Concepts of Probability Theory for Machine Learning

Klaus-Robert Müller, Mikio L. Braun

October 23, 2008

Why Probability Theory?

- Mathematical formalism for expressing uncertainty, randomness and noisy measurements.
- Useful theory what happens as number of samples tends to infinity.
- Theory for tracking uncertainty in view of data.
- Much more than simply counting the number of possibilities.

Focus of this Introduction

- Introducing fundamental concepts.
- Understand, how these concepts are formalized mathematically.

Probability

A *probability* is a number between 0 and 1:

- "0" means almost never
- "1" means almost surely

For example:

- The probability that it rains tomorrow is 30% ($= 0.3$)
- The probability that a fair coin produces a head is 0.5.

Events

Probabilities are defined for so-called "events".

Events can be combined logically:

- Negation: "the dice does not produce 1"
- Intersection: "it rains tomorrow and the temperature is below 15°C"
- Union: "it rains tomorrow or it's cloudy"

Probability Measures

A probability spaces is a three-tuple $(\Omega, \mathcal{A}, P)$ with

- Ω is the set of all events.
- $\mathcal{A}$ is a collection of subsets of Ω ("σ-Algebra")
- A probability measure is $P\colon \mathcal{A} \to \mathbb{R}$ such that
  - $P(\Omega) = 1$
  - $P(\overline{A}) = 1 - P(A)$.
  - $P(\dot{\bigcup}_i^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

In words:

- Events are sets and can be freely combined (up to some technical restrictions, almost never practically relevant).
- The combination of all possible events is called Ω
- A probability is a function on subsets of Ω.

$\rightsquigarrow$ Probabilities are like measuring the volume of a set.

Example: Fair coin

- Events $\Omega = \{\text{head}, \text{tail}\}$.
- Admissible subsets $= \mathcal{P}(\Omega)$ (always possible for discrete sets).
- Probability measure: $P(\{\text{head}\}) = P(\{\text{tail}\}) = 0.5$.

Note: Often, the double parenthesis after P are dropped:

$$P(\{\text{head}\}) \text{ becomes } P\{\mathit{head}\} \text{ or even } P(\text{head}).$$

Example: Loaded dice

- Events $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- Admissible subsets again all subsets of Ω.
- Probability measure defined by

  $\mathrm{P}(\{1\}) = 0.15$
  $\mathrm{P}(\{2\}) = 0.15$
  $\mathrm{P}(\{3\}) = 0.1$
  $\mathrm{P}(\{4\}) = 0.2$
  $\mathrm{P}(\{5\}) = 0.2$
  $\mathrm{P}(\{6\}) = 0.2$

  and $P(\{1, 2\}) = P(\{1\}) + P(\{2\})$, etc.

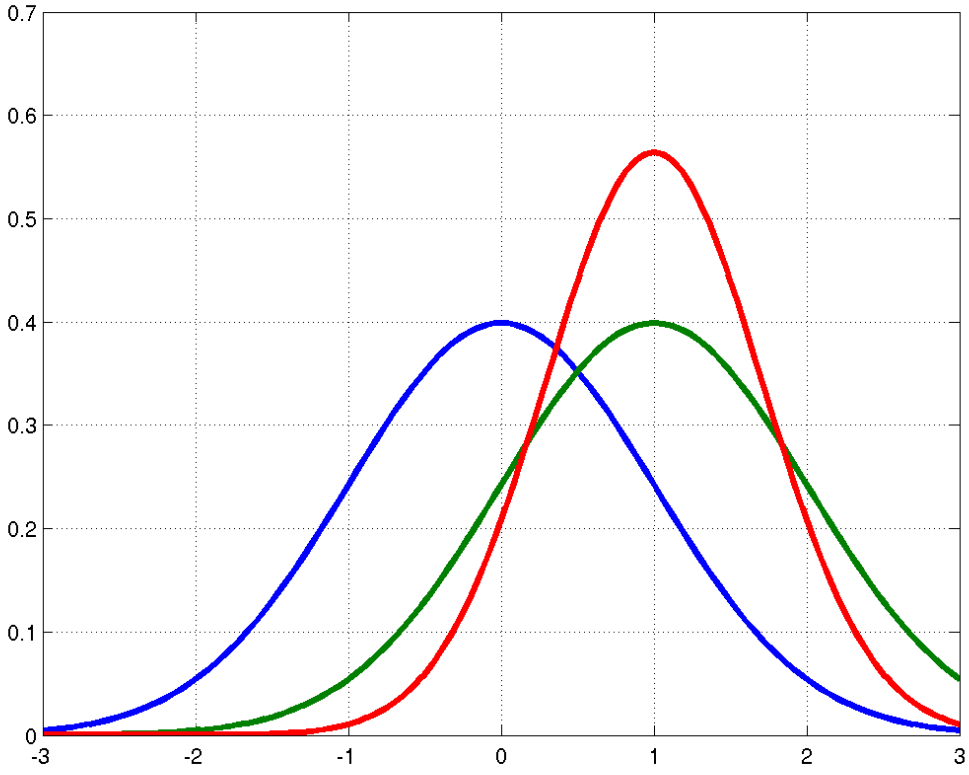
Concrete probabilities

"Normally", probabilities exist mostly in two forms:

- **Discrete probabilities**
  Ω is a finite set, $\mathcal{A}$ is the set of all subsets, P is defined by probabilities for single elements.
- **Probabilities with densities**
  Ω is $\mathbb{R}$, or $\mathbb{R}^d$, $\mathcal{A}$ are what can sensibly be constructed from open and closed intervals, and P is defined by a *density*:

$$P(A) = \int_A p(x)dx$$

Example: Gaussian Distribution



$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Has two parameters μ, σ^2 which control the location and the width of the bump.

Why probability of zero means "almost surely"

If $P(A) = 0$, we say that A happens almost never. Why the "almost"?

For a distribution with a density (like the Gaussian distribution), for each $x \in \mathbb{R}$,

$$P\{x\} = 0,$$

although each time a certain x is realized.

Therefore, we say that the probability that we hit a certain real number x happens "almost never" but on each realization, we get a certain x.

Random Variables

Random variables represent random, noisy or uncertain quantities.

For example:

- The number of pips on a thrown die.
- Outside temperature.
- Number of students in a class.
- Color of a ball drawn at random from an urn.

Mathematical Definition of Random Variables

Mathematically, random variables are functions defined on probability spaces.

This means:

- Random variables are not special mathematical objects.
- There is no inherent randomness in a random variable, it comes from the underlying probability space.

An example

Consider the random variable

$$X = \text{sum of outcome of two dice.}$$

It could be constructed as follows:

- $\Omega = \{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(6,6)\}$
- $P(\omega) = 1/36$ for all single elements of Ω.
- $X(\omega) = a + b$ for $\omega = (a,b) \in \Omega$,

Shorthand notation for events

One can generate new events for a random variable, by considering the event that $X \in A$.

For example: "the sum of two thrown dice is 6"

- X is constructed like before,
- The event $A = \{\omega = (a, b) \in \Omega | a + b = 6\}$.

Often, the notation is further simplified by dropping the elements of the originial probability space.

$$\{\omega \in \Omega | X(\omega) \in A\} \text{ becomes } \{X \in A\}.$$

Likewise: $\{X > a\}$, $\{X = a\}$, etc.

Probability Spaces vs. Random Variables

Probability spaces and random variables are closely related:

- Every random variable has an associated probability measure

  $$P_X(A) = P\{X \in A\} \quad [= P(\{\omega \in \Omega | X(\omega) \in A\}) = P(X^{-1}(A))]$$

- You can directly specify the distribution of the random variable without referring to the underlying space Ω. ("Let X, Y be two Gaussian random variables").

In fact, often, the underlying $(P, \mathcal{A}, \Omega)$ is not specified at all! It is assumed that it is rich enough to allow all the specified random variables.

Expectation

The *expectation* of a random variable can be thought of as the weighted average over all its possible realizations.

It's also called the mean (but it's not the most probable outcome).

Expectation is defined as follows:

$$E(X) = \sum_x xp(X = x) \qquad \text{(discrete case)}$$

or

$$E(X) = \int xp(x)dx \qquad \text{(continuous case)}$$

where x varies over all possible outcomes of x

Example

Discrete probability distribution

1	2	3
0.1	0.5	0.4

Expectation:

$$\sum_x xP(X = x) = 1 \times 0.1 + 2 \times 0.5 + 3 \times 0.4$$

$$= 0.1 + 1 + 1.2 = 2.3$$

Computing with Expectations

Since expectations are basically integrals, computing expectations can often be hard.

However, a few rules exist:

- $E(aX) = aE(X)$
- $E(X + Y) = E(X) + E(Y)$
- $\min_{\omega \in \Omega} X(\omega) \leq E(X) \leq \max_{\omega \in \Omega} X(\omega)$.

Unfortunately, $E(XY) \neq E(X)E(Y)$ in general!

Variance

While the expectation characterizes the *location* of a random variable, the *variance* characterizes its spread around the mean.

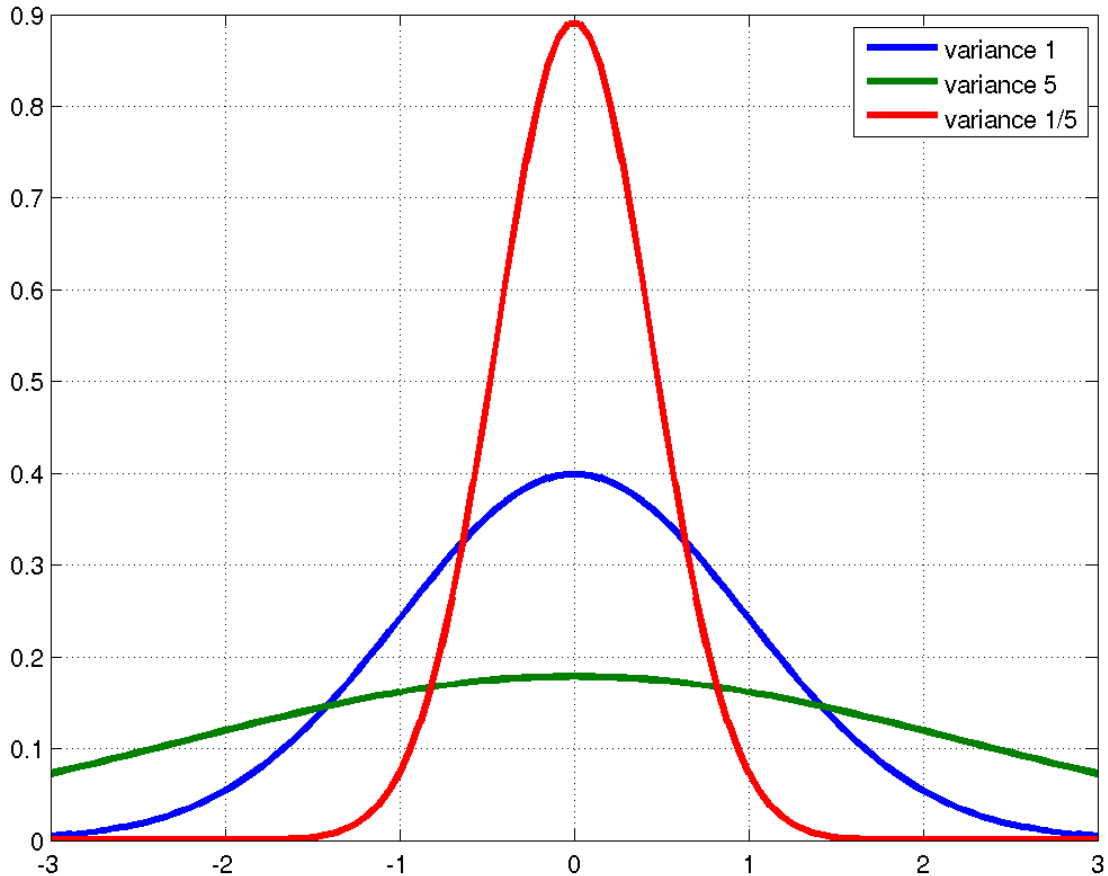
It is defined as the average quadratic deviation from the mean:

$$\mathrm{Var}(X) = E(X - EX)^2.$$

Can also be written as

$$\begin{aligned} E(X - EX)^2 &= E[X^2 - 2X\,EX + (EX)^2] \\ &= E(X^2) - 2(EX)^2 + (EX)^2 \\ &= E(X^2) - (EX)^2. \end{aligned}$$

Example: Gaussian distribution



For Gaussian distributions, σ^2 directly parameterizes the variance.

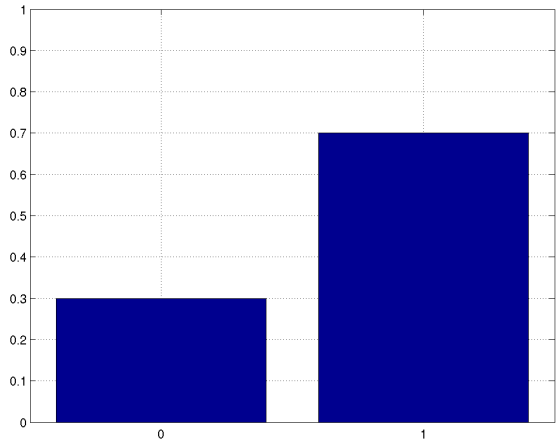
Special Distributions

There exist many special distributions which re-occur often.

They typically model some simple experiment or have other interesting features which.

Some of them are computationally more advantageous than others.

Bernoulli Distribution



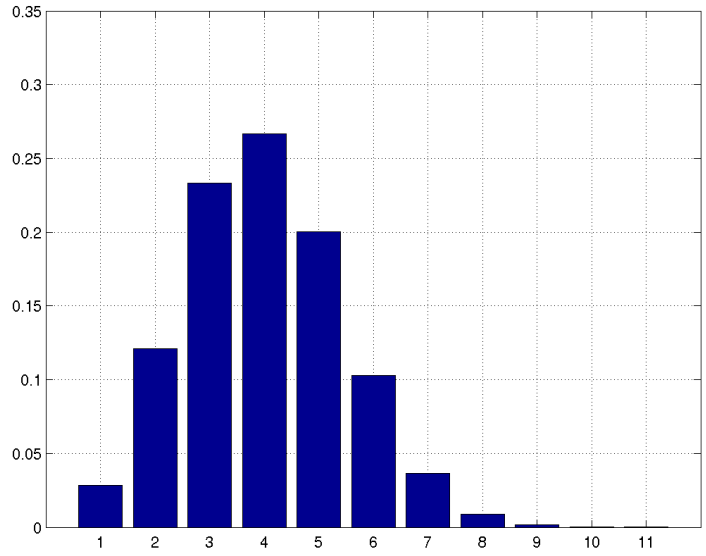
Distribution on $\{0, 1\}$.

Parameter $p = P\{1\}$.

Mean: p.

Variance: $p(1 - p)$.

Binomial Distribution



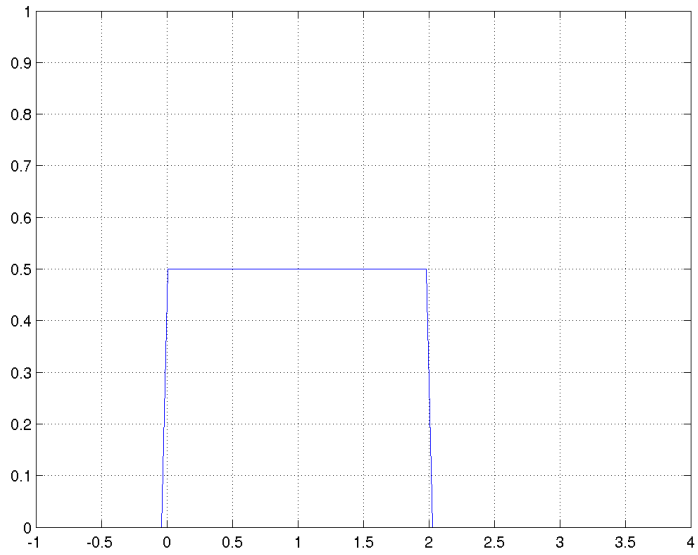
Distribution on $\{0, \ldots, n\}$, occurs as sums of Bernoulli variables.

Parameter $0 \leq p \leq 1$.

$$P\{x\} = \binom{n}{k} p^k (1-p)^{n-k}$$

Mean: np.
Variance: $np(1-p)$.

Uniform Distribution

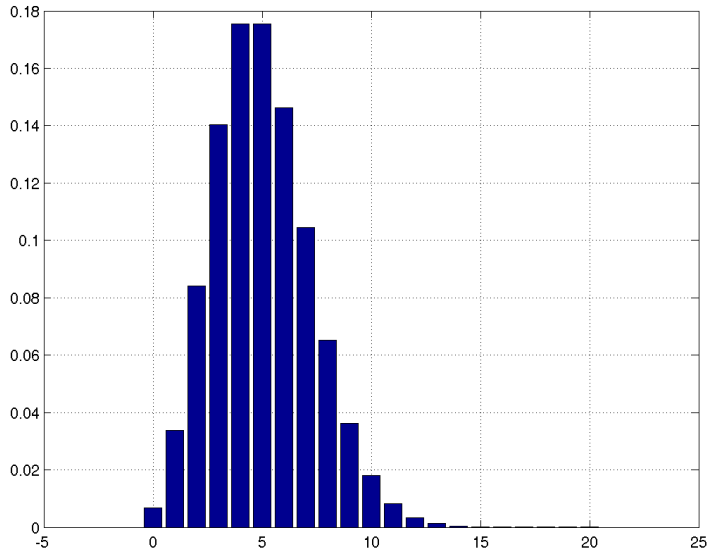


Distribution on $\mathbb{R}$, Parameters: $a < b$

$$p(x) = \begin{cases} 1/(b-a) & a \leq x \leq b \\ 0 & \text{else} \end{cases}.$$

Mean: $(b+a)/2$, Variance: $(b-a)^2/12$.

Poisson Distribution



Distribution on $\mathbb{N}$. Parameter: $\lambda > 0$.

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

Mean: λ. Variance: λ.

Joint Probabilities

The probability of the intersection of two events A, B is also called the *joint probability* of A and B.

For example,

- $A =$ it rains tomorrow.
- $B =$ I forget to take my umbrella with me.

Then, the joint probability of A and B is simply $P(A \cap B)$, also simply written $P(A, B)$.

Conditional Probabilities

Conditional probabilities are like joint probabilities, only that I assume that one of the two events has actually taken place.

For example: "what is the probability that it rains given that I forgot to take my umbrella with me?"

Conditional probabilities are written $P(A|B)$.

They are defined as

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

in order to normalize the probability of B happening.

More notational conventions

- Joint and conditional probabilities are similarily written for more than two events, for example

$$P(A, B, C|D, E) = \frac{P(A, B, C, D, E)}{P(D, E)}$$

- Likewise, events containing random variables are written in a similar manner, for example

$$P(X \geq a|Y = c)$$

- For densities, the convention is used that lower case letters correspond to upper case random variables:
That is, $p(x)$ is the density of X, $p(y)$ is the density of Y, and the conditional density is $p(x|y)$.

An example

Event A is "Drinking hot lemon tea."

Event B is "Cured from a cold in less than 3 days."

Then:

- $P(A, B)$ = probability that somebody was cured from a cold in less than 3 days and has been drinking hot lemon tea.
- $P(B|A)$ = probability that somebody is cured from a cold in less than 3 days who has been drinking hot lemon tea.
- $P(A|B)$ = probability that somebody drank hot lemon tea who has been cured from a cold in less than 3 days.

Marginalization

You can "remove" a random variable from a joint or conditional density by a procedure called *marginalization*, since

$$p(x) = \sum_y p(x, y), \qquad \text{(discrete case)}$$

$$p(x) = \int p(x, y) dy, \qquad \text{(densities)}.$$

More abstractly, this amounts to considering all possible realizations of Y: Let Υ be all possible valules for Y, then

$$P(\{X \in A\} \cap \{Y \in \Upsilon\}) = P(\{X \in A\} \cap \Omega) = P\{X \in A\}.$$

An example, cont'd

Random variable X = Has been drinking hot lemon tea

Random variable Y = Has been cured from a cold in less than 3 days.

X↓, Y→	< 3 days	> 3 days
tea	0.3	0.2
no tea	0.1	0.4

Then:

- $P(\text{tea}) = 0.3 + 0.2 = 0.5$
- $P(< 3\text{days}) = 0.4$
- $P(< 3\text{days}|\text{tea}) = 0.3/0.5 = 0.6$
- $P(> 3\text{days}|\text{tea}) = 0.2/0.5 = 0.4$
- $P(\text{tea}| < 3\text{days}) = 0.3/0.4 = 0.75$

Independent Events

Two events are independent if

$$P(A|B) = P(A)$$

or, put differently,

$$P(A, B) = P(A)P(B).$$

Independent Random Variables

Two random variables X, Y are independent if for all admissible sets A, B

$$P(X \in A, Y \in B) = P(X \in A)P(X \in B).$$

Or, alternatively, if the densities factorize:

$$p(x, y) = p(x)p(y).$$

An example, cont'd

Random variable X = Has been drinking hot lemon tea

Random variable Y = Has been cured from a cold in less than 3 days.

X↓, Y→	< 3 days	> 3 days
tea	0.3	0.2
no tea	0.1	0.4

We have already compute the marginals:
$P(\text{tea}) = 0.3 + 0.2 = 0.5$, and $P(< 3\text{days}) = 0.4$.

Let's reconstruct the joint probabilities based on the marginals:

X↓, Y→	< 3 days	> 3 days
tea	0.2	0.3
no tea	0.2	0.3

Do not seem to be independent!

Covariance and Correlation

If two random variables are independent,

$$E(XY) = E(X)E(Y).$$

On the other hand, two random variables are *uncorrelated*, if the *covariance* is zero.

$$\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y) = 0.$$

Normalized by the variances, we get the *correlation coefficient*

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}.$$

An example, cont'd

Encode "yes" as 1, "no" as 0.

X↓, Y→	1	0
1	0.3	0.2
0	0.1	0.4

Then:

- $P(XY = 1) = 0.3$
- $E(X) = P(X = 1) = 0.3 + 0.2 = 0.5$
- $E(Y) = P(Y = 1) = 0.3 + 0.1 = 0.4$
- $E(XY) = 0.3$
- $\mathrm{Cov}(X, Y) = 0.3 - 0.4 \times 0.5 = 0.1$

Variables are also not uncorrelated!

Covariance and Variance

Another way to write correlation:

$$\text{Cov}(X, Y) = E\left[(X - E(X))(Y - E(Y))\right],$$

because:

$$\begin{aligned}
&E\left[(X - E(X))(Y - E(Y))\right] \\
&= E\left[XY - E(X)Y - XE(Y) + E(X)E(Y)\right] \\
&= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) \\
&= E(XY) - E(X)E(Y).
\end{aligned}$$

Which means that $\text{Cov}(X, X) = \text{Var}(X)$.

I.i.d. Random Variables

"i.i.d." = Independent and Identically Distributed

Examples for using i.i.d. random variables.

- For example, i.i.d. variables are used to model repeatable independent experiments.
- Used to model joint probabilities by building from "independent" factors.

Example: Coin Tosses

Repeatedly throwing a coin can be modeled as by i.i.d. random variables.

Gives, for example: T, H, H, H, T, H, H, H, T, H, T, ... H, H, H, H, T

Probability of a certain sequence $S = (T, H, H, H, T, ...)$

$$\begin{aligned} P(X = S) &= P(X_1 = T, X_2 = H, X_3 = H, X_4 = H, X_5 = T, ...) \\ &= \prod_i P(X_i = S_i) \end{aligned}$$

Example: Coin Tosses

For example, what is the probability that the kth toss is the first "head?" (let's call this H_k?)

$$\begin{aligned} P(H_k) = P(&X_1 = T, \ldots, X_{k-1} = T, \\ &X_k = H, \\ &X_{k+1} = H \text{ or } T, X_{k+2} = H \text{ or } T, \ldots) \end{aligned}$$

(To be precise, we are first marginalizing all $X_{k+1}, X_{k+2}, \ldots$)

Then,

$$P(H_k) = P(X_1 = T)^{k-1} P(X_1 = H) = (1-q)^{k-1} q.$$

Then, we could compute the expected number of "tails" before we see a "head."

Sums of Random Variables

Let $X_1, X_2, \ldots$ be i.i.d. random variables.

Consider

$$S_n = \sum_{i=1}^{n} X_i.$$

We can compute its expectation:

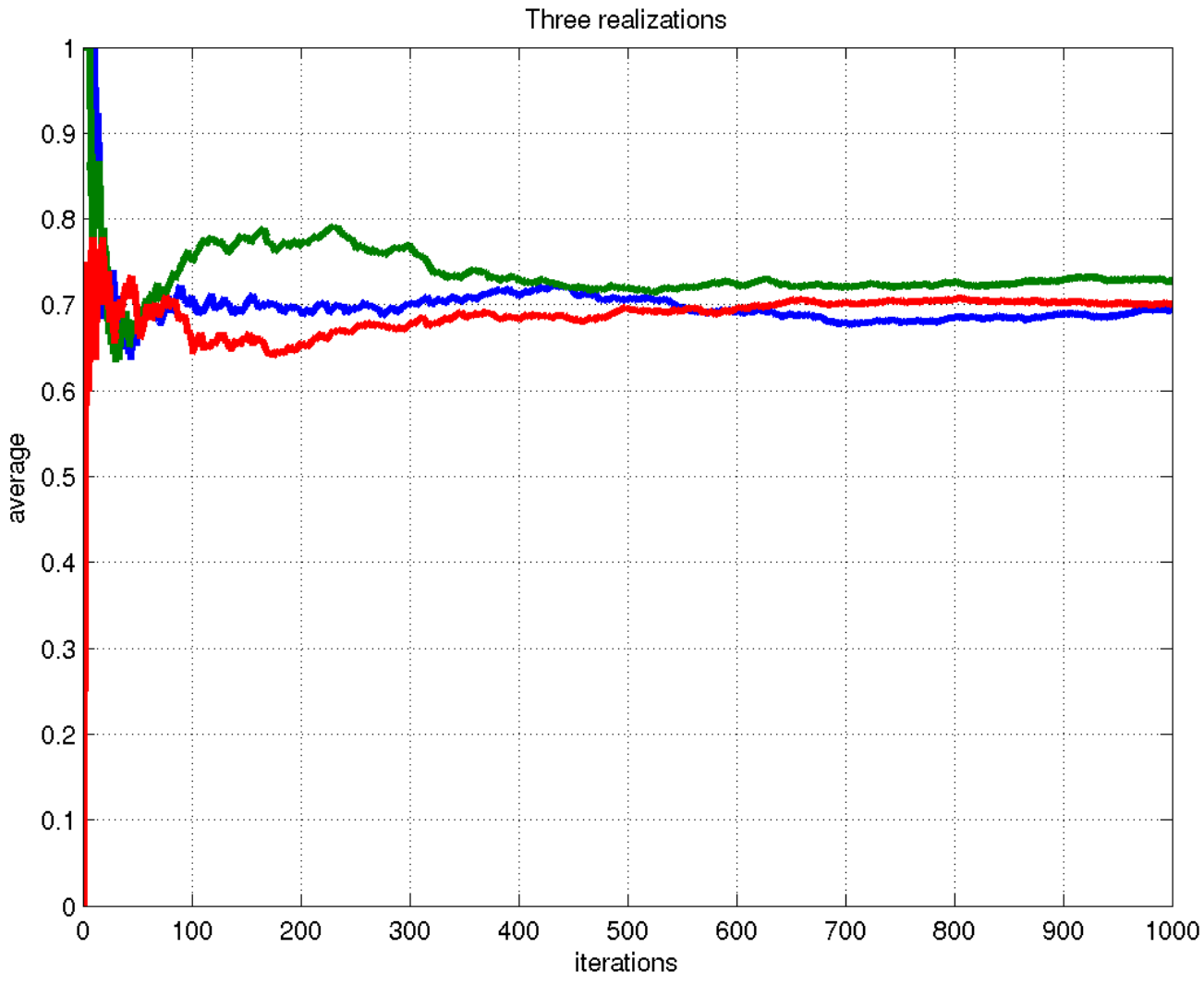
$$E(S_n/n) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = E(X_1).$$

And the variance:

$$\mathrm{Var}(S_n/n) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{1}{n} \mathrm{Var}(X_1).$$

Variance decreases with n!

An example: Averages of Bernoulli Variables with p = 0.7



Convergence is rather slow!

Laws of Large Numbers

- **Weak/Strong Law of Large Numbers**
  Basically,
  $$\frac{1}{n}S_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to E(X_1)$$
  in probability/almost surely.
- **Central Limit Theorem**
  S_n becomes more Gaussian as $n \to \infty$.
  More concretely:
  $$\frac{S_n - n\mu}{\sqrt{\sigma^2 n}} \to N(0,1),$$
  where $\mu = E(X_1)$, $\sigma^2 = \mathrm{Var}(X_1)$.

How is convergence measured?

Convergence of Random Variables

Convergence usually measured with respect to some measure, for example:

A sequence $a_1, a_2, \ldots$ converges to some number a if

$$|a_n - a| \to 0.$$

But for random variables?

Convergence of Random Variables, cont'd

- **Convergence in Probability/weak convergence**

$$\lim_{n\to\infty} P\{|X_n - X| > \epsilon\} \to 0 \qquad \text{for all } \epsilon > 0.$$

- **Convergence in d-mean**

$$\int |X_n - X|^d dP \to 0.$$

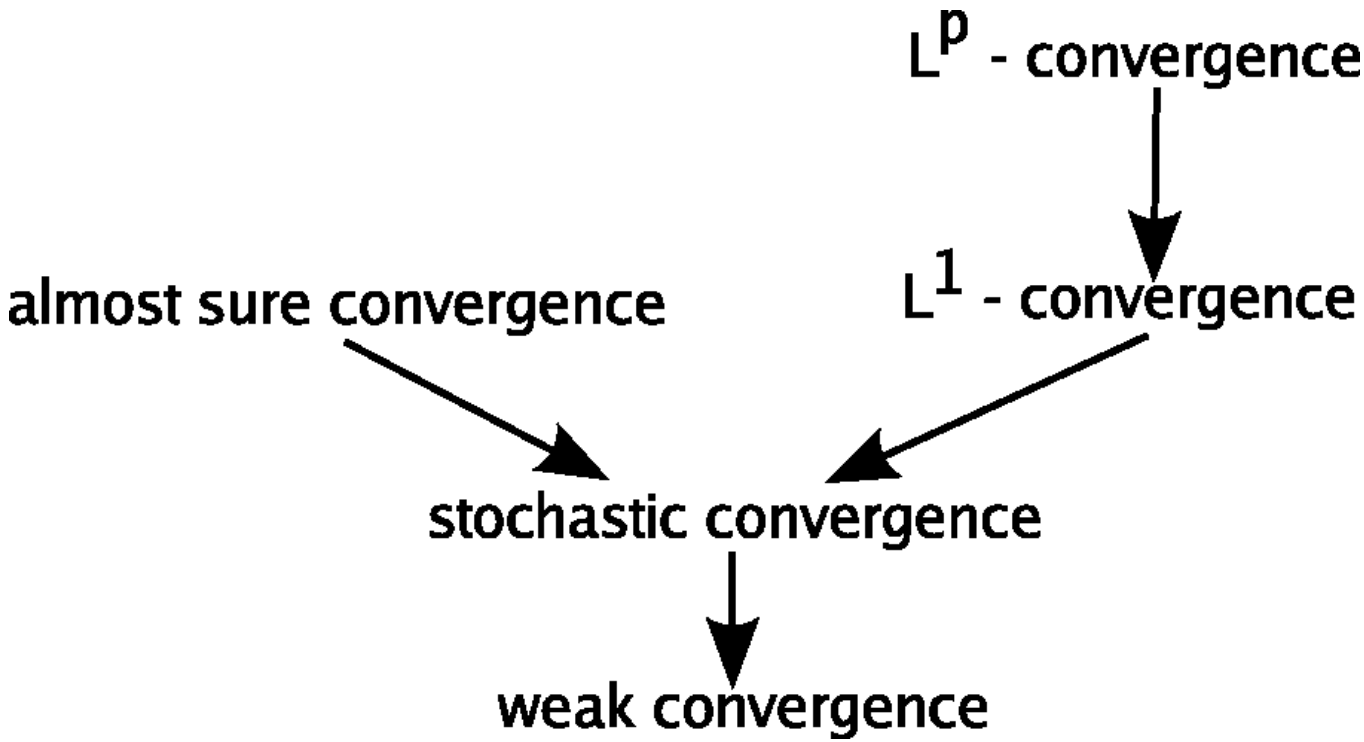
- **Strong convergence/almost sure convergence**

$$P\{|X_n - X| < \epsilon \text{ eventually}\} = 1.$$

- **Convergence in distribution**

$P_X \to P_Z$, meaning $E_X(f) \to E_Z(f)$ for all continuous functions.

Implications Between Types of Convergence

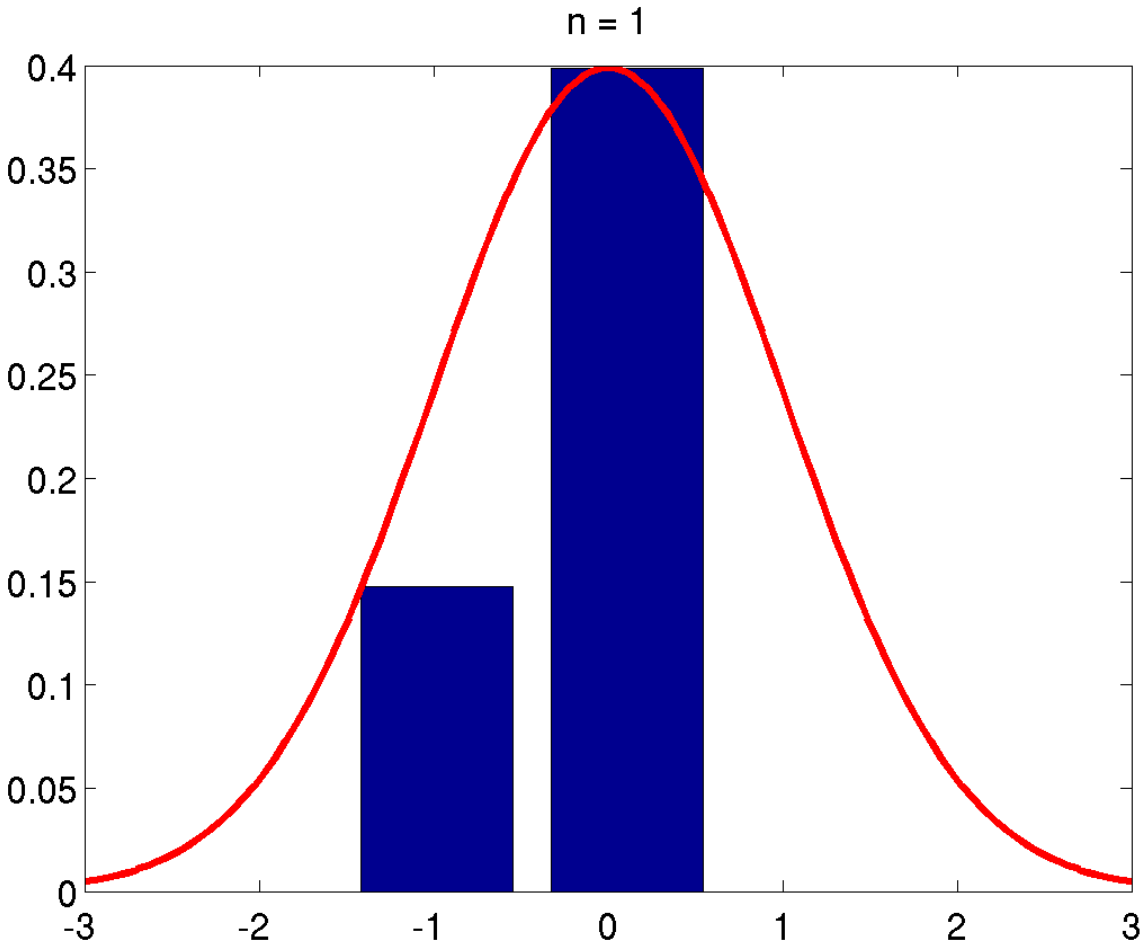


Laws of Large Numbers (Again)

- **Weak/Strong Law of Large Numbers**
  Basically,
  $$\frac{1}{n}S_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to E(X_1)$$
  in probability/almost surely.
- **Central Limit Theorem**
  S_n becomes more Gaussian as $n \to \infty$.
  More concretely:
  $$\frac{S_n - n\mu}{\sqrt{\sigma^2 n}} \to N(0,1),$$
  where $\mu = E(X_1)$, $\sigma^2 = \mathrm{Var}(X_1)$.

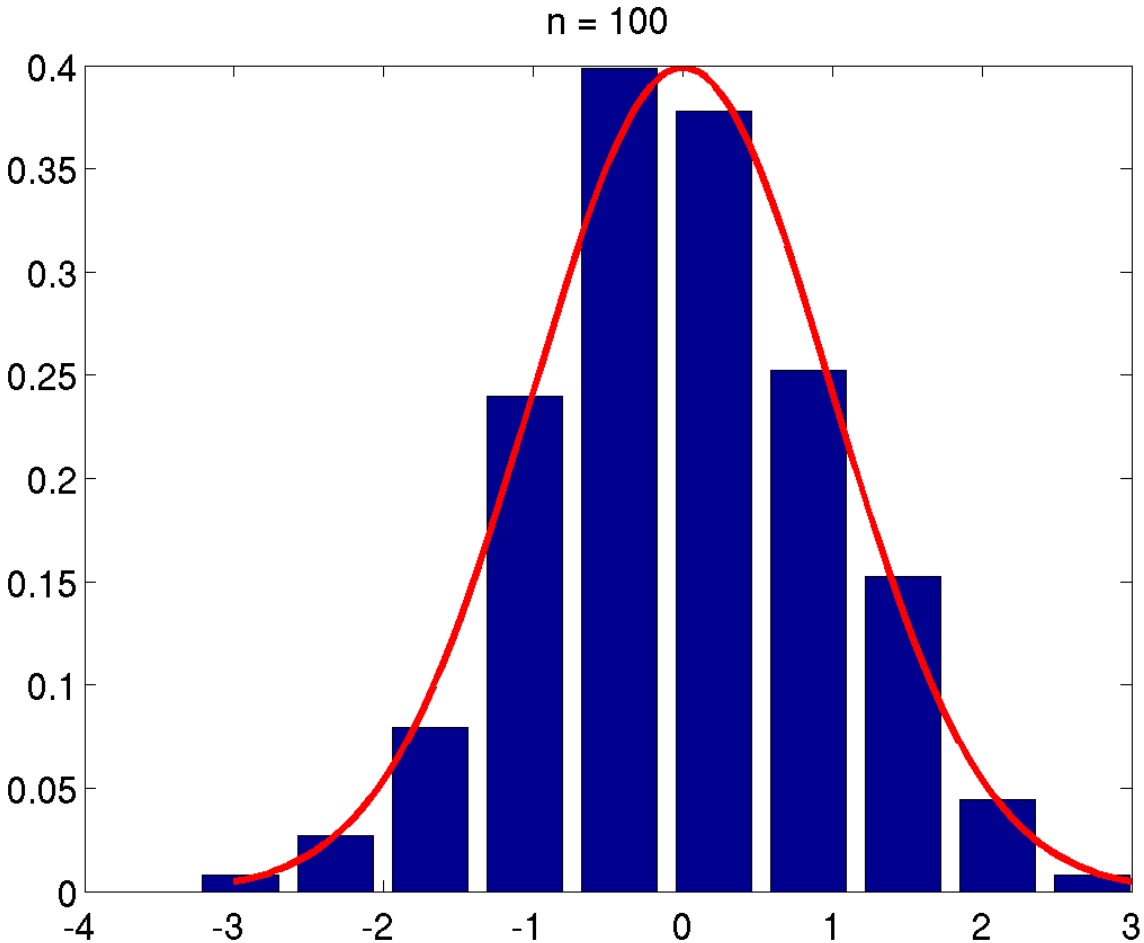
Example for the Central Limit Theorem

Sums of Bernoulli Variables with $p = 0.7$



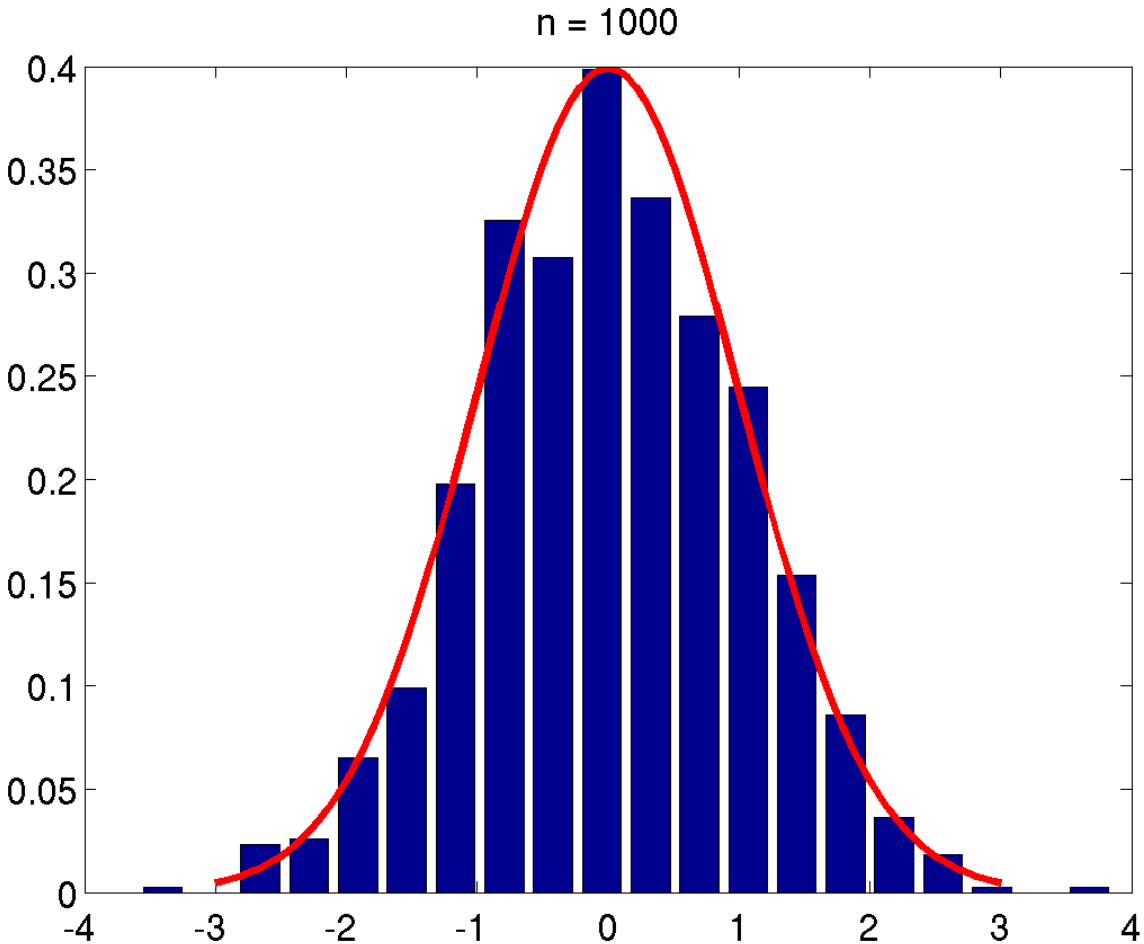
Example for the Central Limit Theorem

Sums of Bernoulli Variables with $p = 0.7$



Example for the Central Limit Theorem

Sums of Bernoulli Variables with $p = 0.7$



Practical Consequences of the Law of Large Numbers

- Often, it makes sense to assume that examples are obtained by drawing from an i.i.d. source.
- *Estimators* for quantities like the mean, variance, correlation, etc.
- Possibilities to extract parameters with small error of the data from a large number of data sets.
- In ensurance, the LoLN is also called the “Produktionsgesetz der Versicherungswirtschaft”.

Conditional Expectation

Written as $E(Y|X)$.

First, we define the expectation if we have densities for a fixed realization of x:

$$E(Y|X = x) = \int yp(y|x)dy.$$

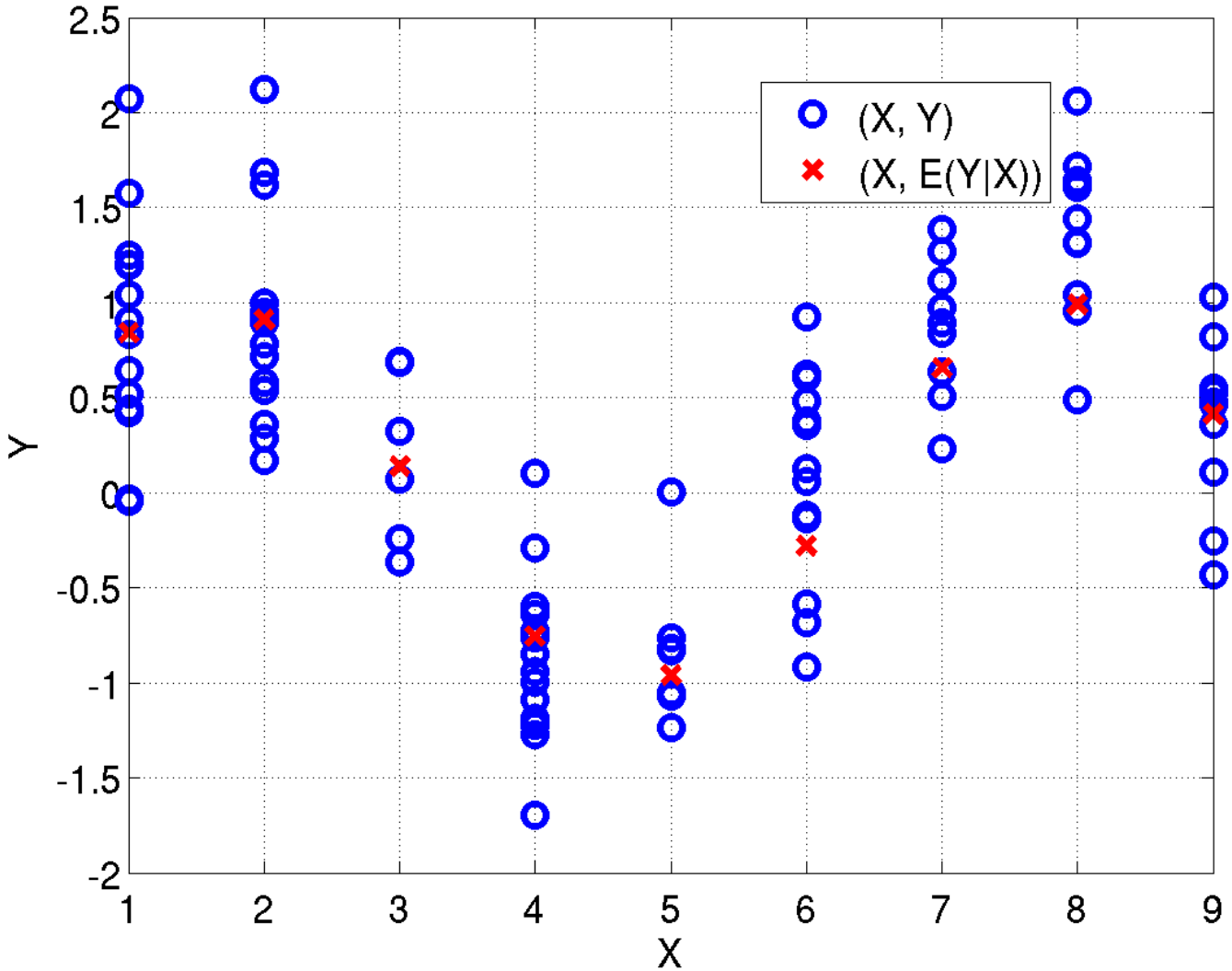
Then, $E(Y|X)$ is $E(Y|X = x)$ where we plug in the values of X for x.

This means, *that conditional expectations are again random variables*.

Conditional expectation can also be understood as *smoothing out the "excess randomness" in Y beyond X.*

Without densities, definition is much more involved.

Example



Markov Chains

One step away from completely independent models is assuming that there is some (limited) interaction between random variables.

Typically, one considers a sequence of random variables $X_1, X_2, \ldots$, called a random process.

Then, the process is said to fulfill the Markov property if

$$P(X_k|X_{k-1}, \ldots, X_1) = P(X_k|X_{k-1}).$$

That is, given X_{k-1}, X_k is *conditionally independent* of all the rest.

Other variant exist, but basically, X_k only depends on its "neighbors."

Examples

Let X_i be i.i.d. random variables which take values ± 1 with probability 0.5.

Then, the $S_n = \sum_{i=1}^{n} X_i$ are a Markov chain, because

$$
\begin{aligned}
& P\{S_n = x | S_{n-1} = x_{-1}, S_{n-2} = x_{-2}, \ldots\} \\
= & \begin{cases} 0.5 & \text{if } x = x_{-1} \pm 1, \\ 0 & \text{else} \end{cases} \\
= & P\{S_n = x | S_{n-1} = x_{-1}\}.
\end{aligned}
$$

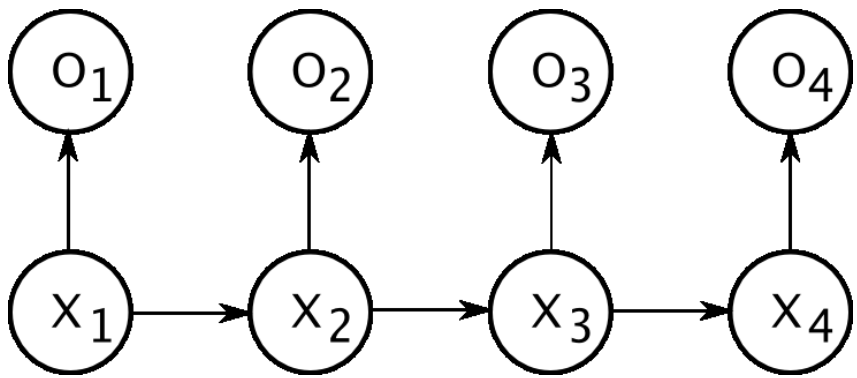
Higher Order Markov Models and Hidden Markov Models

Higher Order Markov Models

Transition probabilities only depend on previous k steps.

Hidden Markov Models

States (that is, values of X_i) cannot be observed, only outputs O_i which only depend on the value of X_i.



Used for example, for modelling natural languages.

Probability Theory and Machine Learning

The Supervised Learning Setting

- The goal is to learn a function $g\colon \mathcal{X} \to \mathcal{Y}$ from examples.
- Examples are assumed to be drawn i.i.d. from a probability distribution $P_{X\times Y}$ on $\mathcal{X} \times \mathcal{Y}$.
- Fitness of a function g is measured by a *loss function* $L\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. The goal is to minimize the *expected risk*

$$R(g) = E(L(g(X), Y)).$$

Classification and Regression

Classification

- When $\mathcal{Y} = \{\pm 1\}$ or $\{1, \ldots, K\}$
- Typical choice for L is the 0-1-loss:

$$L(y, y') = \begin{cases} 0 & y = y' \\ 1 & \text{else} \end{cases}$$

Regression

- $\mathcal{Y} = \mathbb{R}$.
- Typical choice for L is the squared loss:

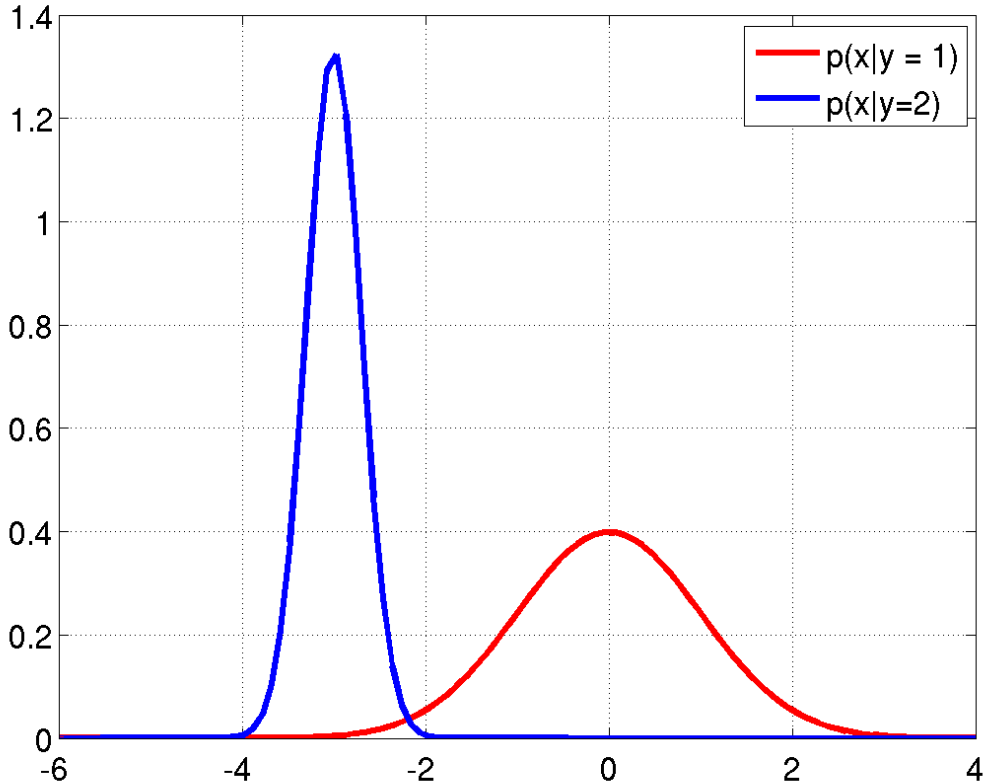
$$L(y, y') = (y - y')^2$$

Some Implications for the Supervised Learning Setting

- We assume that the X are observed with a certain probability, some more often than others. We get the distribution of X by marginalizing $P_{X \times Y}$.
- Also, we do not expect to see all possible values of Y with the same probability (marginalize $P_{X \times Y}$ by Y).
- For a given fixed x, the observed Y are not fixed, but have a probability distribution $P(Y|X)$. This means there may also exist errors in the output examples.
- By the law of large numbers, we can estimate the risk $R(g)$ for a fixed g by its *empirical risk*

$$\hat{R}(g) = \frac{1}{n} \sum_{i=1}^{n} L(g(X), Y).$$

Classification Example



For each possible y, $p(x|y)$ is the distribution of class y.

Bayes Formula

From the definition of the conditional probability, we get

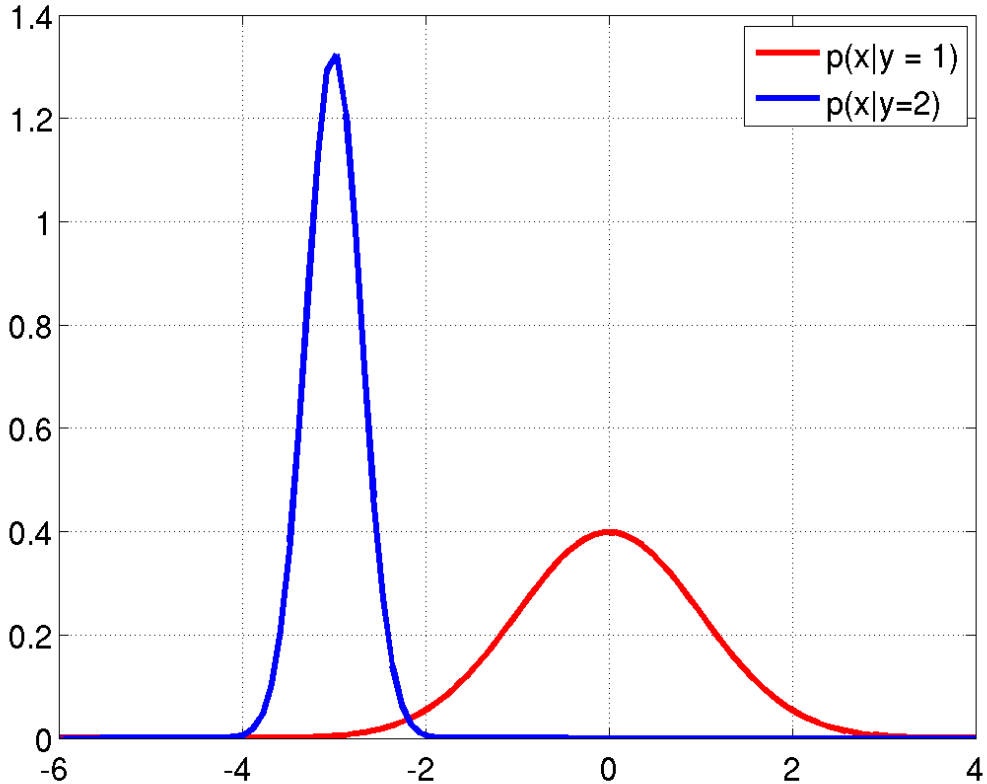
$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x|y)p(y)}{p(y)} = \frac{p(x|y)p(y)}{\sum_x p(x|y)p(y)}$$

This formula is the basis for Bayes theory.

The different parts are called:

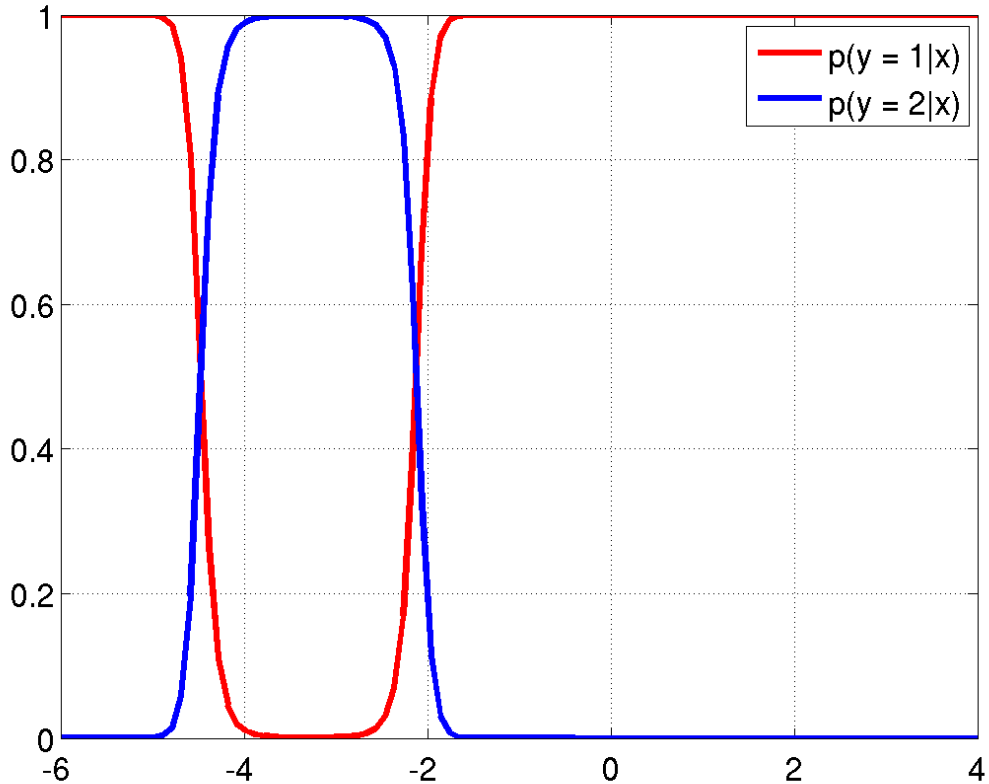
$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

Classification Example



For each possible y, $p(x|y)$ is the distribution of class y.

Classification Example



For each possible y, $p(x|y)$ is the distribution of class y.

Summary

- Events, Probabilities
- Random Variables
- Expectation, Variance
- Independence
- Laws of Large Numbers

- Supervised Learning Setting
- Bayes Formula