EE211: Robotic Perception and Intelligence

Lecture 9 Probability Distributions

Jiankun WANG

Department of Electronic and Electrical Engineering
Southern University of Science and Technology

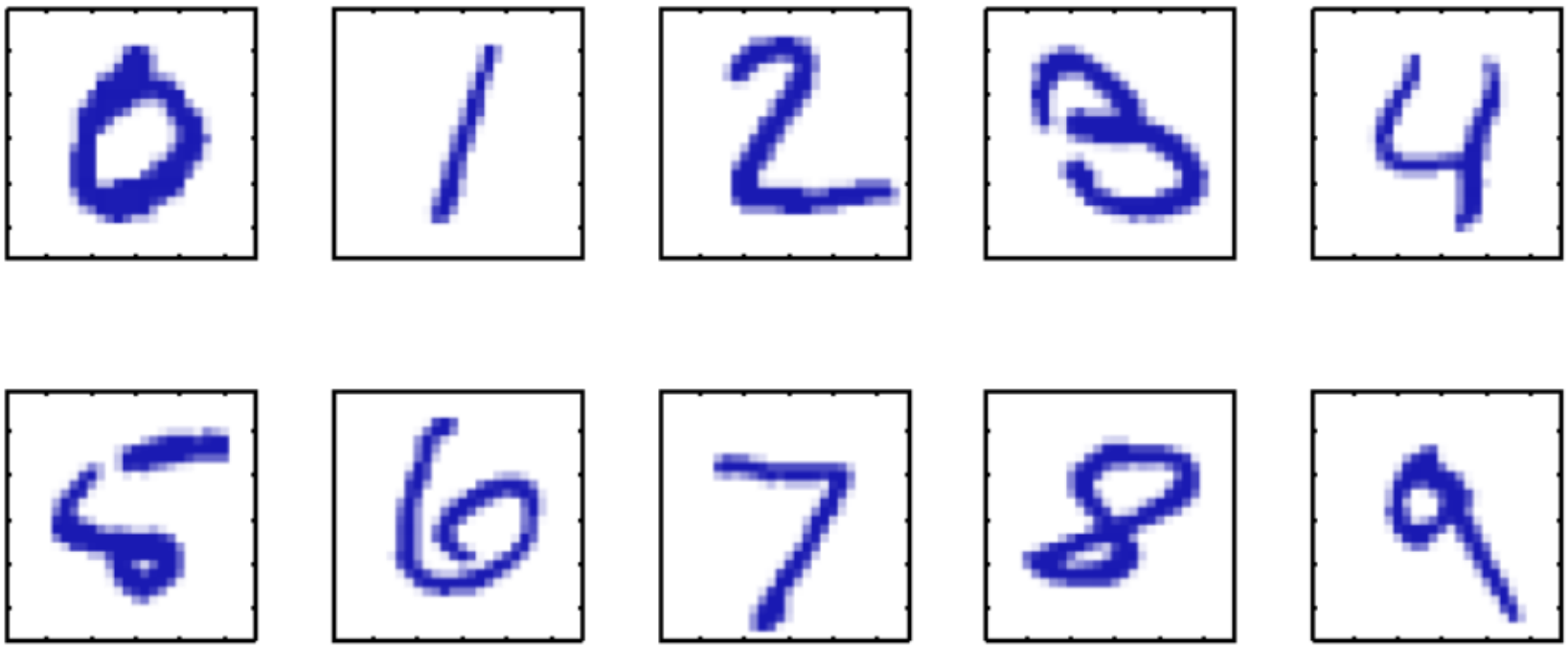
Undergraduate Course, Nov 2024

Outline

Outline

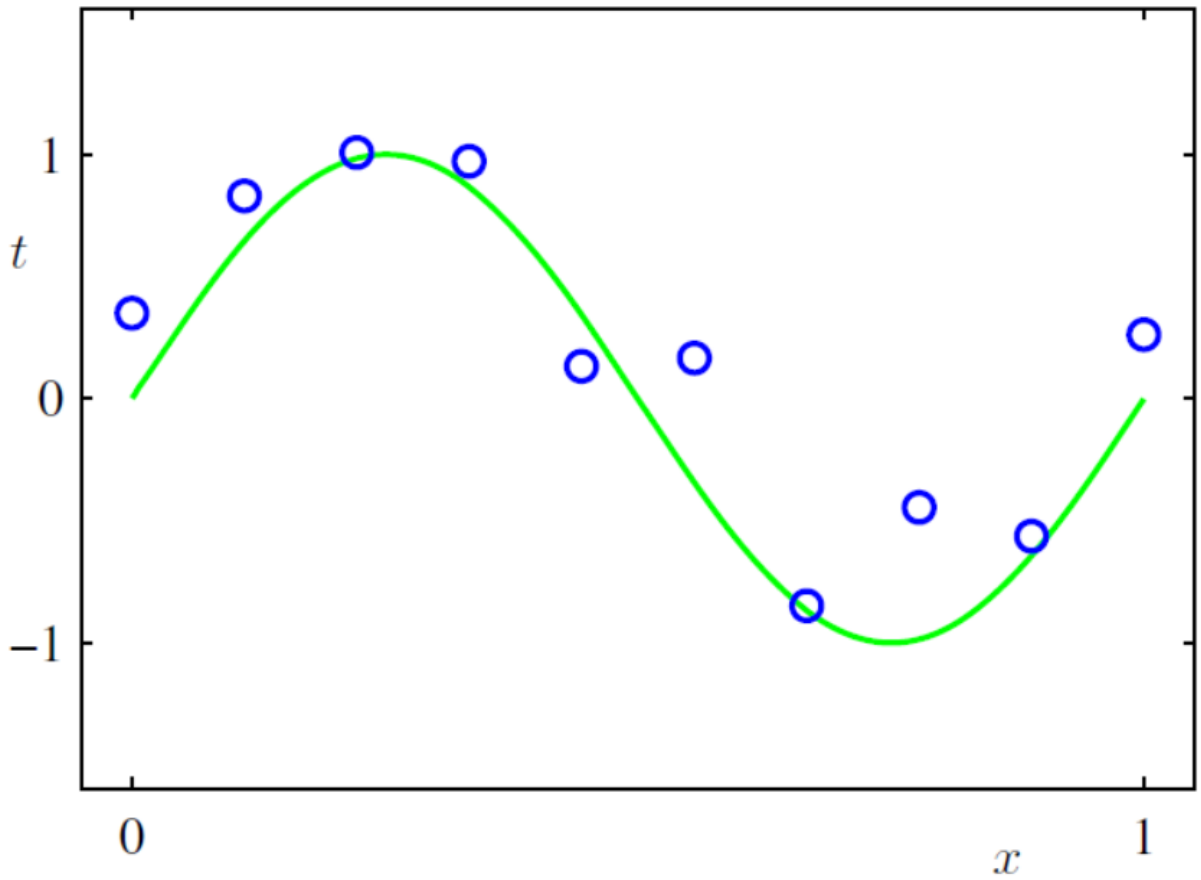
Example

Handwritten Digit Recognition



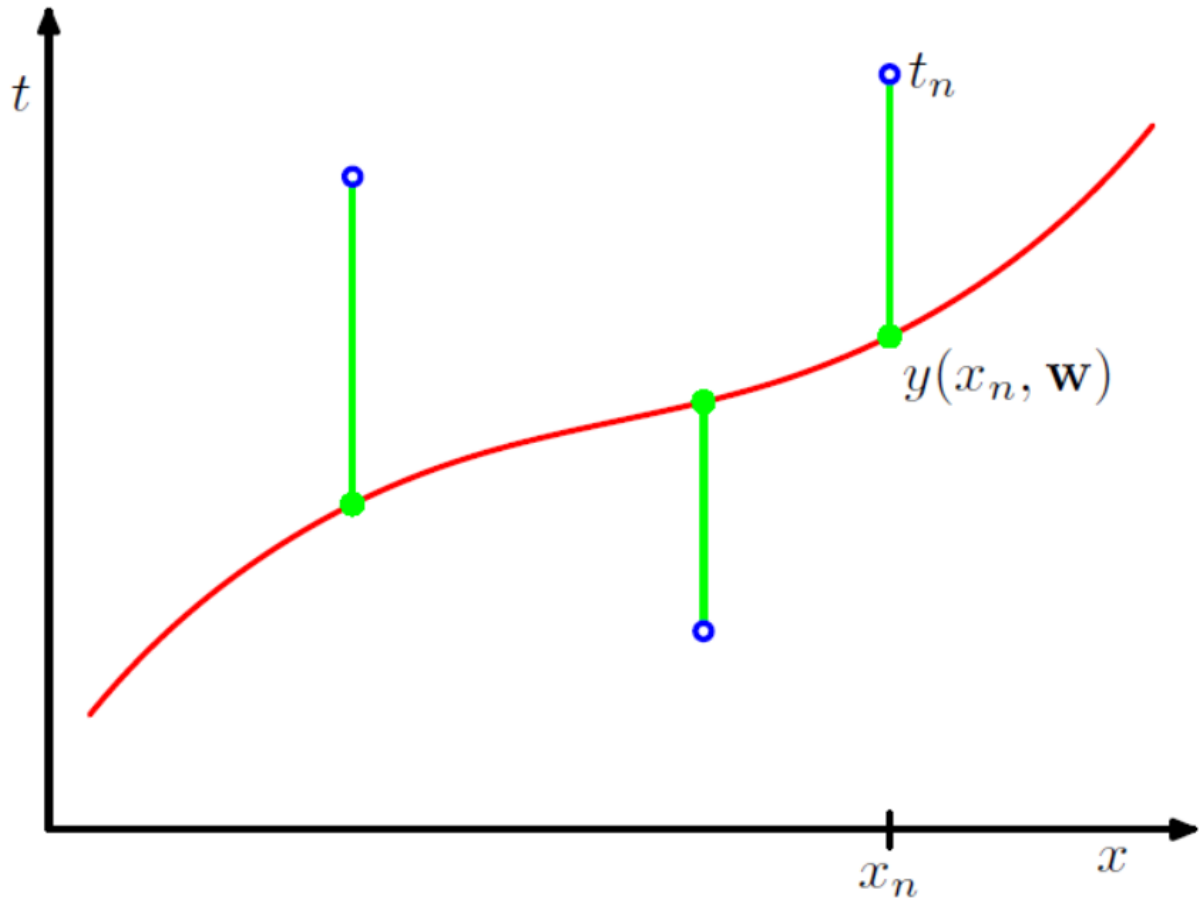
Chris Bishop, Pattern Recognition and Machine Learning.

Polynomial Curve Fitting



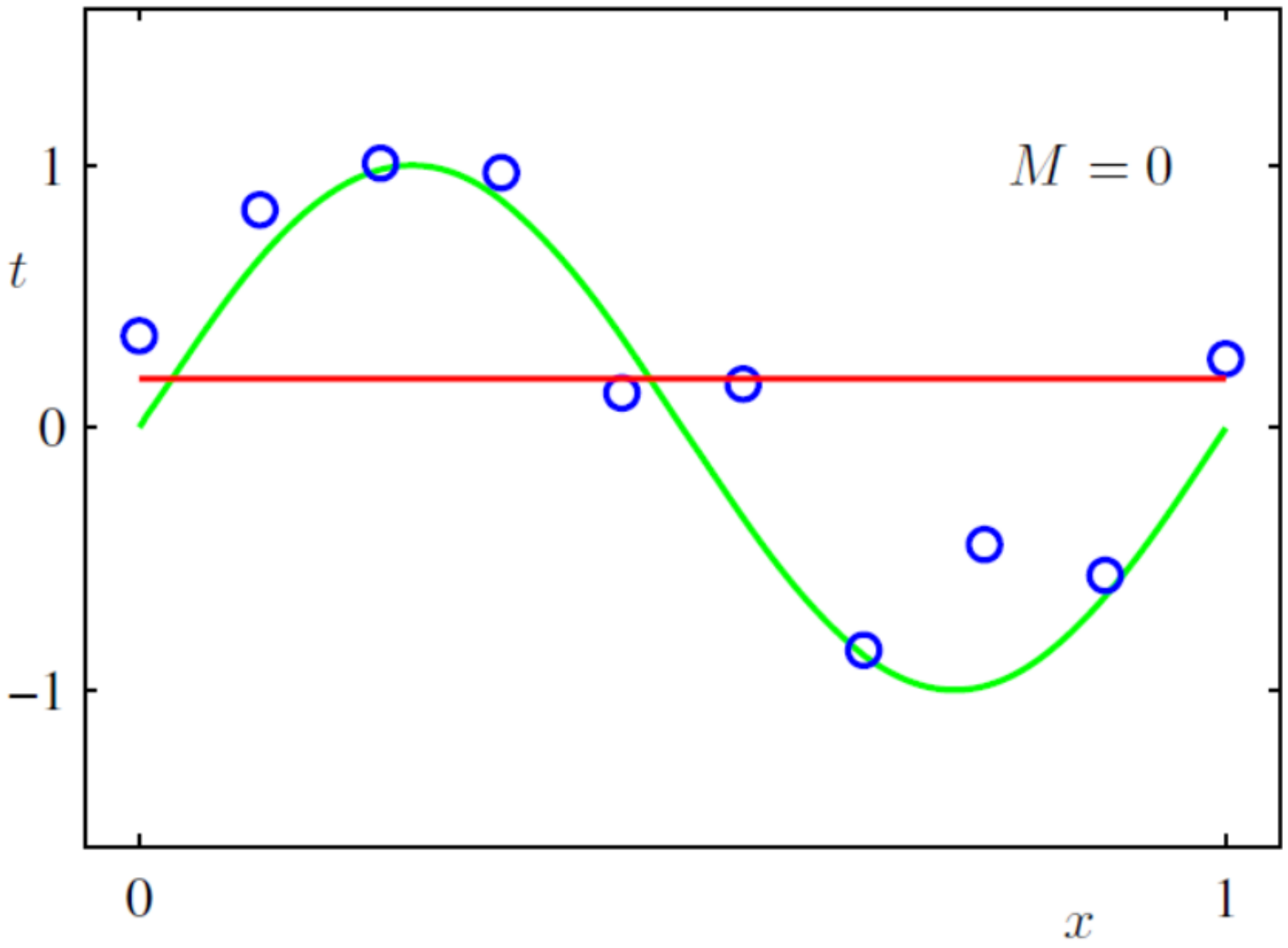
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Sum-of-Squares Error Function

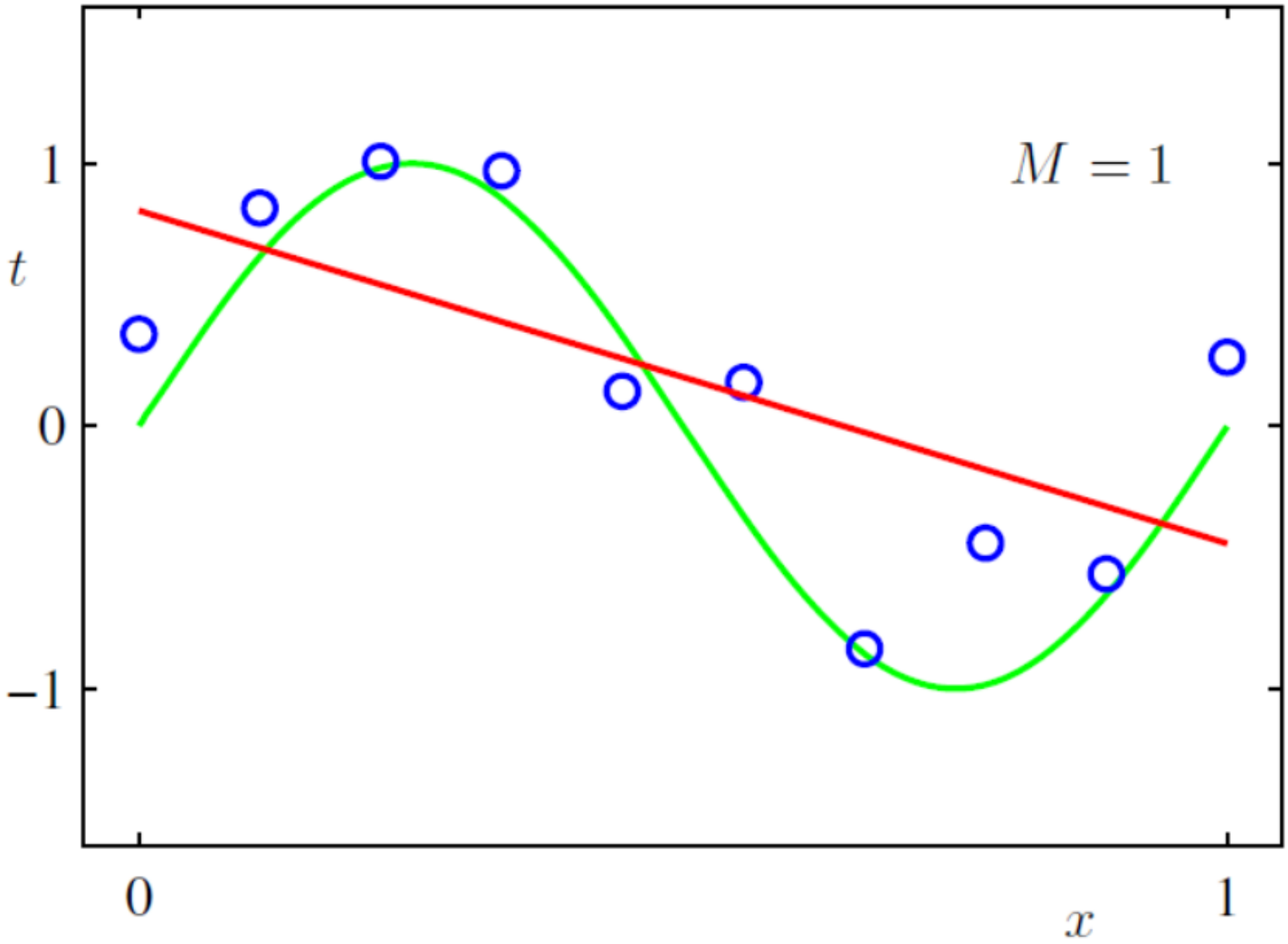


$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

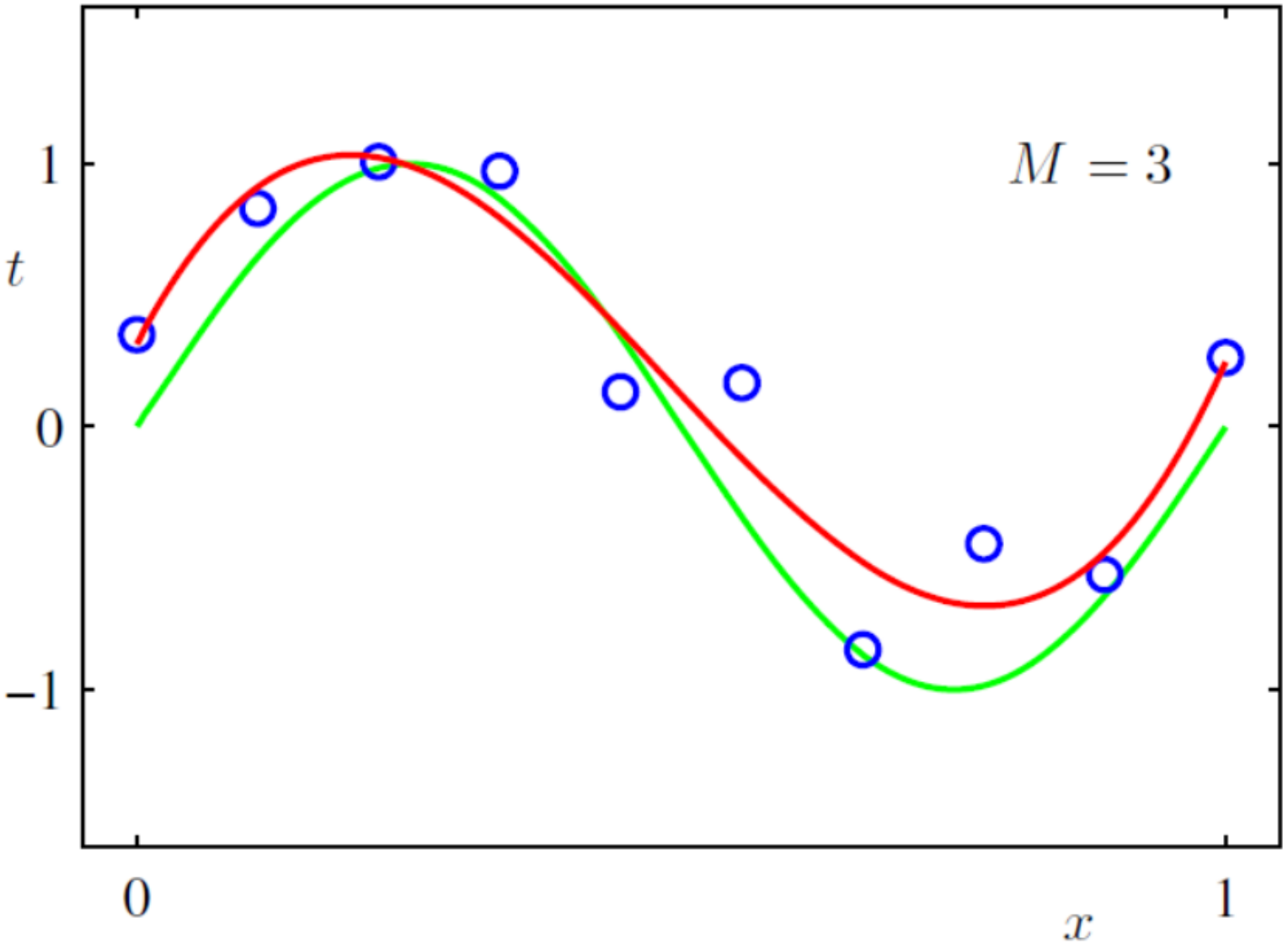
0^{th} Order Polynomial



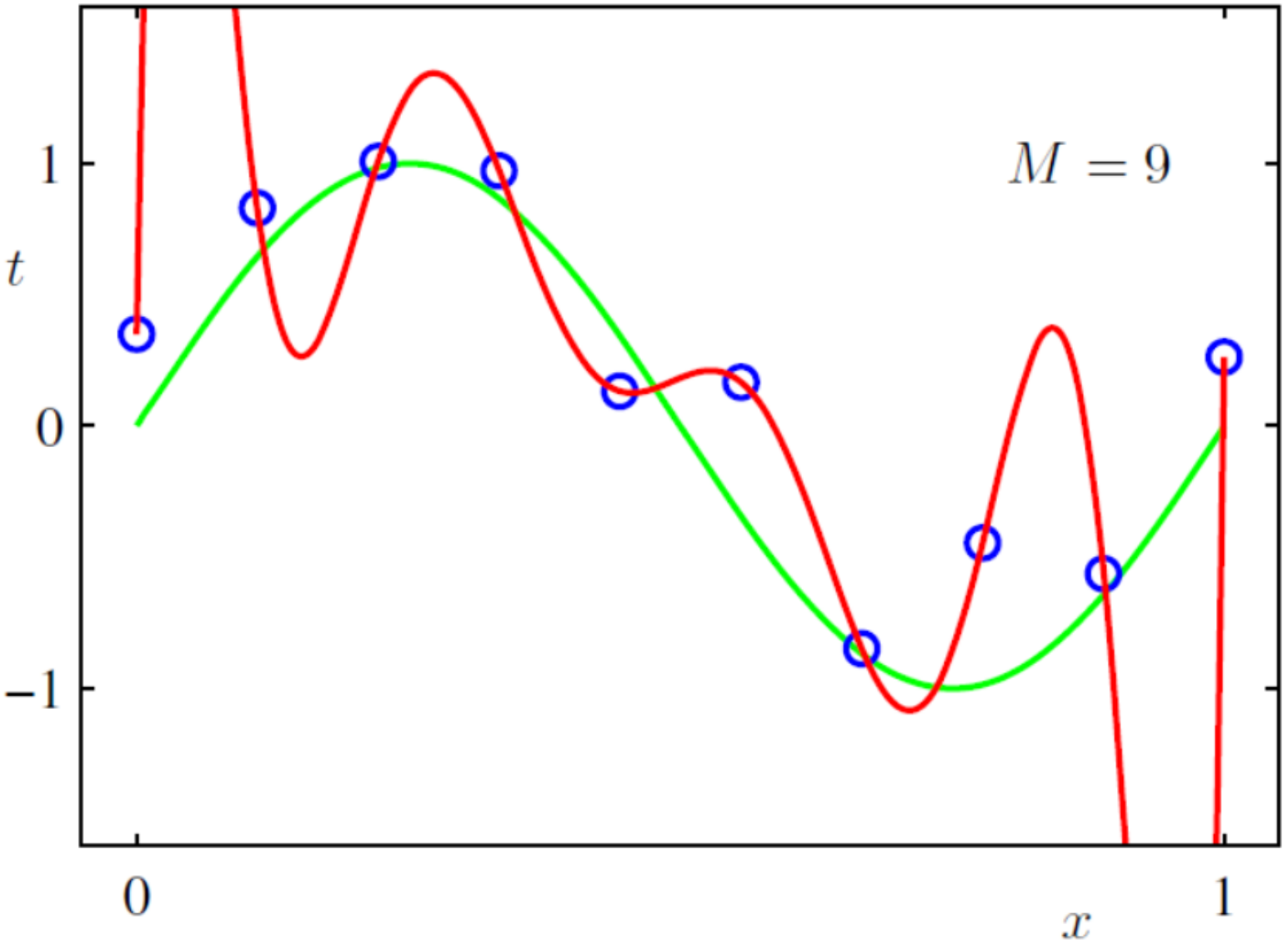
1st Order Polynomial



3rd Order Polynomial

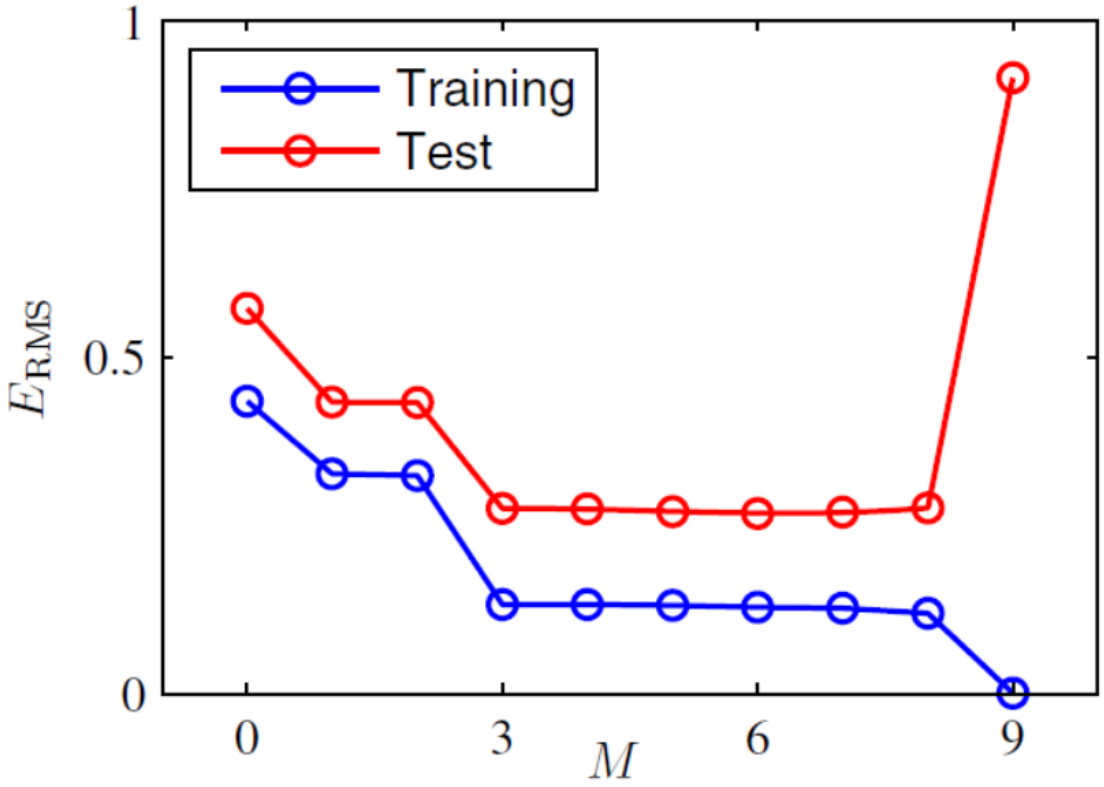


9^{th} Order Polynomial



Over-fitting

Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E\left(\mathbf{w}^{\star}\right)/N}$



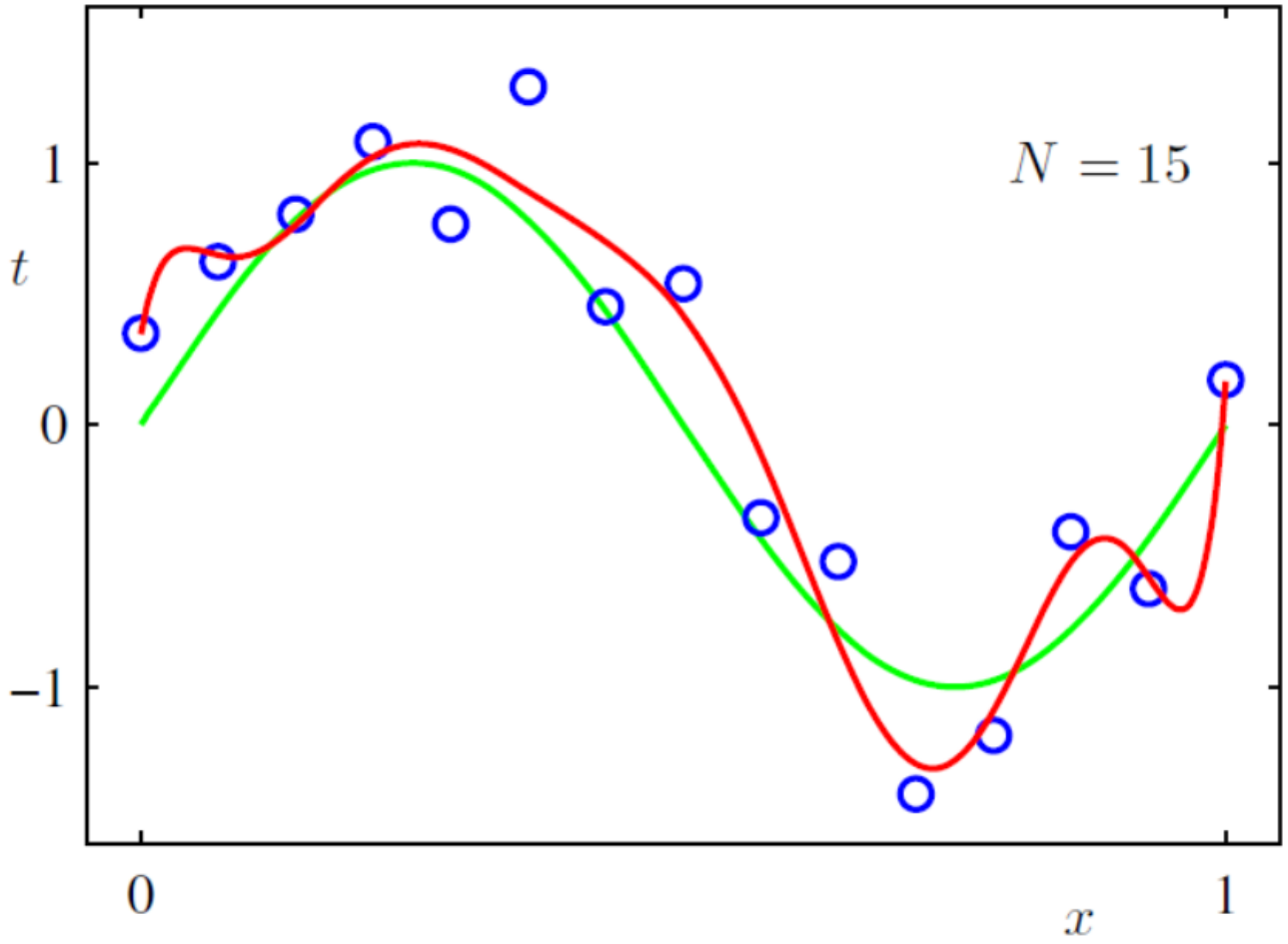
The division by N allows us to compare different sizes of data sets on an equal footing, and the square root ensures that E_{RMS} is measured on the same scale (and in the same units) as the target variable t.

Polynomial Coefficients

	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

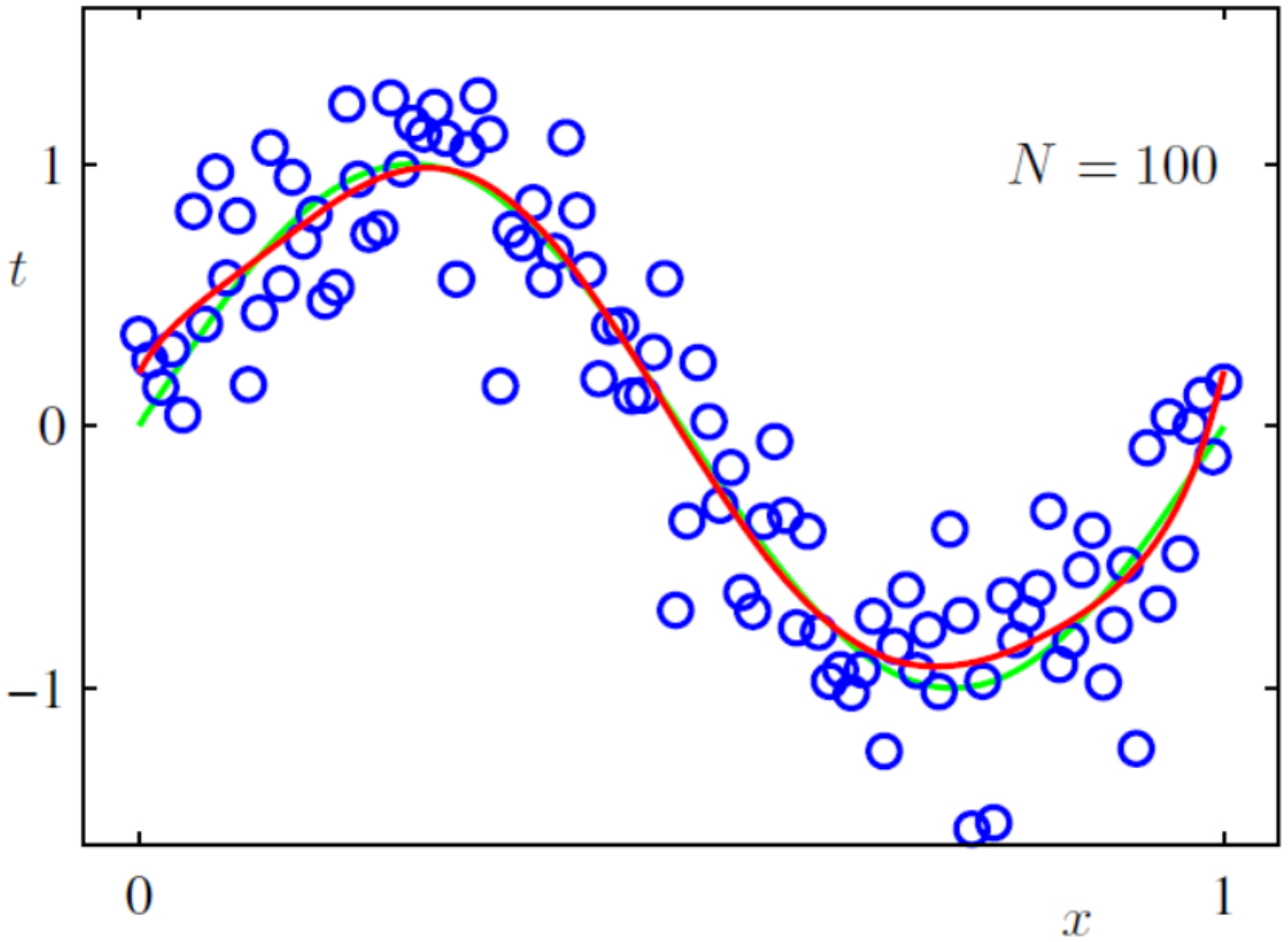
Data Set Size: $N = 15$

9^{th} Order Polynomial



Data Set Size: $N = 100$

9^{th} Order Polynomial

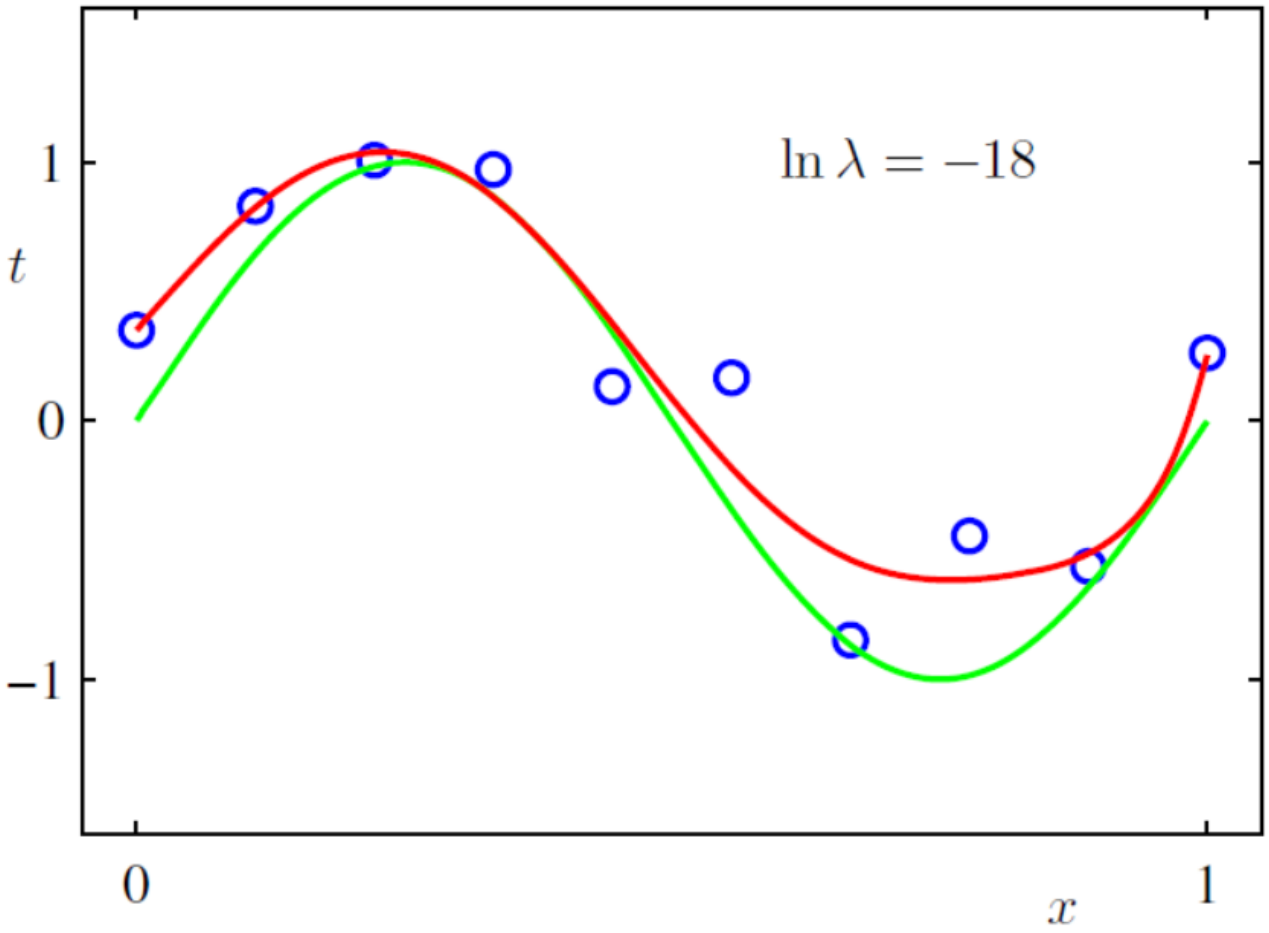


Regularization

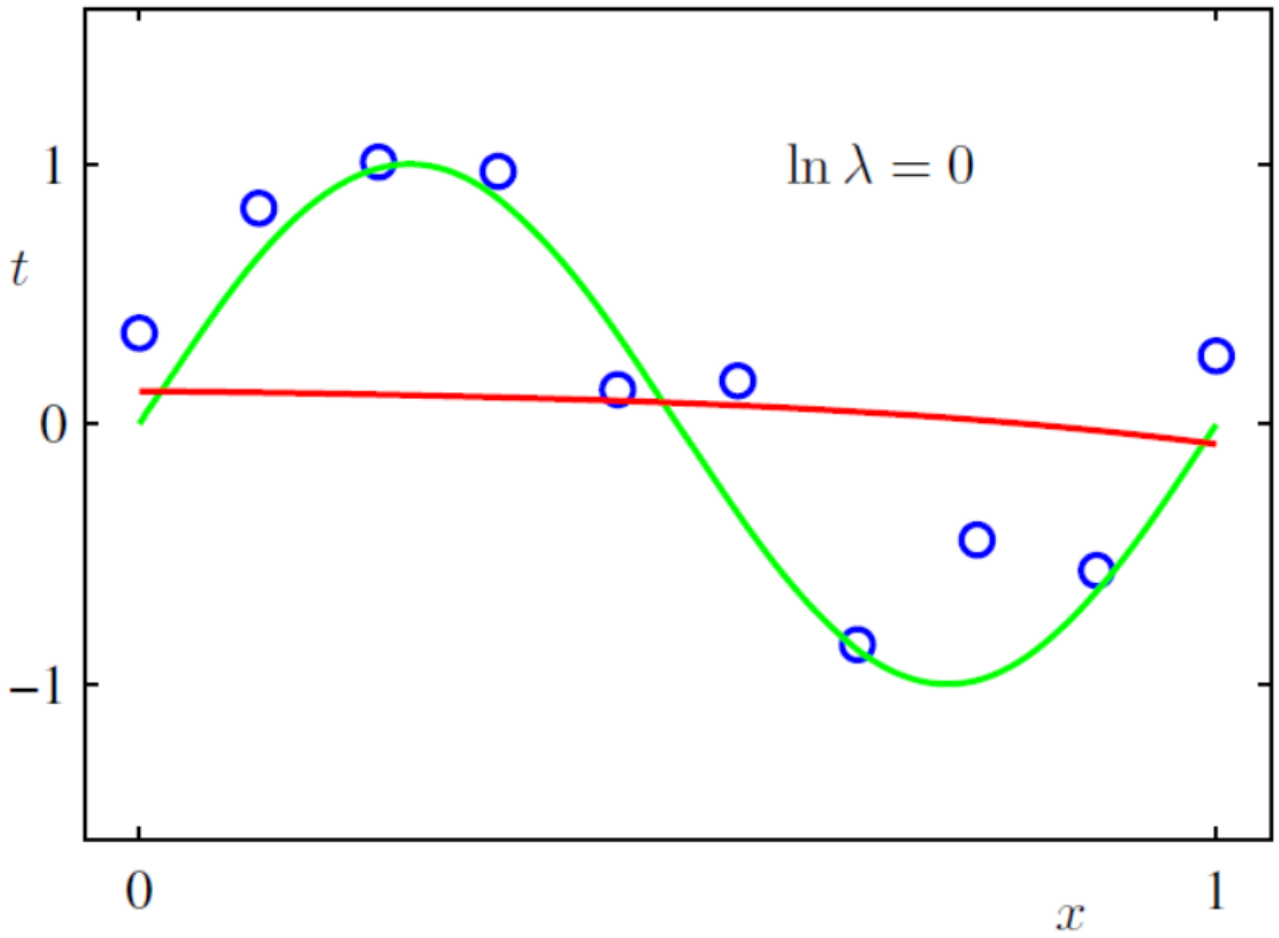
Penalize large coefficient values

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

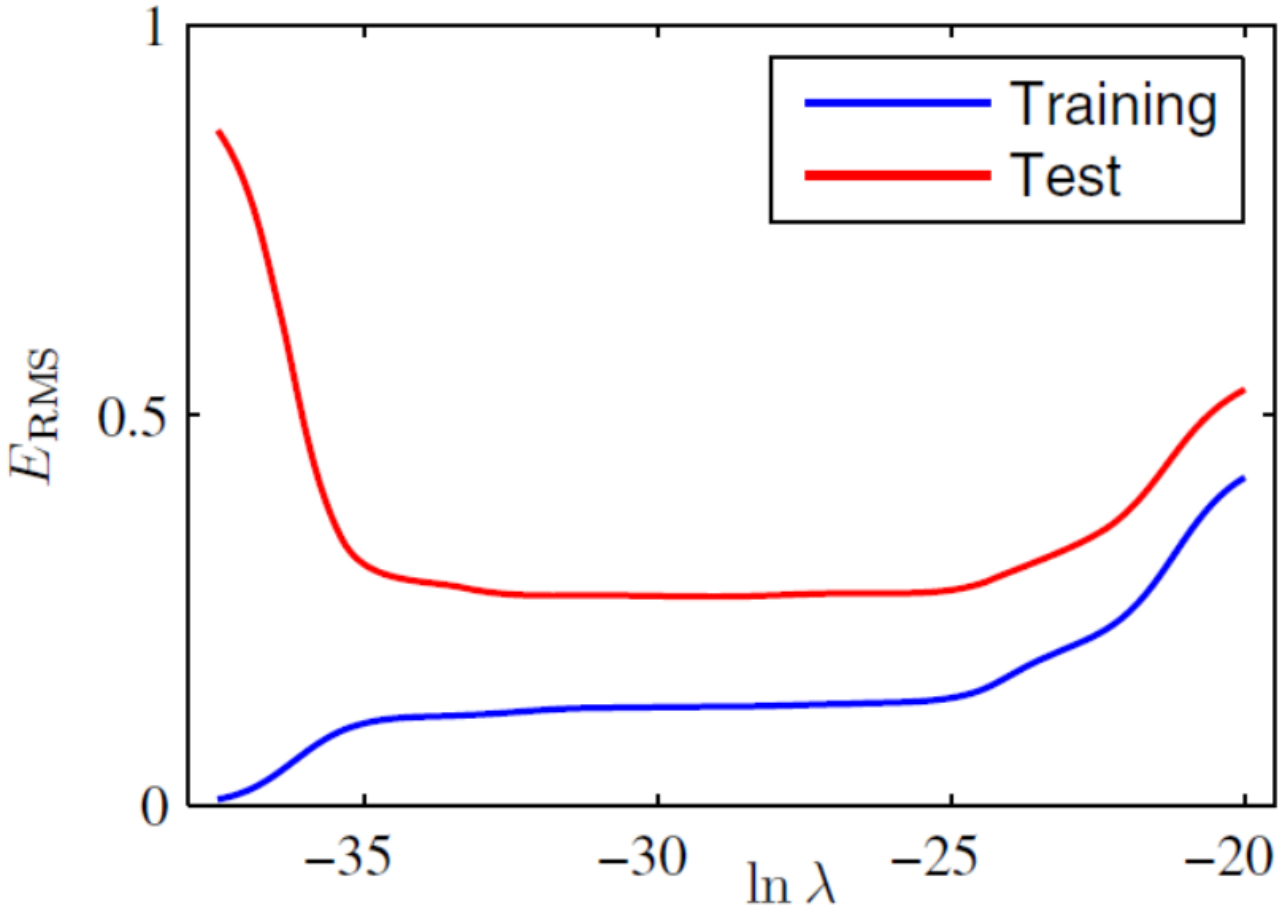
Regularization: $\ln \lambda = -18$



Regularization: $\ln \lambda = 0$



Regularization: E_{RMS} vs. $\ln \lambda$

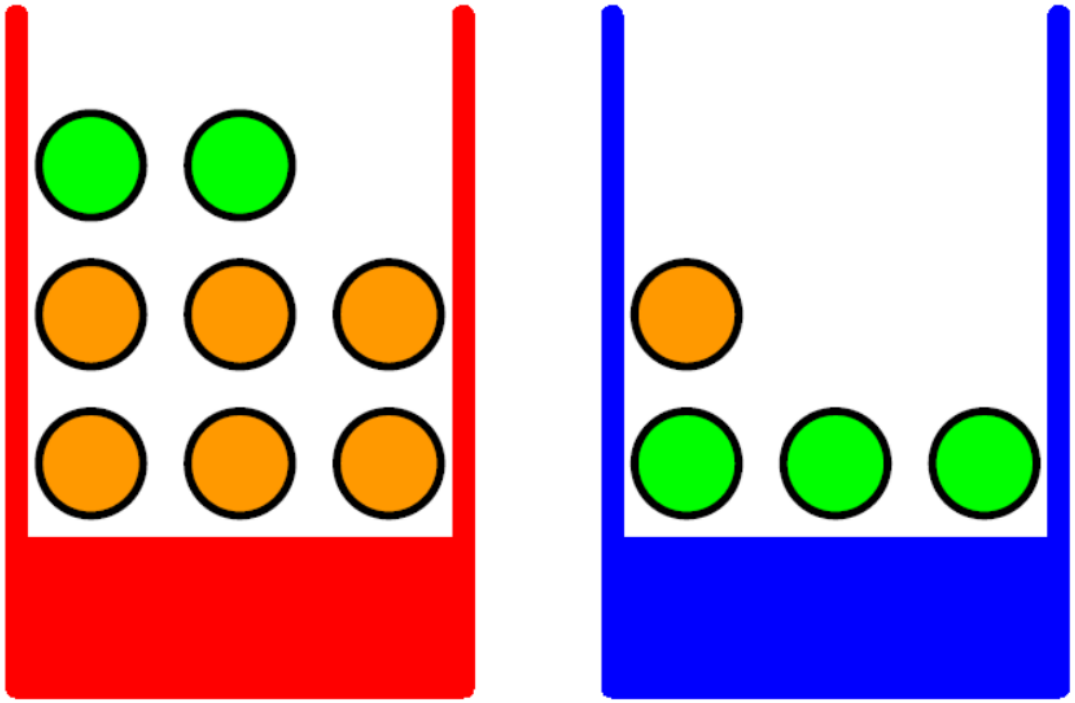


Polynomial Coefficients

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

Outline

Apples and Oranges



Probability Theory

Joint Probability:

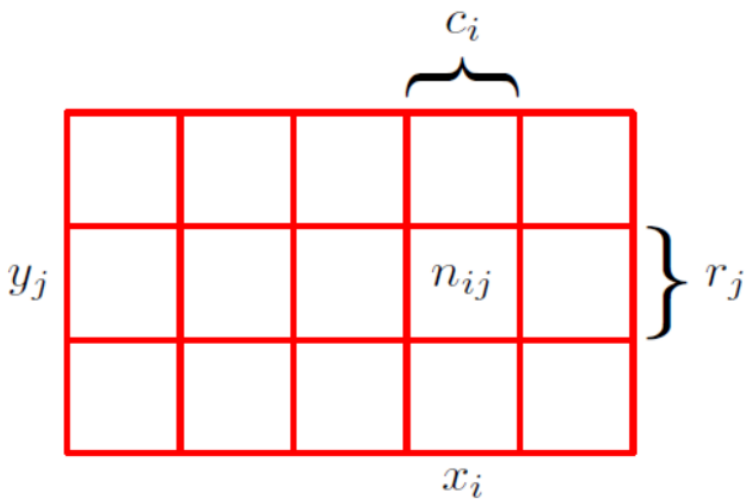
$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Marginal Probability:

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

Conditional Probability:

$$p(Y = y_j \mid X = x_i) = \frac{n_{ij}}{c_i}$$



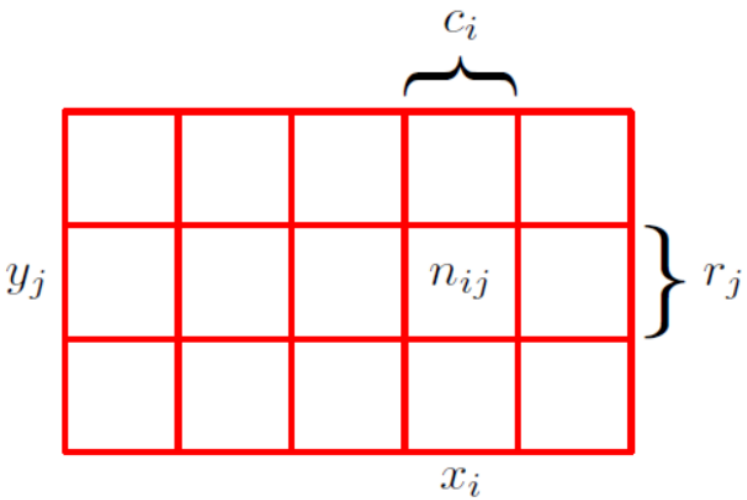
Probability Theory

Sum Rule:

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N} \sum_{j=1}^{L} n_{ij}$$
$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

Product Rule:

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$
$$= p(Y = y_j \mid X = x_i)\, p(X = x_i)$$



The Rules of Probability

Sum Rule

$$p(X) = \sum_{Y} p(X, Y)$$

Product Rule

$$p(X, Y) = p(Y \mid X)p(X)$$

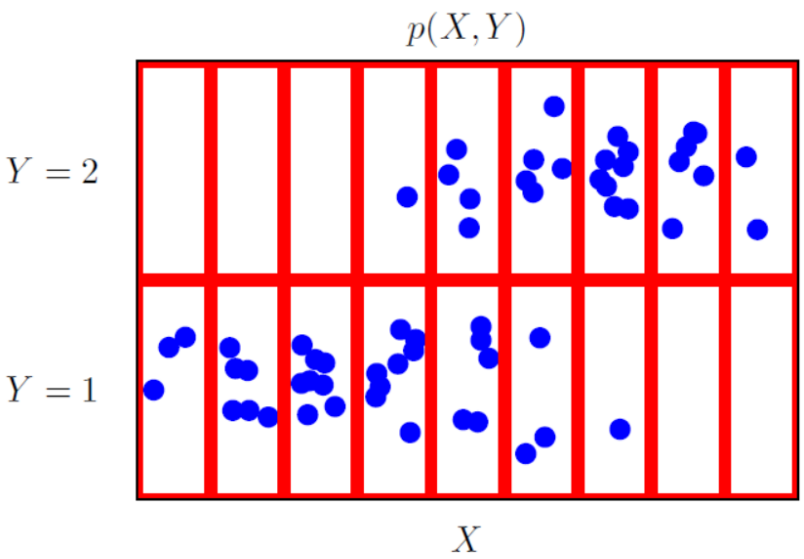
Bayes' Theorem

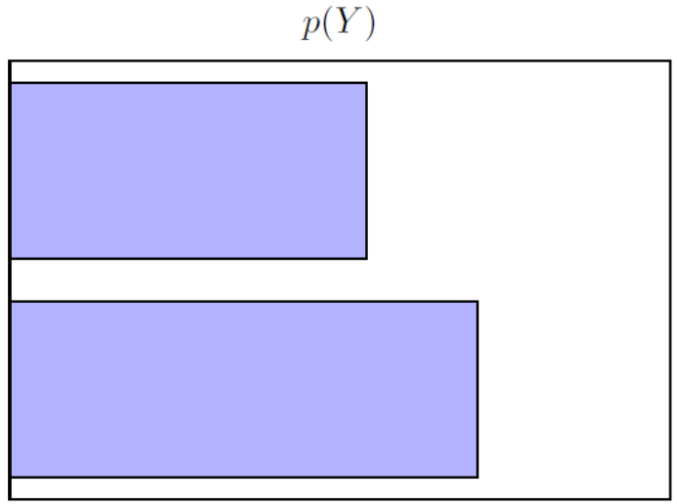
$$p(Y \mid X) = \frac{p(X \mid Y)p(Y)}{p(X)}$$

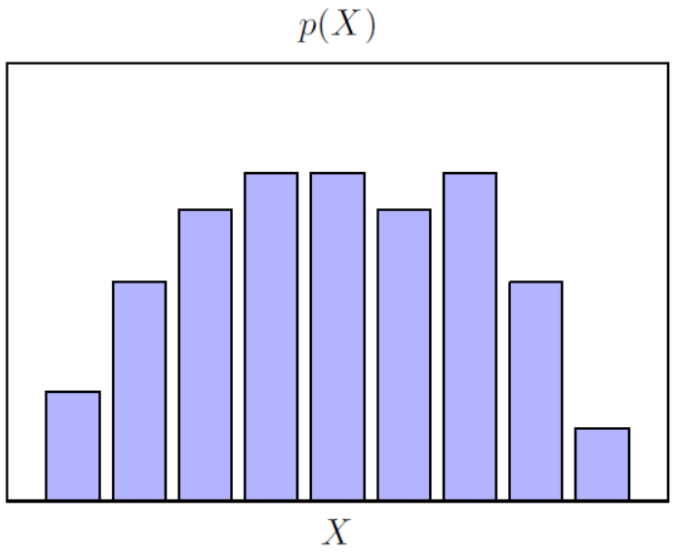
$$p(X) = \sum_{Y} p(X \mid Y)p(Y)$$

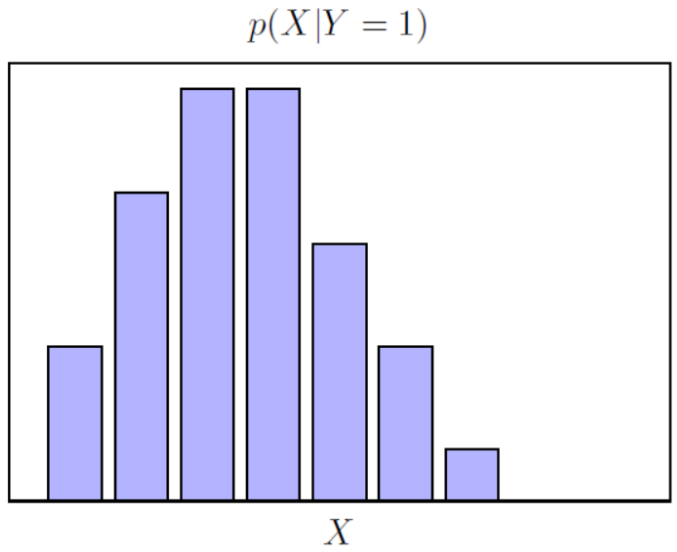
posterior $\propto$ likelihood $\times$ prior

Probability Densities

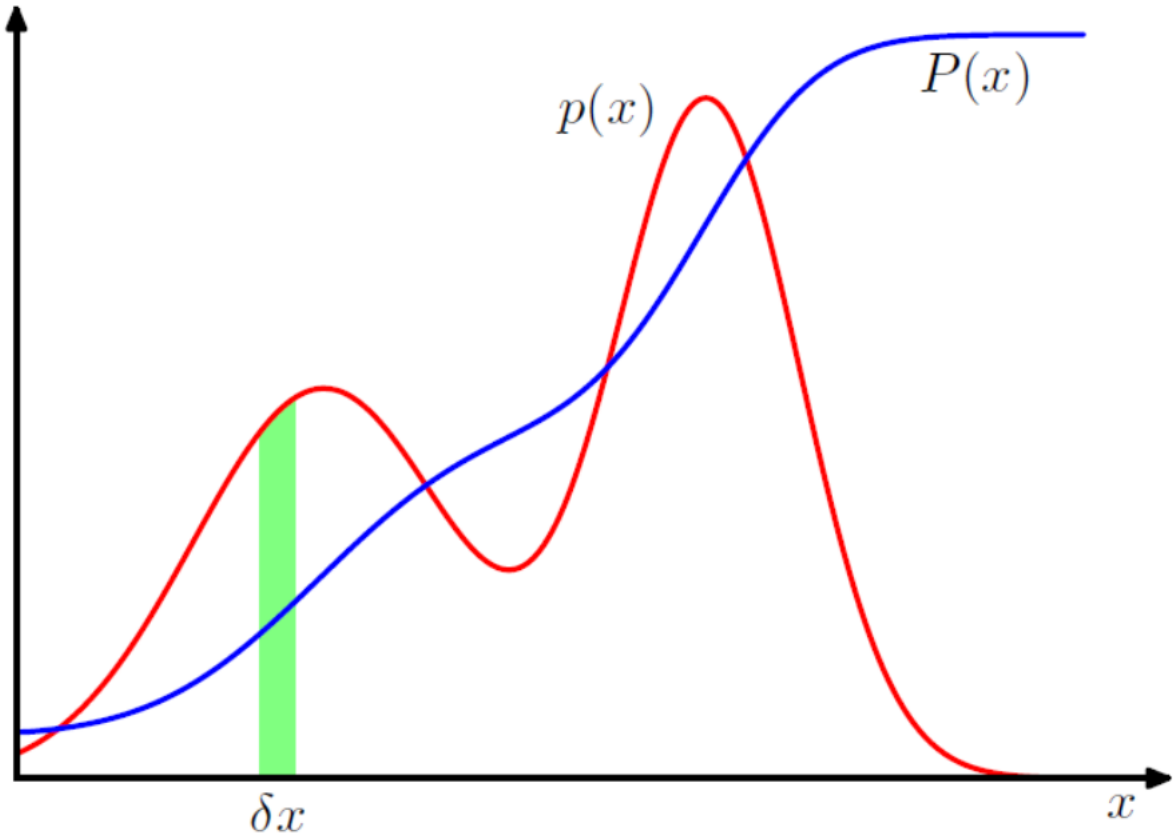








Probability Densities



Probability density: $p(x \in (a, b)) = \int_a^b p(x)\mathrm{d}x$.
$p(x)$ satisfies two conditions: $p(x) \geq 0, \quad \int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1$.
Cumulative distribution function: $P(z) = \int_{-\infty}^{z} p(x)\mathrm{d}x$.

Expectations

- Discrete Expectation: $\mathbb{E}[f] = \sum_x p(x)f(x)$
- Continuous Expectation: $\mathbb{E}[f] = \int p(x)f(x)\mathrm{d}x$
- Approximate Expectation: $\mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n)$
- Conditional Expectation: $\mathbb{E}_x[f \mid y] = \sum p(x \mid y)f(x)$

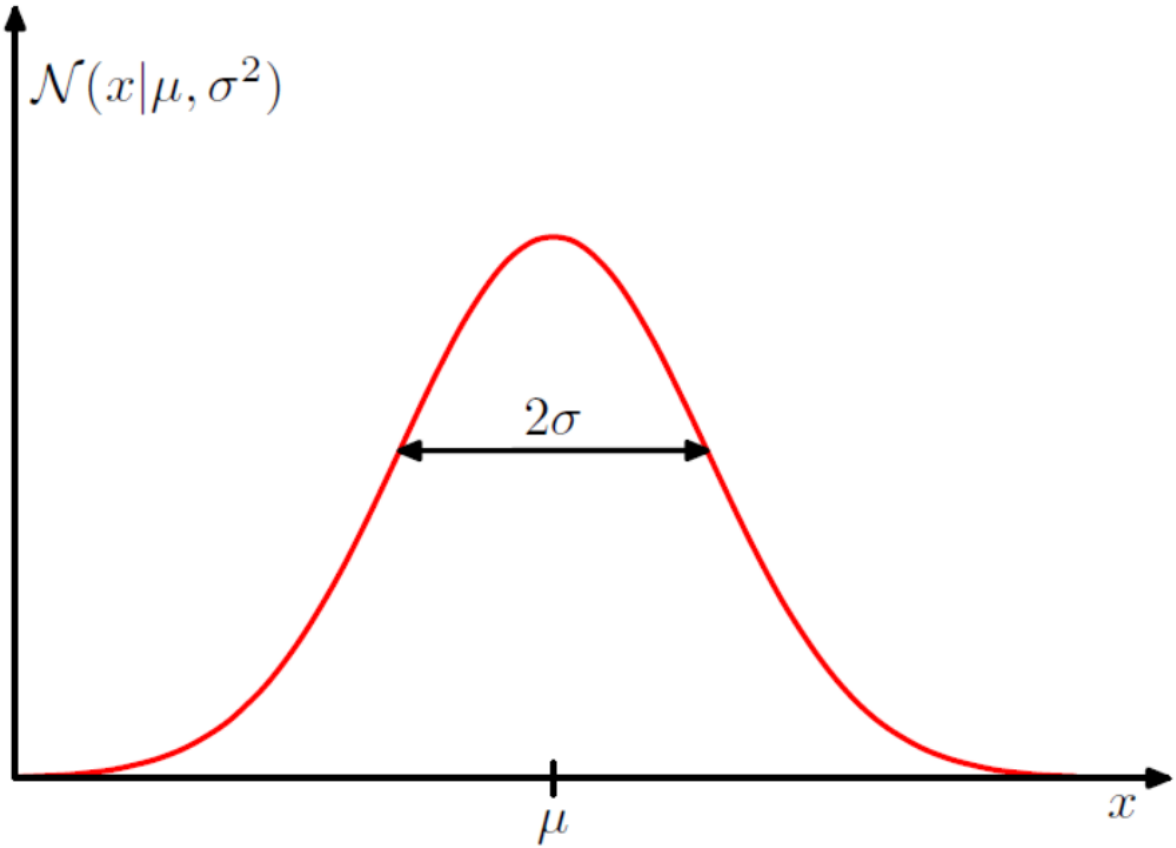
Variances and Covariances

- Variance of f: $\text{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right]$
- $\text{var}[f] = \mathbb{E}\left[f(x)^2\right] - \mathbb{E}[f(x)]^2$
- In particular, we can consider the variance of the variable x itself, which is given by $\text{var}[x] = \mathbb{E}\left[x^2\right] - \mathbb{E}[x]^2$
- Covariance of x and y: $\text{cov}[x, y] = \mathbb{E}[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}]$
- $\text{cov}[x, y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]$

Exercises

- Use the definition $\text{var}[x] = \mathbb{E}\left[(x - \mathbb{E}[x])^2\right]$ to show that $\text{var}[x]$ satisfies $\text{var}[x] = \mathbb{E}\left[x^2\right] - \mathbb{E}[x]^2$.
- Show that if two variables x and y are independent, then their covariance is zero.

The Gaussian Distribution



$$\mathcal{N}\left(x \mid \mu, \sigma^2\right) = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$\mathcal{N}\left(x \mid \mu, \sigma^2\right) > 0, \quad \int_{-\infty}^{\infty} \mathcal{N}\left(x \mid \mu, \sigma^2\right) \mathrm{d}x = 1$$

Gaussian Mean and Variance

- Mean: $\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x \mid \mu, \sigma^2\right) x \, \mathrm{d}x = \mu$
- Second order moment: $\mathbb{E}\left[x^2\right] = \int_{-\infty}^{\infty} \mathcal{N}\left(x \mid \mu, \sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2$
- Variance: $\mathrm{var}[x] = \mathbb{E}\left[x^2\right] - \mathbb{E}[x]^2 = \sigma^2$
- For a D-dimensional vector:

$$\mathcal{N}(\mathrm{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathrm{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathrm{x} - \boldsymbol{\mu})\right\}$$

Exercises

- For a given univariate Gaussian distribution, show that $\mathbb{E}[x] = \mu$ and $\text{var}[x] = \sigma^2$.
- For a given univariate Gaussian distribution, show that the maximum of the Gaussian distribution is obtained when $x = \mu$.

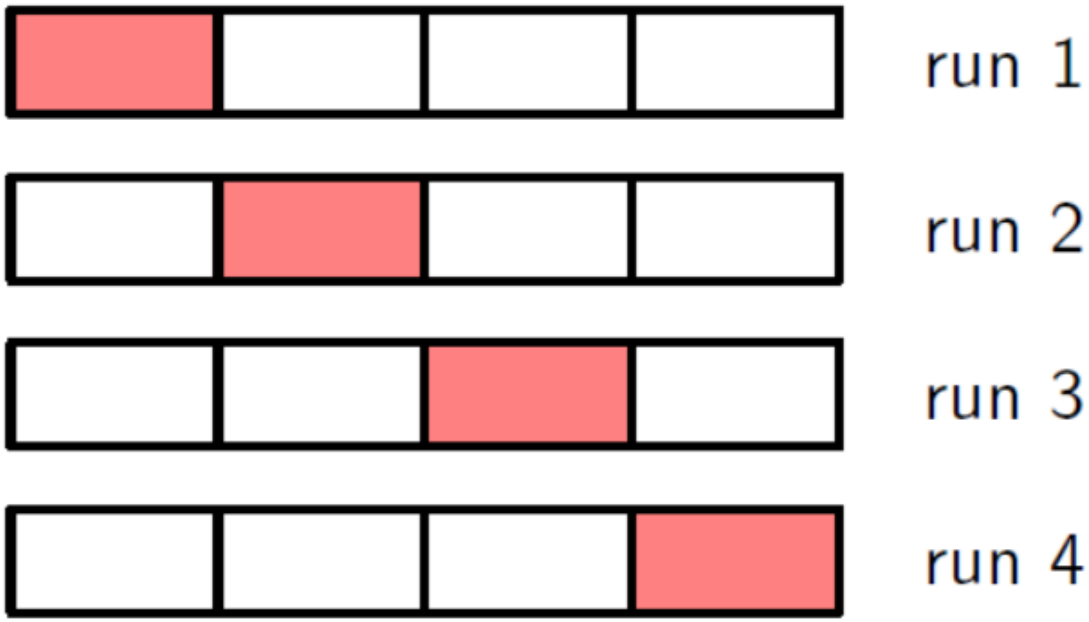
Outline

Model Selection

- Which is the optimal order of the polynomial that gives the best generalization?
- Train a range of models and test them on an independent validation set.
- Cross-validation: use a subset for training and the whole set for assessing the performance.

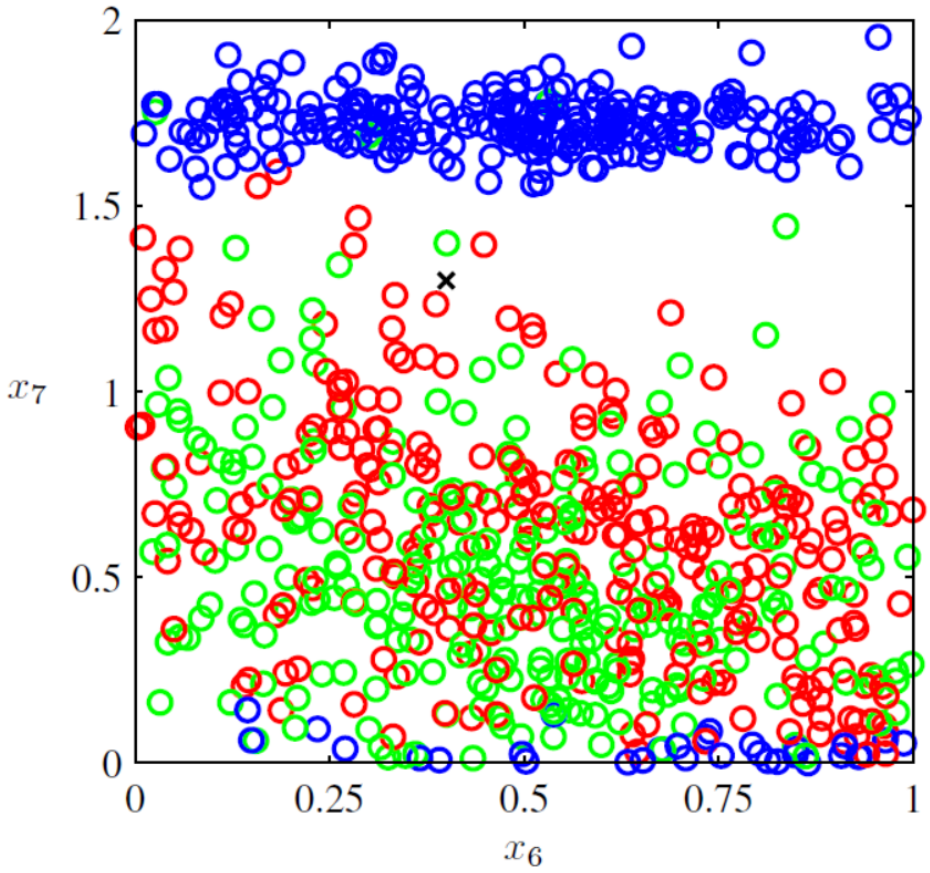
Model Selection

Cross-Validation



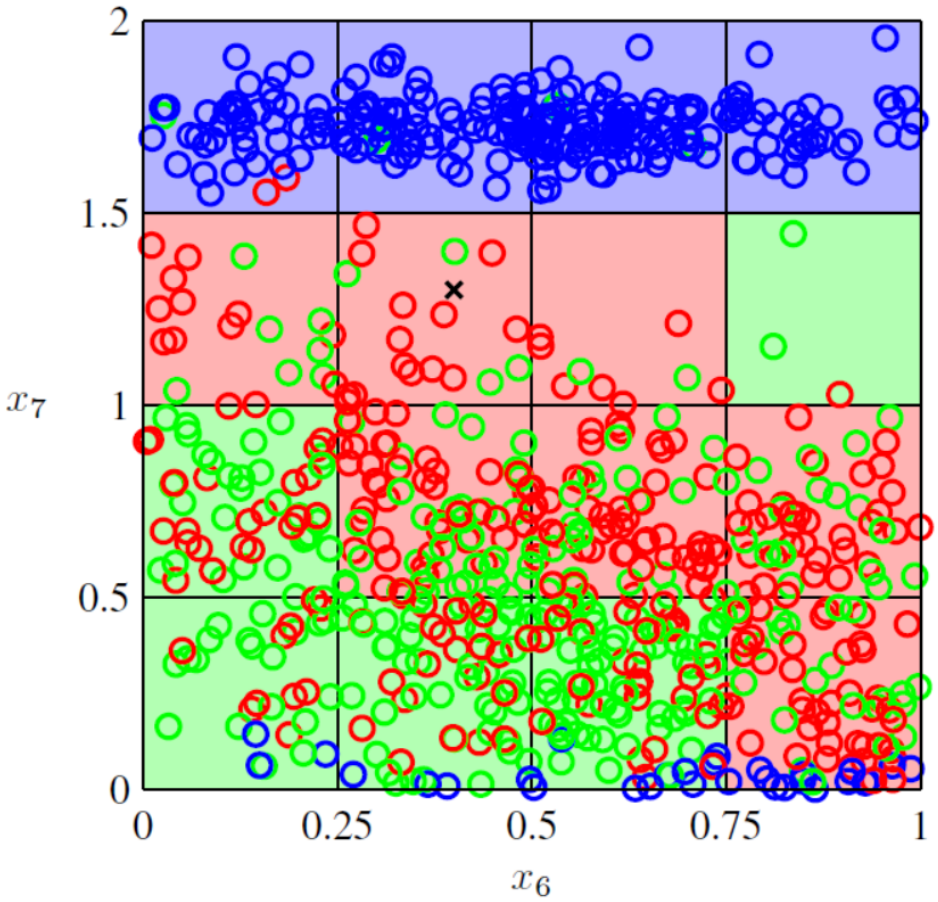
Outline

Curse of Dimensionality



Scatter plot of the oil flow data for input variables x_6 and x_7. Our goal is to classify the new test point denoted by '$\times$'.

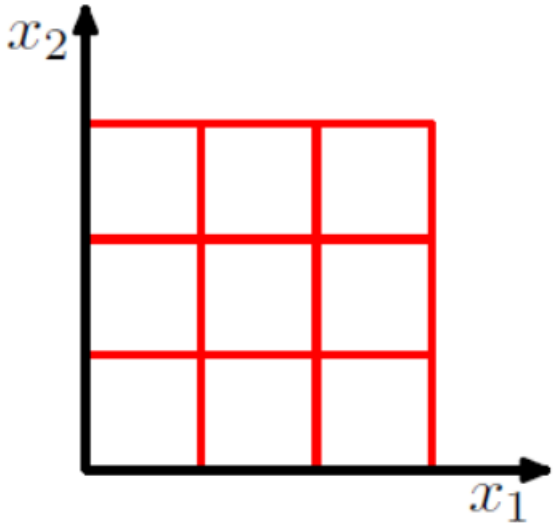
Curse of Dimensionality



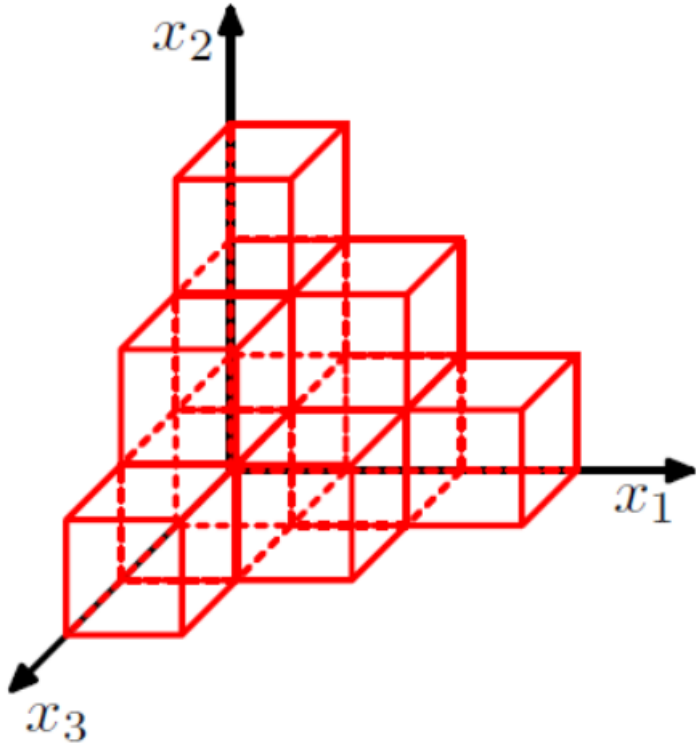
The input space is divided into cells and any new test point is assigned to the class that has a majority number of representatives in the same cell as the test point.

Curse of Dimensionality

$D = 1$



$D = 2$



$D = 3$

Curse of Dimensionality

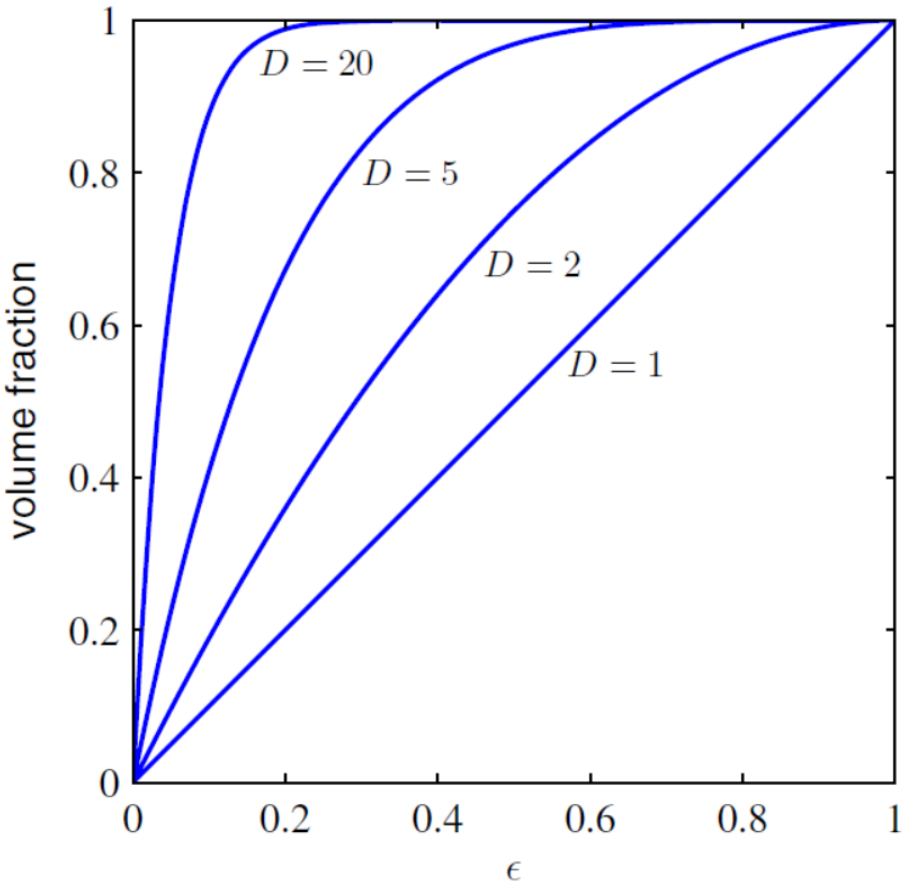
- Consider a sphere of radius $r = 1$ in a space of D dimensions, and ask what is the fraction of the volume of the sphere that lies between radius $r = 1 - \epsilon$ and $r = 1$.
- We can evaluate this fraction by noting that the volume of a sphere of radius r in D dimensions must scale as r^D, and so we write

$$V_D(r) = K_D r^D,$$

  where the constant K_D depends only on D.
- Thus the required fraction is given by

$$\frac{V_D(1) - V_D(1-\epsilon)}{V_D(1)} = 1 - (1-\epsilon)^D.$$

Curse of Dimensionality



Plot of the fraction of the volume of a sphere lying in the range $r = 1 - \epsilon$ to $r = 1$ for various values of the dimensionality D.

Possible Solutions

- First, real data will often be confined to a region of the space having lower effective dimensionality, and in particular the directions over which important variations in the target variables occur may be so confined.

- Second, real data will typically exhibit some smoothness properties (at least locally) so that for the most part small changes in the input variables will produce small changes in the target variables, and so we can exploit local interpolation-like techniques to allow us to make predictions of the target variables for new values of the input variables.

Outline

Example: Medical Diagnosis Problem

- Consider, for example, a medical diagnosis problem in which we have taken an X-ray image of a patient, and we wish to determine whether the patient has cancer or not.
- In this case, the input vector $\mathbf{x}$ is the set of pixel intensities in the image, and output variable t will represent the presence of cancer, which we denote by the class $\mathcal{C}_1$, or the absence of cancer, which we denote by the class $\mathcal{C}_2$.
- We might, for instance, choose t to be a binary variable such that $t = 0$ corresponds to class $\mathcal{C}_1$ and $t = 1$ corresponds to class $\mathcal{C}_2$.

Example: Medical Diagnosis Problem

- When we obtain the X-ray image $\mathbf{x}$ for a new patient, our goal is to decide which of the two classes to assign to the image.
- We are interested in the probabilities of the two classes given the image, which are given by $p(\mathcal{C}_k|\mathbf{x})$.
- Using Bayes' theorem, these probabilities can be expressed in

$$p\left(\mathcal{C}_k \mid \mathbf{x}\right) = \frac{p\left(\mathbf{x} \mid \mathcal{C}_k\right) p\left(\mathcal{C}_k\right)}{p(\mathbf{x})}$$

- Note that any of the quantities appearing in Bayes' theorem can be obtained from the joint distribution $p(\mathbf{x}, \mathcal{C}_k)$ by either marginalizing or conditioning with respect to the appropriate variables.

Example: Medical Diagnosis Problem

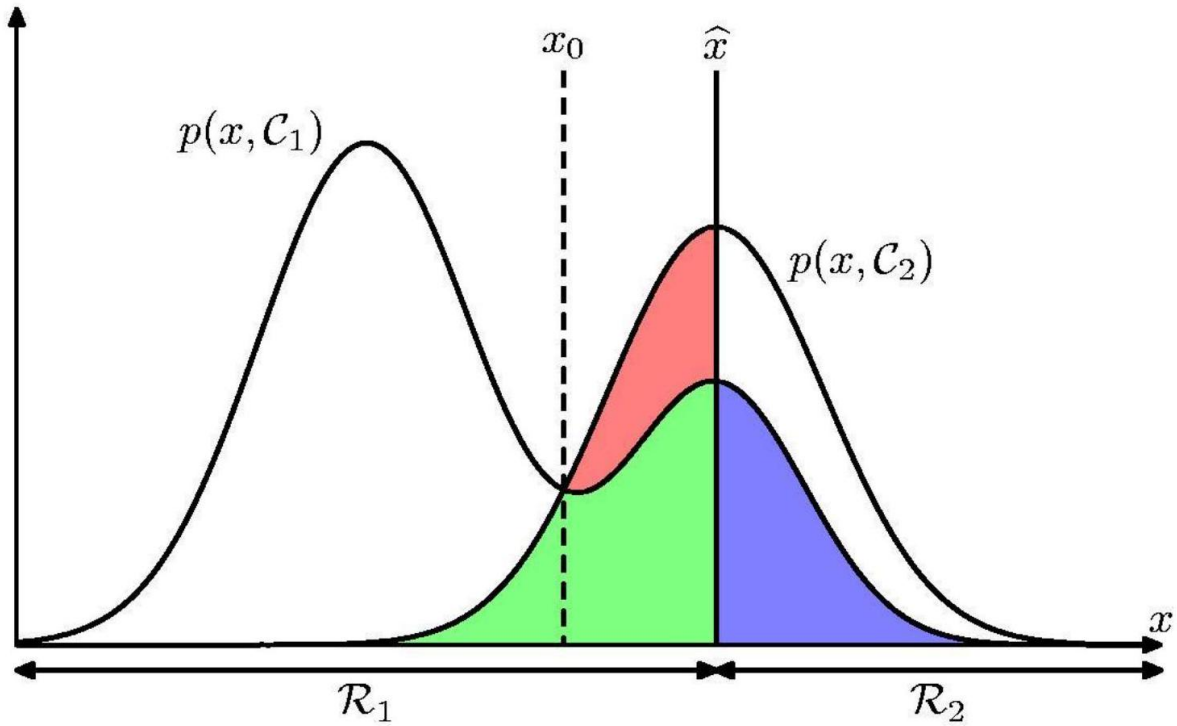
$$p\left(\mathcal{C}_k \mid \mathbf{x}\right) = \frac{p\left(\mathbf{x} \mid \mathcal{C}_k\right) p\left(\mathcal{C}_k\right)}{p(\mathbf{x})}$$

- We can now interpret $p(\mathcal{C}_k)$ as the prior probability for the class $\mathcal{C}_k$, and $p(\mathcal{C}_k \mid \mathbf{x})$ as the corresponding posterior probability.
- Thus $p(\mathcal{C}_1)$ represents the probability that a person has cancer, before we take the X-ray measurement. Similarly, $p(\mathcal{C}_k \mid \mathbf{x})$ is the corresponding probability, revised using Bayes' theorem in light of the information contained in the X-ray.
- If our aim is to minimize the chance of assigning $\mathbf{x}$ to the wrong class, then intuitively we would choose the class having the higher posterior probability.

Minimum Misclassification Rate

- Suppose that our goal is simply to make as few misclassifications as possible. We need a rule that assigns each value of $\mathbf{x}$ to one of the available classes.
- Such a rule will divide the input space into regions $\mathcal{R}_k$ called decision regions, one for each class, such that all points in $\mathcal{R}_k$ are assigned to class $\mathcal{C}_k$.
- The boundaries between decision regions are called decision boundaries or decision surfaces.

Minimum Misclassification Rate



$$
\begin{aligned}
p(\text{mistake}) &= p\left(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2\right) + p\left(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1\right) \\
&= \int_{\mathcal{R}_1} p\left(\mathbf{x}, \mathcal{C}_2\right) \mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p\left(\mathbf{x}, \mathcal{C}_1\right) \mathrm{d}\mathbf{x}
\end{aligned}
$$

Loss-sensitive Decision

Example: classify medical images as 'cancer' or 'normal'

$$
\begin{array}{cc}
 & \text{Decision} \\
 & \begin{array}{cc} \text{cancer} & \text{normal} \end{array} \\
\text{Truth} \begin{array}{c} \text{cancer} \\ \text{normal} \end{array} & \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix}
\end{array}
$$

Loss-sensitive Decision

- Cost/Loss of a decision: L_{kj} = predict $\mathcal{C}_j$ while truth is $\mathcal{C}_k$.
- Loss-sensitive decision $\Rightarrow$ minimize the expected loss:

$$\mathbb{E}[L] = \sum_j \int_{\mathcal{R}_j} \left(\sum_k L_{kj} p\left(\mathbf{x}, \mathcal{C}_k\right) \right) \mathrm{d}\mathbf{x}.$$

- Solution: for each $\mathbf{x}$, choose the class $\mathcal{C}_j$ that minimizes:

$$\sum_k L_{kj} p\left(\mathbf{x}, \mathcal{C}_k\right) \propto \sum_k L_{kj} p\left(\mathcal{C}_k \mid \mathbf{x}\right)$$

$\Rightarrow$ straightforward when we know $p\left(\mathcal{C}_k \mid \mathbf{x}\right)$

Loss-sensitive Decision

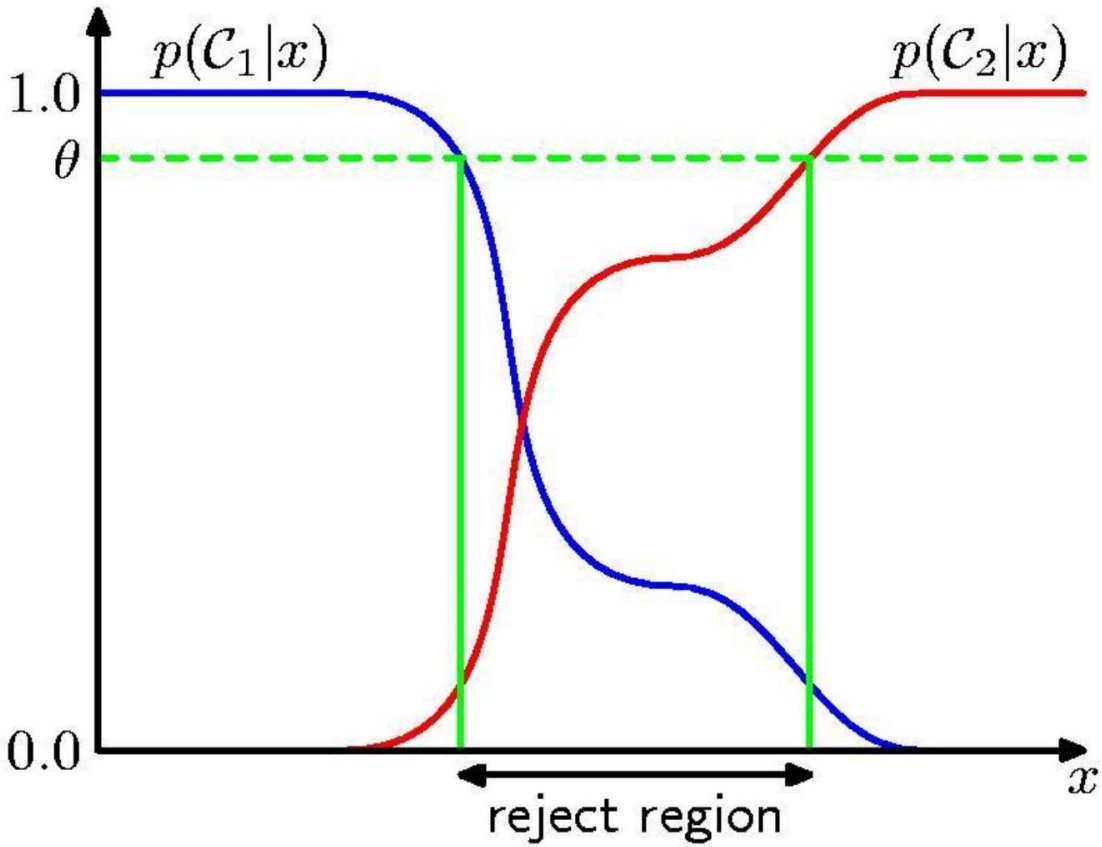
- Typical example = medical diagnosis:
  - $\mathcal{C}_k = \{1, 2\} \Leftrightarrow \{\text{cancer, normal}\}$
  - $L = \begin{bmatrix} 0 & 1000 \\ 1 & 0 \end{bmatrix} \Rightarrow$ strong cost of "missing" a diseased person
- Expected loss:

$$\begin{aligned} \mathbb{E}[L] &= \int_{\mathcal{R}_2} L_{1,2} p\left(\mathbf{x}, \mathcal{C}_1\right) \mathrm{d}\mathbf{x} + \int_{\mathcal{R}_1} L_{2,1} p\left(\mathbf{x}, \mathcal{C}_2\right) \mathrm{d}\mathbf{x} \\ &= \int_{\mathcal{R}_2} 1000 p\left(\mathbf{x}, \mathcal{C}_1\right) \mathrm{d}\mathbf{x} + \int_{\mathcal{R}_1} p\left(\mathbf{x}, \mathcal{C}_2\right) \mathrm{d}\mathbf{x} \end{aligned}$$

- Note: minimizing the probability of misclassification:

$$\begin{aligned} p(\text{mistake}) &= p\left(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2\right) + p\left(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1\right) \\ &= \int_{\mathcal{R}_1} p\left(\mathbf{x}, \mathcal{C}_2\right) \mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p\left(\mathbf{x}, \mathcal{C}_1\right) \mathrm{d}\mathbf{x} \end{aligned}$$

Reject Option



Inference and Decision

- Inference step: we use training data to learn a model for $p(\mathcal{C} \mid \mathbf{x})$.
- Decision step: we use these posterior probabilities to make optimal class assignments.
- An alternative possibility would be to solve both problems together and simply learn a function that maps inputs x directly into decisions. Such a function is called a discriminant function.

Approaches to Decision Problems

- Rely on a probabilistic model, with 2 flavours:
  - Generative: (1) use a generative model to infer $p(x \mid C_k)$, (2) combine with priors $p(C_k)$ to get $p(x, C_k)$ and eventually $p(C_k \mid x)$
  - Discriminative: infer directly $p(C_k \mid x)$
- Learn a discriminant function $f(x)$: (1) directly map input to class labels, (2) for binary classification, $f(x)$ is typically defined as the sign $(+1/-1)$ of an auxiliary function

Approaches to Decision Problems

Pros and Cons:

- Probabilistic generative models:
  - pros: access to $p(x) \rightarrow$ easy detection of outliers, i.e., low-confidence predictions
  - cons: estimating the joint probability $p(x, C_k)$ can be computational and data demanding
- Probabilistic discrimative models:
  - pros: less demanding than the generative approach
- Discriminant functions:
  - pros: a single learning problem (vs inference + decision)
  - cons: no access to $p(C_k \mid x)$

Why Separate Inference and Decision?

- Minimizing risk (loss matrix may change over time)
- Reject option
- Unbalanced class priors
- Combining models

Q & A

1. Example: Polynomial Curve Fitting
2. Probability Theory
3. Model Selection
4. Curse of Dimensionality
5. Decision Theory