Outline of the lecture

This lecture introduces us to the topic of **supervised learning**. Here the data consists of **input**-**output** pairs. Inputs are also often referred to as **covariates**, **predictors** and **features**; while outputs are known as **variates** and **labels**. The goal of the lecture is for you to:

- ❑ Understand the supervised learning setting.
- ❑ Understand linear regression (aka **least squares**)
- ❑ Understand how to apply linear regression models to make predictions.
- ❑ Learn to derive the least squares estimate by optimization.

Linear supervised learning

- Many real processes can be approximated with linear models.
- Linear regression often appears as a module of larger systems.
- Linear problems can be solved analytically.
- Linear prediction provides an introduction to many of the core concepts of machine learning.

We are given a training dataset of n instances of input-ouput pairs $\{\mathbf{x}_{1:n}, \mathbf{y}_{1:n}\}$. Each input $\mathbf{x}_i \in \mathbb{R}^{1\times d}$ is a vector with d attributes. The inputs are also known as predictors or covariates. The output, often referred to as the target, will be assumed to be univariate, $\mathbf{y}_i \in \mathbb{R}$, for now.

$X_{1:n} = \{x_1, x_2, \ldots x_n\}$ ≈

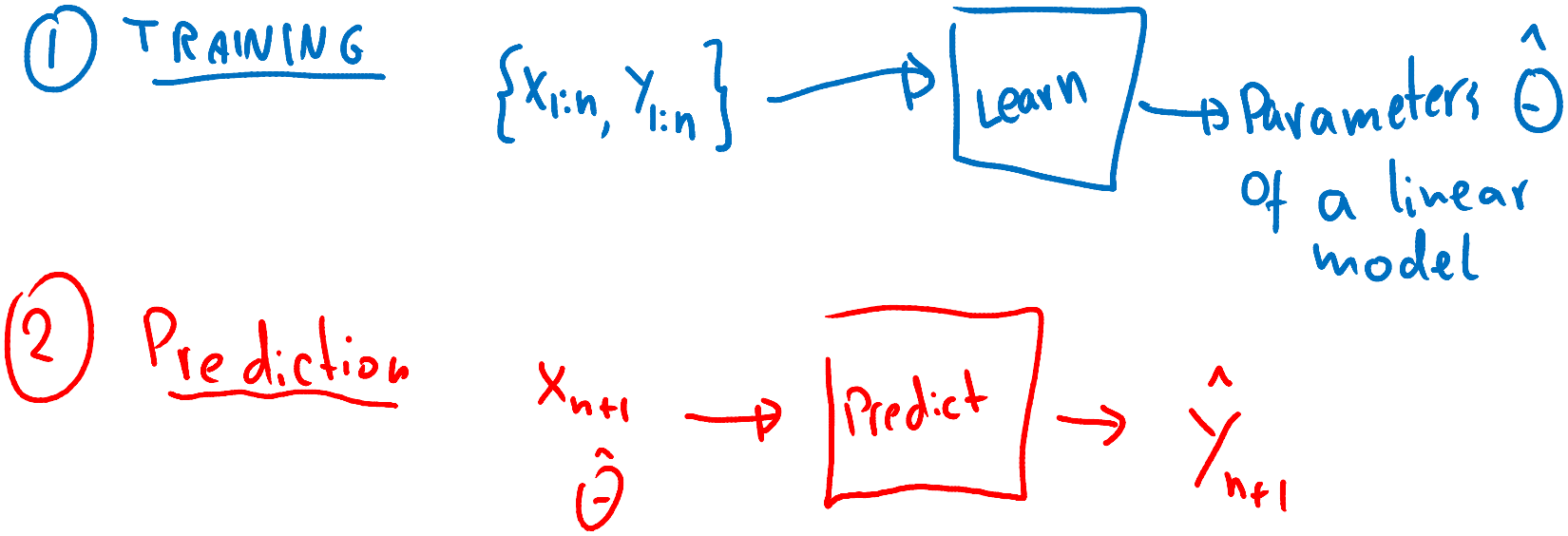
A typical dataset with $n = 4$ instances and 2 attributes would look like the following table: $d=2$

	Wind speed	People inside building	Energy requirement
1	100	2	5
2	50	42	25
3	45	31	22
$n=4$	60	35	18

$x_1 = [100 \quad 2]$ $\quad y_1 = [5]$

Energy demand prediction

Given the training set $\{\mathbf{x}_{1:n}, \mathbf{y}_{1:n}\}$, we would like to learn a model of how the inputs affect the outputs. Given this model and a new value of the input $\mathbf{x}_{n+1}$, we can use the model to make a prediction $\widehat{y}(\mathbf{x}_{n+1})$.



Prostate cancer example

❑ Goal: Predict a prostate-specific antigen (log of lpsa) from a number of clinical measures in men who are about to receive a radical prostatectomy.

❑The inputs are:

- Log cancer volume (lcavol)
- Log prostate weight (lweight)
- Age
- Log of the amount of benign prostatic hyperplasia (lbph)
- Seminal vesicle invasion (svi) - *binary*
- Log of capsular penetration (lcp)
- Gleason score (gleason) – *ordered categorical*
- Percent of Gleason scores 4 or 5 (pgg45)

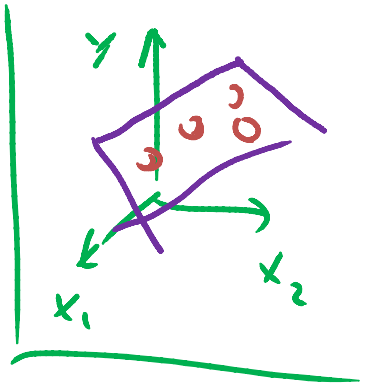
Which inputs are more important?

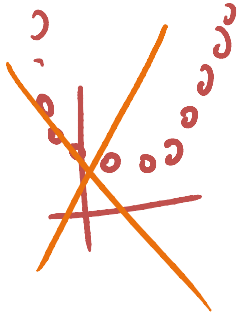
[Hastie, Tibshirani & Friedman book]

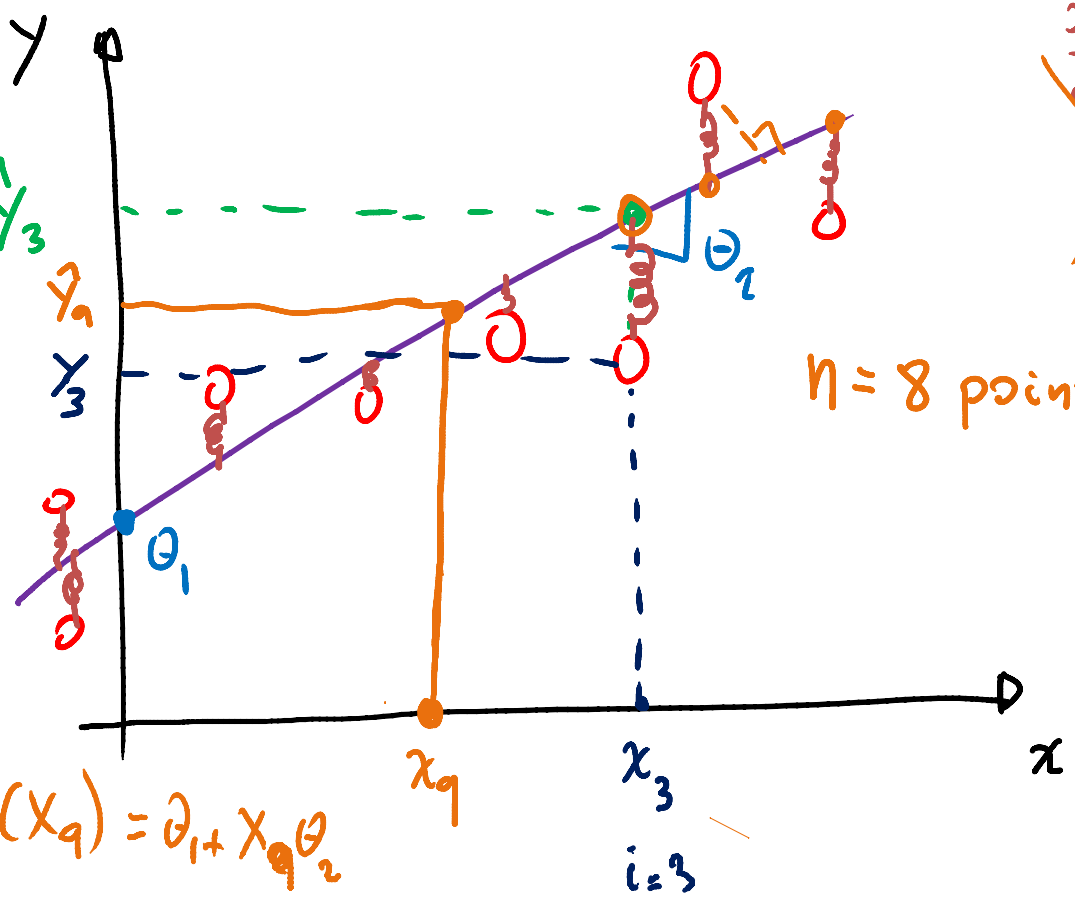
$$\hat{y}(\mathbf{x}_i) = \theta_1 + x_i\theta_2 = \hat{y}_i$$

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \theta_1 - x_i\theta_2)^2$$

↑ objective, cost, loss, energy, error function.







$\hat{y}(x_3) = \hat{y}_3$

$\hat{y}_4 = \hat{y}(x_4) = \theta_1 + x_4\theta_2$

$n = 8$ points

- training
- Test

Linear prediction

In general, the linear model is expressed as follows:

$$\widehat{y}_i = \sum_{j=1}^{d} x_{ij}\theta_j,$$

where we have assumed that $x_{i1} = 1$ so that θ_1 corresponds to the intercept of the line with the vertical axis. θ_1 is known as the bias or offset.

In matrix form, the expression for the linear model is:

$$\widehat{\mathbf{y}} = \mathbf{X}\boldsymbol{\theta},$$

with $\widehat{\mathbf{y}} \in \mathbb{R}^{n\times 1}$, $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{\theta} \in \mathbb{R}^{d\times 1}$. That is,

$$\begin{bmatrix} \widehat{y}_1 \\ \vdots \\ \widehat{y}_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nd} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_d \end{bmatrix}$$

Wind speed	People inside building	Energy requirement
100	2	5
50	42	25
45	31	22
60	35	18

For our energy prediction example, we would form the following matrices with $n = 4$ and $d = 3$:

$$\mathbf{y} = \begin{bmatrix} 5 \\ 25 \\ 22 \\ 18 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 100 & 2 \\ 1 & 50 & 42 \\ 1 & 45 & 31 \\ 1 & 60 & 35 \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}.$$

Suppose that $\boldsymbol{\theta} = [1 \quad 0 \quad 0.5]^T$. Then, by multiplying $\mathbf{X}$ times $\boldsymbol{\theta}$, we would get the following predictions on the training set:

$$\widehat{\mathbf{y}} = \begin{bmatrix} 2 \\ 22 \\ 16.5 \\ 18.5 \end{bmatrix} = \begin{bmatrix} 1 & 100 & 2 \\ 1 & 50 & 42 \\ 1 & 45 & 31 \\ 1 & 60 & 35 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0.5 \end{bmatrix}.$$

Predictions on the training set

Linear prediction

Likewise, for a point that we have never seen before, say x = [50 20], we generate the following prediction:

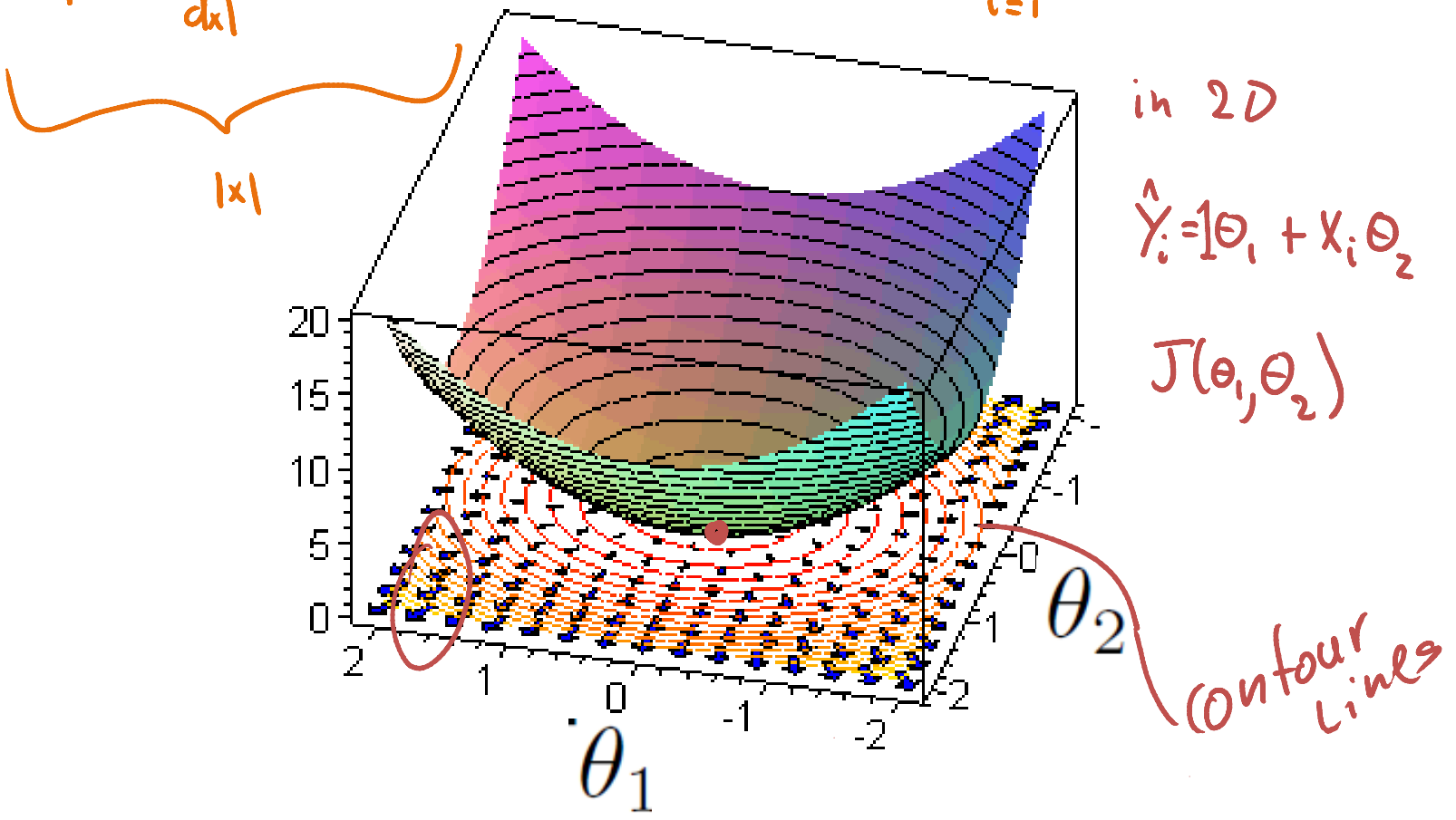
$$\hat{y}(x) = [1\ 50\ 20] \begin{bmatrix} 1 \\ 0 \\ 0.5 \end{bmatrix} = 1 + 0 + 10 = 11.$$

Optimization approach

Our aim is to mininimise the quadratic cost between the output labels and the model predictions

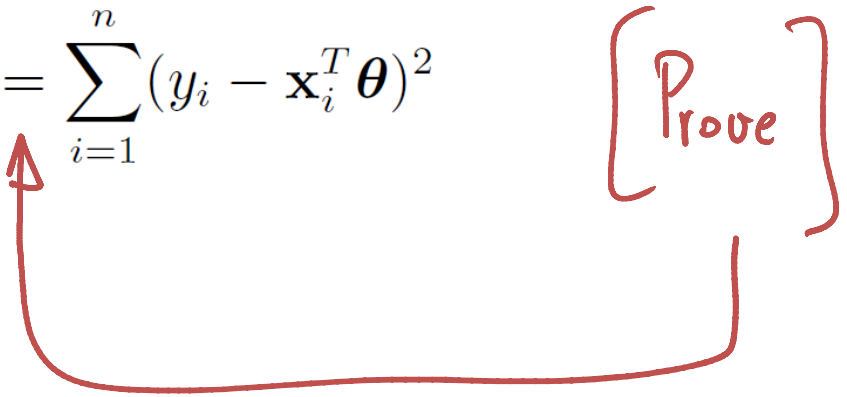
$$J(\boldsymbol{\theta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\theta})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$n \times d$, $n \times 1$, $d \times 1$, 1×1, 1×1



Optimization approach

$$J(\boldsymbol{\theta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\theta})^2$$



Optimization: Finding the minimum

$$J(\boldsymbol{\theta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\theta})^2$$

Suppose $n=3$, $d=2$

$$J(\theta) = (Y_1 - \theta_1 - x_1\theta_2)^2 + (Y_2 - \theta_1 - x_2\theta_2)^2 + (Y_3 - \theta_1 - x_3\theta_2)^2$$

$$= \sum_{i=1}^{3}(Y_i - \theta_1 - x_i\theta_2)^2$$

$$\frac{\partial J(\theta)}{\partial \theta_1} = \sum_{i=1}^{3} 2(Y_i - \theta_1 - x_i\theta_2)(-1) = -2\sum_{i=1}^{3}(Y_i - \theta_1 - x_i\theta_2)$$

Optimization

$$J(\boldsymbol{\theta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$$

n×1 n×d d×1

We will need the following results from matrix differentiation:

$$\frac{\partial \mathbf{A}\theta}{\partial \boldsymbol{\theta}} = \mathbf{A}^T \text{ and } \frac{\partial \boldsymbol{\theta}^T \mathbf{A}\boldsymbol{\theta}}{\partial \boldsymbol{\theta}} = 2\mathbf{A}^T\boldsymbol{\theta}$$

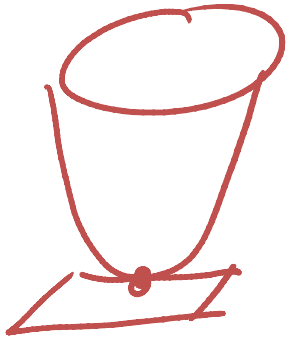
$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \theta}\left[Y^TY + \theta^T X^T X\theta - 2Y^TX\theta\right]$$

1×1

1×n n×d d×1

A A'

$$= 0 + 2X^TX\theta - 2X^TY$$



Least squares estimates

$2X^TX\Theta = 2X^Ty$ Normal eq.

$$\hat{\Theta} = (X^TX)^{-1}X^Ty$$

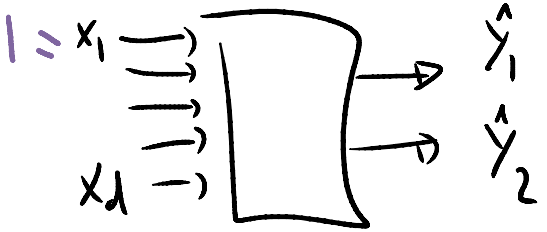
L.S. estimate

$\hat{y} = X\hat{\Theta} = \underbrace{X(X^TX)^{-1}X^T}_{\text{HAT}}y = Hy$

Multiple outputs

If we have several outputs $\mathbf{y}_i \in \mathbb{R}^c$, our linear regression expression becomes:

eg $c=2$



$$\begin{bmatrix} \hat{y}_{11} & \hat{y}_{12} \\ \vdots & \vdots \\ \hat{y}_{n1} & \hat{y}_{n2} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ 1 & & & \\ 1 & & \ddots & \\ \vdots & & & \\ 1 & & & x_{nd} \end{bmatrix} \begin{bmatrix} \theta_{11} & \theta_{12} \\ \vdots & \vdots \\ \theta_{d1} & \theta_{d2} \end{bmatrix}$$