Statistical Learning II

Lecture 11 - PCA & Feature maps

Bruno Loureiro
@ CSD, DI-ENS & CNRS

brloureiro@gmail.com

DL3 IASO, Université Paris Dauphine-PSL
04.12.2024

Principal component analysis (PCA)

Variance reduction

- As we saw in Lecture 6, the OLS estimator suffers from high-variance in directions with small singular values.

$$\hat{\boldsymbol{\theta}}_{OLS}(\boldsymbol{X}, \boldsymbol{y}) = \boldsymbol{\theta}_\star + \sum_{j=1}^{d} \frac{1}{\sigma_j} \langle \boldsymbol{u}_j, \boldsymbol{\varepsilon} \rangle \boldsymbol{v}_j$$

Variance reduction

- As we saw in Lecture 6, the OLS estimator suffers from high-variance in directions with small singular values.

$$\hat{\boldsymbol{\theta}}_{OLS}(\boldsymbol{X}, \boldsymbol{y}) = \boldsymbol{\theta}_{\star} + \sum_{j=1}^{d} \frac{1}{\sigma_j} \langle \boldsymbol{u}_j, \boldsymbol{\varepsilon} \rangle \boldsymbol{v}_j$$

- In Lectures 7 to 10, we studied regularisation as a form to mitigate this problem. For instance, ridge regression:

$$\hat{\boldsymbol{\theta}}_{\lambda}(\boldsymbol{X}, \boldsymbol{y}) = \boldsymbol{\theta}_{\star} - \sum_{j=1}^{\mathrm{rank}(\boldsymbol{X})} \frac{\lambda}{\sigma_j^2 + n\lambda} \langle \boldsymbol{v}_j, \boldsymbol{\theta}_{\star} \rangle \boldsymbol{v}_j \sum_{j=1}^{\mathrm{rank}(\boldsymbol{X})} \frac{\sigma_j}{\sigma_j^2 + n\lambda} \langle \boldsymbol{u}_j, \boldsymbol{\varepsilon} \rangle \boldsymbol{v}_j$$

Variance reduction

- As we saw in Lecture 6, the OLS estimator suffers from high-variance in directions with small singular values.

$$\hat{\boldsymbol{\theta}}_{OLS}(\boldsymbol{X}, \boldsymbol{y}) = \boldsymbol{\theta}_\star + \sum_{j=1}^{d} \frac{1}{\sigma_j} \langle \boldsymbol{u}_j, \boldsymbol{\varepsilon} \rangle \boldsymbol{v}_j$$

- In Lectures 7 to 10, we studied regularisation as a form to mitigate this problem. For instance, ridge regression:

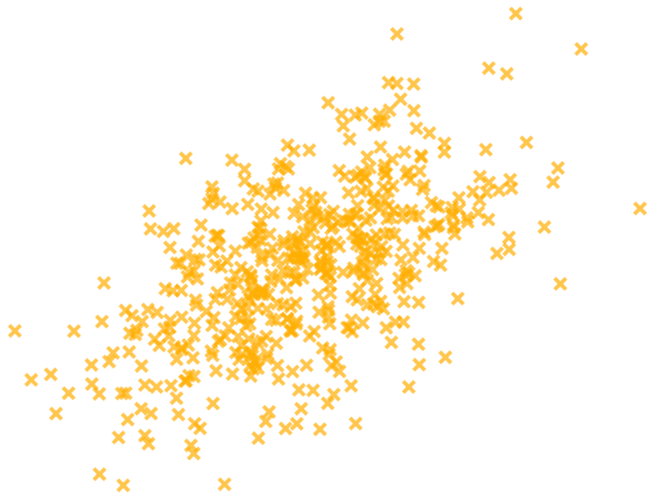
$$\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y}) = \boldsymbol{\theta}_\star - \sum_{j=1}^{\mathrm{rank}(\boldsymbol{X})} \frac{\lambda}{\sigma_j^2 + n\lambda} \langle \boldsymbol{v}_j, \boldsymbol{\theta}_\star \rangle \boldsymbol{v}_j \sum_{j=1}^{\mathrm{rank}(\boldsymbol{X})} \frac{\sigma_j}{\sigma_j^2 + n\lambda} \langle \boldsymbol{u}_j, \boldsymbol{\varepsilon} \rangle \boldsymbol{v}_j$$

🤔 What about getting rid of these directions directly?

Principal component analysis

Let $\boldsymbol{x}_i \in \mathbb{R}^d$ denote $i = 1,\ldots,n$ i.i.d. covariates. Define $\boldsymbol{X} \in \mathbb{R}^{n\times d}$.

Without loss of generality, assume data is centred.



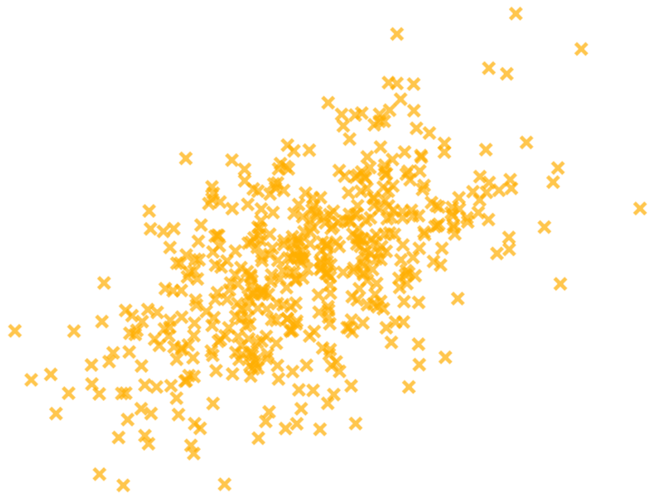
$n = 500$

$d = 2$

Principal component analysis

Let $\boldsymbol{x}_i \in \mathbb{R}^d$ denote $i = 1,\ldots,n$ i.i.d. covariates. Define $\boldsymbol{X} \in \mathbb{R}^{n\times d}$.

Without loss of generality, assume data is centred.



$n = 500$

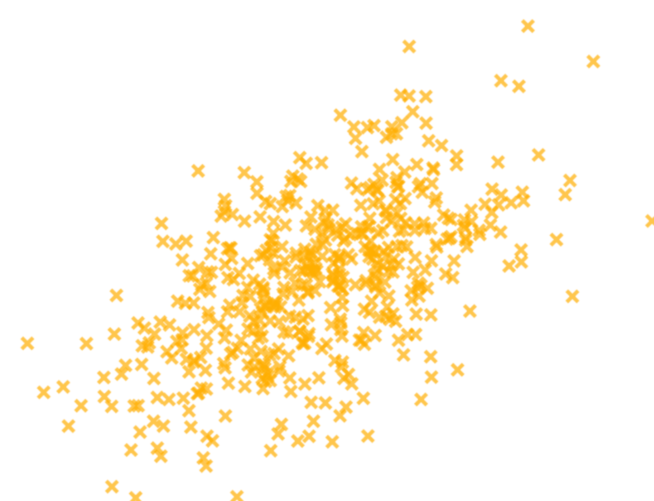
$d = 2$

Goal: find a lower dimensional approximation of $\boldsymbol{X}$.

Principal component analysis

Let $\boldsymbol{x}_i \in \mathbb{R}^d$ denote $i = 1, \ldots, n$ i.i.d. covariates. Define $\boldsymbol{X} \in \mathbb{R}^{n \times d}$.

Without loss of generality, assume data is centred.



$n = 500$

$d = 2$

Goal: find a lower dimensional approximation of $\boldsymbol{X}$.

Simplest case: find best k-dimensional linear approximation of $\boldsymbol{X}$.

Principal component analysis

Mathematically:

- Let $\boldsymbol{z}_i \in \mathbb{R}^d$, $i = 1, \ldots, n$ such that $\operatorname{span}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = \mathbb{R}^k$ with $k \leq d$

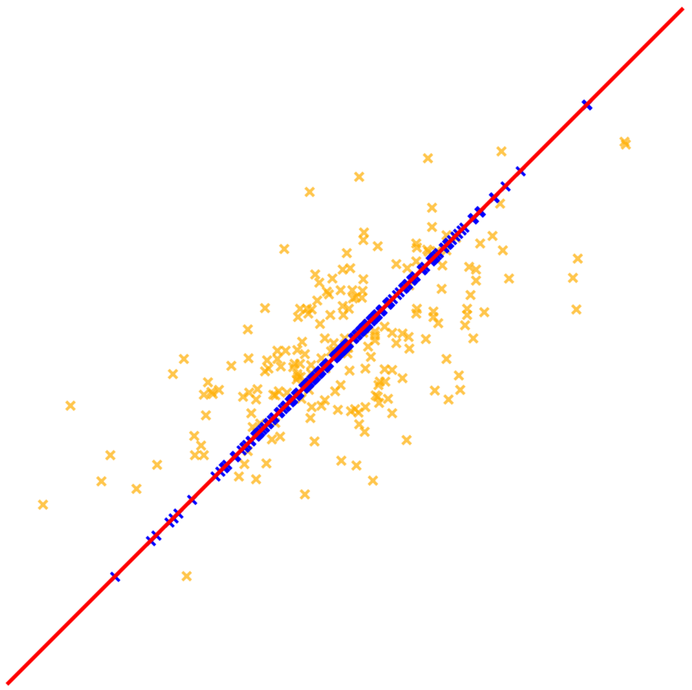
Principal component analysis

Mathematically:

- Let $z_i \in \mathbb{R}^d$, $i = 1,\ldots,n$ such that $\mathrm{span}(z_1, \ldots, z_n) = \mathbb{R}^k$ with $k \leq d$

The PCA problem consists of:

$$\min_{z_1,\ldots,z_n} \sum_{i=1}^{n} ||z_i - \boldsymbol{x}_i||_2^2$$



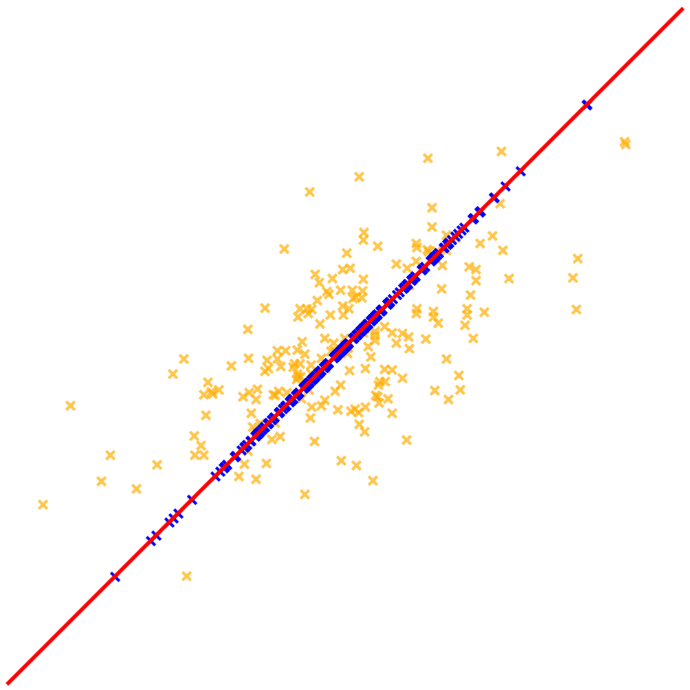
Principal component analysis

Mathematically:

- Let $z_i \in \mathbb{R}^d, i = 1,\ldots,n$ such that $\mathrm{span}(z_1, \ldots, z_n) = \mathbb{R}^k$ with $k \leq d$

The PCA problem consists of:

$$\min_{z_1,\ldots,z_n} \sum_{i=1}^{n} ||z_i - \boldsymbol{x}_i||_2^2$$



This can be equivalently written as:

$$\min_{\substack{\boldsymbol{Z} \in \mathbb{R}^{n\times d}, \\ \mathrm{rank}(Z) \leq k}} ||\boldsymbol{Z} - \boldsymbol{X}||_F^2$$

Principal component analysis

As we saw in Lecture 1, the solution to this problem is the SVD!

$$\underset{\substack{\boldsymbol{Z} \in \mathbb{R}^{n \times d}, \\ \text{rank}(Z) \leq k}}{\text{argmin}} \ ||\boldsymbol{Z} - \boldsymbol{X}||_F^2 = \sum_{j=1}^{k} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j$$

Principal component analysis

As we saw in Lecture 1, the solution to this problem is the SVD!

$$\underset{\substack{\boldsymbol{Z} \in \mathbb{R}^{n \times d}, \\ \text{rank}(Z) \le k}}{\text{argmin}} \ ||\boldsymbol{Z} - \boldsymbol{X}||_F^2 = \sum_{j=1}^{k} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j$$

In other words, the best k-dimensional linear approximation to the data consists of retaining only the top k singular values.

Principal component analysis

As we saw in Lecture 1, the solution to this problem is the SVD!

$$\underset{\substack{\boldsymbol{Z} \in \mathbb{R}^{n \times d}, \\ \operatorname{rank}(Z) \leq k}}{\operatorname{argmin}} \; ||\boldsymbol{Z} - \boldsymbol{X}||_F^2 = \sum_{j=1}^{k} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j$$

In other words, the best k-dimensional linear approximation to the data consists of retaining only the top k singular values.

This result holds for more general norms, and is known as the Eckart-Young-Minsky Theorem

Principal component analysis

As we saw in Lecture 1, the solution to this problem is the SVD!

$$\underset{\substack{\boldsymbol{Z} \in \mathbb{R}^{n \times d}, \\ \mathrm{rank}(Z) \leq k}}{\mathrm{argmin}} \; ||\boldsymbol{Z} - \boldsymbol{X}||_F^2 = \sum_{j=1}^{k} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j$$

In other words, the best k-dimensional linear approximation to the data consists of retaining only the top k singular values.

This result holds for more general norms, and is known as the Eckart-Young-Minsky Theorem

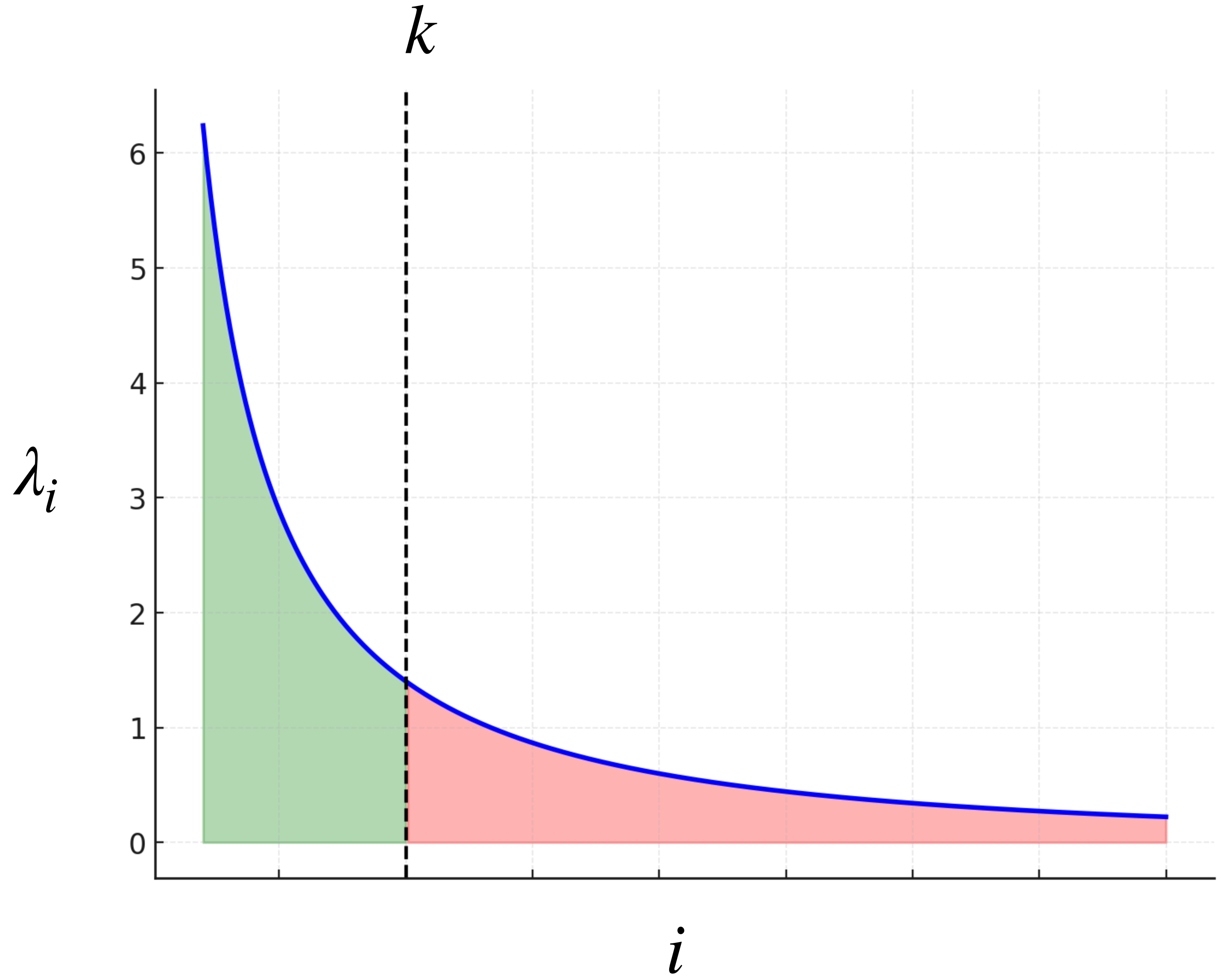
Remark: This is equivalent to keeping the k directions with largest variance in the data.

$$\mathrm{Tr}(\hat{\Sigma}_n) = \frac{1}{n} \sum_{i=1}^{\mathrm{rank}(\boldsymbol{X})} \lambda_i^2$$

PCA in practice

In practice, how to choose the k?

Total variance of data given by: $$\mathrm{Tr}(\hat{\Sigma}_n) = \frac{1}{n} \sum_{i=1}^{\mathrm{rank}(\boldsymbol{X})} \lambda_i^2$$



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 $\boldsymbol{x} \in \mathbb{R}^d$.

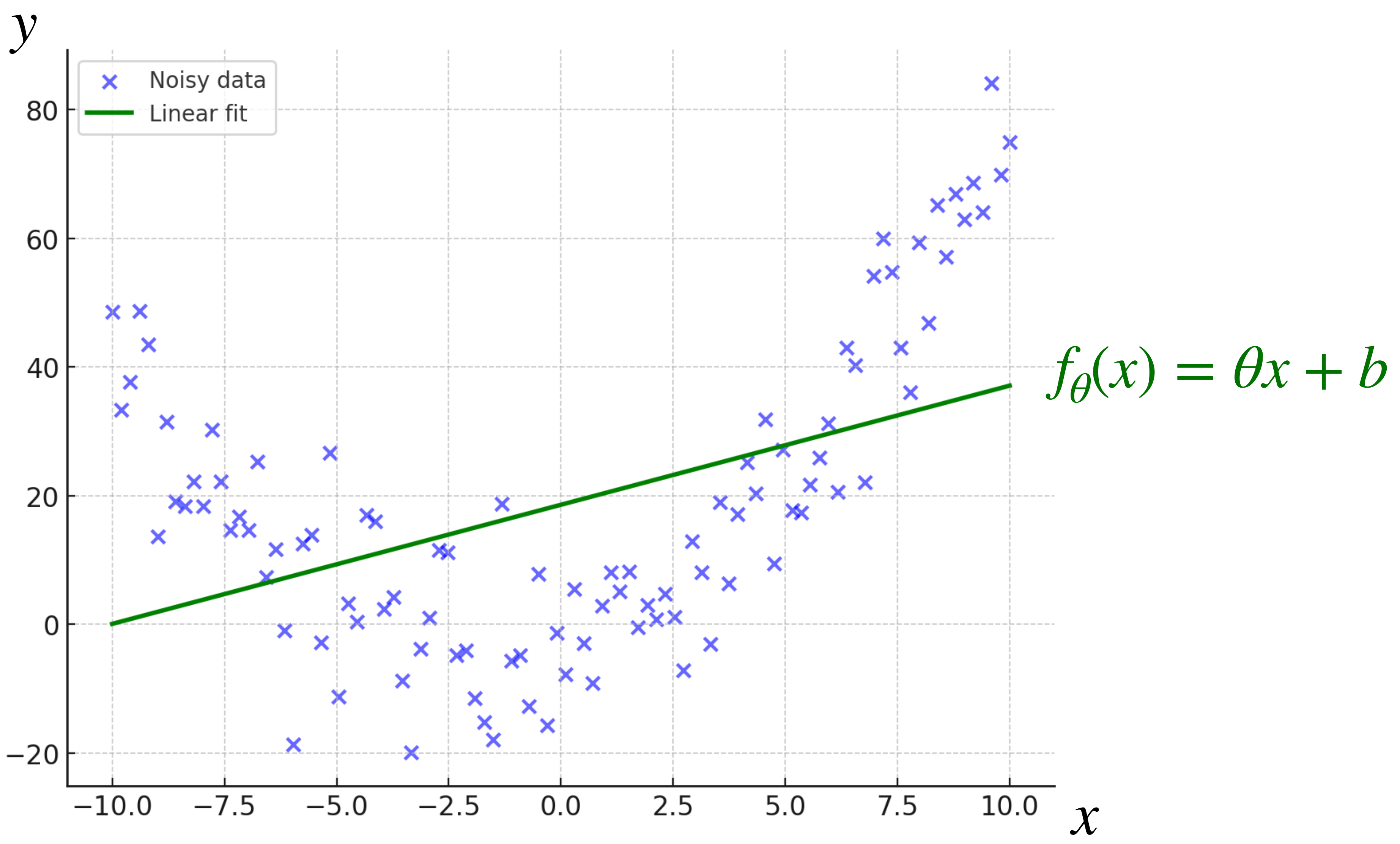
$$f_\theta(\boldsymbol{x}) = \langle \boldsymbol{\theta}, \boldsymbol{x} \rangle$$

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



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

Feature maps

Idea: Introduce a feature map:

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

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:

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

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 $\boldsymbol{\varphi} : \mathcal{X} \to \mathbb{R}^p$

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:

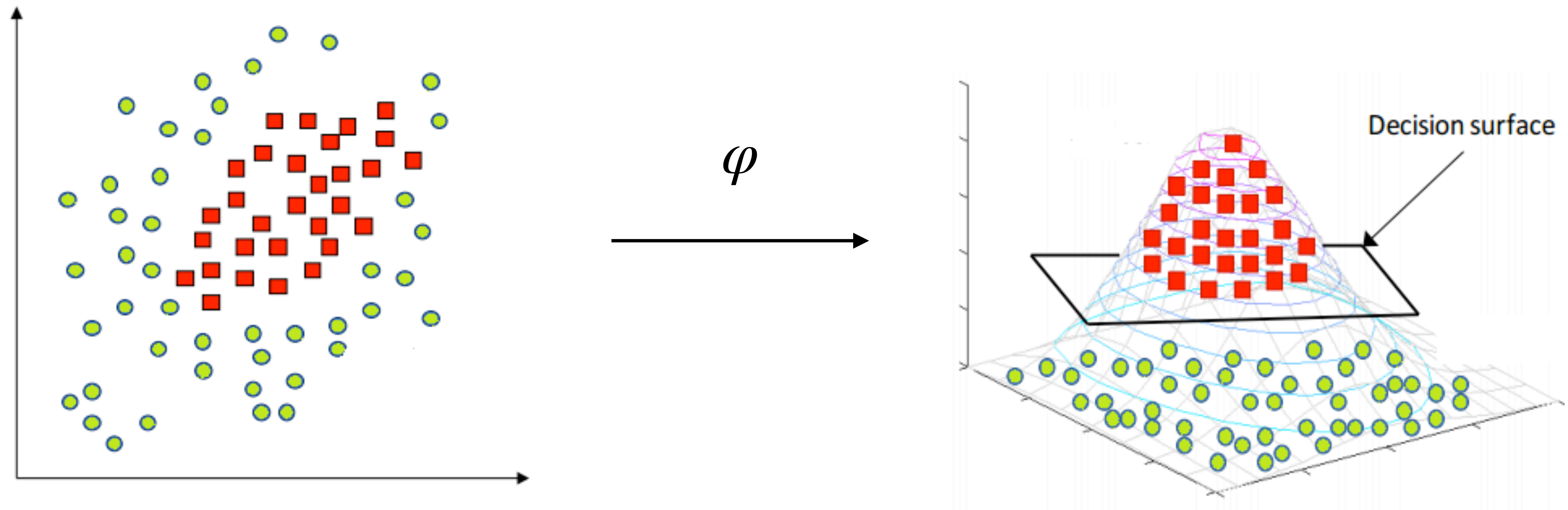
$$f_\theta(\boldsymbol{x}) = \langle \boldsymbol{\theta}, \boldsymbol{\varphi}(\boldsymbol{x}) \rangle$$

- 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 $\boldsymbol{\varphi} : \mathcal{X} \to \mathbb{R}^p$

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

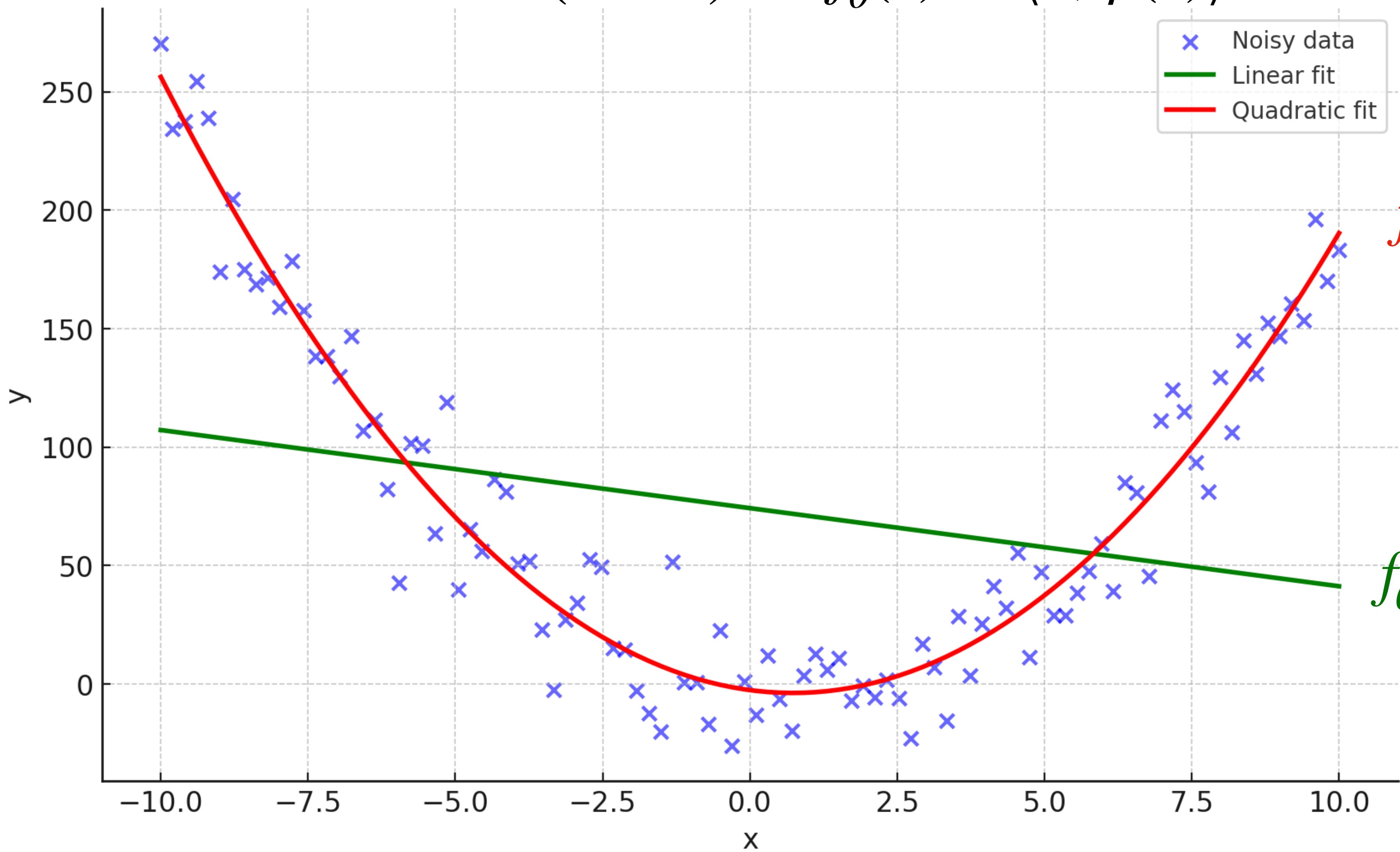
Feature maps

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



Examples: quadratic function

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



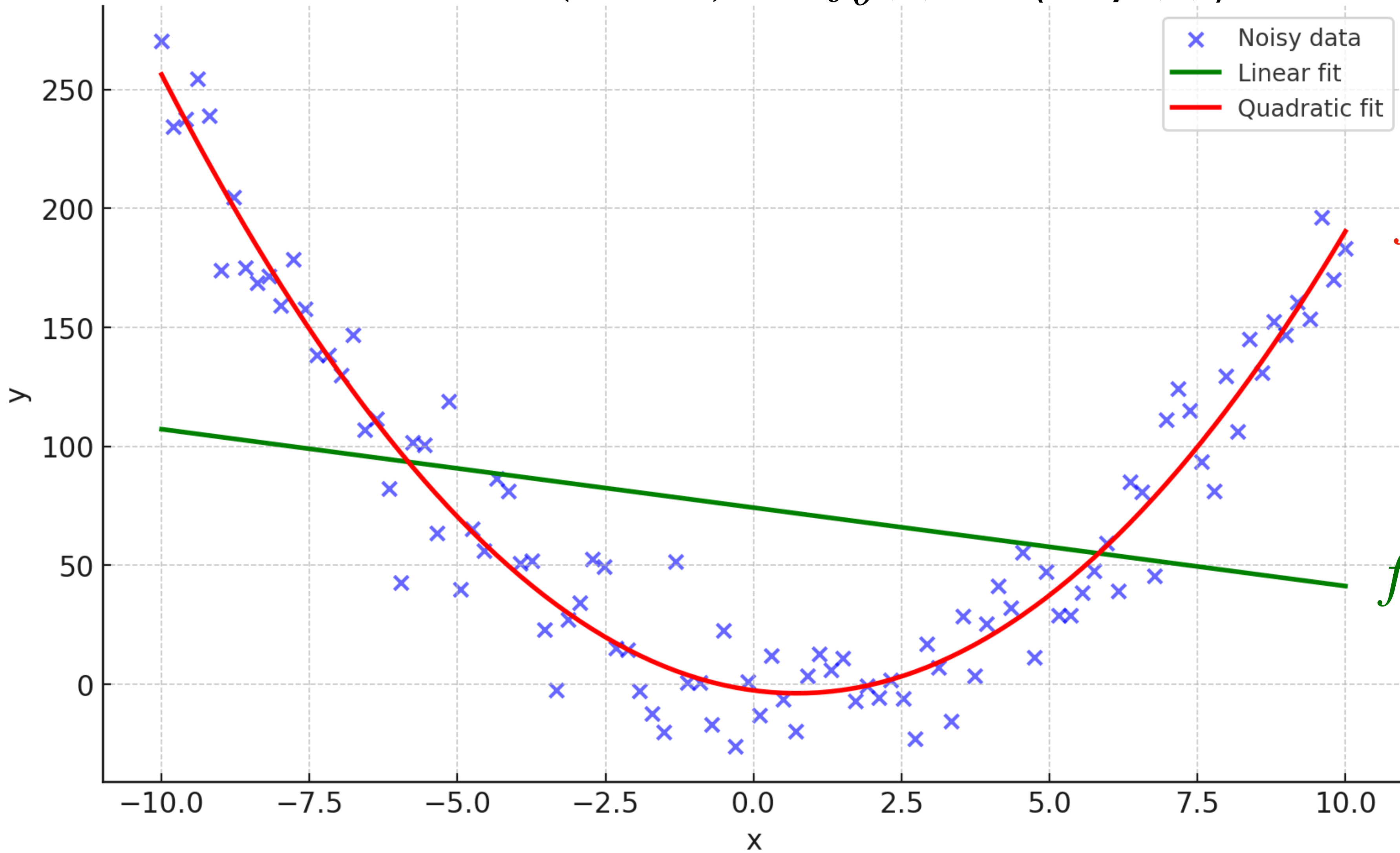
$f_\theta(x) = \theta_1 x^2 + \theta_2 x + b$

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

Question: what is $\boldsymbol{\varphi}(x)$?

Examples: quadratic function

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



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

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

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

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 $\mathbb{R}^k$:

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

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 $\mathbb{R}^d$:

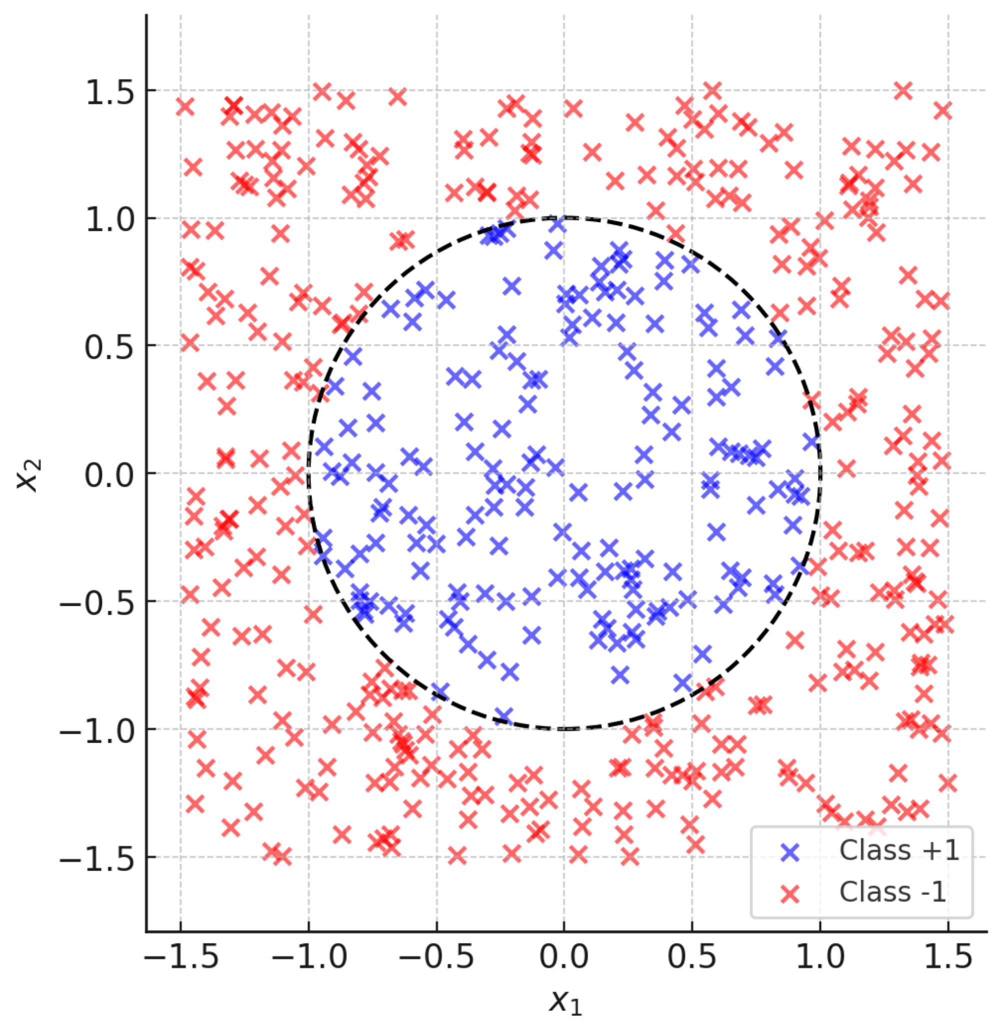
Example $d = 2$:

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

Examples: data in circle

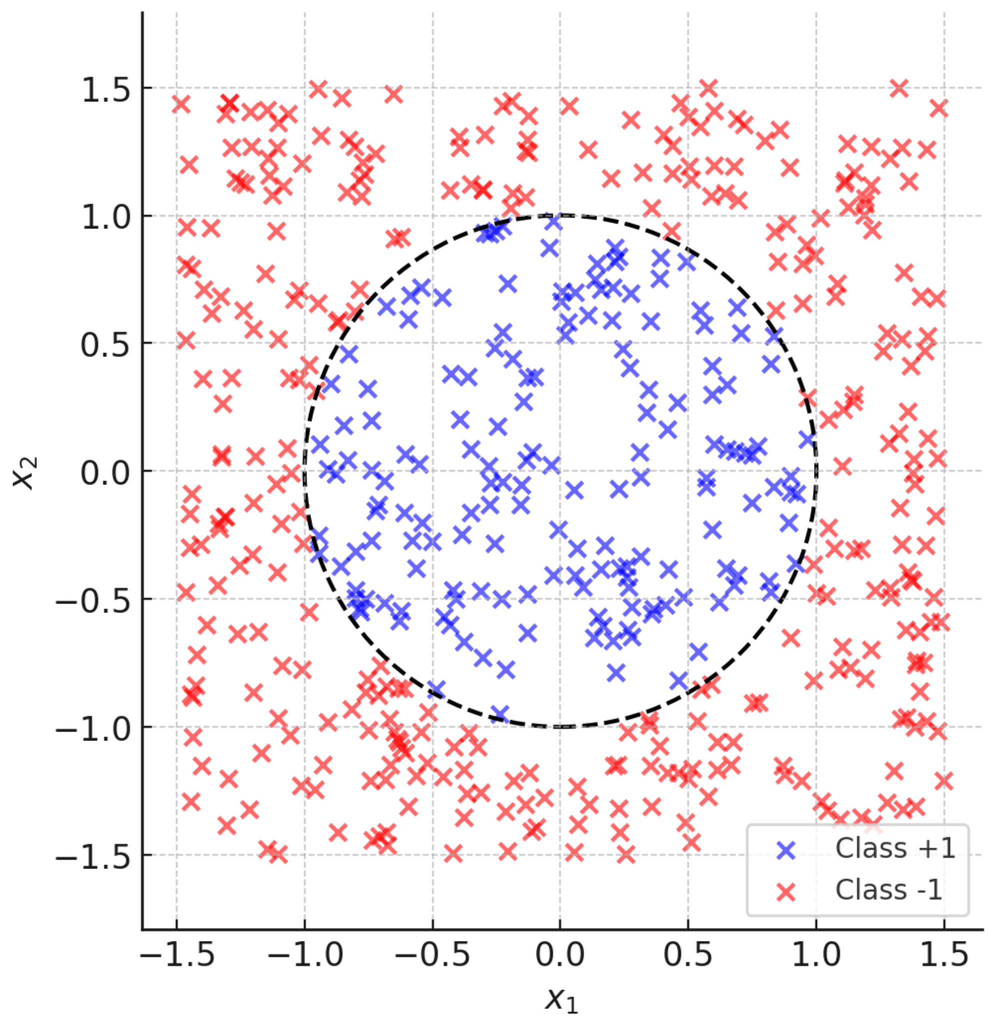
$x \in \mathbb{R}^2 \quad (d = 2)$

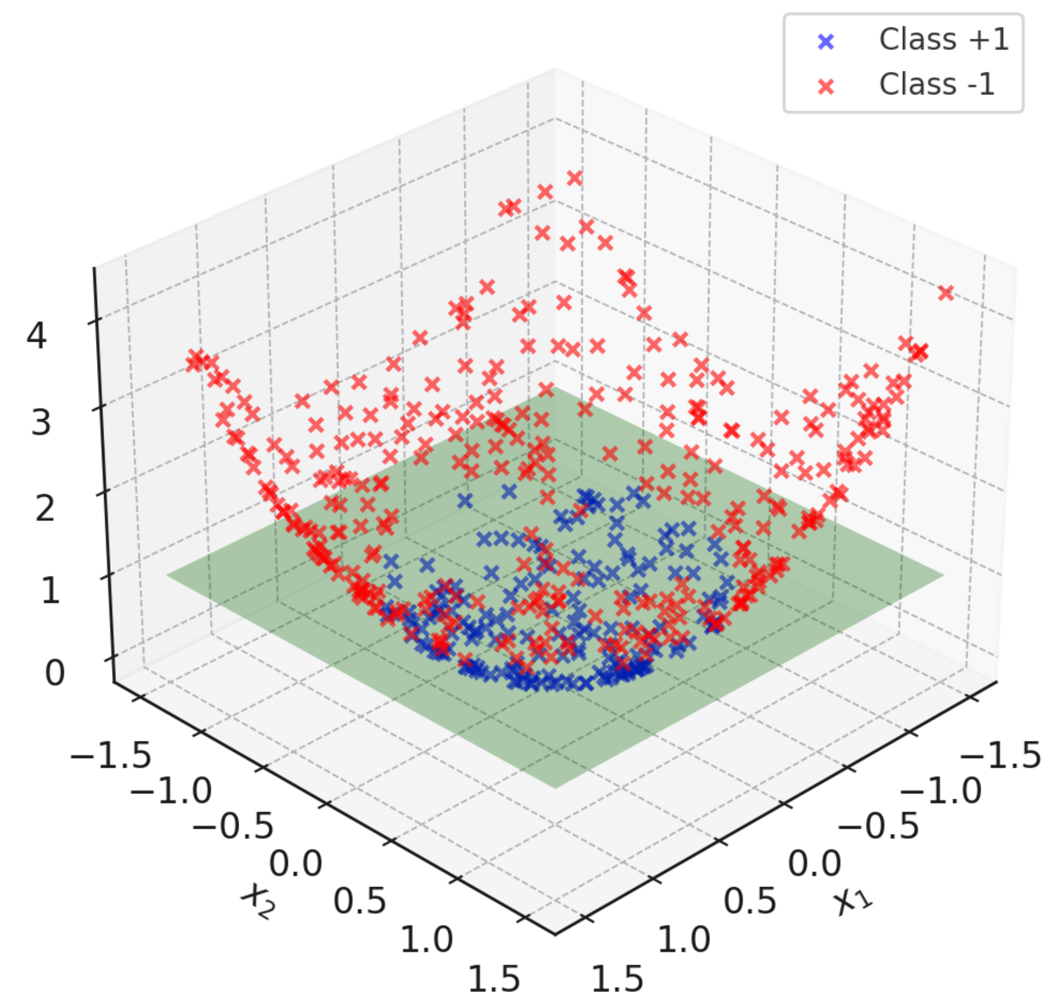
$$y = \begin{cases} +1 & \text{if } x_1^2 + x_2^2 \leq 1 \\ -1 & \text{if } x_1^2 + x_2^2 > 1 \end{cases}$$



Examples: data in circle

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

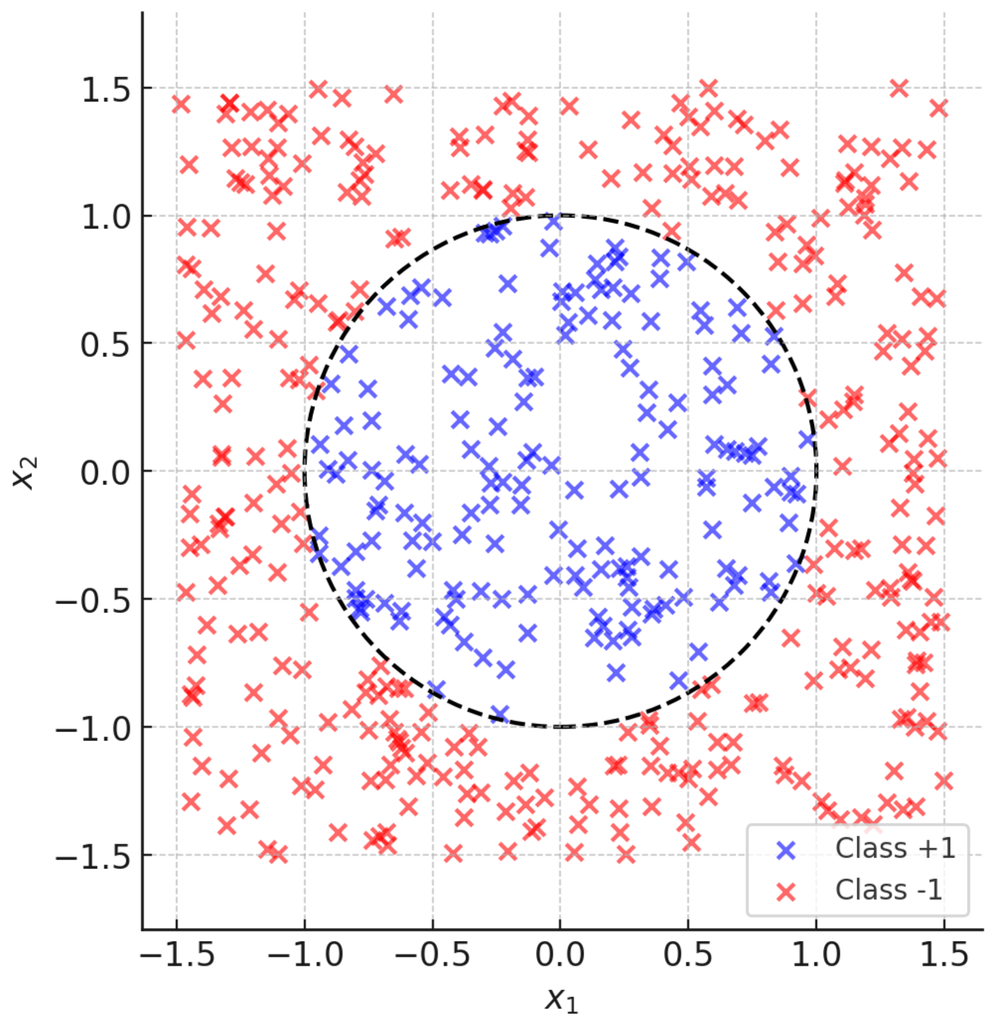


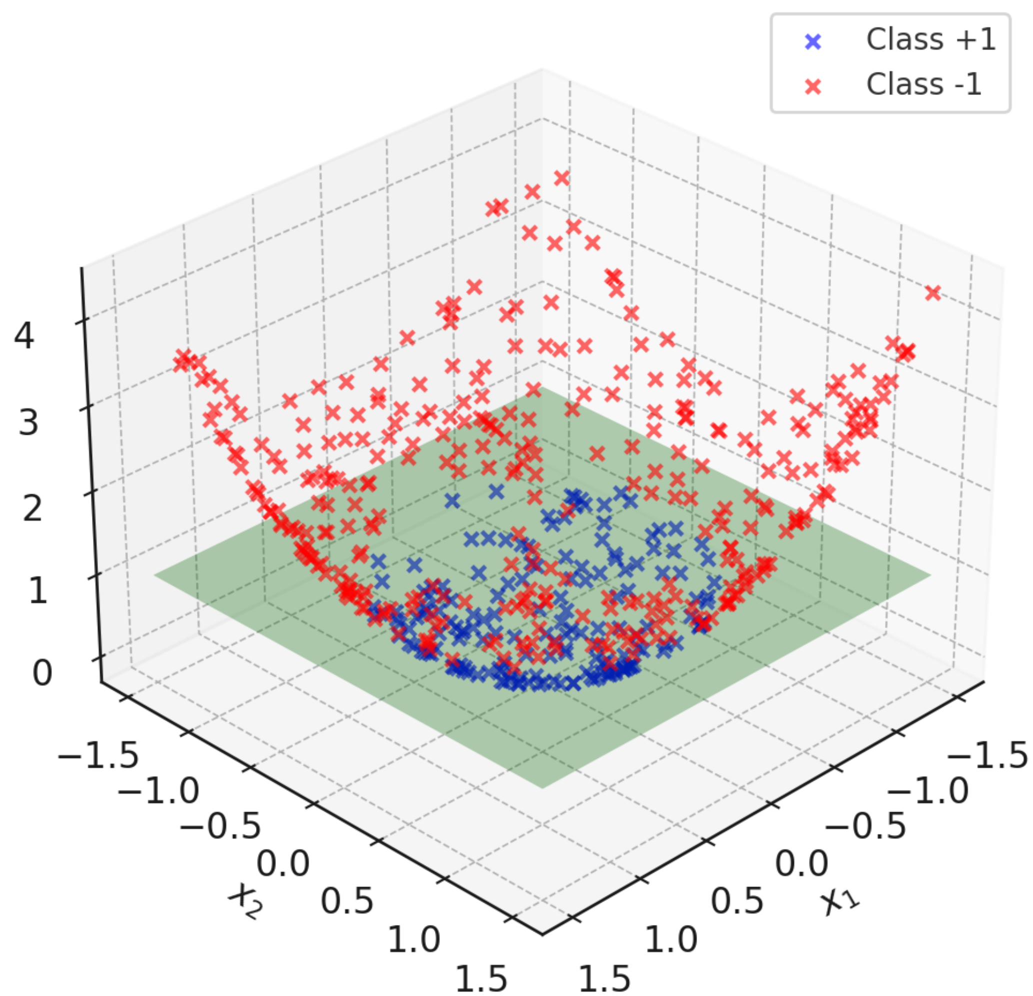


$$\boldsymbol{\varphi}(x) = \begin{bmatrix} x_1 \\ x_2 \\ x_1^2 + x_2^2 \end{bmatrix} \qquad (p = 3)$$

Examples: data in circle

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





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

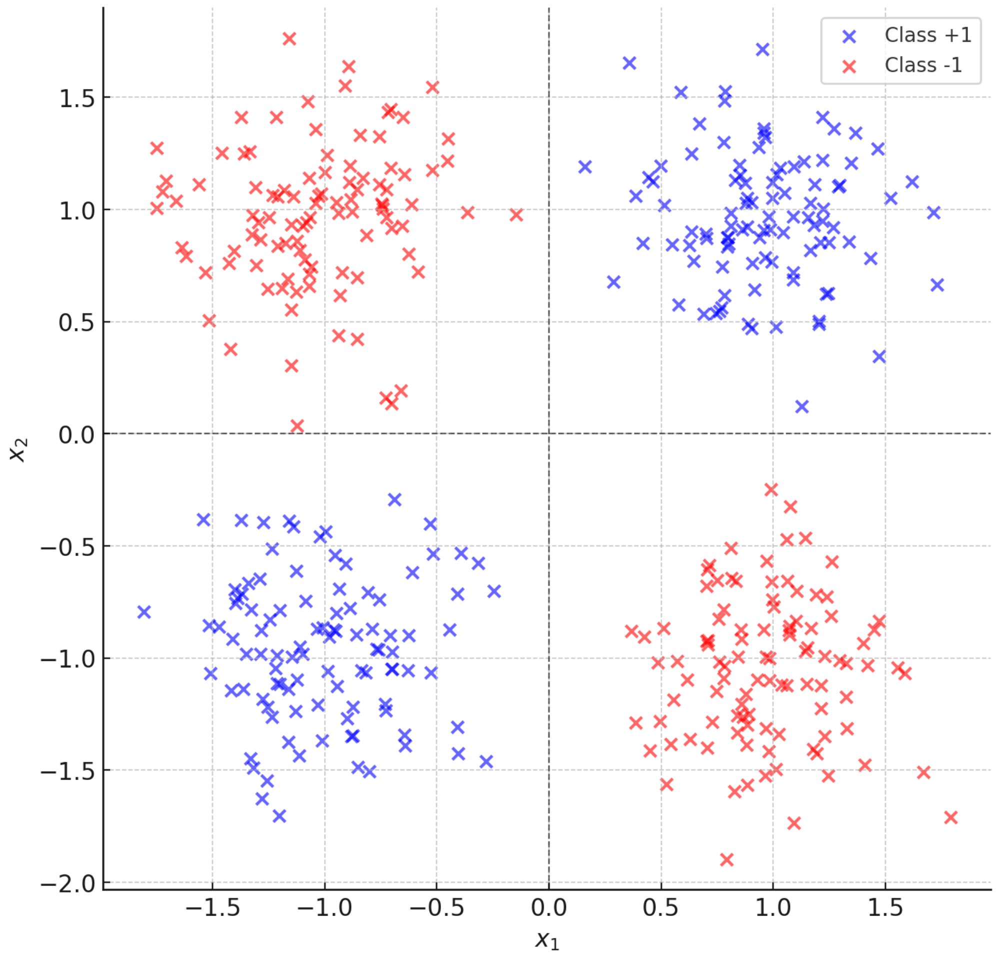
Not unique!

Examples: XOR Gaussian mixture

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

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

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

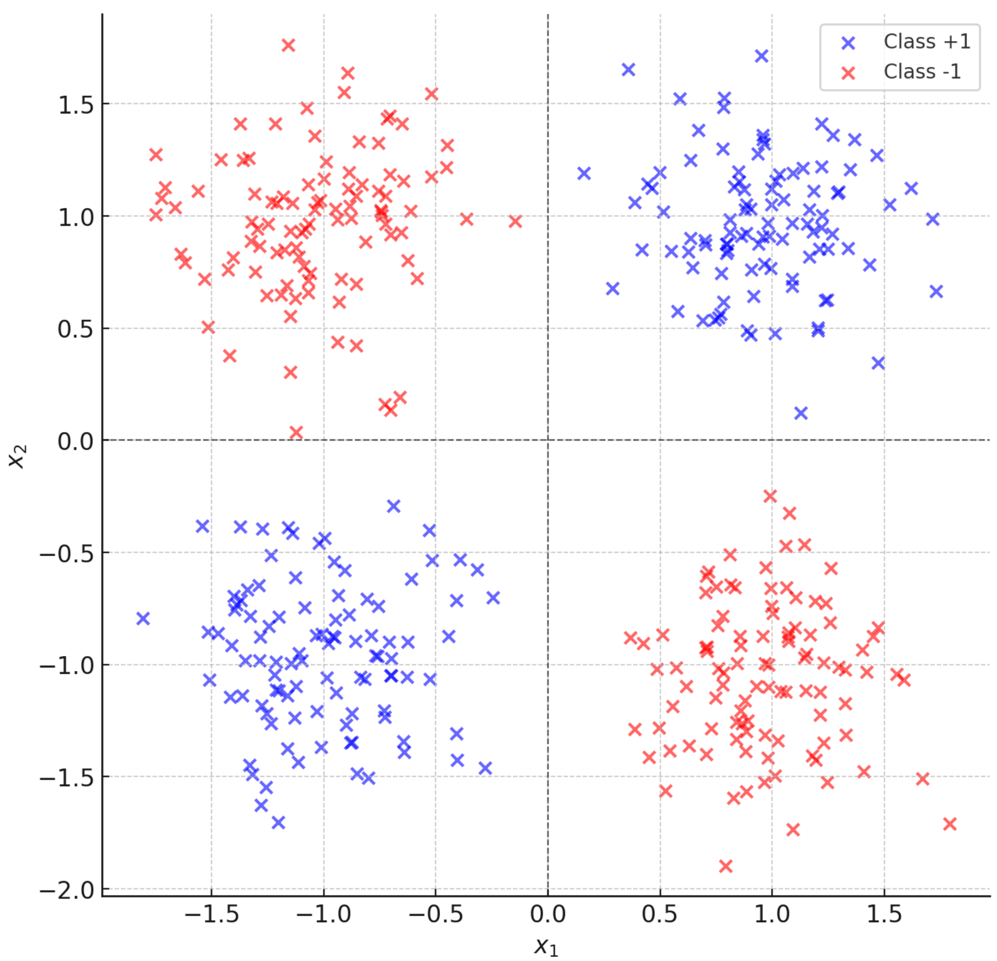


$$\mu_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \qquad \mu_4 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

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:

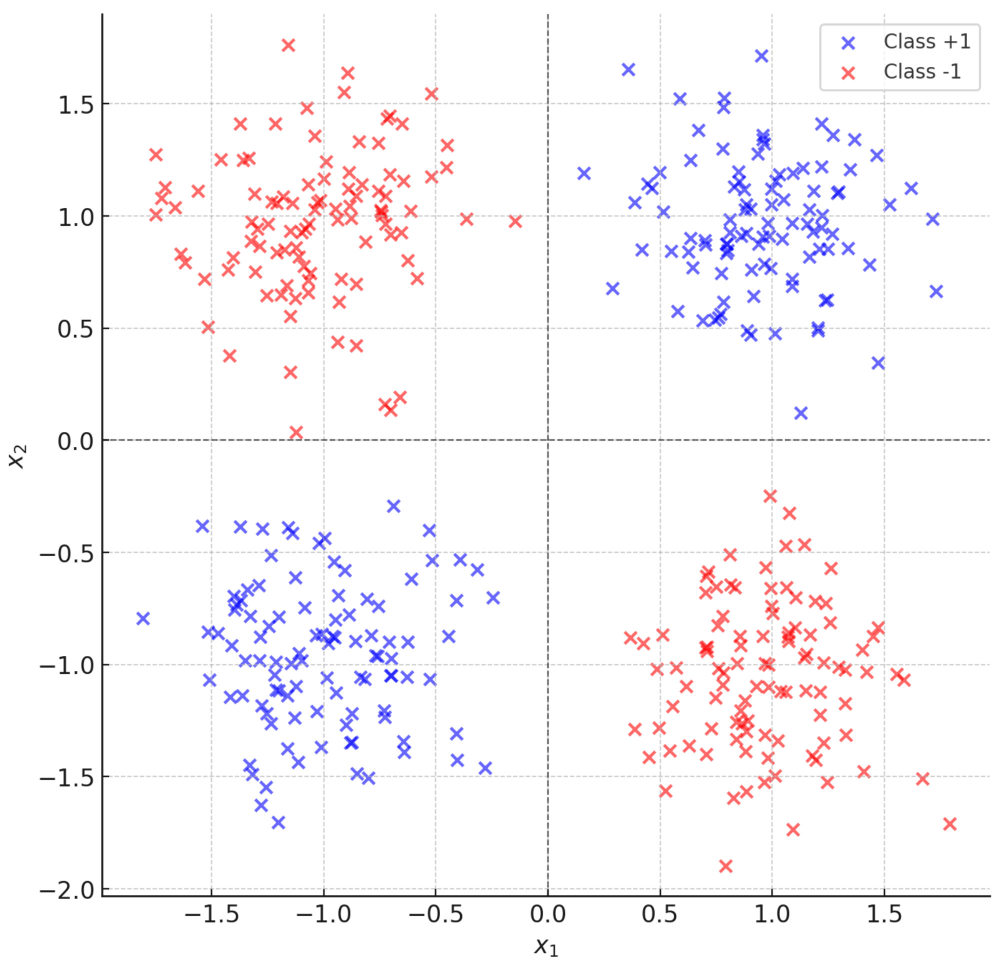
$$y = +1 \qquad (x_1, x_2) \in \{(-1, -1), (1,1)\}$$

$$y = -1 \qquad (x_1, x_2) \in \{(1, -1), (1, -1)\}$$

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:

$$y = +1 \qquad (x_1, x_2) \in \{(-1, -1), (1,1)\}$$

or... $x_1 x_2 = +1$

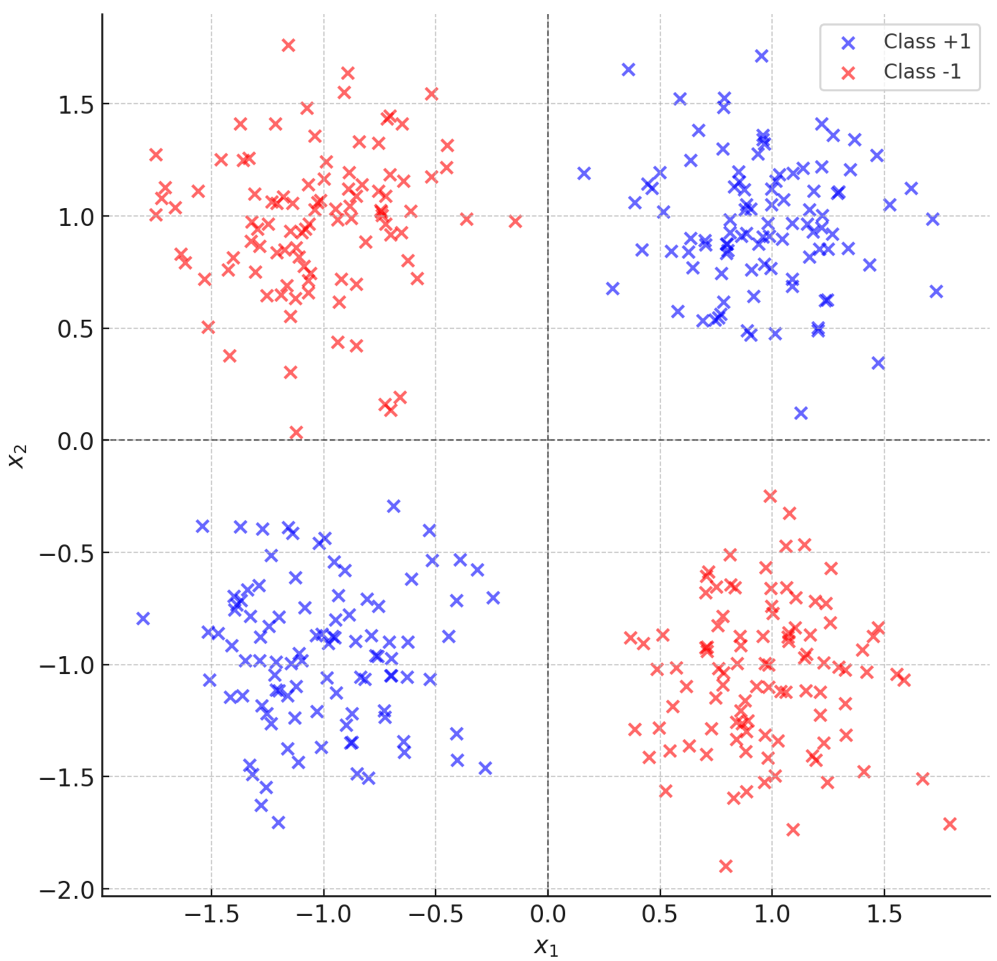
$$y = -1 \qquad (x_1, x_2) \in \{(1, -1), (1, -1)\}$$

or... $x_1 x_2 = -1$

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:

$y = +1 \qquad (x_1, x_2) \in \{(-1, -1), (1,1)\}$

or... $x_1 x_2 = +1$

$y = -1 \qquad (x_1, x_2) \in \{(1, -1), (1, -1)\}$

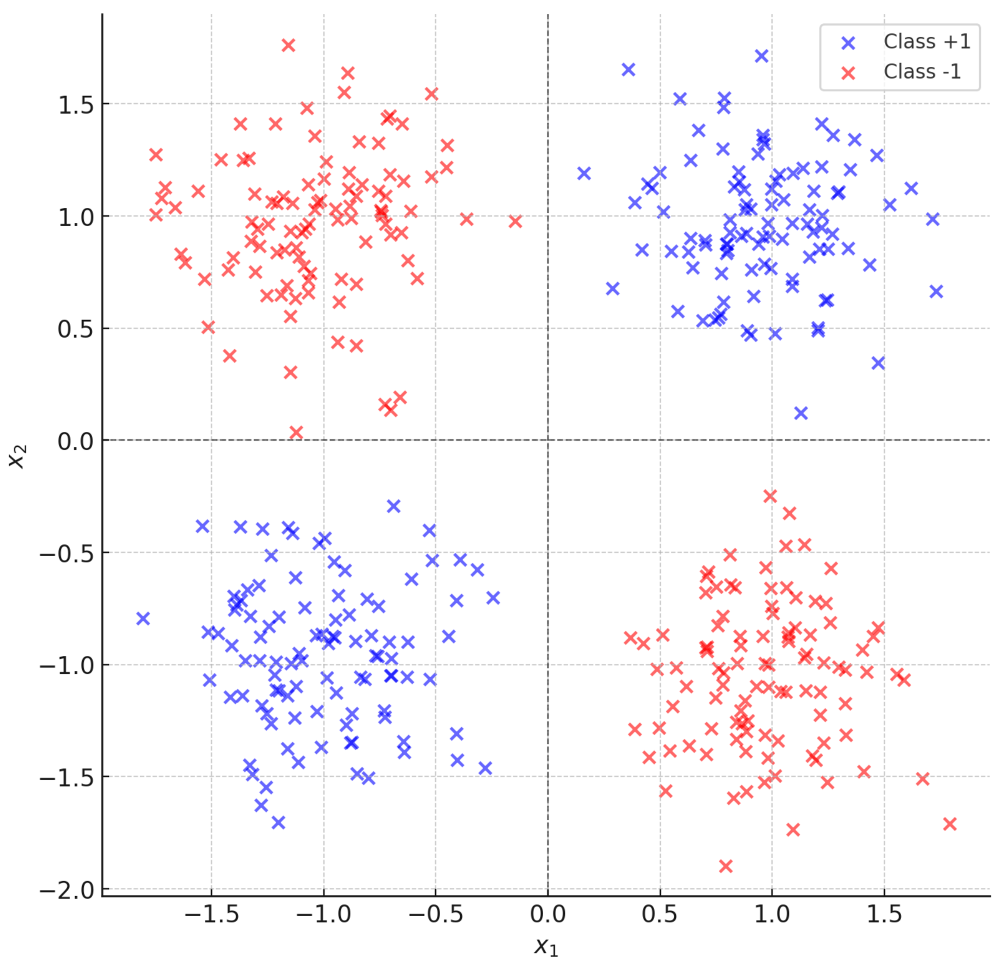
or... $x_1 x_2 = -1$

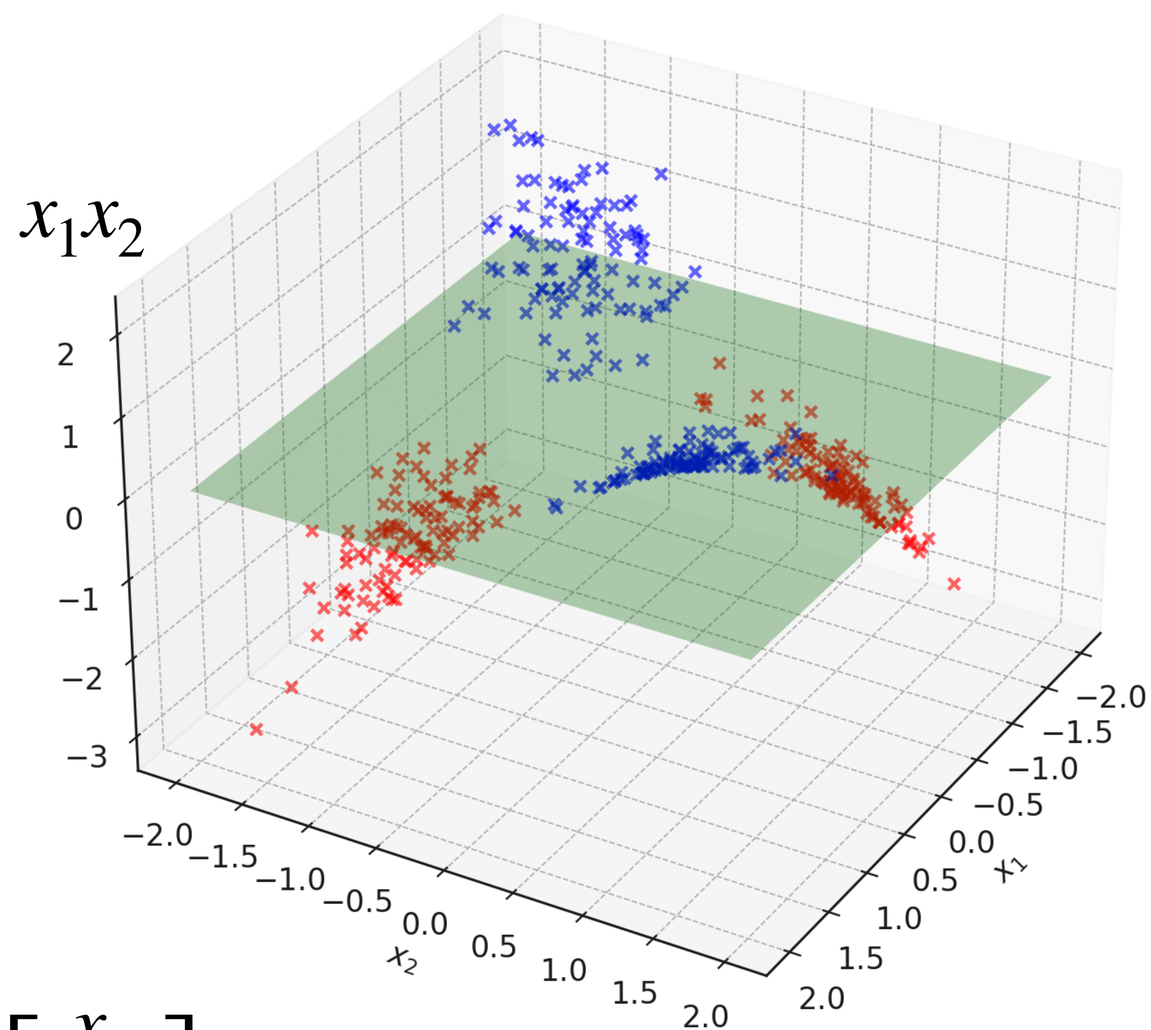
This motivates a choice: $\boldsymbol{\varphi}(\boldsymbol{x}) = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \end{bmatrix} \quad (p = 3)$

Examples: XOR Gaussian mixture

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

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





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