

# Lecture 1: Introduction, Polynomial Curve

## H1 Fitting, and Probability Theory - 13/01/20

### H2 Notations



using *Classifying Hand Written Digits* as example

- A training set of  $N$  digits
- Each digit,  $i$ , is an image, representing as an **input vector of pixel values**  $x_i$
- The category of each digit,  $i$ , is known and expresses as **target vector**  $t_i$
- ML algorithm outputs function  $y(x)$ , which can take new digit input  $x$  and output vector  $y$ , which is a **guess** of the target  $t$ . The precise form of  $y(x)$  is determined during the training phases.
- The ability to categorise new examples that differ from those used for training is called **generalisation**

### H2 Supervised Learning

Problems are ones where the data contains both input and corresponding target vectors.

- **Classification**
- **Regression**

The inputs may be **pre-processed** to reduce variability in the inputs.

### H2 Unsupervised Learning

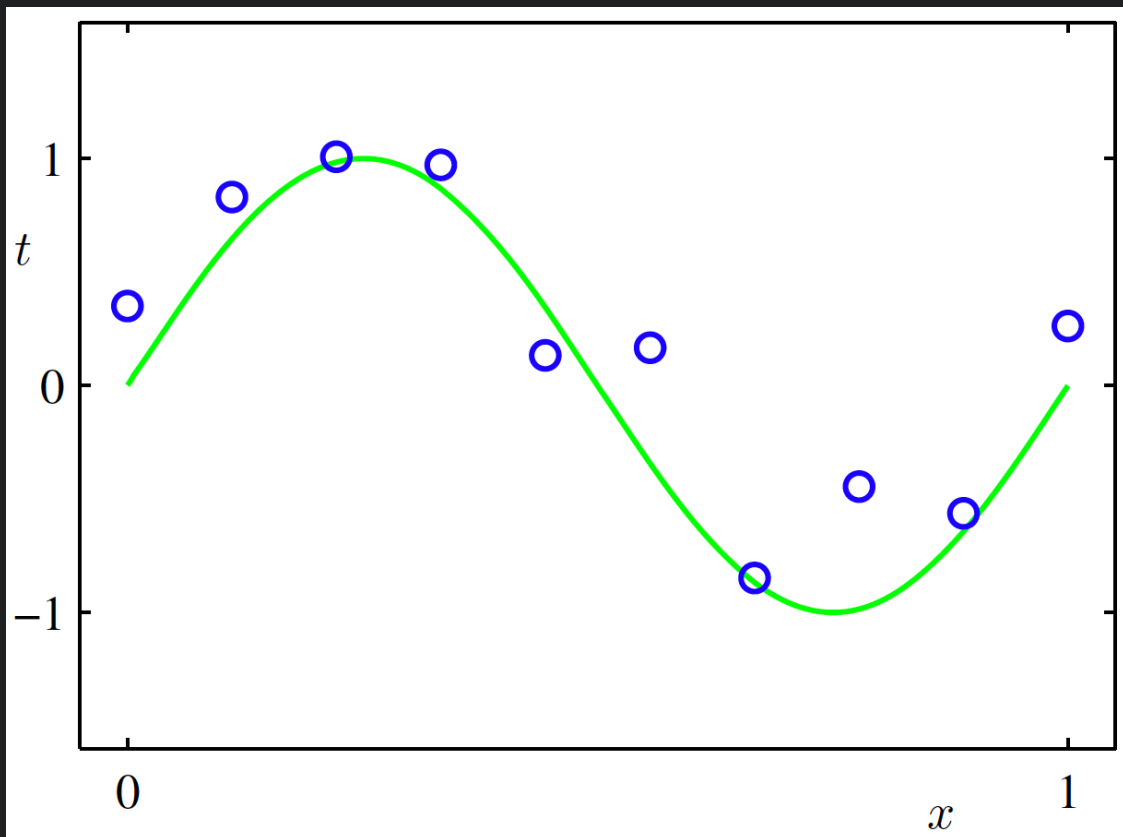
Problems are ones where the data contains only input vectors but no targets.

- **Clustering** - discovering groups of similar examples
- **Density Estimation** - learning how data is distributed
- **Dimensionality Reduction** - representing high dimensional data with just a few variables

## H2 Reinforcement Learning

Problems are ones that interact with an environment by choosing actions and observing changes in state. Actions must act to maximise a **reward signal**. Optimal actions are discovered by **trials and errors**.

## H2 Polynomial Curve Fitting



- **Training inputs**  $x = (x_1, \dots, x_N)^T$
- **Training targets**  $t = (t_1, \dots, t_N)^T$

This is **synthetic data** - we know how it originated

- Each  $x_i$  is sampled uniformly from  $[0, 1]$
- Each  $t_i = \sin(2\pi x_i) + (\text{Gaussian Noise})$

Data tends to have an underlying regularity or structure obscured by noise. Noise can be:

- **intrinsically stochastic** (random)
- resulted of **unobserved** sources of **variability**

## H3 Aim

- Predict a target  $\hat{t}$  for an unseen input  $\hat{x}$
- Discover the **underlying structure**
- Separate it from the **noise**

### H3 Fitting with Linear Model

$$y(x, \mathbf{w}) = w_0 + w_1x + w_2x^2 + \dots + w_Mx^M = \sum_{j=0}^M w_jx^j$$

- $M$  is the **order of the polynomial**
- **Polynomial coefficients**  $w_0, \dots, w_M$  are collected into vector  $\mathbf{w}$
- $y(x, \mathbf{w})$  is non-linear in  $x$ , but it is linear in  $\mathbf{w}$  and so we call this a **linear model**

We estimate values for  $\mathbf{w}$  by fitting the function to training data. Fit the function by **minising** an **error function**

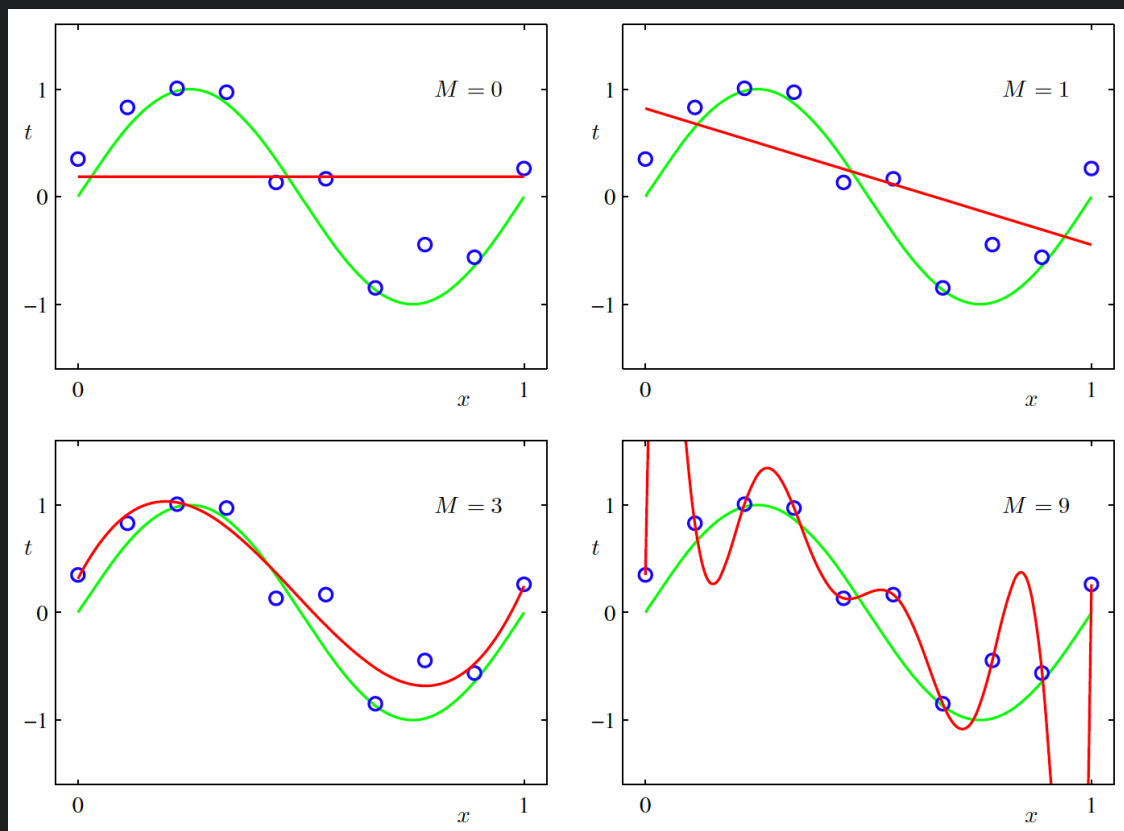
### H3 Error Function

A widely used error function is the **Sum of Square Errors**

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N [y(x_n, \mathbf{w}) - t_n]^2$$

- **Best Fit**  $\mathbf{w}^* = \arg \min_{\mathbf{w}} E(\mathbf{w})$
- **Perfect Fit** if  $E(\mathbf{w}^*) = 0$
- Bigger differences are increasingly **penalised**

### H3 Finding the Best Polynomial Degree



Choosing the best  $M$  is an example of **Model Selection**

- Small values of  $M$  give a poor fit
- Large values of  $M$  appear to **over-fit** - *capture the noise* rather than underlying structure

## H2 Evaluating Fit and Regularisation

We need an objective way to test our fit

**Root Mean Squared Error (RMSE)**

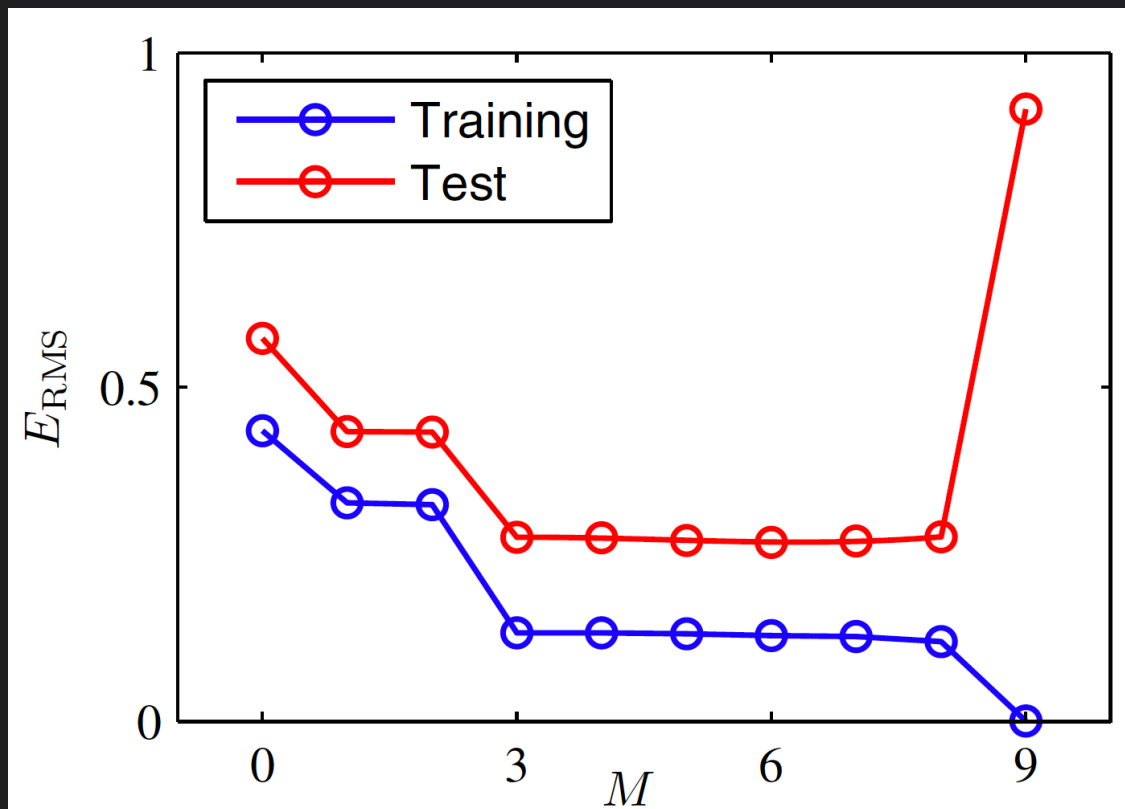
$$E_{RMS} = \sqrt{\frac{2}{N} E(\mathbf{w}^*)}$$

Comparable for different amounts of data

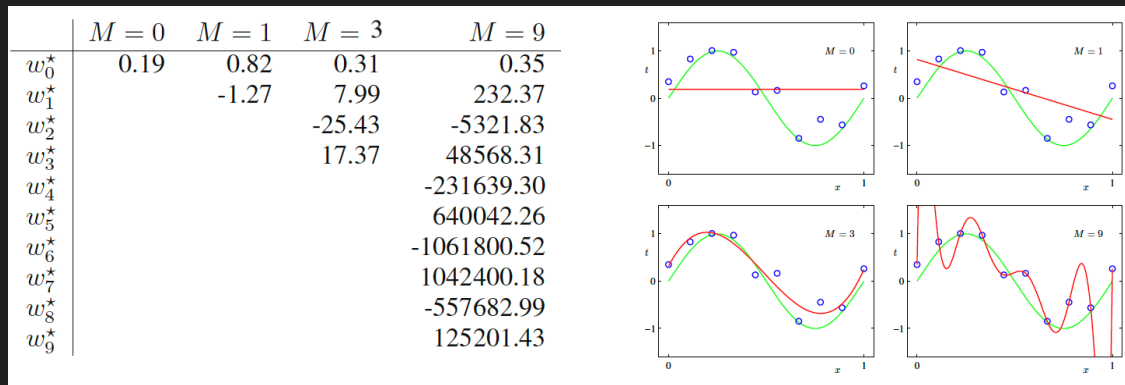
## H2 Avoiding Over-fitting

### H3 Indirect Evidence of Over-fitting

- Dramatic increase of  $E_{RMS}$  of training set and the difference between the  $E_{RMS}$  of training set and testing set as degree gets larger
- Magnitude of  $\mathbf{w}_i^*$  is extremely large



Over-fitting means we fail to **generalise** to un seen data



For  $M = 9$ , the magnitude of some  $w_i^*$  are very large, and the model makes some extreme predictions

**Dilemma:** Complex Models (more expressive) v. Over-fitting

H3 **Solution 1: Use More Data**

H3 **Solution 2: Regularisation**

Using a new **error function** that **penalises** extreme parameter values

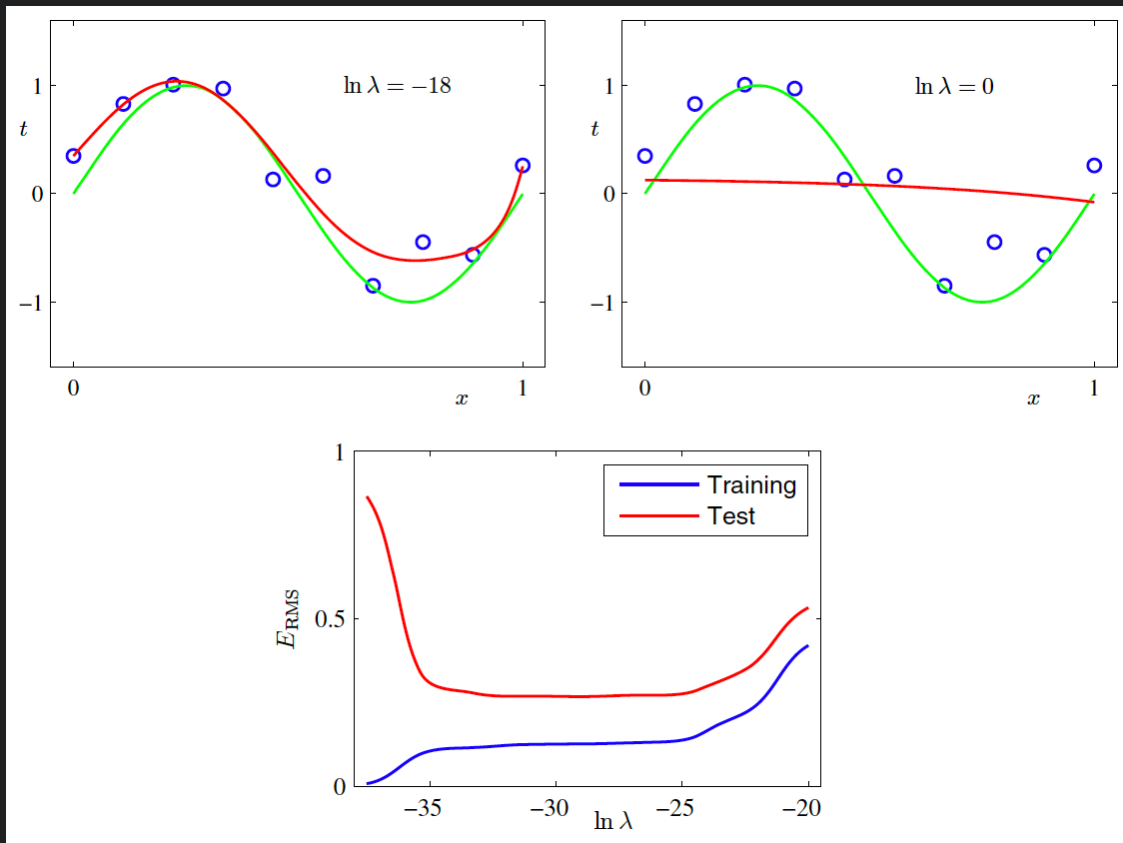
$$\tilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N [y(x_n, \mathbf{w}) - t_n]^2 + \frac{\lambda}{2}$$

Where

$$\|\mathbf{w}\|^2 = \mathbf{w}^T \mathbf{w} = w_0^2 + w_1^2 + \dots + w_M^2$$

Minimising **error function**

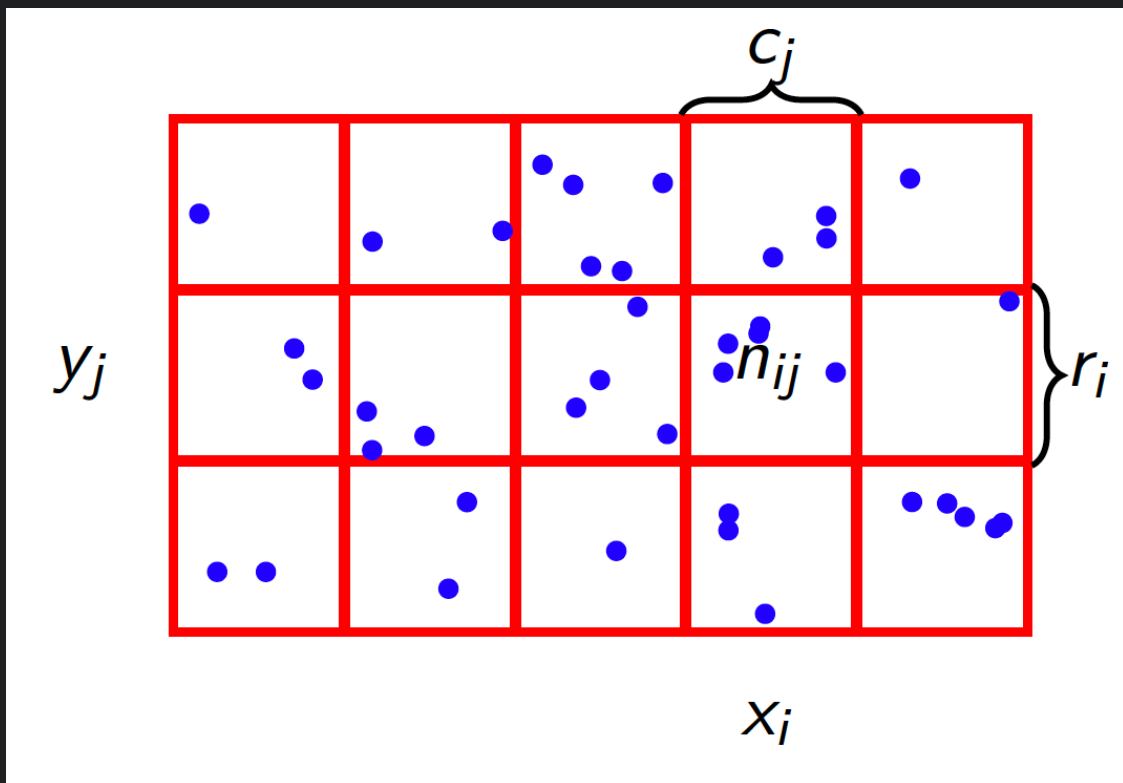
$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \tilde{E}(\mathbf{w})$$



Regularisation appears to control the effective complexity of the model, and hence the degree of overfitting.

## H2 Probability Theory

## H3 Frequentist Probability



$p_{XY}$  is **probability mass function** over values that random variables  $X$  and  $Y$  can take.

$$Pr(X = x_i, Y = y_j) = p_{XY}(x_i, y_j)$$

$X$  is a random variable that can take any value  $x_i$ , so does  $Y$  with  $y_i$

If we sample  $(X, Y)$  a large number of times  $N$ :

- $n_{ij}$  is the number of times  $X = x_i, Y = y_j$
- $c_i$  is the number of times  $X = x_i$
- $r_j$  is the number of times  $Y = y_j$

**Probability mass functions** capture the relative frequency of outcomes

### H3 Probability

**Marginal Probability:**

$$Pr(X = x_i) = p_X(x_i) = \frac{c_i}{N}$$

**Joint Probability:**

$$Pr(X = x_i, Y = y_j) = p_{XY}(x_i, y_j) = \frac{n_{ij}}{N}$$

**Conditional Probability:**

$$Pr(Y = y_j | X = x_i) = p_{Y|X}(y_j | x_i) = \frac{n_{ij}}{c_i}$$

### H3 Rule

**Sum Rule:**

$$p_X(x) = \frac{c_i}{N} = \frac{1}{N} \sum_j n_{ij} = \sum_j p_{xy}(x_i, y_j)$$

**Product Rule:**

$$p_{XY}(x_i, y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \times \frac{c_i}{N} = p_{Y|X}(y_j | x_i) p_X(x_i)$$

### H4 Application in 3 Random Variables Case

$$p(x, y) = \sum_z p(x, y, z)$$

$$\begin{aligned} p(x, y, z) &= p(x, y | z) p(z) \\ &= p(y, z | x) p(x) \end{aligned}$$

If  $p_{XY}(x, y) = p_X(x) p_Y(y)$ , we say  $X$  and  $Y$  are **independent**

### H4 Application of Probability Rule

**Random Variables:**

- **A** disease status ( *ill* or *healthy* )
- **B** blood test ( *+ve* or *-ve* )

$$p_{AB}(a, b) = p_{A|B}(a|b)p_B(b) = p_{B|A}(b|a)p_A(a)$$

$$\begin{aligned} p_A(ill) &= Pr(\text{person has disease}) = 1\% \\ p_B(+ve) &= Pr(\text{person has +ve blood test}) = 10\% \\ p_{B|A}(+ve|ill) &= Pr(\text{blood test is +ve given person is ill}) = 70\% \\ p_{A|B}(ill|+ve) &= Pr(\text{person is ill given blood test is +ve}) = 7\% \end{aligned}$$

### H3 Reasoning

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

$p(x) = \sum_y p(x|y)p(y)$  **normalises** the equation

Practically:

$$\begin{aligned} p(y|x) &\propto p(x|y)p(y) \\ \text{posterior} &\propto \text{likelihood} \times \text{prior} \end{aligned}$$

### H2 Frequentist v. Bayesian

In the frequentist perspective, probability distributions represent **expected outcomes given a large number of trials**, e.g.

$$E[X] = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_n x_n$$

**Bayesian Inference** involves shifting the perspective in order to reason about vents that may **happen only once**, in which **probability is a measure of belief**