Machine Learning with Python

MTH786U/P 2023/24

Week 3: Unstable regression problems

Nicola Perra, Queen Mary University of London (QMUL)

n.perra@qmul.ac.uk

Recap of last week

Mathematical formulation of the regression problem:

Given input/output pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^{s}$ find function f with

$$y_i \approx f(\mathbf{x_i}) \qquad \forall i \in \{1, \ldots, s\}$$

Recap of last week

Mathematical formulation of the regression problem:

Given input/output pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^{s}$ find function f with

$$y_i \approx f(\mathbf{x_i}) \qquad \forall i \in \{1,\ldots,s\}$$

Important to notice how each $\mathbf{x_i}$ is a vector describing d features/variables

$$\mathbf{x_i} = (x_{i1}, \ldots, x_{id})$$

Example: linear regression

$$y_i \approx f(\mathbf{x_i}) \qquad \forall i \in \{1,\ldots,s\}$$

Example: linear regression

$y_i \approx f(\mathbf{x_i}) \qquad \forall i \in \{1,\ldots,s\}$

How do we parametrise f ?

Example: linear regression

$y_i \approx f(\mathbf{x_i}) \qquad \forall i \in \{1,\ldots,s\}$

How do we parametrise f ?

Example:

$$f(\mathbf{x_i}) = w_0 + \sum_{j=1}^{d} x_{ij} w_j$$

Example: linear regression

$y_i \approx f(\mathbf{x_i}) \qquad \forall i \in \{1,\ldots,s\}$

How do we parametrise f ?

Example: $$f(\mathbf{x_i}) = w_0 + \sum_{j=1}^{d} x_{ij} w_j$$

Linear transformation of vector $\mathbf{x_i} = (x_{i1}, \ldots, x_{id})$ with weights $\mathbf{w} \in \mathbb{R}^{d+1}$

Example: linear regression

Notation: $$f(\mathbf{x_i}) = w_0 + \sum_{j=1}^{d} w_j x_{ij} = \langle \mathbf{w}, \mathbf{x_i} \rangle$$

Example: linear regression

Notation: $$f(\mathbf{x_i}) = w_0 + \sum_{j=1}^{d} w_j x_{ij} = \langle \mathbf{w}, \mathbf{x_i} \rangle = \mathbf{w}^\top \mathbf{x_i} = \mathbf{x_i}^\top \mathbf{w}$$

Example: linear regression

Notation: $$f(\mathbf{x_i}) = w_0 + \sum_{j=1}^{d} w_j x_{ij} = \langle \mathbf{w}, \mathbf{x_i} \rangle = \mathbf{w}^\top \mathbf{x_i} = \mathbf{x_i}^\top \mathbf{w}$$

$$\mathbf{x_i} := \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix} \in \mathbb{R}^{d+1}$$

Example: linear regression

Notation: $$f(\mathbf{x_i}) = w_0 + \sum_{j=1}^{d} w_j x_{ij} = \langle \mathbf{w}, \mathbf{x_i} \rangle = \mathbf{w}^\top \mathbf{x_i} = \mathbf{x_i}^\top \mathbf{w}$$

$$\mathbf{x_i} := \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix} \in \mathbb{R}^{d+1} \qquad \mathbf{w} := \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix} \in \mathbb{R}^{d+1}$$

Example: linear regression

Notation: $$f(\mathbf{x_i}) = w_0 + \sum_{j=1}^{d} w_j x_{ij} = \langle \mathbf{w}, \mathbf{x_i} \rangle = \mathbf{w}^\top \mathbf{x_i} = \mathbf{x_i}^\top \mathbf{w}$$

Where this comes from?

$$\mathbf{x_i} := \begin{pmatrix} \boxed{1} \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix} \in \mathbb{R}^{d+1} \qquad \mathbf{w} := \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix} \in \mathbb{R}^{d+1}$$

Example: linear regression

Notation: $$f(\mathbf{x_i}) = w_0 + \sum_{j=1}^{d} w_j x_{ij} = \langle \mathbf{w}, \mathbf{x_i} \rangle = \mathbf{w}^\top \mathbf{x_i} = \mathbf{x_i}^\top \mathbf{w}$$

Where this comes from?

$$\mathbf{x_i} := \begin{pmatrix} \boxed{1} \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix} \in \mathbb{R}^{d+1} \qquad \mathbf{w} := \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix} \in \mathbb{R}^{d+1}$$

How do we choose w such that $y_i \approx f(x_i)$?

Example: linear regression

More in general?

$$\mathbf{y} = \mathbf{X}\mathbf{w}$$

Example: linear regression

More in general? $$\mathbf{y} = \mathbf{X}\mathbf{w}$$

Quality function $$\mathrm{MSE}(\mathbf{w}) = \frac{1}{2s}\sum_{i=1}^{s}\left|(\mathbf{X}\mathbf{w})_i - y_i\right|^2 = \frac{1}{2s}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

Example: linear regression

More in general?

$$\mathbf{y} = \mathbf{X}\mathbf{w}$$

Quality function $\mathrm{MSE}(\mathbf{w}) = \frac{1}{2s}\sum_{i=1}^{s}\left|(\mathbf{X}\mathbf{w})_i - y_i\right|^2 = \frac{1}{2s}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$

The optimal solution is the minimisers of the MSE

Example: linear regression

More in general?

$$\mathbf{y} = \mathbf{X}\mathbf{w}$$

Quality function $\text{MSE}(\mathbf{w}) = \frac{1}{2s}\sum_{i=1}^{s}\left|(\mathbf{X}\mathbf{w})_i - y_i\right|^2 = \frac{1}{2s}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$

The optimal solution is the minimisers of the MSE

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \text{MSE}(\mathbf{w}) = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{\frac{1}{2s}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2\right\}$$

Example: linear regression

More in general?

$$\mathbf{y} = \mathbf{X}\mathbf{w}$$

Quality function $\text{MSE}(\mathbf{w}) = \frac{1}{2s}\sum_{i=1}^{s}\left|(\mathbf{X}\mathbf{w})_i - y_i\right|^2 = \frac{1}{2s}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$

The optimal solution is the minimisers of the MSE

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \text{MSE}(\mathbf{w}) = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{\frac{1}{2s}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2\right\}$$

Due to the convexity of the MSE the minimisers can be found as:

Example: linear regression

More in general?

$$\mathbf{y} = \mathbf{X}\mathbf{w}$$

Quality function $\quad \mathsf{MSE}(\mathbf{w}) = \frac{1}{2s}\sum_{i=1}^{s}\left|(\mathbf{X}\mathbf{w})_i - y_i\right|^2 = \frac{1}{2s}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$

The optimal solution is the minimisers of the MSE

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \mathsf{MSE}(\mathbf{w}) = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{\frac{1}{2s}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2\right\}$$

Due to the convexity of the MSE the minimisers can be found as:

$$\nabla \mathsf{MSE}(\hat{\mathbf{w}}) \stackrel{!}{=} 0 \qquad \Rightarrow \qquad \mathbf{X}^\top\mathbf{X}\hat{\mathbf{w}} = \mathbf{X}^\top\mathbf{y}$$

Example: linear regression

More in general?

$$\mathbf{y} = \mathbf{X}\mathbf{w}$$

Quality function $\quad \mathrm{MSE}(\mathbf{w}) = \frac{1}{2s}\sum_{i=1}^{s}\left|(\mathbf{X}\mathbf{w})_i - y_i\right|^2 = \frac{1}{2s}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$

The optimal solution is the minimisers of the MSE

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \mathrm{MSE}(\mathbf{w}) = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{\frac{1}{2s}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2\right\}$$

Due to the convexity of the MSE the minimisers can be found as:

$$\nabla\mathrm{MSE}(\hat{\mathbf{w}}) \overset{!}{=} 0 \quad \Rightarrow \quad \mathbf{X}^\top\mathbf{X}\hat{\mathbf{w}} = \mathbf{X}^\top\mathbf{y} \quad \Rightarrow \quad \hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

Unstable regression problems

The subject of this lecture is - yet again - the solution of regression problems of the form

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \mathsf{MSE}(\mathbf{w}) = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{2s}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \right\}$$

Unstable regression problems

The subject of this lecture is - yet again - the solution of regression problems of the form

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^{d+1}} \text{MSE}(\mathbf{w}) = \arg \min_{\mathbf{w} \in \mathbb{R}^{d+1}} \left\{ \frac{1}{2s} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \right\}$$

We have established in previous lectures that the solution to this problem is

$$\mathbf{X}^\top \mathbf{X} \hat{\mathbf{w}} = \mathbf{X}^\top \mathbf{y}$$

Unstable regression problems

The subject of this lecture is - yet again - the solution of regression problems of the form

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \mathsf{MSE}(\mathbf{w}) = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{2s} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \right\}$$

We have established in previous lectures that the solution to this problem is

$$\hat{\mathbf{w}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

Unstable regression problems

The subject of this lecture is - yet again - the solution of regression problems of the form

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \mathsf{MSE}(\mathbf{w}) = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{2s} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \right\}$$

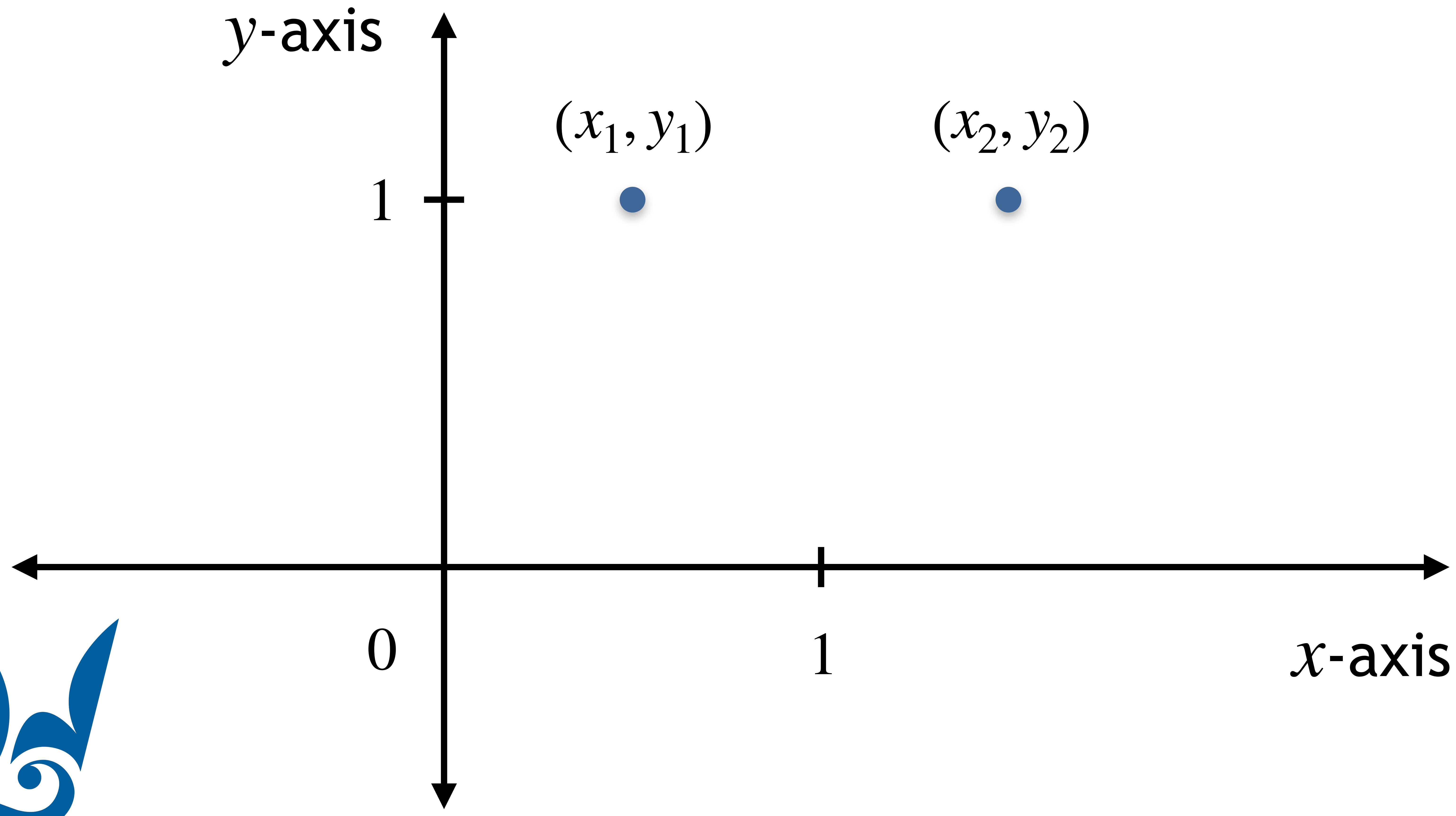
We have established in previous lectures that the solution to this problem is

$$\hat{\mathbf{w}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

In this lecture we investigate if and when such solutions can become unstable (and what this means)

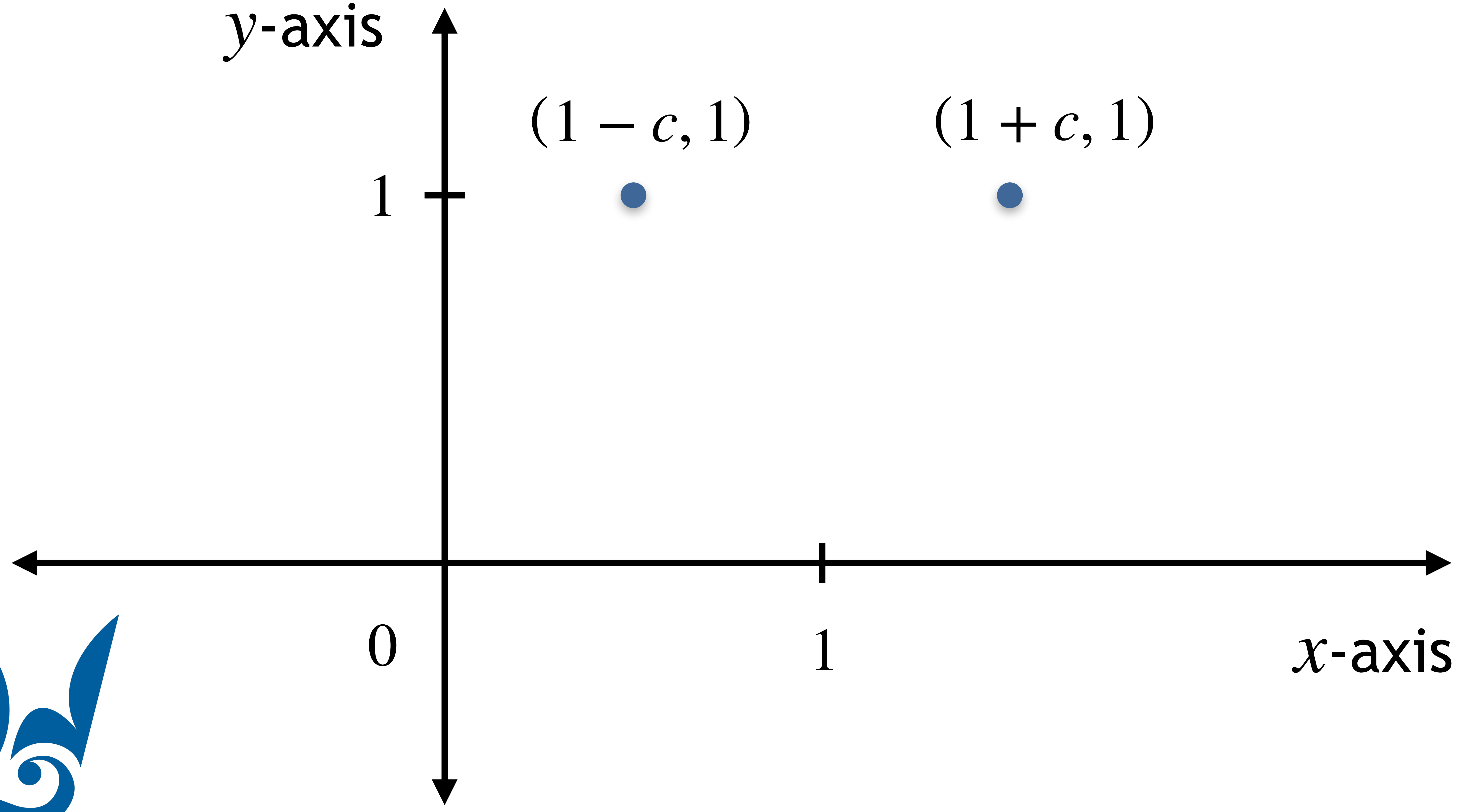
Unstable regression problems

Let us consider the following simple regression problem:



Unstable regression problems

Let us consider the following simple regression problem:



Data points

$x_1 = 1 - c$

$y_1 = 1$

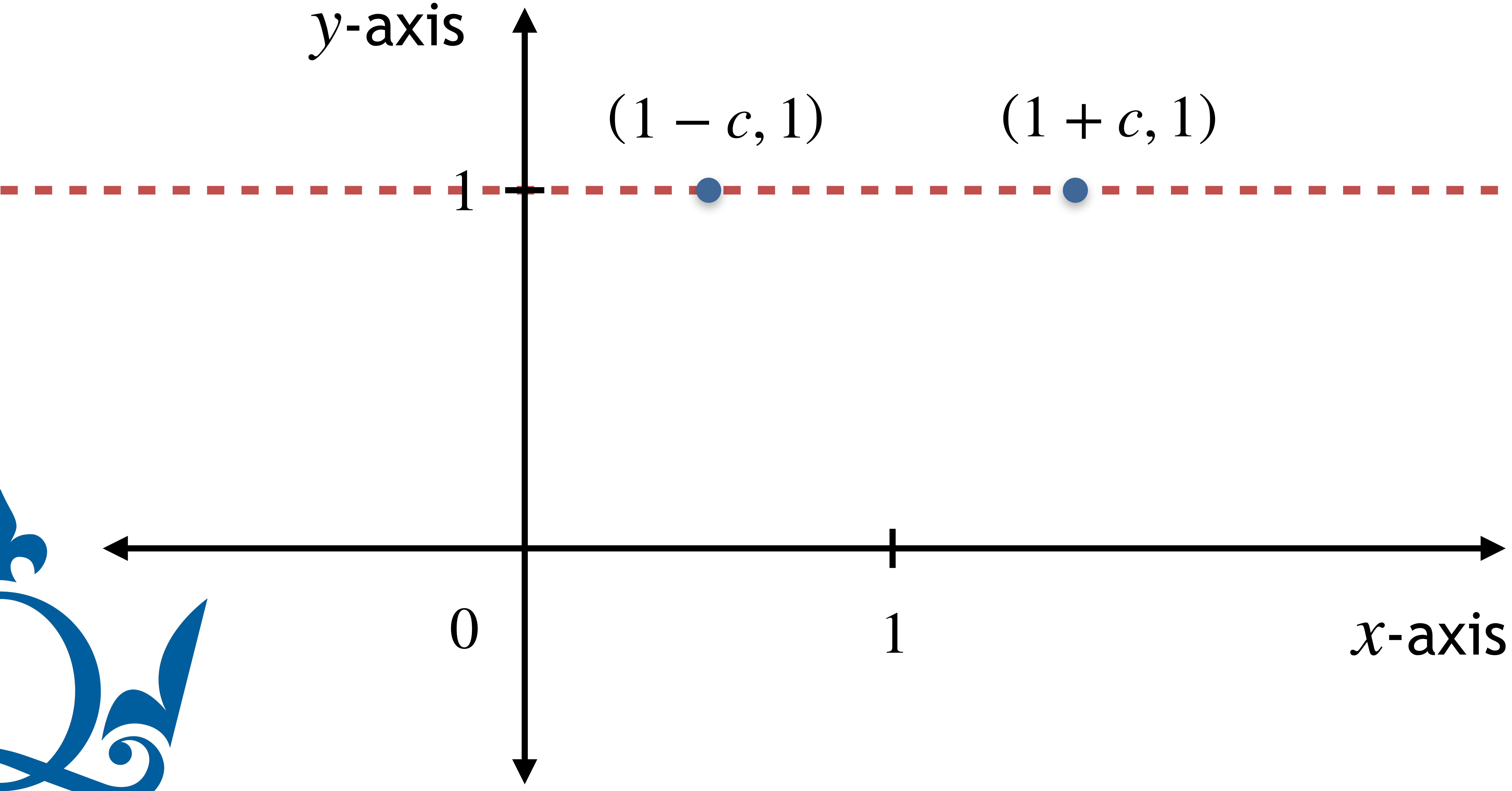
$x_2 = 1 + c$

$y_2 = 1$

for some $c > 0$

Unstable regression problems

Let us consider the following simple regression problem:



$$\hat{\mathbf{w}} = (\hat{w}_0, \hat{w}_1) = (1, 0)$$

Data points

$x_1 = 1 - c$

$y_1 = 1$

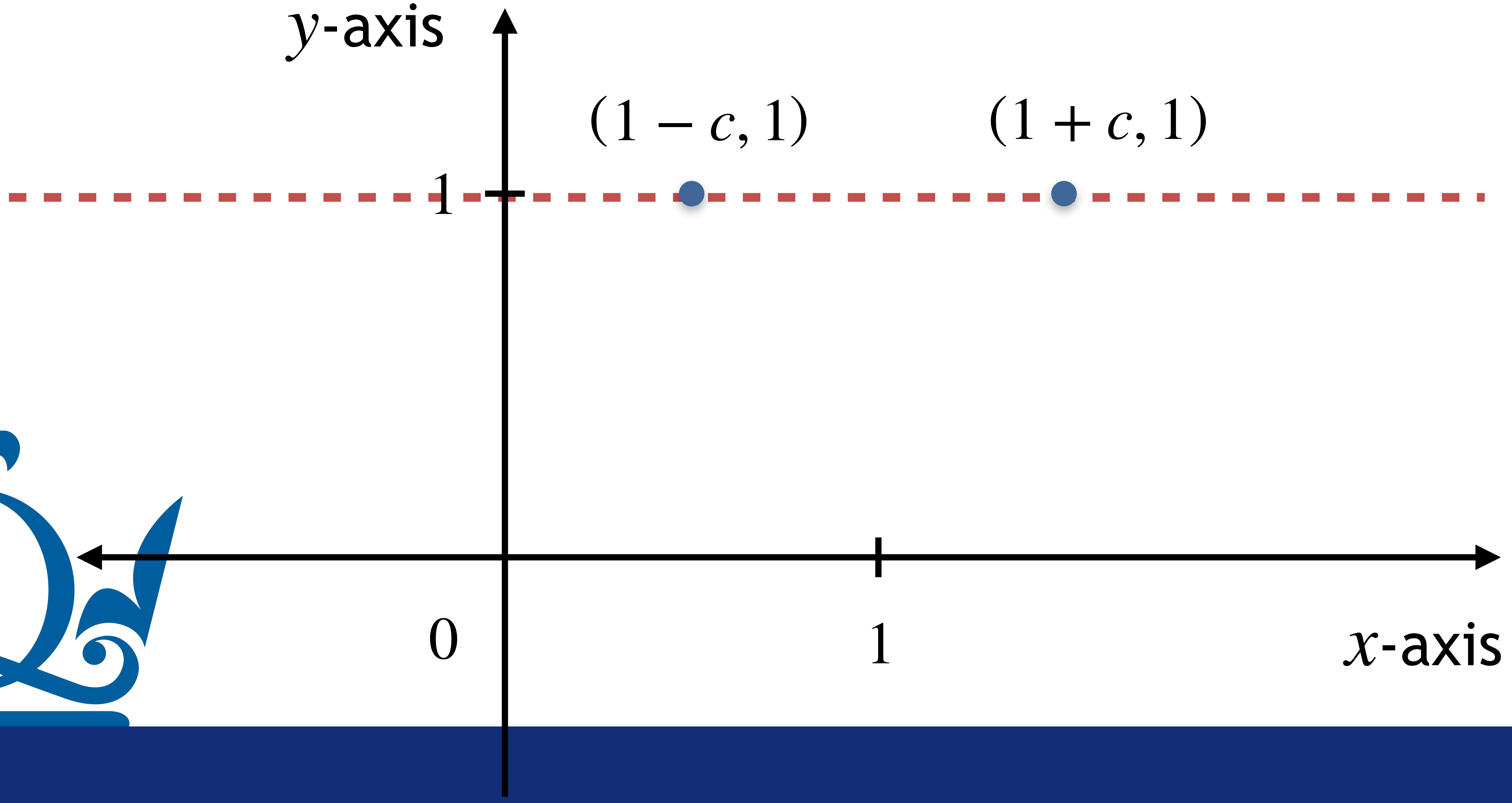
$x_2 = 1 + c$

$y_2 = 1$

for some $c > 0$

Why that's the solution?!

We can just solve the normal equation $\mathbf{X}^\top \mathbf{X} \hat{\mathbf{w}} = \mathbf{X}^\top \mathbf{y}$



Solving the normal equation

Normal equation $\mathbf{X}^\top \mathbf{X} \hat{\mathbf{w}} = \mathbf{X}^\top \mathbf{y}$

Solving the normal equation

Normal equation $\mathbf{X}^\top \mathbf{X}\hat{\mathbf{w}} = \mathbf{X}^\top \mathbf{y}$

How do we solve it?

Solving the normal equation

Normal equation $\mathbf{X}^\top \mathbf{X}\hat{\mathbf{w}} = \mathbf{X}^\top \mathbf{y}$

How do we solve it?

Key: the unknown is…

Solving the normal equation

Normal equation $\mathbf{X}^\top\mathbf{X}\hat{\mathbf{w}} = \mathbf{X}^\top\mathbf{y}$

How do we solve it?

Key: the unknown is... the vector of weights $\hat{\mathbf{w}}$

Solving the normal equation

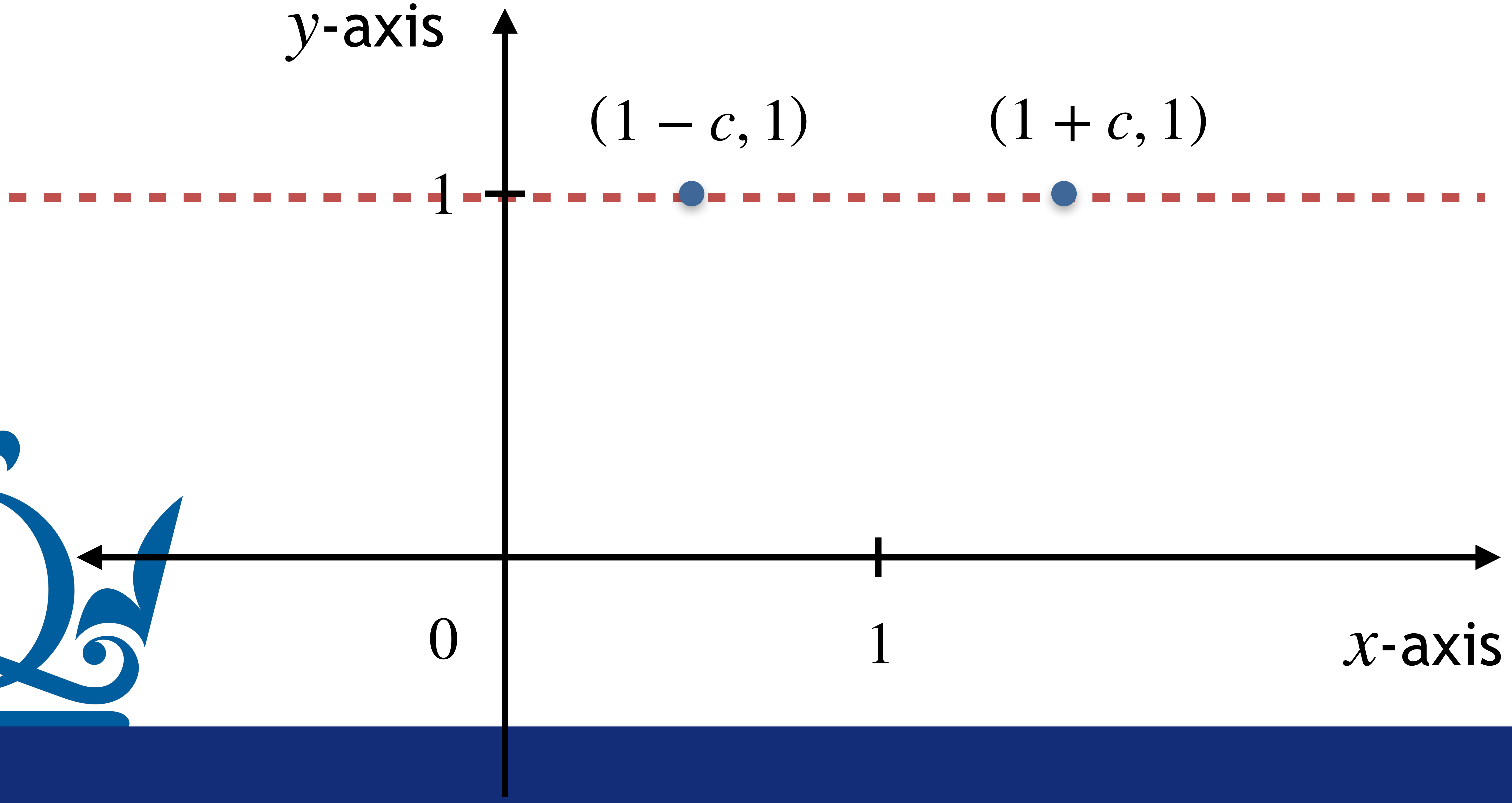
Normal equation $\mathbf{X}^\top \mathbf{X} \hat{\mathbf{w}} = \mathbf{X}^\top \mathbf{y}$

How do we solve it?

Key: the unknown is… the vector of weights $\hat{\mathbf{w}}$

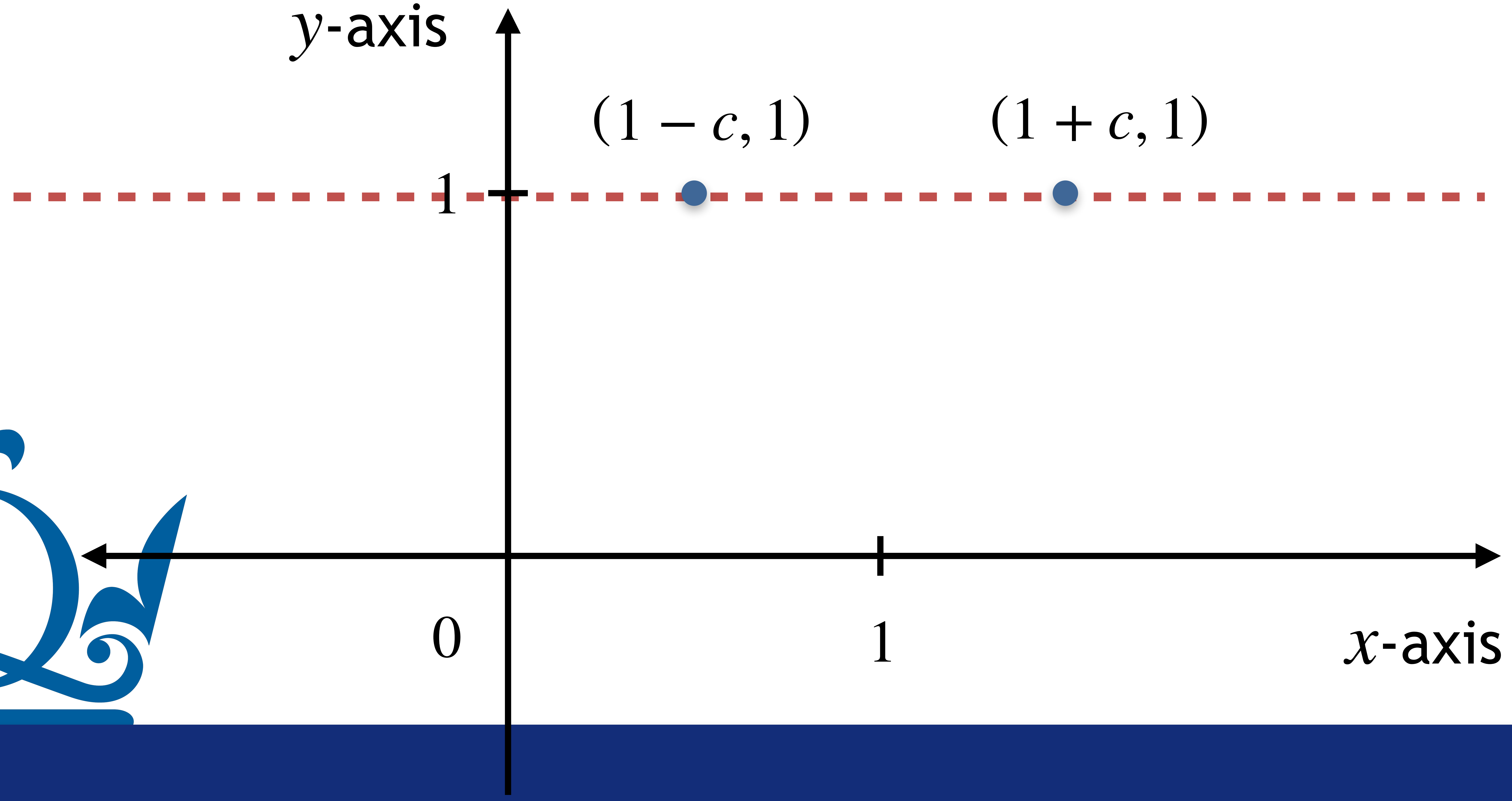
Hence we need just to evaluate $\mathbf{X}^\top \mathbf{X}$ and $\mathbf{X}^\top \mathbf{y}$ then solve the system!

Solving the normal equation



Solving the normal equation

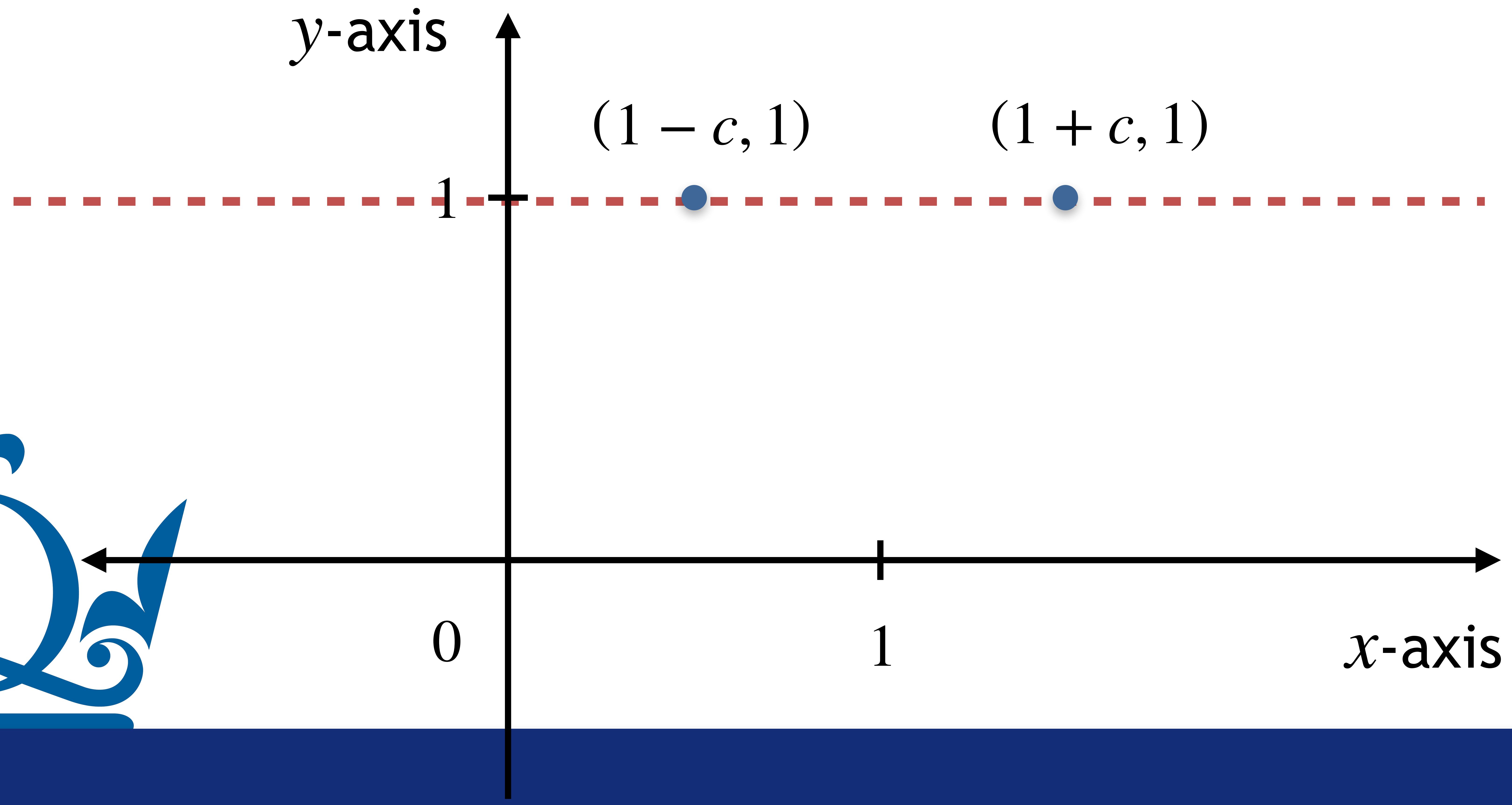
What's the data matrix $\mathbf{X}$



Solving the normal equation

What's the data matrix $\mathbf{X}$

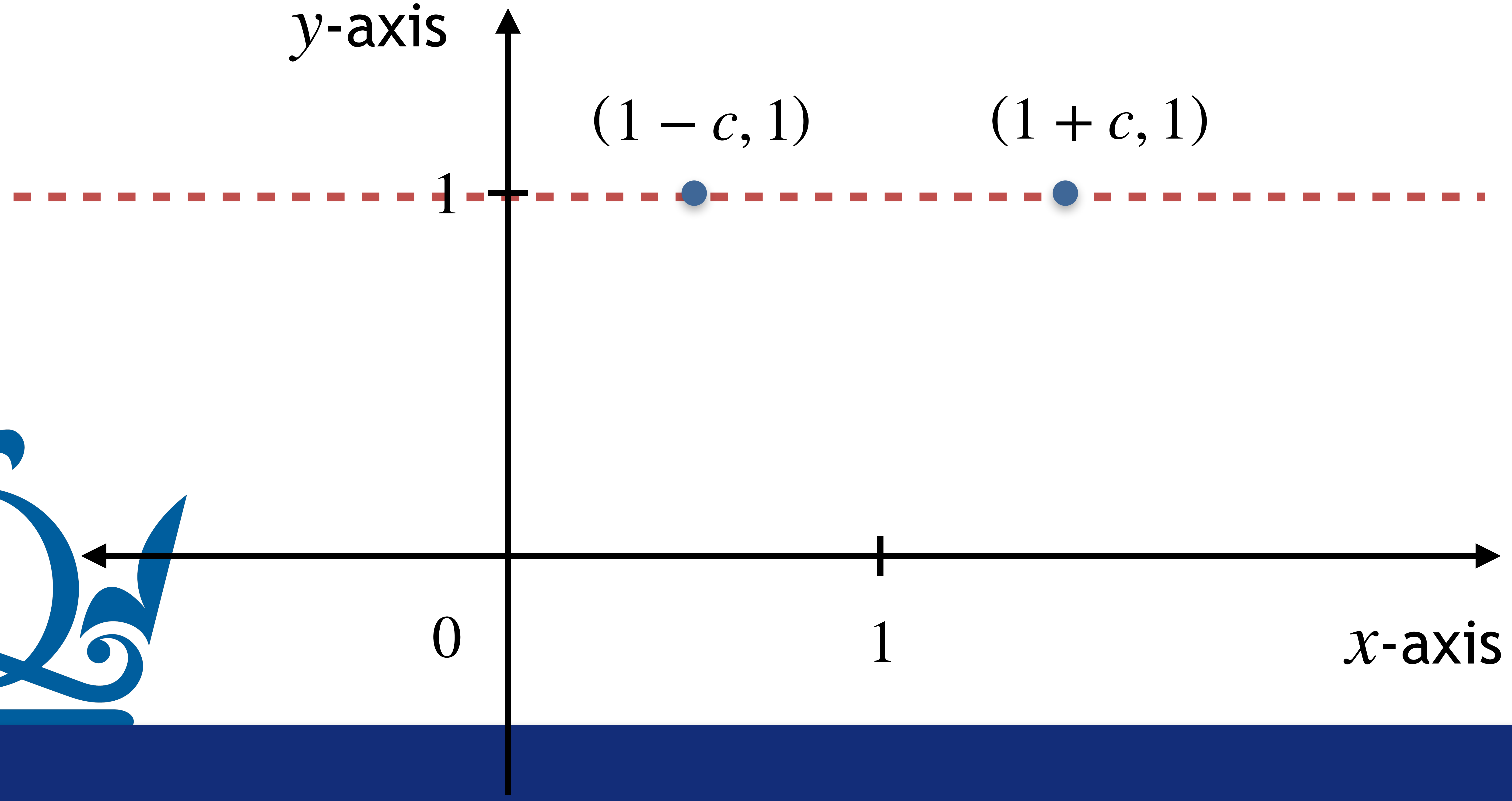
$$\mathbf{X} = \begin{pmatrix} 1 & 1-c \\ 1 & 1+c \end{pmatrix}$$



Solving the normal equation

What's the data matrix $\mathbf{X}$

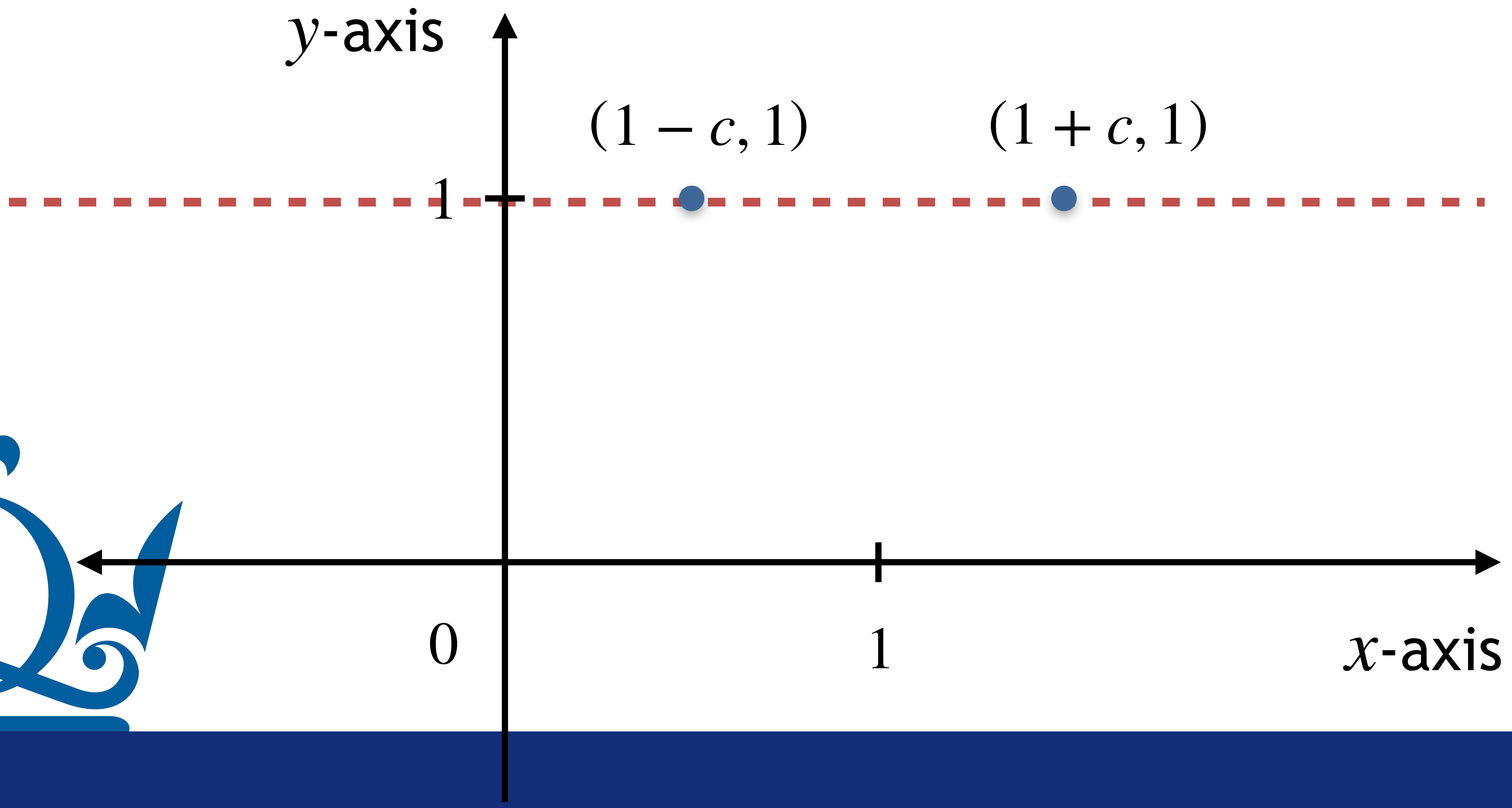
$$\mathbf{X} = \begin{pmatrix} 1 & 1-c \\ 1 & 1+c \end{pmatrix}$$



The transpose?

Solving the normal equation

What's the data matrix $\mathbf{X}$



$$\mathbf{X} = \begin{pmatrix} 1 & 1-c \\ 1 & 1+c \end{pmatrix}$$

The transpose?

$$\mathbf{X}^\top = \begin{pmatrix} 1 & 1 \\ 1-c & 1+c \end{pmatrix}$$

Solving the normal equation

What is the product $\mathbf{X}^\top \mathbf{X}$

$$\mathbf{X}^\top = \begin{pmatrix} 1 & 1 \\ 1-c & 1+c \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & 1-c \\ 1 & 1+c \end{pmatrix}$$

Solving the normal equation

What is the product $\mathbf{X}^\top \mathbf{X}$

$$\mathbf{X}^\top = \begin{pmatrix} 1 & 1 \\ 1-c & 1+c \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & 1-c \\ 1 & 1+c \end{pmatrix}$$

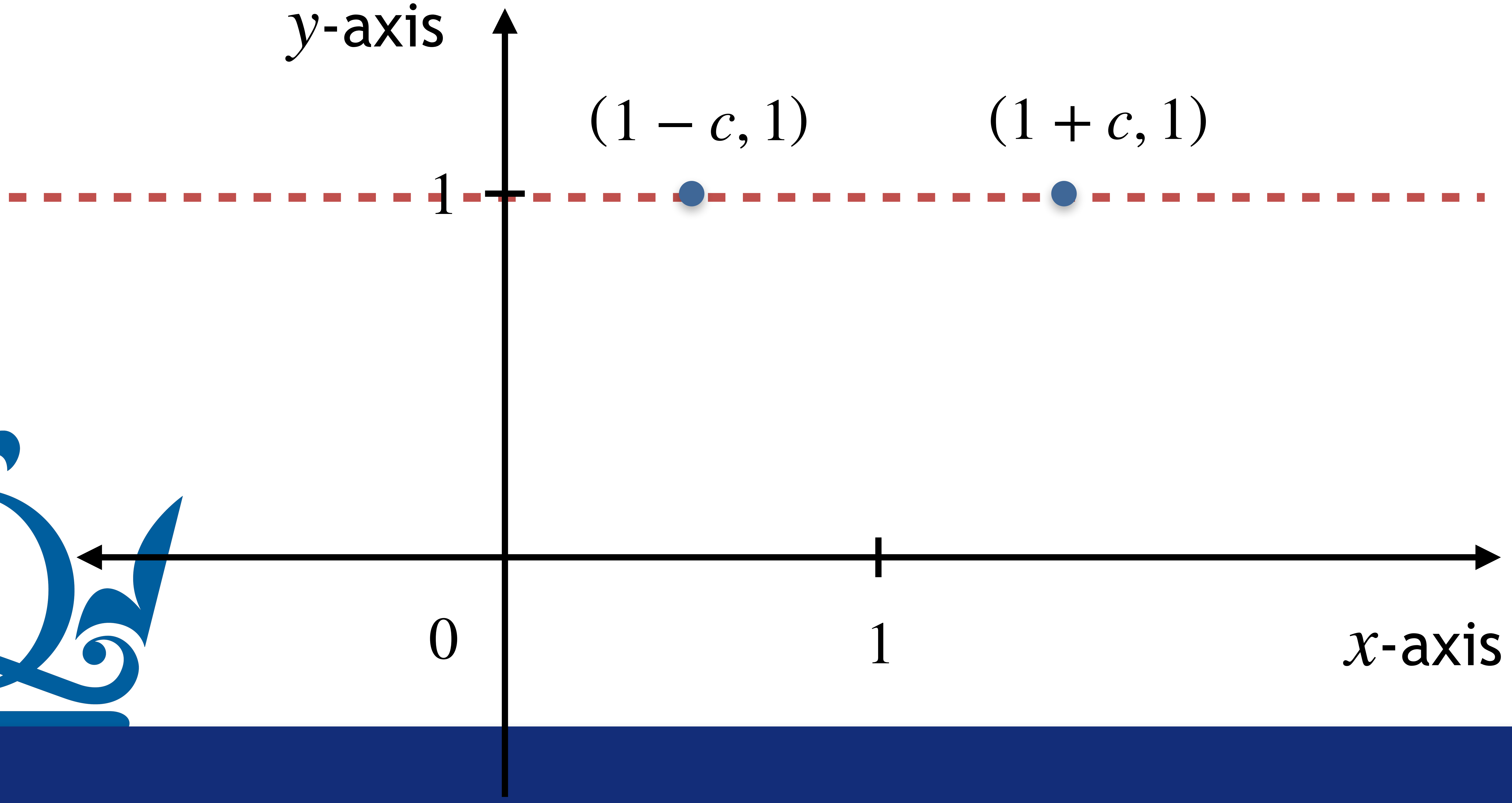
$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 2 & 2 \\ 2 & 2+2c^2 \end{pmatrix}$$

Solving the normal equation

What is the product $\mathbf{X}^\top \mathbf{y}$

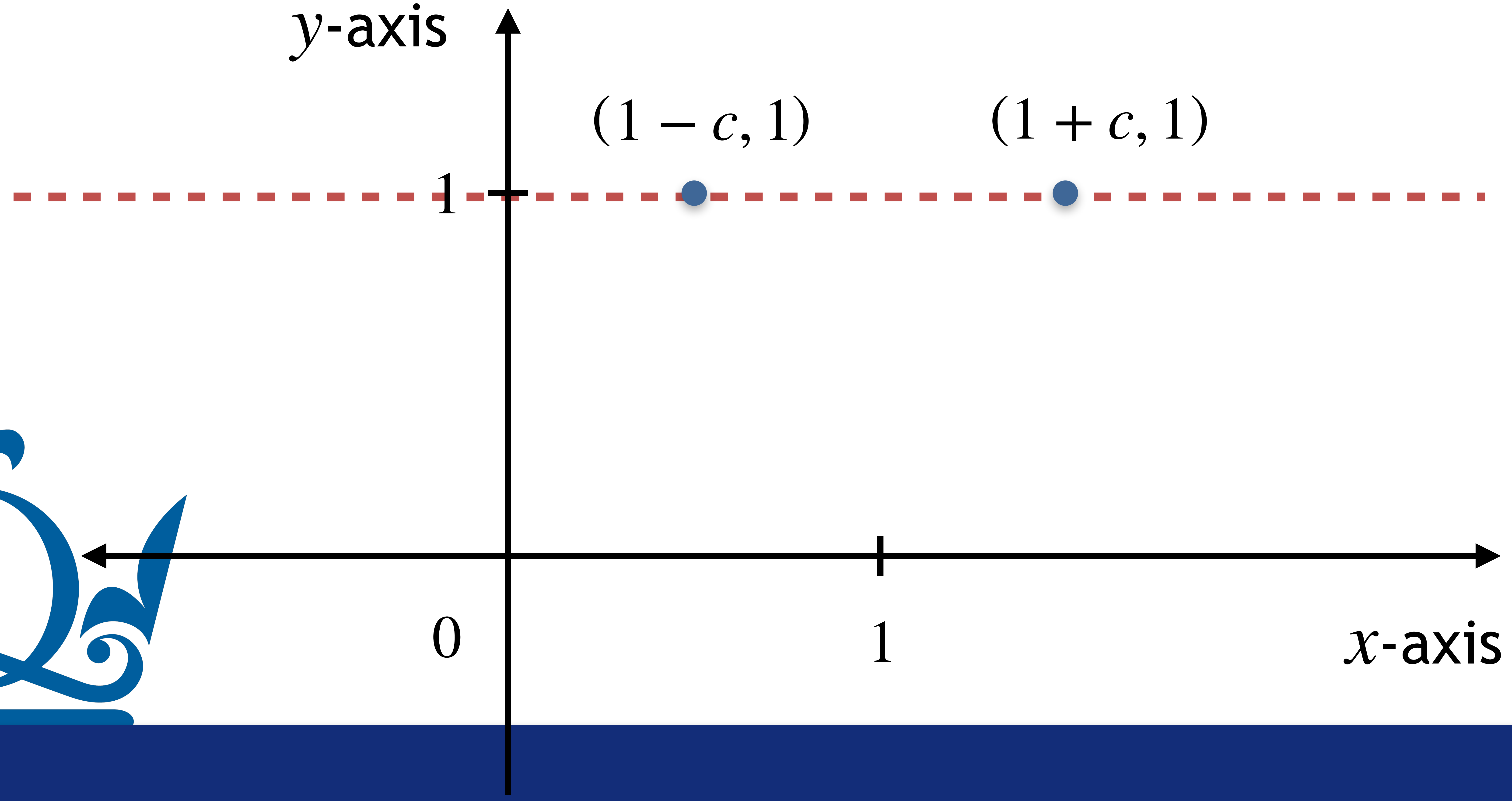
$$\mathbf{X}^\top = \begin{pmatrix} 1 & 1 \\ 1-c & 1+c \end{pmatrix}$$

$$\mathbf{y} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$



Solving the normal equation

What is the product $\mathbf{X}^\top \mathbf{y}$



$$\mathbf{X}^\top = \begin{pmatrix} 1 & 1 \\ 1-c & 1+c \end{pmatrix}$$

$$\mathbf{y} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\mathbf{X}^\top \mathbf{y} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

Solving the normal equation

Almost done! We are solving $\mathbf{X}^\top\mathbf{X}\hat{\mathbf{w}} = \mathbf{X}^\top\mathbf{y}$

$$\mathbf{X}^\top\mathbf{X} = \begin{pmatrix} 2 & 2 \\ 2 & 2+2c^2 \end{pmatrix} \qquad \mathbf{X}^\top\mathbf{y} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

Solving the normal equation

Almost done! We are solving $\mathbf{X}^\top \mathbf{X}\hat{\mathbf{w}} = \mathbf{X}^\top \mathbf{y}$

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 2 & 2 \\ 2 & 2 + 2c^2 \end{pmatrix} \qquad \mathbf{X}^\top \mathbf{y} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

Hence, we can write

Solving the normal equation

Almost done! We are solving $\mathbf{X}^\top \mathbf{X}\hat{\mathbf{w}} = \mathbf{X}^\top \mathbf{y}$

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 2 & 2 \\ 2 & 2 + 2c^2 \end{pmatrix} \qquad \mathbf{X}^\top \mathbf{y} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

Hence, we can write
$$\begin{pmatrix} 1 & 1 \\ 1 & 1 + c^2 \end{pmatrix} \begin{pmatrix} \hat{w}_1 \\ \hat{w}_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Solving the normal equation

Almost done! We are solving $\mathbf{X}^\top \mathbf{X} \hat{\mathbf{w}} = \mathbf{X}^\top \mathbf{y}$

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 2 & 2 \\ 2 & 2 + 2c^2 \end{pmatrix} \qquad \mathbf{X}^\top \mathbf{y} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

Hence, we can write

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 + c^2 \end{pmatrix} \begin{pmatrix} \hat{w}_1 \\ \hat{w}_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

What is the solution?

Solving the normal equation

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 + c^2 \end{pmatrix} \begin{pmatrix} \hat{w}_1 \\ \hat{w}_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Solving the normal equation

$$\begin{pmatrix} 1 & 1 \\ 1 & 1+c^2 \end{pmatrix} \begin{pmatrix} \hat{w}_1 \\ \hat{w}_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Implies

Solving the normal equation

$$\begin{pmatrix} 1 & 1 \\ 1 & 1+c^2 \end{pmatrix} \begin{pmatrix} \hat{w}_1 \\ \hat{w}_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Implies

$$\Bigg\{$$

Solving the normal equation

$$\begin{pmatrix} 1 & 1 \\ 1 & 1+c^2 \end{pmatrix} \begin{pmatrix} \hat{w}_1 \\ \hat{w}_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Implies

$$\begin{cases} \hat{w}_1 + \hat{w}_2 = 1 \\ \end{cases}$$

Solving the normal equation

$$\begin{pmatrix} 1 & 1 \\ 1 & 1+c^2 \end{pmatrix} \begin{pmatrix} \hat{w}_1 \\ \hat{w}_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Implies

$$\begin{cases} \hat{w}_1 + \hat{w}_2 = 1 \\ \hat{w}_1 + (1+c^2)\hat{w}_2 = 1 \end{cases}$$

Solving the normal equation

$$\begin{pmatrix} 1 & 1 \\ 1 & 1+c^2 \end{pmatrix} \begin{pmatrix} \hat{w}_1 \\ \hat{w}_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Implies

$$\begin{cases} \hat{w}_1 + \hat{w}_2 = 1 \\ \hat{w}_1 + (1+c^2)\hat{w}_2 = 1 \end{cases} \qquad \left\{ \vphantom{\begin{matrix} a \\ b \end{matrix}} \right.$$

Solving the normal equation

$$\begin{pmatrix} 1 & 1 \\ 1 & 1+c^2 \end{pmatrix} \begin{pmatrix} \hat{w}_1 \\ \hat{w}_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Implies

$$\begin{cases} \hat{w}_1 + \hat{w}_2 = 1 \\ \hat{w}_1 + (1+c^2)\hat{w}_2 = 1 \end{cases}$$

$$\begin{cases} \hat{w}_1 = 1 - \hat{w}_2 \\ \end{cases}$$

Solving the normal equation

$$\begin{pmatrix} 1 & 1 \\ 1 & 1+c^2 \end{pmatrix} \begin{pmatrix} \hat{w}_1 \\ \hat{w}_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Implies

$$\begin{cases} \hat{w}_1 + \hat{w}_2 = 1 \\ \hat{w}_1 + (1+c^2)\hat{w}_2 = 1 \end{cases} \qquad \begin{cases} \hat{w}_1 = 1 - \hat{w}_2 \\ -\hat{w}_2 + (1+c^2)\hat{w}_2 = 0 \end{cases}$$

Solving the normal equation

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 + c^2 \end{pmatrix} \begin{pmatrix} \hat{w}_1 \\ \hat{w}_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Implies

$$\begin{cases} \hat{w}_1 + \hat{w}_2 = 1 \\ \hat{w}_1 + (1 + c^2)\hat{w}_2 = 1 \end{cases} \qquad \begin{cases} \hat{w}_1 = 1 - \hat{w}_2 \\ -\hat{w}_2 + (1 + c^2)\hat{w}_2 = 0 \end{cases}$$

$$\hat{w}_1 = 1$$

Solving the normal equation

$$\begin{pmatrix} 1 & 1 \\ 1 & 1+c^2 \end{pmatrix} \begin{pmatrix} \hat{w}_1 \\ \hat{w}_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Implies

$$\begin{cases} \hat{w}_1 + \hat{w}_2 = 1 \\ \hat{w}_1 + (1+c^2)\hat{w}_2 = 1 \end{cases}$$

$$\begin{cases} \hat{w}_1 = 1 - \hat{w}_2 \\ -\hat{w}_2 + (1+c^2)\hat{w}_2 = 0 \end{cases}$$

$$\hat{w}_1 = 1$$
$$\hat{w}_2 = 0$$

Solving the normal equation

$$\begin{pmatrix} 1 & 1 \\ 1 & 1+c^2 \end{pmatrix} \begin{pmatrix} \hat{w}_1 \\ \hat{w}_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

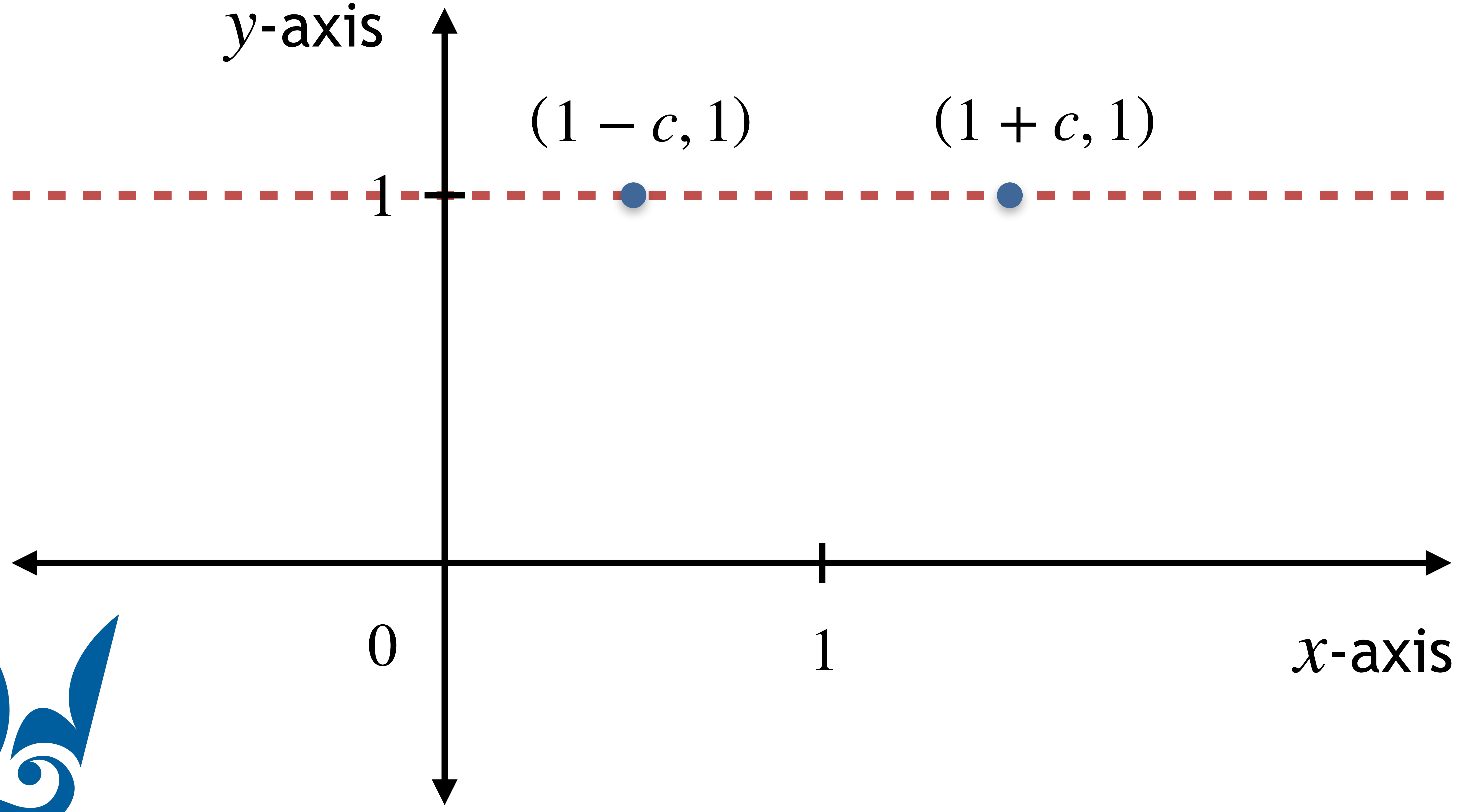
Implies

$$\begin{cases} \hat{w}_1 + \hat{w}_2 = 1 \\ \hat{w}_1 + (1+c^2)\hat{w}_2 = 1 \end{cases} \qquad \begin{cases} \hat{w}_1 = 1 - \hat{w}_2 \\ -\hat{w}_2 + (1+c^2)\hat{w}_2 = 0 \end{cases}$$

$$\begin{aligned} \hat{w}_1 &= 1 \\ \hat{w}_2 &= 0 \end{aligned} \quad \rightarrow \quad \hat{\mathbf{w}} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Unstable regression problems

What if we make a small error when measuring y_1 and y_2?



$$\hat{\mathbf{w}} = (\hat{w}_0, \hat{w}_1) = (1, 0)$$

Data points

$x_1 = 1 - c$

$y_1 = 1$

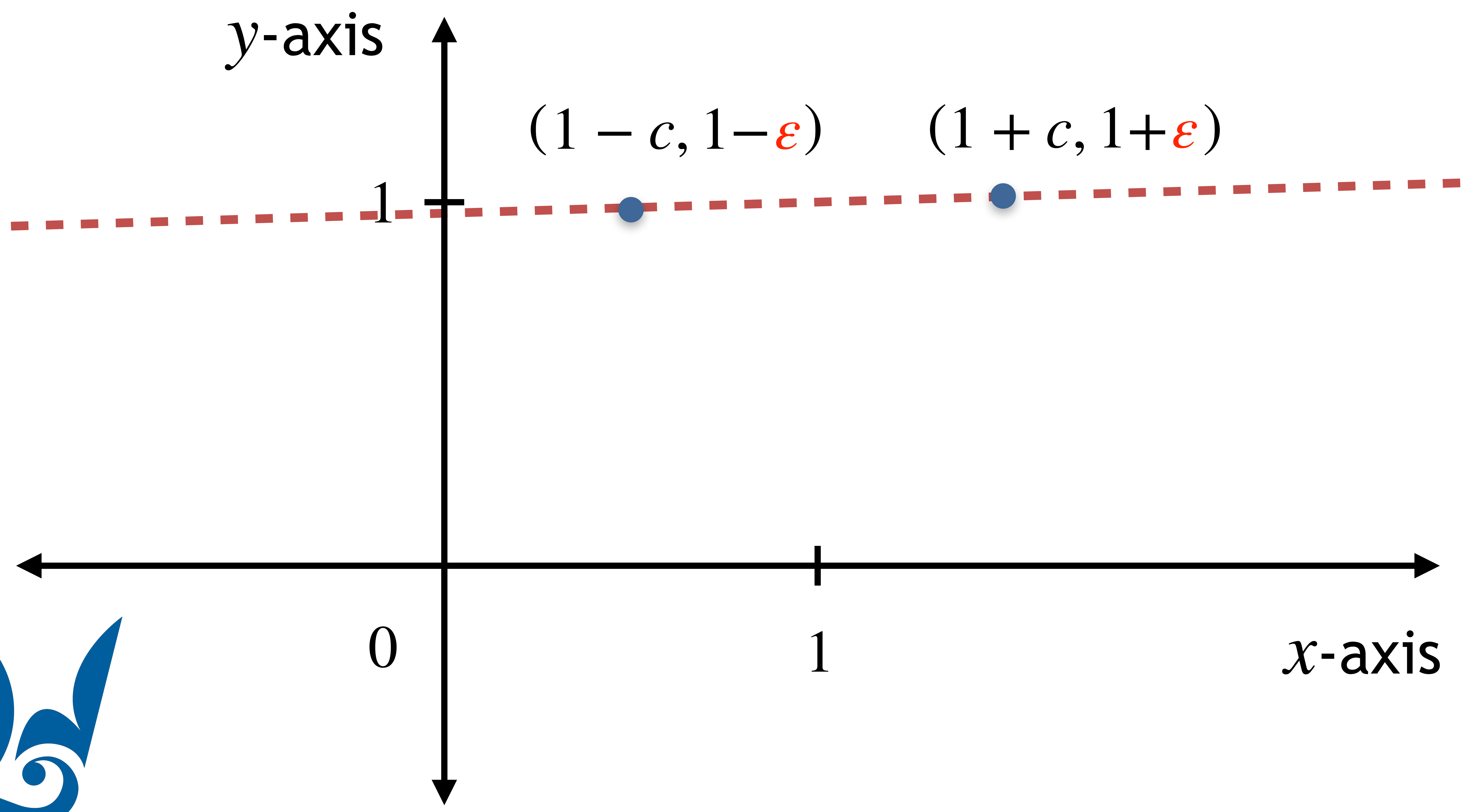
$x_2 = 1 + c$

$y_2 = 1$

for some $c > 0$

Unstable regression problems

What if we make a small error when measuring y_1 and y_2?



$\hat{\mathbf{w}} = (\hat{w}_0, \hat{w}_1) = ?$

Data points

$x_1 = 1 - c$

$y_1 = 1 - \varepsilon$

$x_2 = 1 + c$

$y_2 = 1 + \varepsilon$

for some $c > 0$ and $\varepsilon > 0$

Unstable regression problems

Data matrix $\mathbf{X}$ and data vector $\mathbf{y}$ of our problem read

$$\mathbf{X} = \begin{pmatrix} 1 & 1 - c \\ 1 & 1 + c \end{pmatrix} \qquad \text{and} \qquad \mathbf{y} = \begin{pmatrix} 1 - \varepsilon \\ 1 + \varepsilon \end{pmatrix} ;$$

Unstable regression problems

Data matrix $\mathbf{X}$ and data vector $\mathbf{y}$ of our problem read

$$\mathbf{X} = \begin{pmatrix} 1 & 1-c \\ 1 & 1+c \end{pmatrix} \quad \text{and} \quad \mathbf{y} = \begin{pmatrix} 1 - \varepsilon \\ 1 + \varepsilon \end{pmatrix} ;$$

hence, we compute

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 2 & 2 \\ 2 & 2+2c^2 \end{pmatrix} \quad \text{and} \quad \mathbf{X}^\top \mathbf{y} = \begin{pmatrix} 2 \\ 2 + 2c\varepsilon \end{pmatrix}$$

Unstable regression problems

Data matrix $\mathbf{X}$ and data vector $\mathbf{y}$ of our problem read

$$\mathbf{X} = \begin{pmatrix} 1 & 1-c \\ 1 & 1+c \end{pmatrix} \qquad \text{and} \qquad \mathbf{y} = \begin{pmatrix} 1 - \varepsilon \\ 1 + \varepsilon \end{pmatrix};$$

hence, we compute

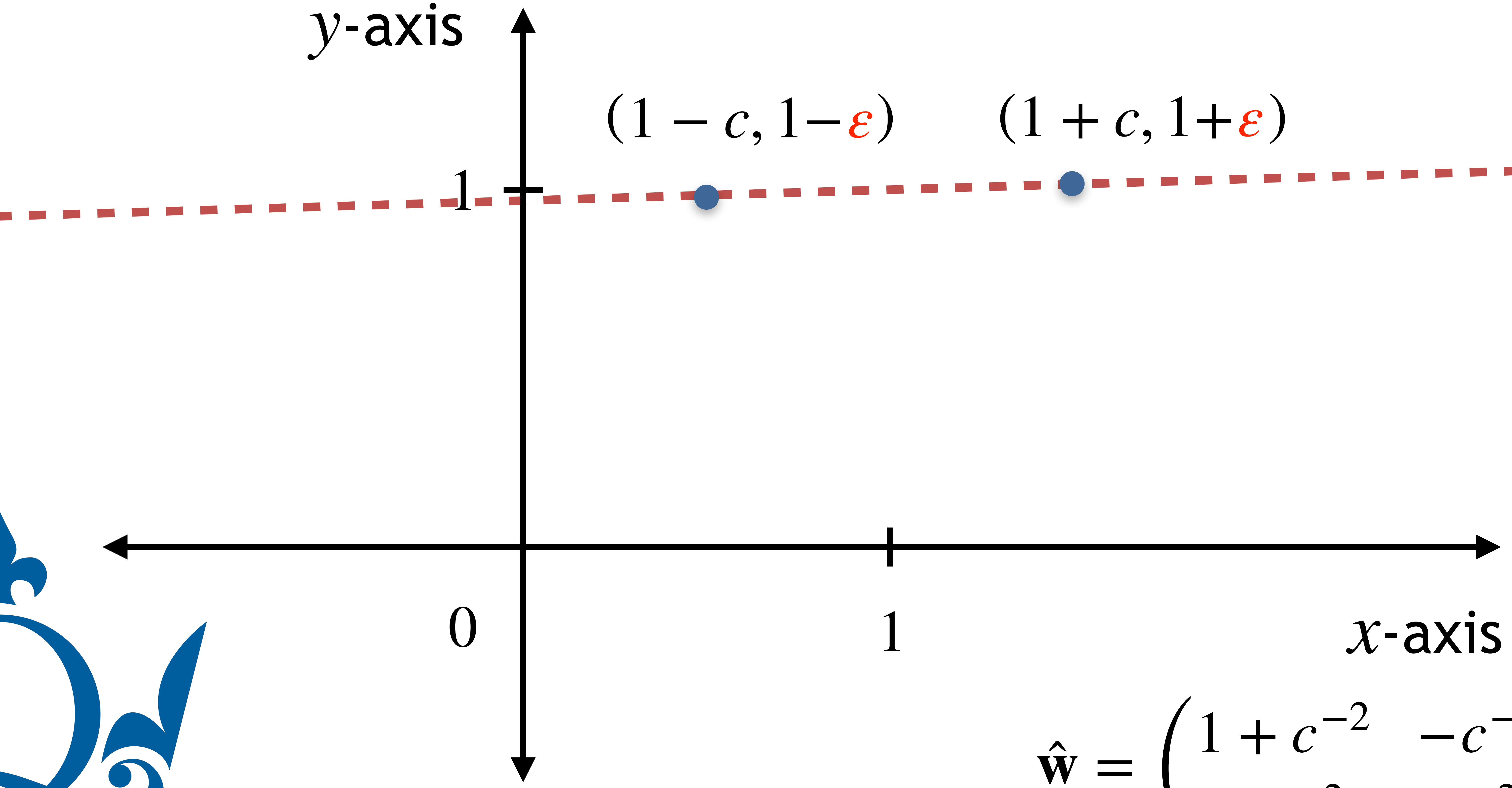
$$\mathbf{X}^\top\mathbf{X} = \begin{pmatrix} 2 & 2 \\ 2 & 2+2c^2 \end{pmatrix} \qquad \text{and} \qquad \mathbf{X}^\top\mathbf{y} = \begin{pmatrix} 2 \\ 2 + 2c\varepsilon \end{pmatrix}$$

and compute $\hat{w}$ via

$$\begin{pmatrix} 1 & 1 \\ 1 & 1+c^2 \end{pmatrix} \hat{\mathbf{w}} = \begin{pmatrix} 1 \\ 1 + c\varepsilon \end{pmatrix} \implies \hat{\mathbf{w}} = \begin{pmatrix} 1+c^{-2} & -c^{-2} \\ -c^{-2} & c^{-2} \end{pmatrix} \begin{pmatrix} 1 \\ 1 + c\varepsilon \end{pmatrix} = \begin{pmatrix} 1 - c^{-1}\varepsilon \\ c^{-1}\varepsilon \end{pmatrix}$$

Unstable regression problems

What if we make a small error when measuring y_1 and y_2?



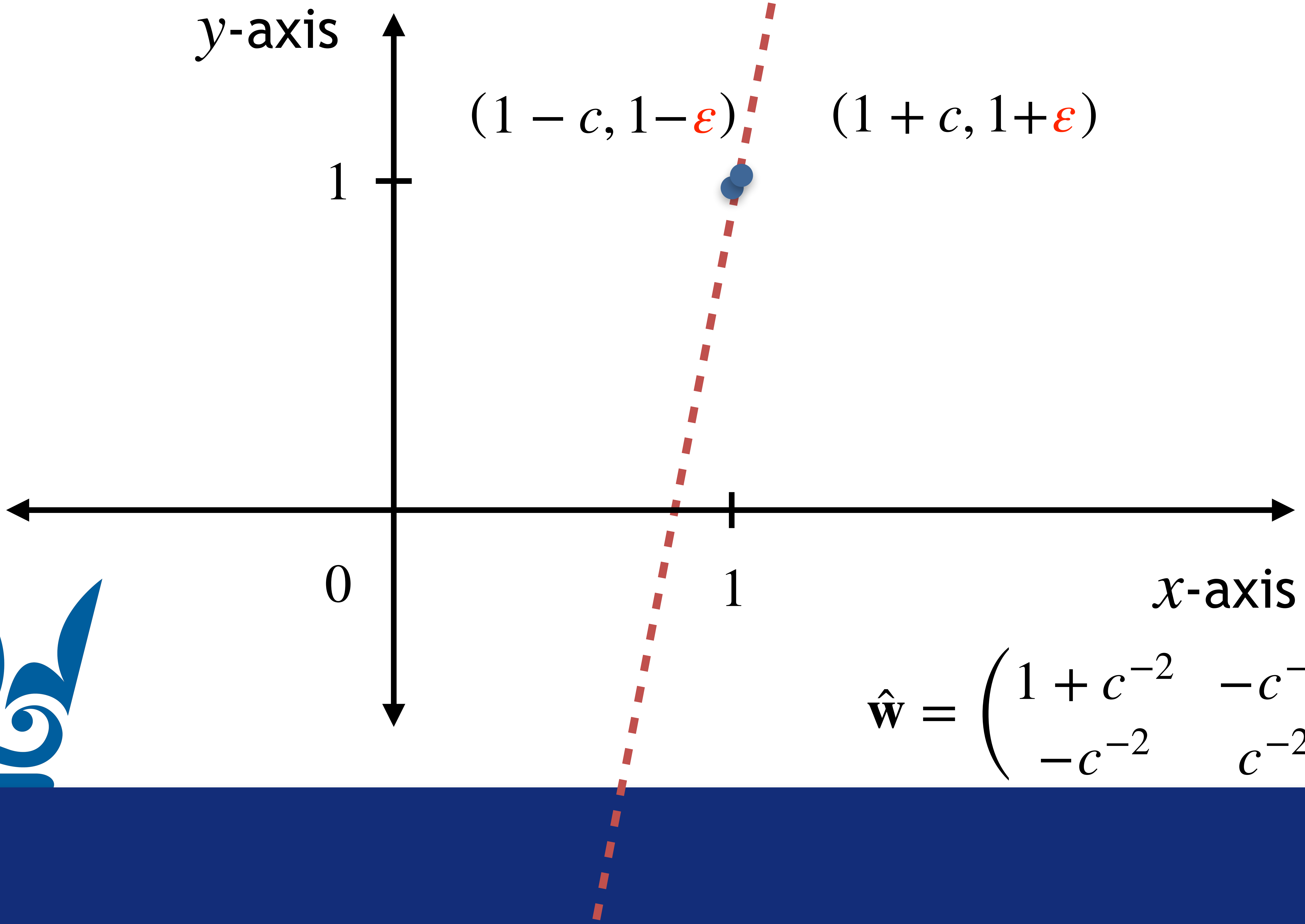
$c = 1/2 \qquad \varepsilon = 1/100$

$$= \begin{pmatrix} \frac{49}{50} \\ \frac{1}{50} \end{pmatrix} = \begin{pmatrix} 0.98 \\ 0.02 \end{pmatrix}$$

$$\hat{\mathbf{w}} = \begin{pmatrix} 1 + c^{-2} & -c^{-2} \\ -c^{-2} & c^{-2} \end{pmatrix} \begin{pmatrix} 1 \\ 1 + c\varepsilon \end{pmatrix} = \begin{pmatrix} 1 - c^{-1}\varepsilon \\ c^{-1}\varepsilon \end{pmatrix}$$

Unstable regression problems

What if we make a small error when measuring y_1 and y_2?



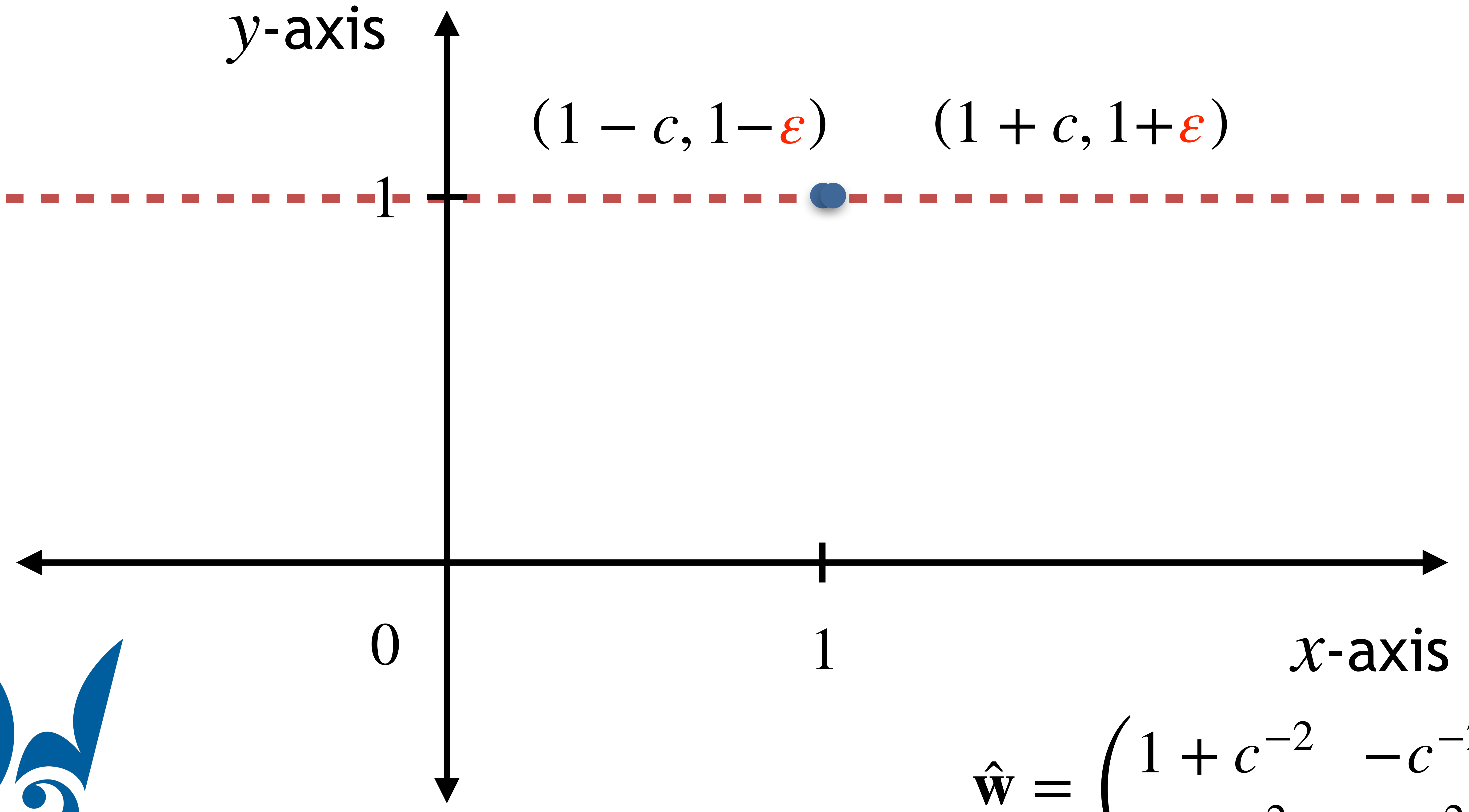
$c = 1/1000 \qquad \varepsilon = 1/100$

$$\hat{\mathbf{w}} = \begin{pmatrix} 1 + c^{-2} & -c^{-2} \\ -c^{-2} & c^{-2} \end{pmatrix} \begin{pmatrix} 1 \\ 1 + c\varepsilon \end{pmatrix} = \begin{pmatrix} 1 - c^{-1}\varepsilon \\ c^{-1}\varepsilon \end{pmatrix}$$

$$= \begin{pmatrix} -9 \\ 10 \end{pmatrix}$$

Unstable regression problems

What if we make a small error when measuring y_1 and y_2?



$c = 1/1000 \qquad \varepsilon = 0$

$$= \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$\hat{\mathbf{w}} = \begin{pmatrix} 1 + c^{-2} & -c^{-2} \\ -c^{-2} & c^{-2} \end{pmatrix} \begin{pmatrix} 1 \\ 1 + c\varepsilon \end{pmatrix} = \begin{pmatrix} 1 - c^{-1}\varepsilon \\ c^{-1}\varepsilon \end{pmatrix}$$

Unstable regression problems

Polynomial regression

$$\hat{\mathbf{w}} = \arg\min_{w \in \mathbb{R}^{d+1}} \mathsf{MSE}(\mathbf{w}) = \arg\min_{\mathbf{w} \in \mathbb{R}^{d+1}} \left\{ \frac{1}{2s} \| \boldsymbol{\Phi}(\boldsymbol{X}) \boldsymbol{w} - \boldsymbol{y} \|^2 \right\}$$

with

$$\boldsymbol{\Phi}(\boldsymbol{X}) = \begin{pmatrix} 1 & x_1 & \cdots & x_1^d \\ 1 & x_2 & \cdots & x_2^d \\ \vdots & \vdots & & \vdots \\ 1 & x_s & \dots & x_s^d \end{pmatrix}$$

Unstable regression problems

Polynomial regression

$$\hat{\mathbf{w}} = \arg\min_{w \in \mathbb{R}^{d+1}} \mathsf{MSE}(\mathbf{w}) = \arg\min_{\mathbf{w} \in \mathbb{R}^{d+1}} \left\{ \frac{1}{2s} \|\boldsymbol{\Phi}(X)\boldsymbol{w} - \boldsymbol{y}\|^2 \right\}$$

with

$$\boldsymbol{\Phi}(X) = \begin{pmatrix} 1 & x_1 & \cdots & x_1^d \\ 1 & x_2 & \cdots & x_2^d \\ \vdots & \vdots & & \vdots \\ 1 & x_s & \ldots & x_s^d \end{pmatrix}$$

Solution:

$$\boldsymbol{\Phi}(X)^\top \boldsymbol{\Phi}(X)\hat{\boldsymbol{w}} = \boldsymbol{\Phi}(X)^\top \boldsymbol{y}$$

Unstable regression problems

Polynomial regression

$$\hat{\mathbf{w}} = \arg\min_{w\in\mathbb{R}^{d+1}} \mathrm{MSE}(\mathbf{w}) = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{2s} \|\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{w} - \boldsymbol{y}\|^2 \right\}$$

with

$$\boldsymbol{\Phi}(\boldsymbol{X}) = \begin{pmatrix} 1 & x_1 & \cdots & x_1^d \\ 1 & x_2 & \cdots & x_2^d \\ \vdots & \vdots & & \vdots \\ 1 & x_s & \ldots & x_s^d \end{pmatrix}$$

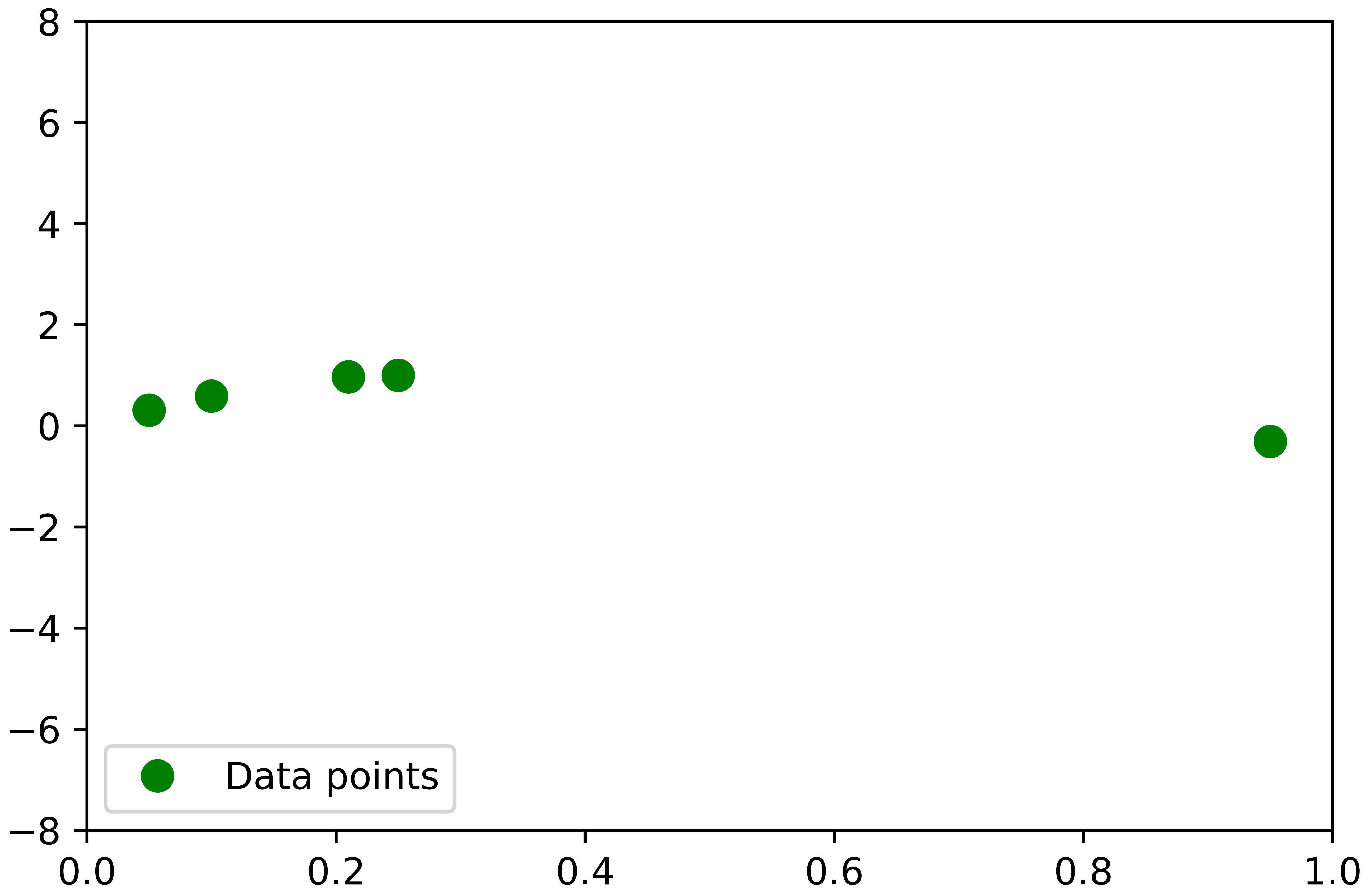
Solution:

$$\boldsymbol{\Phi}(\boldsymbol{X})^\top \boldsymbol{\Phi}(\boldsymbol{X})\hat{\boldsymbol{w}} = \boldsymbol{\Phi}(\boldsymbol{X})^\top \boldsymbol{y}$$

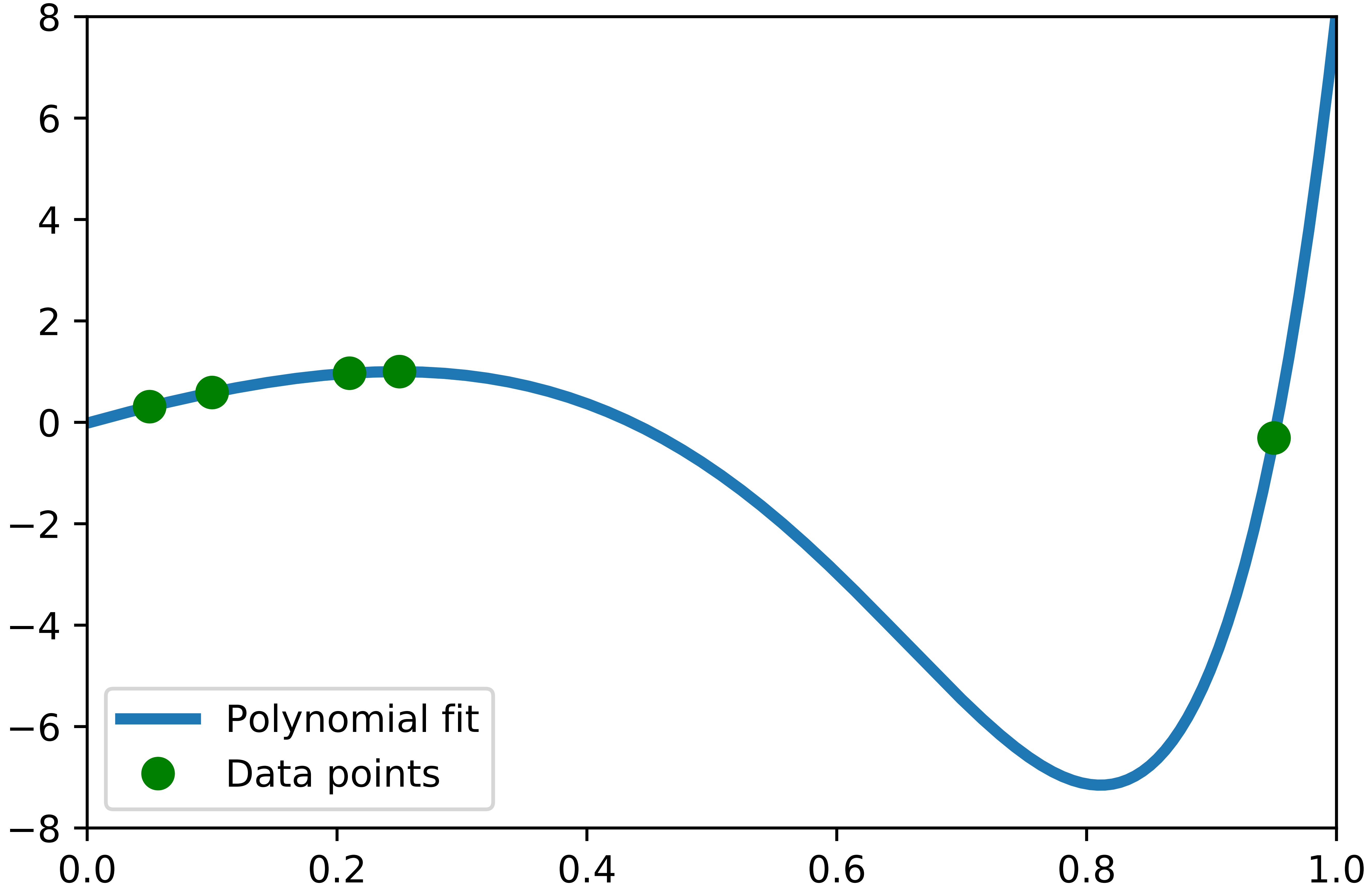
Entries of $\Phi(X)^\top\Phi(X)$:

$$\left(\Phi(X)^\top\Phi(X)\right)_{jk} = \sum_{i=1}^{s} x_i^{j+k-2}$$

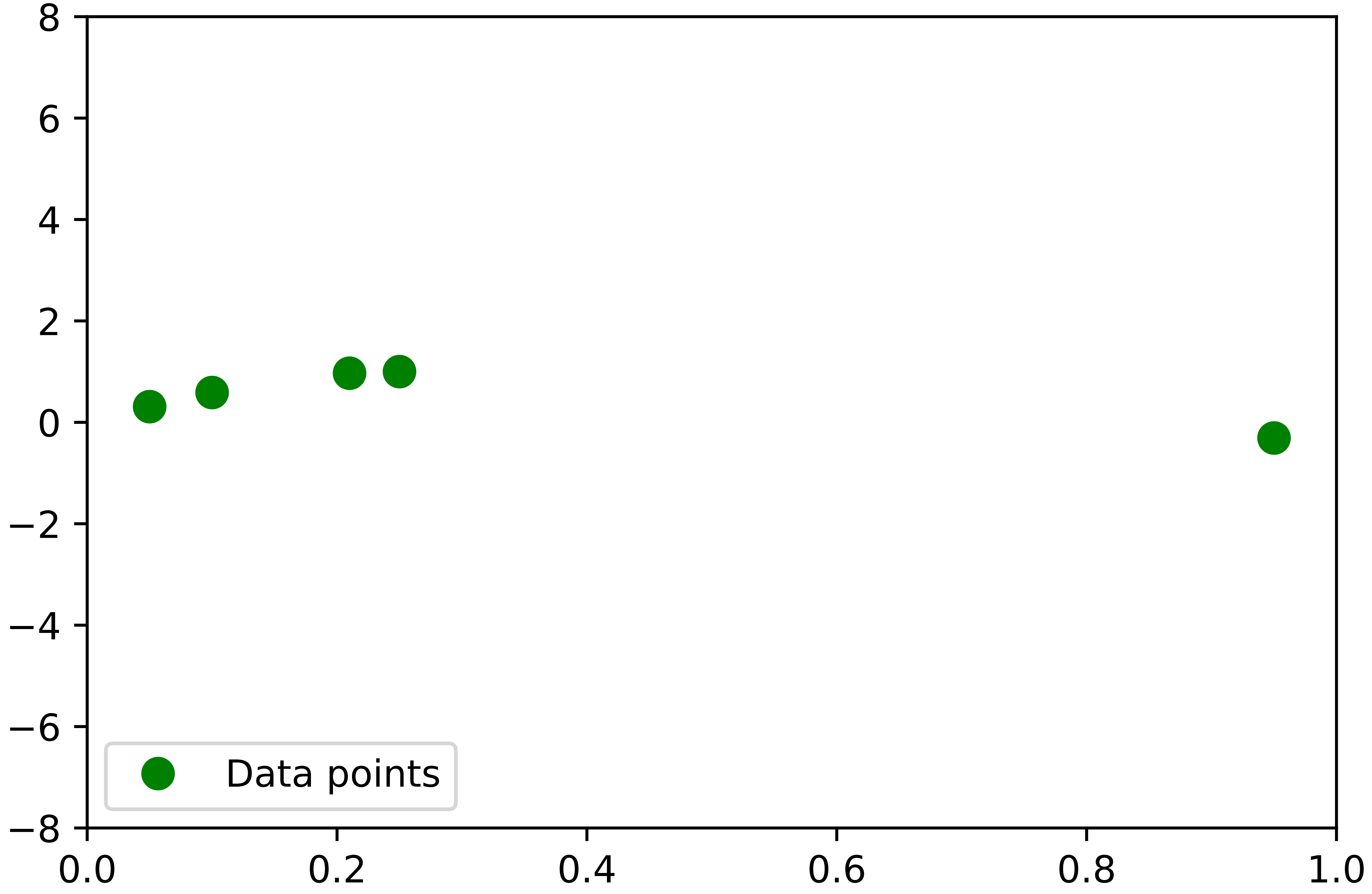
Unstable regression problems



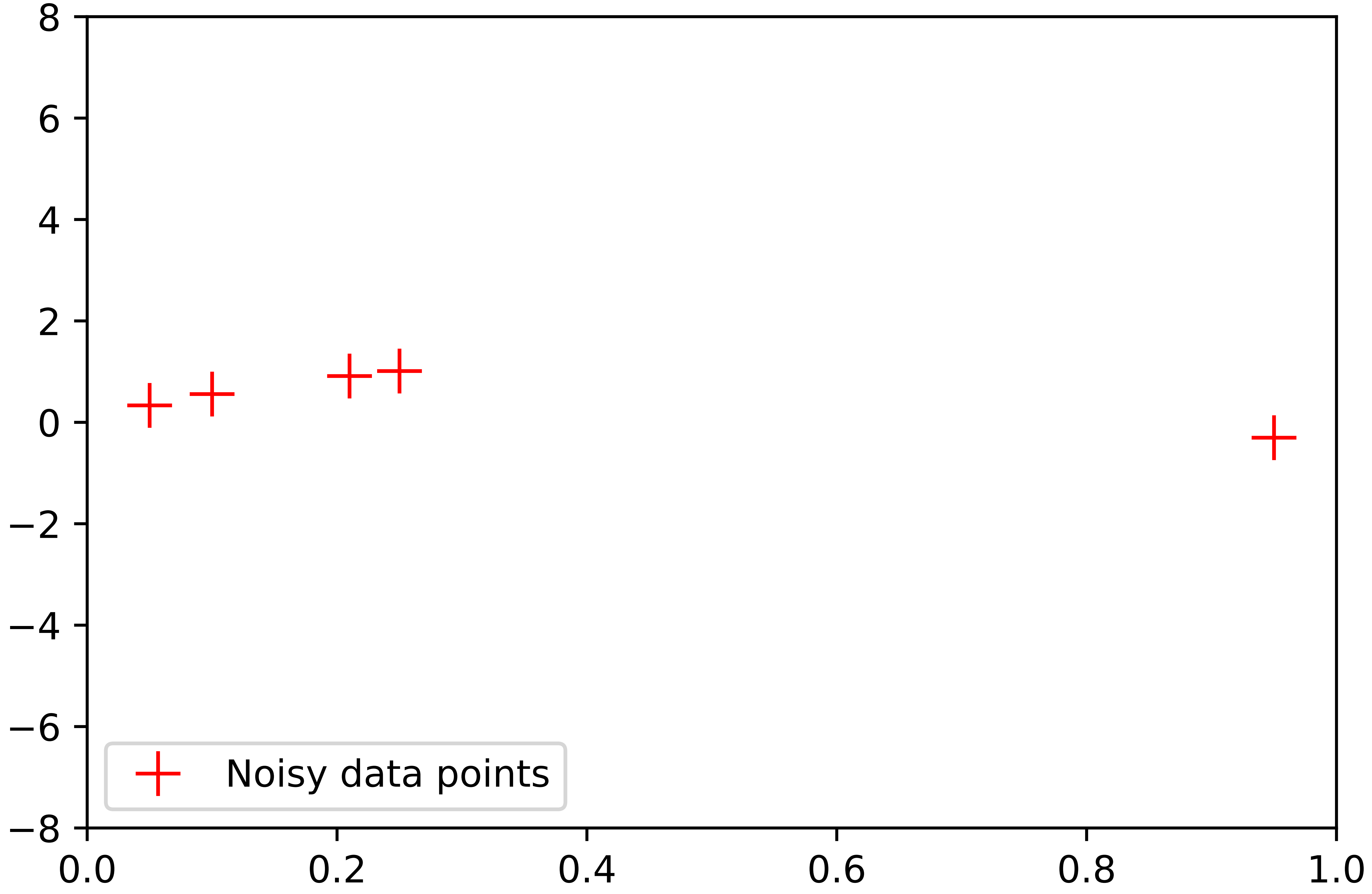
Unstable regression problems



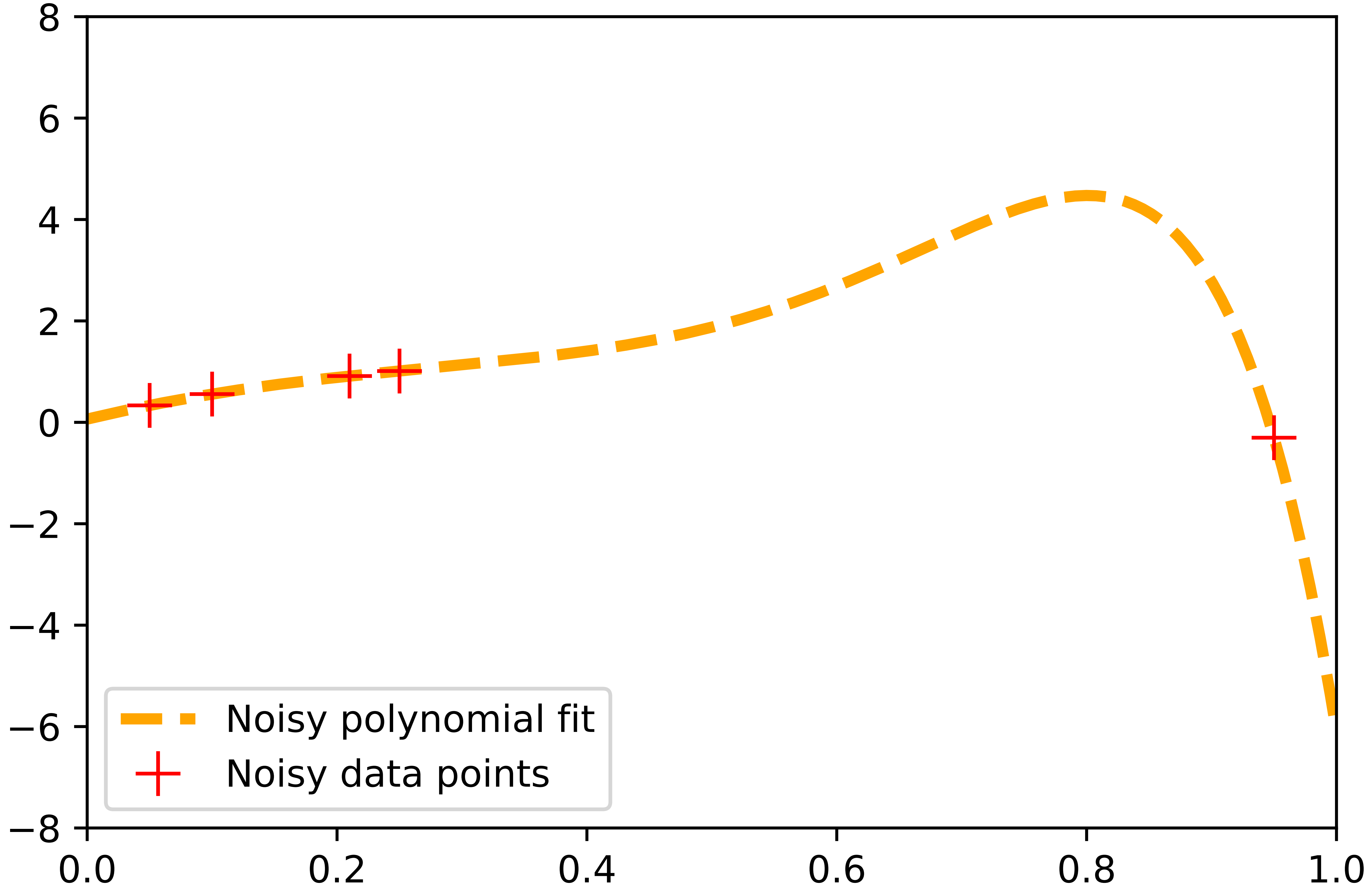
Unstable regression problems



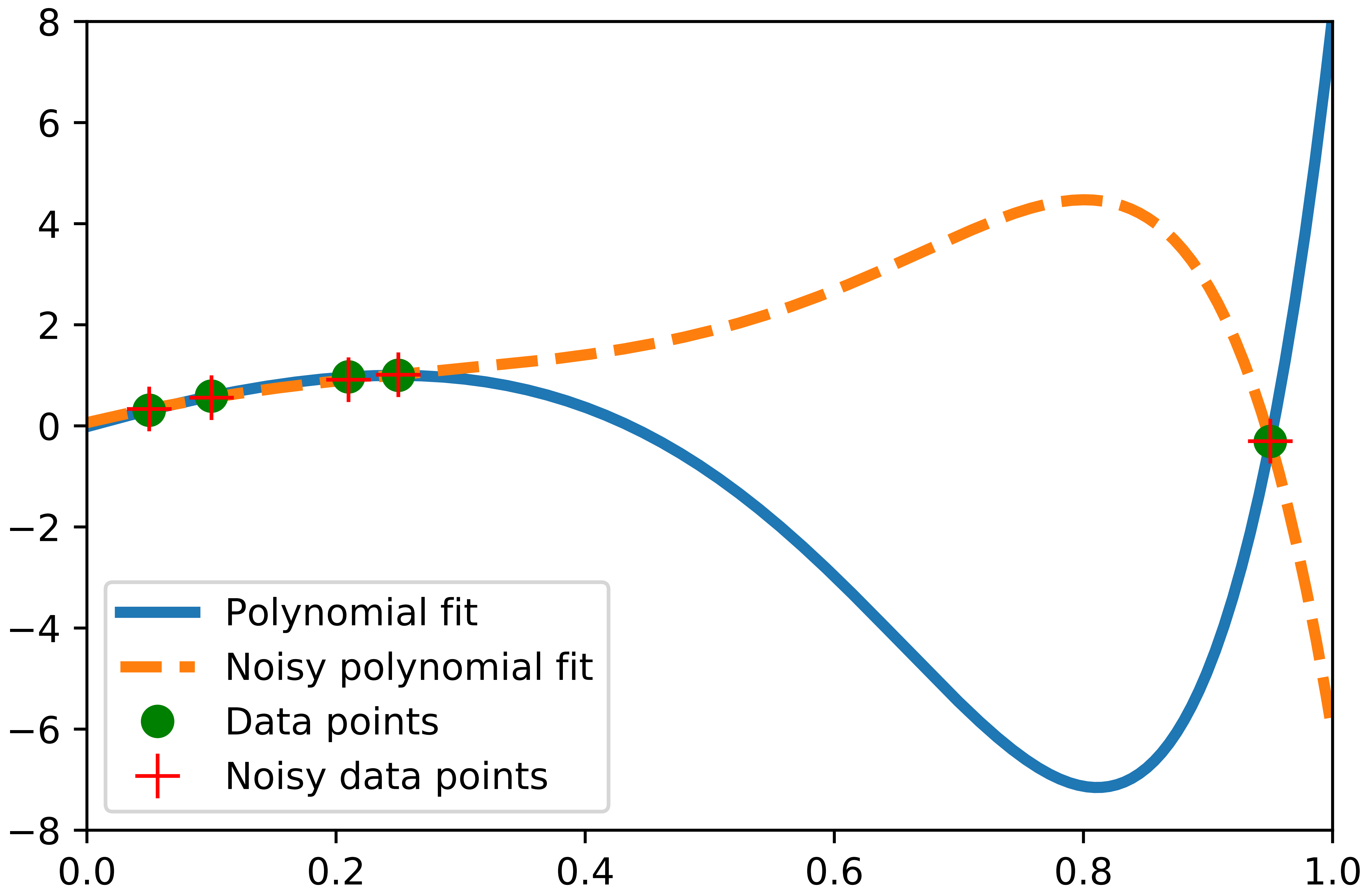
Unstable regression problems



Unstable regression problems



Unstable regression problems



Stability

We have discovered two phenomena:

Stability

We have discovered two phenomena:

1. Having two slightly different outputs $\boldsymbol{y}$ and $\boldsymbol{y}_\delta$ with $\|\boldsymbol{y}-\boldsymbol{y}_\delta\| \leq \delta$, we have seen examples where $\|\hat{\boldsymbol{w}}-\hat{\boldsymbol{w}}_\delta\| \gg \delta$, for

$$\hat{\mathbf{w}} = \arg\min_{\boldsymbol{w}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{2}\|\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{w}-\boldsymbol{y}\|^2 \right\} \quad \text{and} \quad \hat{\mathbf{w}}_\delta = \arg\min_{\boldsymbol{w}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{2}\|\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{w}-\boldsymbol{y}_\delta\|^2 \right\}$$

Stability

We have discovered two phenomena:

1. Having two slightly different outputs $\boldsymbol{y}$ and $\boldsymbol{y}_\delta$ with $\|\boldsymbol{y} - \boldsymbol{y}_\delta\| \leq \delta$, we have seen examples where $\|\hat{\boldsymbol{w}} - \hat{\boldsymbol{w}}_\delta\| \gg \delta$, for

$$\hat{\mathbf{w}} = \arg\min_{\boldsymbol{w}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{2}\|\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{w} - \boldsymbol{y}\|^2 \right\} \quad \text{and} \quad \hat{\mathbf{w}}_\delta = \arg\min_{\boldsymbol{w}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{2}\|\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{w} - \boldsymbol{y}_\delta\|^2 \right\}$$

2. Having two instances $(\boldsymbol{X}_t, \boldsymbol{y}_t)$ and $(\boldsymbol{X}_v, \boldsymbol{y}_v)$ of data samples, we have seen examples where $\|\boldsymbol{\Phi}(\boldsymbol{X}_v)\,\hat{\boldsymbol{w}} - \boldsymbol{y}_v\| \gg \|\boldsymbol{\Phi}(\boldsymbol{X}_t)\,\hat{\boldsymbol{w}} - \boldsymbol{y}_t\|$, for

$$\hat{\mathbf{w}} = \arg\min_{\boldsymbol{w}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{2}\|\boldsymbol{\Phi}(\boldsymbol{X}_t)\,\boldsymbol{w} - \boldsymbol{y}_t\|^2 \right\}$$

Stability

For different values of c and ε we observed

$c > 0$

$\varepsilon = 0$

$$\hat{\mathbf{w}} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Stability

For different values of c and ε we observed

$c > 0$

$\varepsilon = 0$

$$\hat{\mathbf{w}} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$c = 1/2$

$\varepsilon = 1/100$

$$\hat{\mathbf{w}}_\delta = \begin{pmatrix} \frac{49}{50} \\ \frac{1}{50} \end{pmatrix} = \begin{pmatrix} 0.98 \\ 0.02 \end{pmatrix}$$

Stability

For different values of c and ε we observed

$c > 0$

$\varepsilon = 0$

$$\hat{\mathbf{w}} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$c = 1/2$

$\varepsilon = 1/100$

$$\hat{\mathbf{w}}_\delta = \begin{pmatrix} \frac{49}{50} \\ \frac{1}{50} \end{pmatrix} = \begin{pmatrix} 0.98 \\ 0.02 \end{pmatrix}$$

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\| = \frac{1}{\sqrt{1250}}$$

$$\approx 0.0283$$

Stability

For different values of c and ε we observed

$c > 0$

$\varepsilon = 0$

$$\hat{\mathbf{w}} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$c = 1/2$

$\varepsilon = 1/100$

$$\hat{\mathbf{w}}_\delta = \begin{pmatrix} \frac{49}{50} \\ \frac{1}{50} \end{pmatrix} = \begin{pmatrix} 0.98 \\ 0.02 \end{pmatrix}$$

$c = 1/1000$

$\varepsilon = 1/100$

$$\hat{\mathbf{w}}_\delta = \begin{pmatrix} -9 \\ 10 \end{pmatrix}$$

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\| = \frac{1}{\sqrt{1250}}$$

$$\approx 0.0283$$

Stability

For different values of c and ε we observed

$$c > 0$$
$$\varepsilon = 0$$
$$\hat{\mathbf{w}} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$c = 1/2$$
$$\varepsilon = 1/100$$
$$\hat{\mathbf{w}}_\delta = \begin{pmatrix} \frac{49}{50} \\ \frac{1}{50} \end{pmatrix} = \begin{pmatrix} 0.98 \\ 0.02 \end{pmatrix}$$

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\| = \frac{1}{\sqrt{1250}}$$
$$\approx 0.0283$$

$$c = 1/1000$$
$$\varepsilon = 1/100$$
$$\hat{\mathbf{w}}_\delta = \begin{pmatrix} -9 \\ 10 \end{pmatrix}$$

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\| = \sqrt{200}$$
$$\approx 14.14$$

Stability

Let us characterise the error $\|\hat{\boldsymbol{w}} - \hat{\boldsymbol{w}}_\delta\|$ for general problems

Stability

Let us characterise the error $\|\hat{\boldsymbol{w}} - \hat{\boldsymbol{w}}_\delta\|$ for general problems

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{2}\|\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{w} - \boldsymbol{y}\|^2 \right\} \quad \text{and} \quad \hat{\boldsymbol{w}}_\delta = \arg\min_{\boldsymbol{w}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{2}\|\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{w} - \boldsymbol{y}_\delta\|^2 \right\}$$

Stability

The solutions for

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{2}\|\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{w} - \boldsymbol{y}\|^2 \right\} \quad \text{and} \quad \hat{\mathbf{w}}_\delta = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{2}\|\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{w} - \boldsymbol{y}_\delta\|^2 \right\}$$

are

$$\hat{\boldsymbol{w}} = \left(\boldsymbol{\Phi}(\boldsymbol{X})^\top\boldsymbol{\Phi}(\boldsymbol{X})\right)^{-1}\boldsymbol{\Phi}(\boldsymbol{X})^\top\boldsymbol{y} \quad \text{and} \quad \hat{\boldsymbol{w}}_\delta = \left(\boldsymbol{\Phi}(\boldsymbol{X})^\top\boldsymbol{\Phi}(\boldsymbol{X})\right)^{-1}\boldsymbol{\Phi}(\boldsymbol{X})^\top\boldsymbol{y}_\delta$$

Stability

The matrix $\boldsymbol{\Phi}(\boldsymbol{X})^\top\boldsymbol{\Phi}(\boldsymbol{X})$ is a symmetric, positive definite matrix with real entries and

$$\hat{\boldsymbol{w}} = \left(\boldsymbol{\Phi}(\boldsymbol{X})^\top\boldsymbol{\Phi}(\boldsymbol{X})\right)^{-1}\boldsymbol{\Phi}(\boldsymbol{X})^\top\boldsymbol{y} \qquad \text{and} \qquad \hat{\boldsymbol{w}}_\delta = \left(\boldsymbol{\Phi}(\boldsymbol{X})^\top\boldsymbol{\Phi}(\boldsymbol{X})\right)^{-1}\boldsymbol{\Phi}(\boldsymbol{X})^\top\boldsymbol{y}_\delta$$

Using the SVD we can prove that (see notes)

$$\hat{\boldsymbol{w}} = \boldsymbol{V}(\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{U}^T\boldsymbol{y} \qquad \text{and} \qquad \hat{\boldsymbol{w}}_\delta = \boldsymbol{V}(\boldsymbol{\Sigma}^\top)^{-1}\boldsymbol{U}^T\boldsymbol{y}_\delta$$

Stability

Hence

$$\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta = \mathbf{V}(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\mathbf{U}^{\mathbf{T}}(\mathbf{y} - \mathbf{y}_\delta)$$

Stability

Hence

$$\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta = \mathbf{V}(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\mathbf{U}^{\mathbf{T}}(\mathbf{y} - \mathbf{y}_\delta)$$

We can then calculate

Stability

Hence

$$\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta = \mathbf{V}(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\mathbf{U}^\mathbf{T}(\mathbf{y} - \mathbf{y}_\delta)$$

We can then calculate

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\|^2 = \|\mathbf{V}(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\mathbf{U}^\mathbf{T}(\mathbf{y} - \mathbf{y}_\delta)\|^2$$

Stability

Hence

$$\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta = \mathbf{V}(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\mathbf{U}^\mathbf{T}(\mathbf{y} - \mathbf{y}_\delta)$$

We can then calculate

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\|^2 = \|\mathbf{V}(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\mathbf{U}^\mathbf{T}(\mathbf{y} - \mathbf{y}_\delta)\|^2 \leq \ \|\mathbf{V}\|^\mathbf{2}\|(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\|^\mathbf{2}\|\mathbf{U}^\mathbf{T}\|^\mathbf{2}\|(\mathbf{y} - \mathbf{y}_\delta)\|^2$$

Stability

Hence

$$\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta = \mathbf{V}(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\mathbf{U}^{\mathbf{T}}(\mathbf{y} - \mathbf{y}_\delta)$$

We can then calculate

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\|^2 = \|\mathbf{V}(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\mathbf{U}^{\mathbf{T}}(\mathbf{y} - \mathbf{y}_\delta)\|^2 \leq \overbrace{\|\mathbf{V}\|^{\mathbf{2}}}^{1}\|(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\|^{\mathbf{2}}\overbrace{\|\mathbf{U}^{\mathbf{T}}\|^{\mathbf{2}}}^{1}\|(\mathbf{y} - \mathbf{y}_\delta)\|^2$$

Stability

Hence

$$\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta = \mathbf{V}(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\mathbf{U}^\mathbf{T}(\mathbf{y} - \mathbf{y}_\delta)$$

We can then calculate

1

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\|^2 = \|\mathbf{V}(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\mathbf{U}^\mathbf{T}(\mathbf{y} - \mathbf{y}_\delta)\|^2 \leq \|\mathbf{V}\|^\mathbf{2}\|(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\|^\mathbf{2}\|\mathbf{U}^\mathbf{T}\|^\mathbf{2}\|(\mathbf{y} - \mathbf{y}_\delta)\|^2$$

Cauchy-Schwarz

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\|\,\|\mathbf{y}\|$$

Stability

Hence

$$\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta = \mathbf{V}(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\mathbf{U}^\mathbf{T}(\mathbf{y} - \mathbf{y}_\delta)$$

We can then calculate

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\|^2 = \|\mathbf{V}(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\mathbf{U}^\mathbf{T}(\mathbf{y} - \mathbf{y}_\delta)\|^2 \leq \underbrace{\|\mathbf{V}\|^\mathbf{2}}_{1}\|(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\|^\mathbf{2}\underbrace{\|\mathbf{U}^\mathbf{T}\|^\mathbf{2}}_{1}\|(\mathbf{y} - \mathbf{y}_\delta)\|^2$$

Cauchy-Schwarz

$|\langle \mathbf{x}, \mathbf{y}\rangle| \leq \|\mathbf{x}\|\|\mathbf{y}\|$

Stability

Hence

$$\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta = \mathbf{V}(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\mathbf{U}^\mathbf{T}(\mathbf{y} - \mathbf{y}_\delta)$$

We can then calculate

1

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\|^2 = \|\mathbf{V}(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\mathbf{U}^\mathbf{T}(\mathbf{y} - \mathbf{y}_\delta)\|^2 \leq \|\mathbf{V}\|^\mathbf{2}\|(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\|^\mathbf{2}\|\mathbf{U}^\mathbf{T}\|^\mathbf{2}\|(\mathbf{y} - \mathbf{y}_\delta)\|^2$$

def. of norm + diagonal matrix

Cauchy-Schwarz

$|\langle \mathbf{x}, \mathbf{y}\rangle| \leq \|\mathbf{x}\|\,\|\mathbf{y}\|$

Stability

Hence

$$\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta = \mathbf{V}(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\mathbf{U}^{\mathbf{T}}(\mathbf{y} - \mathbf{y}_\delta)$$

We can then calculate

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\|^2 = \|\mathbf{V}(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\mathbf{U}^{\mathbf{T}}(\mathbf{y} - \mathbf{y}_\delta)\|^2 \leq \underbrace{\|\mathbf{V}\|^{\mathbf{2}}}_{1}\|(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\|^{\mathbf{2}}\underbrace{\|\mathbf{U}^{\mathbf{T}}\|^{\mathbf{2}}}_{1}\|(\mathbf{y} - \mathbf{y}_\delta)\|^2$$

$$= \frac{\delta^2}{\sigma_{d+1}^2}$$

def. of norm + diagonal matrix

Cauchy-Schwarz

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\|\|\mathbf{y}\|$$

Stability

We just showed how

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\|^2 \leq \frac{\delta^2}{\sigma_{d+1}^2}$$

Stability

We just showed how

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\|^2 \leq \frac{\delta^2}{\sigma_{d+1}^2}$$

In the worst case the deviation is amplified by the smallest singular value

Stability

From

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\|^2 \leq \|(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\|^{\mathbf{2}}\|(\mathbf{y} - \mathbf{y}_\delta)\|^2$$

Stability

From

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\|^2 \leq \|(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\|^{\mathbf{2}}\|(\mathbf{y} - \mathbf{y}_\delta)\|^2$$

If we assume that $\mathbf{y} = \boldsymbol{\Phi}\mathbf{r}$ and $\mathbf{y}_\delta = \boldsymbol{\Phi}\mathbf{p}$

Stability

From

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\|^2 \leq \|(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\|^{\mathbf{2}}\|(\mathbf{y} - \mathbf{y}_\delta)\|^2$$

If we assume that $\mathbf{y} = \boldsymbol{\Phi}\mathbf{r}$ and $\mathbf{y}_\delta = \boldsymbol{\Phi}\mathbf{p}$

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\|^2 \leq \|(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\|^{\mathbf{2}}\|\boldsymbol{\Phi}(\mathbf{r} - \mathbf{p})\|^{\mathbf{2}} \leq \|(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\|^{\mathbf{2}}\|\boldsymbol{\Phi}\|^{\mathbf{2}}\|(\mathbf{r} - \mathbf{p})\|^2$$

Stability

From

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\|^2 \leq \|(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\|^{\mathbf{2}} \|(\mathbf{y} - \mathbf{y}_\delta)\|^2$$

If we assume that $\mathbf{y} = \boldsymbol{\Phi}\mathbf{r}$ and $\mathbf{y}_\delta = \boldsymbol{\Phi}\mathbf{p}$

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\|^2 \leq \|(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\|^{\mathbf{2}} \|\boldsymbol{\Phi}(\mathbf{r} - \mathbf{p})\|^{\mathbf{2}} \leq \|(\boldsymbol{\Sigma}^\top)^{-\mathbf{1}}\|^{\mathbf{2}} \|\boldsymbol{\Phi}\|^{\mathbf{2}} \|(\mathbf{r} - \mathbf{p})\|^2$$
$$= \frac{\sigma_1^2}{\sigma_{d+1}^2}\epsilon^2 = \kappa(\Phi)^2\epsilon^2$$

Stability

From

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\|^2 \leq \|(\mathbf{\Sigma}^\top)^{-\mathbf{1}}\|^{\mathbf{2}}\|(\mathbf{y} - \mathbf{y}_\delta)\|^2$$

If we assume that $\mathbf{y} = \mathbf{\Phi r}$ and $\mathbf{y}_\delta = \mathbf{\Phi p}$

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\|^2 \leq \|(\mathbf{\Sigma}^\top)^{-\mathbf{1}}\|^{\mathbf{2}}\|\mathbf{\Phi}(\mathbf{r} - \mathbf{p})\|^{\mathbf{2}} \leq \|(\mathbf{\Sigma}^\top)^{-\mathbf{1}}\|^{\mathbf{2}}\|\mathbf{\Phi}\|^{\mathbf{2}}\|(\mathbf{r} - \mathbf{p})\|^2$$
$$= \frac{\sigma_1^2}{\sigma_{d+1}^2}\epsilon^2 = \kappa(\Phi)^2\epsilon^2$$

Where $\kappa = \dfrac{\sigma_1}{\sigma_{d+1}}$ is the condition number that quantifies the amplification of the error in the worst case. A matrix with large kappa is called ill-conditioned

Stability

Back to our initial example for $d = 1$:

$$\mathbf{\Phi}(\mathbf{X}) = \mathbf{X} = \begin{pmatrix} 1 & 1 - c \\ 1 & 1 + c \end{pmatrix}$$

Stability

Back to our initial example for $d = 1$:

$$\mathbf{\Phi}(\mathbf{X}) = \mathbf{X} = \begin{pmatrix} 1 & 1 - c \\ 1 & 1 + c \end{pmatrix}$$

The singular values for this matrix are

$$\sigma_1 = \sqrt{2 + c^2 + \sqrt{c^4 + 4}} \quad \text{and} \quad \sigma_2 = \sqrt{2 + c^2 - \sqrt{c^4 + 4}}$$

Stability

Back to our initial example for $d = 1$:

$$\mathbf{\Phi}(\mathbf{X}) = \mathbf{X} = \begin{pmatrix} 1 & 1 - c \\ 1 & 1 + c \end{pmatrix}$$

The singular values for this matrix are

$$\sigma_1 = \sqrt{2 + c^2 + \sqrt{c^4 + 4}} \qquad \text{and} \qquad \sigma_2 = \sqrt{2 + c^2 - \sqrt{c^4 + 4}}$$

$$\Longrightarrow \qquad \kappa = \frac{\sigma_1}{\sigma_2} = \frac{2 + c^2 + \sqrt{c^4 + 4}}{2c}$$

Stability

$$\kappa = \frac{2 + c^2 + \sqrt{c^4 + 4}}{2c}$$

$$\hat{\mathbf{w}} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Stability

$$c = 1/2$$

$$\varepsilon = 1/100$$

$$\hat{\mathbf{w}}_\delta = \begin{pmatrix} \frac{49}{50} \\ \frac{1}{50} \end{pmatrix} = \begin{pmatrix} 0.98 \\ 0.02 \end{pmatrix}$$

$$\kappa = \frac{2 + c^2 + \sqrt{c^4 + 4}}{2c}$$

$$\hat{\mathbf{w}} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$\kappa \approx 4.3$$

Stability

$$c = 1/2$$

$$\varepsilon = 1/100$$

$$\kappa = \frac{2 + c^2 + \sqrt{c^4 + 4}}{2c}$$

$$\hat{\mathbf{w}}_\delta = \begin{pmatrix} \frac{49}{50} \\ \frac{1}{50} \end{pmatrix} = \begin{pmatrix} 0.98 \\ 0.02 \end{pmatrix}$$

$$\hat{\mathbf{w}} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$\kappa \approx 4.3$$

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\| \leq \kappa \epsilon \approx 0.04$$

Stability

$$\kappa = \frac{2 + c^2 + \sqrt{c^4 + 4}}{2c}$$

$$\hat{\mathbf{w}} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$c = 1/2$

$\varepsilon = 1/100$

$$\hat{\mathbf{w}}_\delta = \begin{pmatrix} \frac{49}{50} \\ \frac{1}{50} \end{pmatrix} = \begin{pmatrix} 0.98 \\ 0.02 \end{pmatrix}$$

$c = 1/1000$

$\varepsilon = 1/100$

$$\hat{\mathbf{w}}_\delta = \begin{pmatrix} -9 \\ 10 \end{pmatrix}$$

$\kappa \approx 4.3$

$\kappa \approx 2000$

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\| \leq \kappa\epsilon \approx 0.04$$

Stability

$$\kappa = \frac{2 + c^2 + \sqrt{c^4 + 4}}{2c}$$

$$\hat{\mathbf{w}} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$c = 1/2$

$\varepsilon = 1/100$

$$\hat{\mathbf{w}}_\delta = \begin{pmatrix} \frac{49}{50} \\ \frac{1}{50} \end{pmatrix} = \begin{pmatrix} 0.98 \\ 0.02 \end{pmatrix}$$

$$\kappa \approx 4.3$$

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\| \leq \kappa \epsilon \approx 0.04$$

$c = 1/1000$

$\varepsilon = 1/100$

$$\hat{\mathbf{w}}_\delta = \begin{pmatrix} -9 \\ 10 \end{pmatrix}$$

$$\kappa \approx 2000$$

$$\|\hat{\mathbf{w}} - \hat{\mathbf{w}}_\delta\| \leq \kappa \epsilon \approx 20$$

REGULARISATION METHODS

Ill-conditioned problems

What can we do in order to be less sensitive towards measurement errors?

$$\mathbf{X}^\top \mathbf{X}\,\mathbf{w} = \mathbf{X}^\top \mathbf{y}$$

Ill-conditioned problems

What can we do in order to be less sensitive towards measurement errors?

$$\left(\mathbf{X}^\top\mathbf{X} + \alpha I\right)\mathbf{w} = \mathbf{X}^\top\mathbf{y} \qquad \alpha > 0$$

We can add a multiple of the identity matrix —> shift of the singular values!

Ill-conditioned problems

What can we do in order to be less sensitive towards measurement errors?

$$(\mathbf{X}^\top \mathbf{X} + \alpha I)\mathbf{w} = \mathbf{X}^\top \mathbf{y} \qquad \alpha > 0$$

We can add a multiple of the identity matrix —> shift of the singular values!

$$\kappa = \sqrt{\frac{\sigma_1^2}{\sigma_{d+1}^2}} \quad \Rightarrow \quad \kappa = \sqrt{\frac{\sigma_1^2 + \alpha}{\sigma_{d+1}^2 + \alpha}}$$

Ill-conditioned problems

What can we do in order to be less sensitive towards measurement errors?

$$\left(\mathbf{X}^\top\mathbf{X} + \alpha I\right)\mathbf{w} = \mathbf{X}^\top\mathbf{y} \qquad \alpha > 0$$

We can add a multiple of the identity matrix —> shift of the singular values!

$$\kappa = \sqrt{\frac{\sigma_1^2}{\sigma_{d+1}^2}} \quad \Rightarrow \quad \kappa = \sqrt{\frac{\sigma_1^2 + \alpha}{\sigma_{d+1}^2 + \alpha}}$$

Example:

$$\sigma_1 = \sqrt{2}$$

$$\sigma_{d+1} = 1/\sqrt{2000000}$$

$$\alpha = 1$$

Ill-conditioned problems

What can we do in order to be less sensitive towards measurement errors?

$$\left(\mathbf{X}^\top \mathbf{X} + \alpha I\right)\mathbf{w} = \mathbf{X}^\top \mathbf{y} \qquad \alpha > 0$$

We can add a multiple of the identity matrix —> shift of the singular values!

$$\kappa = \sqrt{\frac{\sigma_1^2}{\sigma_{d+1}^2}} \quad \Rightarrow \quad \kappa = \sqrt{\frac{\sigma_1^2 + \alpha}{\sigma_{d+1}^2 + \alpha}}$$

Example:

$$\sigma_1 = \sqrt{2}$$

$$\sigma_{d+1} = 1/\sqrt{2000000} \quad \Rightarrow \quad \kappa \approx \sqrt{3} \ll 2000$$

$$\alpha = 1$$

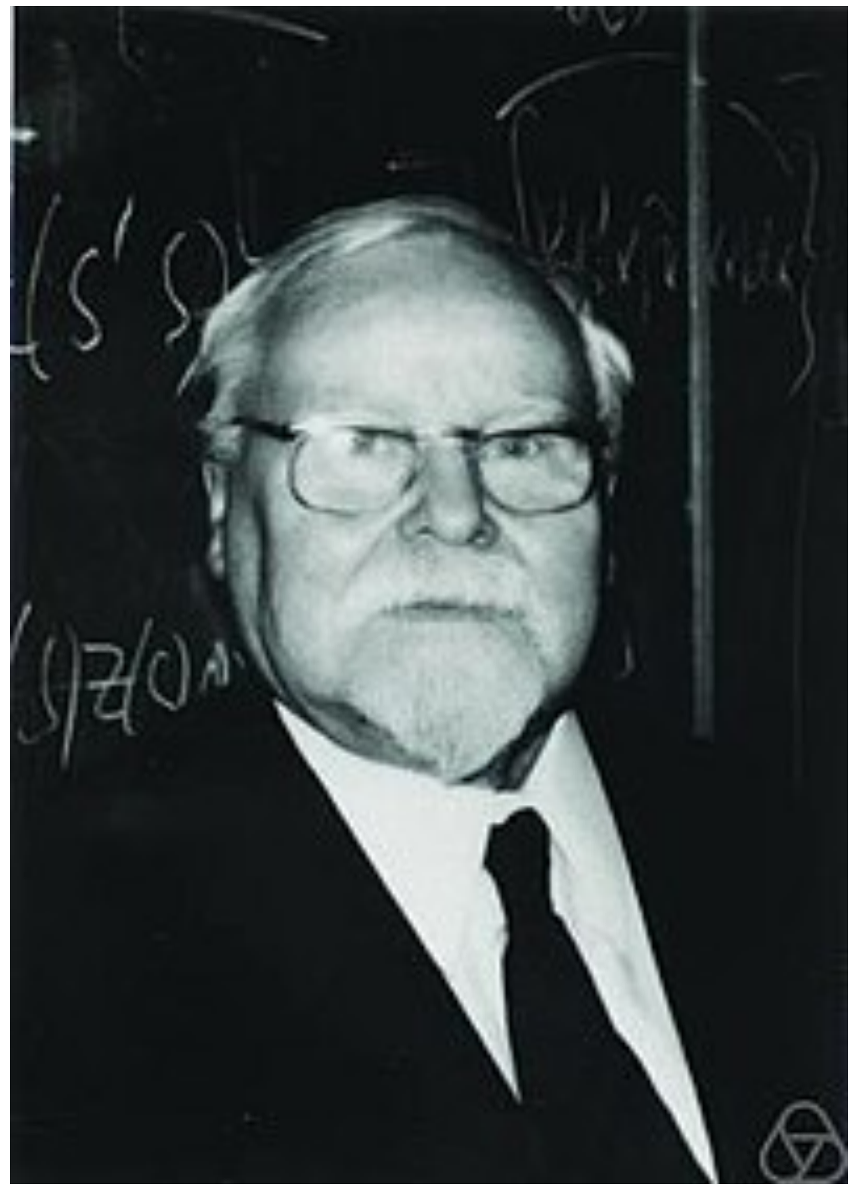
REGULARISATION

Ridge regression / Tikhonov regularisation

The minimisation problem

$$\hat{w} = \arg\min_{w} \left\{ \frac{1}{2} \|Xw - y\|^2 + \frac{\alpha}{2} \|w\|^2 \right\}$$

is also known as *Tikhonov regularisation* or *ridge regression*



Andrey Tikhonov, 1906 - 1993

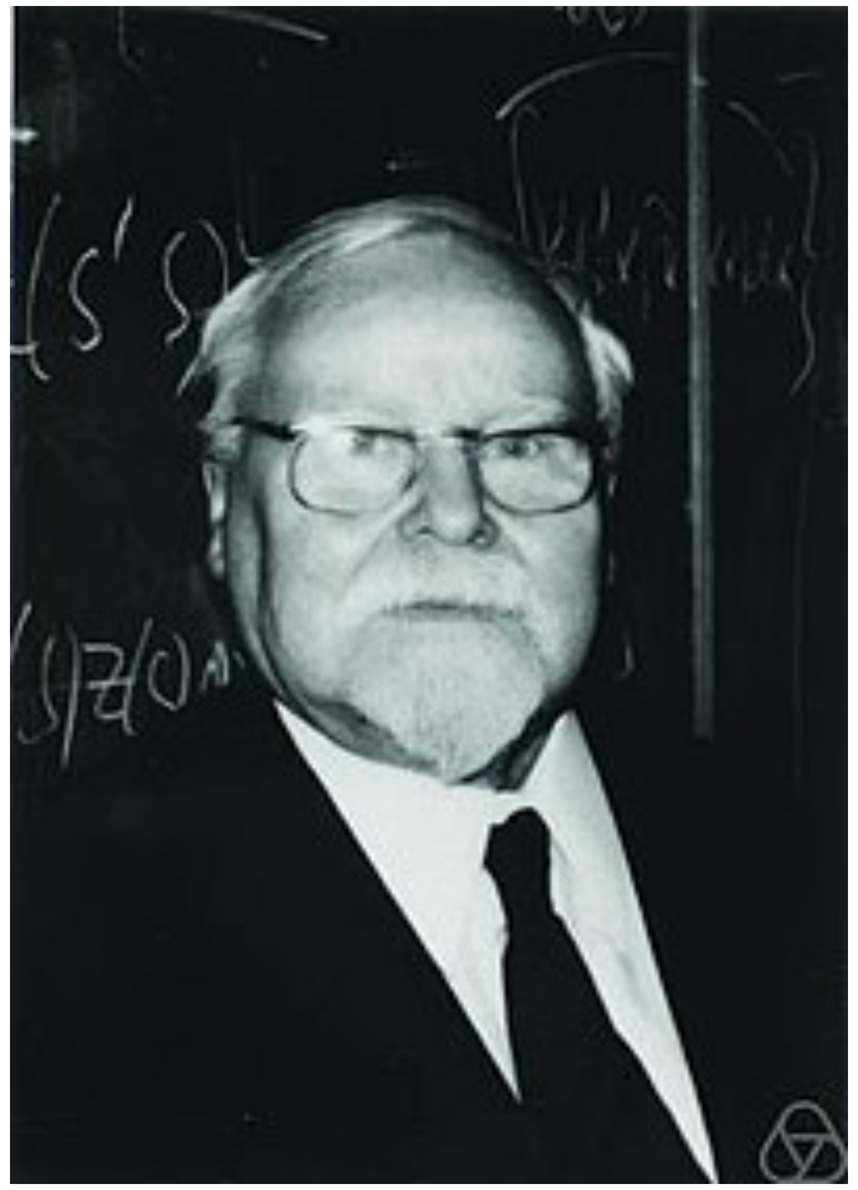
Ridge regression / Tikhonov regularisation

The minimisation problem

$$\hat{w} = \arg\min_{w} \left\{ \frac{1}{2}\|Xw - y\|^2 + \frac{\alpha}{2}\|w\|^2 \right\}$$

Standard regression term

is also known as *Tikhonov regularisation* or *ridge regression*



Andrey Tikhonov, 1906 - 1993

Ridge regression / Tikhonov regularisation

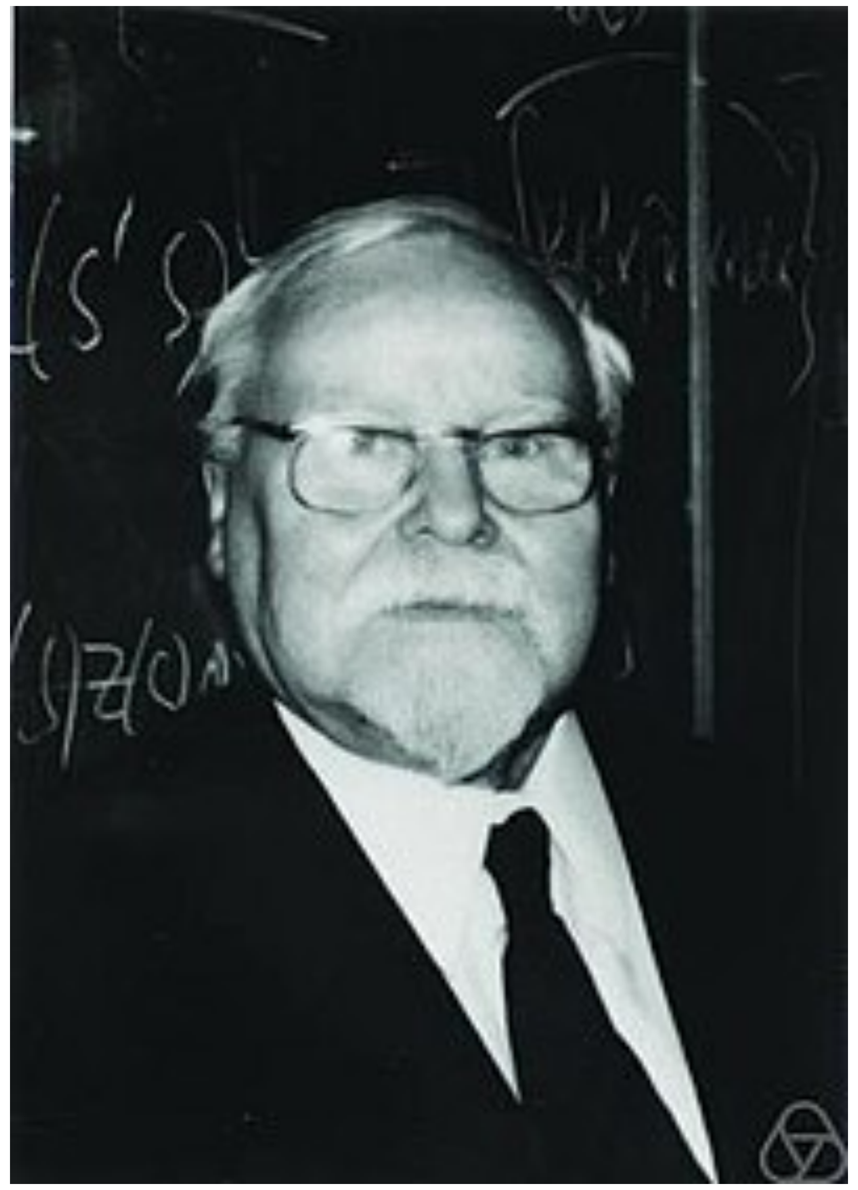
The minimisation problem

$$\hat{w} = \arg\min_{w} \left\{ \frac{1}{2}\|Xw - y\|^2 + \frac{\alpha}{2}\|w\|^2 \right\}$$

Standard regression term

Regularisation term

is also known as *Tikhonov regularisation* or *ridge regression*



Andrey Tikhonov, 1906 - 1993

Ridge regression / Tikhonov regularisation

The minimisation problem

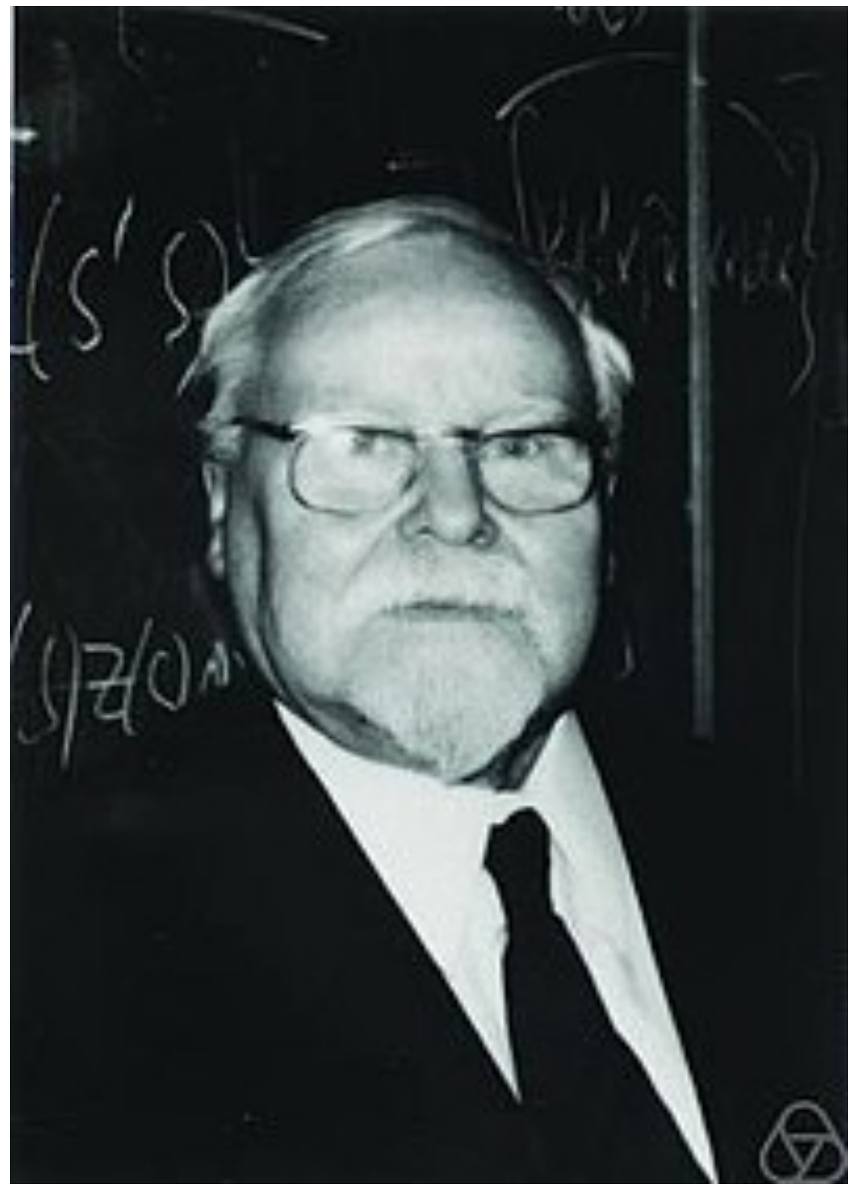
$$\hat{w} = \arg\min_{w} \left\{ \frac{1}{2}\|Xw - y\|^2 + \frac{\alpha}{2}\|w\|^2 \right\}$$

Standard regression term

Regularisation term

Regularisation parameter

is also known as *Tikhonov regularisation* or *ridge regression*



Andrey Tikhonov, 1906 - 1993

Variational regularisation

A more general form of the previous problem is variational regularisation

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \{L(\mathbf{w}) + R(\mathbf{w})\}$$

Variational regularisation

A more general form of the previous problem is variational regularisation

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \{L(\mathbf{w}) + R(\mathbf{w})\}$$

Data term/
Regression term

Variational regularisation

A more general form of the previous problem is variational regularisation

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \{ L(\mathbf{w}) + R(\mathbf{w}) \}$$

Data term/
Regression term

Regularisation
term

Variational regularisation

A more general form of the previous problem is variational regularisation

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left\{ L(\mathbf{w}) + R(\mathbf{w}) \right\}$$

Data term/
Regression term

Regularisation
term

Previous example: $L(\mathbf{w}) = \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$ $R(\mathbf{w}) = \frac{\alpha}{2}\|\mathbf{w}\|^2$

Ill-conditioned problems

Note that $$\left(\mathbf{X}^\top\mathbf{X} + \alpha I\right)\mathbf{w} = \mathbf{X}^\top\mathbf{y}$$

is equivalent to $$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \frac{\alpha}{2}\|\mathbf{w}\|^2 \right\}$$

Ill-conditioned problems

Note that $$\left(\mathbf{X}^\top\mathbf{X} + \alpha I\right)\mathbf{w} = \mathbf{X}^\top\mathbf{y}$$

is equivalent to $$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \frac{\alpha}{2}\|\mathbf{w}\|^2 \right\}$$

Proof sketch:

Ill-conditioned problems

Note that $$\left(\mathbf{X}^{\top}\mathbf{X} + \alpha I\right)\mathbf{w} = \mathbf{X}^{\top}\mathbf{y}$$

is equivalent to $$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \frac{\alpha}{2}\|\mathbf{w}\|^2 \right\}$$

Proof sketch: 1. Compute gradient of $E(w) = \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \frac{\alpha}{2}\|\mathbf{w}\|^2$, set it to zero

and show that this coincides with $(\mathbf{X}^{\top}\mathbf{X} + \alpha\mathbf{I})\mathbf{w} = \mathbf{X}^{\top}\mathbf{y}$

Ill-conditioned problems

Note that $$\left(\mathbf{X}^\top\mathbf{X} + \alpha I\right)\mathbf{w} = \mathbf{X}^\top\mathbf{y}$$

is equivalent to $$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \frac{\alpha}{2}\|\mathbf{w}\|^2 \right\}$$

Proof sketch:

1. Compute gradient of $E(w) = \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \frac{\alpha}{2}\|\mathbf{w}\|^2$, set it to zero and show that this coincides with $(\mathbf{X}^\top\mathbf{X} + \alpha\mathbf{I})\mathbf{w} = \mathbf{X}^\top\mathbf{y}$
2. Show that $E(w) = \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \frac{\alpha}{2}\|\mathbf{w}\|^2$ is convex

Good exercise!

1. Compute gradient of $E(\mathbf{w}) = \frac{1}{2}\|\mathbf{Xw} - \mathbf{y}\|^2 + \frac{\alpha}{2}\|\mathbf{w}\|^2$, set it to zero

Good exercise!

1. Compute gradient of $E(\mathbf{w}) = \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \frac{\alpha}{2}\|\mathbf{w}\|^2$, set it to zero

$$E(\mathbf{w}) = \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \frac{\alpha}{2}\|\mathbf{w}\|^2 = \frac{1}{2}\sum_i \left(\sum_j X_{ij} w_j - y_i \right)^2 + \frac{\alpha}{2}\sum_i w_i^2$$

Good exercise!

1. Compute gradient of $E(\mathbf{w}) = \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \frac{\alpha}{2}\|\mathbf{w}\|^2$, set it to zero

$$E(\mathbf{w}) = \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \frac{\alpha}{2}\|\mathbf{w}\|^2 = \frac{1}{2}\sum_i \left(\sum_j X_{ij}w_j - y_i\right)^2 + \frac{\alpha}{2}\sum_i w_i^2$$

$$\partial_{w_p} E(\mathbf{w}) = \frac{1}{2}\sum_i \partial_{w_p}\left(\sum_j X_{ij}w_j - y_i\right)^2 + \frac{\alpha}{2}\sum_i \partial_{w_p} w_i^2$$

Good exercise!

1. Compute gradient of $E(\mathbf{w}) = \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \frac{\alpha}{2}\|\mathbf{w}\|^2$, set it to zero

$$E(\mathbf{w}) = \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \frac{\alpha}{2}\|\mathbf{w}\|^2 = \frac{1}{2}\sum_i \left(\sum_j X_{ij} w_j - y_i\right)^2 + \frac{\alpha}{2}\sum_i w_i^2$$

$$\partial_{w_p} E(\mathbf{w}) = \frac{1}{2}\sum_i \partial_{w_p} \left(\sum_j X_{ij} w_j - y_i\right)^2 + \frac{\alpha}{2}\sum_i \partial_{w_p} w_i^2$$

$$\partial_{w_p} E(\mathbf{w}) = \sum_i \left(\sum_j X_{ij} w_j - y_i\right) X_{ip} + \alpha w_p$$

Good exercise!

$$\partial_{w_p} E(\mathbf{w}) = \sum_i \left(\sum_j X_{ij} w_j - y_i \right) X_{ip} + \alpha w_p$$

Good exercise!

$$\partial_{w_p} E(\mathbf{w}) = \sum_i \left(\sum_j X_{ij} w_j - y_i \right) X_{ip} + \alpha w_p$$

$$\partial_{w_p} E(\mathbf{w}) = \sum_i \left(\sum_j X_{ij} w_j - y_i \right) X^{\top}_{pi} + \alpha w_p$$

Good exercise!

$$\partial_{w_p} E(\mathbf{w}) = \sum_i \left(\sum_j X_{ij} w_j - y_i \right) X_{ip} + \alpha w_p$$

$$\partial_{w_p} E(\mathbf{w}) = \sum_i \left(\sum_j X_{ij} w_j - y_i \right) X_{pi}^\top + \alpha w_p$$

$$\partial_{w_p} E(\mathbf{w}) = \sum_i \sum_j \left(X_{pi}^\top X_{ij} w_j - X_{pi}^\top y_i \right) + \alpha w_p$$

Good exercise!

$$\partial_{w_p} E(\mathbf{w}) = \sum_i \left(\sum_j X_{ij} w_j - y_i \right) X_{ip} + \alpha w_p$$

$$\partial_{w_p} E(\mathbf{w}) = \sum_i \left(\sum_j X_{ij} w_j - y_i \right) X_{pi}^\top + \alpha w_p$$

$$\partial_{w_p} E(\mathbf{w}) = \sum_i \sum_j \left(X_{pi}^\top X_{ij} w_j - X_{pi}^\top y_i \right) + \alpha w_p$$

$$\partial_{w_p} E(\mathbf{w}) = \left(\mathbf{X}^\top \mathbf{X} \mathbf{w} - \mathbf{X}^\top \mathbf{y} + \alpha \mathbf{w} \right)_p$$

Good exercise!

$$\partial_{w_p} E(\mathbf{w}) = \left(\mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{X}^\top \mathbf{y} + \alpha \mathbf{w}\right)_p$$

Good exercise!

$$\partial_{w_p} E(\mathbf{w}) = \left(\mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{X}^\top \mathbf{y} + \alpha \mathbf{w}\right)_p$$

$$\nabla E(\mathbf{w}) = \mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{X}^\top \mathbf{y} + \alpha \mathbf{w}$$

Good exercise!

$$\partial_{w_p} E(\mathbf{w}) = \left(\mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{X}^\top \mathbf{y} + \alpha \mathbf{w}\right)_p$$

$$\nabla E(\mathbf{w}) = \mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{X}^\top \mathbf{y} + \alpha \mathbf{w}$$

$$\nabla E(\mathbf{w}) = \mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{X}^\top \mathbf{y} + \alpha \mathbf{w} = 0$$

Good exercise!

$$\partial_{w_p} E(\mathbf{w}) = \left(\mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{X}^\top \mathbf{y} + \alpha \mathbf{w}\right)_p$$

$$\nabla E(\mathbf{w}) = \mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{X}^\top \mathbf{y} + \alpha \mathbf{w}$$

$$\nabla E(\mathbf{w}) = \mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{X}^\top \mathbf{y} + \alpha \mathbf{w} = 0$$

$$\mathbf{X}^\top \mathbf{X}\mathbf{w} + \alpha \mathbf{w} = \mathbf{X}^\top \mathbf{y}$$

Maximum-Likelihood-Estimation

Given some data and some model $f(\mathbf{w})$ we can define the maximum likelihood

$$\rho(\mathbf{DATA} \mid \mathbf{w})$$

Maximum-Likelihood-Estimation

Given some data and some model $f(\mathbf{w})$ we can define the maximum likelihood

$$\rho(\mathbf{DATA} \,|\, \mathbf{w})$$

Last week we saw that finding the minimiser of the negative log likelihood is equivalent to minimise the MSE

Maximum-Likelihood-Estimation

This is very general method, to estimate (fit) some parameter of a model we can use the so-called Maximum-Likelihood-Estimation approach

Maximum-Likelihood-Estimation

This is very general method, to estimate (fit) some parameter of a model we can use the so-called Maximum-Likelihood-Estimation approach

$$\hat{\mathbf{w}}_{MLE} = \arg\max_{\mathbf{w}} \rho(\mathbf{DATA} \mid \mathbf{w})$$

Maximum-Likelihood-Estimation

This is very general method, to estimate (fit) some parameter of a model we can use the so-called Maximum-Likelihood-Estimation approach

$$\hat{\mathbf{w}}_{MLE} = \arg\max_{\mathbf{w}} \rho(\mathbf{DATA} \mid \mathbf{w})$$

Assuming that the samples are i.i.d.

Maximum-Likelihood-Estimation

This is very general method, to estimate (fit) some parameter of a model we can use the so-called Maximum-Likelihood-Estimation approach

$$\hat{\mathbf{w}}_{MLE} = \arg\max_{\mathbf{w}} \rho(\mathbf{DATA} \,|\, \mathbf{w})$$

Assuming that the samples are i.i.d.

$$\hat{\mathbf{w}}_{MLE} = \arg\max_{\mathbf{w}} \prod_{i} \rho(\mathbf{DATA_i} \,|\, \mathbf{w})$$

Maximum-Likelihood-Estimation

This is very general method, to estimate (fit) some parameter of a model we can use the so-called Maximum-Likelihood-Estimation approach

$$\hat{\mathbf{w}}_{MLE} = \arg\max_{\mathbf{w}} \rho(\mathbf{DATA} \mid \mathbf{w})$$

Assuming that the samples are i.i.d.

$$\hat{\mathbf{w}}_{MLE} = \arg\max_{\mathbf{w}} \prod_i \rho(\mathbf{DATA_i} \mid \mathbf{w}) = \arg\max_{\mathbf{w}} \log \prod_i \rho(\mathbf{DATA_i} \mid \mathbf{w})$$

Maximum-Likelihood-Estimation

This is very general method, to estimate (fit) some parameter of a model we can use the so-called Maximum-Likelihood-Estimation approach

$$\hat{\mathbf{w}}_{MLE} = \arg\max_{\mathbf{w}} \rho(\mathbf{DATA} \,|\, \mathbf{w})$$

Assuming that the samples are i.i.d.

$$\hat{\mathbf{w}}_{MLE} = \arg\max_{\mathbf{w}} \prod_i \rho(\mathbf{DATA_i} \,|\, \mathbf{w}) = \arg\max_{\mathbf{w}} \log \prod_i \rho(\mathbf{DATA_i} \,|\, \mathbf{w})$$

$$= \arg\max_{\mathbf{w}} \sum_i \rho(\mathbf{DATA_i} \,|\, \mathbf{w})$$

Maximum-A-Posteriori Estimation

Another approach is instead to maximise another quantity the A-Posteriori probability

$$\rho(\mathbf{w} \mid \mathbf{DATA})$$

Maximum-A-Posteriori Estimation

Another approach is instead to maximise another quantity the A-Posteriori probability

$$\rho(\mathbf{w} \,|\, \mathbf{DATA})$$

This is the conditional probability of the model given the data

Maximum-A-Posteriori Estimation

The optimal model is that the maximise this quantity

Maximum-A-Posteriori Estimation

The optimal model is that the maximise this quantity

$$\hat{\mathbf{w}}_{MAP} = \arg\max_{\mathbf{w}} \rho(\mathbf{w} \,|\, \mathbf{DATA})$$

Maximum-A-Posteriori Estimation

The optimal model is that the maximise this quantity

$$\hat{\mathbf{w}}_{MAP} = \arg\max_{\mathbf{w}} \rho(\mathbf{w} \,|\, \mathbf{DATA})$$

We know however that

Maximum-A-Posteriori Estimation

The optimal model is that the maximise this quantity

$$\hat{\mathbf{w}}_{MAP} = \arg\max_{\mathbf{w}} \rho(\mathbf{w} \,|\, \mathbf{DATA})$$

We know however that $$\rho(\mathbf{w} \,|\, \mathbf{DATA}) = \frac{\rho(\mathbf{DATA} \,|\, \mathbf{w})\rho(\mathbf{w})}{\rho(\mathbf{DATA})}$$

Maximum-A-Posteriori Estimation

The optimal model is that the maximise this quantity

$$\hat{\mathbf{w}}_{MAP} = \arg\max_{\mathbf{w}} \rho(\mathbf{w} \,|\, \mathbf{DATA})$$

Prior distribution

We know however that

$$\rho(\mathbf{w} \,|\, \mathbf{DATA}) = \frac{\rho(\mathbf{DATA} \,|\, \mathbf{w})\rho(\mathbf{w})}{\rho(\mathbf{DATA})}$$

Maximum-A-Posteriori Estimation

The optimal model is that the maximise this quantity

$$\hat{\mathbf{w}}_{MAP} = \arg\max_{\mathbf{w}} \rho(\mathbf{w} \,|\, \mathbf{DATA})$$

Prior distribution

We know however that $$\rho(\mathbf{w} \,|\, \mathbf{DATA}) = \frac{\rho(\mathbf{DATA} \,|\, \mathbf{w})\rho(\mathbf{w})}{\rho(\mathbf{DATA})}$$

$$\hat{\mathbf{w}}_{MAP} = \arg\max_{\mathbf{w}} \frac{\rho(\mathbf{DATA} \,|\, \mathbf{w})\rho(\mathbf{w})}{\rho(\mathbf{DATA})}$$

Maximum-A-Posteriori Estimation

The optimal model is that the maximise this quantity

$$\hat{\mathbf{w}}_{MAP} = \arg\max_{\mathbf{w}} \rho(\mathbf{w} \,|\, \mathbf{DATA})$$

Prior distribution

We know however that $$\rho(\mathbf{w} \,|\, \mathbf{DATA}) = \frac{\rho(\mathbf{DATA} \,|\, \mathbf{w})\rho(\mathbf{w})}{\rho(\mathbf{DATA})}$$

$$\hat{\mathbf{w}}_{MAP} = \arg\max_{\mathbf{w}} \frac{\rho(\mathbf{DATA} \,|\, \mathbf{w})\rho(\mathbf{w})}{\rho(\mathbf{DATA})} = \arg\max_{\mathbf{w}} \rho(\mathbf{DATA} \,|\, \mathbf{w})\rho(\mathbf{w})$$

Maximum-A-Posteriori Estimation

The optimal model is that the maximise this quantity

$$\hat{\mathbf{w}}_{MAP} = \arg\max_{\mathbf{w}} \rho(\mathbf{w}\,|\,\mathbf{DATA})$$

Prior distribution

We know however that $$\rho(\mathbf{w}\,|\,\mathbf{DATA}) = \frac{\rho(\mathbf{DATA}\,|\,\mathbf{w})\rho(\mathbf{w})}{\rho(\mathbf{DATA})}$$

$$\hat{\mathbf{w}}_{MAP} = \arg\max_{\mathbf{w}} \frac{\rho(\mathbf{DATA}\,|\,\mathbf{w})\rho(\mathbf{w})}{\rho(\mathbf{DATA})} = \arg\max_{\mathbf{w}} \rho(\mathbf{DATA}\,|\,\mathbf{w})\rho(\mathbf{w})$$

This is very similar to the MLE expression but for the prior distribution!

MLE VS MAP

$$\hat{\mathbf{w}}_{MLE} = \arg\max_{\mathbf{w}} \rho(\mathbf{DATA} \,|\, \mathbf{w})$$

$$\hat{\mathbf{w}}_{MAP} = \arg\max_{\mathbf{w}} \rho(\mathbf{w} \,|\, \mathbf{DATA}) = \arg\max_{\mathbf{w}} \rho(\mathbf{DATA} \,|\, \mathbf{w})\rho(\mathbf{w})$$

MLE VS MAP

$$\hat{\mathbf{w}}_{MLE} = \arg\max_{\mathbf{w}} \rho(\mathbf{DATA} \,|\, \mathbf{w})$$

$$\hat{\mathbf{w}}_{MAP} = \arg\max_{\mathbf{w}} \rho(\mathbf{w} \,|\, \mathbf{DATA}) = \arg\max_{\mathbf{w}} \rho(\mathbf{DATA} \,|\, \mathbf{w})\rho(\mathbf{w})$$

If the prior is "flat" the two are the equivalent, otherwise they differ due to the assumptions on the prior distribution of the parameters of the models

MLE for linear regression

Recall: maximum likelihood estimator for least-squares linear regression

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left\{ -\log \left(\rho(\mathbf{y}, \mathbf{X} \,|\, \mathbf{w}) \right) \right\}$$

MLE for linear regression

Recall: maximum likelihood estimator for least-squares linear regression

$$\begin{aligned}\hat{\mathbf{w}} &= \underset{\mathbf{w}}{\arg\min}\left\{-\log\left(\rho(\mathbf{y},\mathbf{X}\,|\,\mathbf{w})\right)\right\}\\ &= \underset{\mathbf{w}}{\arg\min}\left\{-\log\left(\rho(\mathbf{X}\,|\,\mathbf{w})\rho(\mathbf{y}\,|\,\mathbf{X},\mathbf{w})\right)\right\}\end{aligned}$$

Factoring of likelihood

MLE for linear regression

Recall: maximum likelihood estimator for least-squares linear regression

$$
\begin{aligned}
\hat{\mathbf{w}} &= \arg\min_{\mathbf{w}} \left\{ -\log\left(\rho(\mathbf{y}, \mathbf{X} \mid \mathbf{w}) \right) \right\} \\
&= \arg\min_{\mathbf{w}} \left\{ -\log\left(\rho(\mathbf{X} \mid \mathbf{w}) \rho(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) \right) \right\} && \text{Factoring of likelihood} \\
&= \arg\min_{\mathbf{w}} \left\{ -\log\left(\rho(\mathbf{X}) \rho(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) \right) \right\} && \mathbf{X} \text{ does not depend on } \mathbf{w}
\end{aligned}
$$

MLE for linear regression

Recall: maximum likelihood estimator for least-squares linear regression

$$
\begin{aligned}
\hat{\mathbf{w}} &= \arg\min_{\mathbf{w}} \left\{ -\log\left(\rho(\mathbf{y},\mathbf{X}\,|\,\mathbf{w})\right) \right\} \\
&= \arg\min_{\mathbf{w}} \left\{ -\log\left(\rho(\mathbf{X}\,|\,\mathbf{w})\rho(\mathbf{y}\,|\,\mathbf{X},\mathbf{w})\right) \right\} && \text{Factoring of likelihood} \\
&= \arg\min_{\mathbf{w}} \left\{ -\log\left(\rho(\mathbf{X})\rho(\mathbf{y}\,|\,\mathbf{X},\mathbf{w})\right) \right\} && \mathbf{X} \text{ does not depend on } \mathbf{w} \\
&= \arg\min_{\mathbf{w}} \left\{ -\log\left(\rho(\mathbf{y}\,|\,\mathbf{X},\mathbf{w})\right) \right\} && \rho(\mathbf{X}) \text{ does not depend on } \mathbf{w}
\end{aligned}
$$

MLE for linear regression

Recall: maximum likelihood estimator for least-squares linear regression

$$\begin{aligned}
\hat{\mathbf{w}} &= \arg\min_{\mathbf{w}} \left\{ -\log\left(\rho(\mathbf{y}, \mathbf{X} \,|\, \mathbf{w})\right) \right\} \\
&= \arg\min_{\mathbf{w}} \left\{ -\log\left(\rho(\mathbf{X} \,|\, \mathbf{w})\rho(\mathbf{y} \,|\, \mathbf{X}, \mathbf{w})\right) \right\} && \text{Factoring of likelihood} \\
&= \arg\min_{\mathbf{w}} \left\{ -\log\left(\rho(\mathbf{X})\rho(\mathbf{y} \,|\, \mathbf{X}, \mathbf{w})\right) \right\} && \mathbf{X} \text{ does not depend on } \mathbf{w} \\
&= \arg\min_{\mathbf{w}} \left\{ -\log\left(\rho(\mathbf{y} \,|\, \mathbf{X}, \mathbf{w})\right) \right\} && \rho(\mathbf{X}) \text{ does not depend on } \mathbf{w} \\
&= \arg\min_{\mathbf{w}} \left\{ -\log\left(\prod_{i=1}^{S} \rho(\mathbf{y}_i \,|\, \mathbf{x}_i, \mathbf{w})\right) \right\} && \text{samples are iid}
\end{aligned}$$

MLE for linear regression

Recall: maximum likelihood estimator for least-squares linear regression

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left\{ -\log \left(\prod_{i=1}^{s} \rho(\mathbf{y}_i \,|\, \mathbf{x}_i, \mathbf{w}) \right) \right\}$$

MLE for linear regression

Recall: maximum likelihood estimator for least-squares linear regression

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left\{ -\log \left(\prod_{i=1}^{s} \rho(\mathbf{y}_i \,|\, \mathbf{x}_i, \mathbf{w}) \right) \right\}$$

$$= \arg\min_{\mathbf{w}} \left\{ -\log \left(\prod_{i=1}^{s} \mathcal{N}(\mathbf{y}_i \,|\, \langle \mathbf{x}_i, \mathbf{w} \rangle, \sigma^2) \right) \right\}$$

$\mathcal{N}$ = probability density function for normal distribution

MLE for linear regression

Recall: maximum likelihood estimator for least-squares linear regression

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left\{ -\log \left(\prod_{i=1}^{s} \rho(\mathbf{y}_i \,|\, \mathbf{x}_i, \mathbf{w}) \right) \right\}$$

$$= \arg\min_{\mathbf{w}} \left\{ -\log \left(\prod_{i=1}^{s} \mathcal{N}(\mathbf{y}_i \,|\, \langle \mathbf{x}_i, \mathbf{w} \rangle, \sigma^2) \right) \right\}$$

$\mathcal{N}$ = probability density function for normal distribution

$$= \arg\min_{\mathbf{w}} \left\{ -\log \left(\prod_{i=1}^{s} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mathbf{y}_i - \langle \mathbf{x}_i, \mathbf{w} \rangle)^2}{2\sigma^2} \right) \right) \right\}$$

MLE for linear regression

Recall: maximum likelihood estimator for least-squares linear regression

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left\{ -\log \left(\prod_{i=1}^{s} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mathbf{y}_i - \langle \mathbf{x}_i, \mathbf{w} \rangle)^2}{2\sigma^2} \right) \right) \right\}$$

MLE for linear regression

Recall: maximum likelihood estimator for least-squares linear regression

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left\{ -\log \left(\prod_{i=1}^{s} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mathbf{y}_i - \langle \mathbf{x}_i, \mathbf{w} \rangle)^2}{2\sigma^2} \right) \right) \right\}$$

$$= \arg\min_{\mathbf{w}} \left\{ s \log\left(\sqrt{2\pi\sigma^2} \right) + \frac{1}{2\sigma^2} \sum_{i=1}^{s} \left(\mathbf{y}_j - \langle \mathbf{x}_i, \mathbf{w} \rangle \right)^2 \right\}$$

MLE for linear regression

Recall: maximum likelihood estimator for least-squares linear regression

$$\begin{aligned}
\hat{\mathbf{w}} &= \arg\min_{\mathbf{w}} \left\{ -\log\left(\prod_{i=1}^{s} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mathbf{y}_i - \langle \mathbf{x}_i, \mathbf{w} \rangle)^2}{2\sigma^2} \right) \right) \right\} \\
&= \arg\min_{\mathbf{w}} \left\{ s \log\left(\sqrt{2\pi\sigma^2} \right) + \frac{1}{2\sigma^2} \sum_{i=1}^{s} \left(\mathbf{y}_j - \langle \mathbf{x}_i, \mathbf{w} \rangle \right)^2 \right\} \\
&= \arg\min_{\mathbf{w}} \left\{ \frac{1}{2\sigma^2} \sum_{i=1}^{s} \left(\mathbf{y}_i - \langle \mathbf{x}_i, \mathbf{w} \rangle \right)^2 \right\}
\end{aligned}$$

MLE for linear regression

Recall: maximum likelihood estimator for least-squares linear regression

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left\{ -\log \left(\prod_{i=1}^{s} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mathbf{y}_i - \langle \mathbf{x}_i, \mathbf{w} \rangle)^2}{2\sigma^2} \right) \right) \right\}$$

$$= \arg\min_{\mathbf{w}} \left\{ s \log\left(\sqrt{2\pi\sigma^2}\right) + \frac{1}{2\sigma^2} \sum_{i=1}^{s} \left(\mathbf{y}_j - \langle \mathbf{x}_i, \mathbf{w} \rangle \right)^2 \right\}$$

$$= \arg\min_{\mathbf{w}} \left\{ \frac{1}{2\sigma^2} \sum_{i=1}^{s} \left(\mathbf{y}_i - \langle \mathbf{x}_i, \mathbf{w} \rangle \right)^2 \right\}$$

This is the MSE!

MAP for ridge regression

Maximum a-posteriori estimate for ridge regression:

$$\hat{w} = \arg\min_{w} \left\{ -\log \left(\rho(w \mid X, y) \right) \right\}$$

MAP for ridge regression

Maximum a-posteriori estimate for ridge regression:

$$\hat{w} = \arg\min_{w} \left\{ -\log \left(\rho(w \mid X, y) \right) \right\}$$

$$= \arg\min_{w} \left\{ -\log \left(\frac{\rho(y, X \mid w)\rho(w)}{\rho(y, X)} \right) \right\} \qquad \text{Bayes' Rule/Theorem}$$

MAP for ridge regression

Maximum a-posteriori estimate for ridge regression:

$$\hat{w} = \underset{w}{\arg\min} \left\{ -\log \left(\rho(w \mid X, y) \right) \right\}$$

$$= \underset{w}{\arg\min} \left\{ -\log \left(\frac{\rho(y, X \mid w)\rho(w)}{\rho(y, X)} \right) \right\}$$

Bayes' Rule/Theorem

$$= \underset{w}{\arg\min} \left\{ -\log \left(\rho(y, X \mid w)\rho(w) \right) \right\}$$

MAP for ridge regression

Maximum a-posteriori estimate for ridge regression:

$$
\begin{aligned}
\hat{w} &= \arg\min_{w} \left\{ -\log\left(\rho(w \mid X, y)\right) \right\} \\
&= \arg\min_{w} \left\{ -\log\left(\frac{\rho(y, X \mid w)\rho(w)}{\rho(y, X)}\right) \right\} \qquad \text{Bayes' Rule/Theorem} \\
&= \arg\min_{w} \left\{ -\log\left(\rho(y, X \mid w)\rho(w)\right) \right\} \\
&= \arg\min_{w} \left\{ -\log\left(\rho(y \mid X, w)\rho(w)\right) \right\}
\end{aligned}
$$

MAP for ridge regression

Maximum a-posteriori estimate for ridge regression:

$$\hat{w} = \arg\min_{w} \left\{ -\log \left(\rho(y \,|\, X, w) \rho(w) \right) \right\}$$

MAP for ridge regression

Maximum a-posteriori estimate for ridge regression:

$$
\begin{aligned}
\hat{w} &= \arg\min_{w} \left\{ -\log \left(\rho(y \,|\, X, w) \rho(w) \right) \right\} \\
&= \arg\min_{w} \left\{ -\log \left(\rho(w) \prod_{i=1}^{s} \rho(y_i \,|\, x_i, w) \right) \right\}
\end{aligned}
$$

MAP for ridge regression

Maximum a-posteriori estimate for ridge regression:

$$\begin{aligned}
\hat{w} &= \arg\min_{w} \left\{ -\log \big(\rho(y \,|\, X, w)\rho(w) \big) \right\} \\
&= \arg\min_{w} \left\{ -\log \left(\rho(w) \prod_{i=1}^{s} \rho(y_i \,|\, x_i, w) \right) \right\} \\
&= \arg\min_{w} \left\{ -\log \left[\prod_{j=1}^{d+1} \mathcal{N}\left(w_j \,\middle|\, 0, \frac{\sigma^2}{\alpha} \right) \prod_{i=1}^{s} \mathcal{N}\left(y_i \,\middle|\, \langle x_i, w \rangle, \sigma^2 \right) \right] \right\}
\end{aligned}$$

MAP for ridge regression

Maximum a-posteriori estimate for ridge regression:

$$\hat{w} = \arg\min_{w} \left\{ -\log \left(\prod_{j=1}^{d+1} \mathcal{N}\left(w_j \,\middle|\, 0, \frac{\sigma^2}{\alpha} \right) \prod_{i=1}^{s} \mathcal{N}\left(y_i \,\middle|\, \langle x_i, w \rangle, \sigma^2 \right) \right) \right\}$$

MAP for ridge regression

Maximum a-posteriori estimate for ridge regression:

$$
\begin{aligned}
\hat{w} &= \arg\min_w \left\{ -\log \left(\prod_{j=1}^{d+1} \mathcal{N}\left(w_j \,\middle|\, 0, \frac{\sigma^2}{\alpha} \right) \prod_{i=1}^{s} \mathcal{N}\left(y_i \,\middle|\, \langle x_i, w \rangle, \sigma^2 \right) \right) \right\} \\
&= \arg\min_w \left\{ -\log \left(\prod_{j=1}^{d+1} \sqrt{\frac{\alpha}{2\pi\sigma^2}} \exp\left(-\frac{\alpha}{2\sigma^2} |w_j|^2 \right) \prod_{i=1}^{s} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \langle x_i, w \rangle)^2}{2\sigma^2} \right) \right) \right\}
\end{aligned}
$$

MAP for ridge regression

Maximum a-posteriori estimate for ridge regression:

$$
\begin{aligned}
\hat{w} &= \arg\min_{w} \left\{ -\log \left(\prod_{j=1}^{d+1} \mathcal{N}\left(w_j \,\middle|\, 0, \frac{\sigma^2}{\alpha} \right) \prod_{i=1}^{s} \mathcal{N}\left(y_i \,\middle|\, \langle x_i, w \rangle, \sigma^2 \right) \right) \right\} \\
&= \arg\min_{w} \left\{ -\log \left(\prod_{j=1}^{d+1} \sqrt{\frac{\alpha}{2\pi\sigma^2}} \exp\left(-\frac{\alpha}{2\sigma^2} |w_j|^2 \right) \prod_{i=1}^{s} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \langle x_i, w \rangle)^2}{2\sigma^2} \right) \right) \right\} \\
&= \arg\min_{w} \left\{ \frac{1}{2} \sum_{i=1}^{s} \left(y_i - \langle x_i, w \rangle \right)^2 + \frac{\alpha}{2} \|w\|^2 \right\}
\end{aligned}
$$

MAP for ridge regression

Maximum a-posteriori estimate for ridge regression:

$$
\begin{aligned}
\hat{w} &= \arg\min_w \left\{ -\log \left(\prod_{j=1}^{d+1} \mathcal{N}\left(w_j \,\middle|\, 0, \frac{\sigma^2}{\alpha} \right) \prod_{i=1}^{s} \mathcal{N}\left(y_i \,\middle|\, \langle x_i, w \rangle, \sigma^2 \right) \right) \right\} \\
&= \arg\min_w \left\{ -\log \left(\prod_{j=1}^{d+1} \sqrt{\frac{\alpha}{2\pi\sigma^2}} \exp\left(-\frac{\alpha}{2\sigma^2} |w_j|^2 \right) \prod_{i=1}^{s} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \langle x_i, w \rangle)^2}{2\sigma^2} \right) \right) \right\} \\
&= \arg\min_w \left\{ \frac{1}{2} \sum_{i=1}^{s} \left(y_i - \langle x_i, w \rangle \right)^2 + \frac{\alpha}{2} \|w\|^2 \right\}
\end{aligned}
$$

This is the MSE for the ridge regression