Machine Learning with Python

MTH786U/P 2023/24

Week 2: Regression and minimisers

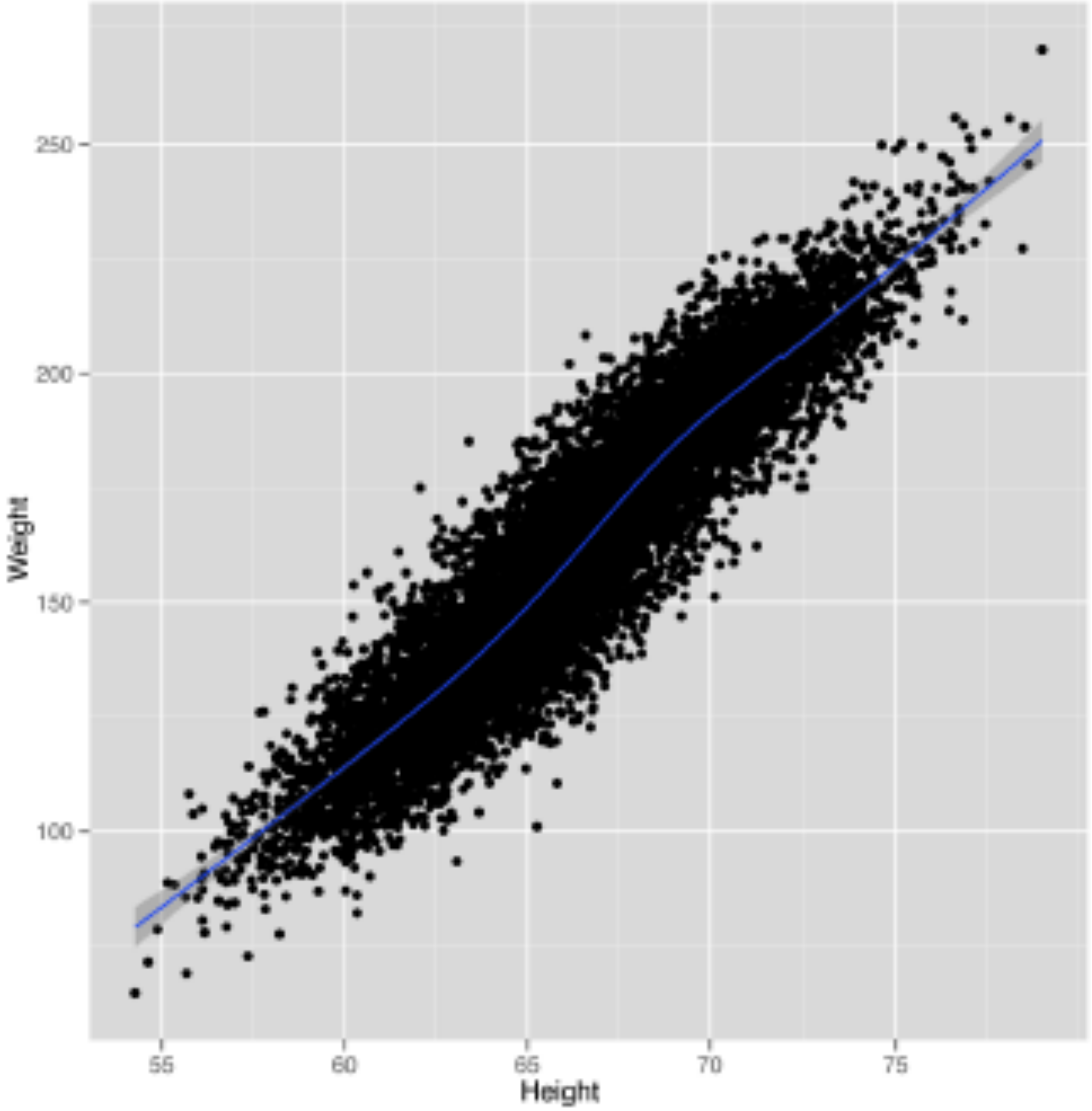
Nicola Perra, Queen Mary University of London (QMUL)

n.perra@qmul.ac.uk

LINEAR REGRESSION

What is regression?

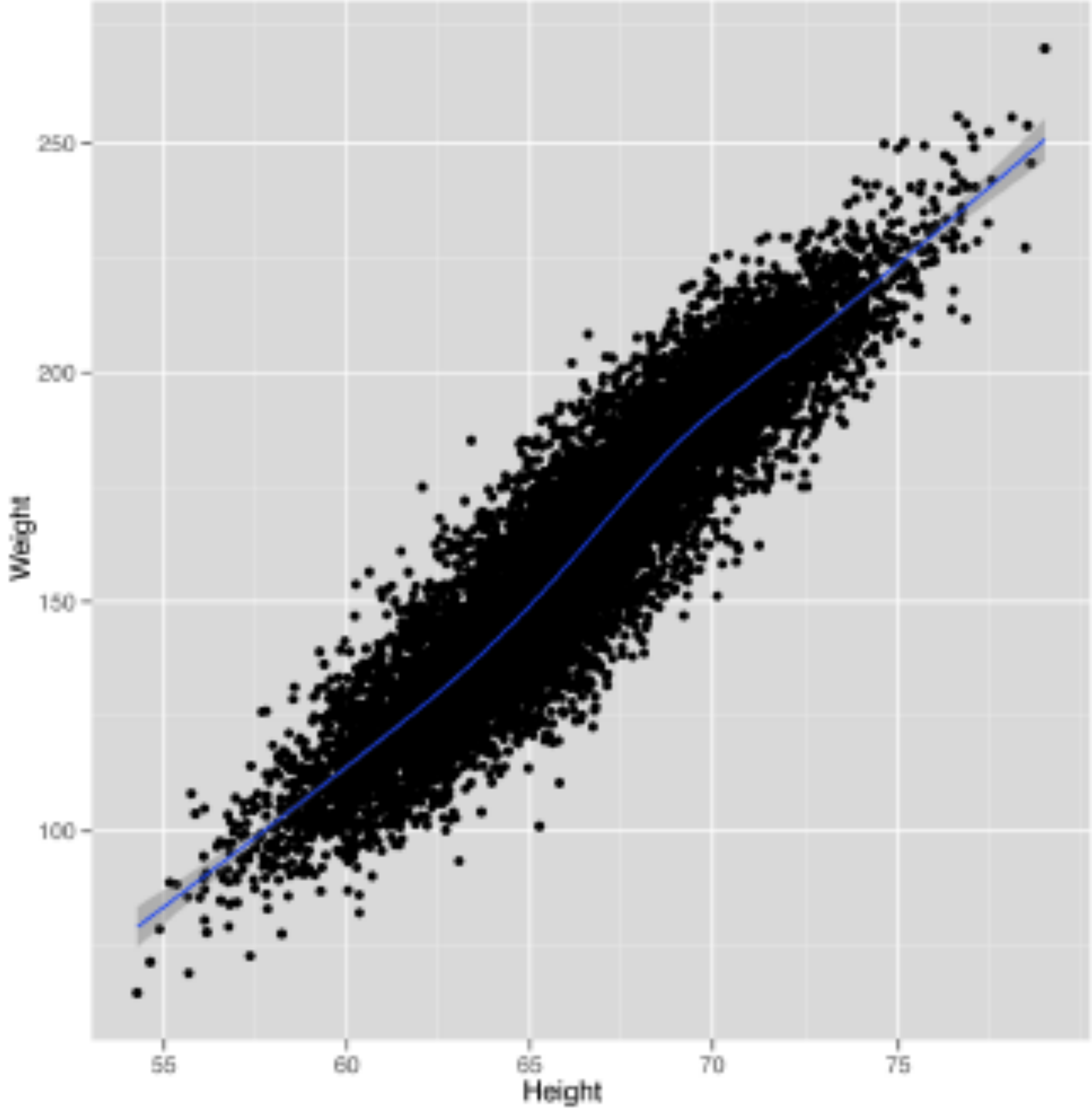
Examples:



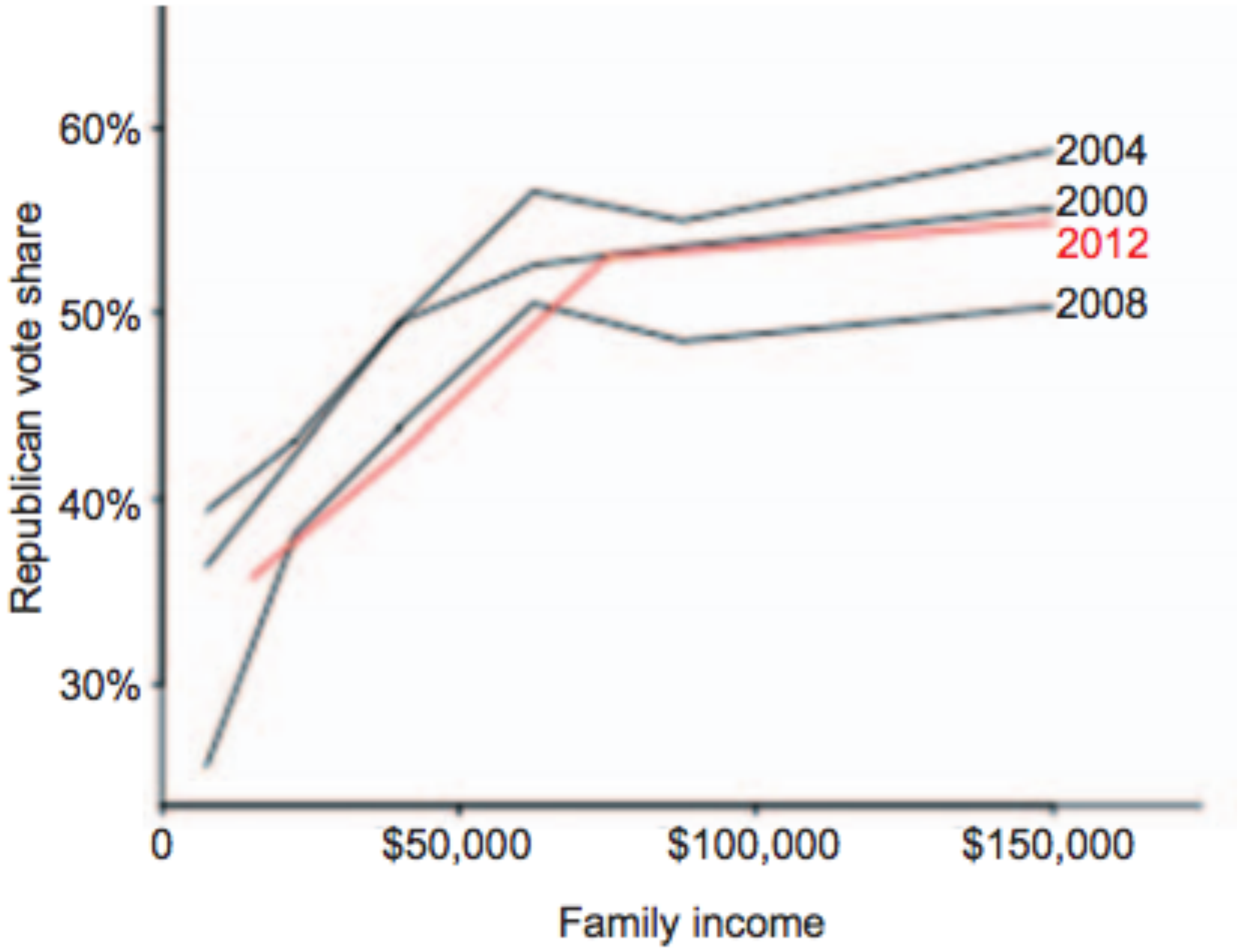
From “Machine Learning for Hackers” by
Conway & White

What is regression?

Examples:



From “Machine Learning for Hackers” by Conway & White



From Avi Feller et al. 2013

What is regression?

Mathematical formulation:

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

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

What is regression?

Mathematical formulation:

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

Example: linear regression

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

How do we parametrise f ?

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

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

Cost function

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

Cost function

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

Cost function

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

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

Cost function

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

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

Cost function

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? →

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

Cost function

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 $y_i \approx f(x_i)$?

Cost function

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

Cost function

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

Imagine $s = 3$ and $d = 2$:

Cost function

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

Imagine $s = 3$ and $d = 2$:

$$\mathbf{y} = (y_1, y_2, y_3)^\top$$

Cost function

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

Imagine $s = 3$ and $d = 2$:

$$\mathbf{y} = (y_1, y_2, y_3)^\top$$

$$\mathbf{x_1} = (x_{11}, x_{12})^\top$$

Cost function

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

Imagine $s = 3$ and $d = 2$:

$$\mathbf{y} = (y_1, y_2, y_3)^\top$$

$$\mathbf{x_1} = (x_{11}, x_{12})^\top$$

$$\mathbf{x_2} = (x_{21}, x_{22})^\top$$

Cost function

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

Imagine $s = 3$ and $d = 2$:

$$\mathbf{y} = (y_1, y_2, y_3)^\top$$

$$\mathbf{x_1} = (x_{11}, x_{12})^\top$$

$$\mathbf{x_2} = (x_{21}, x_{22})^\top$$

$$\mathbf{x_3} = (x_{31}, x_{32})^\top$$

Cost function

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

Imagine $s = 3$ and $d = 2$:

$$\mathbf{y} = (y_1, y_2, y_3)^\top$$

$$\mathbf{x_1} = (x_{11}, x_{12})^\top$$
$$\mathbf{x_2} = (x_{21}, x_{22})^\top$$
$$\mathbf{x_3} = (x_{31}, x_{32})^\top$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \end{pmatrix}$$

Cost function

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

Imagine $s = 3$ and $d = 2$:

$$\mathbf{y} = (y_1, y_2, y_3)^\top$$

$$\mathbf{x_1} = (x_{11}, x_{12})^\top$$

$$\mathbf{x_2} = (x_{21}, x_{22})^\top$$

$$\mathbf{x_3} = (x_{31}, x_{32})^\top$$

$$\mathbf{X} = \left.\begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \end{pmatrix}\right\downarrow s$$

Cost function

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

Imagine $s = 3$ and $d = 2$:

$$\mathbf{y} = (y_1, y_2, y_3)^\top$$

$$\mathbf{x_1} = (x_{11}, x_{12})^\top$$

$$\mathbf{x_2} = (x_{21}, x_{22})^\top$$

$$\mathbf{x_3} = (x_{31}, x_{32})^\top$$

$$\mathbf{X} = \begin{array}{cc} \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \end{pmatrix} & \Big\downarrow s \\ \xrightarrow[d+1]{\qquad\qquad} & \end{array}$$

Cost function

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

Imagine $s = 3$ and $d = 2$:

$$\mathbf{y} = (y_1, y_2, y_3)^\top$$

$$\mathbf{x_1} = (x_{11}, x_{12})^\top$$

$$\mathbf{x_2} = (x_{21}, x_{22})^\top$$

$$\mathbf{x_3} = (x_{31}, x_{32})^\top$$

$$\mathbf{X} = \underset{\xrightarrow[d+1]{\hspace{6em}}}{\begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \end{pmatrix}} \Big\downarrow s$$

$$\in \mathbb{R}^{s \times (d+1)}$$

Cost function

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

Imagine $s = 3$ and $d = 2$:

$$\mathbf{y} = (y_1, y_2, y_3)^\top$$

$$\mathbf{x_1} = (x_{11}, x_{12})^\top$$
$$\mathbf{x_2} = (x_{21}, x_{22})^\top$$
$$\mathbf{x_3} = (x_{31}, x_{32})^\top$$

$$\mathbf{X} = \underset{\xrightarrow[d+1]{\hspace{6em}}}{\begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \end{pmatrix}} \Big\downarrow s$$

$$\in \mathbb{R}^{s\times(d+1)}$$

Cost function

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

Imagine $s = 3$ and $d = 2$:

$$\mathbf{y} = (y_1, y_2, y_3)^\top$$

$$\mathbf{x_1} = (x_{11}, x_{12})^\top$$

$$\mathbf{x_2} = (x_{21}, x_{22})^\top$$

$$\mathbf{x_3} = (x_{31}, x_{32})^\top$$

$$\mathbf{X} = \underset{\xrightarrow{\qquad d+1 \qquad}}{\begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \end{pmatrix}} \Big\downarrow s$$

$$\in \mathbb{R}^{s \times (d+1)}$$

$$\mathbf{w} = (w_0, w_1, \ldots, w_d)^\top$$

Cost function

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

Imagine $s = 3$ and $d = 2$:

$$\mathbf{y} = (y_1, y_2, y_3)^\top$$

$$\mathbf{x_1} = (x_{11}, x_{12})^\top$$

$$\mathbf{x_2} = (x_{21}, x_{22})^\top$$

$$\mathbf{x_3} = (x_{31}, x_{32})^\top$$

$$\mathbf{X} = \underset{\xrightarrow[d+1]{\qquad\qquad}}{\begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \end{pmatrix}} \Big\downarrow s$$

$$\in \mathbb{R}^{s \times (d+1)}$$

$$\mathbf{w} = (w_0, w_1, \ldots, w_d)^\top$$

$$\in \mathbb{R}^{d+1}$$

Cost function

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

Imagine $s = 3$ and $d = 2$:

Cost function

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

Imagine $s = 3$ and $d = 2$:

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

Cost function

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

Imagine $s = 3$ and $d = 2$:

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

$$\mathbf{y} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix} \rightarrow$$

Cost function

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

Imagine $s = 3$ and $d = 2$:

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

$$\mathbf{y} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix} \rightarrow \begin{array}{c} w_0 + x_{11}w_1 + x_{12}w_2 = y_1 \\ w_0 + x_{21}w_1 + x_{22}w_2 = y_2 \\ w_0 + x_{31}w_1 + x_{32}w_2 = y_3 \end{array}$$

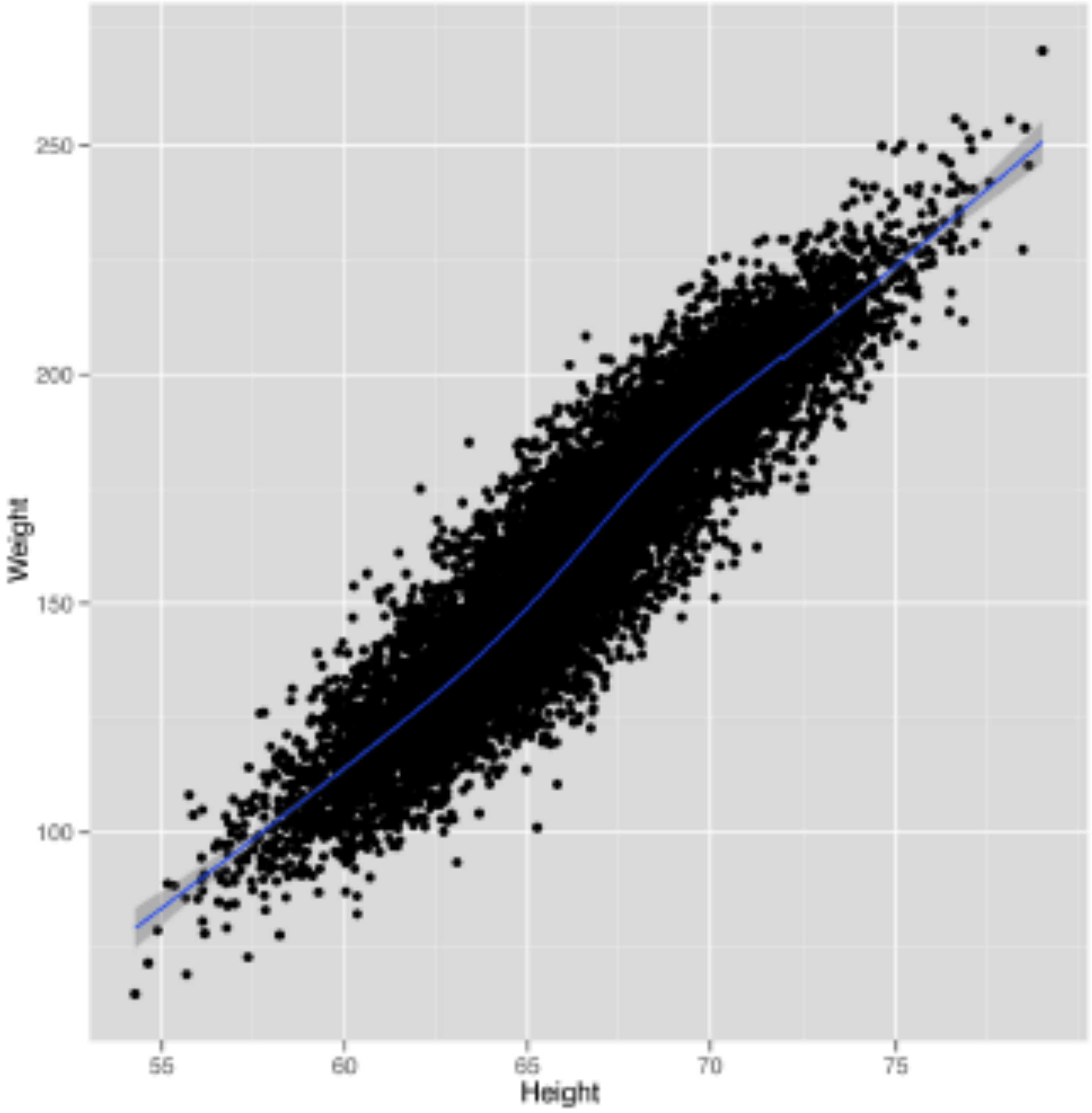
The system of linear equations has a unique solution if…?

The system of linear equations has a unique solution if…?

But is it realistic to assume $s = d + 1$?

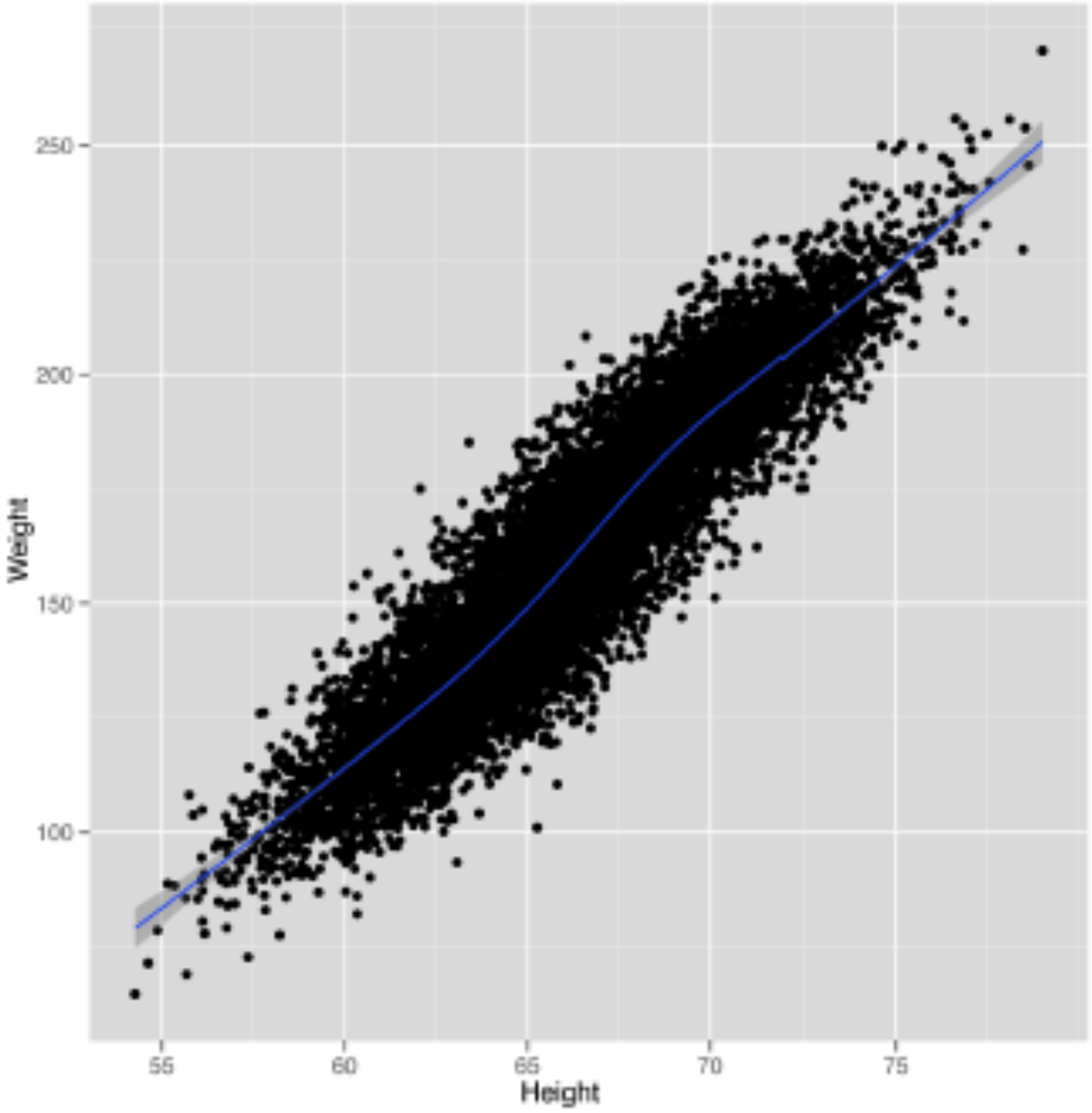
The system of linear equations has a unique solution if…?

But is it realistic to assume $s = d + 1$?



The system of linear equations has a unique solution if…?

But is it realistic to assume $s = d + 1$?



$$s \gg d + 1 = 2$$

Instead we need to find an approximation that is optimal in some sense

Instead we need to find an approximation that is optimal in some sense

Example: Mean-Square Error (MSE)

$$\mathrm{MSE}(\mathbf{w}) := \frac{1}{2s}\sum_{i=1}^{s} |f(\mathbf{x_i}) - y_i|^2$$

Instead we need to find an approximation that is optimal in some sense

Example: Mean-Square Error (MSE)

$$\mathrm{MSE}(\mathbf{w}) := \frac{1}{2s}\sum_{i=1}^{s} |f(\mathbf{x_i}) - y_i|^2$$

Obtain ‘optimal’ parameters $\hat{\mathbf{w}}$ by minimising MSE:

Instead we need to find an approximation that is optimal in some sense

Example: Mean-Square Error (MSE)

$$\mathsf{MSE}(\mathbf{w}) := \frac{1}{2s}\sum_{i=1}^{s} |f(\mathbf{x_i}) - y_i|^2$$

Obtain 'optimal' parameters $\hat{\mathbf{w}}$ by minimising MSE:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathsf{MSE}(\mathbf{w})$$

Instead we need to find an approximation that is optimal in some sense

Example: Mean-Square Error (MSE)

$$\mathrm{MSE}(\mathbf{w}) := \frac{1}{2s} \sum_{i=1}^{s} |f(\mathbf{x_i}) - y_i|^2$$

Obtain 'optimal' parameters $\hat{\mathbf{w}}$ by minimising MSE:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathrm{MSE}(\mathbf{w})$$

How can we do this?

Few remarks

$$\text{MSE(def 1)}(\mathbf{w}) := \frac{1}{2s}\sum_{i=1}^{s} |f(\mathbf{x_i}) - y_i|^2$$

Few remarks

$$\text{MSE(def 1)}(\mathbf{w}) := \frac{1}{2s}\sum_{i=1}^{s} |f(\mathbf{x_i}) - y_i|^2 \qquad \text{MSE(def 2)}(\mathbf{w}) := \frac{1}{s}\sum_{i=1}^{s} |f(\mathbf{x_i}) - y_i|^2$$

Few remarks

$$\text{MSE(def 1)}(\mathbf{w}) := \frac{1}{2s}\sum_{i=1}^{s}|f(\mathbf{x_i}) - y_i|^2 \qquad \text{MSE(def 2)}(\mathbf{w}) := \frac{1}{s}\sum_{i=1}^{s}|f(\mathbf{x_i}) - y_i|^2$$

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \text{MSE(def 1)}(\mathbf{w}) = \arg\min_{\mathbf{w}} \text{MSE(def 2)}(\mathbf{w})$$

Few remarks

$$\text{MSE(def 1)}(\mathbf{w}) := \frac{1}{2s}\sum_{i=1}^{s} |f(\mathbf{x_i}) - y_i|^2 \qquad \text{MSE(def 2)}(\mathbf{w}) := \frac{1}{s}\sum_{i=1}^{s} |f(\mathbf{x_i}) - y_i|^2$$

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \text{MSE(def 1)}(\mathbf{w}) = \arg\min_{\mathbf{w}} \text{MSE(def 2)}(\mathbf{w})$$

To find the arg min, we do not care really for the value of MSE(w), we seek the arguments ws that minimize it! So any constant of ws does not affect the search!

How do we compute $\hat{\mathbf{w}}$?

How do we compute $\hat{\mathbf{w}}$?

Example: $f(\mathbf{x_i}) = w_0 \qquad \forall i \in \{1,\ldots,s\}, \ \ d = 0$

How do we compute $\hat{\mathbf{w}}$?

Example: $f(\mathbf{x_i}) = w_0 \qquad \forall i \in \{1,\ldots,s\},\ \ d = 0$

MSE cost function: $\mathrm{MSE}(w_0) := \frac{1}{2s}\sum_{i=1}^{s} | w_0 - y_i |^2$

How do we compute $\hat{\mathbf{w}}$?

Example: $f(\mathbf{x_i}) = w_0 \qquad \forall i \in \{1,\ldots,s\}, \;\; d = 0$

MSE cost function: $$\text{MSE}(w_0) := \frac{1}{2s}\sum_{i=1}^{s} |\, w_0 - y_i\,|^2$$

We do what we did in school, we compute the derivative and set it to zero: $$\nabla \text{MSE}(\hat{w}_0) = \text{MSE}'(\hat{w}_0) = \frac{1}{s}\sum_{i=1}^{s}(\hat{w}_0 - y_i) \overset{!}{=} 0$$

How do we compute $\hat{\mathbf{w}}$?

Example: $f(\mathbf{x_i}) = w_0 \qquad \forall i \in \{1,\ldots,s\},\ \ d = 0$

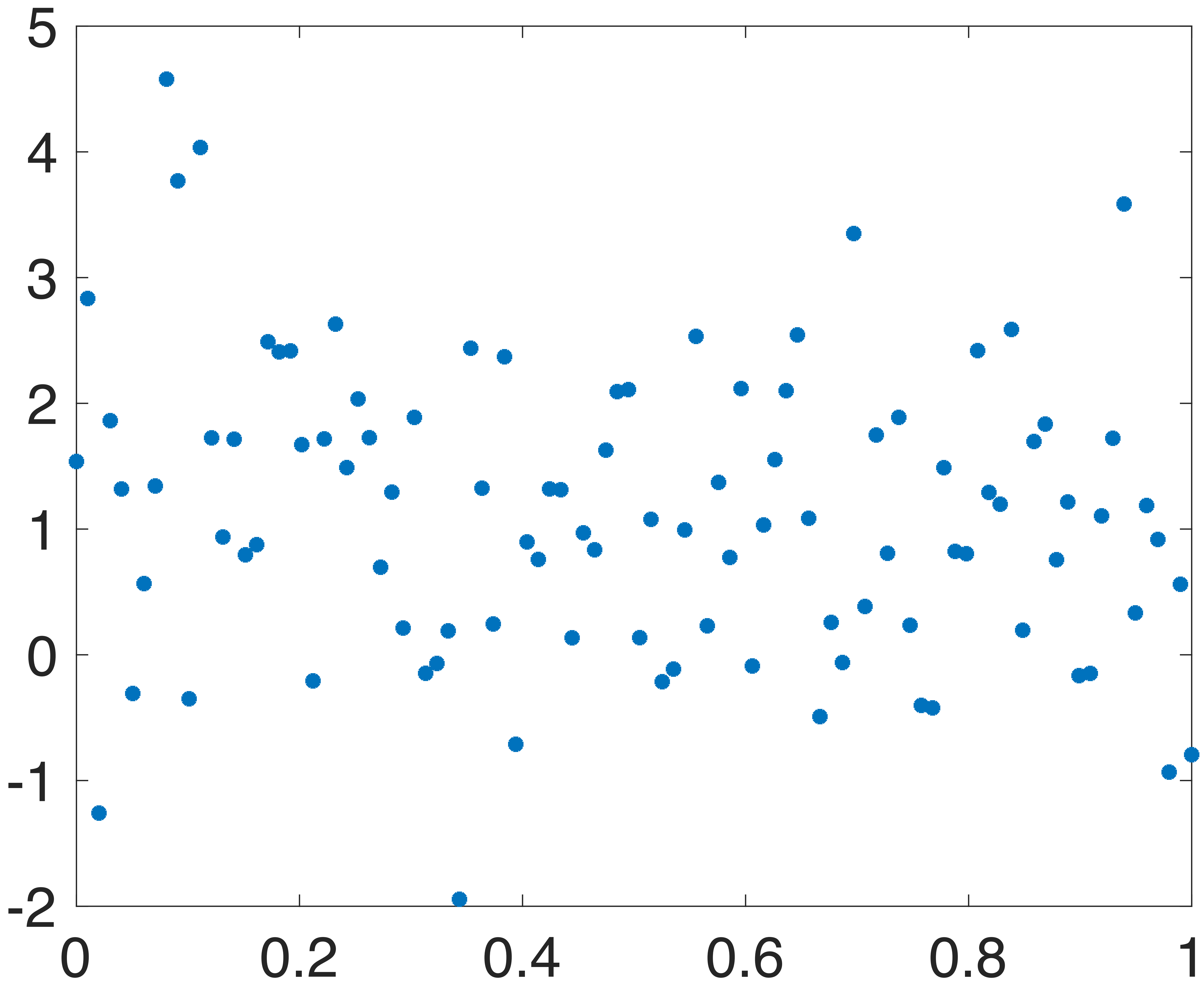
MSE cost function: $\text{MSE}(w_0) := \frac{1}{2s}\sum_{i=1}^{s} |w_0 - y_i|^2$

We do what we did in school, we compute the derivative and set it to zero:

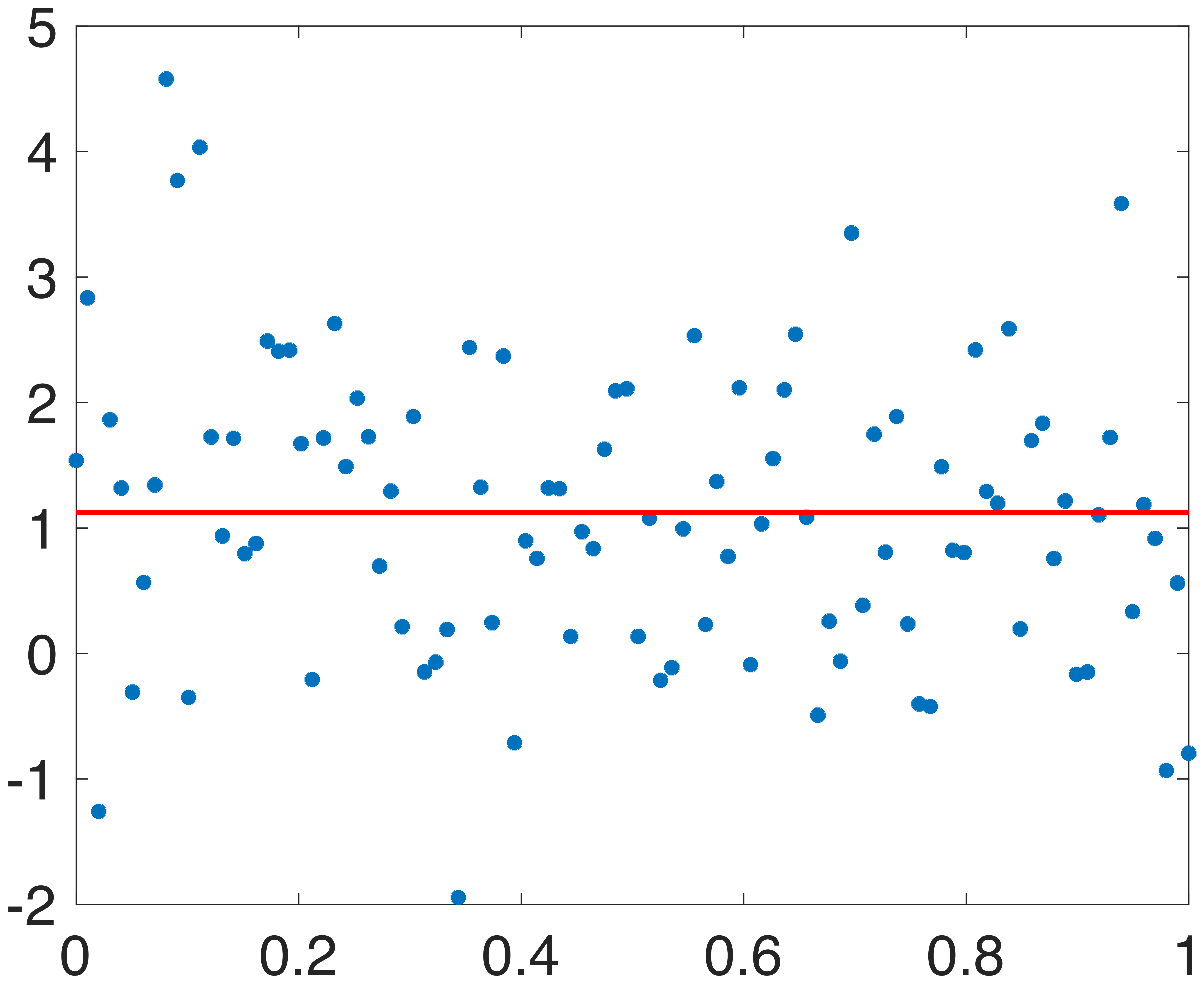
$$\nabla \text{MSE}(\hat{w}_0) = \text{MSE}^{'}(\hat{w}_0) = \frac{1}{s}\sum_{i=1}^{s}(\hat{w}_0 - y_i) \overset{!}{=} 0$$

$$\Rightarrow \qquad \hat{w}_0 = \frac{1}{s}\sum_{i=1}^{s} y_i$$

Example:

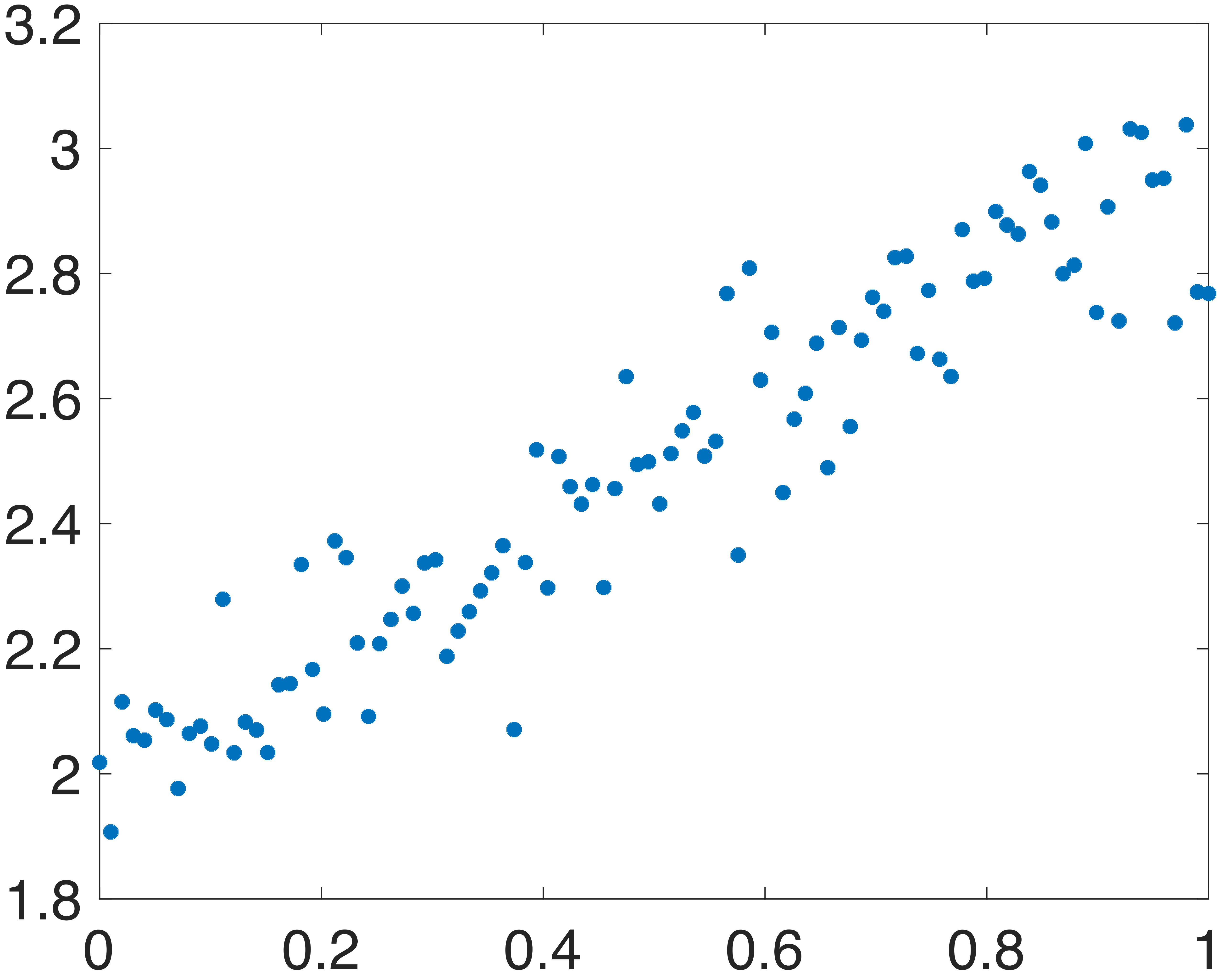


Example:

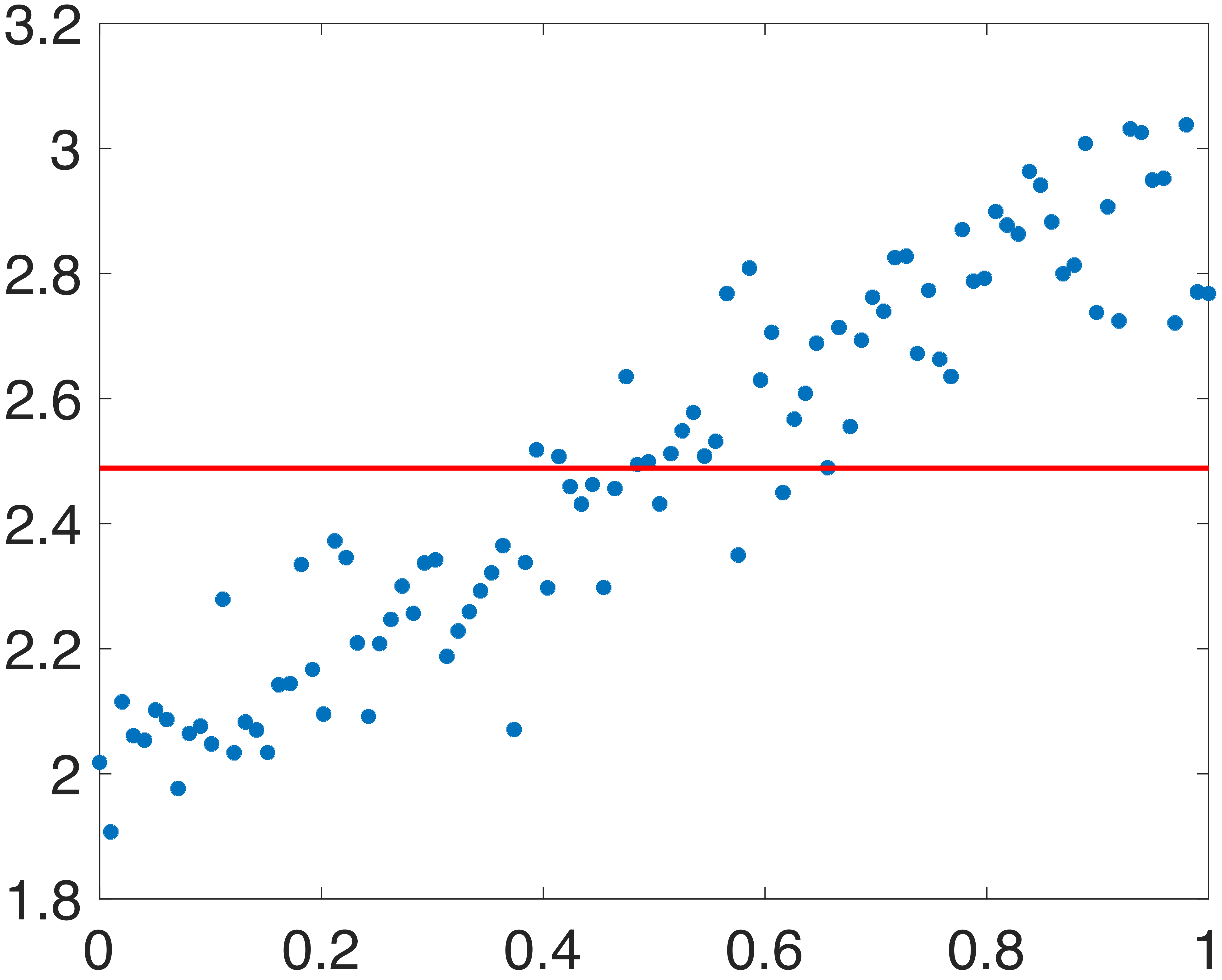


$\hat{w}_0 \approx 1.1231$

Example:



Example:



$\hat{w}_0 \approx 2.4889$

A slightly more complicated example:

$$f(x_i) = w_0 + w_1 x_i \qquad \forall i \in \{1, \ldots, s\}, \;\; d = 1$$

A slightly more complicated example:

$$f(x_i) = w_0 + w_1 x_i \qquad \forall i \in \{1,\ldots,s\},\ \ d = 1$$

MSE cost function: $$\text{MSE}(w_0, w_1) := \frac{1}{2s}\sum_{i=1}^{s} |\, w_0 + w_1 x_i - y_i \,|^2$$

A slightly more complicated example:

$$f(x_i) = w_0 + w_1 x_i \qquad \forall i \in \{1,\ldots,s\},\ \ d = 1$$

MSE cost function: $$\text{MSE}(w_0, w_1) := \frac{1}{2s}\sum_{i=1}^{s} | w_0 + w_1 x_i - y_i |^2$$

$$\Rightarrow \quad \nabla \text{MSE} = \begin{pmatrix} \partial_{w_0} MSE(\mathbf{w}) \\ \partial_{w_1} MSE(\mathbf{w}) \end{pmatrix}$$

A slightly more complicated example:

$$f(x_i) = w_0 + w_1 x_i \qquad \forall i \in \{1,\ldots,s\},\ \ d = 1$$

MSE cost function: $$\text{MSE}(w_0, w_1) := \frac{1}{2s}\sum_{i=1}^{s} | w_0 + w_1 x_i - y_i |^2$$

$$\Rightarrow \ \ \nabla\text{MSE} = \begin{pmatrix} \partial_{w_0} MSE(\mathbf{w}) \\ \partial_{w_1} MSE(\mathbf{w}) \end{pmatrix} \quad \Rightarrow \quad \nabla\text{MSE} = \frac{1}{s}\begin{pmatrix} \sum_{i=1}^{s} (w_0 + w_1 x_i - y_i) \\ \sum_{i=1}^{s} (w_0 + w_1 x_i - y_i) x_i \end{pmatrix}$$

$$\nabla \text{MSE} = \frac{1}{s}\begin{pmatrix} \sum_{i=1}^{s} (w_0 + w_1 x_i - y_i) \\ \sum_{i=1}^{s} (w_0 + w_1 x_i - y_i) x_i \end{pmatrix} \overset{!}{=} \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \Rightarrow$$

$$\nabla \text{MSE} = \frac{1}{s}\begin{pmatrix} \sum_{i=1}^{s}(w_0 + w_1 x_i - y_i) \\ \sum_{i=1}^{s}(w_0 + w_1 x_i - y_i)x_i \end{pmatrix} \overset{!}{=} \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \Rightarrow \quad \begin{gathered} \hat{w}_0 + \bar{x}\hat{w}_1 = \bar{y} \\ \bar{x}\hat{w}_0 + \frac{\|x\|^2}{s}\hat{w}_1 = \frac{\langle y, x\rangle}{s} \end{gathered}$$

$$\nabla \text{MSE} = \frac{1}{s}\begin{pmatrix} \sum_{i=1}^{s}(w_0 + w_1 x_i - y_i) \\ \sum_{i=1}^{s}(w_0 + w_1 x_i - y_i)x_i \end{pmatrix} \overset{!}{=} \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \Rightarrow \quad \begin{aligned} \hat{w}_0 + \bar{x}\hat{w}_1 &= \bar{y} \\ \bar{x}\hat{w}_0 + \frac{\|x\|^2}{s}\hat{w}_1 &= \frac{\langle y, x\rangle}{s} \end{aligned}$$

$$\Rightarrow \quad \begin{aligned} \hat{w}_0 &= \frac{\bar{y}\|x\|^2 - \bar{x}\langle x, y\rangle}{\|x\|^2 - s\bar{x}^2} \\ \hat{w}_1 &= \frac{\langle x, y\rangle - s\bar{x}\bar{y}}{\|x\|^2 - s\bar{x}^2} \end{aligned} \quad \text{for} \quad \|x\|^2 \neq s\bar{x}^2$$

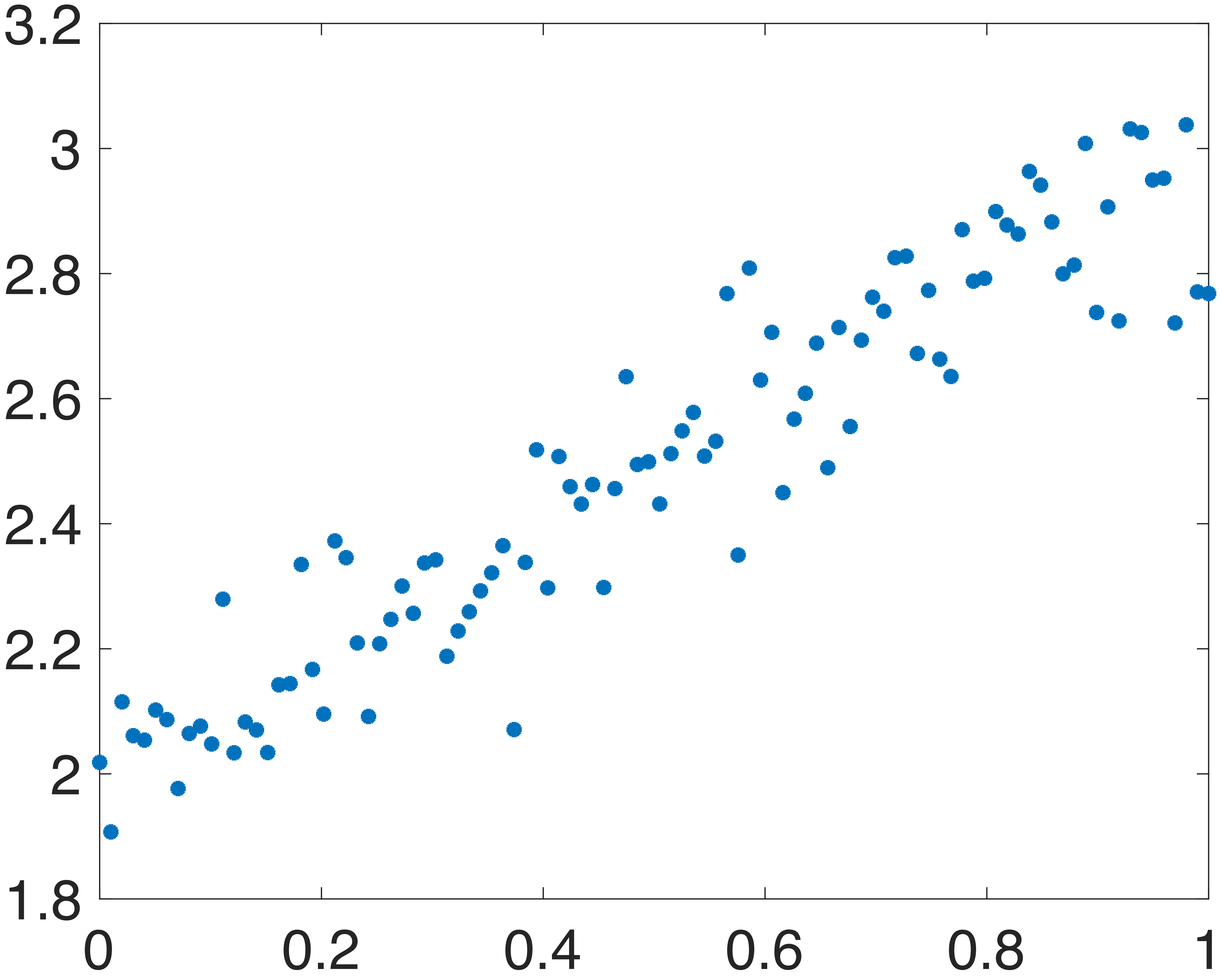
$$\nabla \mathrm{MSE} = \frac{1}{s}\begin{pmatrix} \sum_{i=1}^{s}(w_0 + w_1 x_i - y_i) \\ \sum_{i=1}^{s}(w_0 + w_1 x_i - y_i)x_i \end{pmatrix} \overset{!}{=} \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \Rightarrow \quad \begin{aligned} \hat{w}_0 + \bar{x}\hat{w}_1 &= \bar{y} \\ \bar{x}\hat{w}_0 + \frac{\|x\|^2}{s}\hat{w}_1 &= \frac{\langle y, x\rangle}{s} \end{aligned}$$

$$\Rightarrow \quad \begin{aligned} \hat{w}_0 &= \frac{\bar{y}\|x\|^2 - \bar{x}\langle x, y\rangle}{\|x\|^2 - s\bar{x}^2} \\ \hat{w}_1 &= \frac{\langle x, y\rangle - s\bar{x}\bar{y}}{\|x\|^2 - s\bar{x}^2} \end{aligned} \quad \text{for} \quad \|x\|^2 \neq s\bar{x}^2$$

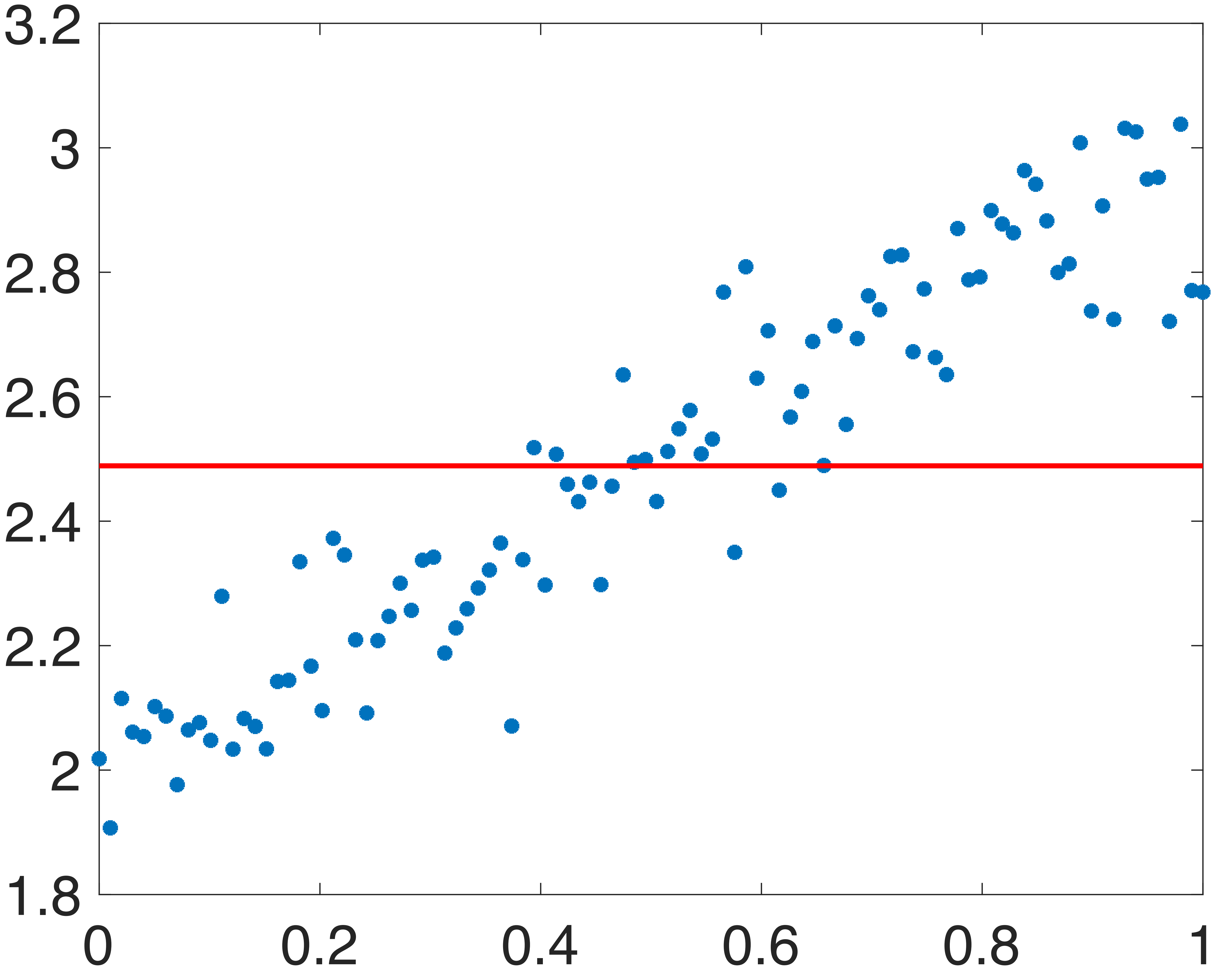
$$\text{for} \quad \bar{x} := \frac{1}{s}\sum_{j=1}^{s} x_j$$

$$\text{and} \quad \bar{y} := \frac{1}{s}\sum_{j=1}^{s} y_j$$

Example:

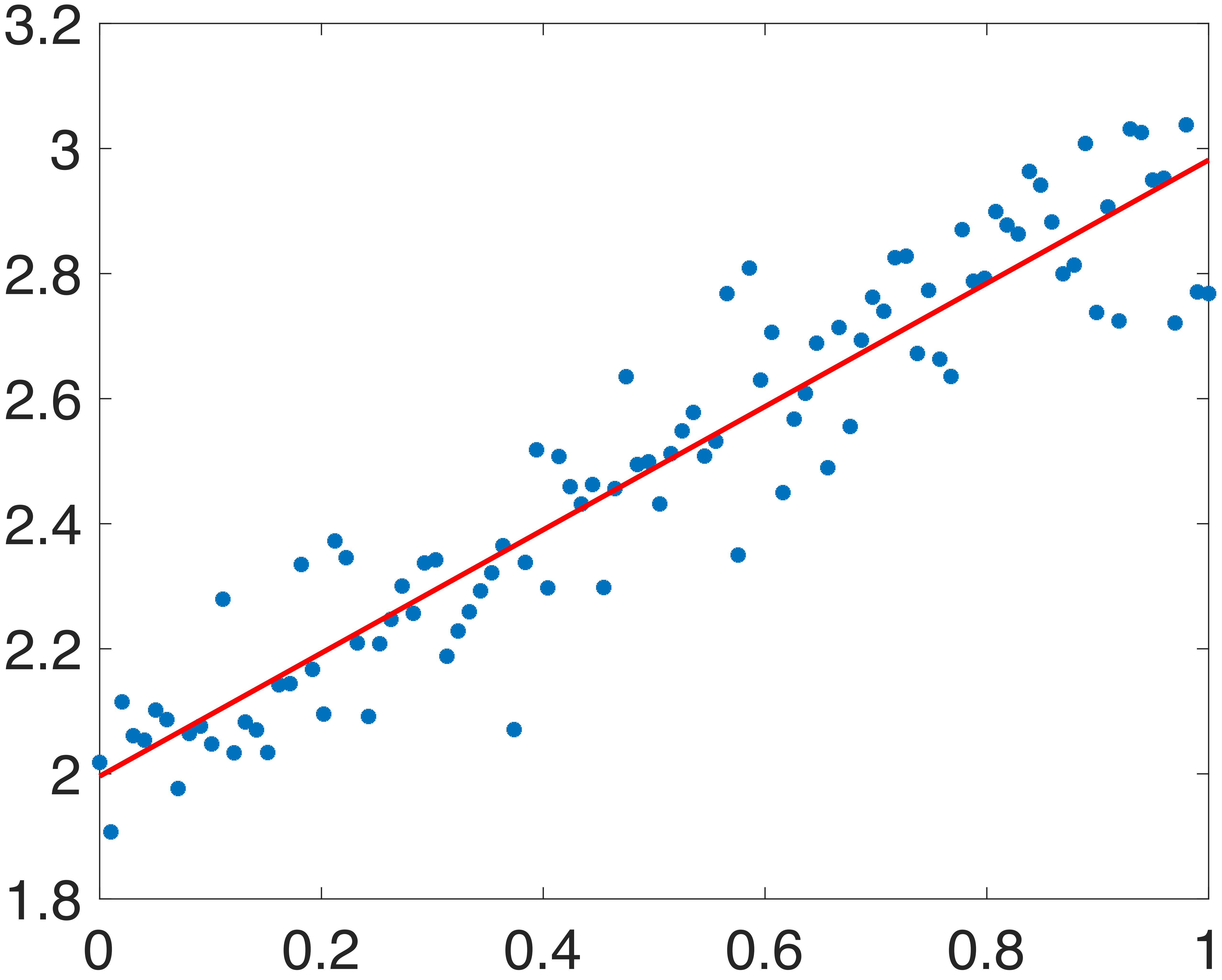


Example:



$\hat{w}_0 \approx 2.4889$

Example:



$$\hat{w}_0 \approx 1.9962$$

$$\hat{w}_1 \approx 0.9854$$

$$f(x_i) = w_0 + w_1 x_i \approx y_i \qquad \forall i \in \{1,\ldots,s\} \qquad \Leftrightarrow \qquad \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_s \end{pmatrix}}_{=:\mathbf{X}} \underbrace{\begin{pmatrix} w_0 \\ w_1 \end{pmatrix}}_{=:\mathbf{w}} \approx \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_s \end{pmatrix}}_{=:\mathbf{y}}$$

$$f(x_i) = w_0 + w_1 x_i \approx y_i \qquad \forall i \in \{1,\dots,s\} \qquad \Leftrightarrow \qquad \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_s \end{pmatrix}}_{=:\mathbf{X}} \underbrace{\begin{pmatrix} w_0 \\ w_1 \end{pmatrix}}_{=:\mathbf{w}} \approx \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_s \end{pmatrix}}_{=:\mathbf{y}}$$

More in general?

$$f(x_i) = w_0 + w_1 x_i \approx y_i \qquad \forall i \in \{1,\ldots,s\} \qquad \Leftrightarrow \qquad \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_s \end{pmatrix}}_{=:\mathbf{X}} \underbrace{\begin{pmatrix} w_0 \\ w_1 \end{pmatrix}}_{=:\mathbf{w}} \approx \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_s \end{pmatrix}}_{=:\mathbf{y}}$$

More in general?

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

$$f(x_i) = w_0 + w_1 x_i \approx y_i \qquad \forall i \in \{1,\ldots,s\} \qquad \Leftrightarrow \qquad \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_s \end{pmatrix}}_{=:\mathbf{X}} \underbrace{\begin{pmatrix} w_0 \\ w_1 \end{pmatrix}}_{=:\mathbf{w}} \approx \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_s \end{pmatrix}}_{=:\mathbf{y}}$$

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

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

$$f(x_i) = w_0 + w_1 x_i \approx y_i \qquad \forall i \in \{1,\ldots,s\} \qquad \Leftrightarrow \qquad \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_s \end{pmatrix}}_{=:\mathbf{X}} \underbrace{\begin{pmatrix} w_0 \\ w_1 \end{pmatrix}}_{=:\mathbf{w}} \approx \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_s \end{pmatrix}}_{=:\mathbf{y}}$$

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

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

$$f(x_i) = w_0 + w_1 x_i \approx y_i \qquad \forall i \in \{1,\ldots,s\} \qquad \Leftrightarrow \qquad \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_s \end{pmatrix}}_{=:\mathbf{X}} \underbrace{\begin{pmatrix} w_0 \\ w_1 \end{pmatrix}}_{=:\mathbf{w}} \approx \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_s \end{pmatrix}}_{=:\mathbf{y}}$$

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

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

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

$$f(x_i) = w_0 + w_1 x_i \approx y_i \qquad \forall i \in \{1,\ldots,s\} \qquad \Leftrightarrow \qquad \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_s \end{pmatrix}}_{=:\mathbf{X}} \underbrace{\begin{pmatrix} w_0 \\ w_1 \end{pmatrix}}_{=:\mathbf{w}} \approx \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_s \end{pmatrix}}_{=:\mathbf{y}}$$

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

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

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

$$f(x_i) = w_0 + w_1 x_i \approx y_i \qquad \forall i \in \{1,\ldots,s\} \qquad \Leftrightarrow \qquad \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_s \end{pmatrix}}_{=:\mathbf{X}} \underbrace{\begin{pmatrix} w_0 \\ w_1 \end{pmatrix}}_{=:\mathbf{w}} \approx \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_s \end{pmatrix}}_{=:\mathbf{y}}$$

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

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

Try to prove this!

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

What about other cost functions?

Mean absolute error: $$\mathrm{MAE}(\mathbf{w}) := \frac{1}{s}\sum_{i=1}^{s}\left|(\mathbf{X}\mathbf{w})_i - y_i\right|$$

What about other cost functions?

Mean absolute error: $$\mathrm{MAE}(\mathbf{w}) := \frac{1}{s}\sum_{i=1}^{s}\left|(\mathbf{X}\mathbf{w})_i - y_i\right|$$

- More robust to outliers

What about other cost functions?

Mean absolute error: $\text{MAE}(\mathbf{w}) := \frac{1}{s}\sum_{i=1}^{s}\left|(\mathbf{X}\mathbf{w})_i - y_i\right|$

- More robust to outliers
- Not differentiable —> more difficult to compute minimiser

A statistical motivation

Why did we come up with the least squares function in order to fit our model function to the data?

A statistical motivation

Why did we come up with the least squares function in order to fit our model function to the data?

Choice was basically arbitrary until now!

A statistical motivation

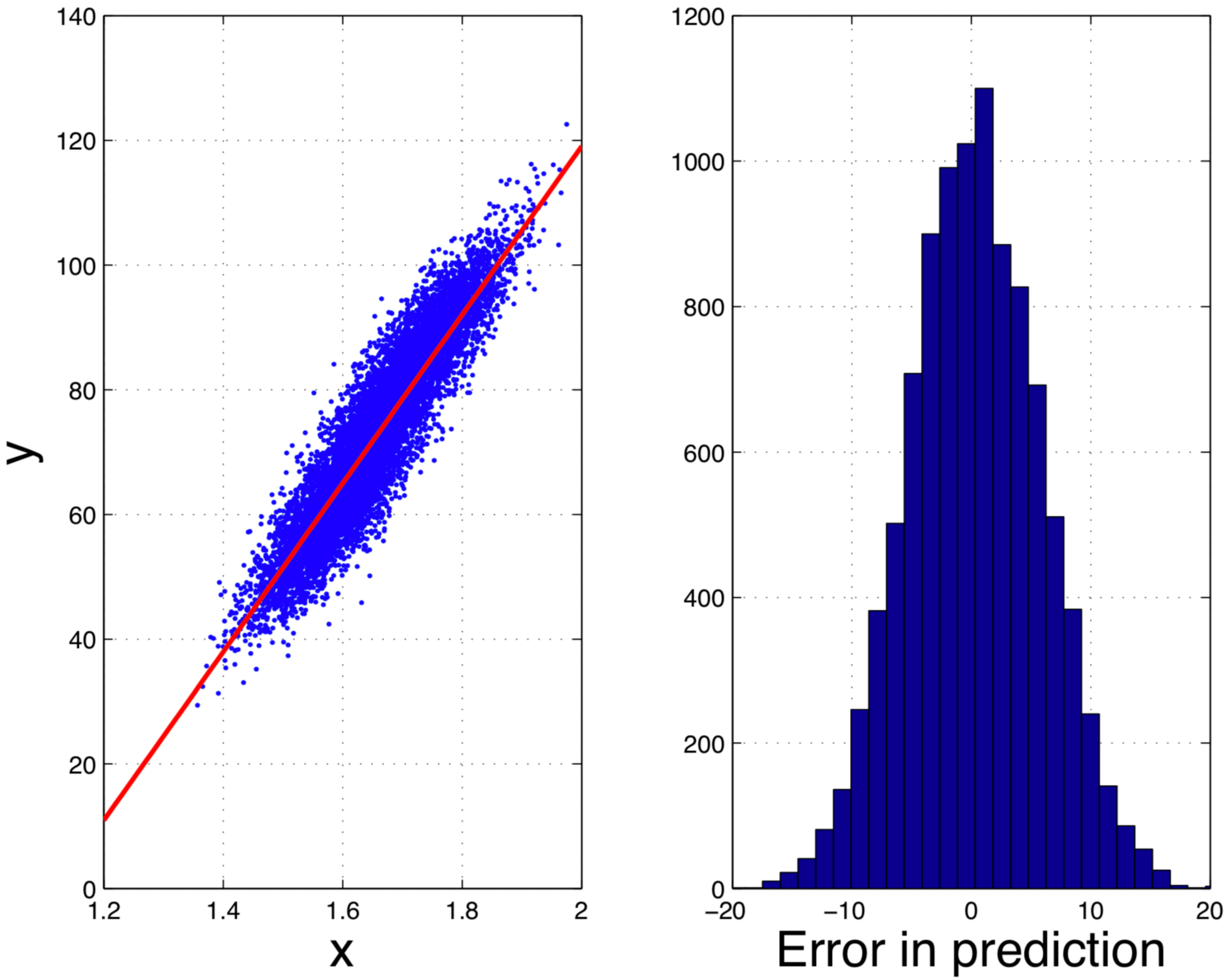
Statistical motivation: we can write

$$y_i = \langle \mathbf{x_i}, \mathbf{w} \rangle + \varepsilon_i$$

Or:

$$\epsilon_i = y_i - \langle \mathbf{x_i}, \mathbf{w} \rangle$$

A statistical motivation



A statistical motivation

Observation: ε_i is an instance of a normal-distributed random variable with mean zero and variance σ^2

Probability density function

$$\rho(\varepsilon_i \,|\, 0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\varepsilon_i^2}{2\sigma^2}}$$

A statistical motivation

Probability density function

$$\rho(\varepsilon_i | 0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\varepsilon_i^2}{2\sigma^2}}$$

A statistical motivation

Probability density function

$$\rho(\varepsilon_i\,|\,0,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\varepsilon_i^2}{2\sigma^2}}$$

Assumption: all ε_i's are i.i.d., i.e.

$$\rho(\varepsilon_i,\varepsilon_j\,|\,0,\sigma^2) = \rho(\varepsilon_i\,|\,0,\sigma^2)\,\rho(\varepsilon_j\,|\,0,\sigma^2) \qquad \text{for } i \neq j.$$

A statistical motivation

$$\rho(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_s \,|\, 0, \sigma^2) = (2\pi\sigma^2)^{-\frac{s}{2}} \prod_{i=1}^{s} e^{-\frac{\varepsilon_i^2}{2\sigma^2}}$$

A statistical motivation

$$\rho(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_s \,|\, 0, \sigma^2) = (2\pi\sigma^2)^{-\frac{s}{2}} \prod_{i=1}^{s} e^{-\frac{\varepsilon_i^2}{2\sigma^2}} = (2\pi\sigma^2)^{-\frac{s}{2}} \prod_{i=1}^{s} e^{-\frac{(y_i - \langle \mathbf{x_i}, \mathbf{w} \rangle)^2}{2\sigma^2}}$$

A statistical motivation

$$\rho(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_s \,|\, 0, \sigma^2) = (2\pi\sigma^2)^{-\frac{s}{2}} \prod_{i=1}^{s} e^{-\frac{\varepsilon_i^2}{2\sigma^2}} = (2\pi\sigma^2)^{-\frac{s}{2}} \prod_{i=1}^{s} e^{-\frac{(y_i - \langle \mathbf{x_i}, \mathbf{w} \rangle)^2}{2\sigma^2}}$$

$$= \rho(y_1, \ldots, y_s \,|\, \langle \mathbf{x_1}, \mathbf{w} \rangle, \ldots, \langle \mathbf{x_s}, \mathbf{w} \rangle, \sigma^2)$$

A statistical motivation

Statistical motivation: $\varepsilon_i = y_i - \langle \mathbf{x_i}, \mathbf{w} \rangle$

A statistical motivation

Statistical motivation: $\varepsilon_i = y_i - \langle \mathbf{x_i}, \mathbf{w} \rangle$

Choose parameters $\mathbf{w} = \hat{\mathbf{w}}$ such that they maximise the likelihood $\rho(y \,|\, \mathbf{X}\mathbf{w}, \sigma^2)$, for

$\mathbf{y} := (y_1, \ldots, y_s)^\top$ and $\mathbf{X} := \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1(d+1)} \\ x_{21} & \ddots & & \vdots \\ \vdots & & & \\ x_{s1} & \ldots & & x_{s(d+1)} \end{pmatrix}.$

A statistical motivation

Statistical motivation: $\varepsilon_i = y_i - \langle \mathbf{x_i}, \mathbf{w} \rangle$

Choose parameters $\mathbf{w} = \hat{\mathbf{w}}$ such that they maximise the likelihood $\rho(y \,|\, \mathbf{X}\mathbf{w}, \sigma^2)$, for

$\mathbf{y} := (y_1, \ldots, y_s)^\top$ and $\mathbf{X} := \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1(d+1)} \\ x_{21} & \ddots & & \vdots \\ \vdots & & & \\ x_{s1} & \ldots & & x_{s(d+1)} \end{pmatrix}.$

Alternative: choose $\hat{\mathbf{w}}$ such that it minimises the negative log-likelihood, i.e.

A statistical motivation

Statistical motivation: $\varepsilon_i = y_i - \langle \mathbf{x_i}, \mathbf{w} \rangle$

Choose parameters $\mathbf{w} = \hat{\mathbf{w}}$ such that they maximise the likelihood $\rho(y \,|\, \mathbf{X}\mathbf{w}, \sigma^2)$, for

$\mathbf{y} := (y_1, \ldots, y_s)^\top$ and $\mathbf{X} := \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1(d+1)} \\ x_{21} & \ddots & & \vdots \\ \vdots & & & \\ x_{s1} & \ldots & & x_{s(d+1)} \end{pmatrix}.$

Alternative: choose $\hat{\mathbf{w}}$ such that it minimises the negative log-likelihood, i.e.

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^{d+1}} \{-\log(\rho(\mathbf{y} \,|\, \mathbf{X}\mathbf{w}, \sigma^2))\}$$

A statistical motivation

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^{d+1}} \{-\log(\rho(\mathbf{y} \,|\, \mathbf{X}\mathbf{w}, \sigma^2))\}$$

A statistical motivation

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \{-\log(\rho(\mathbf{y}\,|\,\mathbf{X}\mathbf{w},\sigma^2))\}$$

$$= \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{-\log\left(\prod_{i=1}^{s}\rho(y_i\,|\,\langle\mathbf{x_i},\mathbf{w}\rangle,\sigma^2)\right)\right\}$$

A statistical motivation

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}\in\mathbb{R}^{d+1}} \{-\log(\rho(\mathbf{y}\,|\,\mathbf{X}\mathbf{w}, \sigma^2))\}$$

$$= \arg \min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{-\log\left(\prod_{i=1}^{s}\rho(y_i\,|\,\langle\mathbf{x_i}, \mathbf{w}\rangle, \sigma^2)\right)\right\}$$

$$= \arg \min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{-\sum_{i=1}^{s}\log\left(\rho(y_i\,|\,\langle\mathbf{x_i}, \mathbf{w}\rangle, \sigma^2)\right)\right\}$$

A statistical motivation

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \{-\log(\rho(\mathbf{y}\,|\,\mathbf{X}\mathbf{w},\sigma^2))\}$$

$$= \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{-\log\left(\prod_{i=1}^{s}\rho(y_i\,|\,\langle\mathbf{x_i},\mathbf{w}\rangle,\sigma^2)\right)\right\}$$

$$= \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{-\sum_{i=1}^{s}\log\left(\rho(y_i\,|\,\langle\mathbf{x_i},\mathbf{w}\rangle,\sigma^2)\right)\right\}$$

$$= \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{\frac{1}{2\sigma^2}\sum_{i=1}^{s}(y_i-\langle\mathbf{x_i},\mathbf{w}\rangle)^2+\text{const}\right\}$$

A statistical motivation

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \{-\log(\rho(\mathbf{y}\,|\,\mathbf{X}\mathbf{w},\sigma^2))\}$$

$$= \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{-\log\left(\prod_{i=1}^{s}\rho(y_i\,|\,\langle\mathbf{x_i},\mathbf{w}\rangle,\sigma^2)\right)\right\}$$

$$= \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{-\sum_{i=1}^{s}\log\left(\rho(y_i\,|\,\langle\mathbf{x_i},\mathbf{w}\rangle,\sigma^2)\right)\right\}$$

$$= \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{\frac{1}{2\sigma^2}\sum_{i=1}^{s}(y_i-\langle\mathbf{x_i},\mathbf{w}\rangle)^2 + \text{const}\right\}$$

$$\rho(y_i\,|\,\langle\mathbf{x_i},\mathbf{w}\rangle,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y_i-\langle\mathbf{x_i},\mathbf{w}\rangle)^2}{2\sigma^2}}$$

A statistical motivation

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^{d+1}} \left\{ \frac{1}{2\sigma^2} \sum_{i=1}^{s} (y_i - \langle \mathbf{x_i}, \mathbf{w} \rangle)^2 + \text{const} \right\}$$

A statistical motivation

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{2\sigma^2} \sum_{i=1}^{s} (y_i - \langle \mathbf{x_i}, \mathbf{w} \rangle)^2 + \text{const} \right\}$$

MSE function:

$$\text{MSE}(\mathbf{w}) = \frac{1}{2s} \sum_{i=1}^{s} (y_i - \langle \mathbf{x_i}, \mathbf{w} \rangle)^2$$

A statistical motivation

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{2\sigma^2} \sum_{i=1}^{s} (y_i - \langle \mathbf{x_i}, \mathbf{w} \rangle)^2 + \text{const} \right\}$$

MSE function:

$$\text{MSE}(\mathbf{w}) = \frac{1}{2s} \sum_{i=1}^{s} (y_i - \langle \mathbf{x_i}, \mathbf{w} \rangle)^2 \quad \Rightarrow \quad \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \{-\log(\rho(\mathbf{y} \,|\, \mathbf{X}\mathbf{w}, \sigma^2)\}$$

$$= \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \text{MSE}(\mathbf{w})$$

Regression revisited

Models can be *too limited* or *too rich*:

Regression revisited

Models can be *too limited* or *too rich*:

Too limited —> we cannot find a function that is a good fit to our data

Regression revisited

Models can be *too limited* or *too rich*:

Too limited —> we cannot find a function that is a good fit to our data

Too rich —> we find a function that fits the data too well

Regression revisited

Models can be *too limited* or *too rich*:

Too limited —> we cannot find a function that is a good fit to our data

Too rich —> we find a function that fits the data too well

Too limited —> function is *underfitting* the data

Too rich —> function is *overfitting* the data

Regression revisited

Models can be *too limited* or *too rich*:

Too limited —> we cannot find a function that is a good fit to our data

Too rich —> we find a function that fits the data too well

Too limited —> function is *underfitting* the data

Too rich —> function is *overfitting* the data

Both are issues, and difficult to address in practice, as we do not know what part of the data is signal and what is noise

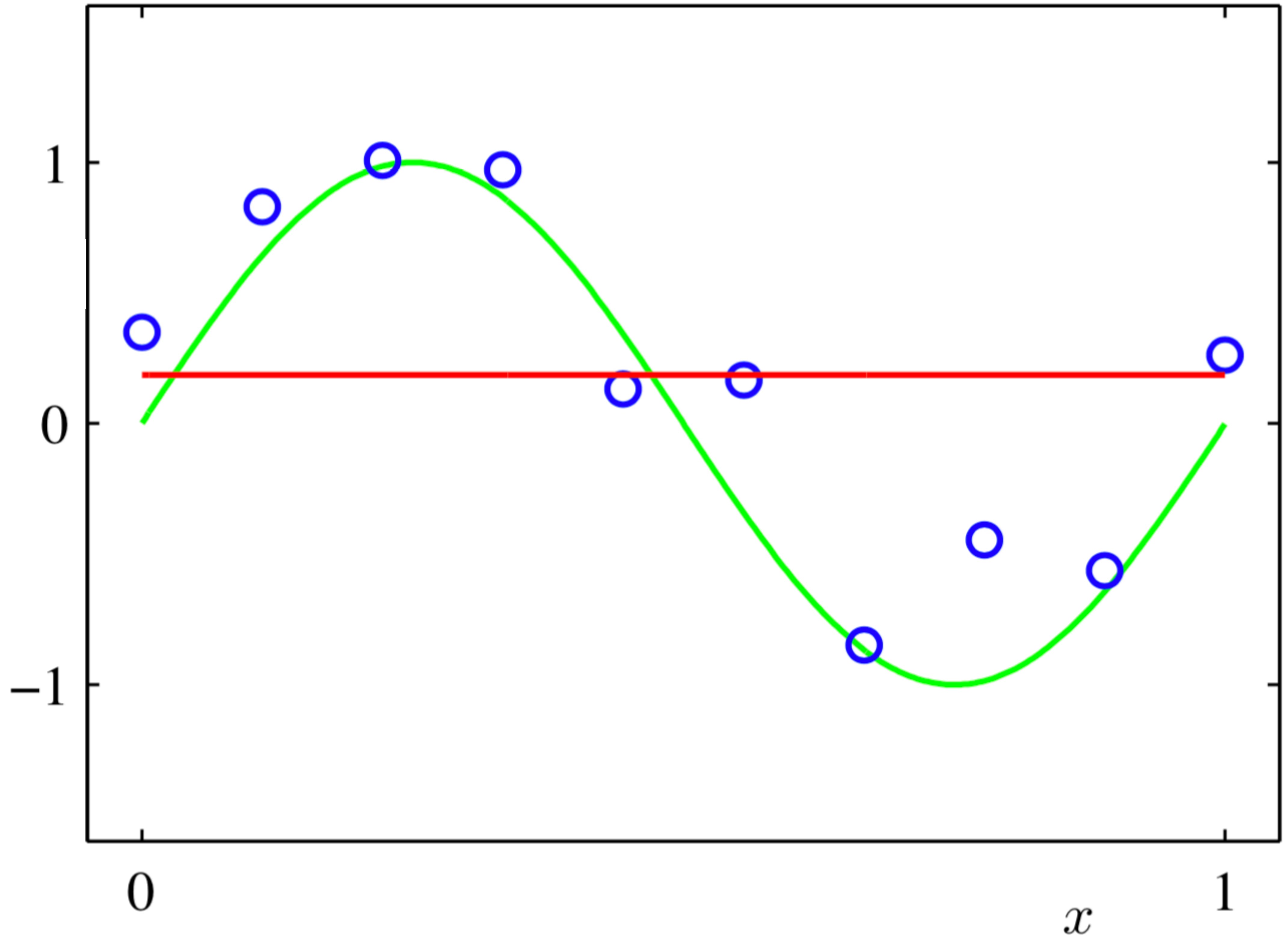
Underfitting

Example:

Underfitting

Example:

Fit one-parameter MSE model to match blue circles

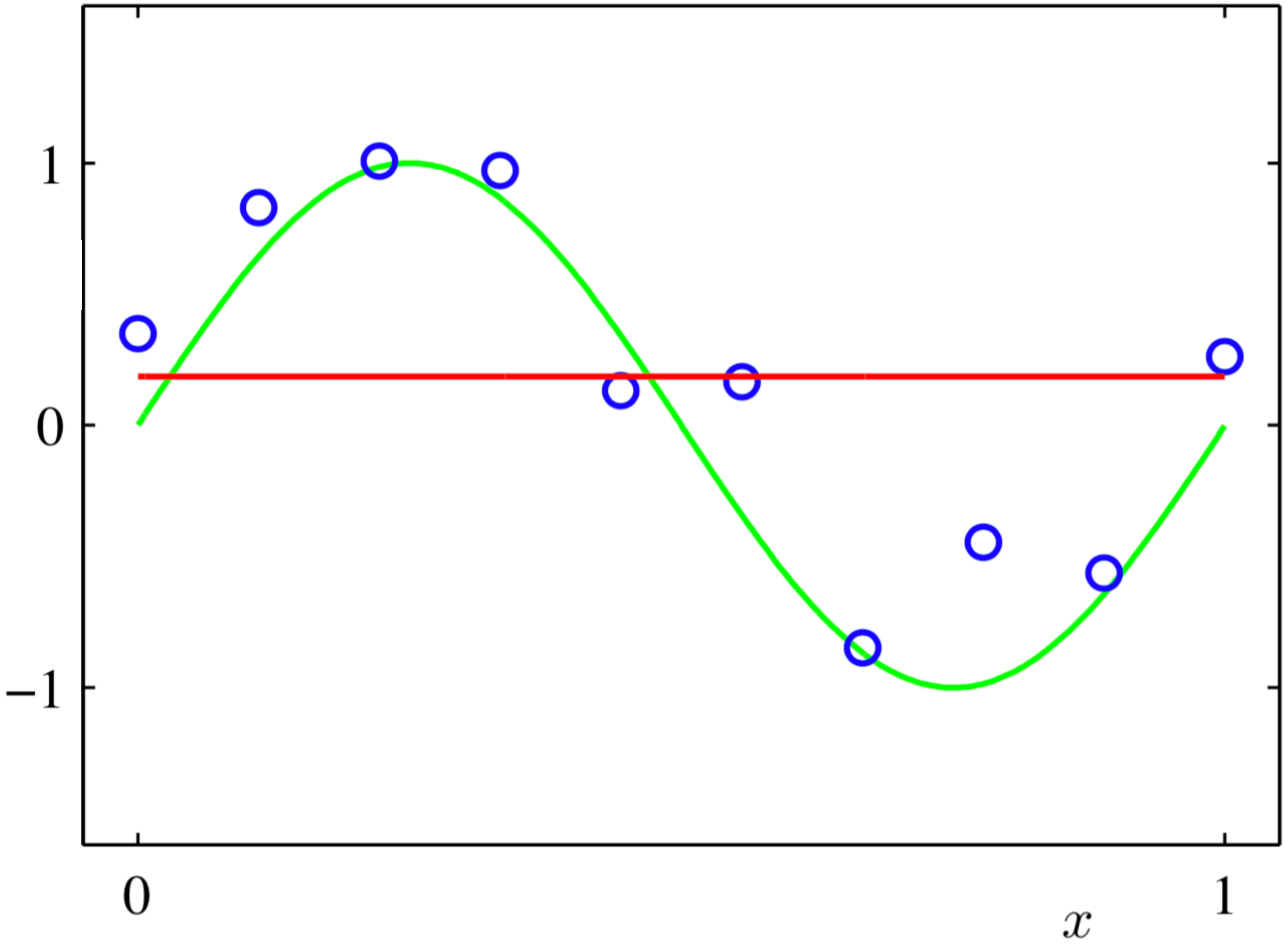


Bishop 2006

Underfitting

Example:

Fit one-parameter MSE model to match blue circles



Bishop 2006

Regardless of how many samples, we will never be able to fit the green curve!

Extended/Augmented feature vectors

The previous example seems to suggest that linear models are often too simple and tend to underfit

Extended/Augmented feature vectors

The previous example seems to suggest that linear models are often too simple and tend to underfit

We will see that quite the opposite is true, but first we discuss a remedy for the underfitting of linear models

Extended/Augmented feature vectors

The previous example seems to suggest that linear models are often too simple and tend to underfit

We will see that quite the opposite is true, but first we discuss a remedy for the underfitting of linear models

Standard trick: augment input with polynomial basis of degree d , i.e.

Extended/Augmented feature vectors

The previous example seems to suggest that linear models are often too simple and tend to underfit

We will see that quite the opposite is true, but first we discuss a remedy for the underfitting of linear models

Standard trick: augment input with polynomial basis of degree d , i.e.

consider $\boldsymbol{\phi}(x_i) = \begin{pmatrix} 1 & x_i & x_i^2 & \ldots & x_i^d \end{pmatrix}^T$

Extended/Augmented feature vectors

The previous example seems to suggest that linear models are often too simple and tend to underfit

We will see that quite the opposite is true, but first we discuss a remedy for the underfitting of linear models

Standard trick: augment input with polynomial basis of degree d , i.e.

consider $\boldsymbol{\phi}(x_i) = \begin{pmatrix} 1 & x_i & x_i^2 & \dots & x_i^d \end{pmatrix}^T$

and the linear model $f(x_i, \mathbf{w}) = \langle \boldsymbol{\phi}(x_i), \boldsymbol{w} \rangle = \sum_{k=0}^{d} x_i^k w_k$

Extended/Augmented feature vectors

The previous example seems to suggest that linear models are often too simple and tend to underfit

We will see that quite the opposite is true, but first we discuss a remedy for the underfitting of linear models

Standard trick: augment input with polynomial basis of degree d , i.e.

consider $\boldsymbol{\phi}(x_i) = \begin{pmatrix} 1 & x_i & x_i^2 & \dots & x_i^d \end{pmatrix}^T$

$x_i \in \mathbb{R}$

$\boldsymbol{w} \in \mathbb{R}^{d+1}$

and the linear model $f(x_i, \mathbf{w}) = \langle \boldsymbol{\phi}(x_i), \boldsymbol{w} \rangle = \sum_{k=0}^{d} x_i^k w_k$

Extended/Augmented feature vectors

$$\boldsymbol{\phi}(x_i) = \begin{pmatrix} 1 & x_i & x_i^2 & \dots & x_i^d \end{pmatrix}^T$$

$$f(x_i, \boldsymbol{w}) = \langle \boldsymbol{\phi}(x_i), \boldsymbol{w} \rangle = \sum_{k=0}^{d} x_i^k w_k$$

Notation: $\boldsymbol{\Phi}(\boldsymbol{X}) = \begin{pmatrix} \boldsymbol{\phi}(x_1)^T \\ \boldsymbol{\phi}(x_2)^T \\ \vdots \\ \boldsymbol{\phi}(x_s)^T \end{pmatrix} \in \mathbb{R}^{s\times(d+1)}$

Extended/Augmented feature vectors

$$\boldsymbol{\phi}(x_i) = \begin{pmatrix} 1 & x_i & x_i^2 & \dots & x_i^d \end{pmatrix}^T$$

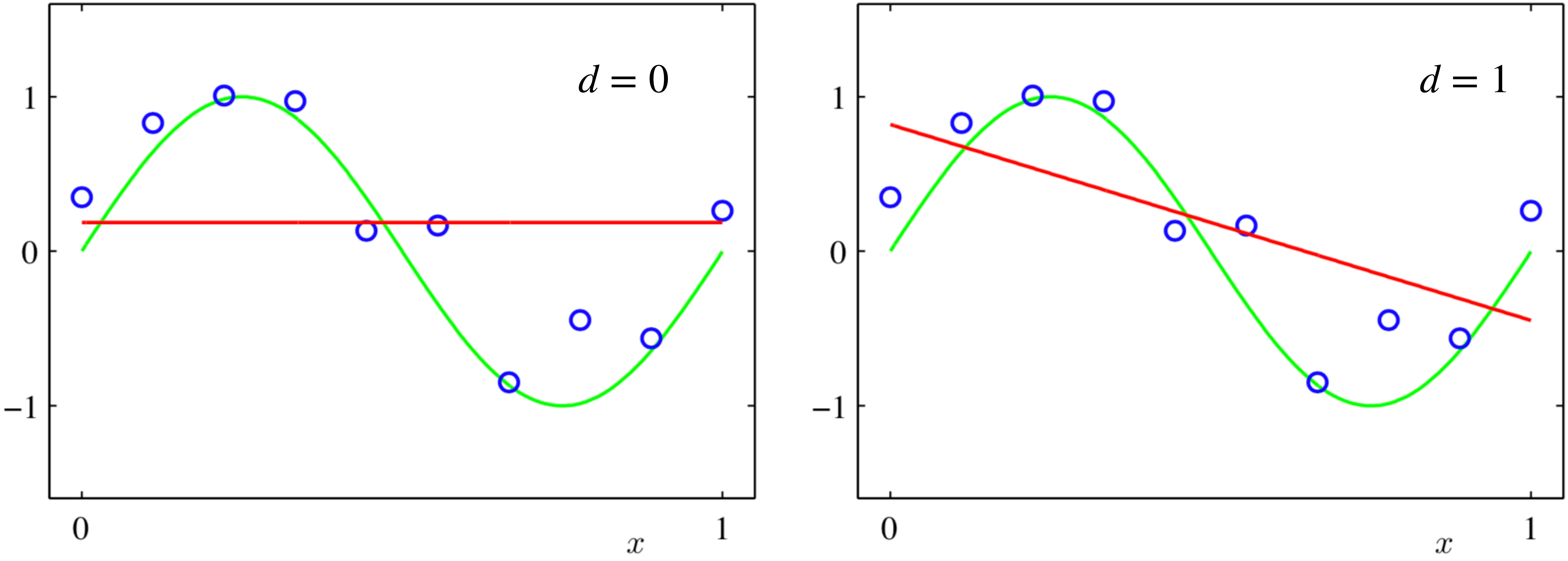
$$f(x_i, \boldsymbol{w}) = \langle \boldsymbol{\phi}(x_i), \boldsymbol{w} \rangle = \sum_{k=0}^{d} x_i^k w_k$$

Notation: $\boldsymbol{\Phi}(\boldsymbol{X}) = \begin{pmatrix} \boldsymbol{\phi}(x_1)^T \\ \boldsymbol{\phi}(x_2)^T \\ \vdots \\ \boldsymbol{\phi}(x_s)^T \end{pmatrix} \in \mathbb{R}^{s \times (d+1)}$

Modified MSE-problem:

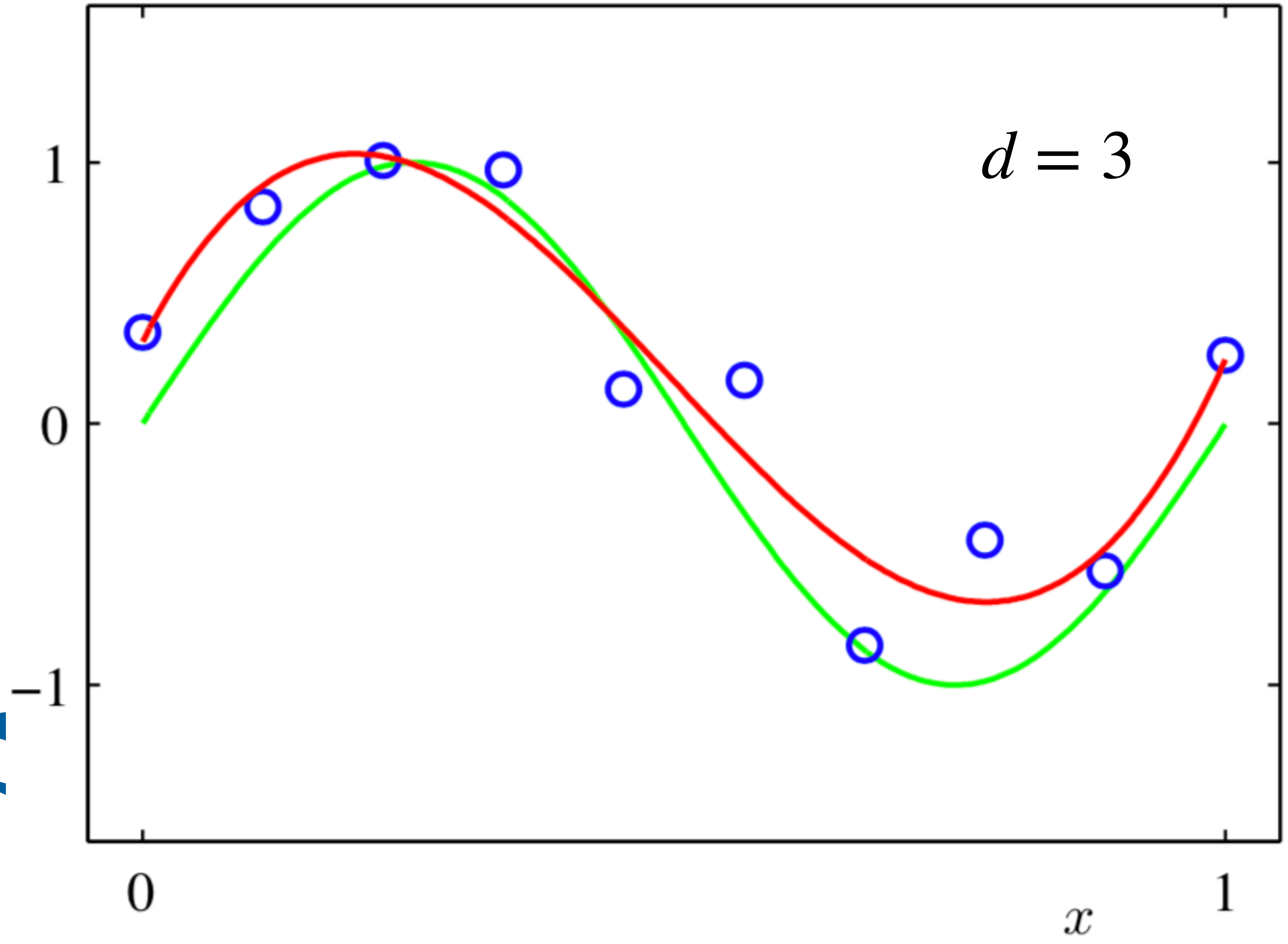
$$\hat{\boldsymbol{w}} = \arg \min_{\boldsymbol{w} \in \mathbb{R}^{d+1}} \left\{ \frac{1}{2s} \parallel \boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{w} - \boldsymbol{y} \parallel^2 \right\}$$

From under- to overfitting



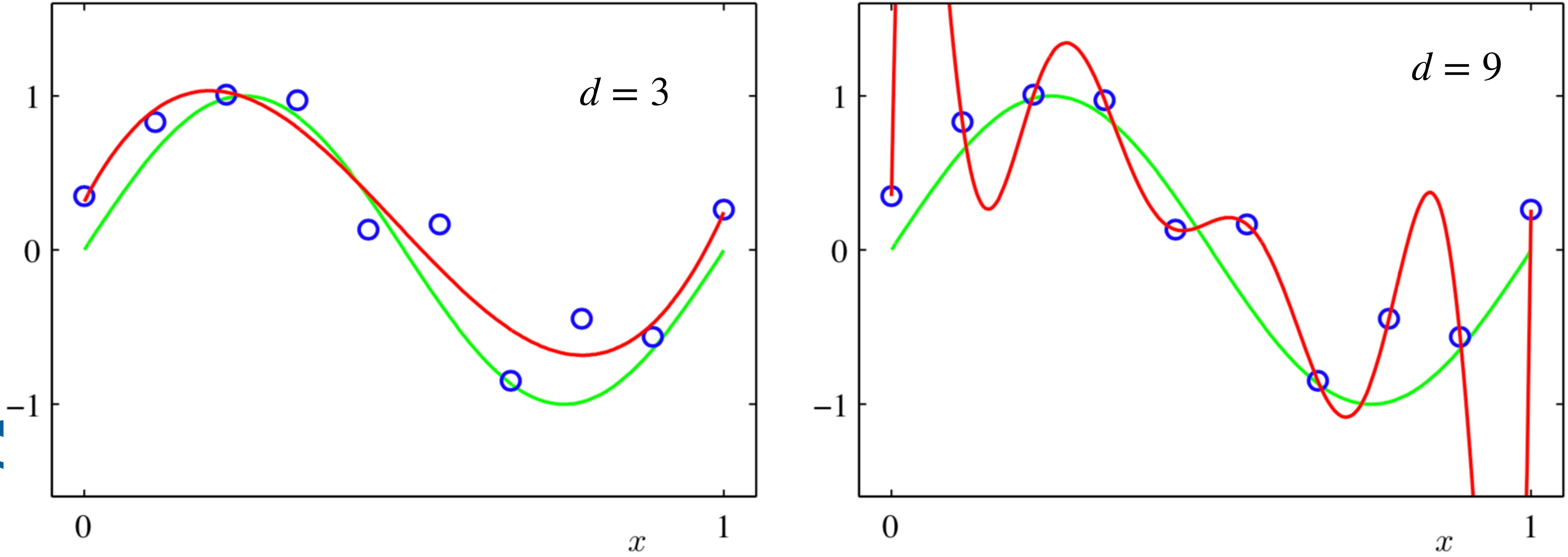
Bishop 2006

From under- to overfitting



Bishop 2006

From under- to overfitting



Bishop 2006

From under- to overfitting

$d = 0$ function is underfitting

$d = 1$ function is underfitting

$d = 3$ function seems to fit reasonably well

$d = 9$ function is overfitting

From under- to overfitting

$d = 0$ function is underfitting

$d = 1$ function is underfitting

$d = 3$ function seems to fit reasonably well

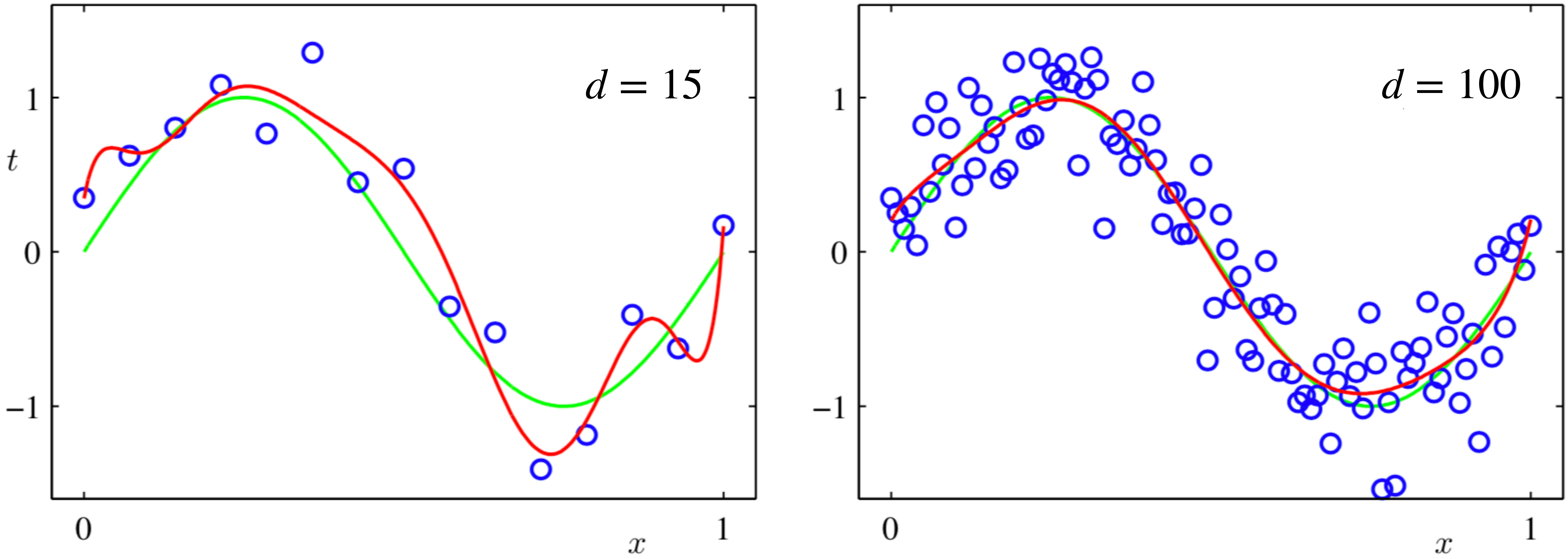
$d = 9$ function is overfitting

What can we do to prevent overfitting?

From under- to overfitting

We could increase the no. of samples s :

Bishop 2006



Or we could use regularisation (next week's topic)

MINIMISERS & THE ROLE OF CONVEXITY

Minimisers & the role of convexity

We have made the following assumption:

Minimisers & the role of convexity

We have made the following assumption:

In order to compute

Minimisers & the role of convexity

We have made the following assumption:

In order to compute

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

Minimisers & the role of convexity

We have made the following assumption:

In order to compute

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

we can solve

Minimisers & the role of convexity

We have made the following assumption:

In order to compute

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

we can solve

$$\nabla \mathsf{MSE}(\hat{\mathbf{w}}) = 0$$

Minimisers & the role of convexity

We have made the following assumption:

In order to compute

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

we can solve

$$\nabla \mathrm{MSE}(\hat{\mathbf{w}}) = 0 \qquad \Leftrightarrow$$

Minimisers & the role of convexity

We have made the following assumption:

In order to compute

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

we can solve

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

Minimisers & the role of convexity

Or

In order to compute

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

we can solve

$$\nabla \mathrm{MSE}(\hat{\mathbf{w}}) = 0 \qquad \Leftrightarrow \qquad \boldsymbol{\Phi}(X)^{\mathsf{T}}\boldsymbol{\Phi}(X)\,\hat{w} = \boldsymbol{\Phi}(X)^{\mathsf{T}}\mathbf{y}$$

Minimisers & the role of convexity

This raises a couple of questions:

Minimisers & the role of convexity

This raises a couple of questions:

1. Why is computing

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^{d+1}} \mathsf{MSE}(\mathbf{w})$$

   equivalent to solving

$$\nabla \mathsf{MSE}(\hat{\mathbf{w}}) = 0 \quad ?$$

Minimisers & the role of convexity

This raises a couple of questions:

1. Why is computing

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^{d+1}} \mathrm{MSE}(\mathbf{w})$$

   equivalent to solving

$$\nabla \mathrm{MSE}(\hat{\mathbf{w}}) = 0 \quad ?$$

2. Does a solution $\hat{\mathbf{w}}$ always exist?

Minimisers & the role of convexity

This raises a couple of questions:

1. Why is computing

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^{d+1}} \mathsf{MSE}(\mathbf{w})$$

equivalent to solving

$$\nabla \mathsf{MSE}(\hat{\mathbf{w}}) = 0 \quad ?$$

2. Does a solution $\hat{\mathbf{w}}$ always exist?

3. Is the solution $\hat{\mathbf{w}}$ unique?

Minimisers & the role of convexity

1. For now we assume the first condition to be true (we will verify this later)

2. Does a solution $\hat{\mathbf{w}}$ always exist?

3. Is the solution $\hat{\mathbf{w}}$ unique?

Minimisers & the role of convexity

1. For now we assume the first condition to be true (we will verify this later)

2. Does a solution $\hat{\mathbf{w}}$ always exist?

3. Is the solution $\hat{\mathbf{w}}$ unique?

Minimisers & the role of convexity

1. For now we assume the first condition to be true (we will verify this later)

2. Does a solution $\hat{\mathbf{w}}$ always exist?

3. Is the solution $\hat{\mathbf{w}}$ unique?

This is equivalent to asking when does a solution to

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

exist?

Minimisers & the role of convexity

1. For now we assume the first condition to be true (we will verify this later)

2. Does a solution $\hat{\mathbf{w}}$ always exist?

3. Is the solution $\hat{\mathbf{w}}$ unique?

This is equivalent to asking when does a solution to

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

exist?

Yes! Proof in the notes, not examinable

Minimisers & the role of convexity

1. For now we assume the first condition to be true

2. Does a solution $\hat{w}$ always exist?

3. Is the solution $\hat{w}$ unique?

Minimisers & the role of convexity

1. For now we assume the first condition to be true

2. Does a solution $\hat{w}$ always exist?

3. Is the solution $\hat{w}$ unique?

Before we can answer this, we need to introduce the concept of convexity first

CONVEXITY

Convexity of a cost function

What is a convex set?

Convexity of a cost function

What is a convex set?

A set C is called *convex* if for all $x, y \in C$ the element

$$z := \lambda x + (1 - \lambda)y$$

is also included in C, i.e. $z \in C$, for any $\lambda \in [0,1]$.

Convexity of a cost function

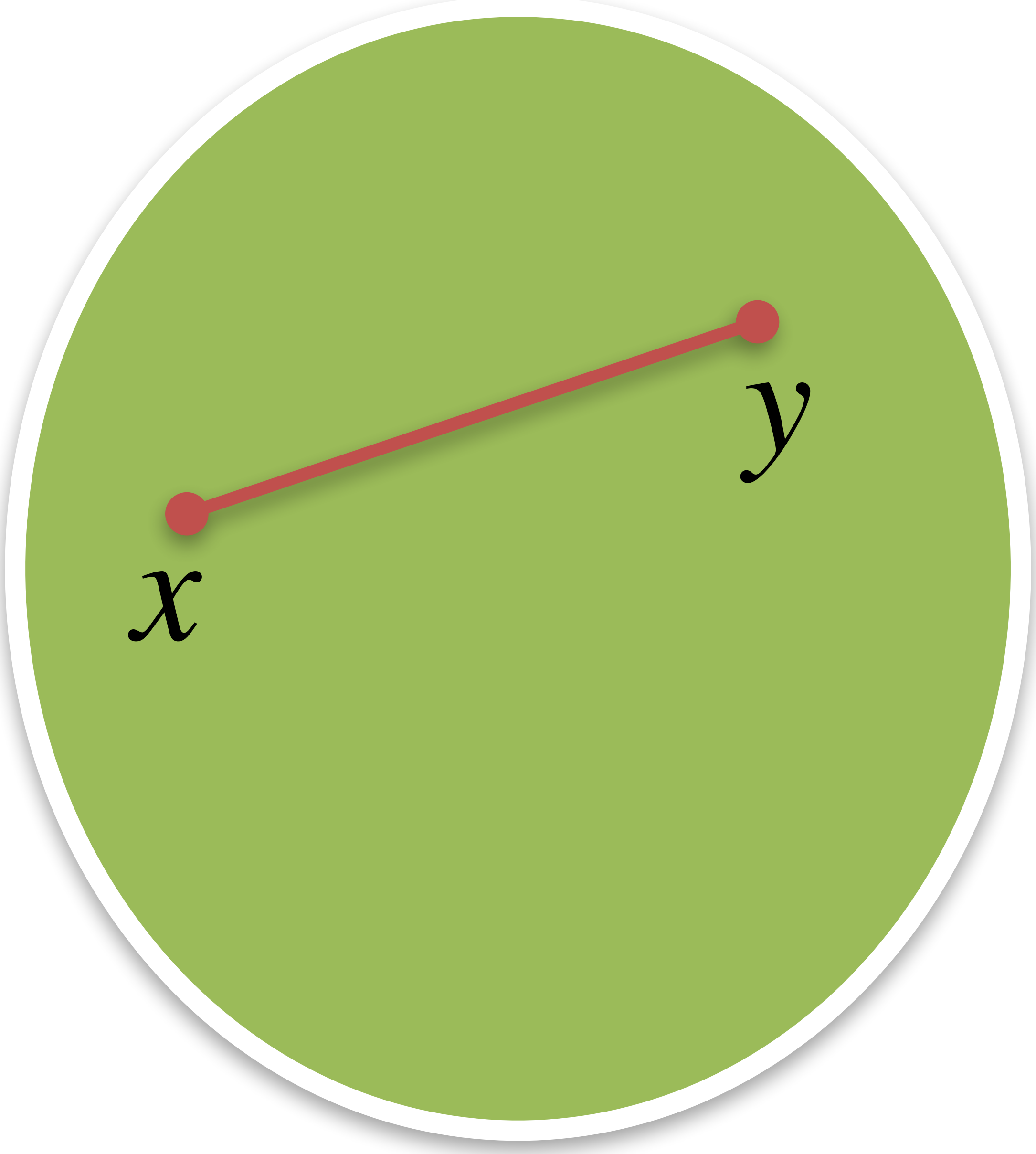
What is a convex set?

A set C is called *convex* if for all $x, y \in C$ the element

$$z := \lambda x + (1 - \lambda) y$$

is also included in C, i.e. $z \in C$, for any $\lambda \in [0,1]$.

A set C



Convexity of a cost function

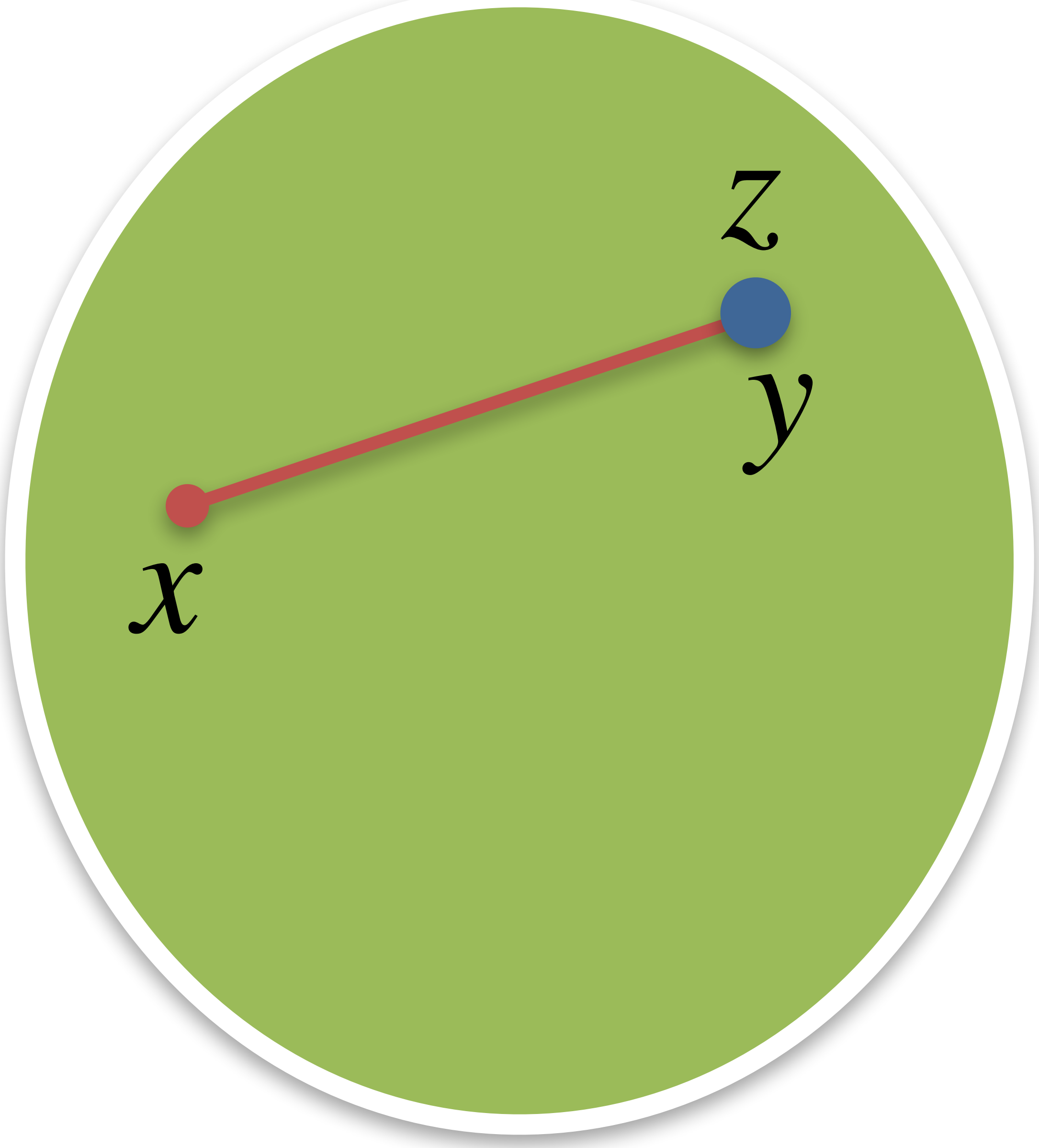
What is a convex set?

A set C is called *convex* if for all $x, y \in C$ the element

$$z := \lambda x + (1 - \lambda) y$$

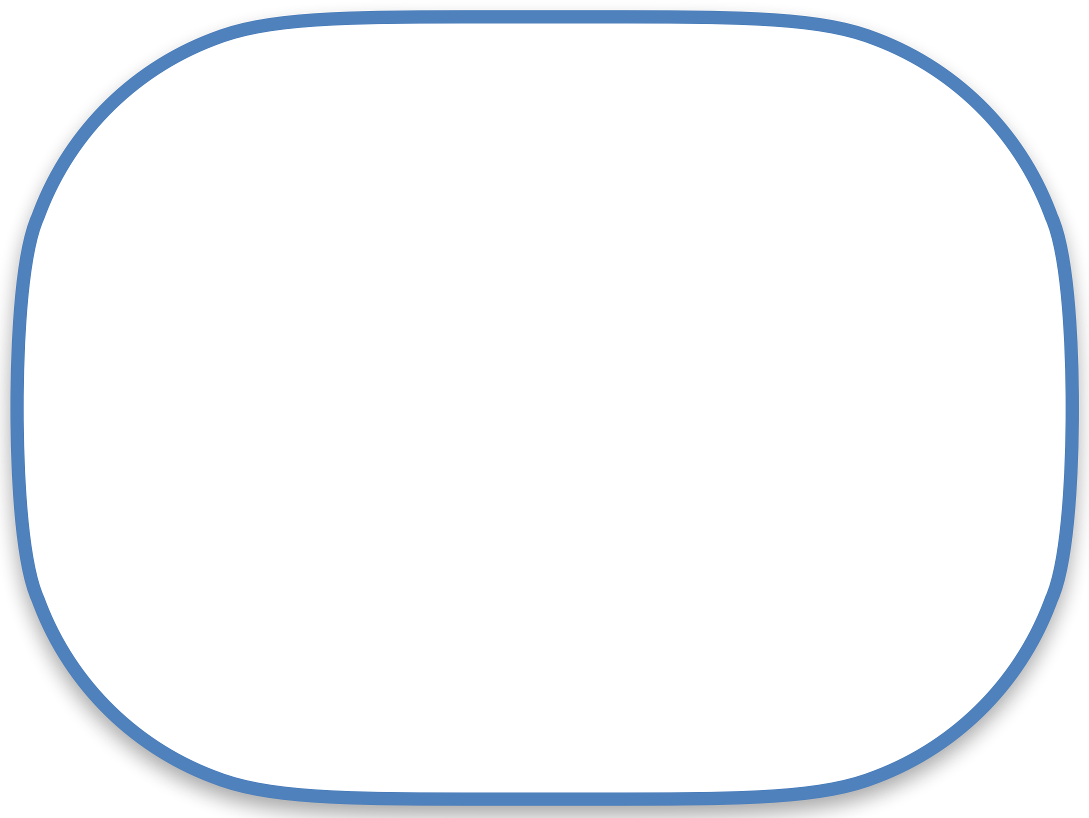
is also included in C, i.e. $z \in C$, for any $\lambda \in [0,1]$.

A set C

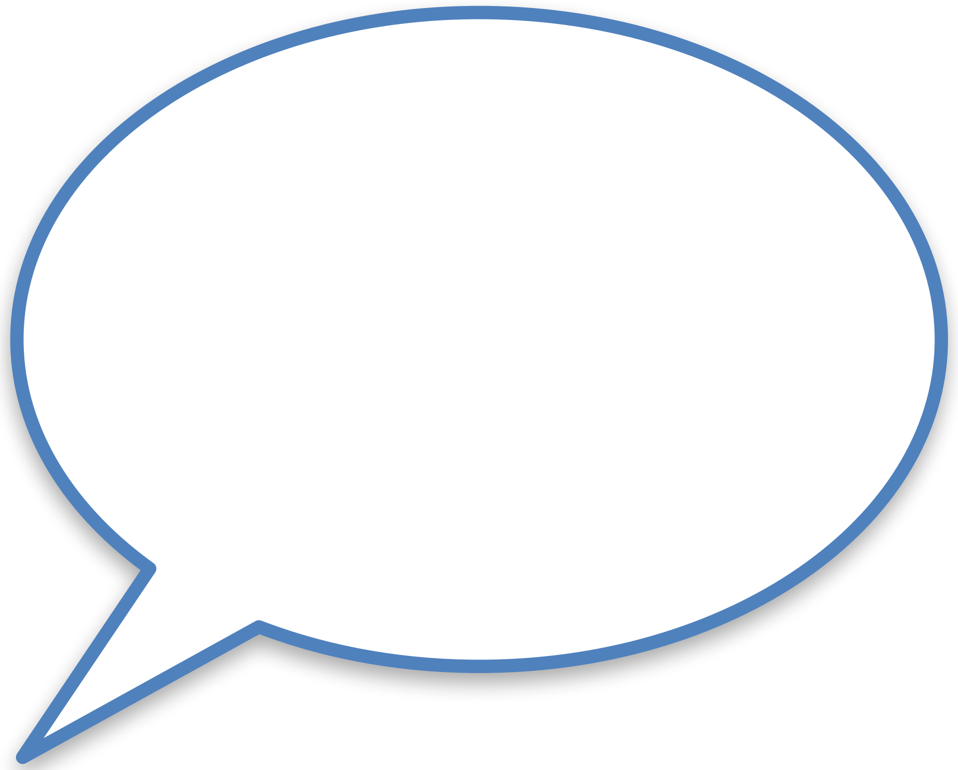


Convexity of a cost function

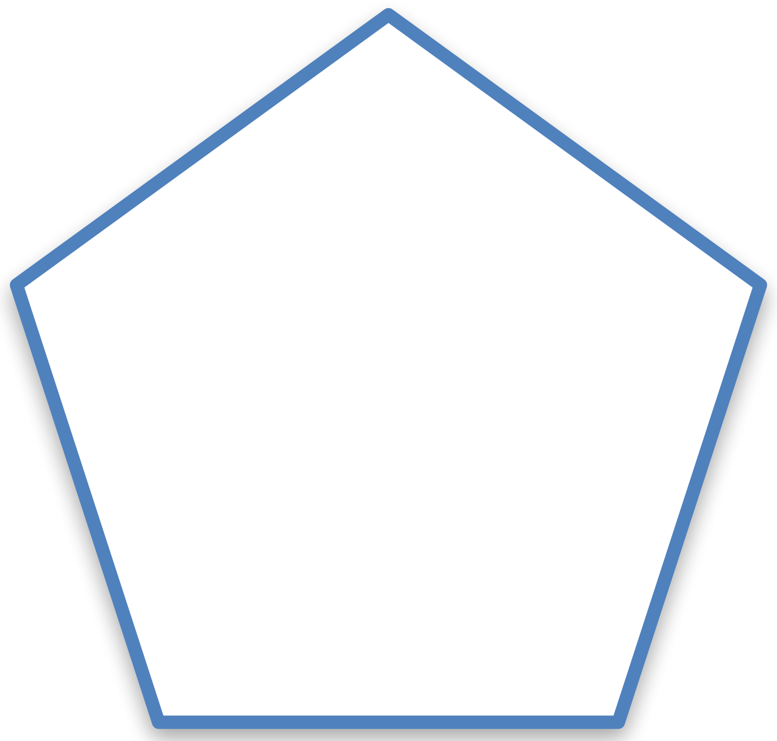
Which sets are convex?



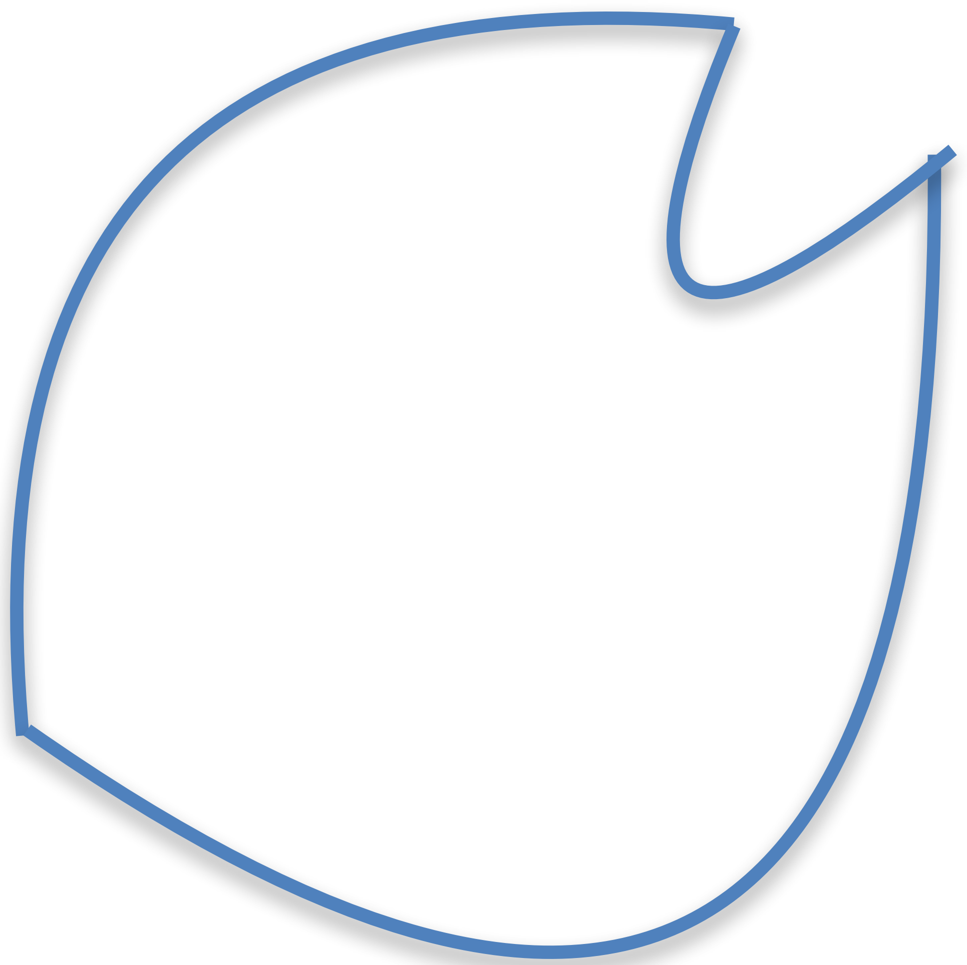
(a)



(b)



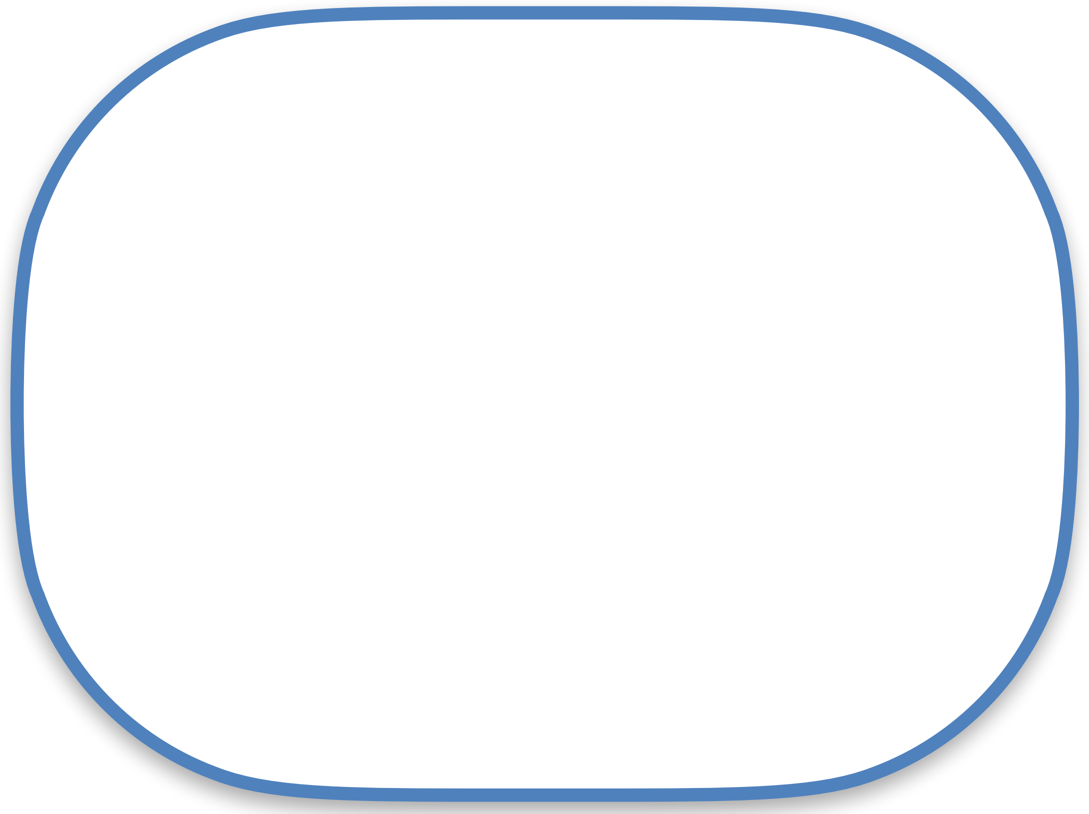
(c)



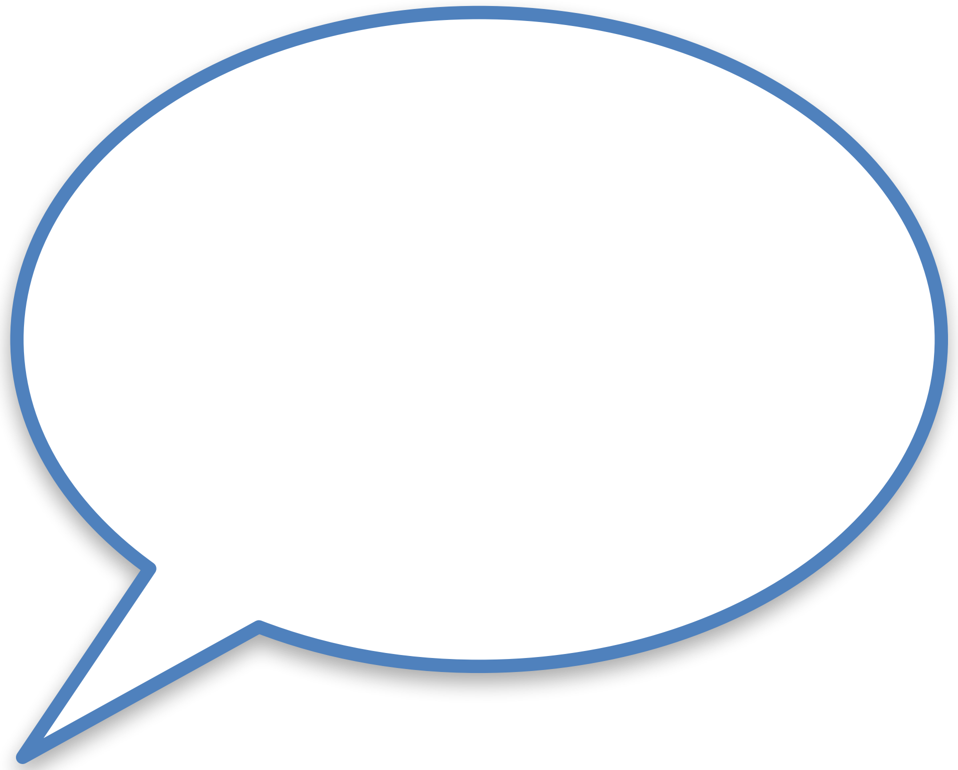
(d)

Convexity of a cost function

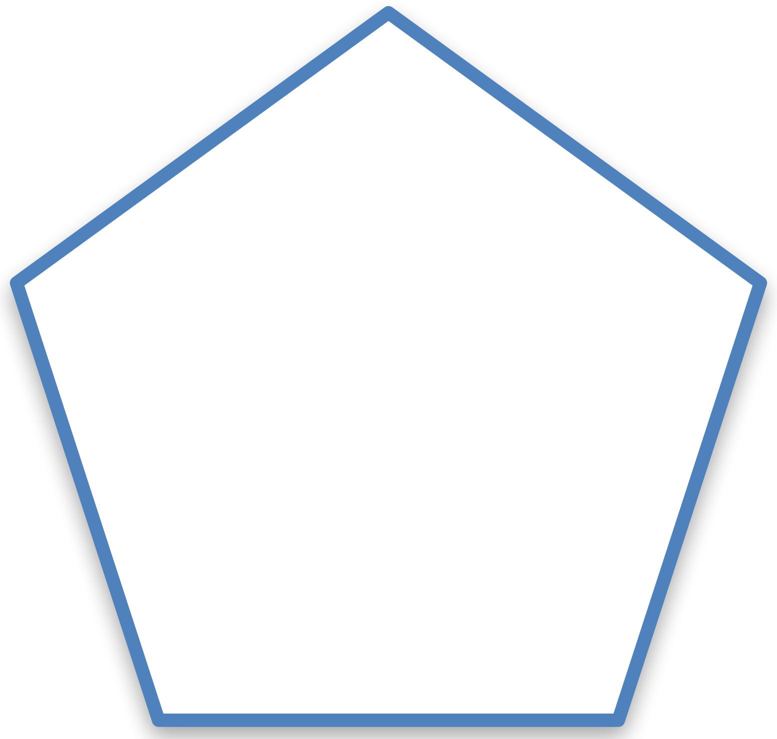
Which sets are convex?



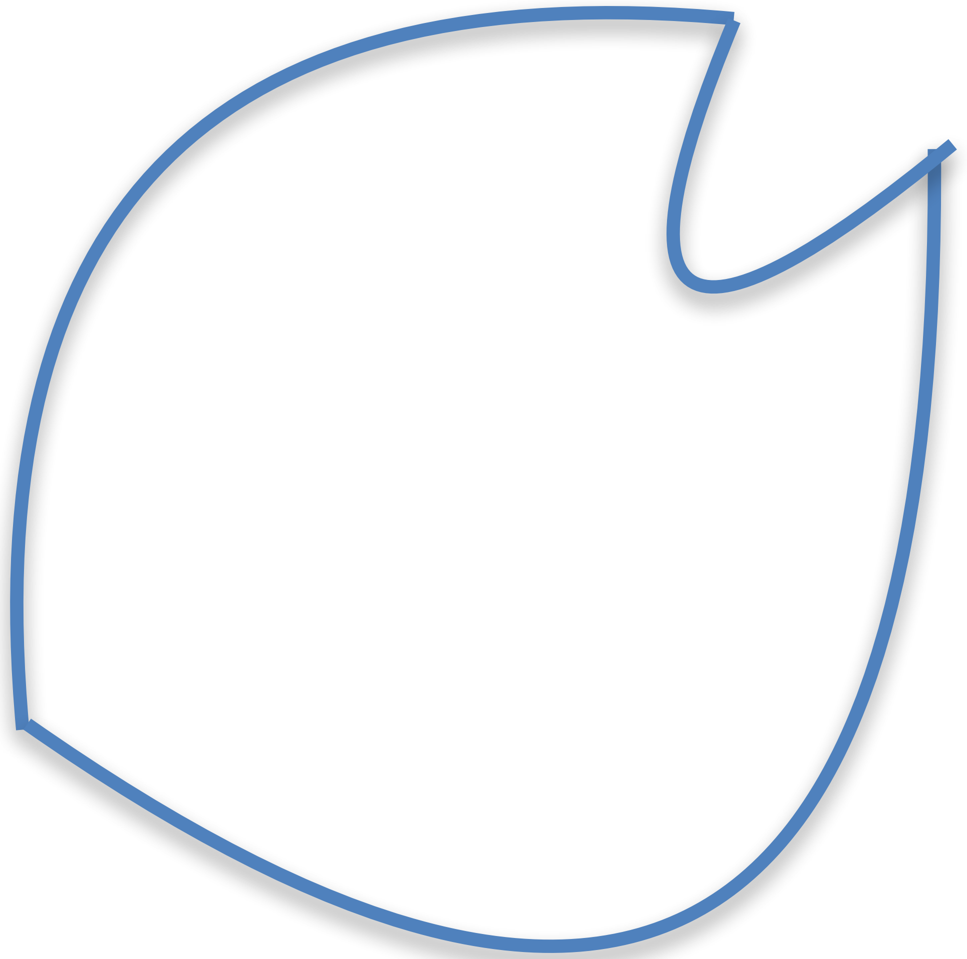
(a)



(b)



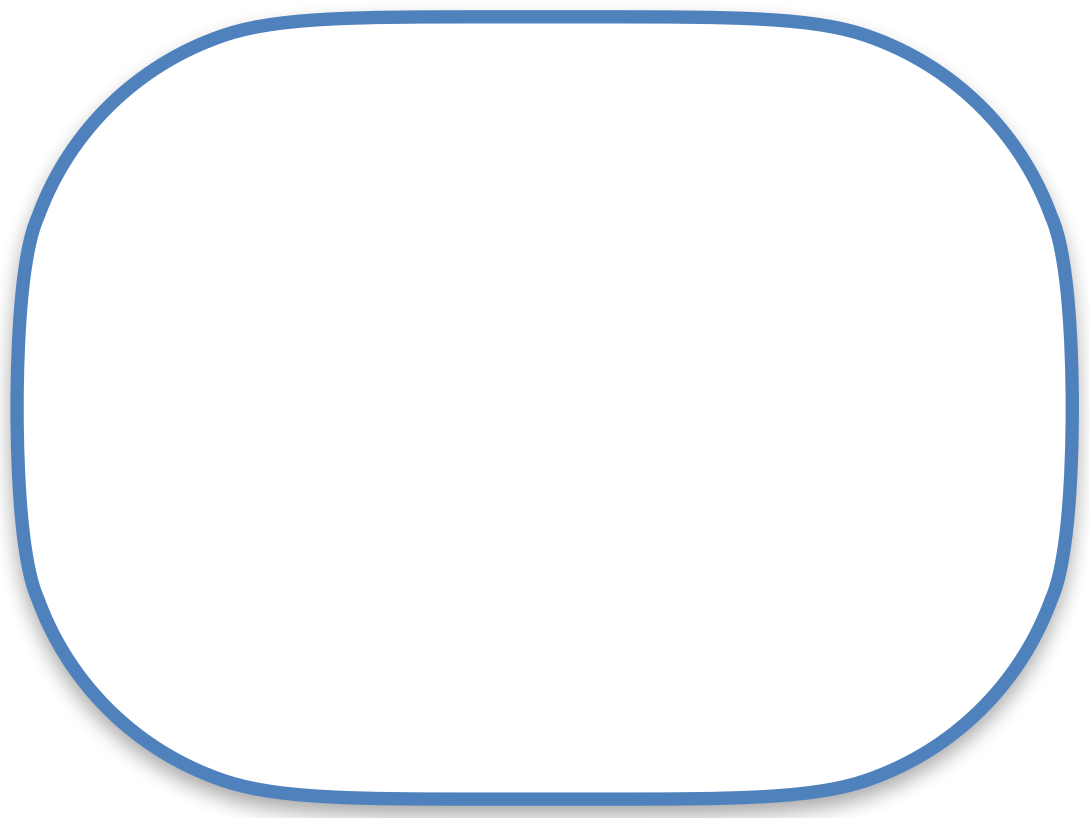
(c)



(d)

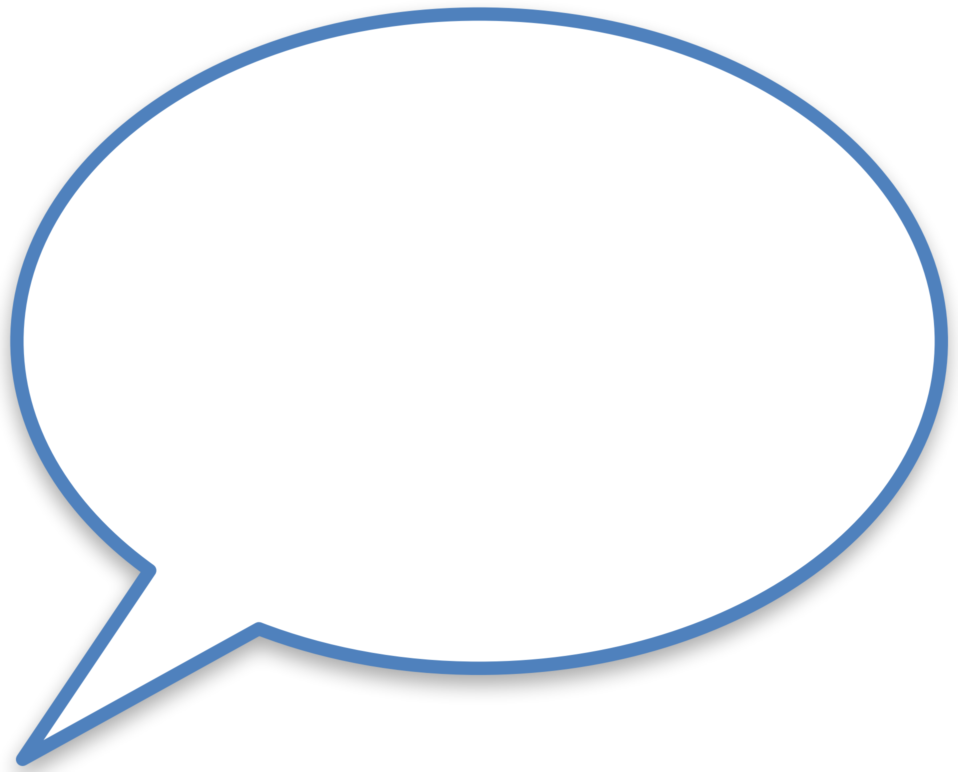
Convexity of a cost function

Which sets are convex?

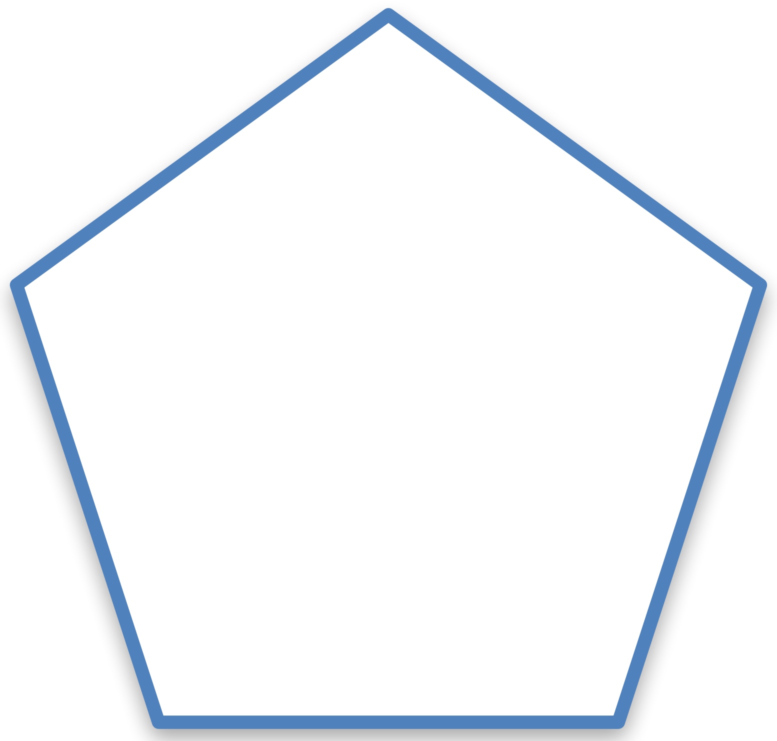


(a)

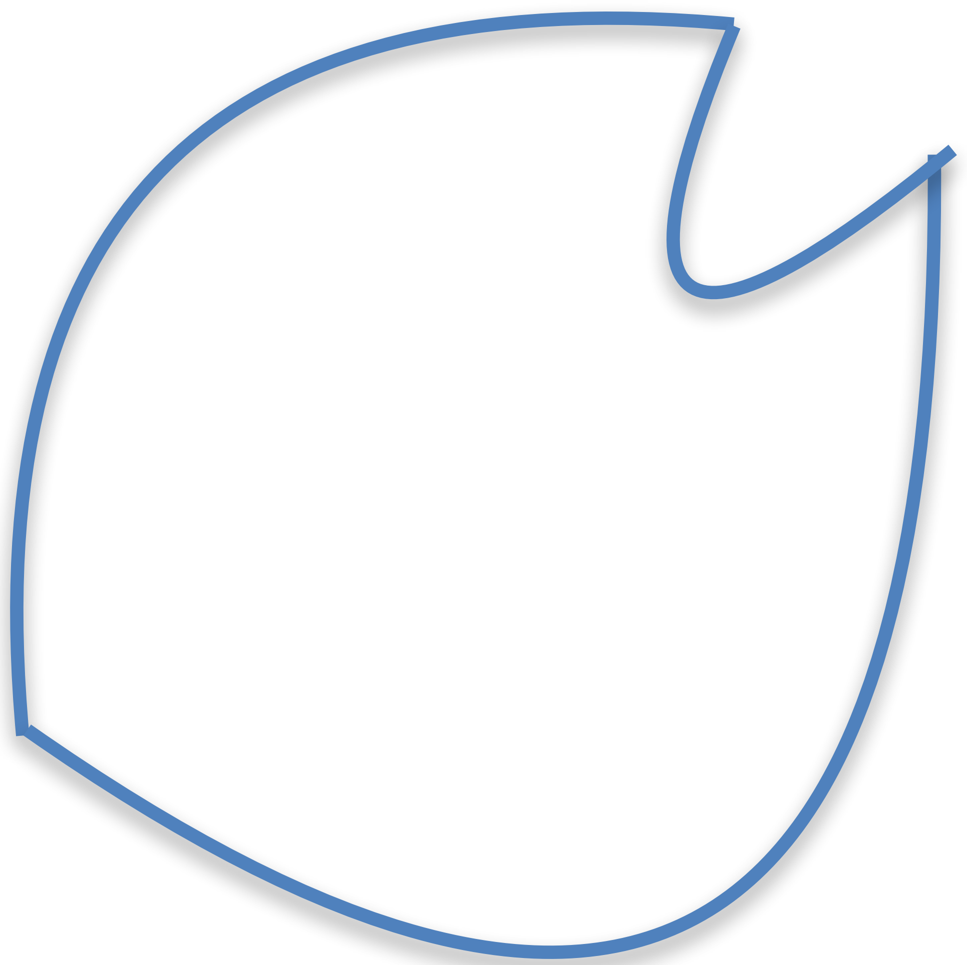
✓



(b)



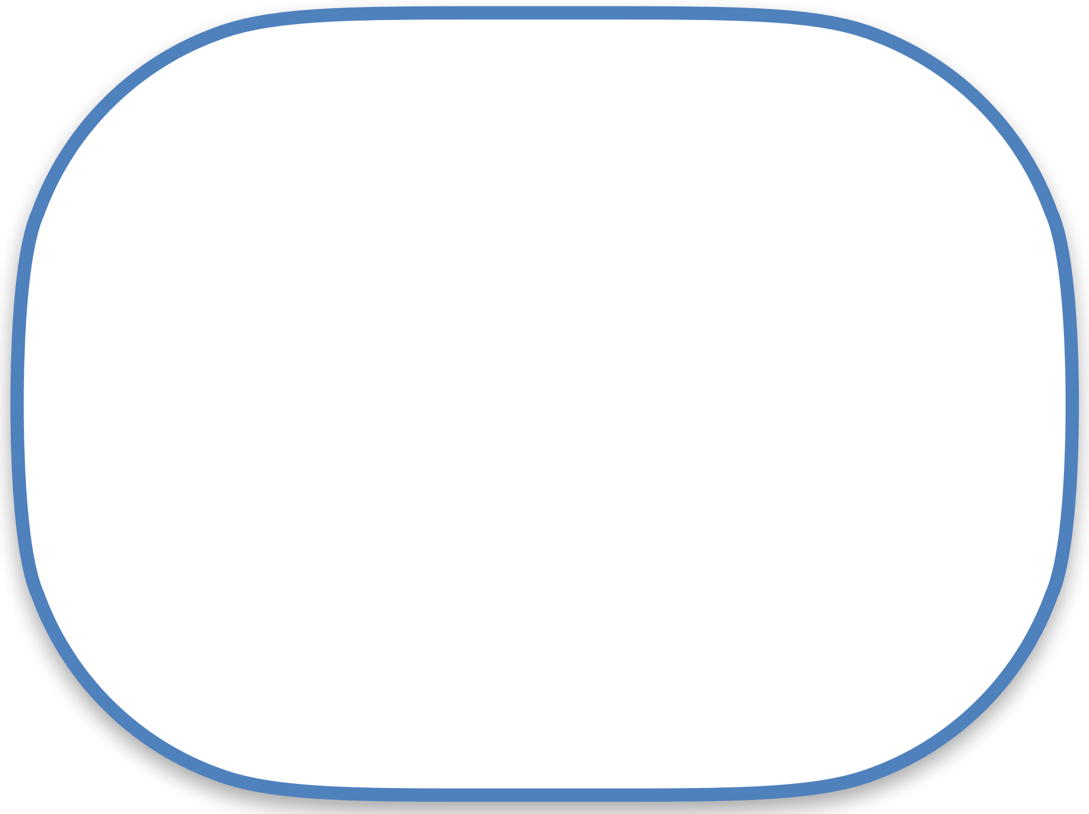
(c)



(d)

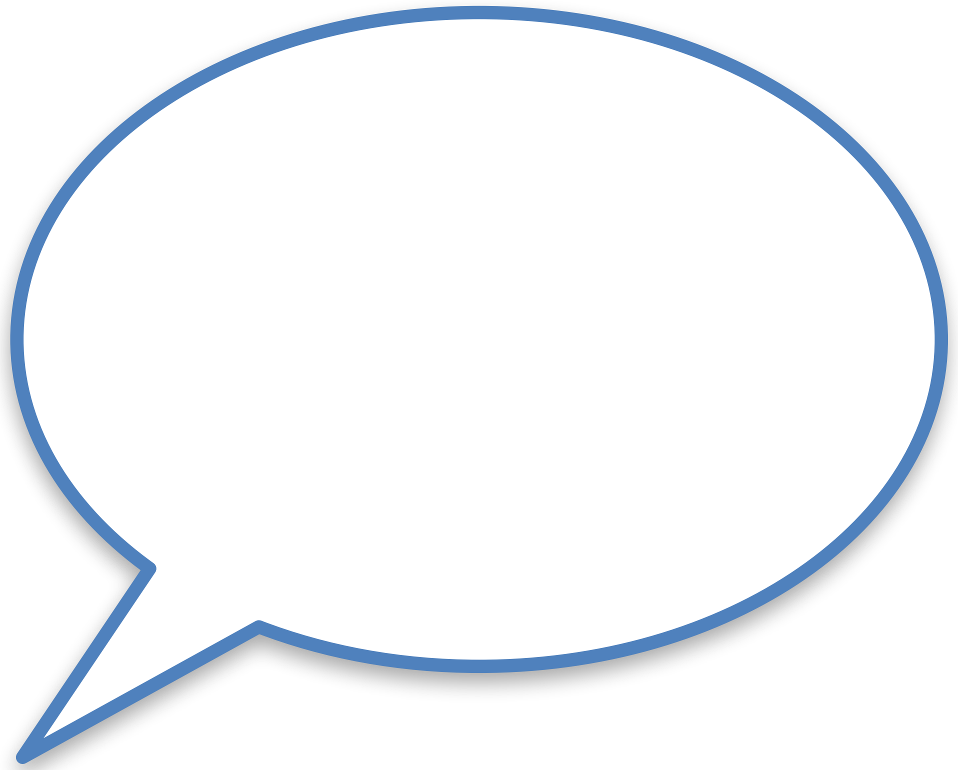
Convexity of a cost function

Which sets are convex?



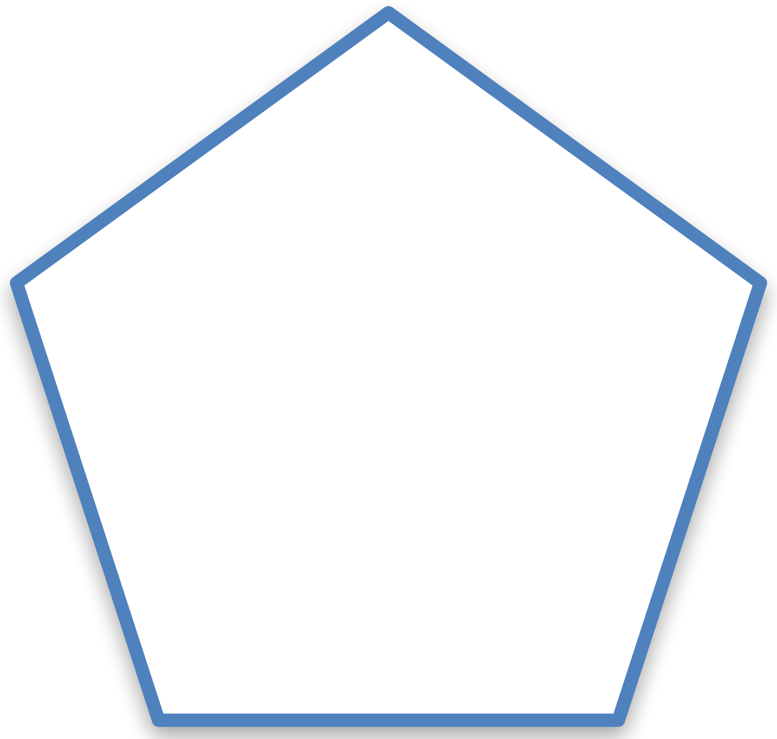
(a)

✓

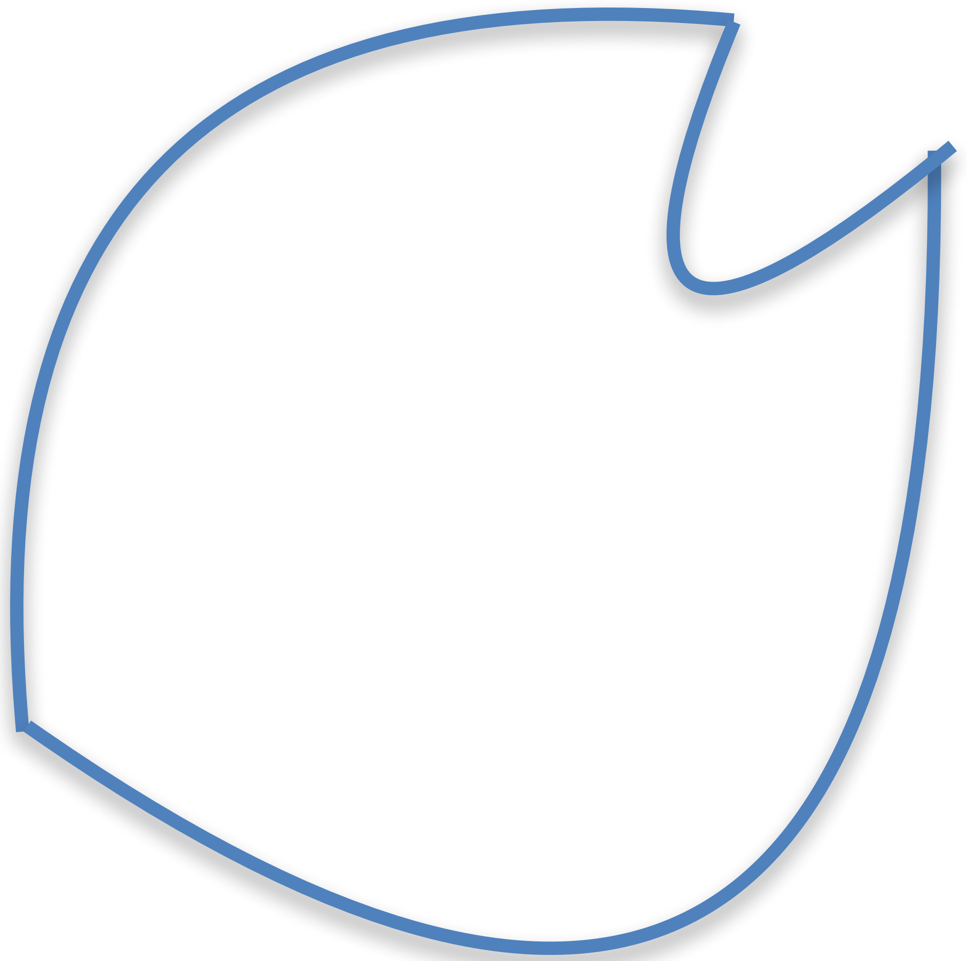


(b)

✗



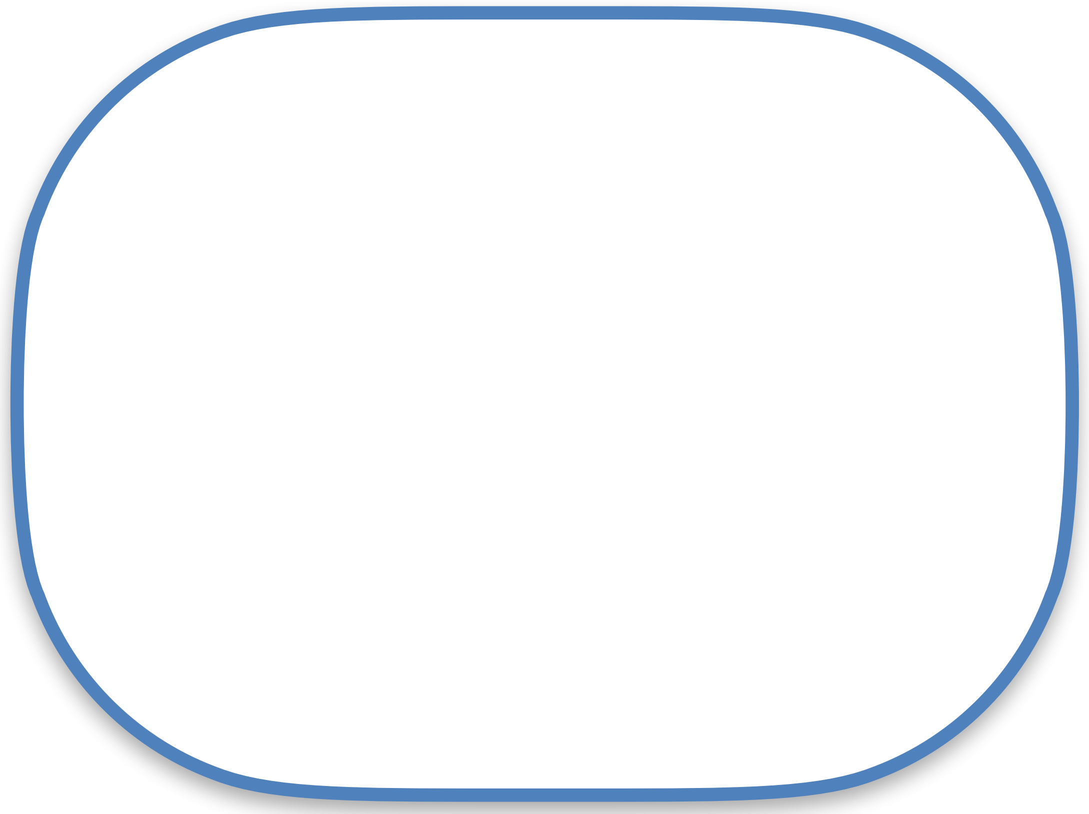
(c)



(d)

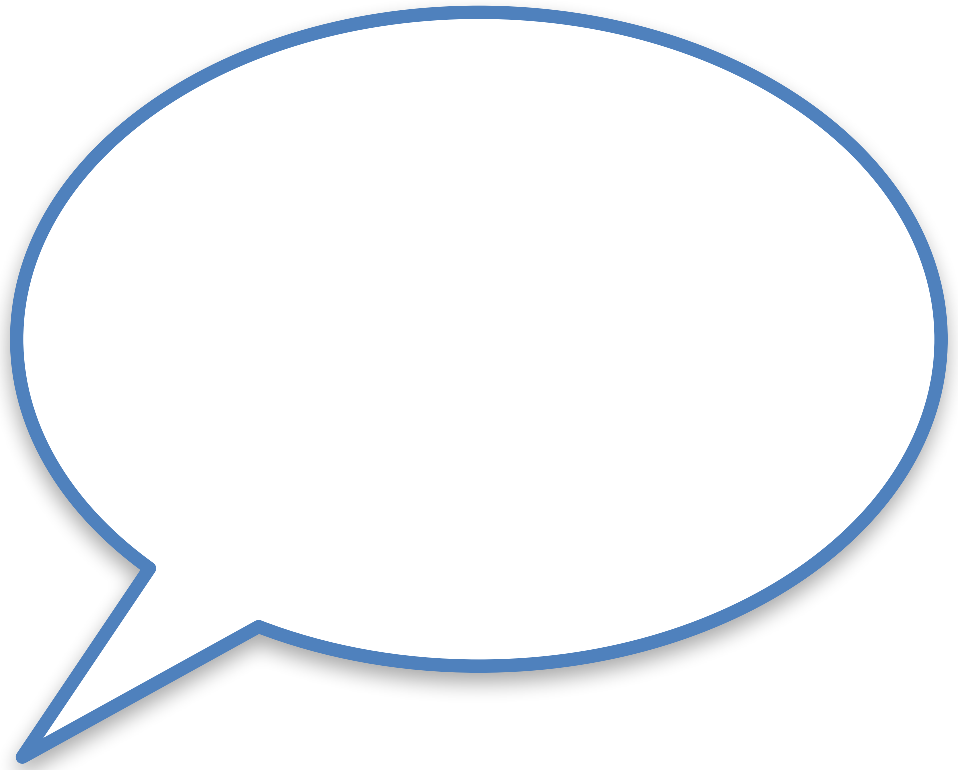
Convexity of a cost function

Which sets are convex?



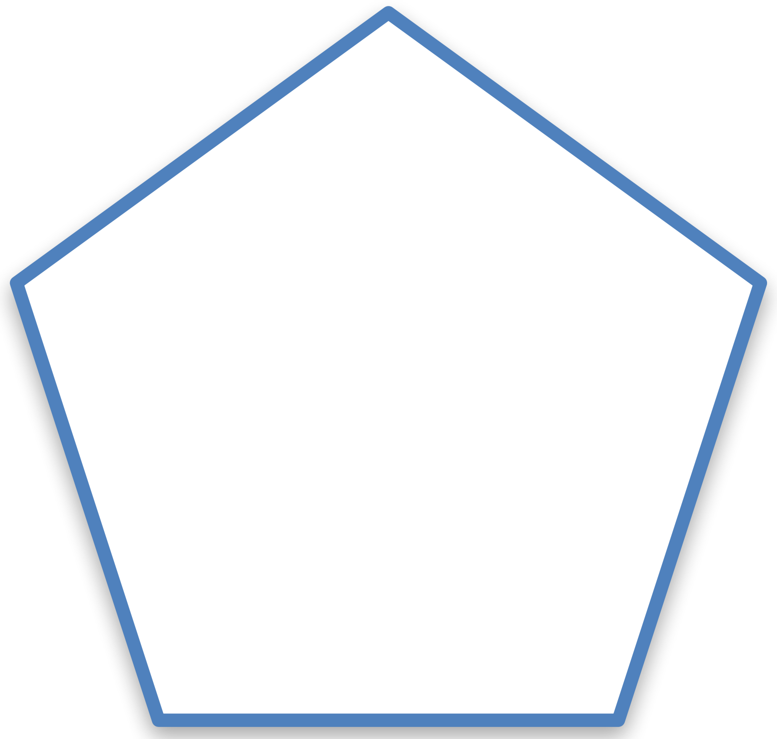
(a)

✓



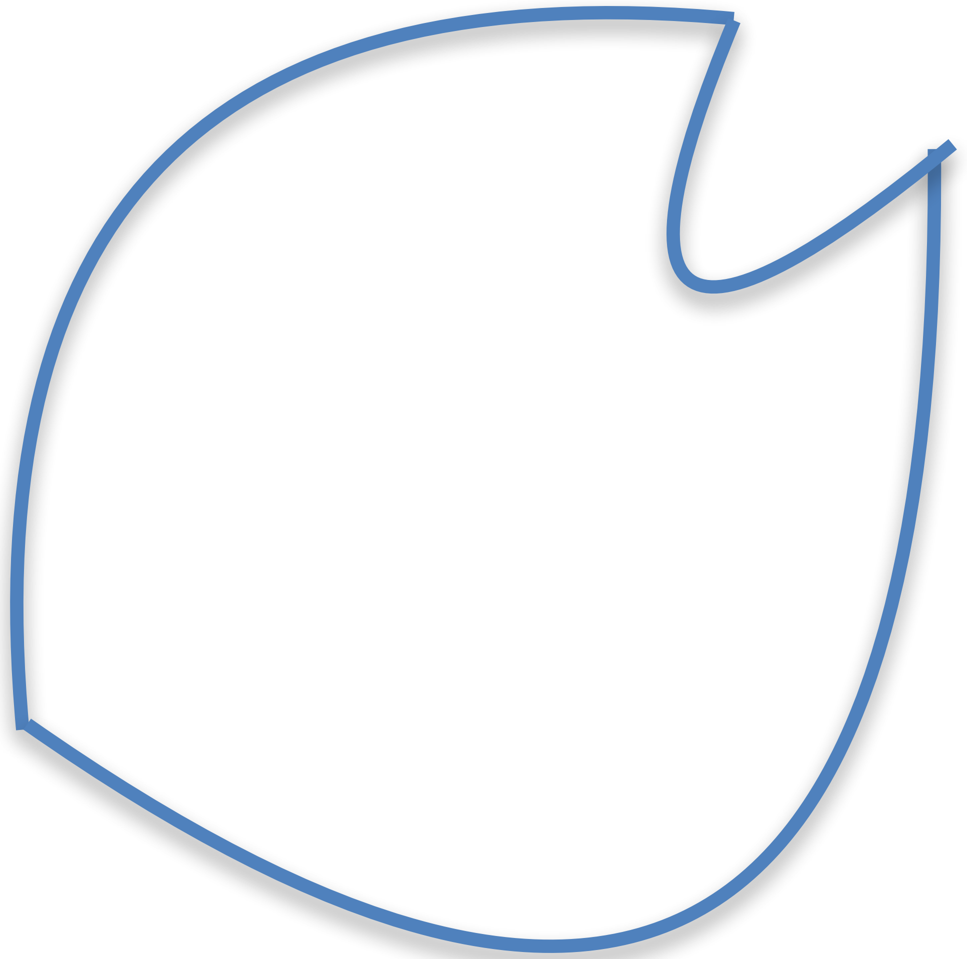
(b)

✗



(c)

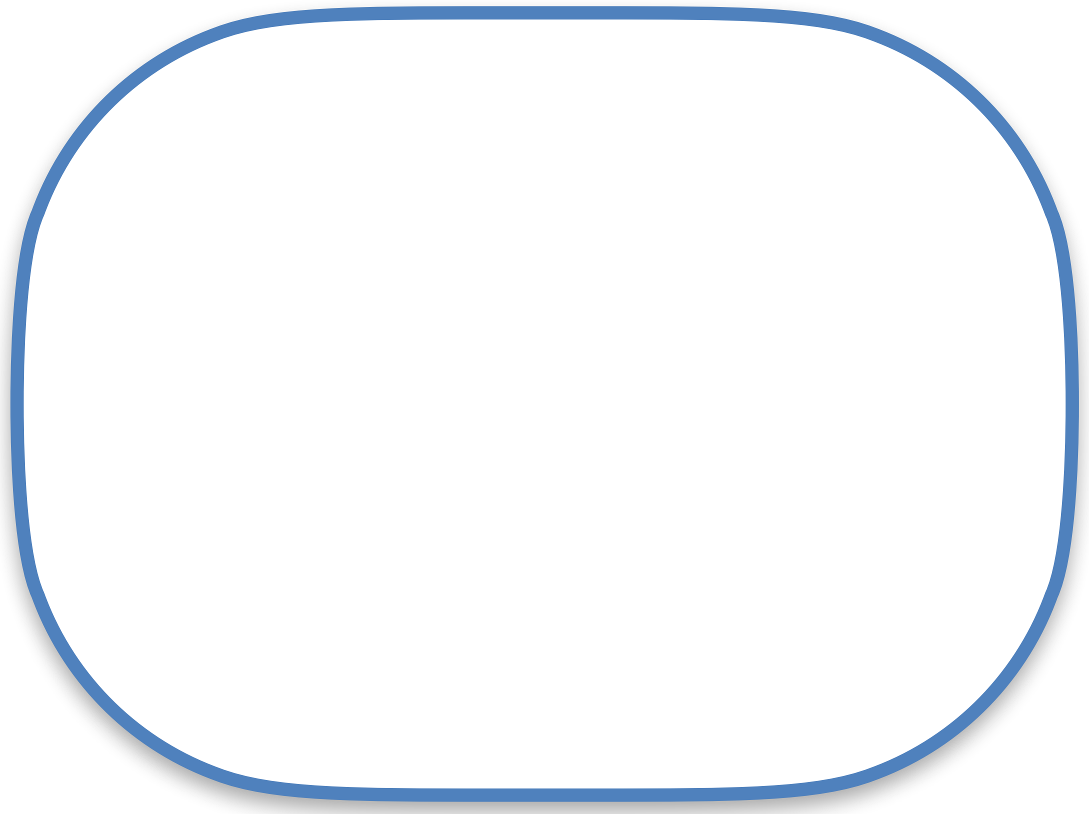
✓



(d)

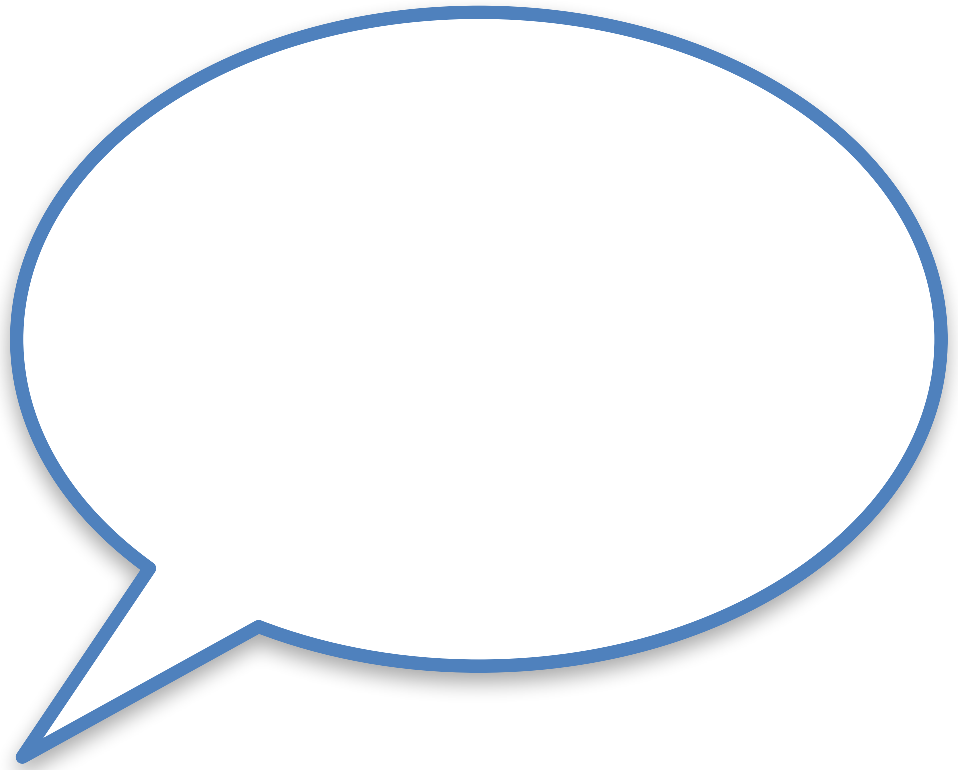
Convexity of a cost function

Which sets are convex?



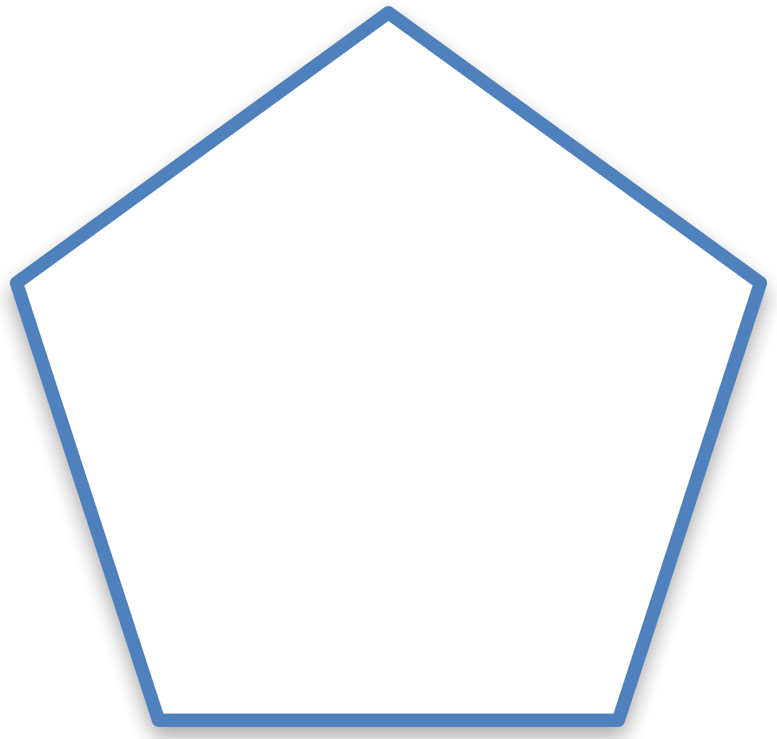
(a)

✓



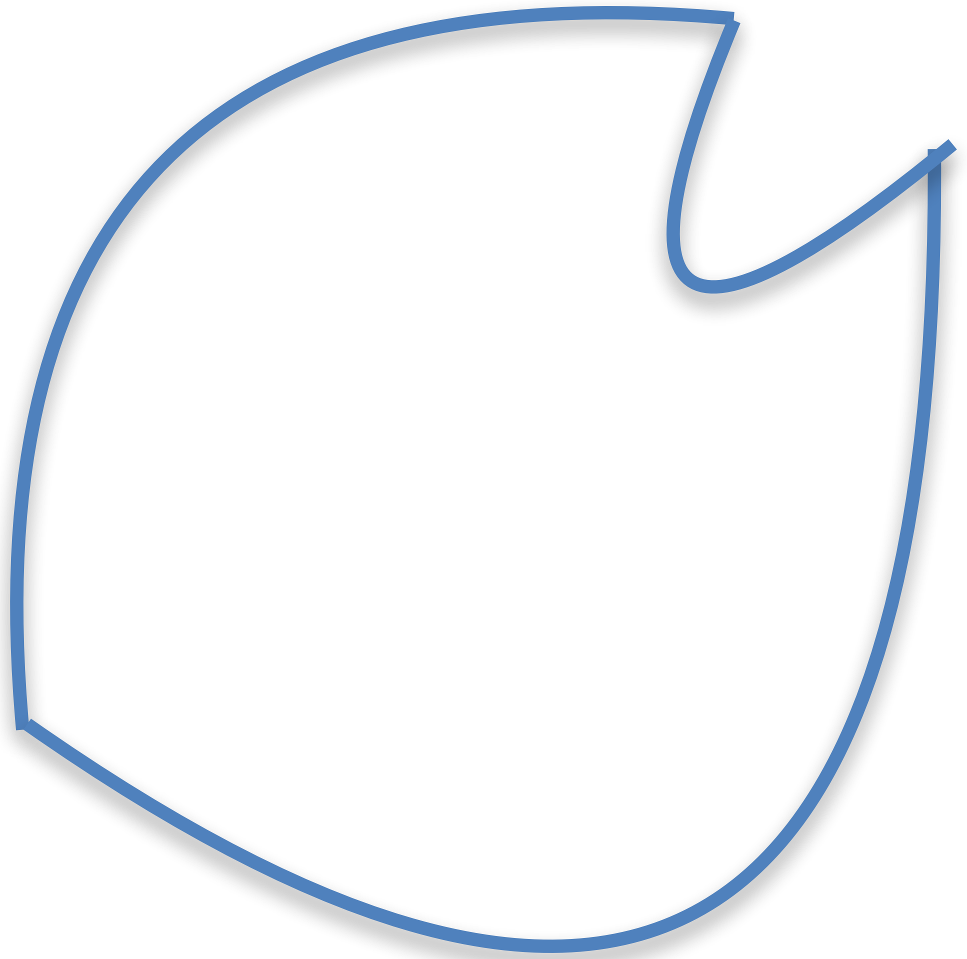
(b)

✗



(c)

✓



(d)

✗

Convexity of a cost function

What is a convex function?

Convexity of a cost function

What is a convex function?

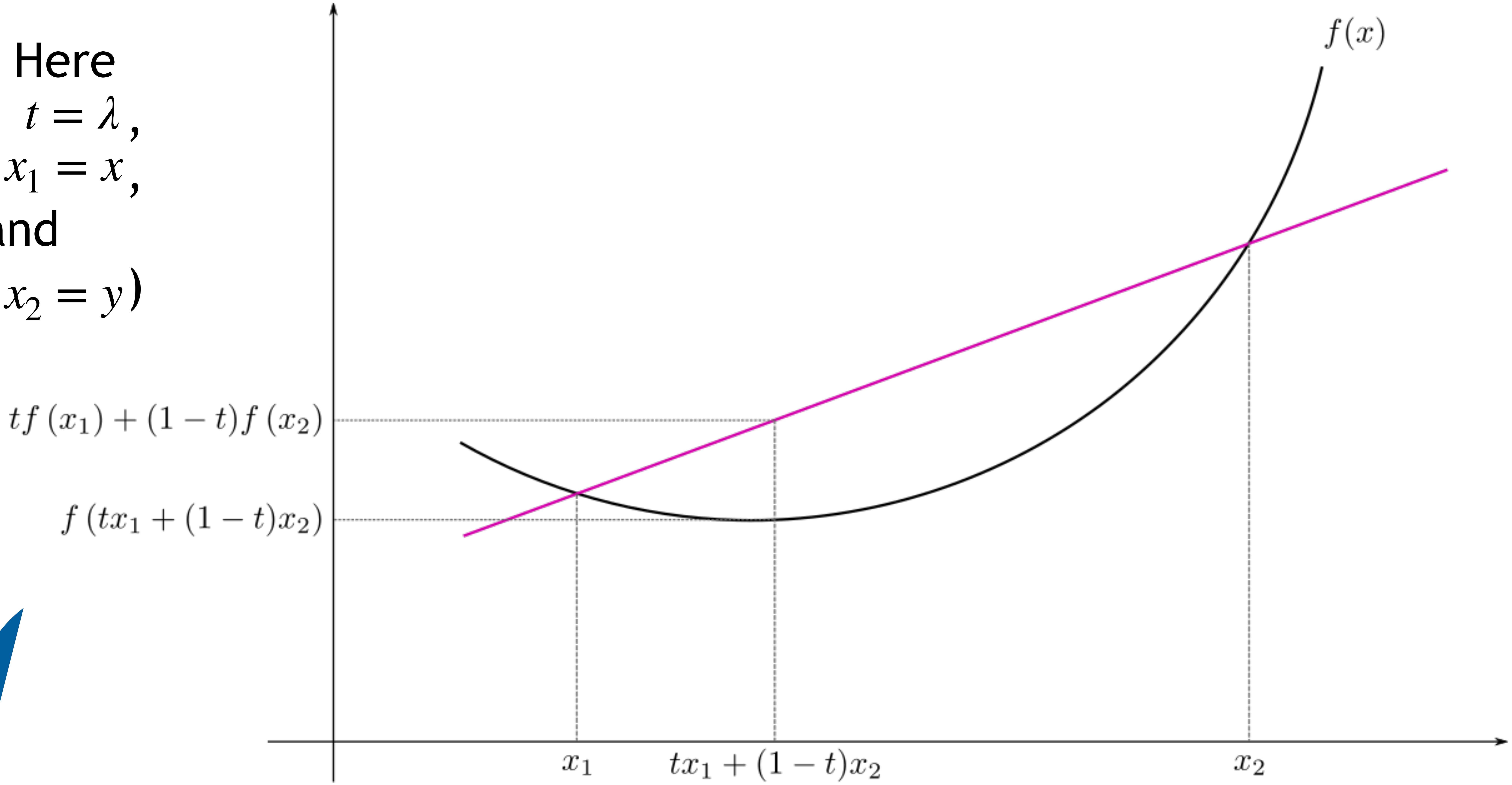
A function $f: C \to \mathbb{R}$ over a convex set C is called *convex* if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$$

is satisfied for all $x, y \in C$ and $\lambda \in [0,1]$.

Convexity of a cost function

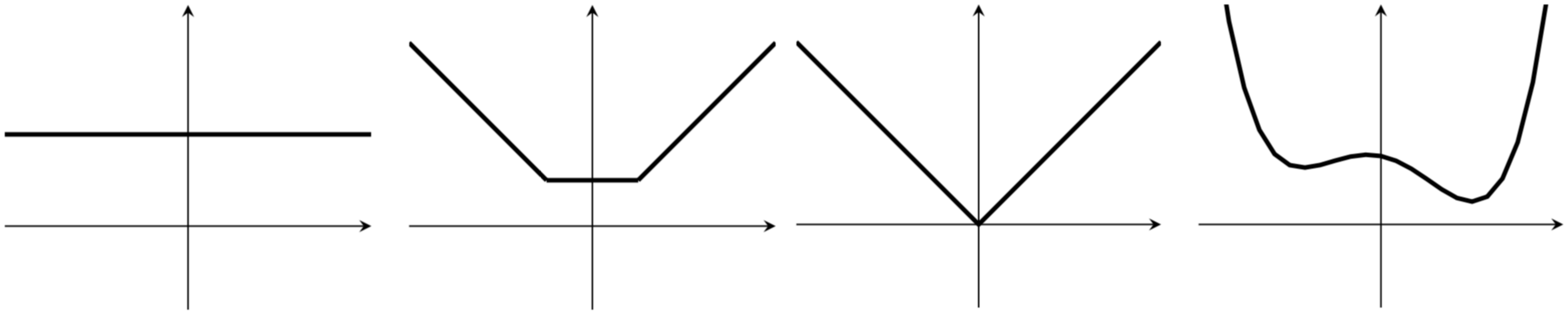
(Here
$t = \lambda$,
$x_1 = x$,
and
$x_2 = y$)



©Wikimedia commons

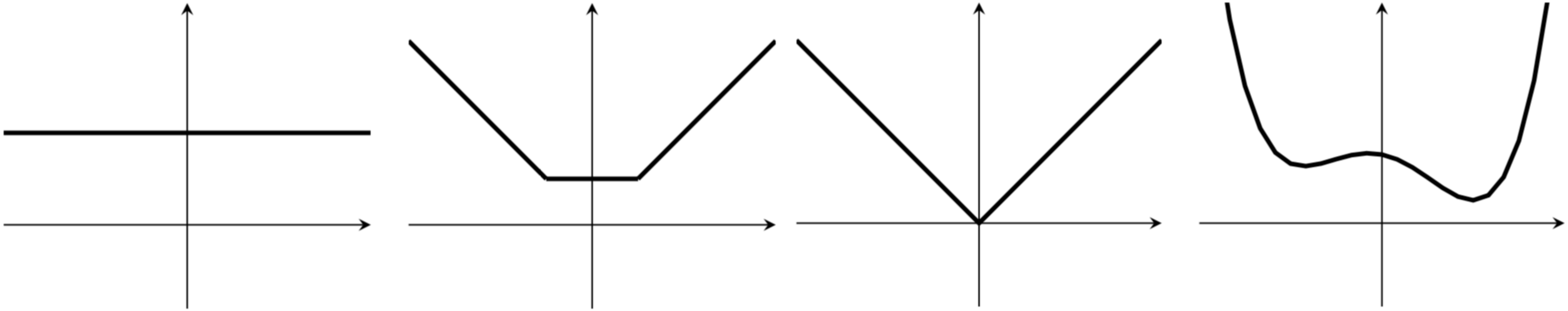
Convexity of a cost function

Examples:



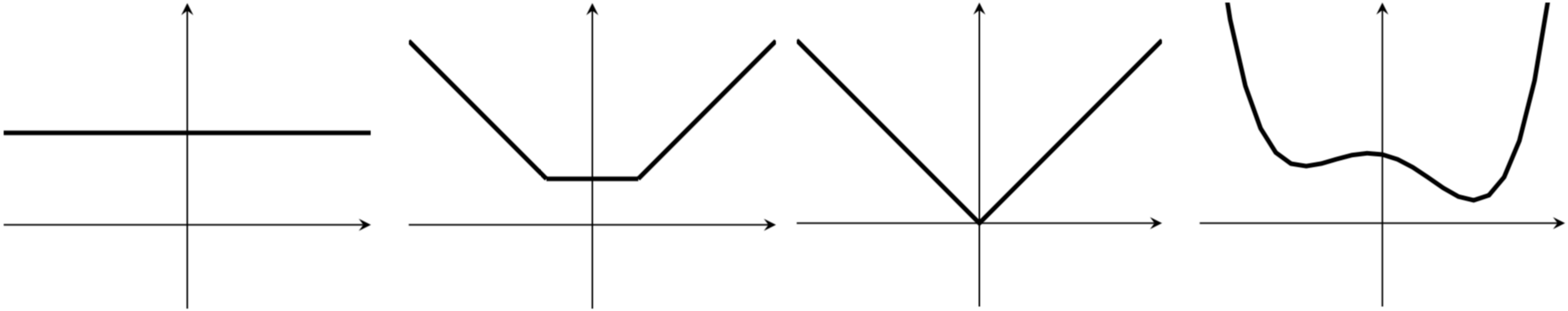
Convexity of a cost function

Examples:



Convexity of a cost function

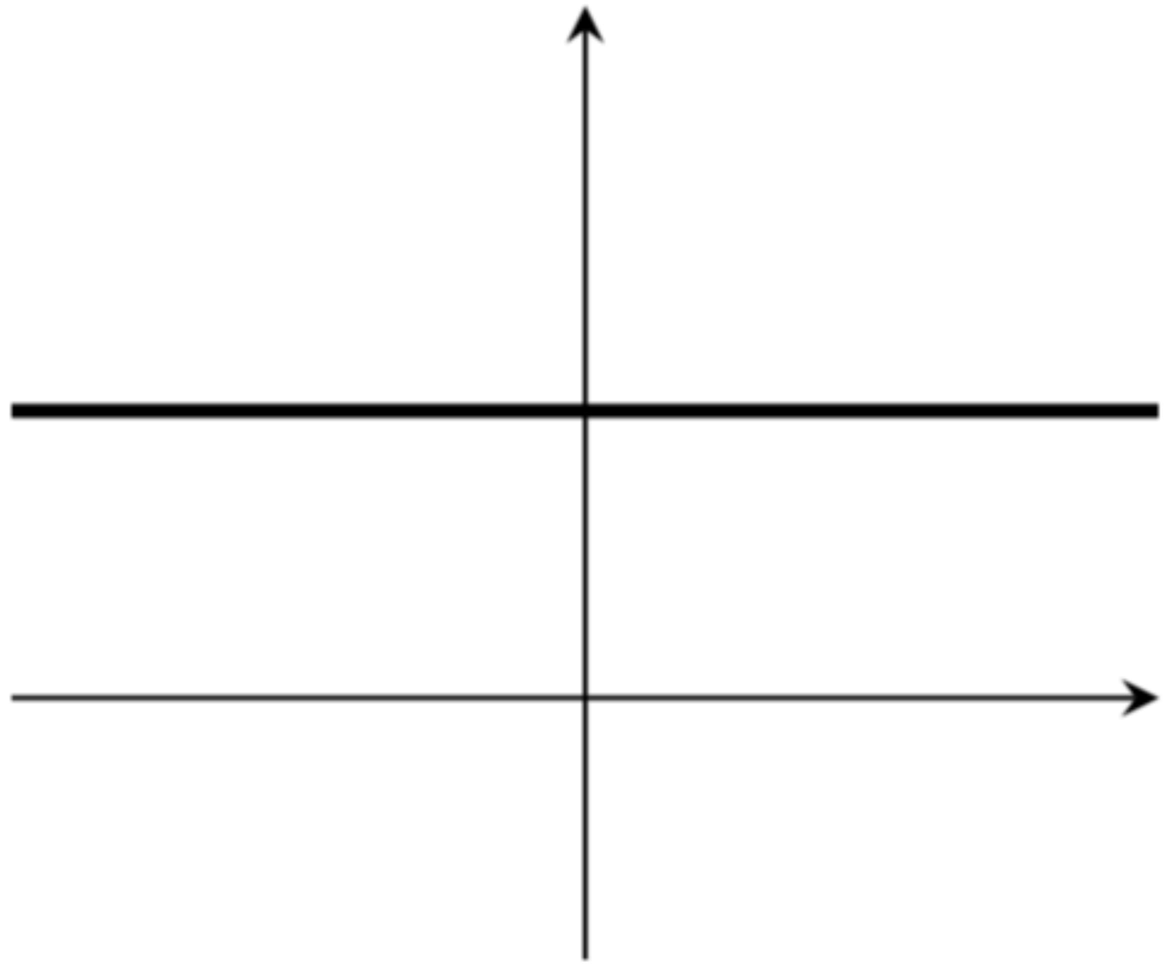
Examples:



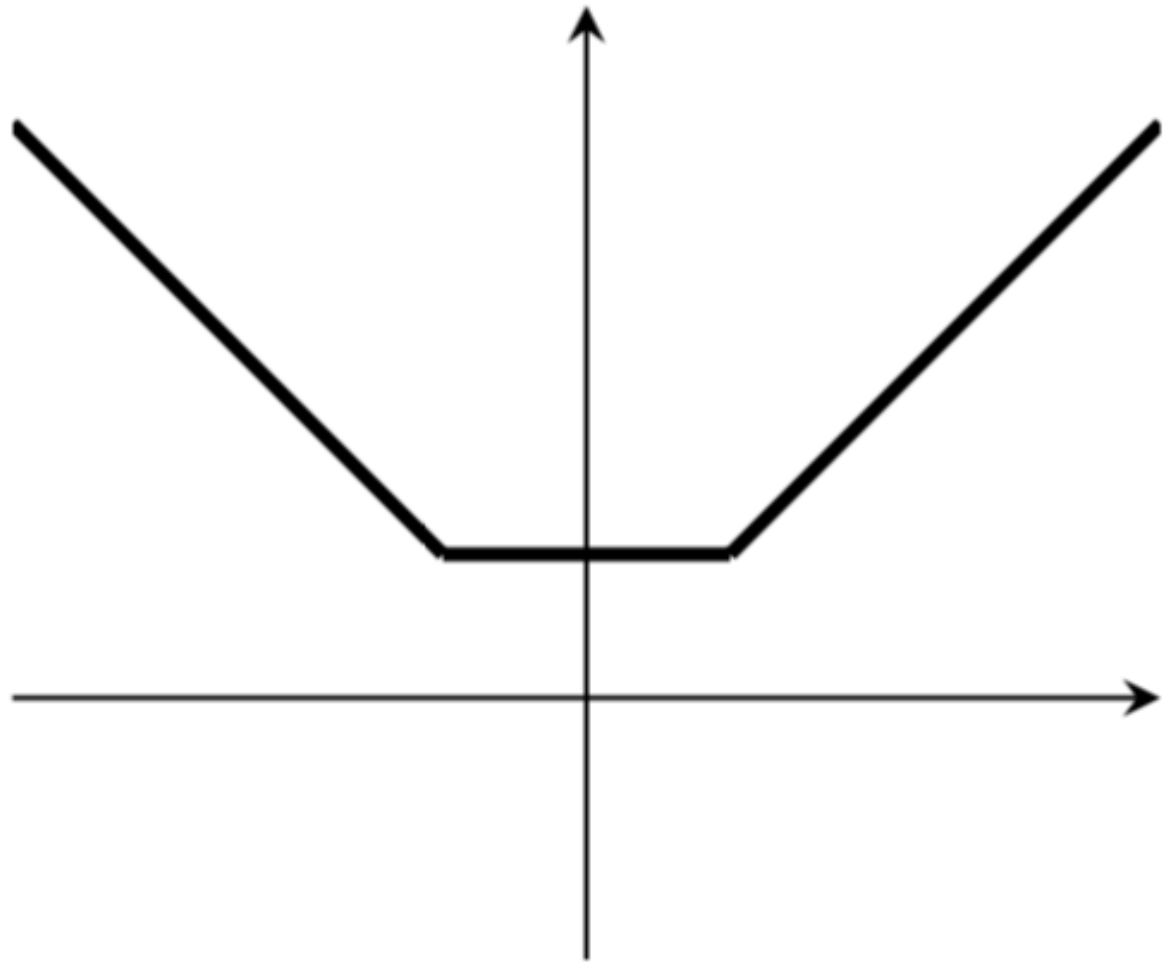
✓

Convexity of a cost function

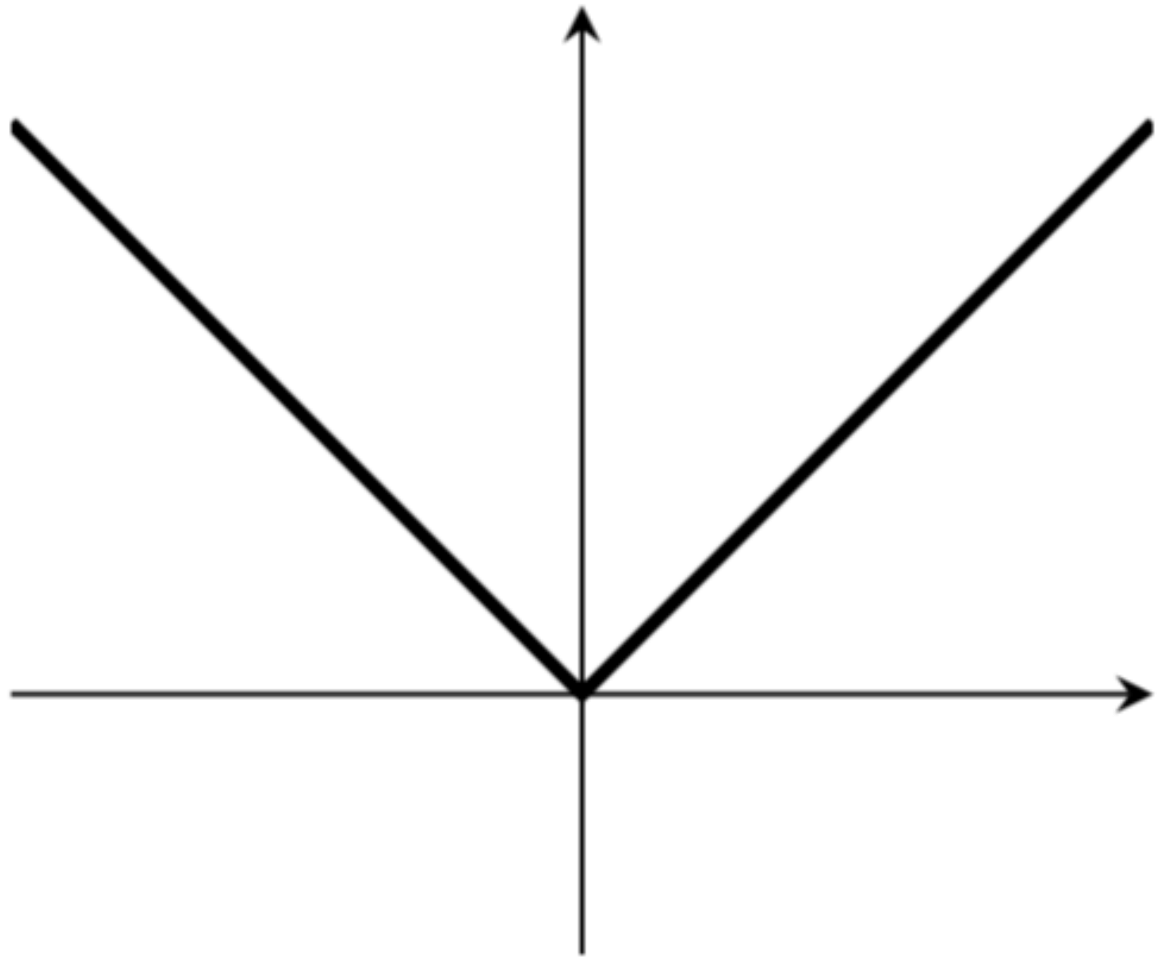
Examples:

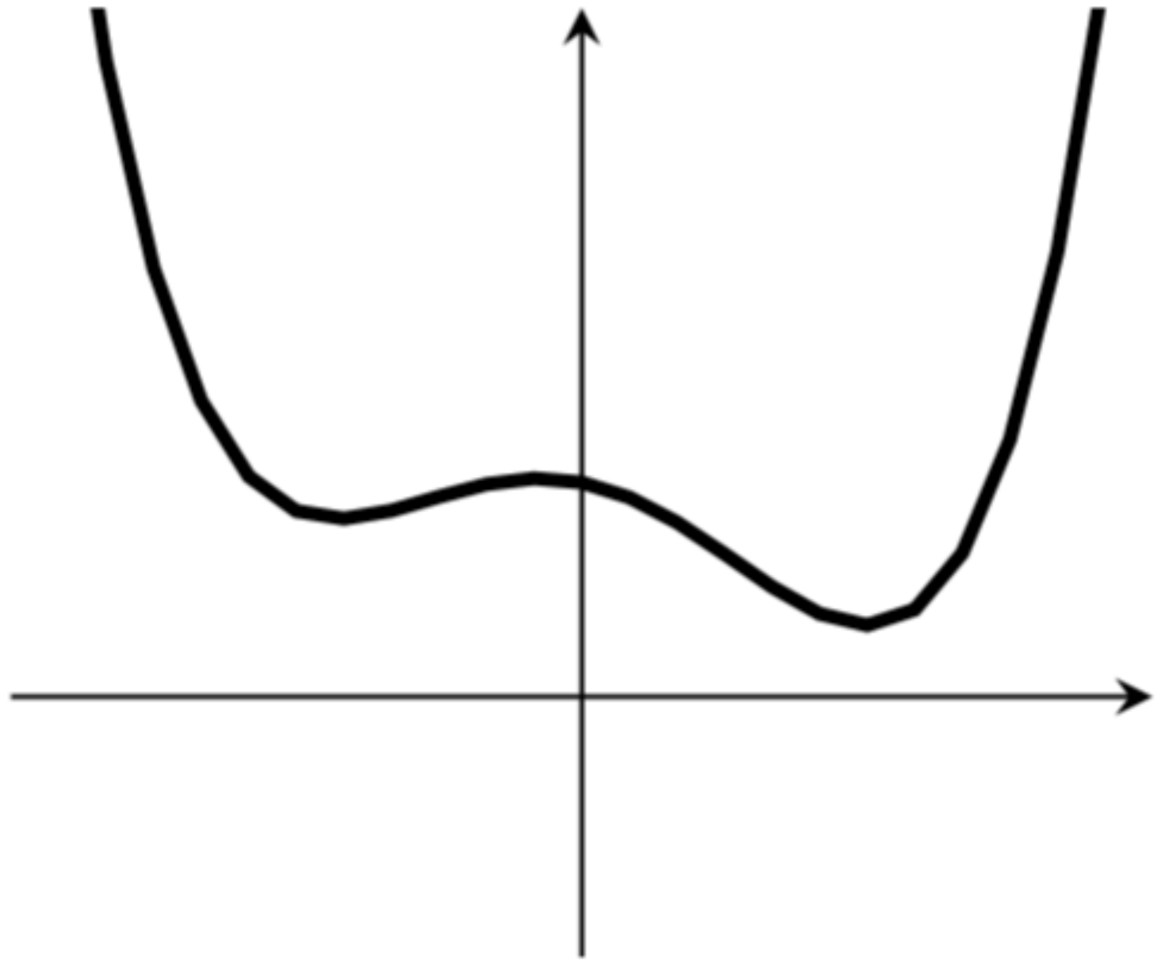


✓



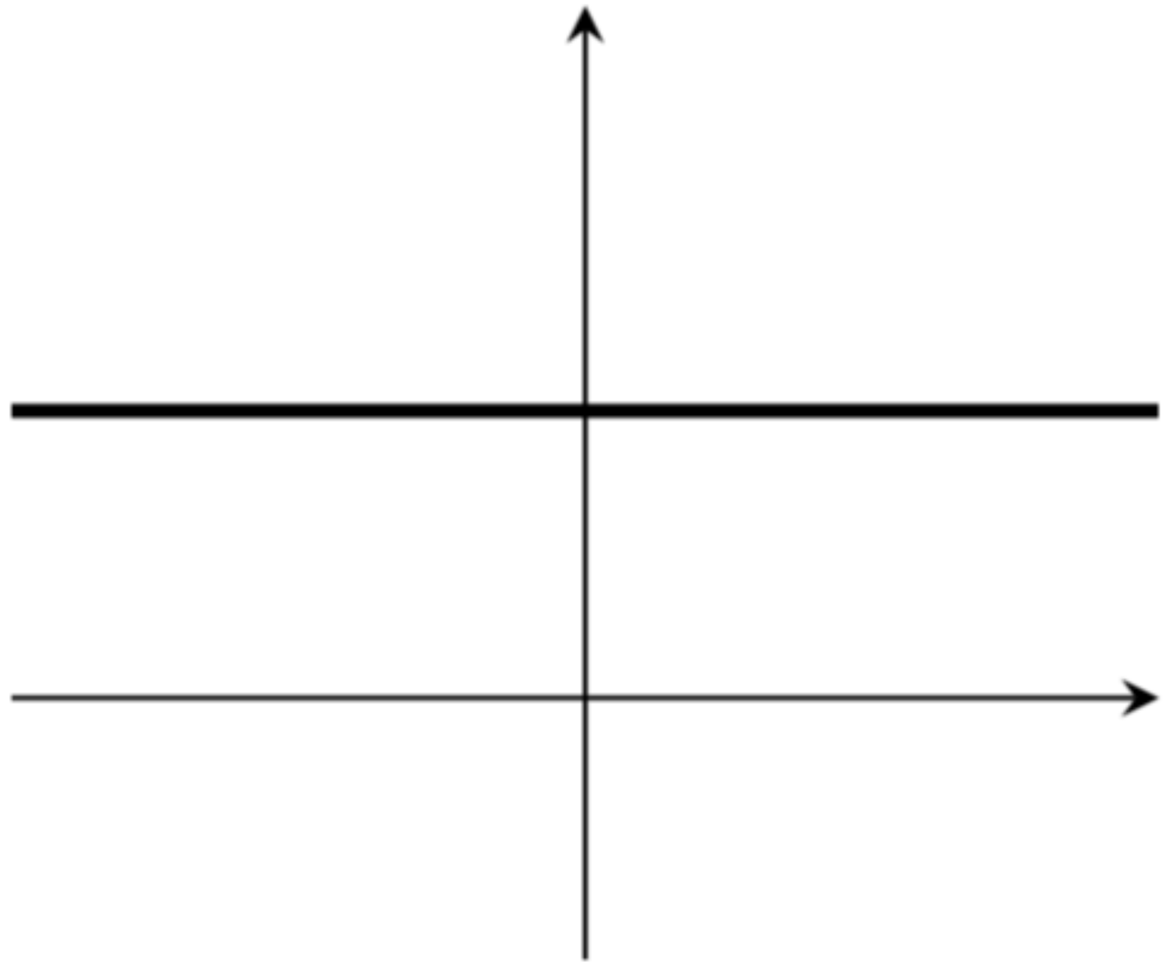
✓



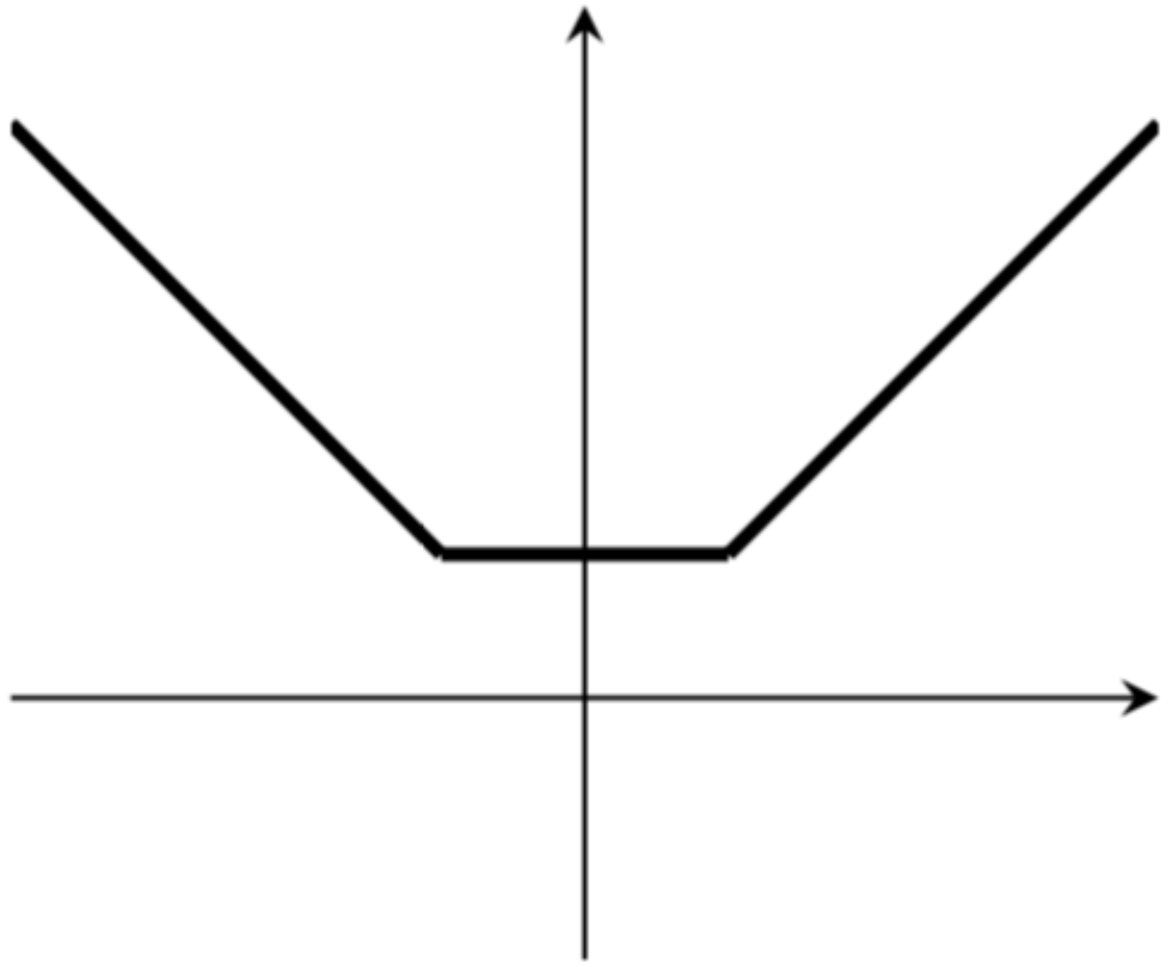


Convexity of a cost function

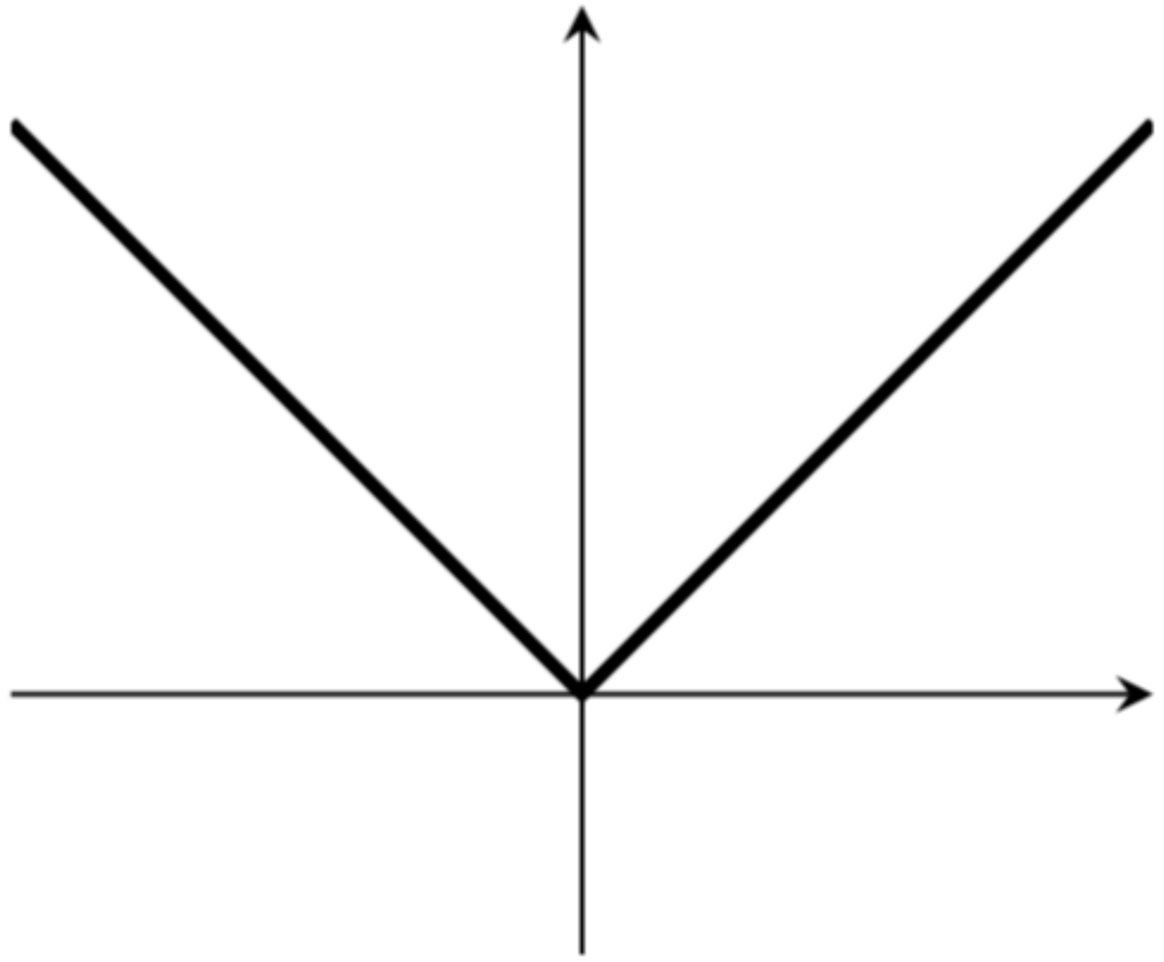
Examples:



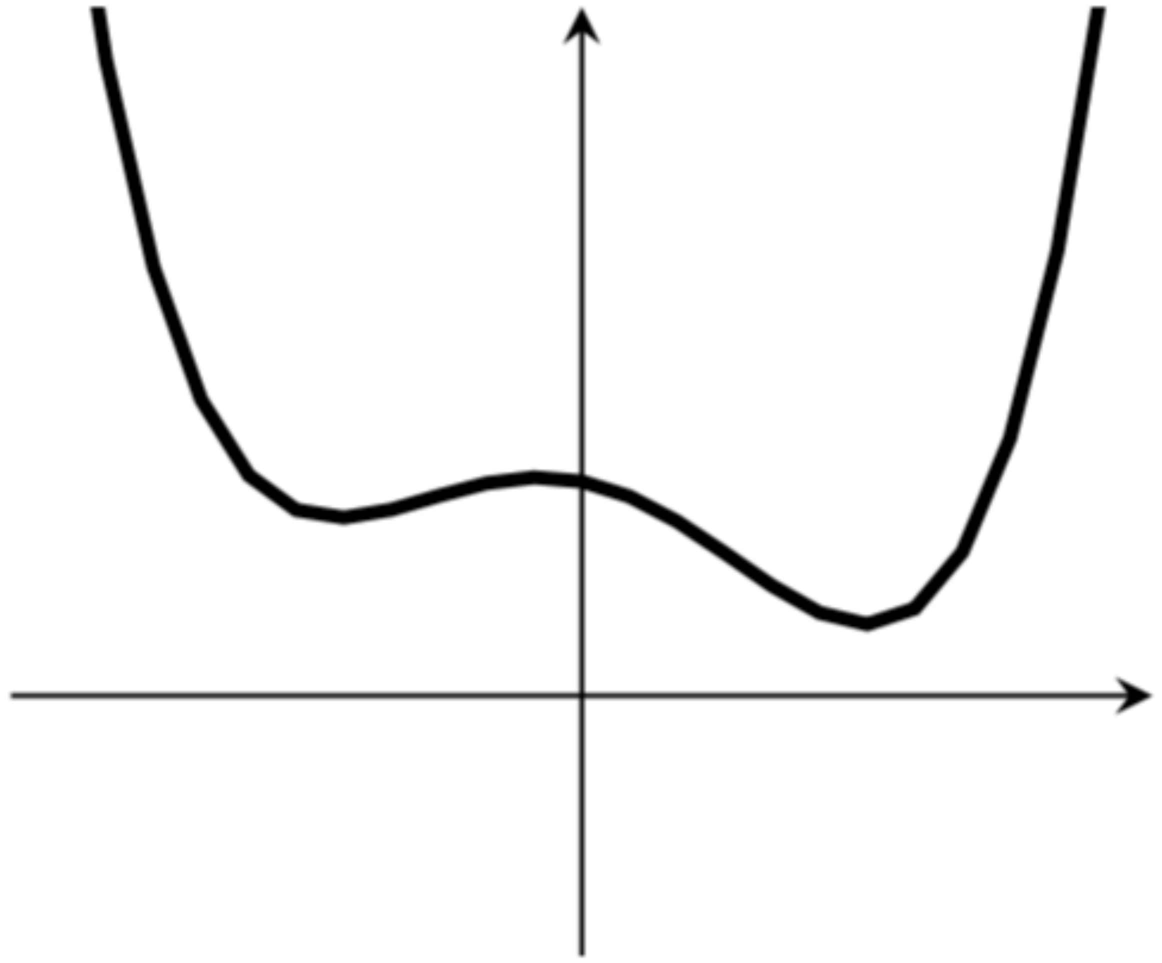
✓



✓

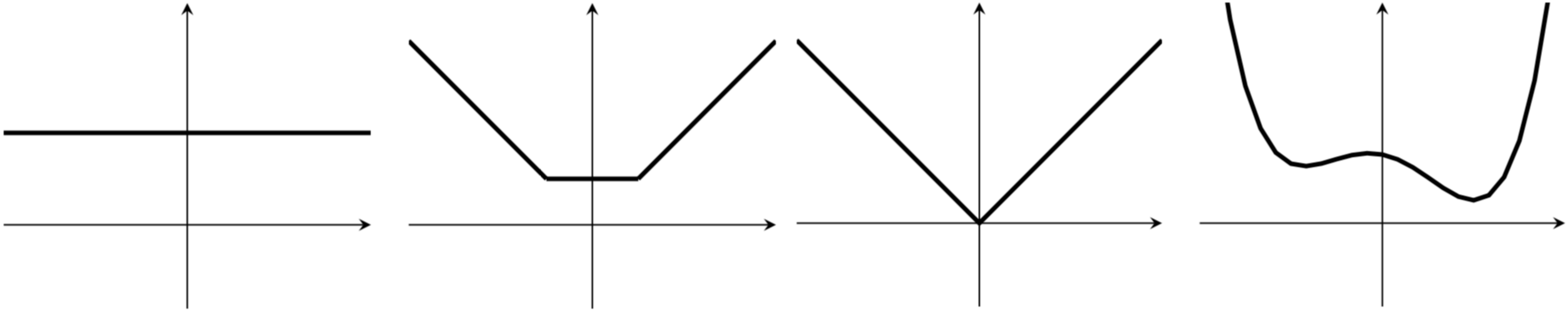


✓



Convexity of a cost function

Examples:



✓ ✓ ✓ ✗

Convexity of a cost function

Why is convexity useful?

Convexity of a cost function

Why is convexity useful?

Suppose $\hat{x}$ with $\nabla f(\hat{x}) = 0$, then

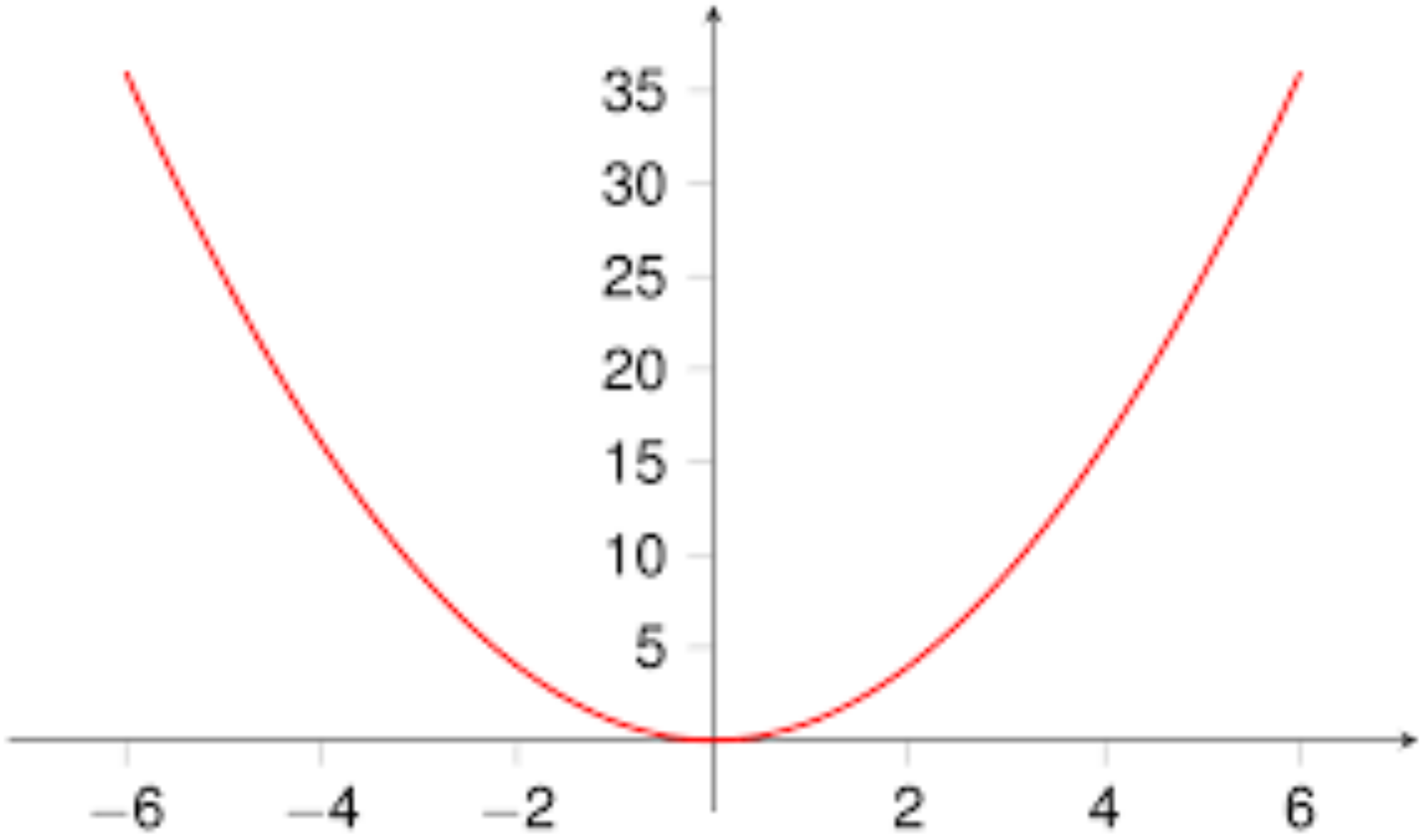
$$f(\hat{x}) \leq f(x) \quad \forall x \in C$$

Convexity of a cost function

Why is convexity useful?

Suppose $\hat{x}$ with $\nabla f(\hat{x}) = 0$, then

$$f(\hat{x}) \leq f(x) \quad \forall x \in C$$



$$f(x) = x^2$$

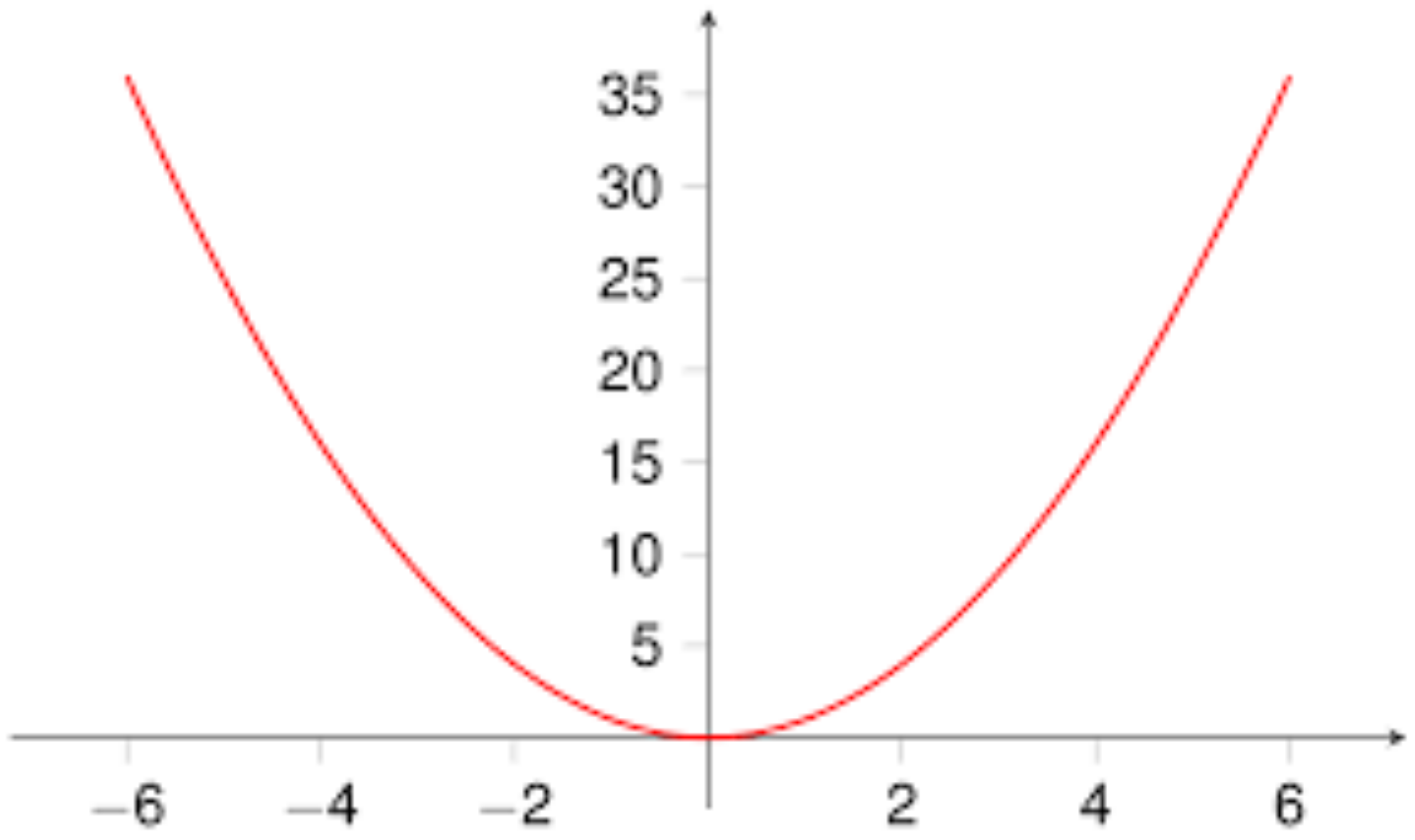
©Wikimedia commons

Convexity of a cost function

Why is convexity useful?

Suppose $\hat{x}$ with $\nabla f(\hat{x}) = 0$, then

$$f(\hat{x}) \leq f(x) \quad \forall x \in C$$



$f(x) = x^2$

©Wikimedia commons

Global minima can be determined by computing $\nabla f(\hat{x}) = 0$

Convexity of a cost function

Why is convexity useful?

Suppose $\hat{x}$ with $\nabla f(\hat{x}) = 0$, then

$$f(\hat{x}) \leq f(x) \quad \forall x \in C$$

Proof in 1D:

Convexity of a cost function

Why is convexity useful?

Suppose $\hat{x}$ with $\nabla f(\hat{x}) = 0$, then

$$f(\hat{x}) \leq f(x) \quad \forall x \in C$$

Proof in 1D: $f(\lambda x + (1-\lambda)\hat{x}) \leq \lambda f(x) + (1-\lambda) f(\hat{x})$

Convexity of a cost function

Why is convexity useful?

Suppose $\hat{x}$ with $\nabla f(\hat{x}) = 0$, then

$$f(\hat{x}) \leq f(x) \quad \forall x \in C$$

Proof in 1D:

$$f(\lambda x + (1 - \lambda)\hat{x}) \leq \lambda f(x) + (1 - \lambda) f(\hat{x})$$

$$\Leftrightarrow \quad f(\hat{x} + \lambda(x - \hat{x})) \leq f(\hat{x}) + \lambda(f(x) - f(\hat{x}))$$

Convexity of a cost function

Why is convexity useful?

Suppose $\hat{x}$ with $\nabla f(\hat{x}) = 0$, then

$$f(\hat{x}) \leq f(x) \quad \forall x \in C$$

Proof in 1D:

$$f(\lambda x + (1-\lambda)\hat{x}) \leq \lambda f(x) + (1-\lambda) f(\hat{x})$$

$$\Leftrightarrow \quad f(\hat{x} + \lambda(x - \hat{x})) \leq f(\hat{x}) + \lambda(f(x) - f(\hat{x}))$$

$$\Leftrightarrow \quad \frac{f(\hat{x} + \lambda(x - \hat{x})) - f(\hat{x})}{\lambda} \leq f(x) - f(\hat{x})$$

Convexity of a cost function

Why is convexity useful?

Suppose $\hat{x}$ with $\nabla f(\hat{x}) = 0$, then

$$f(\hat{x}) \leq f(x) \quad \forall x \in C$$

Proof in 1D:

$$f(\lambda x + (1-\lambda)\hat{x}) \leq \lambda f(x) + (1-\lambda) f(\hat{x})$$

$$\Leftrightarrow \quad f(\hat{x} + \lambda(x - \hat{x})) \leq f(\hat{x}) + \lambda(f(x) - f(\hat{x}))$$

$$\Leftrightarrow \quad \frac{f(\hat{x} + \lambda(x - \hat{x})) - f(\hat{x})}{\lambda} \leq f(x) - f(\hat{x})$$

$$\Rightarrow \quad \lim_{\lambda \to 0} \frac{f(\hat{x} + \lambda(x - \hat{x})) - f(\hat{x})}{\lambda} \leq f(x) - f(\hat{x})$$

Convexity of a cost function

Proof in 1D, continued:

$$\Rightarrow \quad \lim_{\lambda \to 0} \frac{f(\hat{x} + \lambda(x - \hat{x})) - f(\hat{x})}{\lambda} \le f(x) - f(\hat{x})$$

Convexity of a cost function

Proof in 1D, continued:

$$\Rightarrow \quad \underbrace{\lim_{\lambda \to 0} \frac{f(\hat{x} + \lambda(x - \hat{x})) - f(\hat{x})}{\lambda}}_{} \le f(x) - f(\hat{x})$$

$$\lim_{\lambda \to 0} \frac{f(\hat{x} + \lambda(x - \hat{x})) - f(\hat{x})}{\lambda} = \lim_{\lambda \to 0} \frac{\big(f(\hat{x} + \lambda(x - \hat{x})) - f(\hat{x})\big)(x - \hat{x})}{\lambda(x - \hat{x})} = f'(\hat{x})\,(x - \hat{x})$$

Convexity of a cost function

Proof in 1D, continued:

$$\Rightarrow \quad \underbrace{\lim_{\lambda\to 0} \frac{f(\hat{x}+\lambda(x-\hat{x}))-f(\hat{x})}{\lambda}}_{} \le f(x)-f(\hat{x})$$

$$\lim_{\lambda\to 0} \frac{f(\hat{x}+\lambda(x-\hat{x}))-f(\hat{x})}{\lambda} = \lim_{\lambda\to 0} \frac{\big(f(\hat{x}+\lambda(x-\hat{x}))-f(\hat{x})\big)(x-\hat{x})}{\lambda(x-\hat{x})} = f'(\hat{x})\,(x-\hat{x})$$

Hence, we conclude

$$f'(\hat{x})\,(x-\hat{x}) \le f(x)-f(\hat{x}) \qquad\qquad \forall x\in C$$

Convexity of a cost function

Proof in 1D, continued:

$$\Rightarrow \quad \underbrace{\lim_{\lambda \to 0} \frac{f(\hat{x} + \lambda(x - \hat{x})) - f(\hat{x})}{\lambda}}_{} \leq f(x) - f(\hat{x})$$

$$\lim_{\lambda \to 0} \frac{f(\hat{x} + \lambda(x - \hat{x})) - f(\hat{x})}{\lambda} = \lim_{\lambda \to 0} \frac{\big(f(\hat{x} + \lambda(x - \hat{x})) - f(\hat{x})\big)(x - \hat{x})}{\lambda(x - \hat{x})} = f'(\hat{x})\,(x - \hat{x})$$

Hence, we conclude

$$f'(\hat{x})\,(x - \hat{x}) \leq f(x) - f(\hat{x}) \qquad \forall x \in C$$

and $$f'(\hat{x}) = 0 \qquad \Rightarrow \qquad f(\hat{x}) \leq f(x) \qquad \forall x \in C$$

∎

Global minima

Given $$\mathrm{MSE}(\mathbf{w}) = \frac{1}{2s}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

Global minima

Given $$\text{MSE}(\mathbf{w}) = \frac{1}{2s}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

we obtain $$\hat{\mathbf{w}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\mathbf{y}$$

Global minima

Given $$\text{MSE}(\mathbf{w}) = \frac{1}{2s}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

we obtain $$\hat{\mathbf{w}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\mathbf{y}$$

by computing $$\nabla \text{MSE}(\hat{\mathbf{w}}) = 0$$

Global minima

Given $\text{MSE}(\mathbf{w}) = \frac{1}{2s}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$

we obtain $\hat{\mathbf{w}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\mathbf{y}$

by computing $\nabla \text{MSE}(\hat{\mathbf{w}}) = 0$

If MSE is convex, we have $\text{MSE}(\hat{\mathbf{w}}) \leq \text{MSE}(\mathbf{w}) \qquad \forall \mathbf{w} \in \mathbb{R}^n$

Global minima

Given $\mathsf{MSE}(\mathbf{w}) = \frac{1}{2s}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$

we obtain $\hat{\mathbf{w}} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\mathbf{T}\mathbf{y}$

by computing $\nabla \mathsf{MSE}(\hat{\mathbf{w}}) = 0$

If MSE is convex, we have $\mathsf{MSE}(\hat{\mathbf{w}}) \leq \mathsf{MSE}(\mathbf{w}) \qquad \forall \mathbf{w} \in \mathbb{R}^n$

Thus $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathsf{MSE}(\mathbf{w})$

Minimisers & the role of convexity

1. Why is computing

$$\hat{w} = \arg \min_{w \in \mathbb{R}^{d+1}} \mathsf{MSE}(w)$$

equivalent to solving

$$\nabla \mathsf{MSE}(\hat{w}) = 0 \quad ?$$

2. Does a solution $\hat{w}$ always exist?

3. Is the solution $\hat{w}$ unique?

Minimisers & the role of convexity

1. Why is computing

$$\hat{w} = \arg \min_{w \in \mathbb{R}^{d+1}} \mathsf{MSE}(w)$$

equivalent to solving

$$\nabla \mathsf{MSE}(\hat{w}) = 0 \quad ?$$

2. Does a solution $\hat{w}$ always exist?

3. Is the solution $\hat{w}$ unique?

What is left to show?

Exercise:

Show that MSE is convex!

(for linear regression model)

TUTORIAL ON FRIDAY

We will discuss the solutions of Coursework 1

To make the most of these tutorials, attempt completing the coursework before!