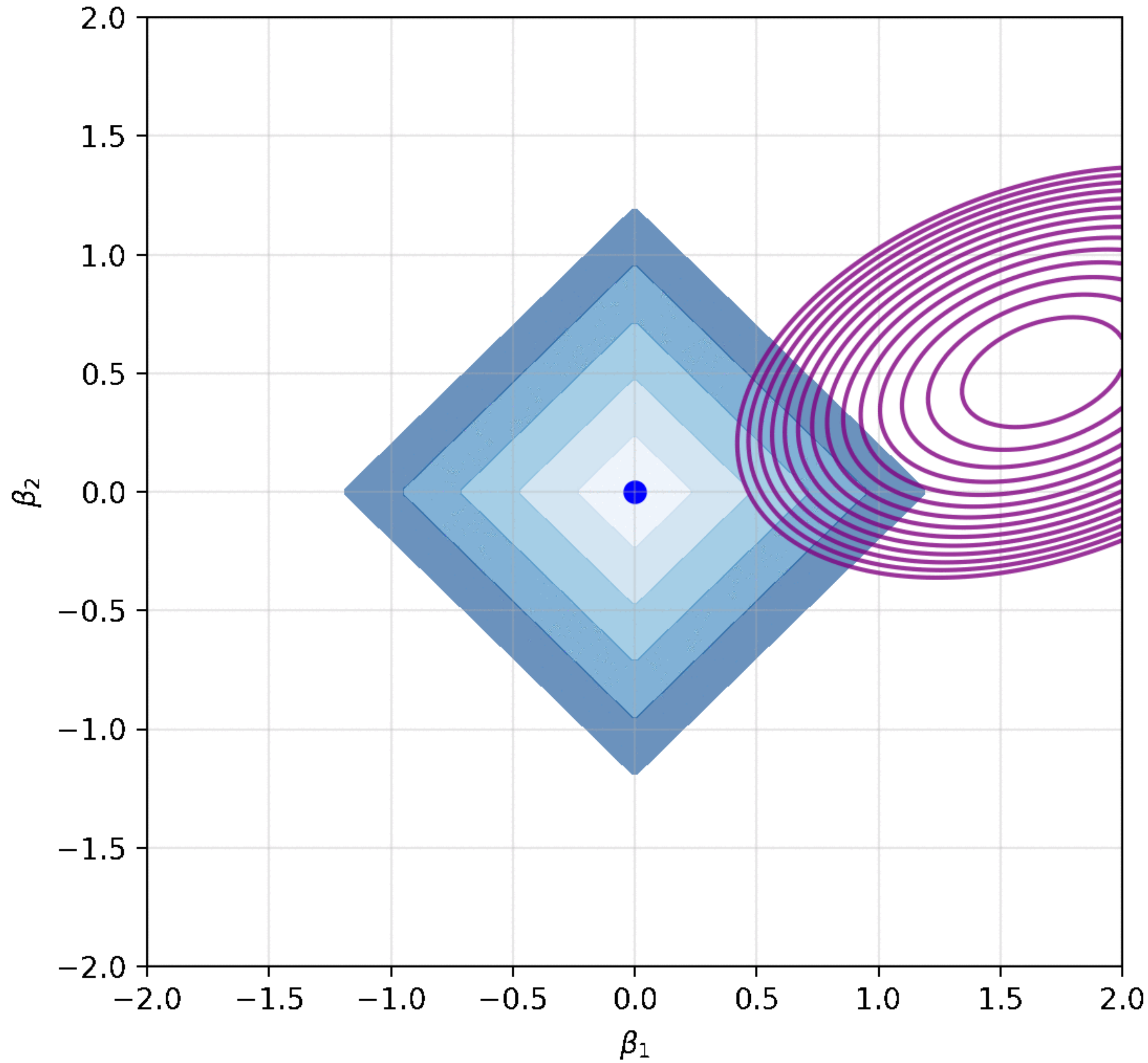
$\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{I}_2)$

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

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

$\eta = 0.1$

$\lambda = 0.1$



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.

Feature maps

Motivation

Up to know, our focus has been on parametric functions $f_\theta(\boldsymbol{x})$ which are linear on both $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{x} \in \mathbb{R}^d$.

$$f_\theta(\boldsymbol{x}) = \langle \boldsymbol{\theta}, \boldsymbol{x} \rangle$$

Motivation

Up to know, our focus has been on parametric functions $f_\theta(\boldsymbol{x})$ which are linear on both $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{x} \in \mathbb{R}^d$.

$$f_\theta(\boldsymbol{x}) = \langle \boldsymbol{\theta}, \boldsymbol{x} \rangle$$

The main convenience of linear functions is that for convex loss functions, the ERM problem is convex:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(\boldsymbol{x}))$$

Motivation

Up to know, our focus has been on parametric functions $f_\theta(\boldsymbol{x})$ which are linear on both $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{x} \in \mathbb{R}^d$.

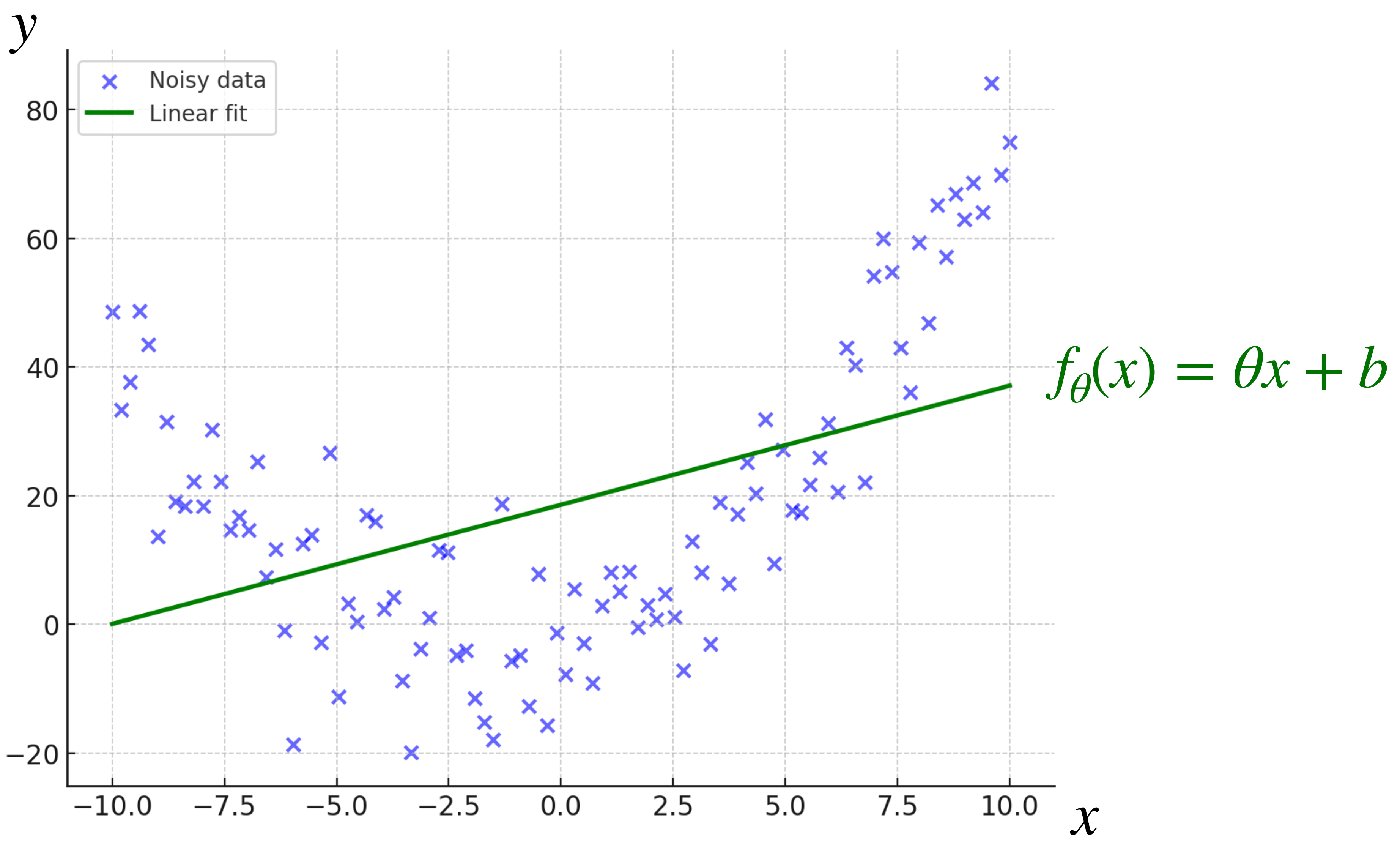
$$f_\theta(\boldsymbol{x}) = \langle \boldsymbol{\theta}, \boldsymbol{x} \rangle$$

The main convenience of linear functions is that for convex loss functions, the ERM problem is convex:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(\boldsymbol{x}))$$

But the main drawback is that we can only express linear relationships between the covariates and the labels...

Motivation



Noisy data
Linear fit
80
60
40
20
0
−20
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
y
x
$f_\theta(x) = \theta x + b$

Feature maps

Idea: Introduce a feature map:

$$
\begin{aligned}
\boldsymbol{\varphi} : \mathbb{R}^d &\to \mathbb{R}^p \\
\boldsymbol{x} &\mapsto \boldsymbol{\varphi}(\boldsymbol{x})
\end{aligned}
$$

And consider a linear predictor in feature space:

$$
f_\theta(\boldsymbol{x}) = \langle \boldsymbol{\theta}, \boldsymbol{\varphi}(\boldsymbol{x}) \rangle
$$

Feature maps

Idea: Introduce a feature map:

$$\begin{aligned} \boldsymbol{\varphi} : \mathbb{R}^d &\to \mathbb{R}^p \\ \boldsymbol{x} &\mapsto \boldsymbol{\varphi}(\boldsymbol{x}) \end{aligned}$$

And consider a linear predictor in feature space:

$$f_\theta(\boldsymbol{x}) = \langle \boldsymbol{\theta}, \boldsymbol{\varphi}(\boldsymbol{x}) \rangle$$

- Now we have $\boldsymbol{\theta} \in \mathbb{R}^p$.
- f_θ still a linear function of $\boldsymbol{\theta}$.
- Typically $p > d$.
- More generally, we can consider $\boldsymbol{\varphi} : \mathscr{X} \to \mathbb{R}^p$

Feature maps

Idea: Introduce a feature map:

$$\begin{aligned} \boldsymbol{\varphi} : \mathbb{R}^d &\to \mathbb{R}^p \\ \boldsymbol{x} &\mapsto \boldsymbol{\varphi}(\boldsymbol{x}) \end{aligned}$$

And consider a linear predictor in feature space:

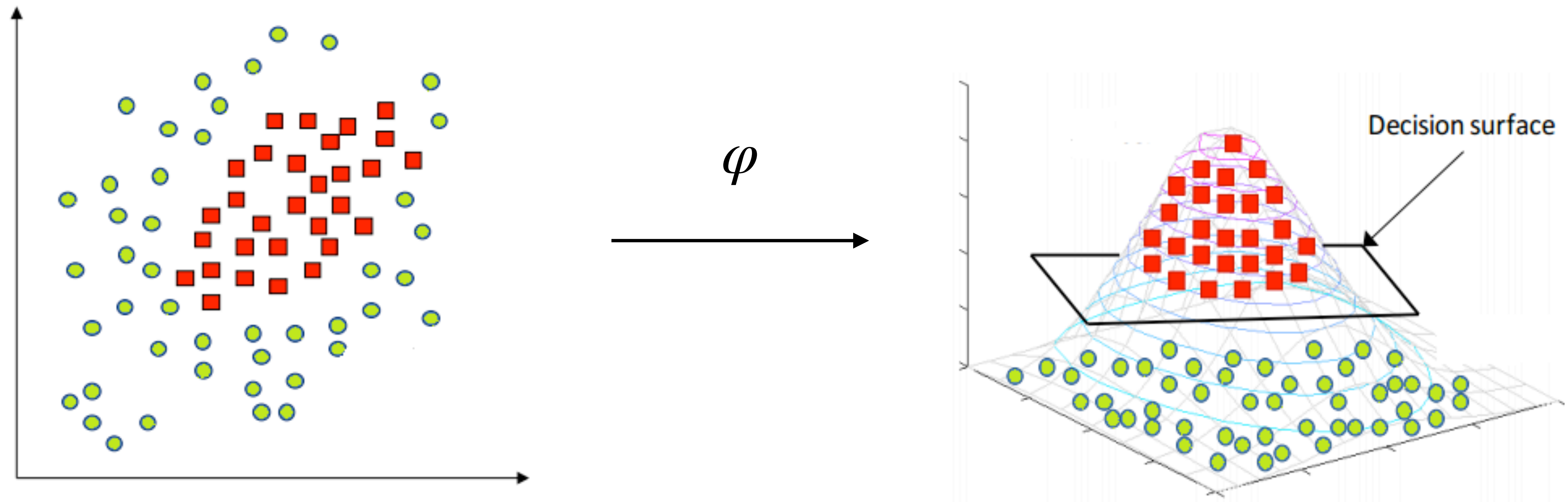
$$f_\theta(\boldsymbol{x}) = \langle \boldsymbol{\theta}, \boldsymbol{\varphi}(\boldsymbol{x}) \rangle$$

- Now we have $\boldsymbol{\theta} \in \mathbb{R}^p$.
- f_θ still a linear function of $\boldsymbol{\theta}$.
- Typically $p > d$.
- More generally, we can consider $\boldsymbol{\varphi} : \mathscr{X} \to \mathbb{R}^p$

Example: $\mathscr{X}$ a collection of books.

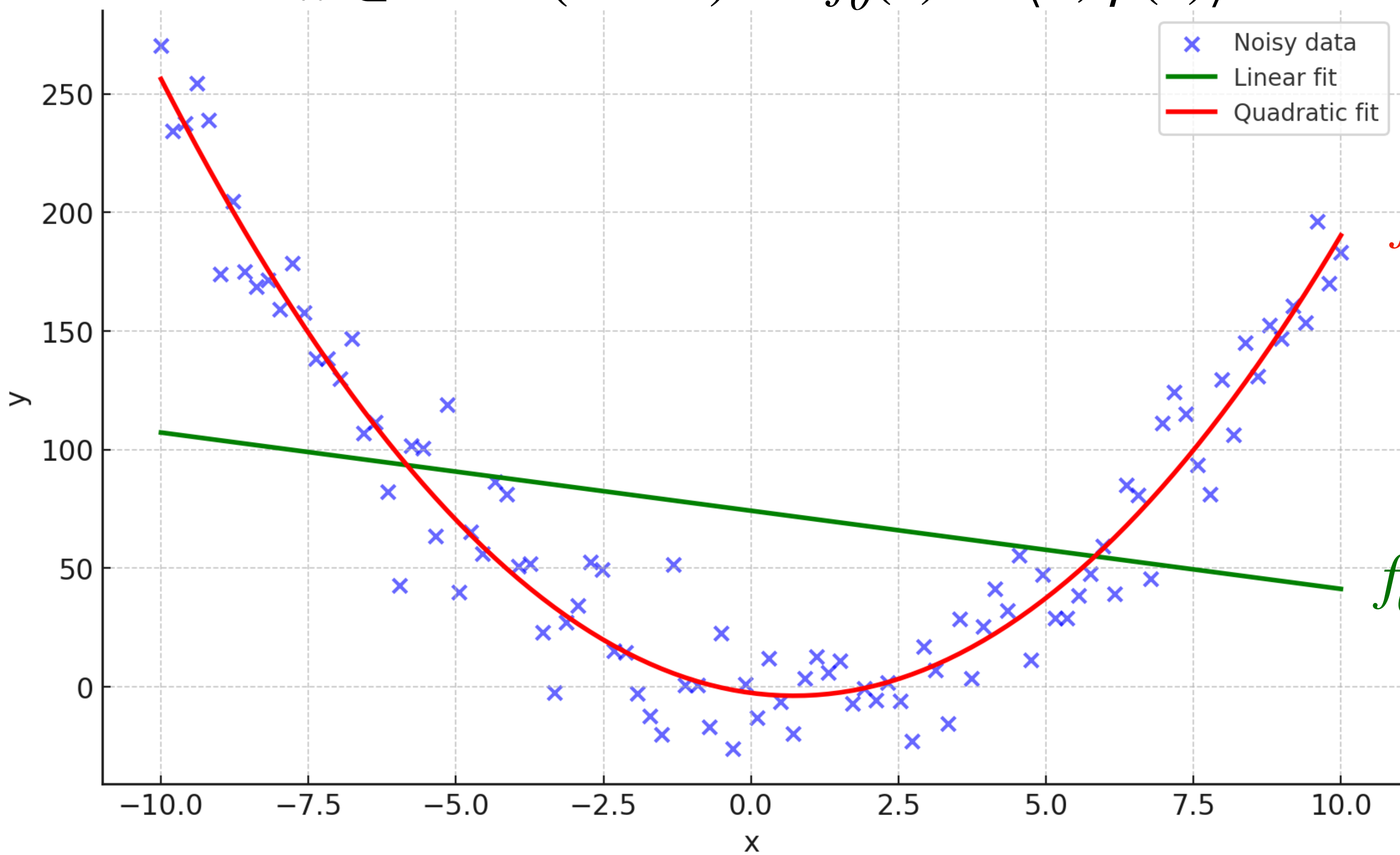
Feature maps

Intuition: Typically easier to linearly separate data in higher-dimensions



Examples: quadratic function

$x \in \mathbb{R} \qquad (d = 1) \qquad f_\theta(x) = \langle \theta, \varphi(x) \rangle + b$



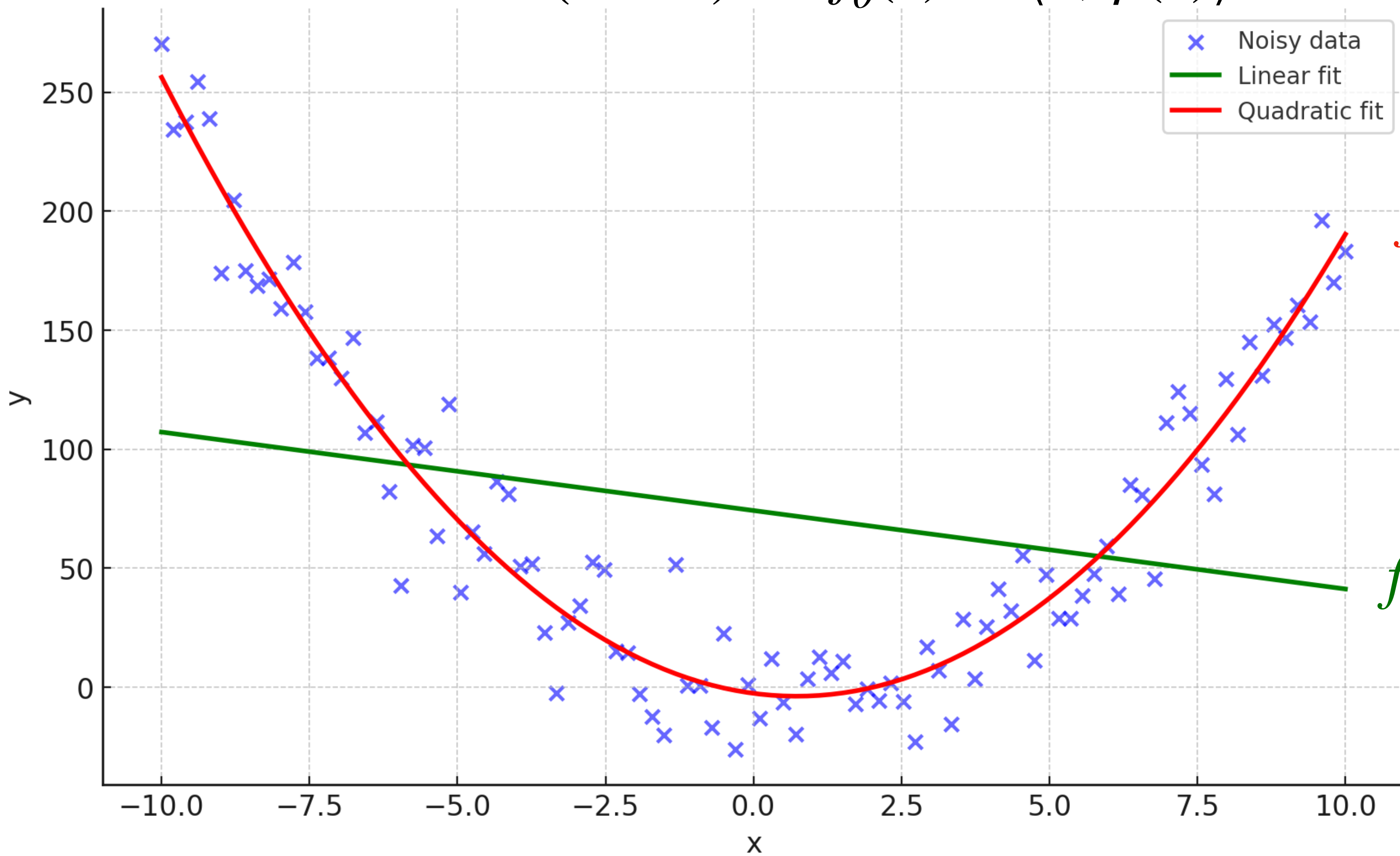
$f_\theta(x) = \theta_1 x^2 + \theta_2 x + b$

$f_\theta(x) = \theta x + b$

<u>Question</u>: what is $\boldsymbol{\varphi}(x)$?

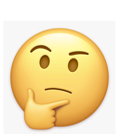
Examples: quadratic function

$x \in \mathbb{R} \qquad (d = 1) \qquad f_\theta(x) = \langle \theta, \varphi(x) \rangle + b$



$f_\theta(x) = \theta_1 x^2 + \theta_2 x + b$

$f_\theta(x) = \theta x + b$



<u>Question</u>: what is $\boldsymbol{\varphi}(x)$? $\quad \boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \qquad \boldsymbol{\varphi}(x) = \begin{bmatrix} x^2 \\ x \end{bmatrix} \qquad (p = 2)$

Polynomial regression

More generally, any polynomial of degree $k \in \mathbb{N}$ over $\mathbb{R}$

$$p(x) = \sum_{j=1}^{k} \theta_j x^j + b = \theta_k x^k + \theta_{k-1} x^{k-1} + \ldots + \theta_1 x + b$$

Polynomial regression

More generally, any polynomial of degree $k \in \mathbb{N}$ over $\mathbb{R}$

$$p(x) = \sum_{j=1}^{k} \theta_j x^j + b = \theta_k x^k + \theta_{k-1} x^{k-1} + \ldots + \theta_1 x + b$$

Can be written as a linear function in $\mathbb{R}^k$:

$$p(x) = \langle \boldsymbol{\theta}, \boldsymbol{\varphi}(x) \rangle + b \qquad\qquad \boldsymbol{\varphi}(x) = \begin{bmatrix} x \\ x^2 \\ \vdots \\ x^k \end{bmatrix} \in \mathbb{R}^k$$

Polynomial regression

More generally, any polynomial of degree $k \in \mathbb{N}$ over $\mathbb{R}$

$$p(x) = \sum_{j=1}^{k} \theta_j x^j + b = \theta_k x^k + \theta_{k-1} x^{k-1} + \ldots + \theta_1 x + b$$

Can be written as a linear function in $\mathbb{R}^k$:

$$p(x) = \langle \boldsymbol{\theta}, \boldsymbol{\varphi}(x) \rangle + b \qquad \boldsymbol{\varphi}(x) = \begin{bmatrix} x \\ x^2 \\ \vdots \\ x^k \end{bmatrix} \in \mathbb{R}^k$$

We can generalise this to degree k polynomials in $\mathbb{R}^d$:

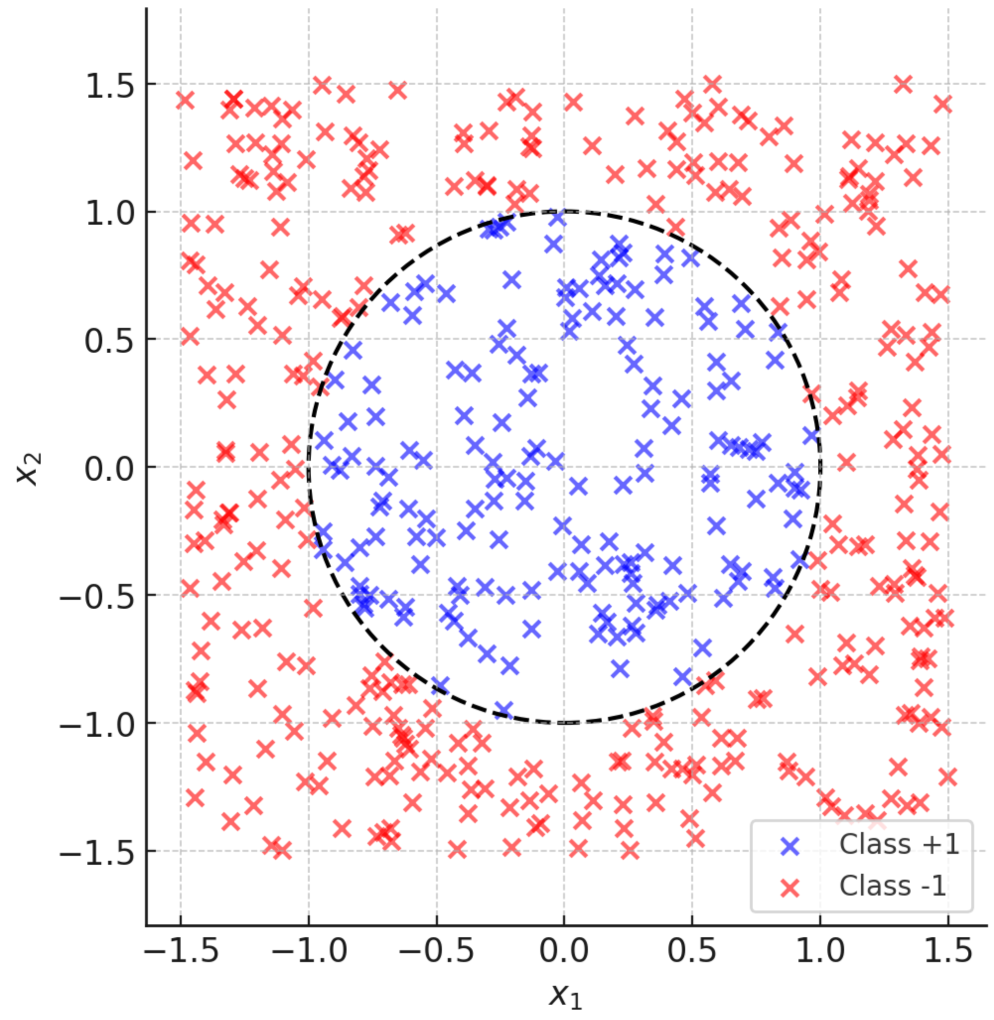
Example $d = 2$:

$$p(\boldsymbol{x}) = \langle \boldsymbol{\theta}, \boldsymbol{\varphi}(\boldsymbol{x}) \rangle + b \qquad \boldsymbol{\varphi}(x) = \begin{bmatrix} x_1 \\ x_2 \\ x_1^2 \\ x_1 x_2 \\ x_2^2 \end{bmatrix} \in \mathbb{R}^5$$

Examples: data in circle

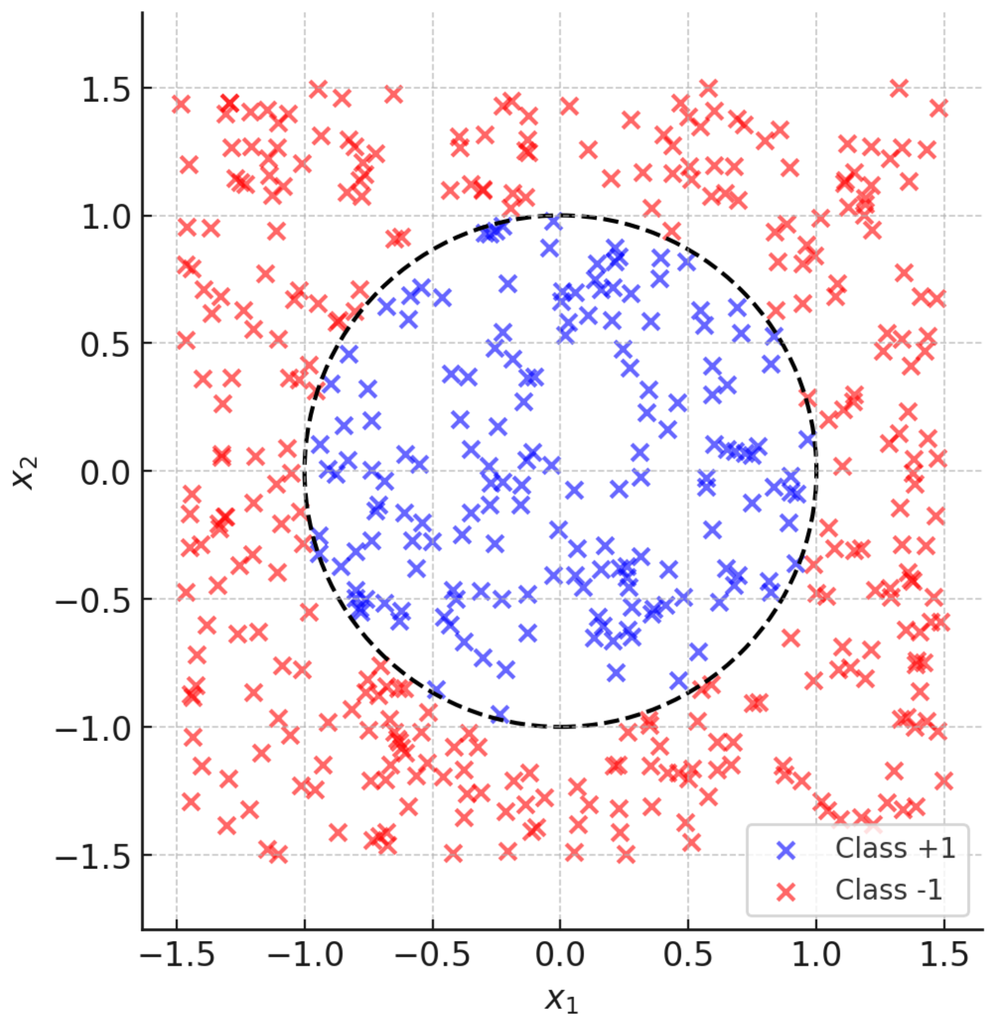
$$x \in \mathbb{R}^2 \quad (d = 2)$$

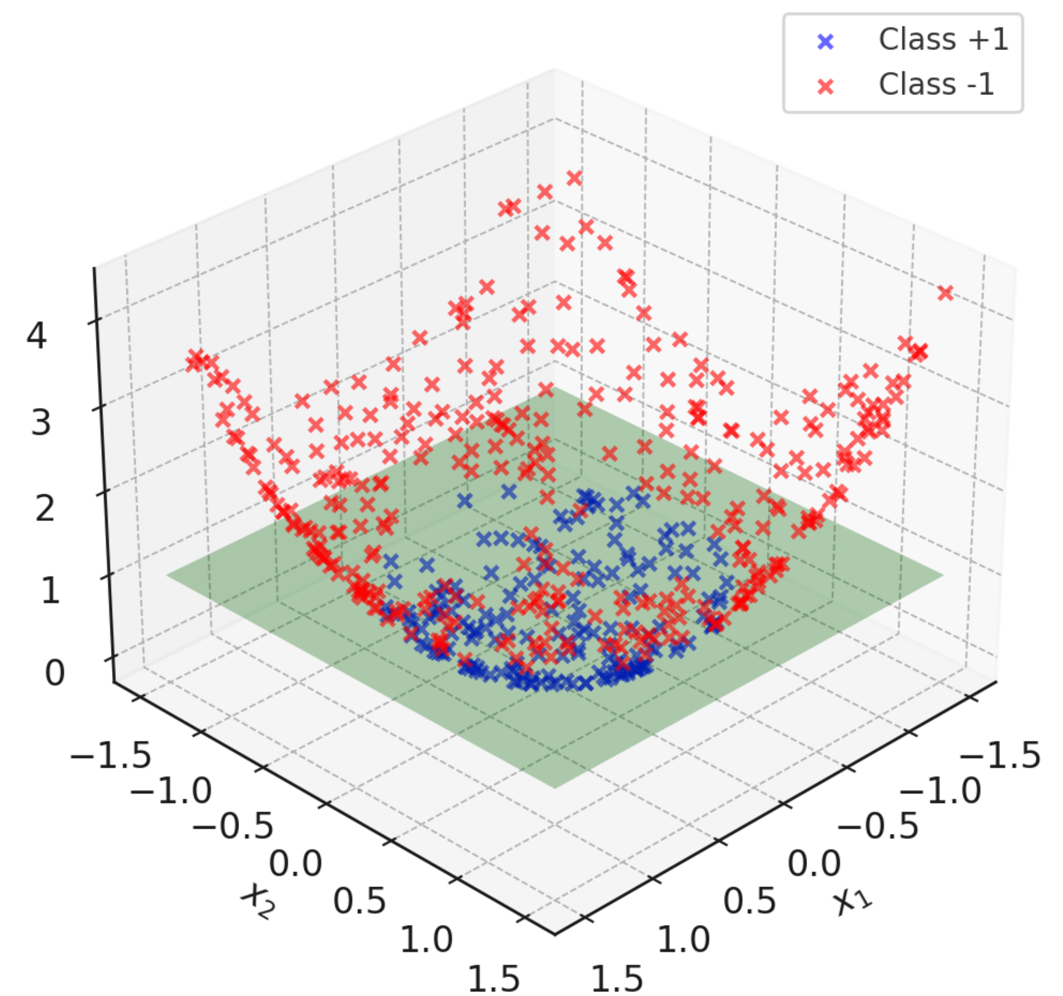
$$y = \begin{cases} +1 & \text{if } x_1^2 + x_2^2 \leq 1 \\ -1 & \text{if } x_1^2 + x_2^2 > 1 \end{cases}$$



Examples: data in circle

$$x \in \mathbb{R}^2 \quad (d = 2) \qquad y = \begin{cases} +1 & \text{if } x_1^2 + x_2^2 \le 1 \\ -1 & \text{if } x_1^2 + x_2^2 > 1 \end{cases}$$

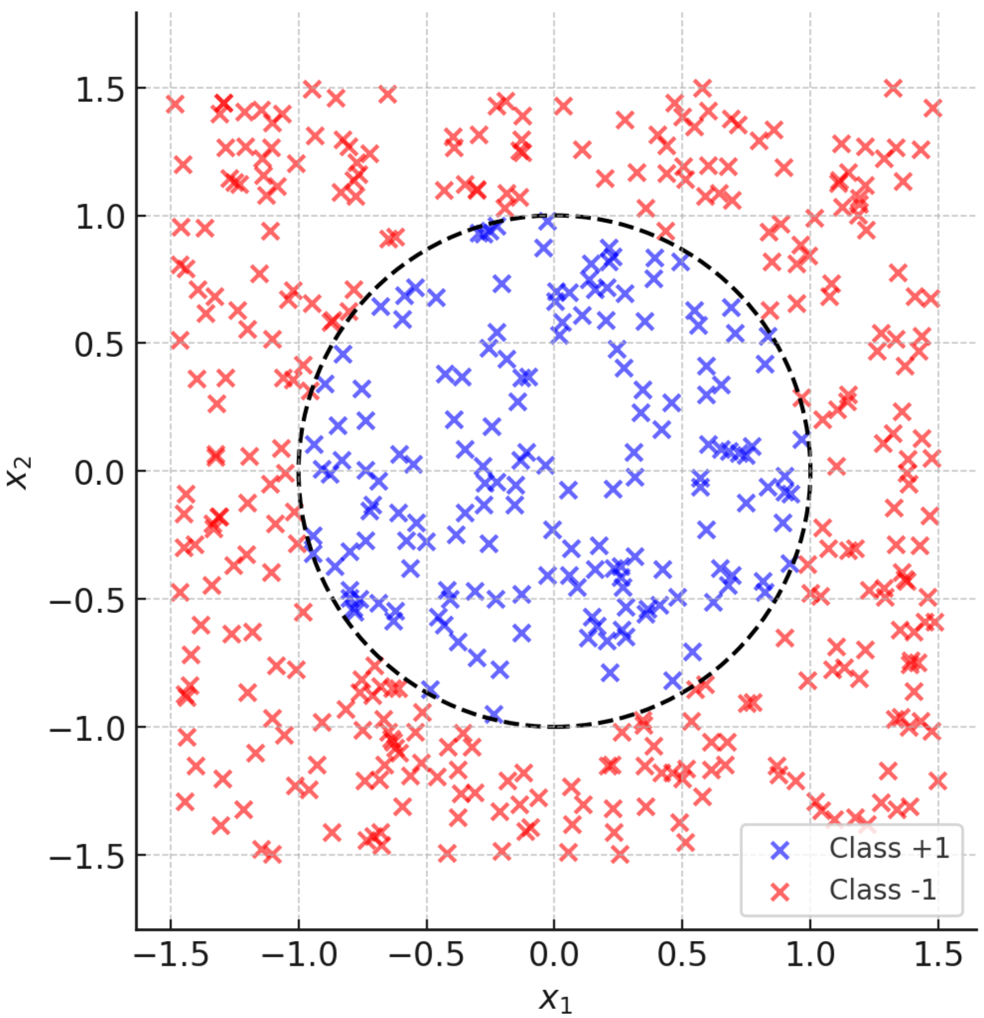


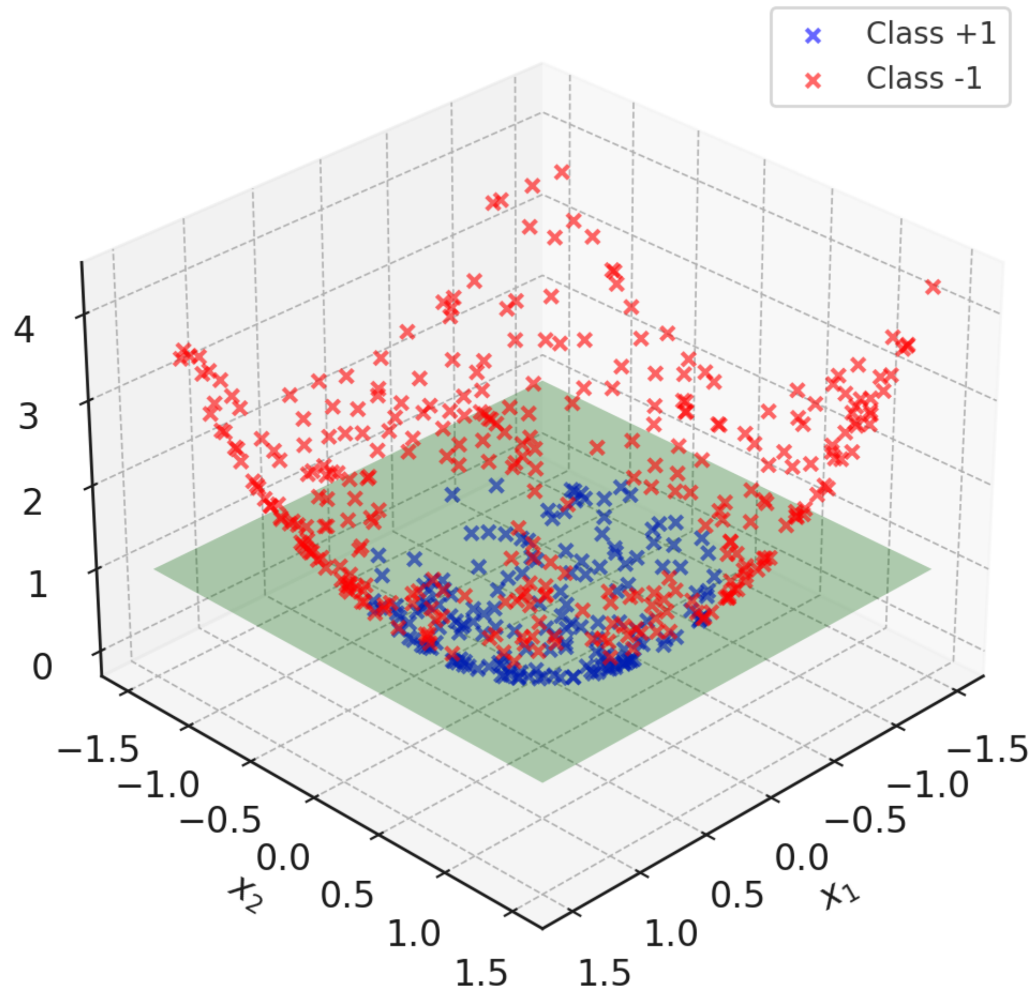


$$\boldsymbol{\varphi}(x) = \begin{bmatrix} x_1 \\ x_2 \\ x_1^2 + x_2^2 \end{bmatrix} \qquad (p = 3)$$

Examples: data in circle

$$x \in \mathbb{R}^2 \quad (d = 2) \qquad y = \begin{cases} +1 & \text{if } x_1^2 + x_2^2 \leq 1 \\ -1 & \text{if } x_1^2 + x_2^2 > 1 \end{cases}$$





$$\boldsymbol{\varphi}(x) = \begin{bmatrix} x_1 \\ x_2 \\ x_1^2 + x_2^2 \end{bmatrix} \qquad (p = 3)$$

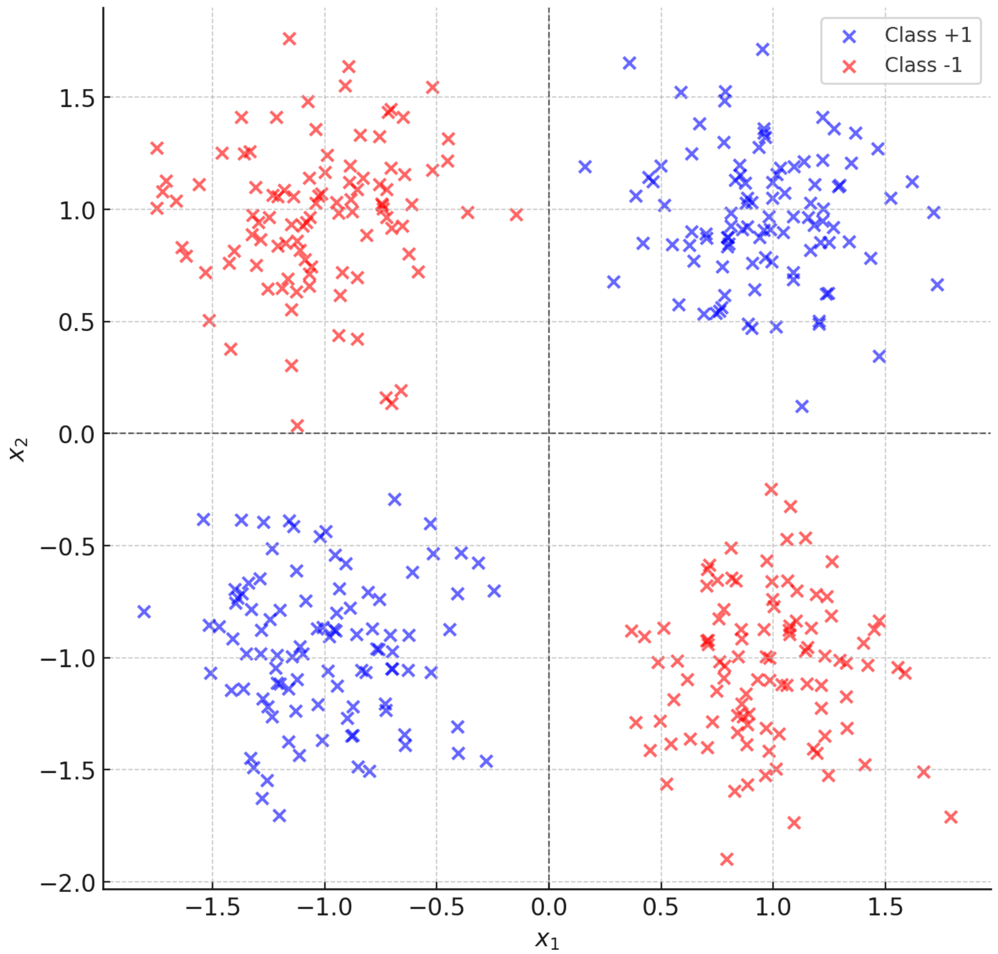
⚠ Not unique!

Examples: XOR Gaussian mixture

$x \in \mathbb{R}^2 \quad (d = 2)$

$$p(\boldsymbol{x}) = \frac{1}{4} \sum_{k=1}^{4} \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{I}_2)$$

$$\mu_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \qquad \mu_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

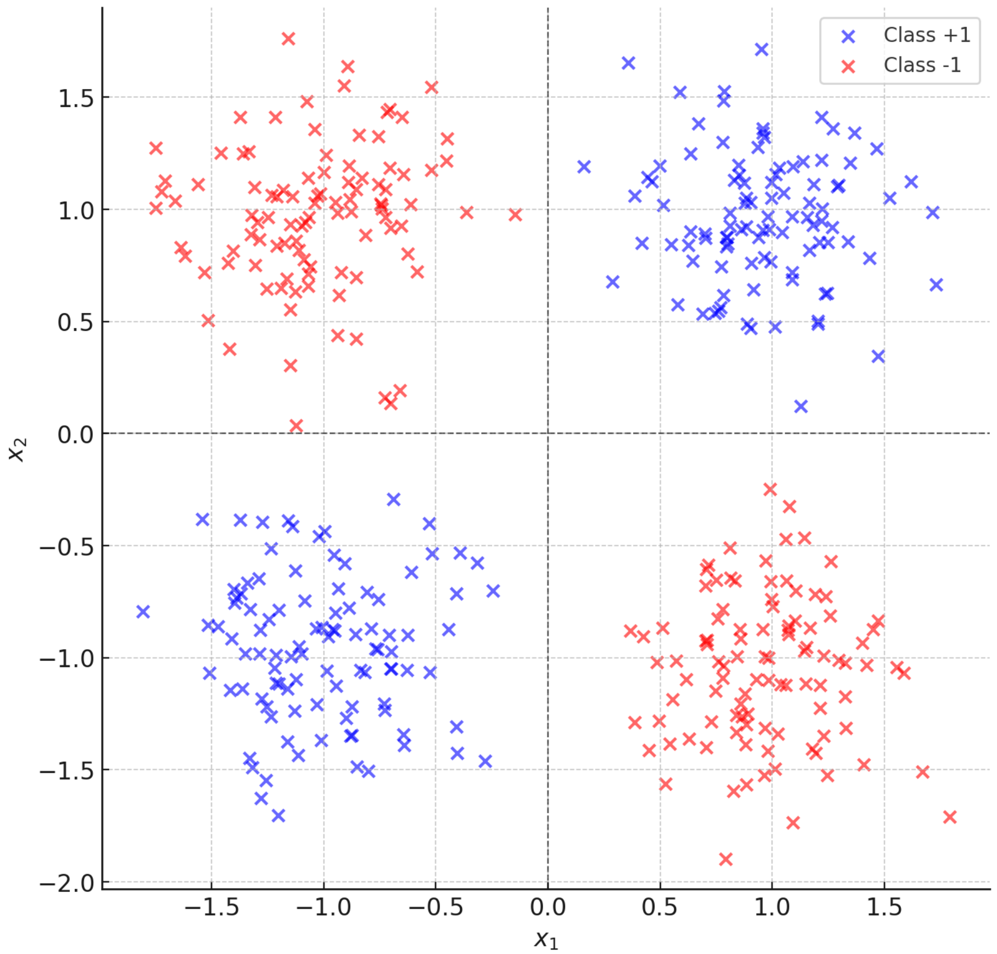


$$\mu_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \qquad \mu_4 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Examples: XOR Gaussian mixture

$x \in \mathbb{R}^2 \quad (d = 2)$

$$p(\boldsymbol{x}) = \frac{1}{4} \sum_{k=1}^{4} \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{I}_2)$$



Note that:

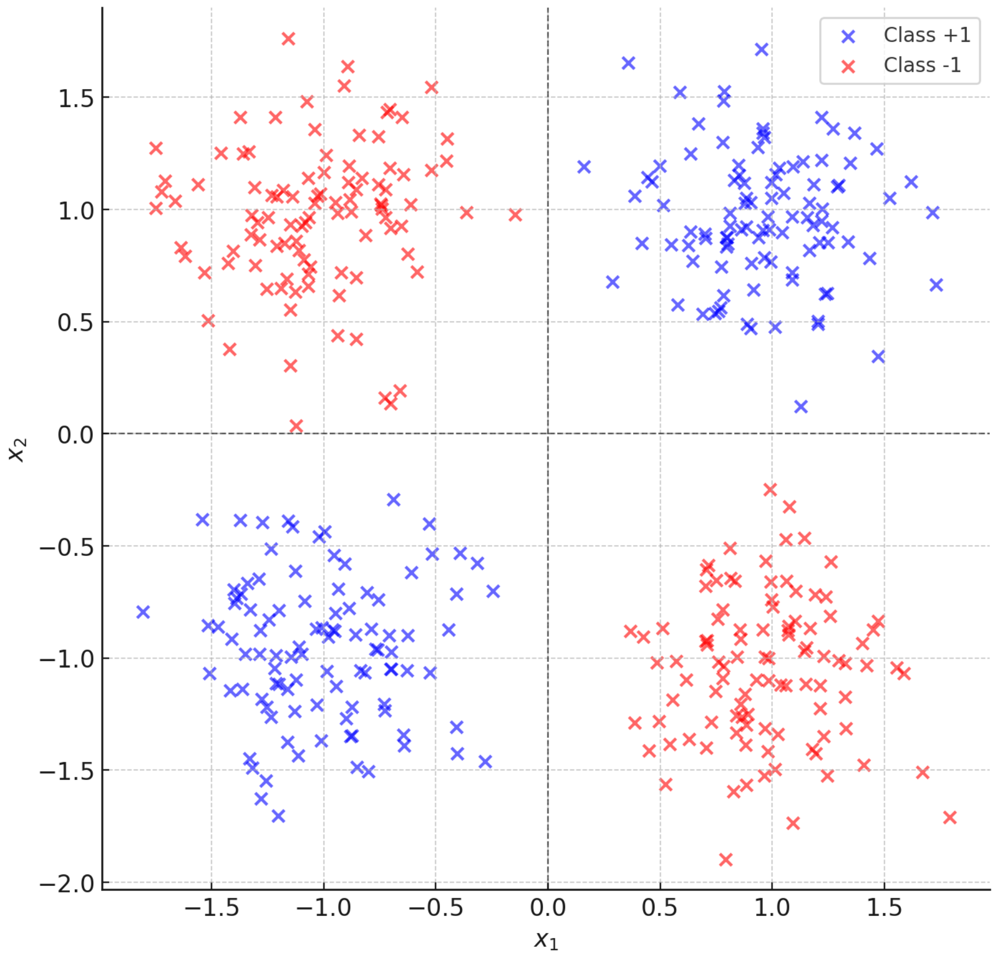
$y = +1$ $\quad x_1, x_2 > 0$ or $x_1, x_2 < 0$

$y = -1$ $\quad x_1 > 0$ and $x_2 < 0$ or
$x_1 < 0$ and $x_2 > 0$

Examples: XOR Gaussian mixture

$x \in \mathbb{R}^2 \quad (d = 2)$

$$p(\boldsymbol{x}) = \frac{1}{4}\sum_{k=1}^{4} \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{I}_2)$$



Note that:

$y = +1 \qquad x_1, x_2 > 0$ or $x_1, x_2 < 0$

or… $\quad x_1 x_2 > 0$

$y = -1 \qquad x_1 > 0$ and $x_2 < 0$ or

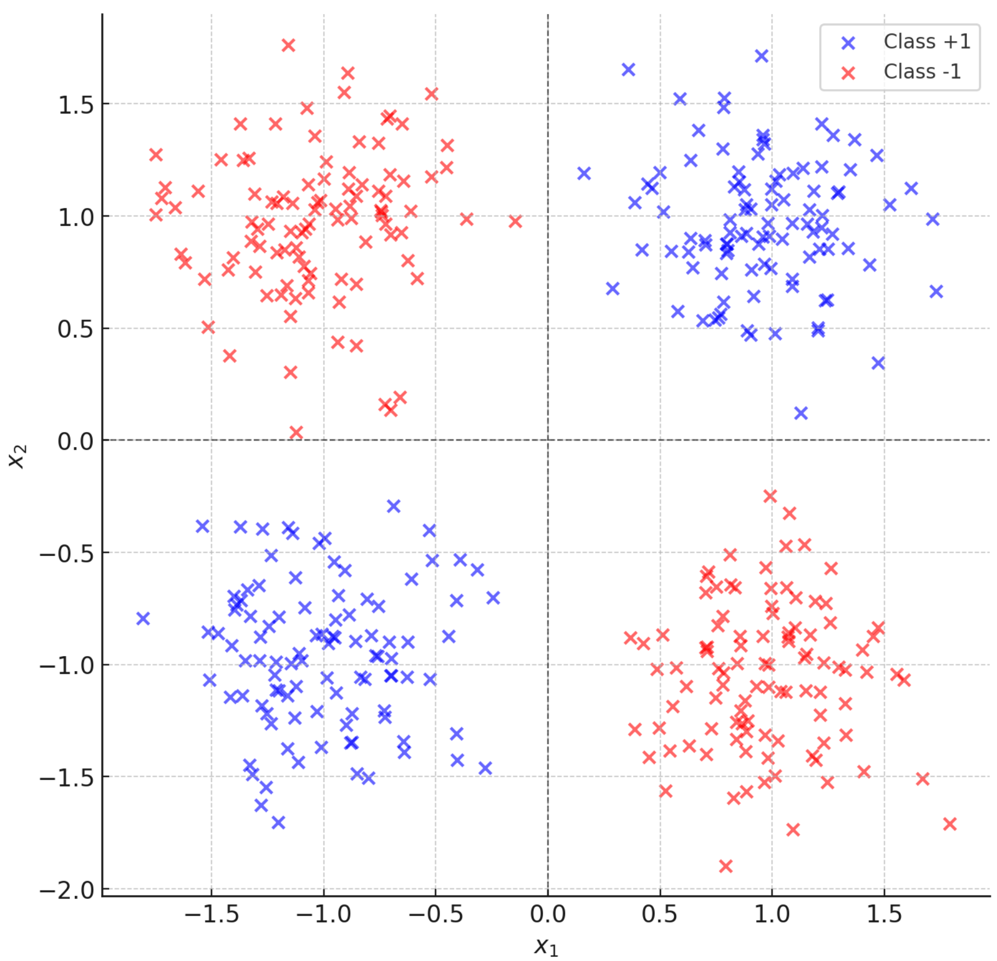
$x_1 < 0$ and $x_2 > 0$

or… $\quad x_1 x_2 < 0$

Examples: XOR Gaussian mixture

$x \in \mathbb{R}^2 \quad (d = 2)$

$$p(\boldsymbol{x}) = \frac{1}{4} \sum_{k=1}^{4} \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{I}_2)$$



Note that:

$y = +1 \qquad x_1, x_2 > 0$ or $x_1, x_2 < 0$

or... $\quad x_1 x_2 > 0$

$y = -1 \qquad x_1 > 0$ and $x_2 < 0$ or

$x_1 < 0$ and $x_2 > 0$

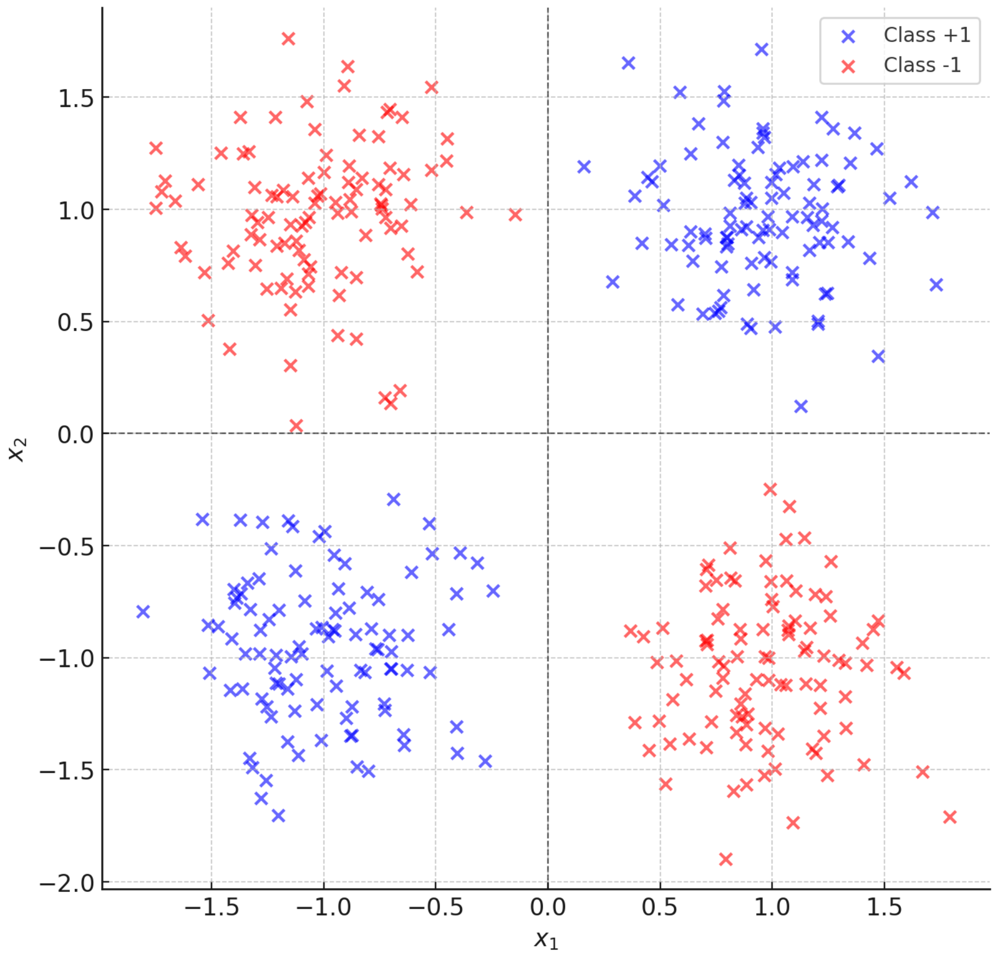
or... $\quad x_1 x_2 < 0$

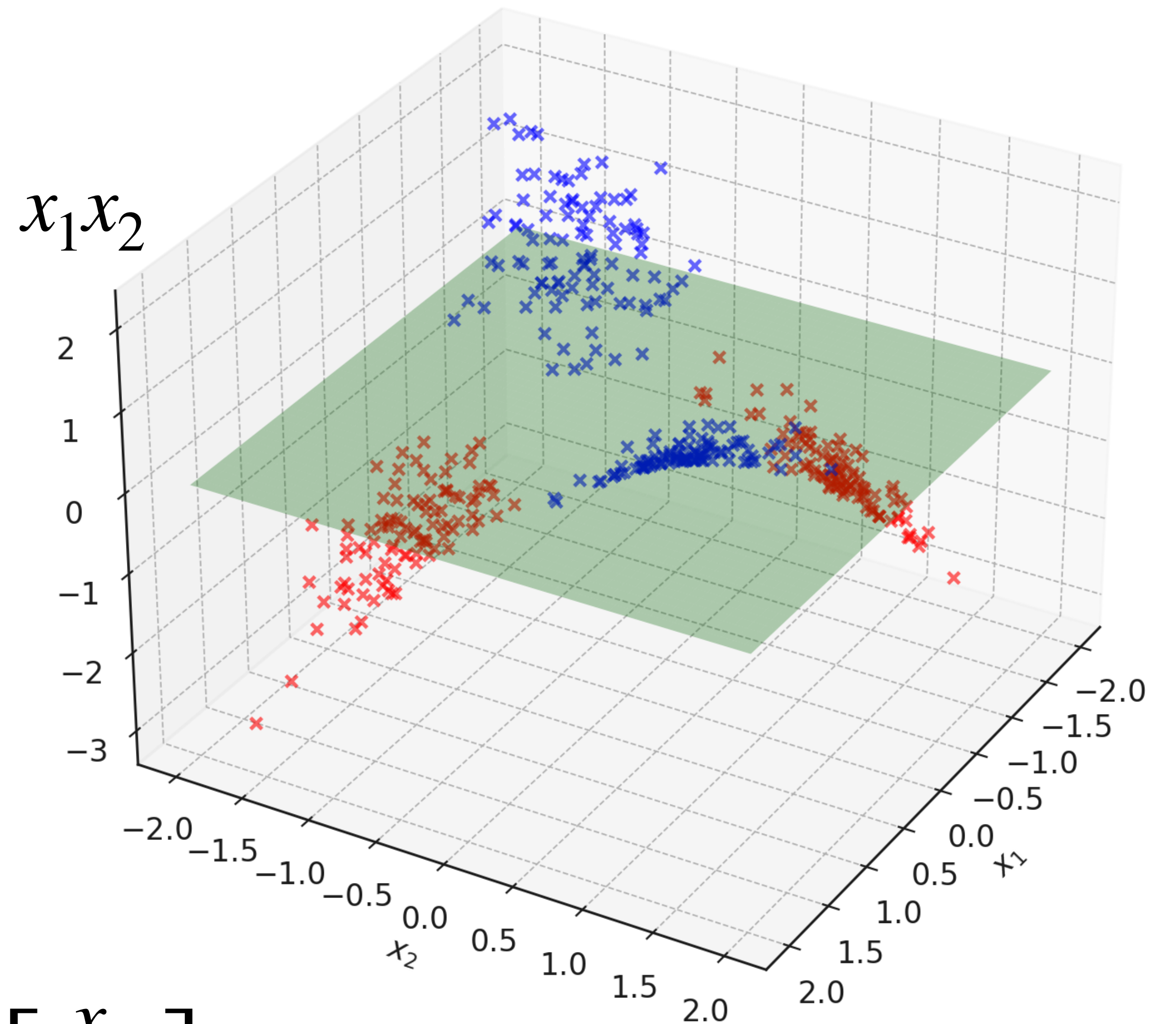
This motivates a choice: $\quad \boldsymbol{\varphi}(\boldsymbol{x}) = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \end{bmatrix} \quad (p = 3)$

Examples: XOR Gaussian mixture

$x \in \mathbb{R}^2 \quad (d = 2)$

$$p(\boldsymbol{x}) = \frac{1}{4}\sum_{k=1}^{4} \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{I}_2)$$





$$\boldsymbol{\varphi}(\boldsymbol{x}) = \begin{bmatrix} x_1 \\ x_2 \\ x_1x_2 \end{bmatrix}$$