06-20416 and 06-12412
(Intro to) Neural Computation

02 – Linear Regression

Per Kristian Lehre

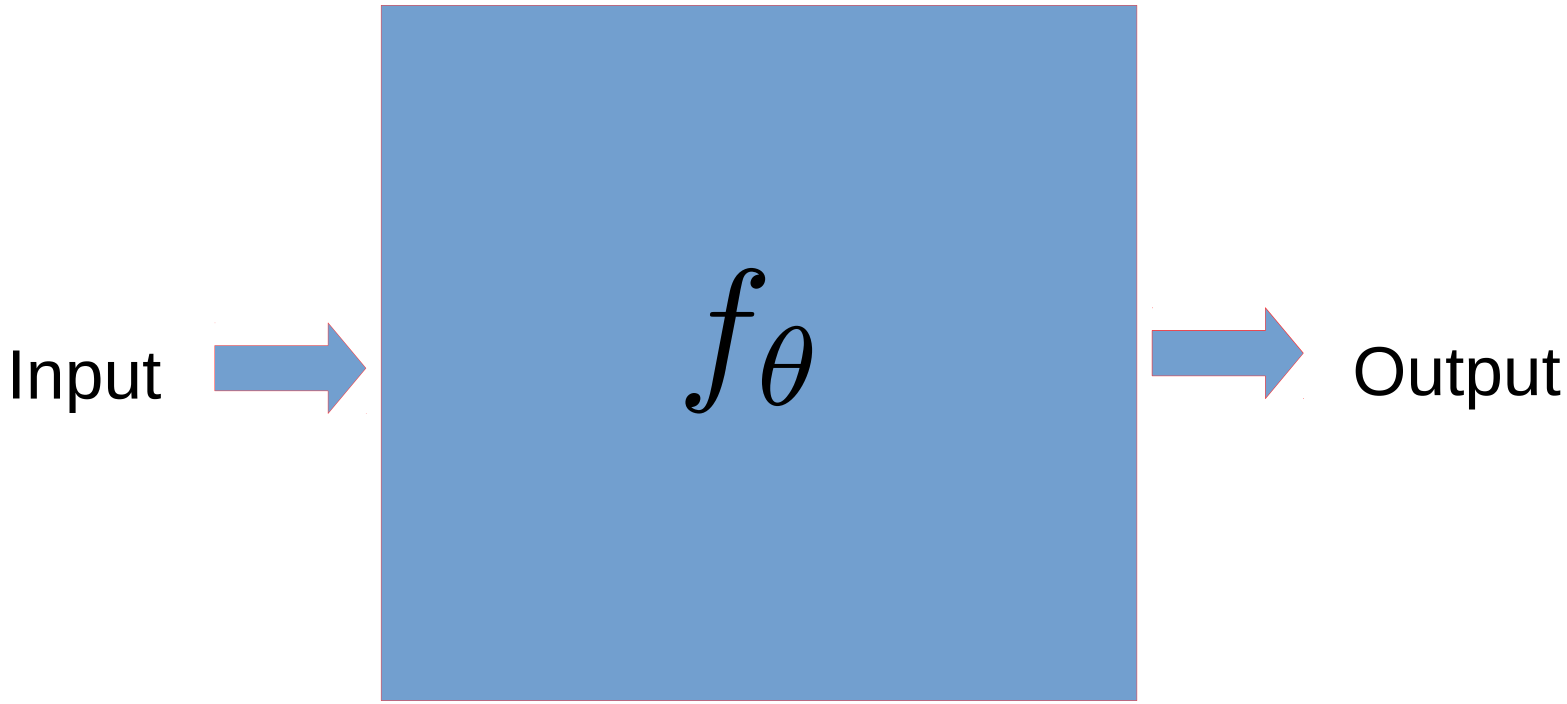
Last lecture

- A definition of machine learning
  - Performance *P* of algorithm at task *T* improves with experience *E*
- Machine learning tasks *T*
  - regression, classification, transcription, translation, synthesis and sampling
- Performance measure *P*
  - Depends on learning tasks, e.g., accuracy for classification
- Experience *E*
  - supervised learning, unsupervised learning, reinforcement learning

Outline

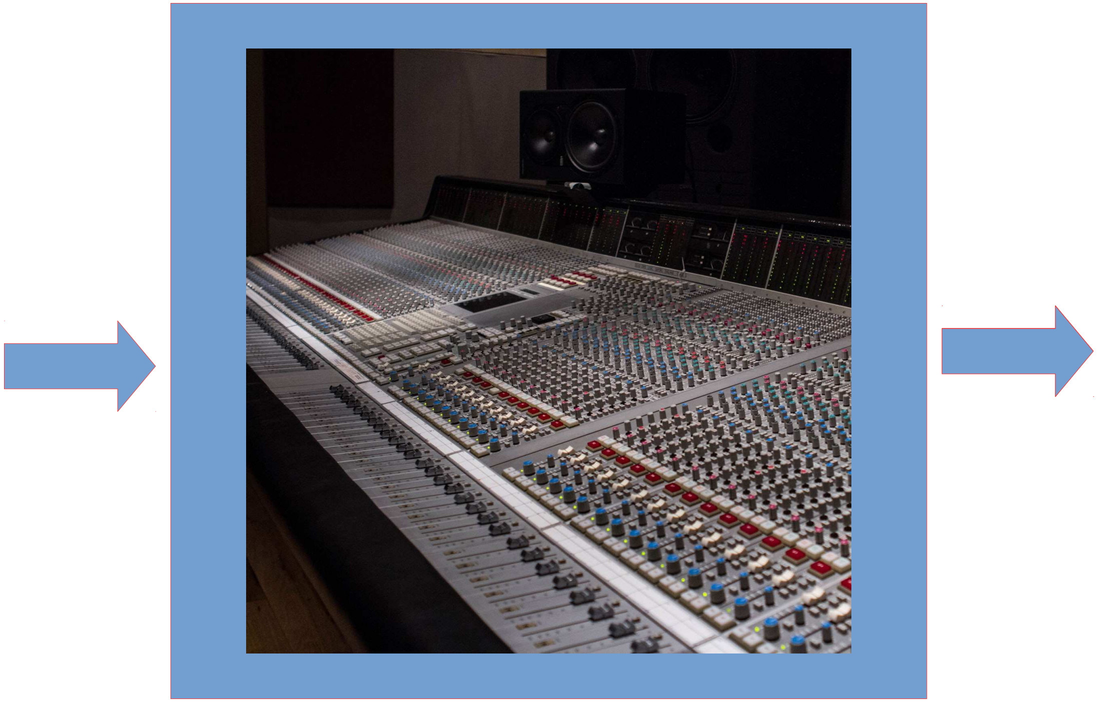
- Linear regression models
  - model *linear* relationship between input and output
  - Mean square error as cost function
- Optimisation
- Derivatives
  - The chain rule
- Ordinary Least Square (OLS)
- Gradient Descent

Cartoon picture of ML



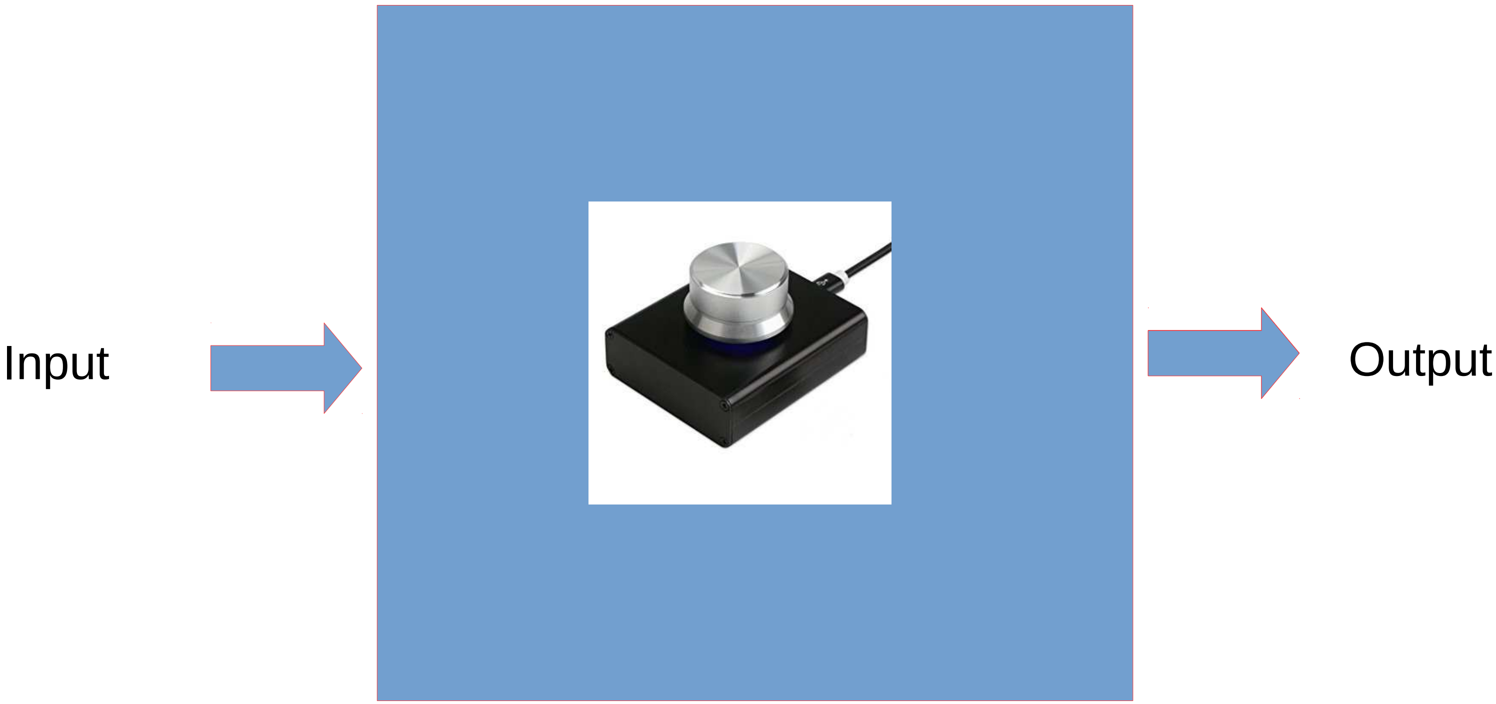
Cartoon picture of ML

Input

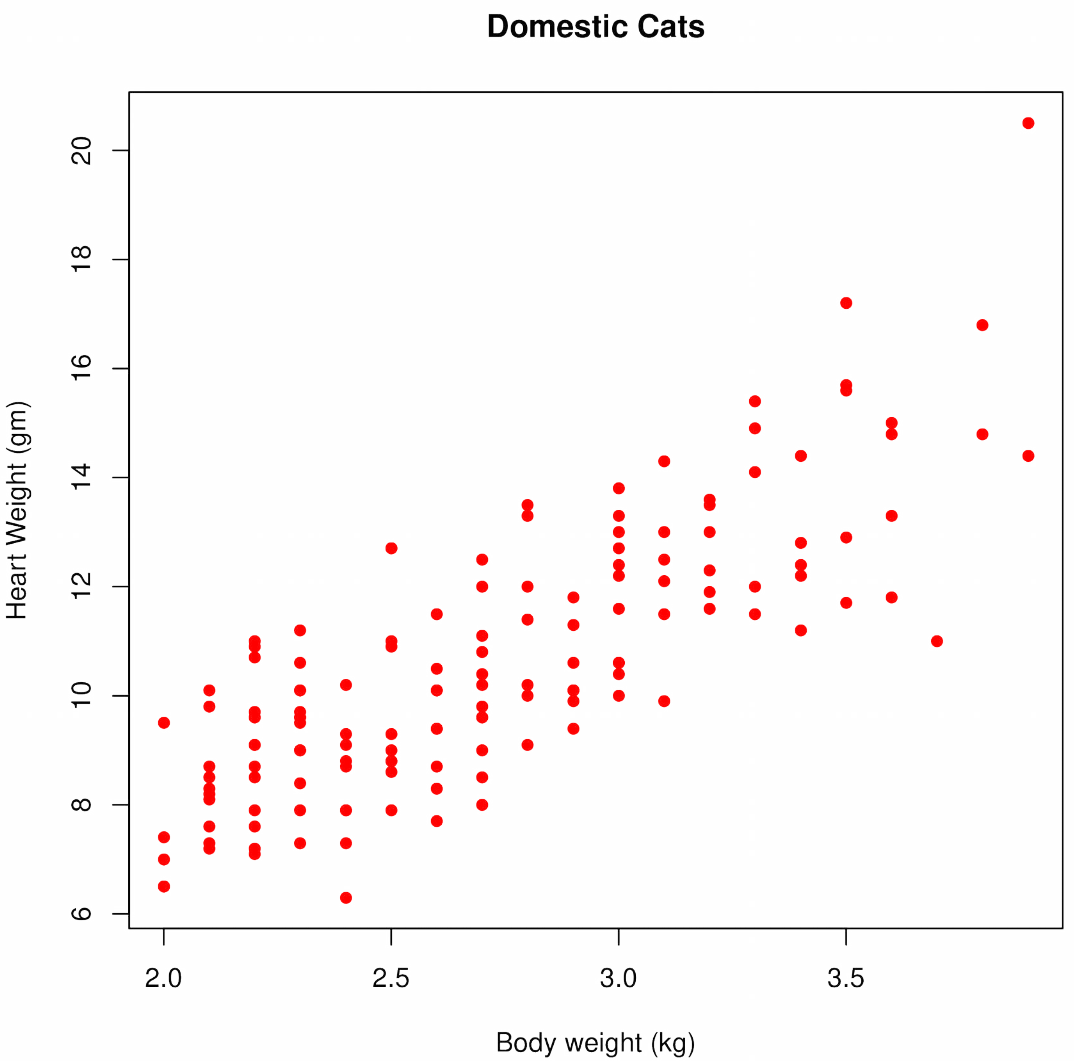


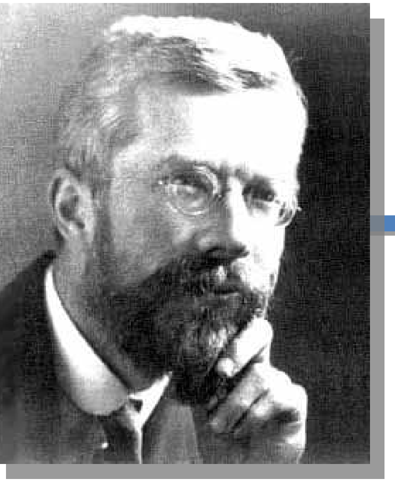
Output

Today: Adjusting a single knob



Example: Predicting the weight of cat hearts





R. A. Fisher

THE ANALYSIS OF COVARIANCE METHOD FOR THE RELATION BETWEEN A PART AND THE WHOLE

R. A. Fisher
University of Cambridge

At the suggestion of Dr. C. I. Bliss and by the courtesy of Professor H. G. O. Holck, whose data I shall use, the following note may serve to illustrate the extreme simplicity with which the technique derived from the analysis of covariance may be applied to problems concerned with the relation of a part to the whole, such as are constantly arising in many fields.

The data consist of the body weights in kilograms and the heart weights in grams of 144 cats used in a group of digitalis assays.[1] Of these 47 were females and 97 males. These data are presented in the table at the end of this note. To simplify the calculations only one decimal place was used for each value. Thus we have:

TABLE 1

Total Weights

	Females	*Males*
Number	47	97
Total body weight	110.9 Kg.	281.3 Kg.
Total heart weight	432.5 g.	1098.2 g.
Heart as fraction of entire body	.3900%	.3904%

The observed variation in these two measurements can, of course, be expressed by means of the sums of squares and products, as in the following tables. The rather intimidating phrase "spurious correlation" used in the earlier literature sometimes prevents workers from taking the simplest course. Obviously it would be easy to derive from

[1] Holck, Harald G. O., Kazuo K. Kimura, and Barbara Bartels, "Effect of the Anesthetic and the Rate of Injection of Digitalis upon Its Lethal Dose in Cats," *Journal of the American Pharmaceutical Assn.* 35: 366–370 (1946).

Experience E

The dataset consists of n data points

$(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \mathbb{R}$ where

$x^{(i)} \in \mathbb{R}^d$ is the "input" for the i-th data point as a feature vector with d elements

(e.g., $d=1$, the body weight of the i-th cat)

$y^{(i)} \in \mathbb{R}$ is the "output" for the i-th data point

(e.g., the weight of the heart of the i-th cat)

Linear Regression Task T

Find a "model", i.e., a function

$$f : \mathbb{R}^d \to \mathbb{R}$$

such that on future observations, i.e.,

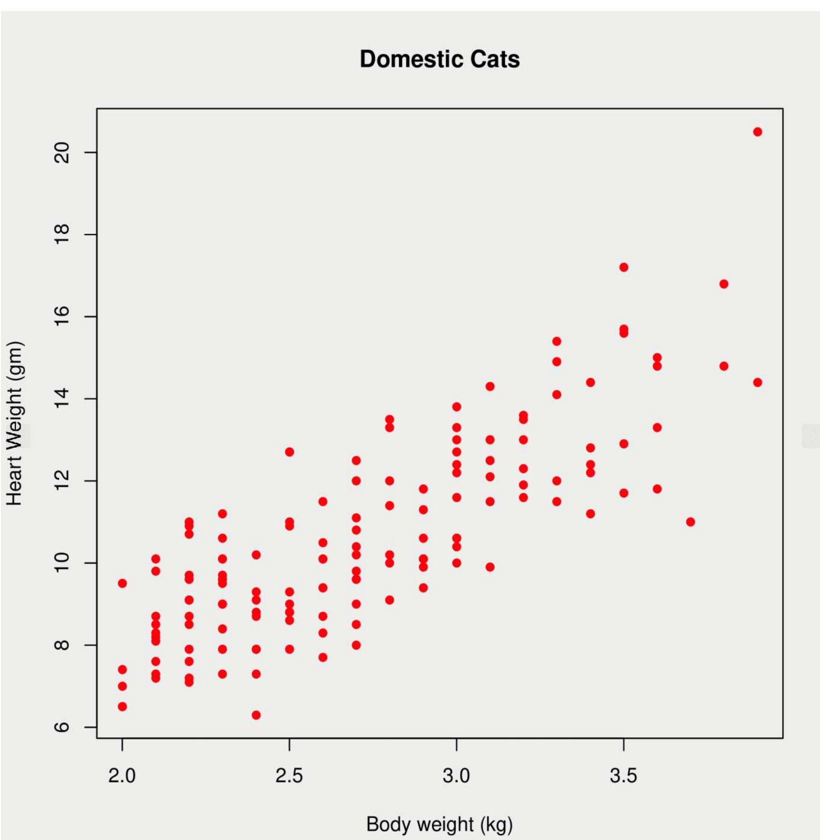
The unknown distribution which the data comes from (see lecture 1).

input

$$(X, Y) \sim \mathcal{D}$$

output

the predicted output $f(X)$ is "close to" the true output Y.



Visualisation of the data indicates a linear relationship between the input (body weight) and the output (heart weight).

Linear Regression Model

A linear regression model has the form

$$f(x) = \sum_{i=1}^{d} w_i x_i + b$$

where

$x \in \mathbb{R}^d$ is the input vector (features)

$w \in \mathbb{R}^d$ is a weight vector (parameters)

$b \in \mathbb{R}$ is a bias parameter

$f(x) \in \mathbb{R}$ is the predicted output

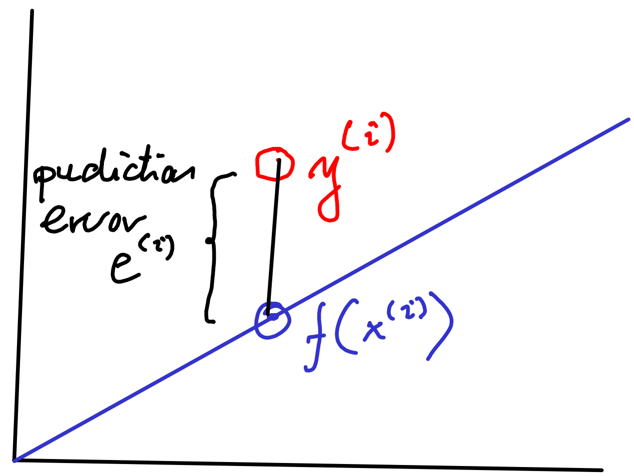
In our cat example, we have

- $d = 1$ because "body weight" is only feature
- $b = 0$ why?

$\Longrightarrow$ Our model has one parameter: w

Performance Measure P

- Want a function $J(w)$ which quantifies the error in the predictions for a given parameter w.



- The prediction "error" on the i-th data point can be defined as

$$e^{(i)} = y^{(i)} - f(x^{(i)})$$
$$= y^{(i)} - w x^{(i)}$$

- The following empirical loss function J takes into account the errors for all n data points

weight parameter

$$J(w) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \left(y^{(i)} - w x^{(i)} \right)^2$$

By squaring the error, we
1) ignore the sign of the errors
2) penalise large errors more (assuming $e^{(i)} > 1$)

These are constants that will be useful later

Idea: Find the parameter w which minimises the loss $J(w)$.

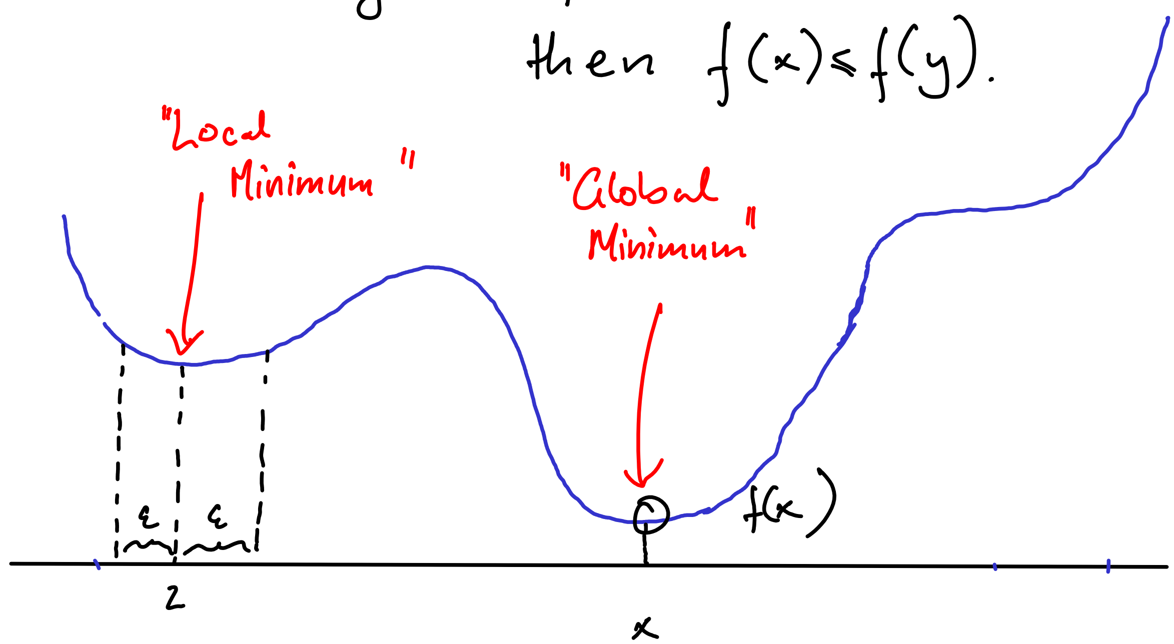
Unconstrained Optimisation (minimisation)

Given a function

$$f : \mathbb{R}^d \to \mathbb{R} \qquad \text{called the } \underline{\text{loss function}},$$

an element $x \in \mathbb{R}^d$ is called

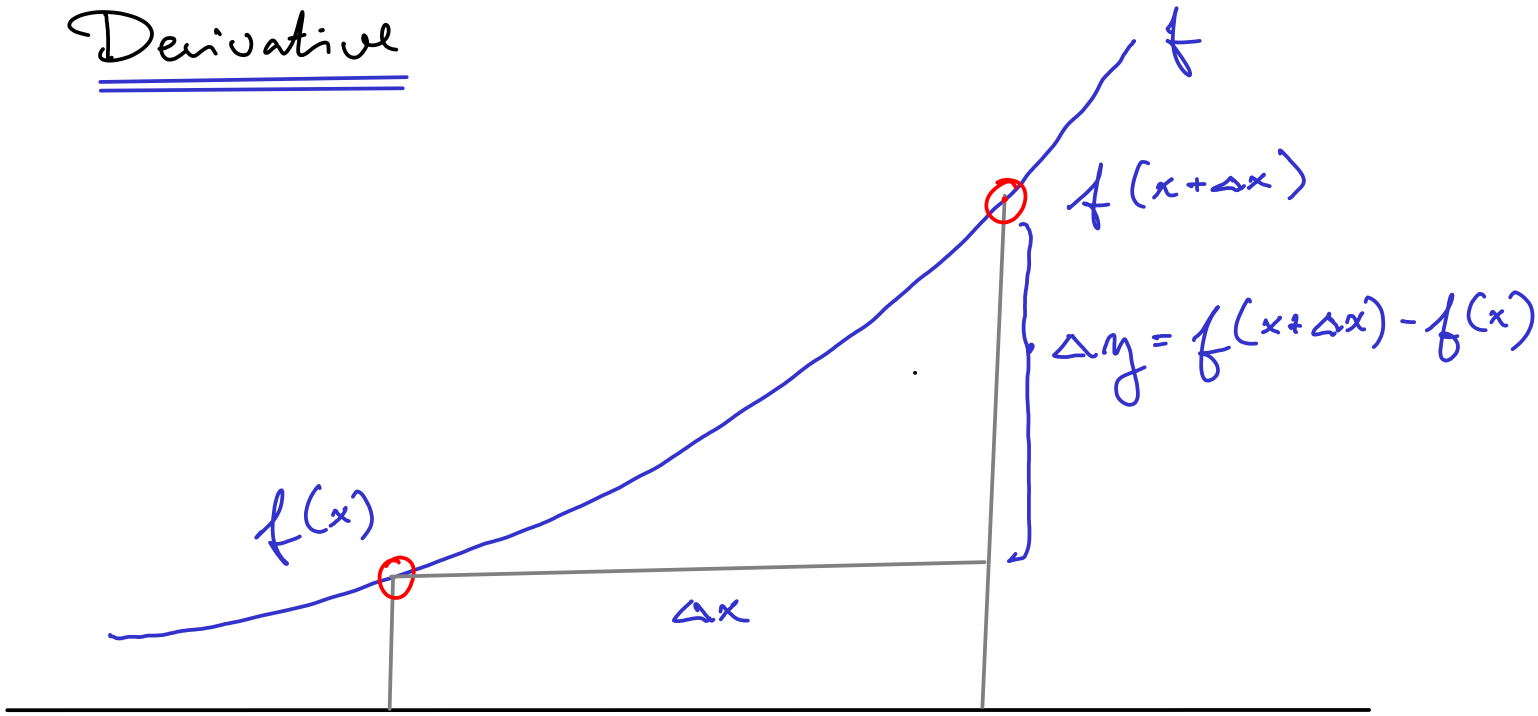
- a <u>global minimum</u> of f if
  $\forall y \in \mathbb{R}^d \quad f(x) \leq f(y)$
- a <u>local minimum</u> of f if
  $\exists \varepsilon > 0 \ \forall y \in \mathbb{R}^d$ if $\forall i \in \{1, \dots, d\} \ |x_i - y_i| < \varepsilon$
  then $f(x) \leq f(y)$.



Theorem

For any continuous function $f : \mathbb{R} \to \mathbb{R}$,
if x is a local optimum, then $f'(x) = 0$.

Derivative



Definition

The first derivative of a function $f: \mathbb{R} \to \mathbb{R}$ is

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$$

Leibniz' notation

If $y = f(x)$ then

$$\frac{dy}{dx} = f'(x)$$

Differentiation Rules

$$(c f(x))' = c f'(x)$$

$$(x^k)' = k x^{k-1} \quad \text{if } k \neq 0$$

$$(f(x) + g(x))' = f'(x) + g'(x)$$

$$(f(g(x)))' = f'(g(x)) \cdot g'(x)$$ Chain Rule

Chain Rule in Leibniz' notation

Assume that

$$x \longrightarrow h \longrightarrow z \longrightarrow g \longrightarrow y$$

$$z = h(x)$$

$$y = g(z) = g(h(x))$$

then $$\frac{dy}{dx} = \frac{dy}{dz} \cdot \frac{dz}{dx}$$

NB! Important

Chain Rule: Example

What is the derivative of

$$f(w) = \frac{1}{2}(y - wx)^2$$

Define the functions

$$\left.\begin{aligned} g(e) &= \frac{1}{2} e^2 \\ h(w) &= y - wx \end{aligned}\right\} f(w) = g(h(w))$$

$$w \rightarrow h \rightarrow e \rightarrow g \rightarrow z$$

$$e = h(w)$$

$$z = g(e) = g(h(w))$$

In Leibniz' notation

$$\frac{de}{dw} = -x \qquad \frac{dz}{de} = e$$

The chain rule gives

$$f'(w) = \frac{dz}{dw} = \frac{dz}{de} \cdot \frac{de}{dw} = -x \cdot e = x(wx - y)$$

Approach 1: Ordinary Least Squares (OLS)

Optimise J by solving $J'(\omega)=0$

$$J(\omega) = \frac{1}{2n} \sum_{i=1}^{n} \left(y^{(i)} - \omega x^{(i)}\right)^2$$

$$J'(\omega) = \frac{1}{n} \sum_{i=1}^{n} \left(\omega x^{(i)} - y^{(i)}\right) x^{(i)}$$

$$J'(\omega) = 0$$

$$\frac{1}{n} \sum_{i=1}^{n} \left(\omega x^{(i)} - y^{(i)}\right) x^{(i)} = 0$$

$$\omega \sum_{i=1}^{n} \left(x^{(i)}\right)^2 = \sum_{i=1}^{n} x^{(i)} y^{(i)}$$

$$\boxed{\omega = \frac{\sum_{i=1}^{n} x^{(i)} y^{(i)}}{\sum_{i=1}^{n} \left(x^{(i)}\right)^2}}$$

$\Rightarrow$ One solution to $J'(\omega)=0$, hence globally optimal.

Approach 2:

Often difficult or impossible to solve $J'(w)=0$ for non-linear models with many parameters such as neural networks.

Idea:

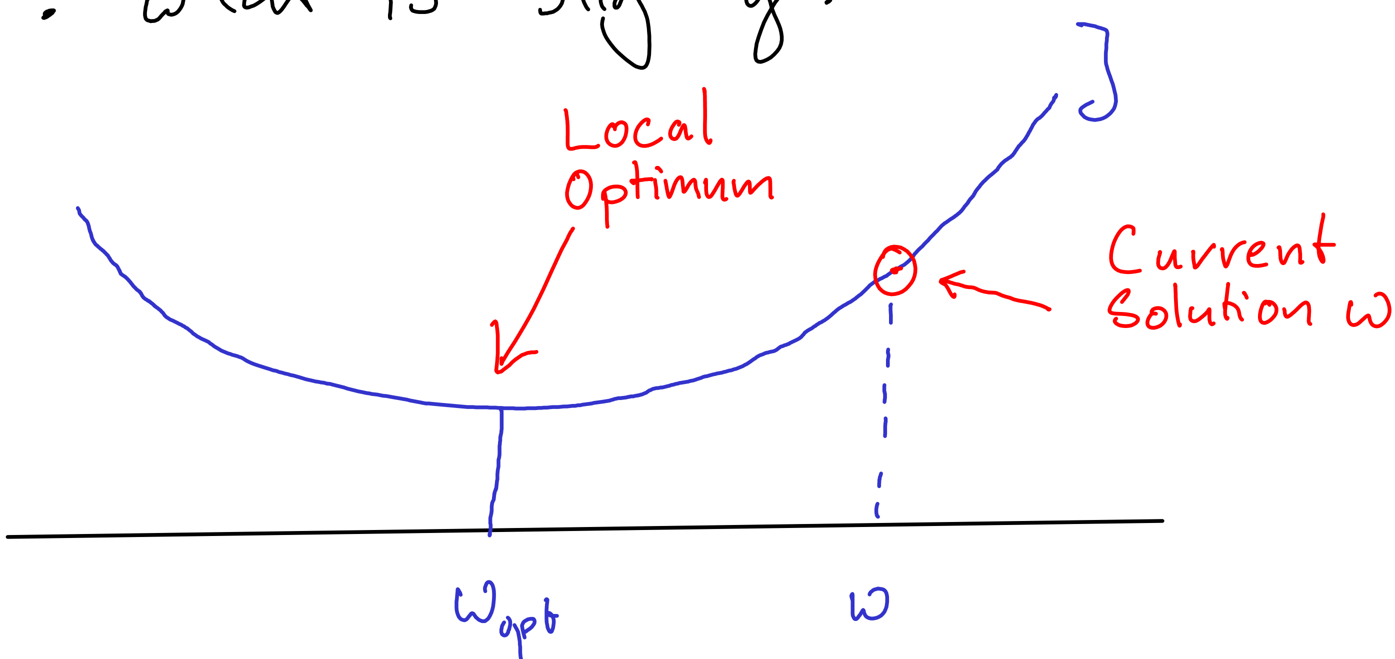
Start with an initial guess w

while $J'(w) \neq 0$

  move w "slightly" in the "right direction"

To make the idea concrete, need to clarify

- what is the right direction?
- what is slightly?



Attempt 1: (Failed attempt)

$w \leftarrow$ initial weight

repeat

    if $J'(w) < 0$

        $w \leftarrow w + \varepsilon$ i.e., move right

    else if $J'(w) > 0$

        $w \leftarrow w - \varepsilon$ i.e., move left

Problem with this attempt:

- w may oscillate in the interval $[w_{opt} - \varepsilon, w_{opt} + \varepsilon]$
- w fails to converge

Attempt 2: Gradient Descent (1D)

$w \leftarrow$ initial weight

repeat

$$w \leftarrow w - \varepsilon J'(w)$$

ε – learning rate (a hyperparameter)

Summary

- Linear regression models
  - model *linear* relationship between input and output
  - Mean square error as cost function
- Optimisation
- Derivatives
  - The chain rule
- Ordinary Least Square (OLS)
- Gradient Descent

Next week

- Maximum Likelihood
  - How to construct good loss functions