MTH786: Machine Learning with Python

Week 2: Linear and polynomial regressions. Convexity. Regularisation.

m.poplavskyi@qmul.ac.uk

14/06/2021

Linear regression

What is a regression?

The main task of **regression analysis** consist of finding a mapping $f : \mathbb{R}^n \to \mathbb{R}^m$ between s known input arguments $\{\mathbf{x}^{(j)}\}_{j=1}^{s}$, with $\mathbf{x}^{(j)} \in \mathbb{R}^n$ for all $1 \leq j \leq s$, onto a collection of s output elements $\{\mathbf{y}^{(j)}\}_{j=1}^{s}$, with $\mathbf{y}^{(j)} \in \mathbb{R}^m$ for all $1 \leq j \leq s$ such that

$$f\left(\mathbf{x}^{(j)}\right) \approx \mathbf{y}^{(j)}, \quad \forall j \in \{1,\ldots,s\}\,.$$

How to find a suitable function $f(x)$?

1. Consider a class of functions $\mathscr{F}$ parametrised by a set of parameters $\overrightarrow{w}$.
2. Use the training data provided to find an optimal value of parameters $\overrightarrow{w}$.
3. Validate the result using validation data.

What does optimal mean? ⟶ **smallest error**

m.poplavskyi@qmul.ac.uk

Example: linear regression

How a linear function $f : \mathbb{R}^n \to \mathbb{R}^m$ could be parametrised?

$$f_{\mathbf{W}}(\mathbf{x}) = \mathbf{w}^{(0)} + \sum_{j=1}^{n} \mathbf{w}^{(j)} x_j,$$

where $\mathbf{w}^{(j)} \in \mathbb{R}^m$ for all $0 \leq j \leq n$ are vector-valued unknown weights.

Remark: unless other agreement is discussed we use superscript for object numbering and subscript for vector indexing. Vectors and matrices are **bold.**

Note: The sum in above goes over vector coordinates not vector labels, i.e.

$$f_{\mathbf{W}}\left(\mathbf{x}^{(j)}\right) = \mathbf{w}^{(0)} + \sum_{k=1}^{n} \mathbf{w}^{(k)} x_k^{(j)}, \quad \left(f_{\mathbf{W}}\left(\mathbf{x}^{(j)}\right)\right)_{\ell} = w_{\ell}^{(0)} + \sum_{k=1}^{n} w_{\ell}^{(k)} x_k^{(j)}.$$

In the case of our main interest $m = 1$ and one can write

$f_{\mathbf{W}}(\mathbf{x}) = \left\langle \mathring{\mathbf{x}}, \mathbf{W} \right\rangle = \mathring{\mathbf{x}}^T \mathbf{W}$, where $\mathring{\mathbf{x}} = \left(1, x_1, \ldots, x_n\right)^T, \mathbf{W} = \left(w_0, w_1, \ldots, w_n\right)^T.$

m.poplavskyi@qmul.ac.uk

Example: linear regression

$$f_{\mathbf{W}}(\mathbf{x}) = \left\langle \mathring{\mathbf{x}}, \mathbf{W} \right\rangle = \mathring{\mathbf{x}}^T \mathbf{W}, \text{ where } \mathring{\mathbf{x}} = \left(1, x_1, \ldots, x_n\right)^T, \mathbf{W} = \left(w_0, w_1, \ldots, w_n\right)^T.$$

Can we fit the data?

Question: for the training data $\{(\mathbf{x}^{(j)}, \mathbf{y}^{(j)})\}_{j=1}^{s}$ find $f_{\mathbf{W}}$ such that $f_{\mathbf{W}}\left(\mathbf{x}^{(j)}\right) = \mathbf{y}^{(j)}$

Example: let $s = 3$ and $n = 2$, then

$$\begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} \\ 1 & x_1^{(3)} & x_2^{(3)} \end{pmatrix} \begin{pmatrix} w_1^{(0)} \\ w_1^{(1)} \\ w_1^{(2)} \end{pmatrix} = \begin{pmatrix} y_1^{(1)} \\ y_1^{(2)} \\ y_1^{(3)} \end{pmatrix}$$

Is the system uniquely solvable? ⟶ **Yes, for a "good" input**

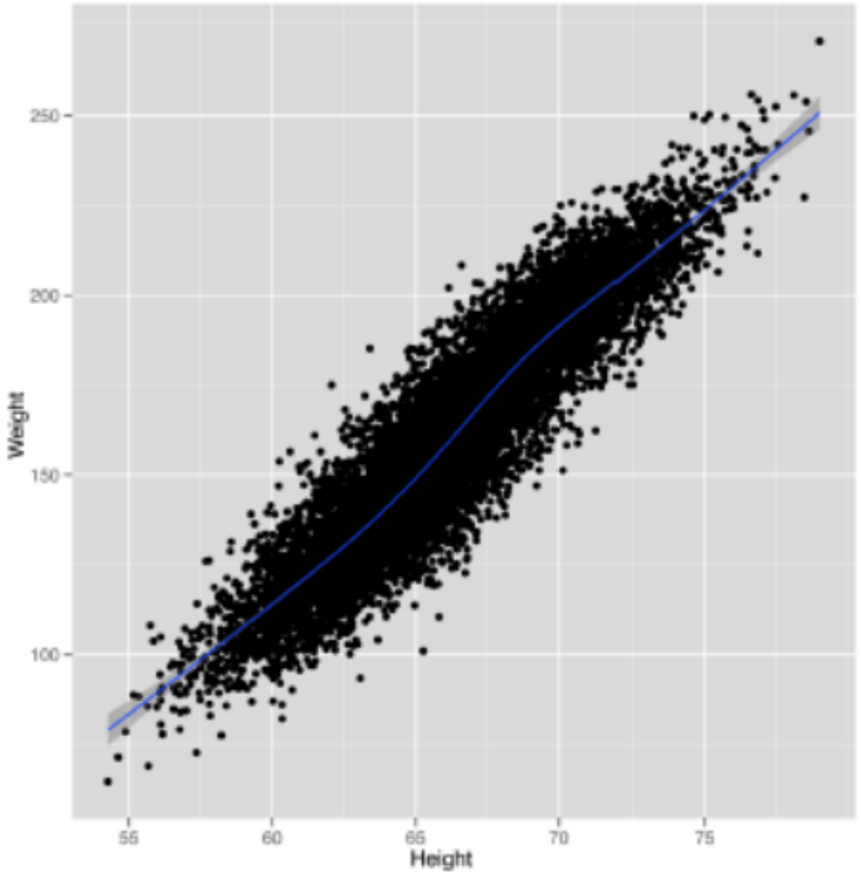
What would happen in general case? ⟶ sm **eqs,** $(n+1)m$ **vars**

Only if $s = n + 1$

m.poplavskyi@qmul.ac.uk

Example: linear regression

Only if $s = n + 1$. Is it realistic?



Example: Prediction of a person's weight from its height by using linear regression.

Dimensionality: $n = 1$ and $s \gg 1$.
Question: how to measure an error of approximation when exact fit is not possible?

Idea: we need to measure how far is the prediction out of expected value in average.

Remind us the *variance.* For $n = 1$

$$\mathrm{MSE}\,(\mathbf{W}) = \frac{1}{2s}\sum_{j=1}^{s}\left|f_{\mathbf{W}}\left(x^{(j)}\right) - y^{(j)}\right|^2$$

Mean-squared error

m.poplavskyi@qmul.ac.uk

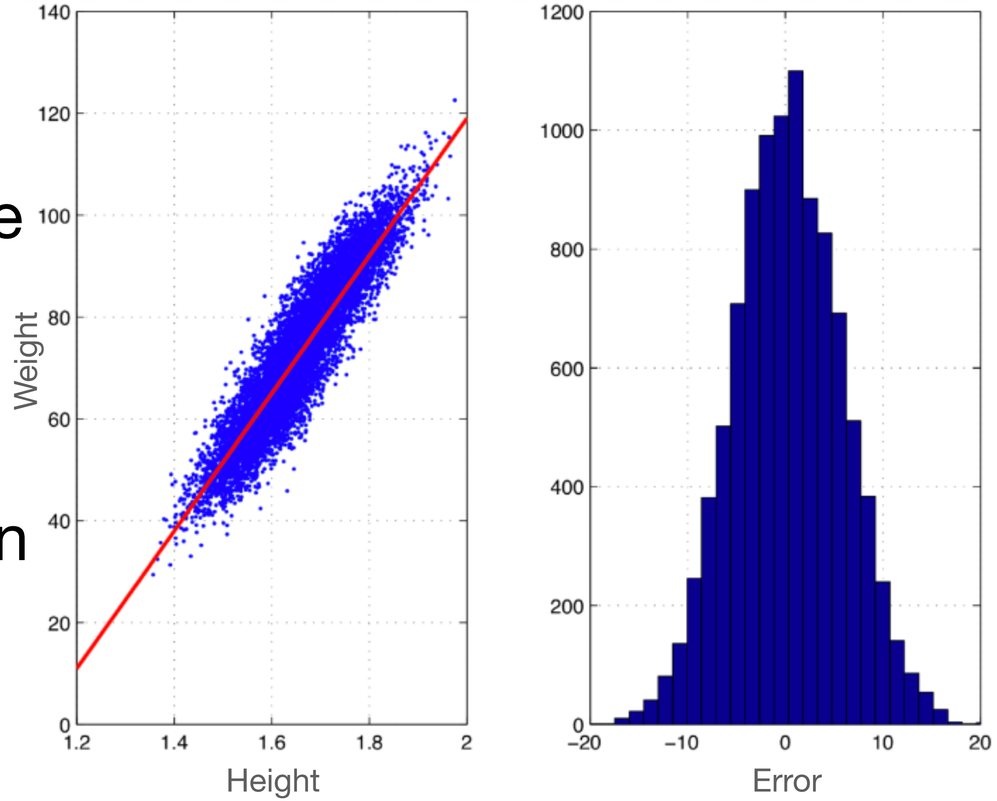
Statistical motivation of MSE

Question: how did we come up with MSE?

Motivation: let us consider the distribution of errors $\varepsilon^{(j)} := f_{\mathbf{W}}\left(\mathbf{x}^{(j)}\right) - \mathbf{y}^{(j)}$. If we believe the model has been chosen right then $\varepsilon^{(j)}$ are r.v. Moreover, $\{\varepsilon^{(1)}, \ldots, \varepsilon^{(s)}\}$ are independent.

Observation: the distribution is Gaussian.

Assumption: $\{\varepsilon^{(1)}, \ldots, \varepsilon^{(s)}\}$ are i.i.d. Gaussian random variables with mean 0 and variance σ^2, i.e.



$$\rho\left(\varepsilon^{(j)}|0,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{\left(\varepsilon^{(j)}\right)^2}{2\sigma^2}}$$

Problem: find optimal values of $\mathbf{W}$ to maximise the likelihood of sampling $\mathbf{y}^{(j)}$.

m.poplavskyi@qmul.ac.uk

Statistical motivation of MSE

Remark: below we work with $m = 1$. For $m > 1$ the method still works.

Likelihood

$$\rho(\varepsilon^{(1)}, \varepsilon^{(2)}, \ldots, \varepsilon^{(s)} | 0, \sigma^2) = \prod_{j=1}^{s} (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{\left(\varepsilon^{(j)}\right)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-\frac{s}{2}} \prod_{j=1}^{s} e^{-\frac{\left(\mathbf{y}^{(j)} - f\left(\mathbf{x}^{(j)}\right)\right)^2}{2\sigma^2}}$$

$$= \rho\left(\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(s)} | f\left(\mathbf{x}^{(1)}\right), \ldots, f\left(\mathbf{x}^{(s)}\right), \sigma^2\right)$$

Maximisation

$$\hat{W} = \arg\min_{W \in \mathbb{R}^{n+1}} \{-\log(\rho(\mathbf{Y}|f_{\mathbf{W}}(\mathbf{X}), \sigma^2))\} = \arg\min_{W \in \mathbb{R}^{n+1}} \left\{-\log\left(\prod_{j=1}^{s} \rho(\mathbf{y}^{(j)}|f_{\mathbf{W}}\left(\mathbf{x}^{(j)}\right), \sigma^2)\right)\right\} = \arg\min_{W \in \mathbb{R}^{n+1}} \left\{-\sum_{j=1}^{s} \log \rho(\mathbf{y}^{(j)}|f_{\mathbf{W}}\left(\mathbf{x}^{(j)}\right), \sigma^2)\right\}$$

$$= \arg\min_{W \in \mathbb{R}^{n+1}} \left\{\frac{1}{2\sigma^2} \sum_{j=1}^{s} \left(\mathbf{y}^{(j)} - f_{\mathbf{W}}\left(\mathbf{x}^{(j)}\right)\right)^2 + \frac{s}{2} \log(2\pi\sigma^2)\right\} = \arg\min_{W \in \mathbb{R}^{n+1}} \left\{\frac{2s}{\sigma^2} \mathrm{MSE}(\mathbf{W}) + \frac{s}{2} \log(2\pi\sigma^2)\right\} = \arg\min_{W \in \mathbb{R}^{n+1}} \{\mathrm{MSE}(\mathbf{W})\}$$

Question 1: How to find a minimiser?

Question 2: Does the minimiser exist?

Question 3: Is the minimiser unique?

m.poplavskyi@qmul.ac.uk

Example: 1+1 dimension

Problem: For a given set of input-output pairs $\left\{ \left(x^{(j)}, y^{(j)}\right) \right\}_{j=1}^{s}$ find a linear function $f_{\mathbf{W}} : \mathbb{R} \to \mathbb{R}$ such that $\mathrm{MSE}\,(\mathbf{W})$ is minimised over all possible $\mathbf{W}$.

Zero approximation

Let $f_{\mathbf{W}}(x) = w^{(0)}$ for all $x \in \mathbb{R}$. Then

$$\mathrm{MSE}\left(w^{(0)}\right) = \frac{1}{2s} \sum_{j=1}^{s} \left(w^{(0)} - y^{(j)}\right)^2$$

$$\frac{\partial \mathrm{MSE}\left(w^{(0)}\right)}{\partial w^{(0)}} = \frac{1}{s} \sum_{j=1}^{s} \left(w^{(0)} - y^{(j)}\right) \stackrel{?}{=} 0$$

$$\hat{w}_0 = \frac{y^{(1)} + \ldots + y^{(s)}}{s} =: \overline{y}$$

Real linear approximation

Let $f_{\mathbf{W}}(x) = w^{(0)} + w^{(1)}x$, $x \in \mathbb{R}$. Then

$$\mathrm{MSE}\,(\mathbf{W}) = \frac{1}{2s} \sum_{j=1}^{s} \left(w^{(0)} + w^{(1)}x^{(j)} - y^{(j)}\right)^2$$

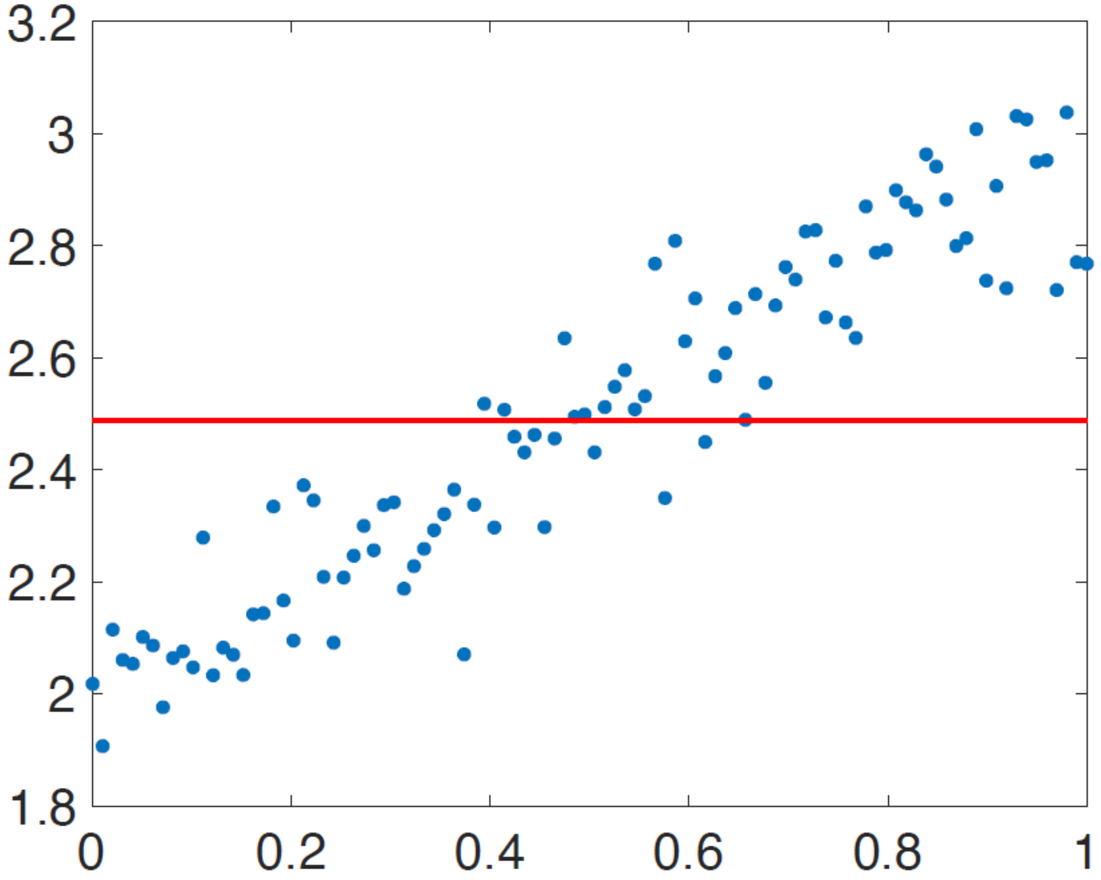
$$\frac{\partial \mathrm{MSE}\,(\mathrm{W})}{\partial w^{(0)}} = \frac{1}{s} \sum_{j=1}^{s} \left(w^{(0)} + w^{(1)}x^{(j)} - y^{(j)}\right) \stackrel{?}{=} 0$$

$$\frac{\partial \mathrm{MSE}\,(\mathrm{W})}{\partial w^{(1)}} = \frac{1}{s} \sum_{j=1}^{s} x^{(j)} \left(w^{(0)} + w^{(1)}x^{(j)} - y^{(j)}\right) \stackrel{?}{=} 0$$

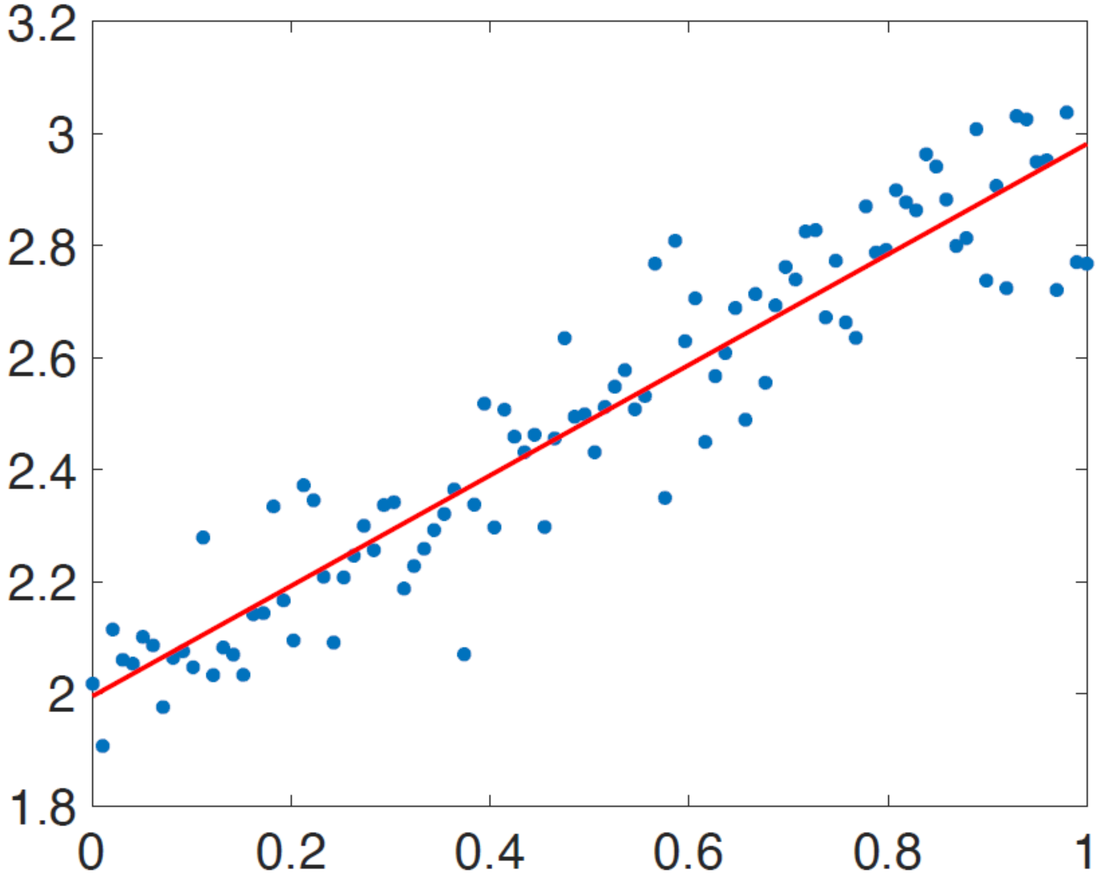
$$\hat{w}^{(0)} = \frac{\overline{y}\,\overline{x^2} - \overline{x}\,\overline{xy}}{\overline{x^2} - \overline{x}^2}, \qquad \hat{w}^{(1)} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}$$

m.poplavskyi@qmul.ac.uk

Example: 1+1 dimension



$\hat{w}^{(0)} = 2.4489$



$\hat{w}^{(0)} = 1.9962$
$\hat{w}^{(1)} = 0.9854$

m.poplavskyi@qmul.ac.uk

Example: $n+1$ dimensions

Problem: For a given set of input-output pairs $\left\{\left(\mathbf{x}^{(j)}, y^{(j)}\right)\right\}_{j=1}^{s}$ find a linear function $f_{\mathbf{W}}: \mathbb{R}^n \to \mathbb{R}$ such that $\mathrm{MSE}(\mathbf{W})$ is minimised over all possible $\mathbf{W}$.

Let $f_{\mathbf{W}}(\mathbf{x}) = w^{(0)} + \sum_{j=1}^{n} w^{(j)}\mathbf{x}_j = \mathring{\mathbf{x}}^T\mathbf{W}$, where $\mathbf{W} = \left(w^{(0)}, \ldots, w^{(n)}\right)^T$.

$$\mathrm{MSE}(\mathbf{W}) = \frac{1}{2s}\sum_{j=1}^{s}\left(\left[\mathbf{X}\mathbf{W}\right]_j - y^{(j)}\right)^2 = \frac{1}{2s}\left\|\mathbf{X}\mathbf{W} - \mathbf{Y}\right\|^2.$$

$$\nabla\mathrm{MSE}(\hat{\mathbf{W}}) = 0 \Rightarrow \mathbf{X}^T\mathbf{X}\hat{\mathbf{W}} = \hat{\mathbf{X}}^T\mathbf{Y}. \text{ (Exercise!)}$$

$$\hat{\mathbf{W}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}.$$

m.poplavskyi@qmul.ac.uk

Discussion of further development

Question 1: Are there any other cost-functions except of $\mathrm{MSE}\,(\mathbf{W})$?
Answer 1: Indeed there is a wide variety of cos-functions, for example in the case of $m = 1$ on can consider mean absolute error (MAE)

$$\mathrm{MAE}\,(\mathrm{W}) = \frac{1}{2s}\sum_{j=1}^{s}\left|f\left(\mathbf{x}^{(j)}\right) - y^{(j)}\right|.$$

+: robust to outliers (see Assignment 2) **-:** non-differentiable function

Question 2: How the same approach would work when allowing nonlinearity?
Answer 2: We can use the same approach for nonlinear functions $f\,(\mathbf{W})$.

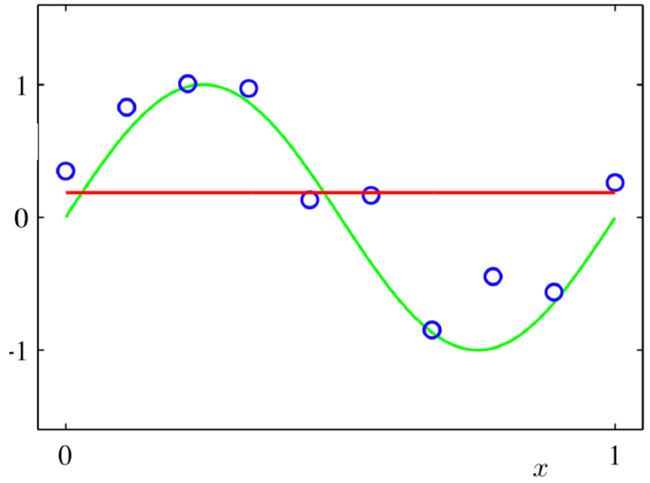
Question 3: Whether the minimiser we have found is indeed a minimiser?
Answer 3: It can be checked via evaluation of higher derivatives.

m.poplavskyi@qmul.ac.uk

Polynomial regression

Polynomial regression

In this section we work in $1+1$ dimensions only. Data: $\left\{x^{(j)}, y^{(j)}\right\}_{j=1}^{s}$.



The green curve is the plot of a function. Blue dots are points from the curve perturbed by some random forces. Red line is a line of linear regression. **It does not fit well.**

How to adapt a function $f(x)$?

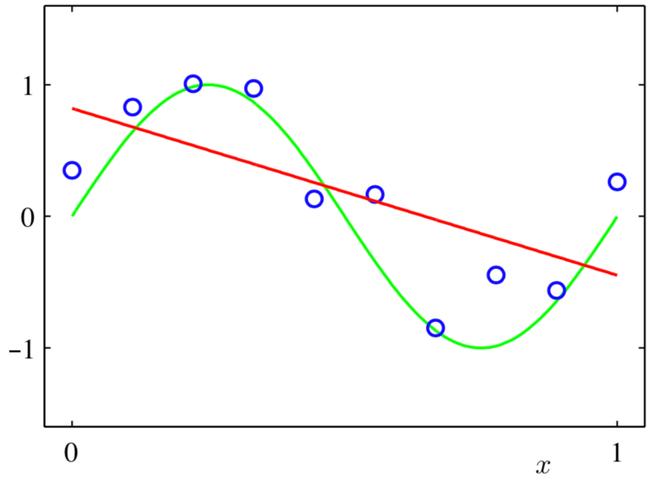
The next simplest class of functions are polynomials, i.e.

$$f_{\mathbf{W}}(x) = w^{(0)} + w^{(1)}x + \ldots + w^{(d)}x^d,$$

for some unknown vector $\mathbf{W} = \left(w^{(0)}, \ldots, w^{(d)}\right)^T$. Let

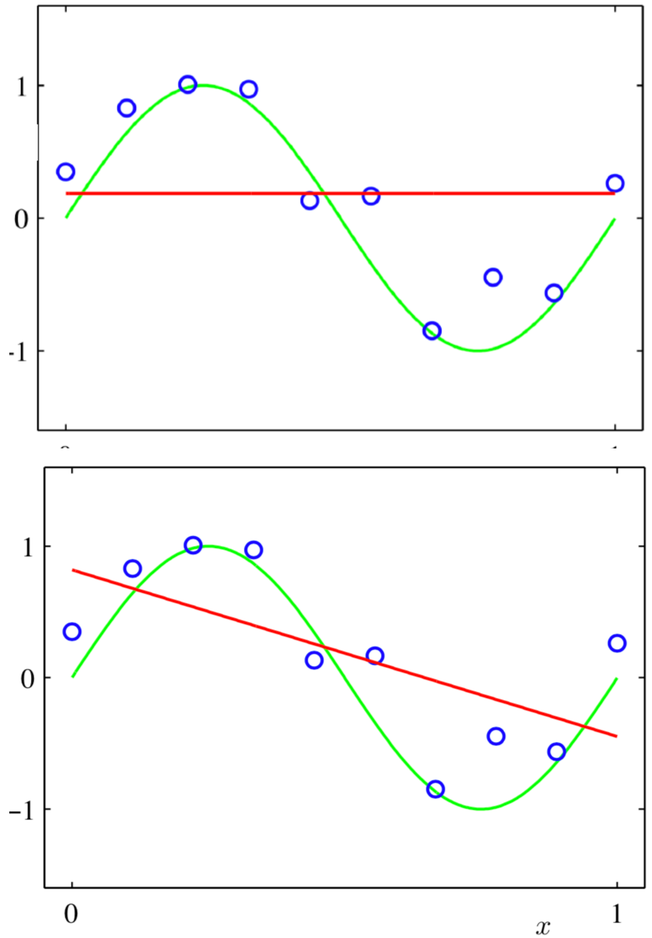
$$\phi(x) = \left(1, x, \ldots, x^d\right)^T, \text{ then } f_{\mathbf{W}}(x) = \langle \phi(x), \mathbf{W} \rangle.$$

Augmented/Extended feature vectors



m.poplavskyi@qmul.ac.uk

Polynomial regression



How to adapt a function $f(x)$?

For a new problem $f_{\mathbf{W}}(\mathbf{x}) = \langle \phi(x), \mathbf{W} \rangle$. Modified MSE-problem is

$$\hat{\mathbf{W}} = \arg \min_{\mathbf{W} \in \mathbb{R}^{d+1}} \left\{ \frac{1}{2s} \left\| \Phi(\mathbf{X}) \mathbf{W} - \mathbf{Y} \right\|^2 \right\},$$

where $\Phi(\mathbf{X}) = \begin{pmatrix} \phi\left(x^{(1)}\right) \\ \cdots \\ \phi\left(x^{(s)}\right) \end{pmatrix}$. Solution to the problem is:

$$\hat{\mathbf{W}} = \left(\Phi(\mathbf{X})^T \Phi(\mathbf{X}) \right)^{-1} \Phi(\mathbf{X})^T \mathbf{Y}$$

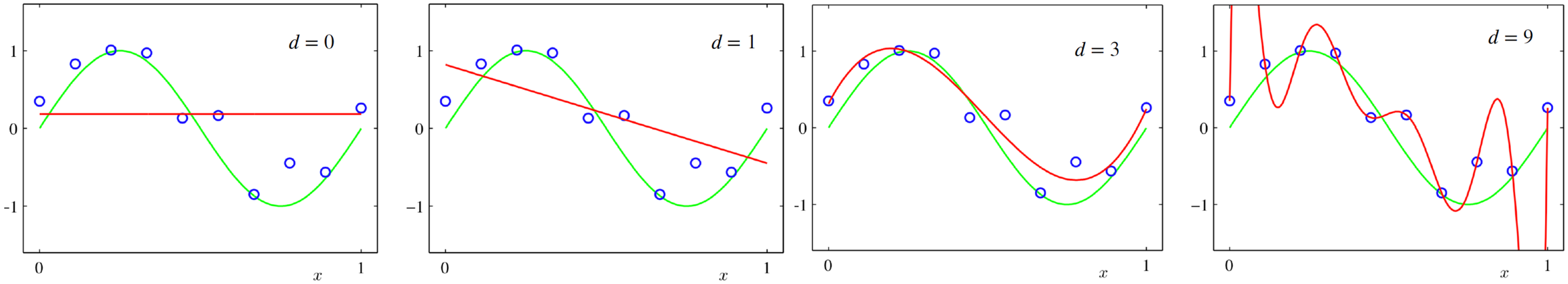
m.poplavskyi@qmul.ac.uk

Under- and over- fitting

The models we consider can be too limited or too rich:

- We say that the function is **underfitting** the data if we can't find a function that is a good fit to our data.

- We say that the function is **overfitting** the data if we find a function that fits the data too well.



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.

m.poplavskyi@qmul.ac.uk

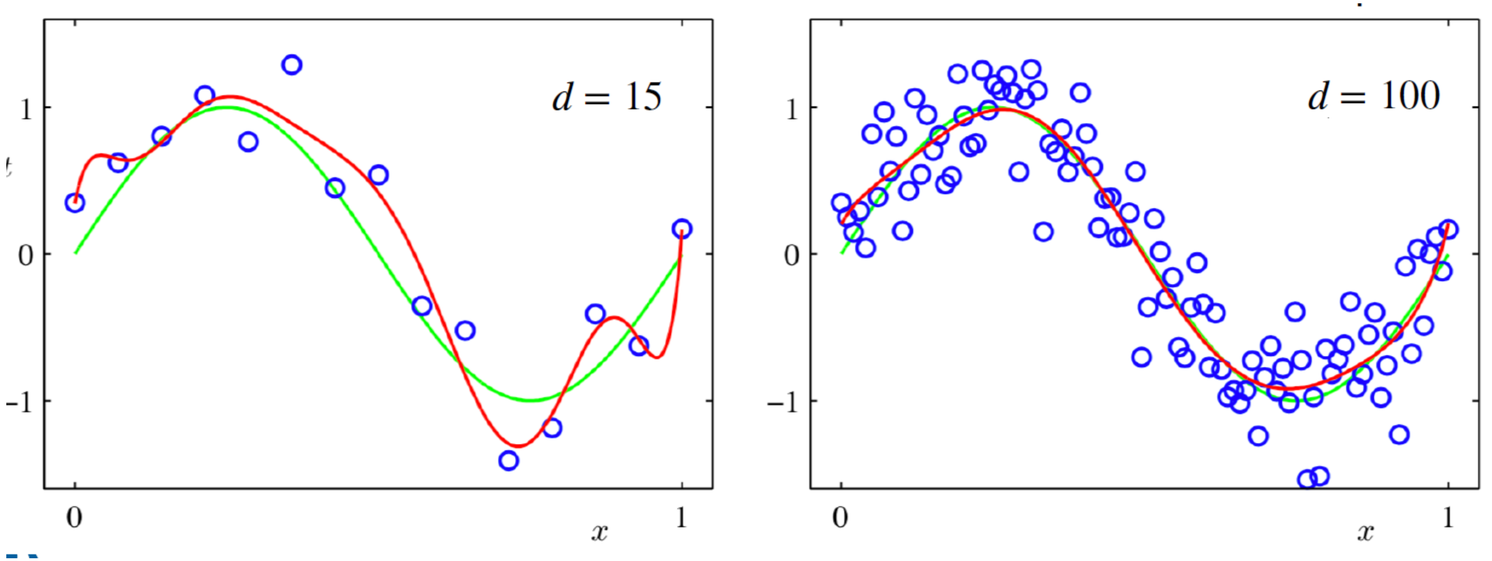
Under- and over- fitting

The models we consider can be too limited or too rich:

- We say that the function is **underfitting** the data if we can't find a function that is a good fit to our data.

- We say that the function is **overfitting** the data if we find a function that fits the data too well.

How to tackle the problem of overfitting?



- Increase the number of samples
- Use regularisation

m.poplavskyi@qmul.ac.uk

Other types of regressions

Instead of a **power basis** one may consider other basis functions, for example:

- Trigonometric functions

  Let us consider functions $a_k(x) = \sin(\pi k x)$ and $b_k(x) = \cos(\pi k x)$ and corresponding augmented vectors
  $\phi(x) = \left(b_0(x), a_1(x), b_1(x), \ldots, a_d(x), b_d(x)\right)$

- Radial basis functions

  One can also consider bell functions either $\psi(x) = \mathrm{e}^{-\frac{x^2}{2\sigma^2}}$ or
  $\psi(x) = \left(1 - \frac{1}{\alpha}|x|\right) 1_{|x| \le \alpha}$ and $\phi(x) = \left(\psi(x - \mu_1), \ldots, \psi(x - \mu_d)\right)$

m.poplavskyi@qmul.ac.uk

Minimisers and convexity

Minimisation problem

Previously we have seen that the problem of linear/polynomial regression is:

$$\text{MSE}\left(\mathbf{W}\right) = \frac{1}{2s}\sum_{j=1}^{s}\left\| f\left(\mathbf{x}^{(j)}\right) - \mathbf{y}^{(j)}\right\|^{2} \to \min$$

And in case of linear regression this can be solved via:

$$\nabla \text{MSE}\left(\hat{\mathbf{W}}\right) = 0 \Leftrightarrow \mathbf{X}^{T}\mathbf{X}\hat{\mathbf{W}} = \mathbf{X}^{T}\mathbf{Y}$$

While in case of polynomial regression this can be solved via:

$$\nabla \text{MSE}\left(\hat{\mathbf{W}}\right) = 0 \Leftrightarrow \Phi\left(\mathbf{X}\right)^{T}\Phi\left(\mathbf{X}\right)\hat{\mathbf{W}} = \Phi\left(\mathbf{X}\right)^{T}\mathbf{Y}$$

Can we be sure that this is the minimiser?

m.poplavskyi@qmul.ac.uk

3 questions to answer

There are 3 questions/concerns to raise:

1. Why solving of the equation $\nabla \mathrm{MSE}\left(\hat{\mathrm{W}}\right) = 0$ is equivalent to the problem $\hat{\mathrm{W}} = \arg \min_{W \in \mathbb{R}^{(d+1)\times m}} \mathrm{MSE}\,(W)$.
2. Why does a solution to minimisation problem exist?
3. Why is a solution to minimisation problem unique?

Question 2: Linear algebra

Question 3: Convexity

m.poplavskyi@qmul.ac.uk

Existence of a minimiser

Solving $\nabla \mathrm{MSE}(\mathbf{W}) = 0$ in the case of linear/polynomial regression is equivalent to:

$$\Phi(\mathbf{X})^T \Phi(\mathbf{X}) \mathbf{W} = \Phi(\mathbf{X})^T \mathbf{Y} \qquad (1)$$

Lemma: For any matrix $\mathbf{M} \in \mathbb{R}^{n \times m}$ we have

$$\ker\left(\mathbf{M}^T\mathbf{M}\right) = \ker \mathbf{M} \qquad \mathrm{ran}\left(\mathbf{M}^T\mathbf{M}\right) = \mathbf{M}^{\mathrm{T}}$$

Proof of the lemma: see lecture notes (non-examinable)

Application of the lemma: the right hand side of (1) belongs to $\ker \Phi(\mathbf{X})^T$. By the lemma this belongs to $\ker\left(\mathbf{\Phi}(\mathbf{X})^{\mathbf{T}} \mathbf{\Phi}(\mathbf{X})\right)$, i.e. there exist a vector $\hat{\mathbf{W}}$ solving (1).

The minimisation problem is thus solvable.

m.poplavskyi@qmul.ac.uk

Convexity

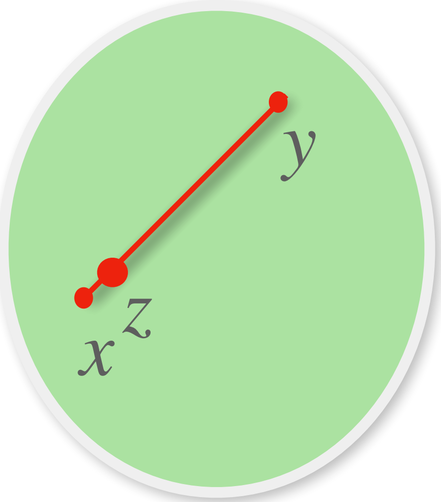
In this section we will show that the cost function is a convex one and then will use the theory of convex functions to conclude that the gradient of cost function is equal to zero only at the point of minimum.

Definition. A subset C of a vector space is said to be **convex set** if for any $\lambda \in [0,1]$ and $\mathbf{x}, \mathbf{y} \in C$ the element

$$\mathbf{z}_\lambda := \lambda \mathbf{x} + (1-\lambda)\,\mathbf{y} \in C.$$

Definition. A function f that is defined on a convex set C is said to be a **convex function** if for any $\lambda \in [0,1]$ and $\mathbf{x}, \mathbf{y} \in C$

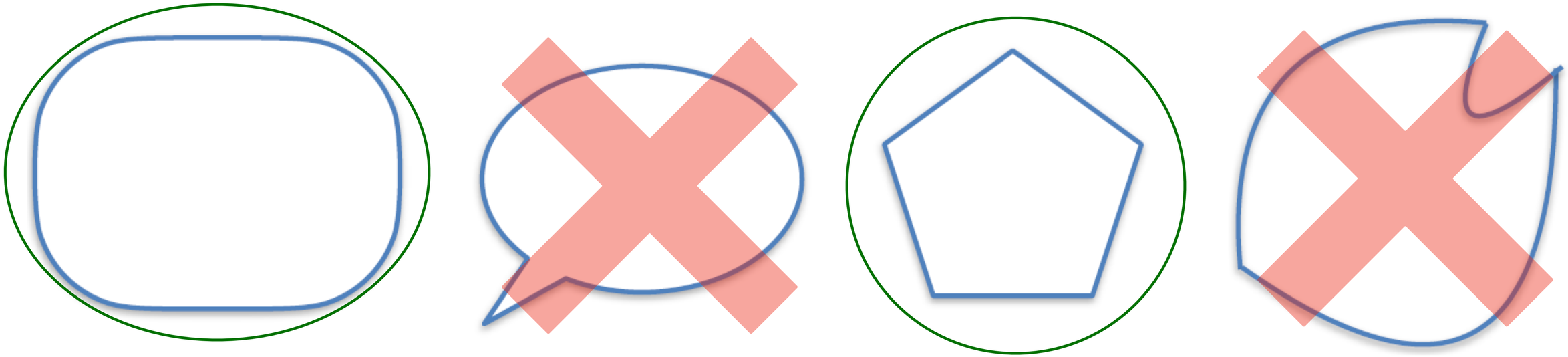
$$f\left(\lambda \mathbf{x} + (1-\lambda)\,\mathbf{y}\right) \leq \lambda f(\mathbf{x}) + (1-\lambda) f\left(\mathbf{y}\right).$$



m.poplavskyi@qmul.ac.uk

Convexity quiz

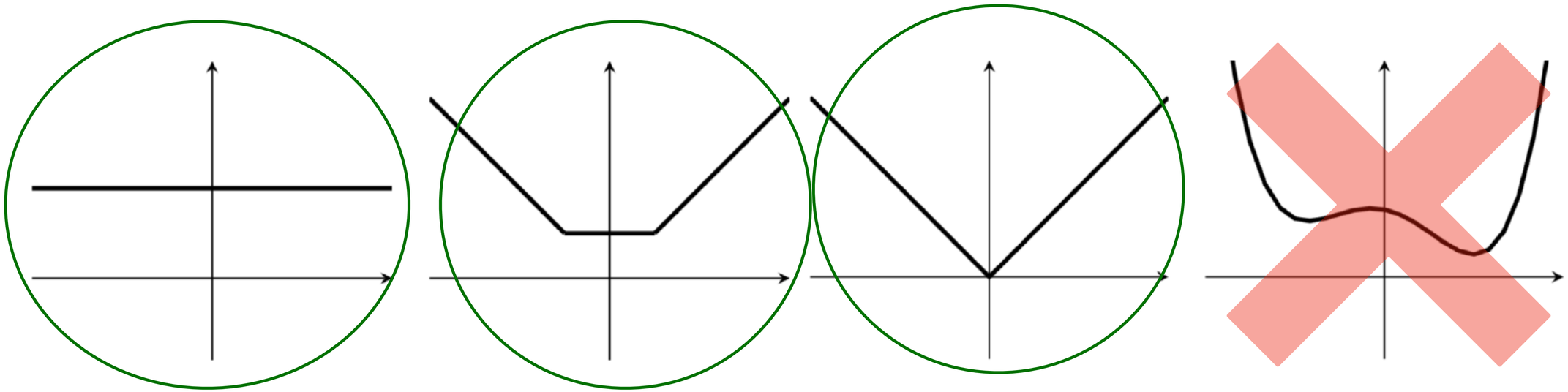
Which of these sets are convex?



m.poplavskyi@qmul.ac.uk

Convexity quiz

Which of these functions are convex?



m.poplavskyi@qmul.ac.uk

How the convexity can help?

Let $f : \mathscr{C} \to \mathbb{R}$ be a convex function defined on a convex set $\mathscr{C}$.

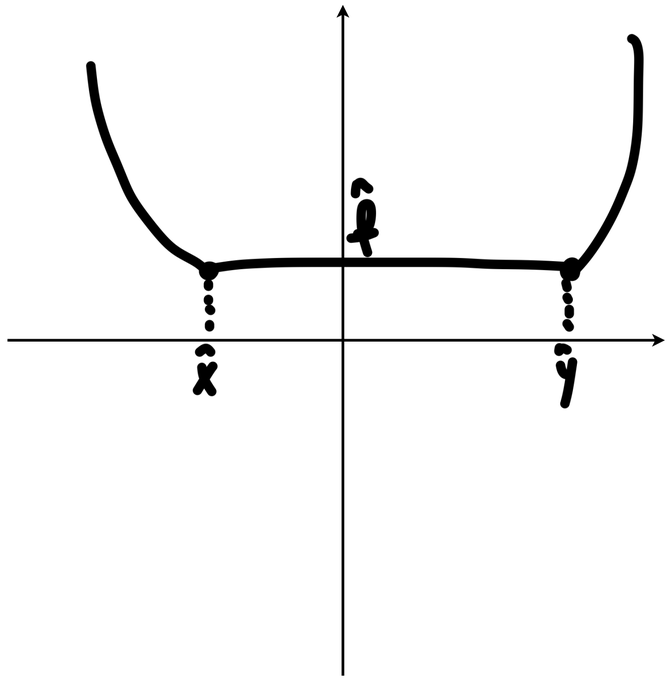
Question: Could a convex function have two minimisers?

Let $\mathbf{x}, \mathbf{y}$ be two different minima, $f(\mathbf{x}) = f(\mathbf{y}) = \hat{f} = \min f$. Then $\forall \lambda \in (0,1)$

$$f\left(\lambda \mathbf{x} + (1-\lambda)\,\mathbf{y}\right) \leq \lambda f(\mathbf{x}) + (1-\lambda) f\left(\mathbf{y}\right) = \lambda \hat{f} + (1-\lambda)\,\hat{f} = \hat{f} \Rightarrow$$

$$\forall \lambda \in [0,1] \quad f\left(\lambda \mathbf{x} + (1-\lambda)\,\mathbf{y}\right) = \hat{f}$$

Answer: Convex function can have more than one minimiser if it is a constant on some interval.



m.poplavskyi@qmul.ac.uk

How the convexity can help?

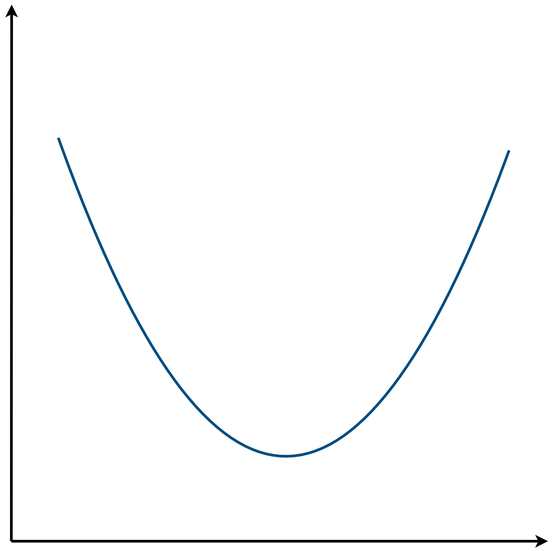
Lemma: Let $f : \mathscr{C} \subset \mathbb{R}^n \to \mathbb{R}$ is a convex function that is bounded from below and also differentiable. Then an argument $\hat{\mathbf{w}}$ satisfies $\nabla f\left(\hat{\mathbf{w}}\right) = 0$ if and only if $\hat{\mathbf{w}}$ is a global minimiser.

Motivation:

We need to prove 2 results

If $\hat{\mathbf{w}}$ is a minimiser $\Rightarrow \nabla f\left(\hat{\mathbf{w}}\right) = 0.$

If $\nabla f\left(\hat{\mathbf{w}}\right) = 0 \Rightarrow \hat{\mathbf{w}}$ is a minimiser.



m.poplavskyi@qmul.ac.uk

How the convexity can help?

Lemma: Let $f : \mathscr{C} \subset \mathbb{R}^n \to \mathbb{R}$ is a convex function that is bounded from below and also differentiable. Then an argument $\hat{\mathbf{w}}$ satisfies $\nabla f\left(\hat{\mathbf{w}}\right) = 0$ if and only if $\hat{\mathbf{w}}$ is a global minimiser.

Proof 1D: minimiser ⇒ tangent

Let $\hat{w}$ be a minimiser. What can we say about $f'\left(\hat{w}\right)$?

$$f'\left(\hat{w}\right) = \lim_{w \to \hat{w}+} \frac{\overbrace{f(w) - f\left(\hat{w}\right)}^{\geq 0}}{\underbrace{w - \hat{w}}_{>0}} \geq 0 \qquad f'\left(\hat{w}\right) = \lim_{w \to \hat{w}-} \frac{\overbrace{f(w) - f\left(\hat{w}\right)}^{\geq 0}}{\underbrace{w - \hat{w}}_{<0}} \leq 0 \Rightarrow f'\left(\hat{w}\right) = 0$$

m.poplavskyi@qmul.ac.uk

How the convexity can help?

Lemma: Let $f : \mathscr{C} \subset \mathbb{R}^n \to \mathbb{R}$ is a convex function that is bounded from below and also differentiable. Then an argument $\hat{\mathbf{w}}$ satisfies $\nabla f\left(\hat{\mathbf{w}}\right) = 0$ if and only if $\hat{\mathbf{w}}$ is a global minimiser.

Proof 1D: tangent ⇒ minimiser

Let $\hat{w}$ be an argument such that $f'\left(\hat{w}\right) = 0$. For any $\lambda \in \left[0,1\right]$ and $w \in \mathscr{C}$

$$f\left(\lambda w + (1-\lambda)\,\hat{w}\right) \leq \lambda f(w) + (1-\lambda) f\left(\hat{w}\right)$$

$$f\left(\hat{w} + \lambda\left(w - \hat{w}\right)\right) - f\left(\hat{w}\right) \leq \lambda\left(f(w) - f\left(\hat{w}\right)\right)$$

$$f(w) - f\left(\hat{w}\right) \geq \frac{f\left(\hat{w} + \lambda\left(w - \hat{w}\right)\right) - f\left(\hat{w}\right)}{\lambda} \to_{\lambda \to 0} f'\left(\hat{w}\right)\left(w - \hat{w}\right) = 0$$

m.poplavskyi@qmul.ac.uk

How the convexity can help?

Conclusions:

- Solving a regression problem is equivalent to a minimisation of cost function.
- If the cost function is convex, then the above is equivalent to finding a point of zero gradient.
- The solution to normal equation exists.
- This solution is unique except of very special cases

Question: What is left?

Answer: We need to show the convexity of a cost function

m.poplavskyi@qmul.ac.uk