AARHUS UNIVERSITY

Data Science 1

Chris Mathys

Master's Degree Programme in Cognitive Science

Spring 2022

Data Science – Overview

- Analysis of large, complex, and naturalistic data sets
- Advanced statistical and computational methods
- Relate these methods to
  - Experimental and lab methodology
  - Theories of cognitive functions
- Start out with core principles
- Foundations of building and training artificial neural networks
- Analysis of time series data

Materials and Literature

Core Principles

A fantastic resource on the foundations of all of data science is the textbook by Chris Bishop:

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer-Verlag New York Inc.

This is available as a free PDF.

Artificial Neural Networks (ANNs)

We will use the online textbook by Michael Nielsen:

Nielsen, M. A. (2015). *Neural Networks and Deep Learning.* Determination Press.

Time Series Analysis

For this part of the course, we will use the online textbook FPP3 by Rob J Hyndman and George Athanasopoulos:

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice (3rd edition).* OTexts.

Additional Materials

In the later parts of the course, we will also look at scientific papers in the fields of time series analysis and cognitive neuroscience.

Schedule

Course week	Week of year	Topics and readings
1	6	Introduction, basics of data science (Bishop 1.1, 1.2.{1-3})
2	7	Basics of data science, continued (Bishop 1.2.4, 1.3, 1.4)
3	8	ANNs: foundations (Nielsen 1)
4	9	ANNs: backpropagation and training of networks (Nielsen 2,3)
5	10	ANNs: deep learning (Nielsen 5,6)
6	12	Time series: intro, graphics, decomposition (FPP3 1,2,3)
7	13	Time series: forecaster's toolbox, regression (FPP3, 4,5,7)
8	14	Time series: ETS, ARIMA, Bayesian structural models (FPP3, 7,8,9)
9	17	Time series: dynamic regression models, advanced methods (FPP3, 10,11,12)
10	18	Time series: Bayesian structural models
11	19	Belief updating: cognitive processing of sequential inputs
12	20	Belief updating: hierarchical Gaussian filters

Exam

Format and Deadlines

The format is very simple: **you choose a dataset, analyze it using current data scientific methods, and write a paper on the results.**

- The product associated with your paper is the software you produce for the analysis. The whole analysis pipeline has to be submitted and is an integral part of the exam project.
- Your software may (and is expected to) rely on available tools, i.e., you don't have to start from scratch
- Your chosen dataset may be publicly available, newly acquired, or available only to you
- By 3 April, you decide who (if anybody) you will work with
- By 24 April, you send me an abstract of your proposed paper (maximum 250 words) for approval
- By the date specified in the exam plan, you submit your exam project.
- You submit a GitHub repository containing your text and product via GitHub Classroom. Additionally, you submit the same items via the Digital Exam system.

Exam

Formal requirements

As specified in the course description, ordinary examination and re-examination are as follows:

The examination consists of an individual take-home assignment on a topic of the student's choice and a related practical product.

The scope and nature of the product must be relevant in relation to the content of the course and is subject to the approval of the teacher. It must be possible to submit the product digitally in a documented form which can be accessed by the supervisor and co-examiner. The product must be accompanied by a take-home assignment on a topic of the student's choice, in which the student explains the relevance and methodological and theoretical basis of the product. Assessment is based on an overall assessment of the take-home assignment and the practical product.

The assignment can be written individually or in groups of up to 3 students. Group assignments must be written in such a way that the contribution of each student, except for the introduction, thesis statement and conclusion, can form the basis of individual assessment. The assignment should clearly state which student is responsible for which section.

Length for one student: 10-15 standard pages

Length for two students: 15-20 standard pages

Length for three students: 20-25 standard pages

The take-home assignment must be handed in for assessment in the Digital Exam system by the date specified in the exam plan.

Additional resources

The web is full of resources on data science

- https://www.kaggle.com/
- https://towardsdatascience.com/
- http://neuralnetworksanddeeplearning.com/index.html
- etc.

Questions for you

Choose 3 of the following questions and take 10 minutes to give a short written answer to each. You won't have to share what you write.

- Do you have a definition of data science?
- Why are there data science courses today, but not a few decades ago (at least not under that name)? What has changed?
- Is data science a branch of mathematics, a kind of engineering, or actually a science?
- What is a model to you? And explain how you compare different models.
- Explain the tension between explanation and prediction
- What datasets are you most interested in?

Questions for you

Choose 3 of the following questions and take 10 minutes to give a short written answer to each. You won't have to share what you write.

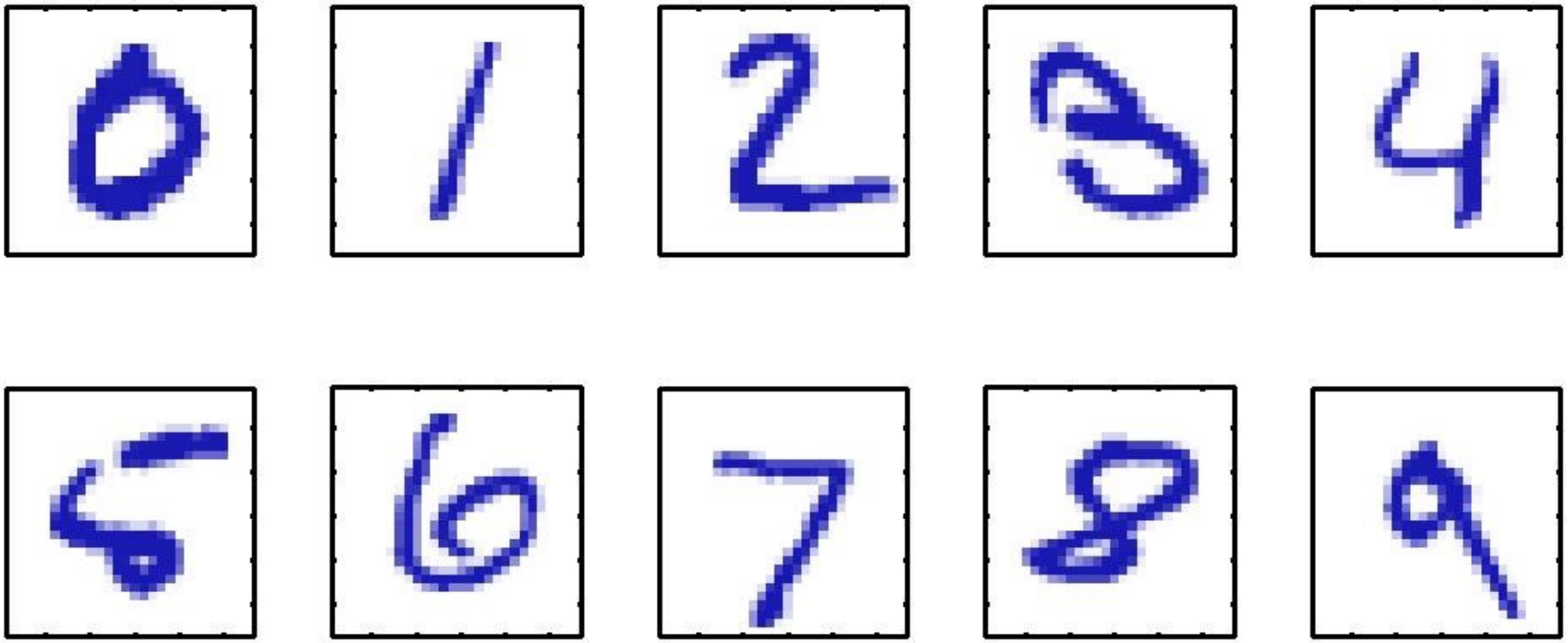
- Do you have a definition of data science?
- Why are there data science courses today, but not a few decades ago (at least not under that name)? What has changed?
- Is data science a branch of mathematics, a kind of engineering, or actually a science?
- What is a model to you? And explain how you compare different models.
- Explain the tension between explanation and prediction
- What datasets are you most interested in?

PATTERN RECOGNITION AND MACHINE LEARNING

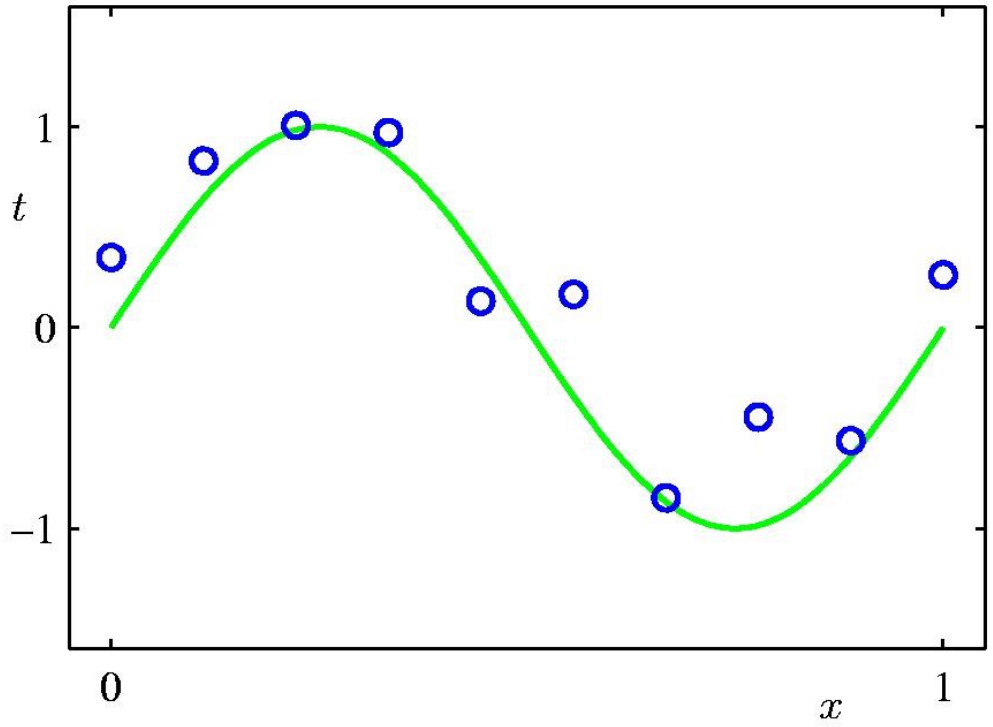
CHAPTER 1: INTRODUCTION

Example

Handwritten Digit Recognition

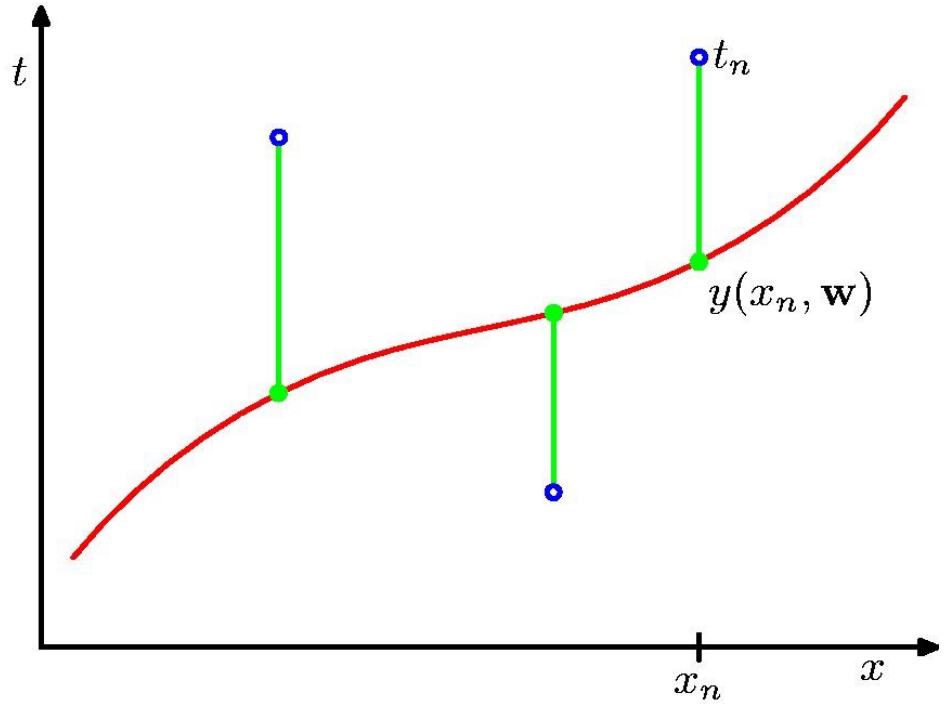


Polynomial Curve Fitting



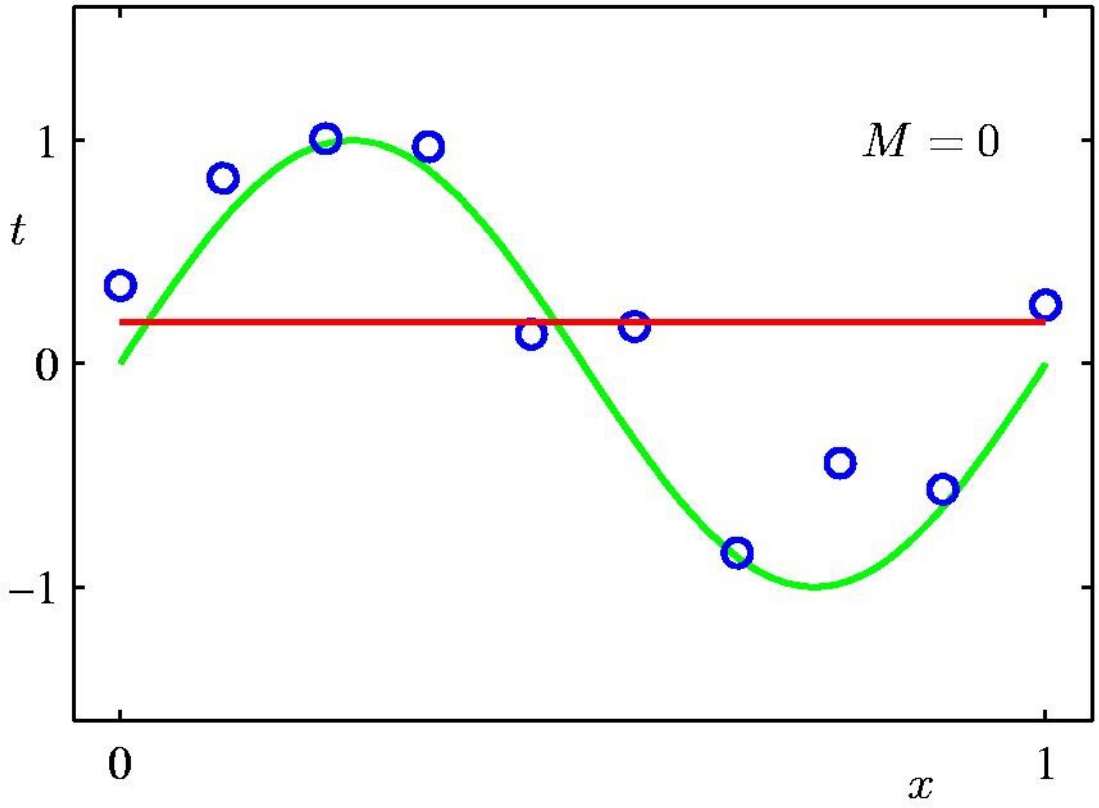
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Sum-of-Squares Error Function

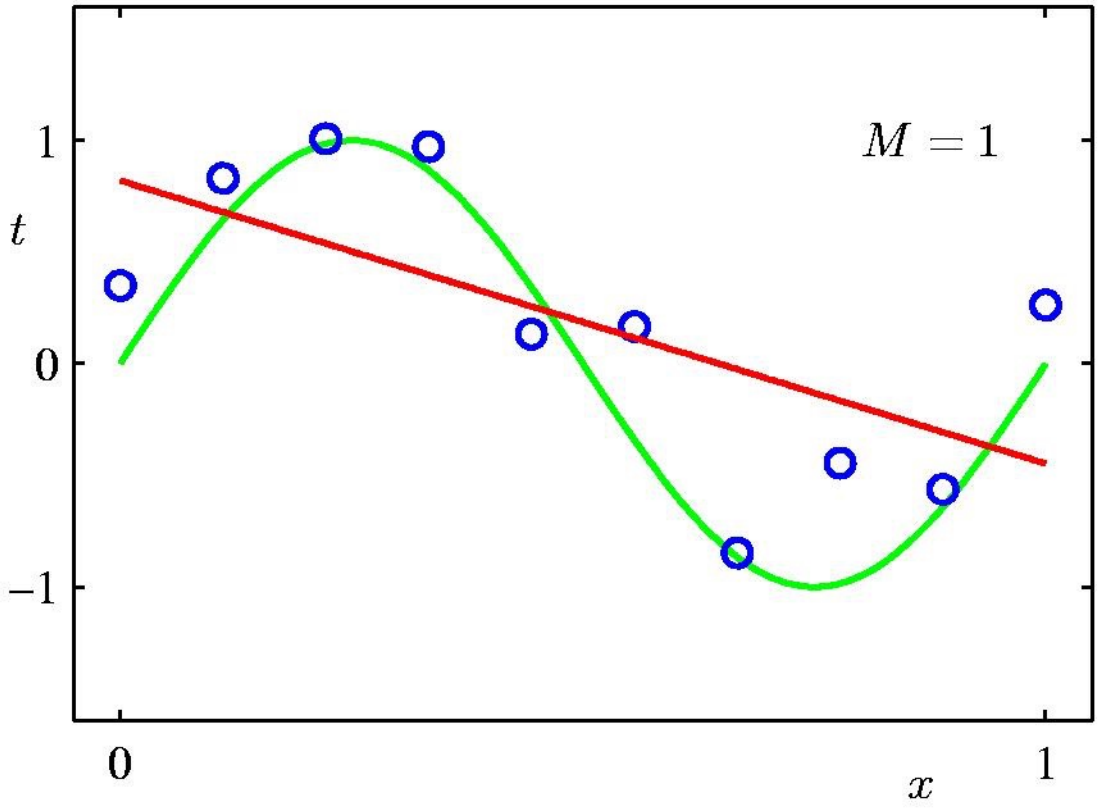


$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

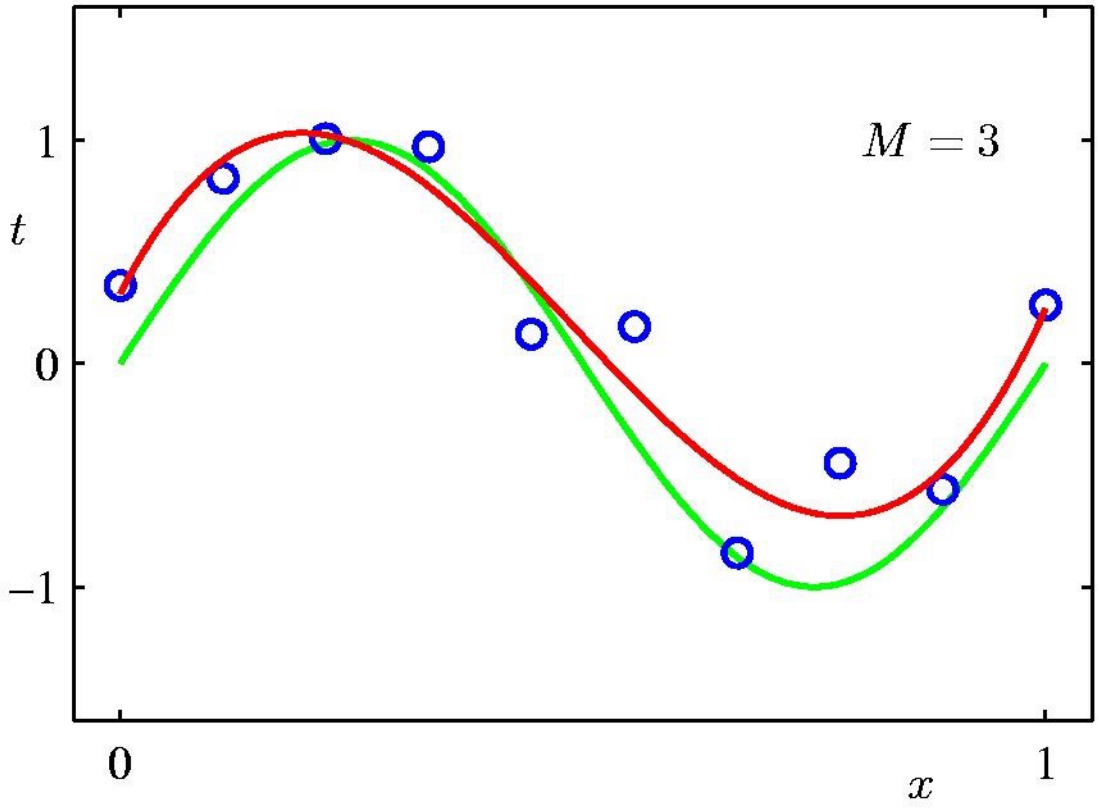
0^{th} Order Polynomial



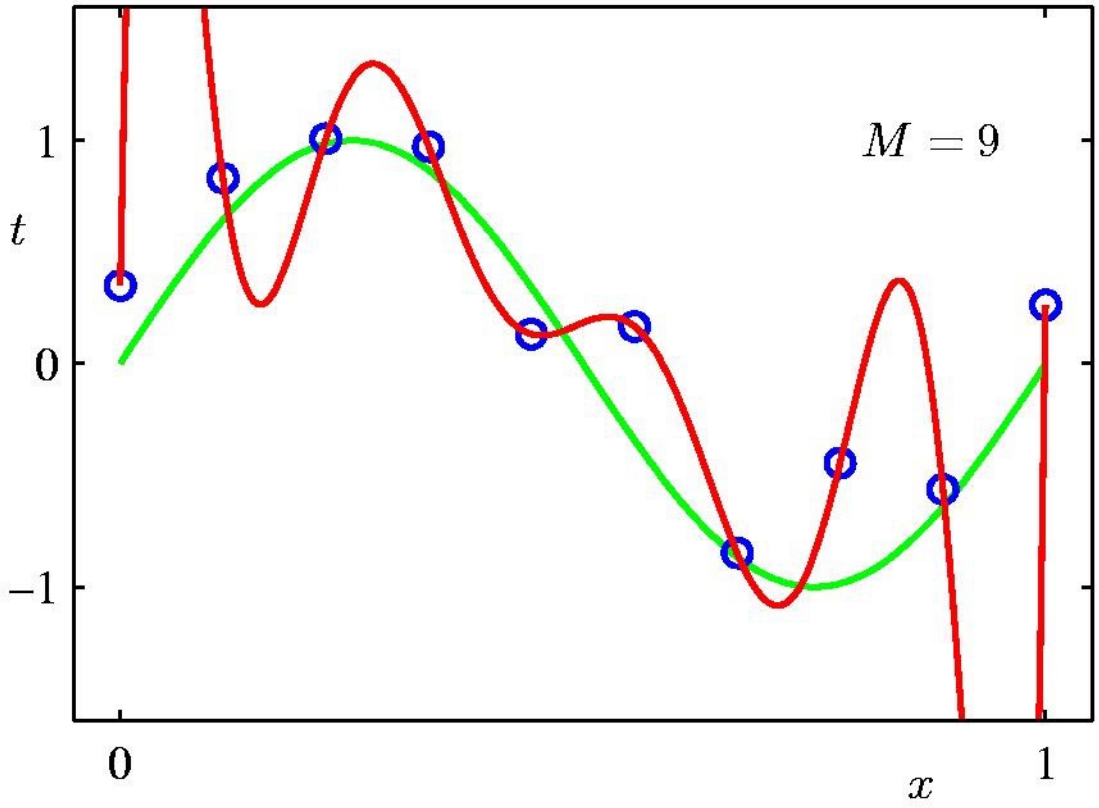
1st Order Polynomial



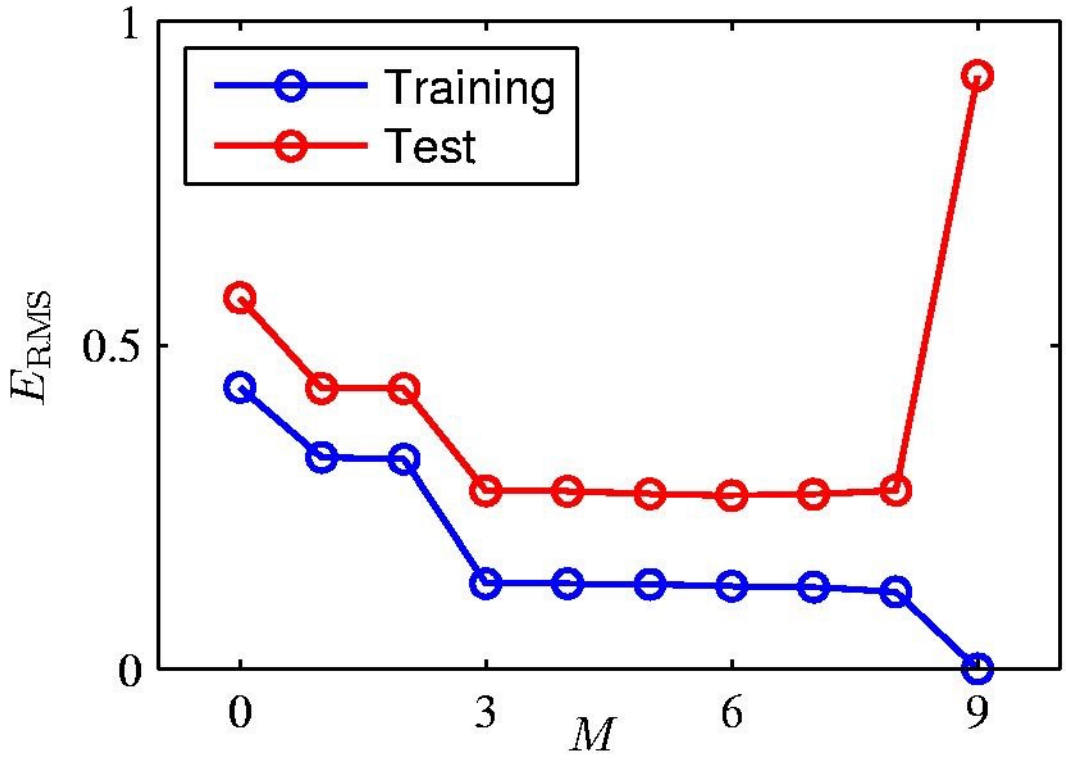
3rd Order Polynomial



9th Order Polynomial



Over-fitting



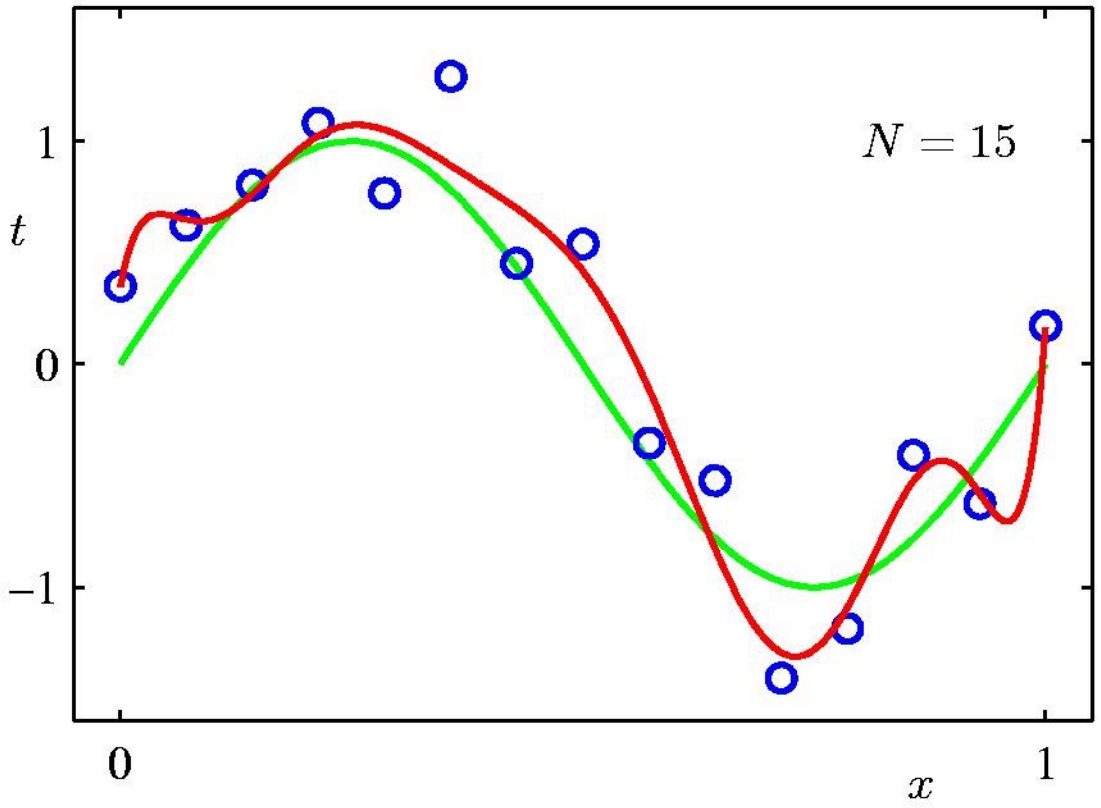
Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^{\star})/N}$

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

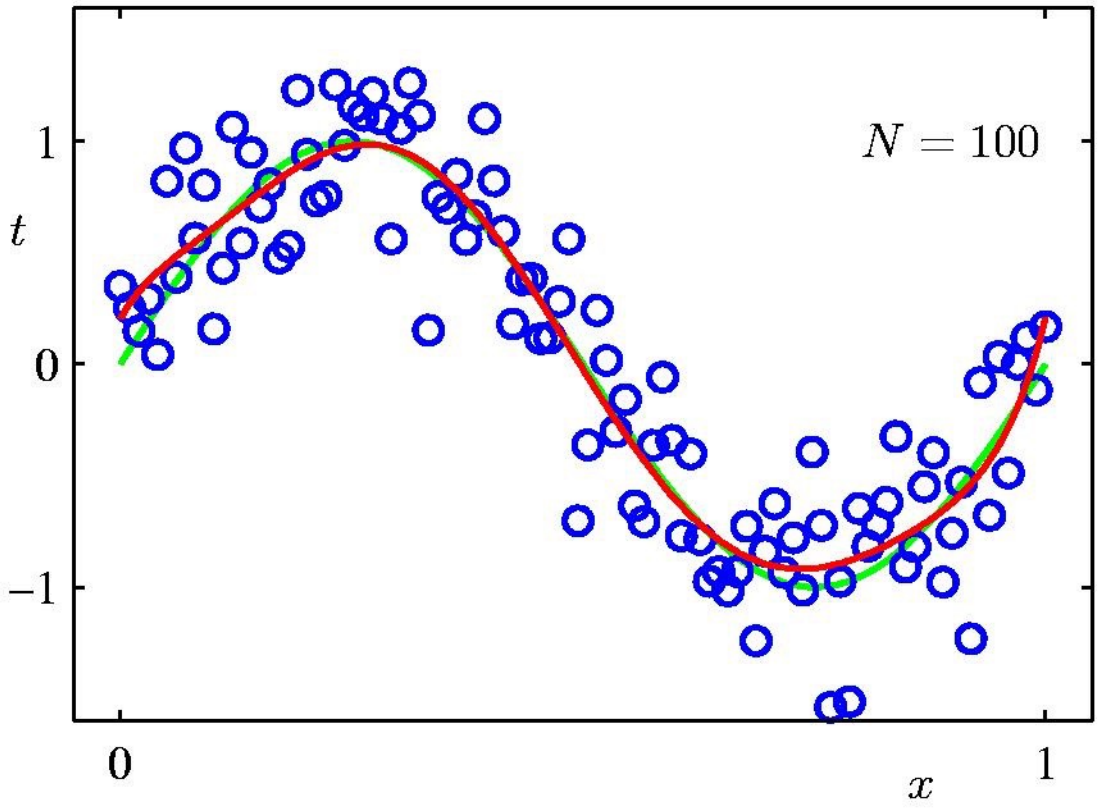
Data Set Size: $N = 15$

9th Order Polynomial



Data Set Size: $N = 100$

9th Order Polynomial

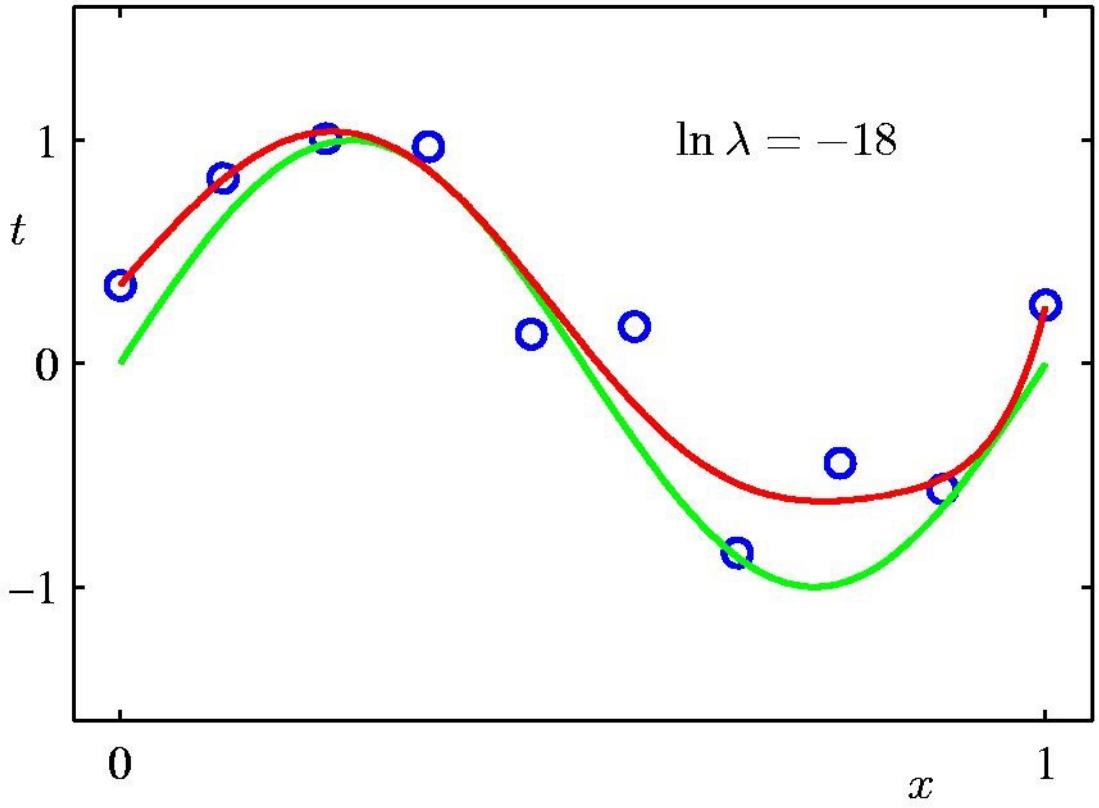


Regularization

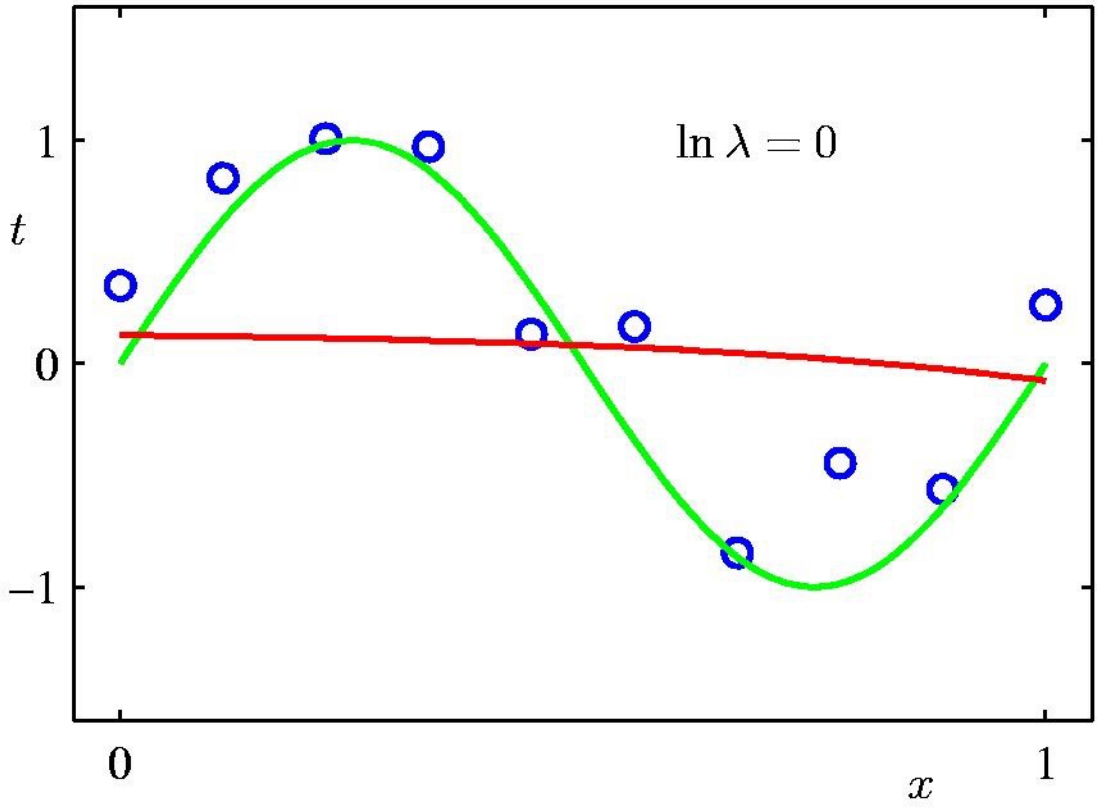
Penalize large coefficient values

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

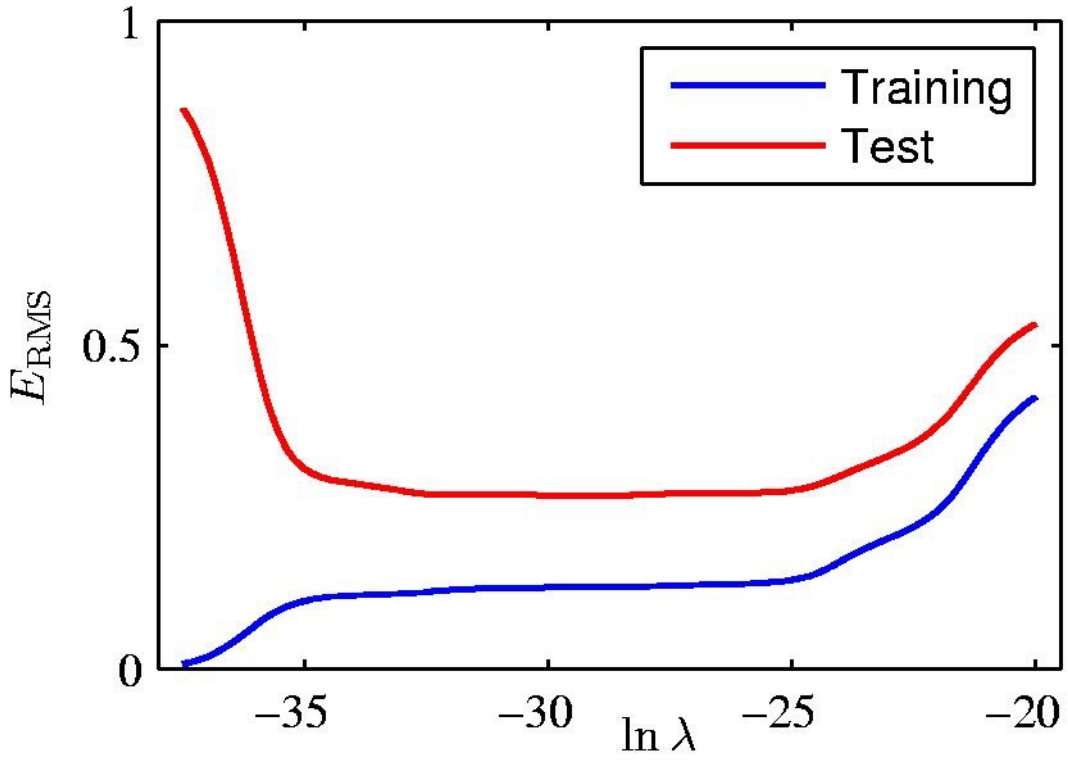
Regularization: $\ln \lambda = -18$



Regularization: $\ln \lambda = 0$



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

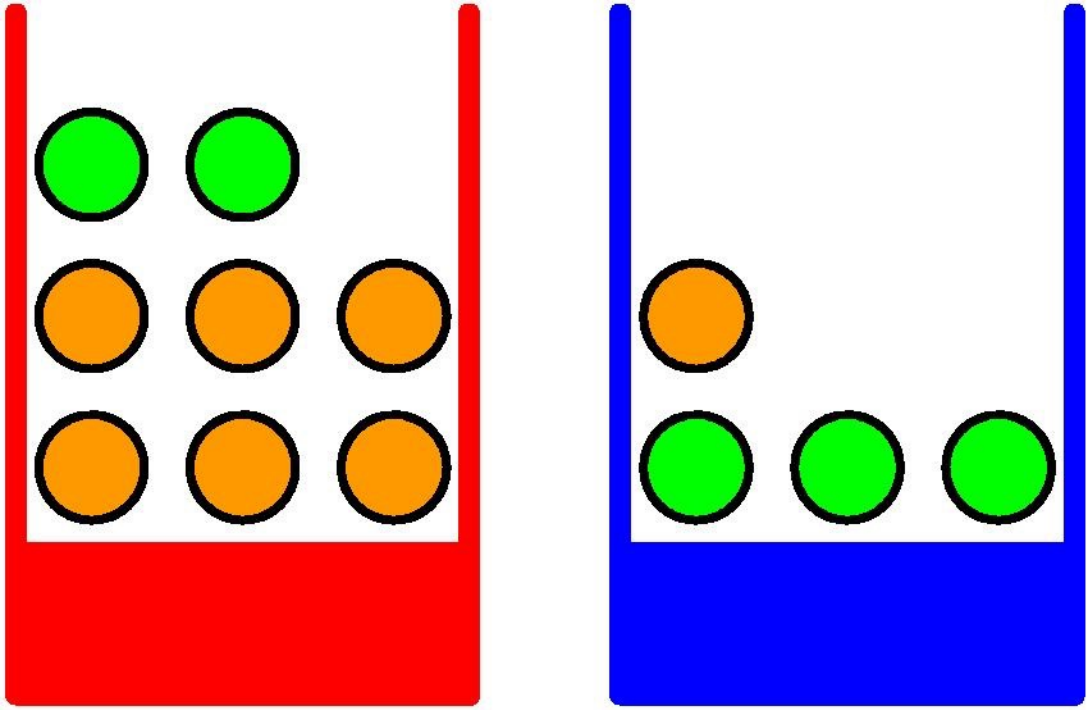


Polynomial Coefficients

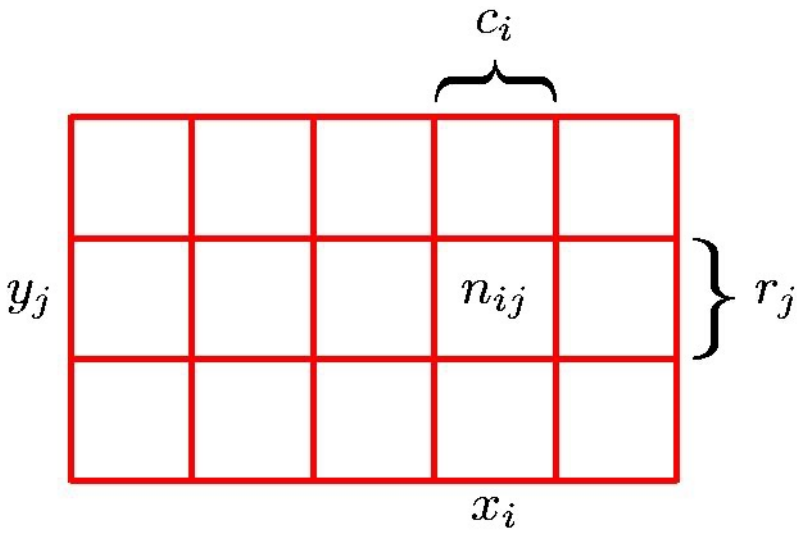
	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$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

Probability Theory

Apples and Oranges



Probability Theory



Marginal Probability

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

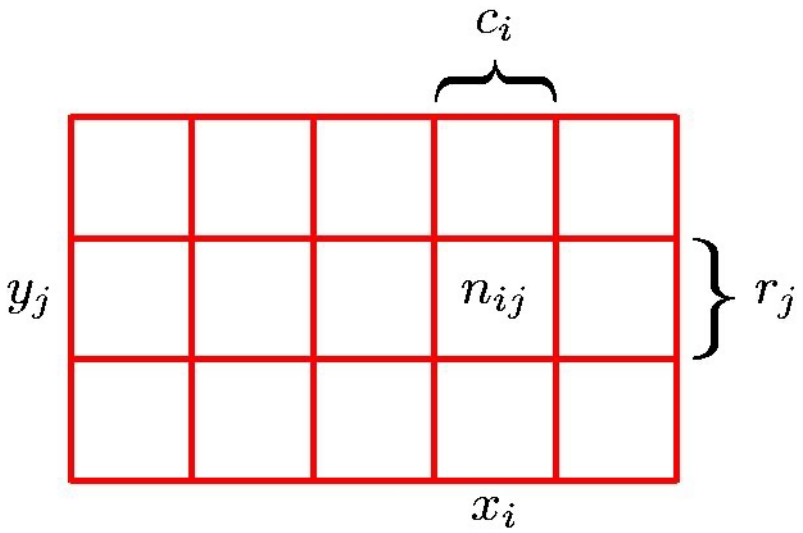
Joint Probability

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

Conditional Probability

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

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

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

$$= p(Y = y_j | X = x_i)p(X = x_i)$$

The Rules of Probability

Sum Rule

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

Product Rule

$$p(X, Y) = p(Y|X)p(X)$$

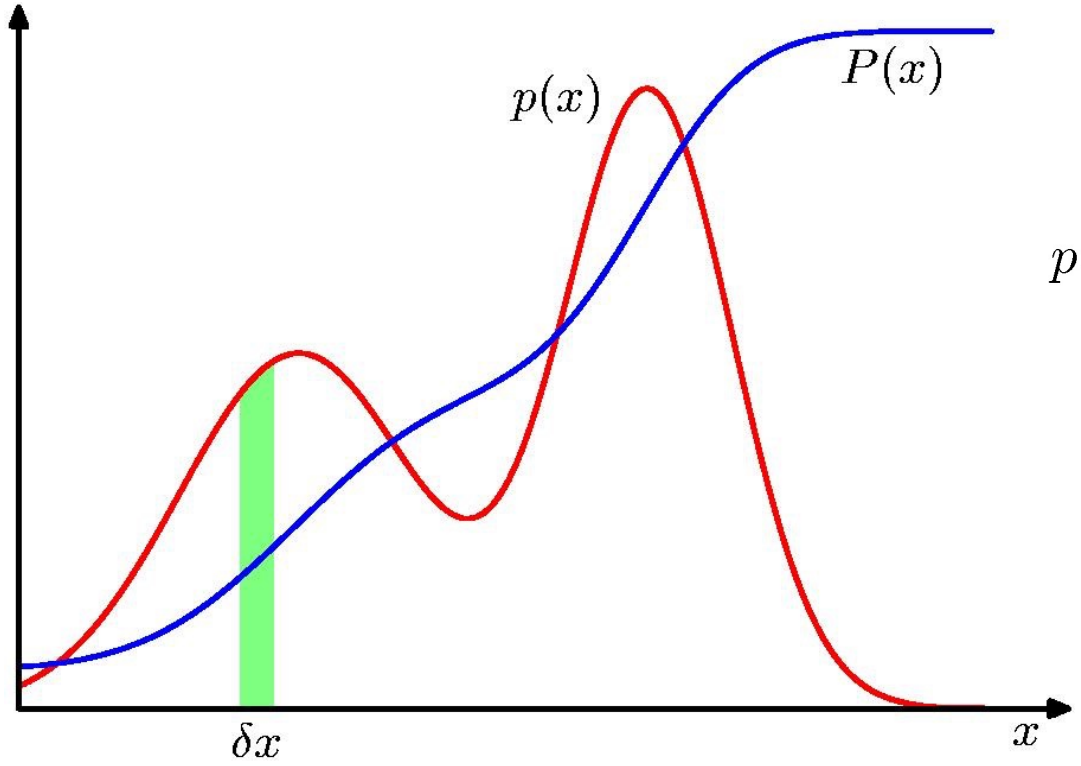
Bayes' Theorem

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

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

posterior $\propto$ likelihood $\times$ prior

Probability Densities



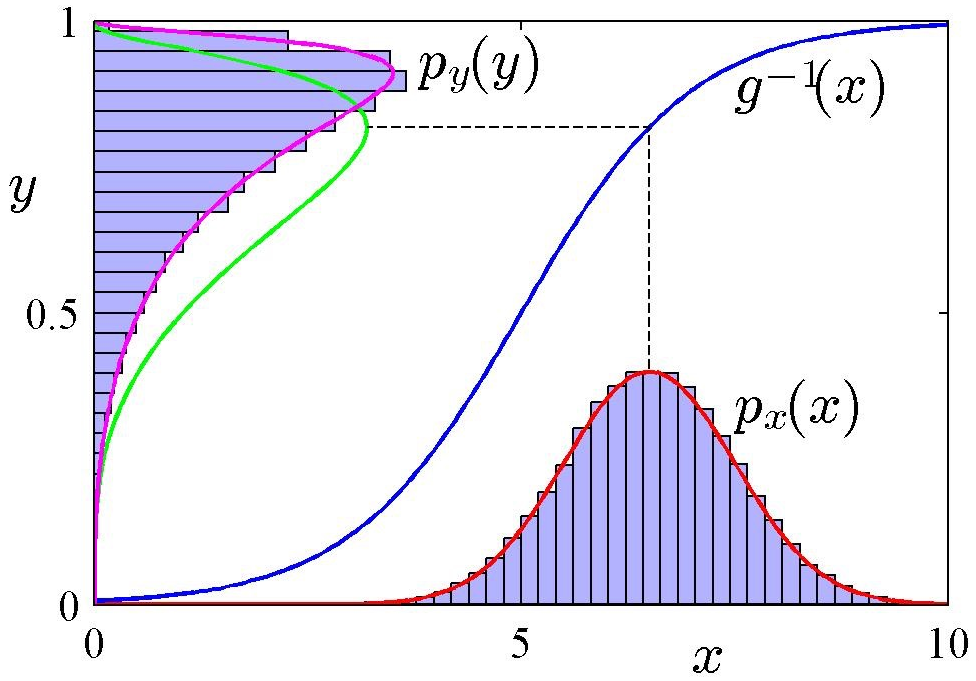
$$p(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x$$

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

$$p(x) \geqslant 0$$

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

Transformed Densities



$$
\begin{aligned}
p_y(y) &= p_x(x) \left| \frac{\mathrm{d}x}{\mathrm{d}y} \right| \\
&= p_x(g(y)) \, |g'(y)|
\end{aligned}
$$

Expectations

$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

$$\mathbb{E}[f] = \int p(x) f(x) \, \mathrm{d}x$$

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x)$$

Conditional Expectation (discrete)

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

Approximate Expectation (discrete and continuous)

Variances and Covariances

$$\mathrm{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

$$\begin{aligned}
\mathrm{cov}[x, y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\
&= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]
\end{aligned}$$

$$\begin{aligned}
\mathrm{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\
&= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}]
\end{aligned}$$