Part IV

Probability theory

Recommended text book:

Jacod / Protter: Probability essentials.

If you want to see all the maths behind it, I recommend

Shiryaev: Probability

σ-Algebra

Motivation

Given some set X, want to "measure" the size of sets.

- define notion of a "volume", eg to define the Lebesgue integral
- define notion of a "probability" of a set

Ideally, would like to assign a "measure" to each subset of X.
Then the measure might be a mapping

$$m : \underbrace{\mathcal{P}(X)}_{\text{power set, all subsets}} \to \mathbb{R}^+$$

However, for many domains, including $X = \mathbb{R}$, this leads to mathematical problems. Need to restrict the set of "measurable" subsets.

σ - Algebra

<u>Def</u> X non-empty set. A <u>σ-algebra on X</u> is a non-empty collection $\mathcal{F}$ of subsets of X such that

(i) $\mathcal{F}$ is closed under taking complements:

$$A \in \mathcal{F} \Rightarrow X \setminus A \in \mathcal{F}$$

(ii) $\mathcal{F}$ is closed under countable unions:

$$A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$$

(iii) $X \in \mathcal{F}$

Examples

- **Trivial σ-algebras:** Given X, we can define two (pretty useless) σ-algebras:
  - $\mathcal{F}_1 = \{\emptyset, X\}$
  - $\mathcal{F}_2 = \mathcal{P}(X) = \{A \mid A \subset X\}$.

$A \subset B :\Leftrightarrow$
$a \in A \Rightarrow a \in B$

- Given X, let $\mathcal{G}$ be any collection of subsets of X. The σ-algebra generated by $\mathcal{G}$ is the smallest σ-algebra $\mathcal{F}$ such that $\mathcal{G} \subset \mathcal{F}$. Notation: $\sigma(\mathcal{G})$

  Remark: Existence can be proved easily by an explicit construction:

$$\sigma(\mathcal{G}) = \bigcap \{\Sigma : \Sigma \text{ is a } \sigma\text{-Algebra that contains } \mathcal{G}\}$$

Borel - σ - algebra

- Consider a metric space (X, d) and let $\mathcal{G}$ be the collection of open subsets of X. Then the **Borel - σ - algebra** is defined as $\sigma(\mathcal{G})$.

Measures

Measurable space

Def A **measurable space** consists of a set X and a σ-algebra $\mathcal{F}$ over X. Notation: $(X, \mathcal{F})$. The sets in $\mathcal{F}$ are called **measurable**.

Measure

Def Given a measurable space $(\mathcal{X}, \mathcal{F})$, a measure is a map $\mu : \mathcal{F} \to [0, \infty]$ such that

(i) $\mu(\emptyset) = 0$.

(ii) For any countable collection of disjoint subsets $(S_i)_{i \in \mathbb{N}}$ with $S_i \in \mathcal{F}$ we have

$$\mu\Big(\bigcup_{i \in \mathbb{N}} S_i\Big) = \sum_{i \in \mathbb{N}} \mu(S_i).$$

⚠ measure is a function on $\mathcal{F}$, not on $\mathcal{X}$!

Measure space

<u>Def</u> A measurable space $(X, \mathcal{F})$ endowed with a measure μ is called a <u>measure space</u> $(X, \mathcal{F}, \mu)$.

Example: Discrete measure

<u>Discrete measure</u>: Given a finite (or countable) space $X = \{x_1, x_2, x_3, \dots\}$, define the σ-algebra $\mathcal{F}$ to be the set of all subsets of X. Consider a sequence $(m_i)_{i \in \mathbb{N}} \subset \mathbb{R}_{\geq 0}$ such that $\sum_{i=1}^{\infty} m_i$ is finite.

(Example: $m_i = \frac{1}{i^2}$)

Want to define $\mu : \mathcal{F} \to \mathbb{R}$. Proceed as follows:

$$\mu(\{x_i\}) := m_i$$ define μ on all "elementary" sets

For all other sets $A \in \mathcal{F}$ we can now deduce the measure (due to countability) by

$$\mu(A) = \sum_{x_i \in A} \mu(\{x_i\}) = \sum_{x_i \in A} m_i$$

Lebesgue measure on $\mathbb{R}$

Consider the space $\mathbb{R}$ with its Borel-σ-algebra: the σ-algebra $\mathcal{B}$ induced by the open sets $]a,b[$ $(a,b \in \mathbb{R},\ a<b)$.

To each such interval assign the volume

$$\mathrm{vol}(]a,b[) = b-a.$$

One can prove that this notion of volume can be extended to the whole σ-algebra $\mathcal{B}$. The resulting measure on $(\mathbb{R},\mathcal{B})$ is called the Lebesgue-measure and is often denoted with the letter λ.

Lebesgue measure on $\mathbb{R}^d$

Similarly to the 1-dim case, one can also construct a measure on $\mathbb{R}^d$ with the Borel-σ-algebra $\mathcal{B}$:

Consider sets of the form $]a_1, b_1[\times]a_2, b_2[\times \ldots \times]a_d, b_d[$ and assign them the volume $(b_1 - a_1) \cdot \ldots \cdot (b_d - a_d)$.

Can extend this consistently to the whole σ-algebra

$\rightsquigarrow$ multi-dim Lebesgue measure λ_d

A funny measure on $\mathbb{R}$

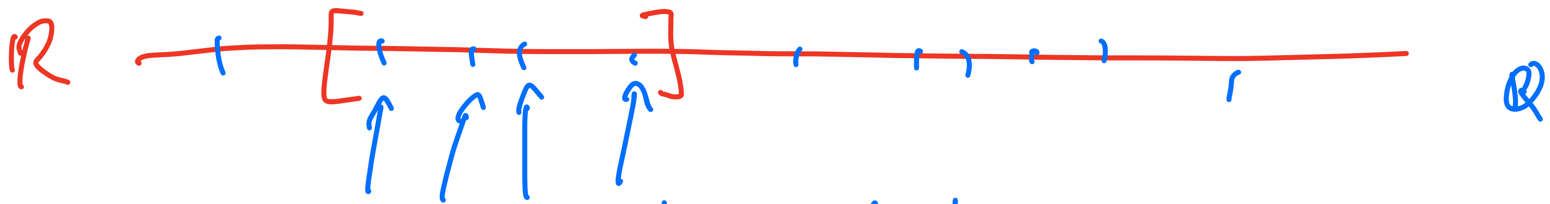
Let $\mathcal{F}$ be the Borel-σ-algebra on $\mathbb{R}$. Want to define a measure that just assigns mass to rational numbers.

Let $\underbrace{q_1, q_2, q_3, \dots}_{\mathbb{Q}}$ be all rational numbers. Countably many.

Consider $(m_i)_{i \in \mathbb{N}}$ as before: $m_i = \frac{1}{i^2}$.

$$\mu(\{q_i\}) = m_i$$

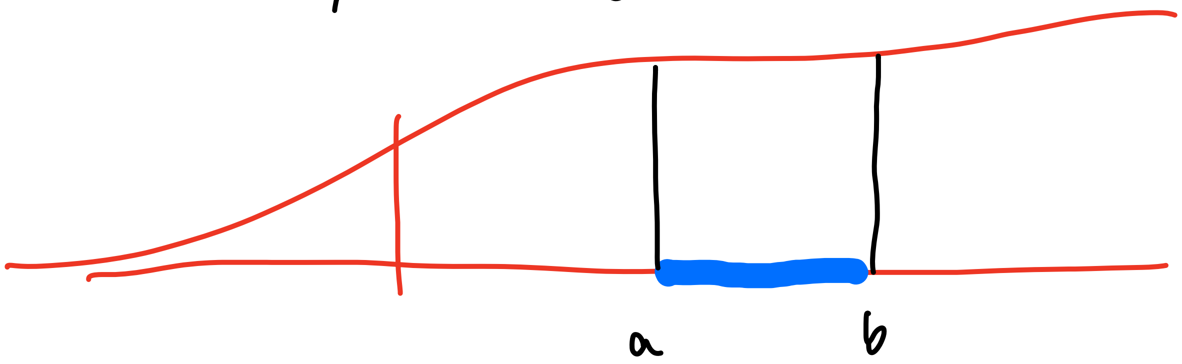
$$\mu([a,b]) = \sum_{q_i \in [a,b]} \mu(\{q_i\}) = \sum_{q_i \in [a,b]} m_i$$



sum up over all the rational points in the interval (infinitely many points!)

More general measures on $\mathbb{R}$

$X = \mathbb{R}$, $\mathcal{F}$ Borel-σ-algebra. Let $F: \mathbb{R} \to \mathbb{R}$ be monotonically increasing, continuous.



For an interval $]a,b]$ define its measure as

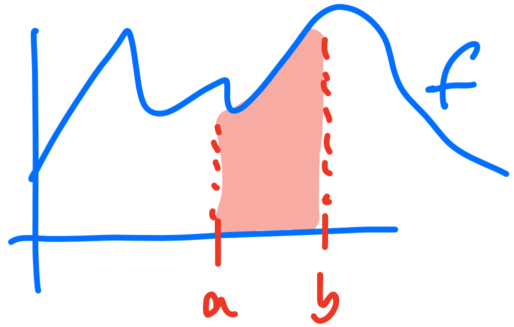
$$\mu(]a,b]) = F(b) - F(a)$$

Can prove that one can extend this "measure" to the whole σ-algebra, making it a well-defined measure on $\mathbb{R}$.

Measure with a density

Consider a function $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ that is integrable.

For intervals $[a,b]$ define



$$\nu([a,b]) := \int_a^b f(x)\, dx$$

This can be extended to a proper measure on $(\mathbb{R}, \mathcal{B})$

Then ν is the measure with density f with respect to the Lebesgue measure.

More keywords: Radon - Nikodym theorem

Carathéodory extension theorem

Have seen several instances where we defined a "measure" on intervals $[a,b]$ or $]a,b[$ or $]a,b]$ and concluded that it "extends" to the whole σ-algebra.

Mathematical basis for this approach is the Carathéodory theorem.

(skipped here)

Measures with continuous and discrete parts

Observe: measures can have "continuous and discrete" parts.

Example: If λ is the Lebesgue measure on $[0,1]$ with Borel-σ-algebra. Let δ_0 be the discrete measure that assigns mass 1 to the set $\{0\}$.

Then we can define a measure $\nu = \lambda + \delta_0$.

Dirac measure

$$A \subset [0,1]; \quad \nu(A) = \lambda(A) + \delta_0(A) = \begin{cases} \lambda(A) & \text{if } 0 \notin A \\ \lambda(A)+1 & \text{if } 0 \in A \end{cases}$$

More keywords: Lebesgue decomposition theorem of measures

Null sets

Consider a measure space $(\mathcal{X}, \mathcal{F}, \mu)$.

A subset $N \in \mathcal{F}$ is called a **null set** if $\mu(N) = 0$.

We say that a property holds **almost everywhere** if it holds for all $x \in \mathcal{X}$ except for x in a null set N.

(in probability theory, we say "**almost surely**").

Measurable mappings

Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a measure space and

and $(\mathcal{W}, \mathcal{A})$ be a measurable space.

A mapping $f : \mathcal{X} \to \mathcal{W}$ is called measurable if

$\forall A \in \mathcal{A}$, $f^{-1}(A) \in \mathcal{F}$ ("pre-images of measurable sets

are measurable").

The mapping f induces a measure ν on $(\mathcal{W}, \mathcal{A})$ via

$\nu(A) = \mu(f^{-1}(A))$.

Probability measure

Probability measure

$\mathcal{A} = S_1, S_2, \dots \subset \Omega$

$\Omega \in \mathcal{A}$

$P : \mathcal{A} \to \mathbb{R}$

A measure P on a measurable space $(\Omega, \mathcal{A})$ is called a **probability measure** if $P(\Omega) = 1$.

The elements in $\mathcal{A}$ are called **events**.

$(\Omega, \mathcal{A}, P)$ is called a **probability space**.

Example: throw a die

$\Omega = \{1, 2, \dots 6\}$, $\mathcal{A} = \mathfrak{P}(\Omega)$ (σ-algebra generated by the "elementary events" $\{1\}, \{2\}, \dots, \{6\}$).

P can be defined uniquely by assigning

$$P(\{1\}) = P(\{2\}) = \dots = P(\{6\}) = \frac{1}{6}$$

For example $P(\{1, 5\}) = P(\{1\}) + P(\{5\}) = \frac{1}{3}$

Throw two dice:

$$\Omega = \{1, 2, \dots 6\} \times \{1, 2, \dots, 6\} = \{(1,1), (1,2), (1,3), \dots\}$$

first die

second

(i,j)

$$\mathcal{A} = \mathfrak{P}(\Omega)$$

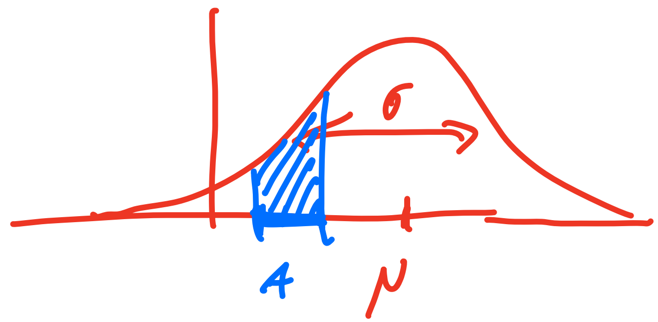
$$P(\{(i,j)\}) = \frac{1}{36}$$

Example: normal distribution

$\Omega = \mathbb{R}$

$\mathcal{A} =$ Borel-σ-algebra

$$f_{\mu,\sigma} : \mathbb{R} \to \mathbb{R},$$
$$x \longmapsto \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



$$P : \mathcal{A} \to [0,1]$$
$$P(A) := \int_A f_{\mu,\sigma}(x)\, dx$$

P is the probability measure on $(\mathbb{R}, \mathcal{A})$ with density $f_{\mu,\sigma}$.

More examples

... see later...

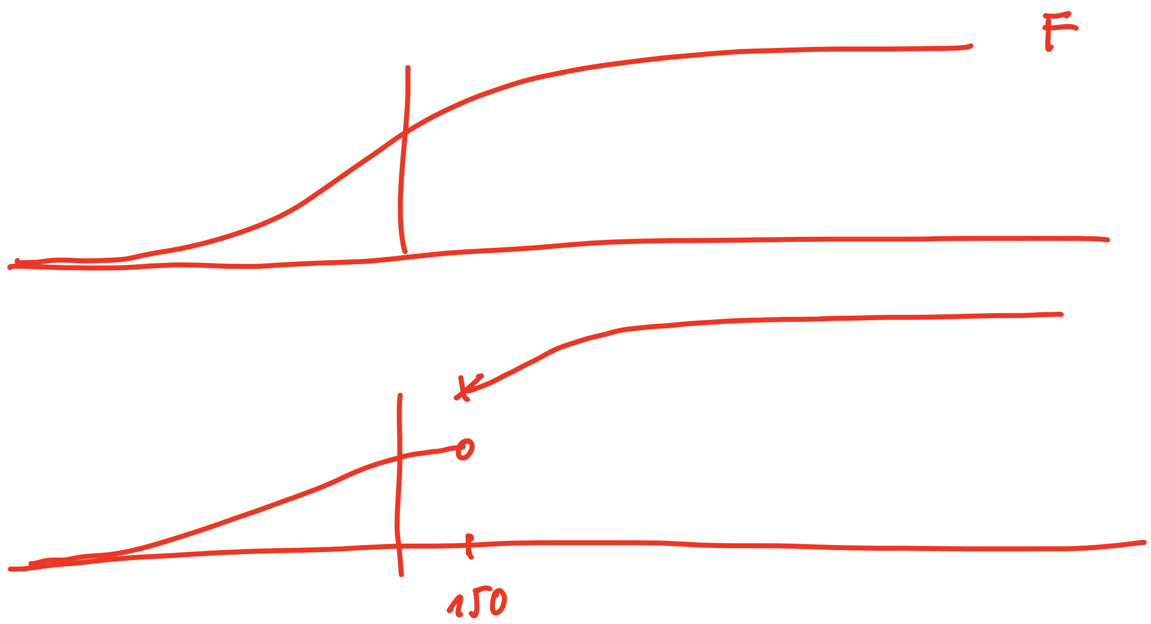
Cumulative distribution function (cdf)

Cumulative distribution function

Let P be a prob. measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

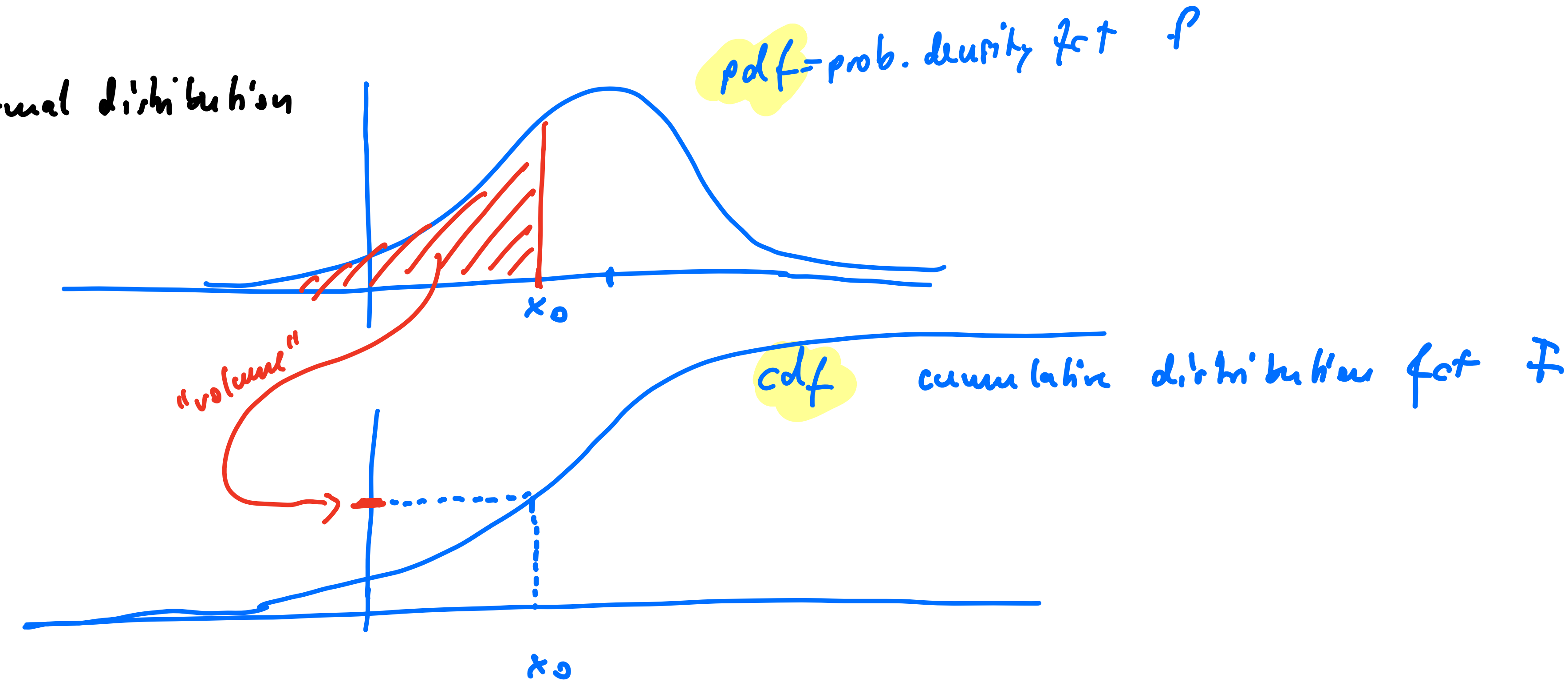
Define the function $F: \mathbb{R} \to \mathbb{R}$, $x \mapsto P(]-\infty, x])$.

We say that F is a cumulative distribution function (cdf).



cdf vs pdf

Example: normal distribution



Properties of the cdf

A cdf satisfies the following properties:

(i) F is monotonically increasing with

$$\lim_{x \to -\infty} F(x) = 0 \quad \text{and} \quad \lim_{x \to +\infty} F(x) = 1.$$

(ii) F is continuous from the right:

$(x_n)_n$ sequence with $x_n \searrow x$

(i.e. $x_n \geq x_{n+1}$ and $x_n \to x$) then also

$F(x_n) \to F(x)$.

The other way round? Is it true that a function that satisfies (i) and (ii) is always the cdf of a prob. measure? $\leadsto$ Yes:

"Cdf" gives rise to a unique prob. measure

Proposition:

Let $F : \mathbb{R} \to \mathbb{R}$ be a function with properties (i) and (ii).

Then there exist a unique prob. measure P on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$

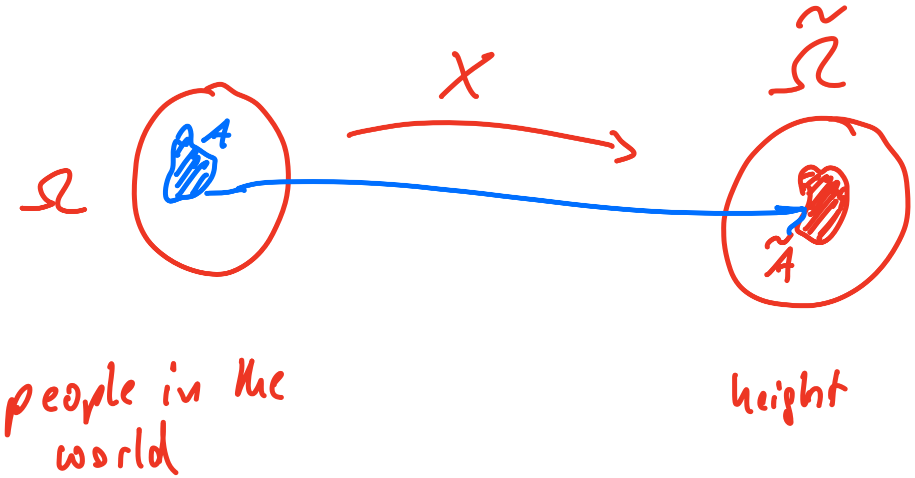
such that $P(]-\infty, x]) := F(x)$.

Random variable

Random variable

<u>Def</u> Let $(\Omega, \mathcal{A}, P)$ be a probability space, $(\tilde{\Omega}, \tilde{\mathcal{A}})$ be another measurable space. A mapping: $X : \Omega \to \tilde{\Omega}$ is called a random variable if X is measurable, i.e.

$$\forall \tilde{A} \in \tilde{\mathcal{A}} : \quad X^{-1}(\tilde{A}) := \{\omega \in \Omega \mid X(\omega) \in \tilde{A}\} \in \mathcal{A}.$$



$\tilde{A}$ = "at least 170"

A = "persons that are at least 170cm tall"

$P(A) = 0.1$

Example

Example: sum of two dice

$$\Omega = \{ (i,j) \mid i,j \in \{1, \dots 6\} \}$$

$$\mathcal{A} = \mathcal{P}(\Omega)$$

$$P(\{(i,j)\}) = \frac{1}{36}$$

$$\tilde{\Omega} = \{2, \dots, 12\}$$

$$\tilde{\mathcal{A}} = \mathcal{P}(\tilde{\Omega})$$

X "sum of the two values"

$$X : \Omega \to \underbrace{\{2, \dots, 12\}}_{\tilde{\Omega}} , \quad (i,j) \longmapsto i+j$$

is measurable.

$$P(\text{sum} = 12) = P(\{ (i,j) \mid i+j = 12 \}) = P(\{(6,6)\}) = \frac{1}{36}$$

Distribution of a rv

Def A random variable $X : (\Omega, \mathcal{A}, P) \to (\tilde{\Omega}, \tilde{\mathcal{A}})$ induces a measure on the target space:

For $\tilde{A} \in \tilde{\mathcal{A}}$ we define

$$P_X(\tilde{A}) := P(X^{-1}(\tilde{A}))$$

This is a probability measure on $(\tilde{\Omega}, \tilde{\mathcal{A}})$ and it is called the distribution of X.

Induced σ-algebra

$(\Omega, \mathcal{A}, P)$, $(\tilde{\Omega}, \tilde{\mathcal{A}})$. $X : \Omega \to \tilde{\Omega}$

Def. $X : (\Omega, \mathcal{A}, P) \to (\tilde{\Omega}, \tilde{\mathcal{A}})$. Then the family

$$\sigma(X) := \{ X^{-1}(\tilde{A}) \mid \tilde{A} \in \tilde{\mathcal{A}} \}$$

is a σ-algebra on Ω and it is called the σ-algebra induced by X

(it is the smallest σ-algebra on Ω that makes X measurable).

$$X^{-1}(\tilde{A}) = \{ \omega \in \Omega \mid X(\omega) \in \tilde{A} \}$$

Expectation

Expectation (finite case)

Consider a finite random variable $X : \Omega \to \mathbb{R}$ (that is, $X(\Omega)$ is finite).

<u>Definition</u> $(\Omega, \mathcal{A}, P)$ prob. space, $S \subset \mathbb{R}$ finite, $X : \Omega \to S$ random variable. Then

$$\mathbb{E}(X) := \sum_{r \in S} r \cdot P(X = r)$$ is called the expectation of X.

(sometimes people write EX, $\mathbb{E}X$ or $\mathbb{E}(X)$).

A rv is called "centred" if $E(X) = 0$.

Examples

- Toss a coin. $\Omega = \{\text{head}, \text{tail}\}$, $\mathcal{A} = \mathcal{P}(\Omega)$, $P(\text{head}) = p$,

  $P(\text{tail}) = 1-p$, $0 \leq p \leq 1$.

  $X : \Omega \to \{0, 1\}$, head $\mapsto 1$, tail $\mapsto 0$.

  $$E(X) = 0 \cdot \underbrace{P(X=0)}_{1-p} + 1 \cdot \underbrace{P(X=1)}_{p} = p.$$

In ML:

- Train error: expected error wrt empirical distribution which assigns prob $\frac{1}{n}$ to all the n training points.

- Test error of a classifier: expected error wrt to the (unknown) underlying data distribution

Expectation is linear

<u>Prop</u>: Let $X, Y : (\Omega, \mathcal{A}, P) \to \mathbb{R}$ be two (discrete) random variables. Then

$$E(\underbrace{a}_{\in\mathbb{R}} \cdot \overset{rv}{X} + \underbrace{b}_{\in\mathbb{R}} \cdot \overset{rv}{Y}) = a \cdot E(X) + b \cdot E(Y)$$

<u>Proof sketch</u> (discrete case)

$$E(aX+bY) = \sum_{x,y} (ax + by)\, P(X=x, Y=y) =$$

$$= a \sum_{x,y} x \cdot P(X=x, Y=y) + b \sum_{x,y} y\, P(X=x, Y=y) =$$

$$= a\Big(\sum_x x \cdot \underbrace{\sum_y P(X=x, Y=y)}_{P(X=x)}\Big) + b\,(\ \dots\ )$$

law of total prob.

$$= a \sum_x x \cdot P(X=x) + b \sum_y y\, P(Y=y) = a \cdot E(X) + b \cdot E(Y).$$

∎

Expectation of a general rv

Motivation

Let $(\Omega, \mathcal{A}, P)$ be a measure space, and $X: \Omega \to \mathbb{R}$ a r.v.

Want to define $E(X) :=$ " $\int X \, dP$ "

If P has a probability density p and $X: \mathbb{R} \to \mathbb{R}$, we might do something similar as the Riemann integral:

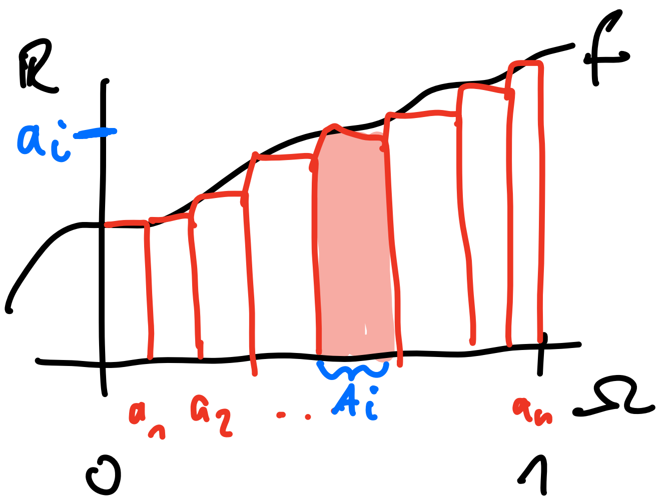
$$E(X) \approx \sum_i X(A_i) \cdot P(A_i)$$

$$\xrightarrow[\lim]{} \int x \, p(x) \, dx$$

Motivation

Consider a probability space $(\Omega, \mathcal{A}, P)$ and a measurable function $f: X \to \mathbb{R}$. Want to compute "$\int f(x)\, dP$", the integral of the function wrt probability measure P.

① In case $\Omega = [0,1]$ with the uniform distribution, this is easy: it is the same as computing the "normal integral":

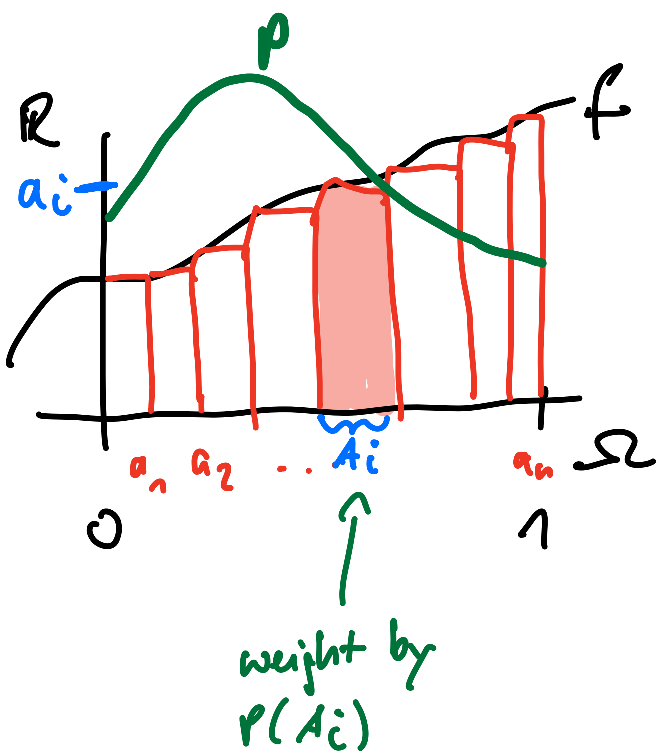


Intuitively, we partition the space $\Omega = \mathbb{R}$ into sets $A_1, A_2, \dots, A_n$ and compute the approximate integral:

$$\sum_{i=1}^{n} f(a_i) \cdot \underline{\underline{\text{length}}}(A_i)$$

Then we let $n \to \infty$ and (hopefully) converge to $\int_0^1 f(x)\, dx$.

② Now consider a case of Ω with a prob. distribution with density p:



We now want to pursue a similar approach.

But instead of using the length of A_i as volume, we now want to use the probability of A_i as notion of volume:

$$\sum_{i=1}^{n} f(a_i) \underline{\underline{P}}(A_i)$$

Then we let $n \to \infty$ and (hopefully) converge to

$$\int_0^1 f(x) \underbrace{p(x)}_{\text{weight by the prob. density}} dx$$

③ But what do we do if we don't have a density ρ? Or a more general input space Ω?

We cannot describe the "weight" by a function ρ, so we need a more general construction.

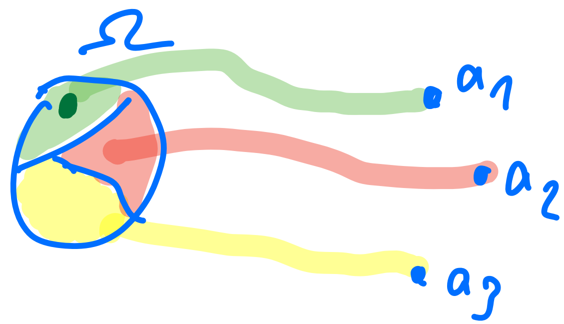
As it will turn out, we will turn this process upside down:

Instead of starting with a partition of the input space Ω and refining it (which might be difficult because Ω does not have any "structure" beyond a σ-algebra), we use a partition of the target space $\mathbb{R}$:

Construction of an integral : simple rv

Step 1 : Assume the random variable $X : \Omega \to \mathbb{R}$ only takes finitely many values : $X \in \{a_1, \dots, a_k\}$, that is there exist some sets $A_1, \dots, A_k$ such that

$$\mathbb{1}_{A_i}(\omega) = \begin{cases} 1 & \text{if } \omega \in A_i \\ 0 & \text{otherwise} \end{cases} \qquad X(\omega) = \sum_{i=1}^{k} a_i \, \mathbb{1}_{A_i}(\omega)$$

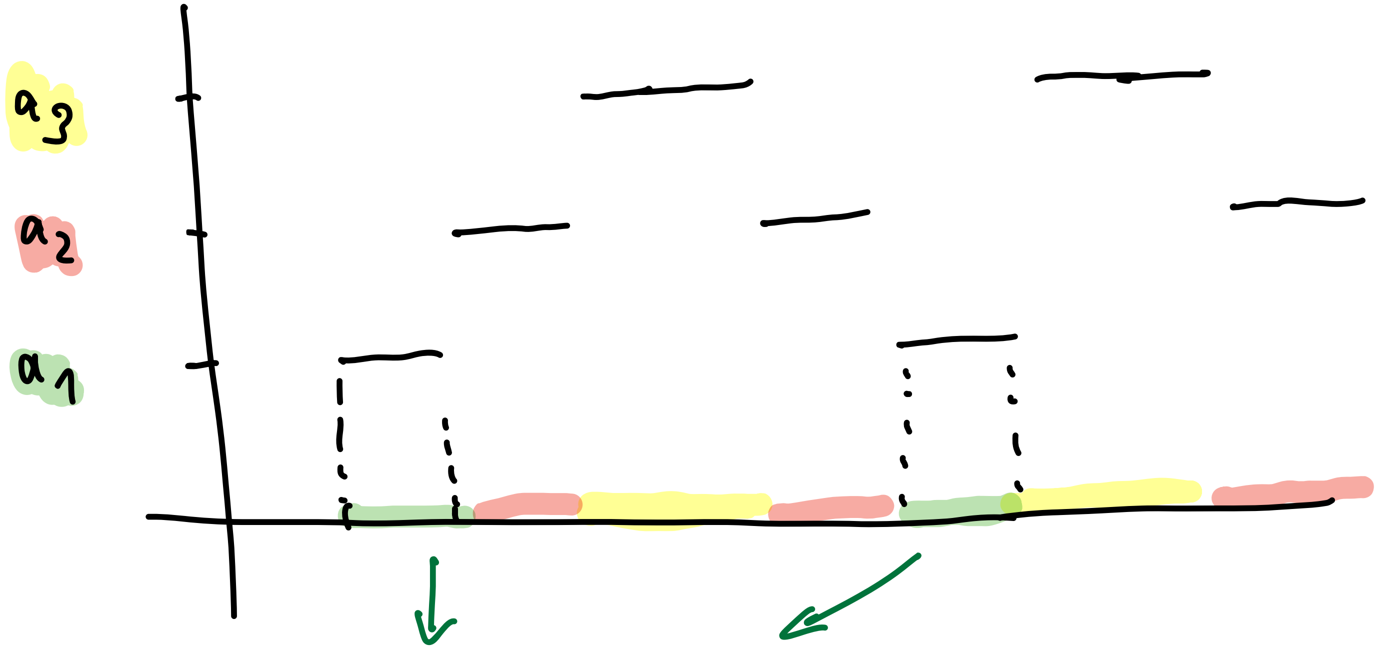


Note that because X is a rv, the sets A_i need to be ext.

We call such a rv "simple". We now define

$$E(X) := \sum_{i=1}^{k} a_i \, P(A_i)$$

(Note that this coincides with the previous def. in the discrete case).



together these two intervals are the set A_1

Construction of an integral: non-negative rv

<u>Step 2</u>: Assume the rv takes values in $[0, \infty[$. We define

$$\text{"}E(X)\text{"} := \sup \{ E(Y) \mid Y \text{ simple rv with } \underbrace{0 \le Y \le X}_{Y \le X \iff \forall \omega \in \Omega:\ Y(\omega) \le X(\omega)} \}$$

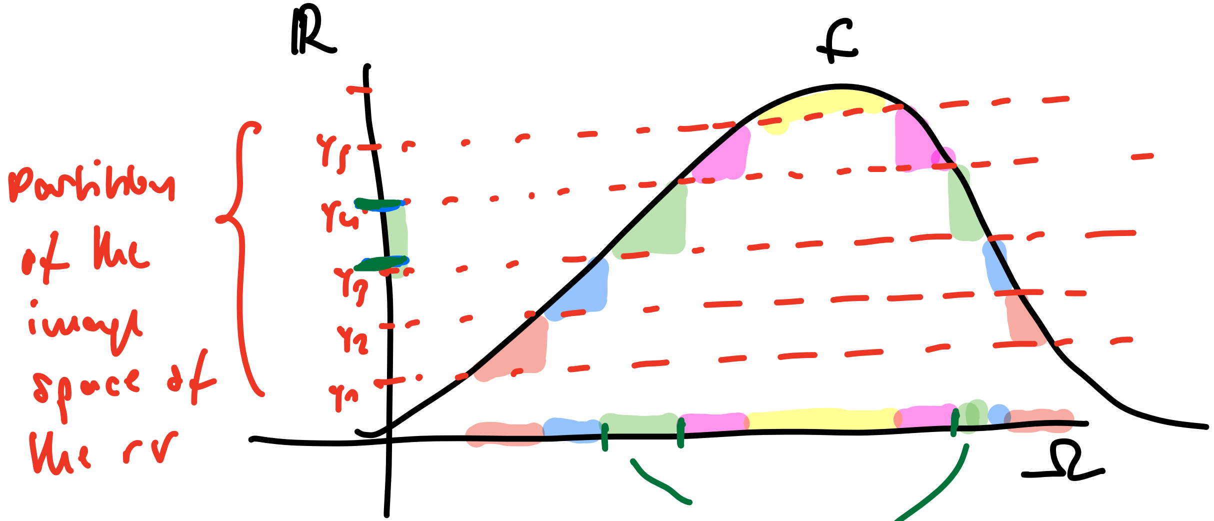
If the supremum, exists and is finite, we call the resulting number the expectation $E(X)$ of the non-neg. rv X.

Intuitively:

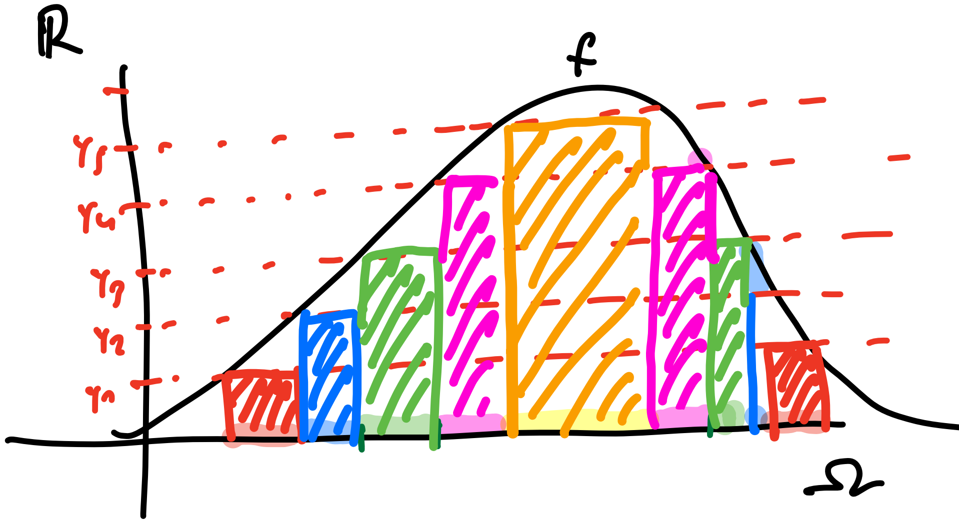
- we "discretize" the output of the rv into finitely many values.
- we look at the corr. partitions $A_1, \dots, A_n$ of the input space Ω
- we use these partitions to "approximate" X from below.
- By taking the supremum over all such partitions, we implicitly use finer and finer partitions (without the explicit need to define what "refinements" of partitions are)

Illustration of Step 2

Consider the case of a rv $f : \mathbb{R} \to \mathbb{R}$



$$A_3 = \{ x \mid f(x) \in [y_3, y_4[\}$$

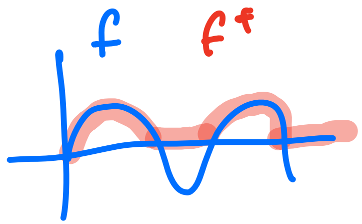


approximation of f based on the partition induced by the output space

Construction of an integral: general case

Step 3: Let X be a general rv $X : \Omega \to \mathbb{R}$.

Define $X^+ := \max\{X, 0\}$ and $X^- := -\min\{X, 0\}$.



Then $X = X^+ - X^-$, and both X^+ and X^- are non-negative rvs.

Now assume that both X^+ and X^- have finite expectations.

Then we define

$$E(X) = E(X^+) - E(X^-).$$

Notation: $E(X) = \int_\Omega X(\omega)\, dP(\omega) = \int X \, dP$

Important properties of the expectation

We introduce the notation $L^1(\Omega, \mathcal{A}, P)$ to denote all random variables for which a finite expectation exists.

- Consider two rvs X, Y on Ω such that $X(\omega) \leq Y(\omega)$ $\forall \omega$ and $X, Y \in L^1$. Then $E(X) \leq E(Y)$.

- $X \in L^1(\Omega, \mathcal{A}, P) \iff |X| \in L^1(\Omega, \mathcal{A}, P)$

  In this case: $E(X) \leq E(|X|)$

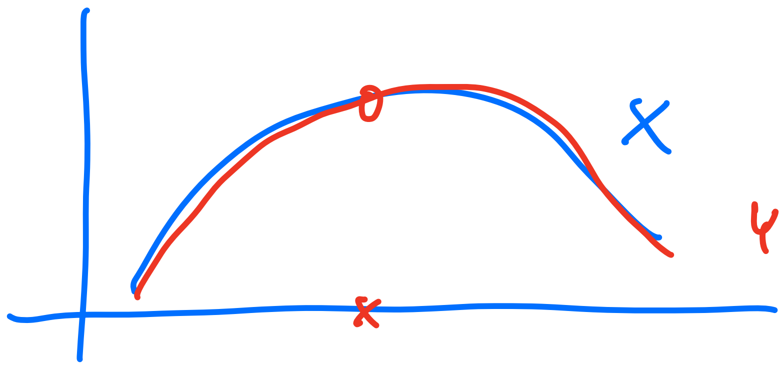
- Any bounded rv possesses an expectation:
  - X bounded $\iff$ $\forall \omega \in \Omega : |X(\omega)| \leq c$ for some $c \in \mathbb{R}$
  - "possess an expectation": have a well-defined, finite $E(X)$

- $X = Y$ a.s. $\Rightarrow E(X) = E(Y)$

Recall: $X = Y$ a.s. $\Leftrightarrow \exists N \in \mathcal{A}$ with $P(N) = 0$ such that $\forall \omega \in \Omega \setminus N$:

$$X(\omega) = Y(\omega)$$

"functions agree everywhere except on a set of measure 0"

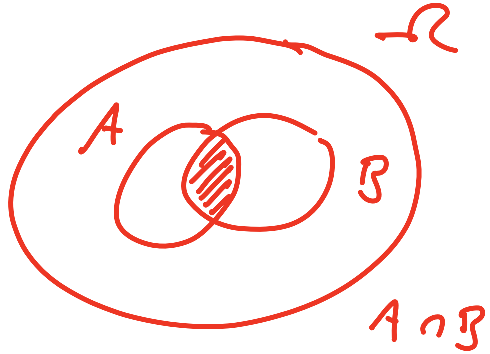


Conditional probability

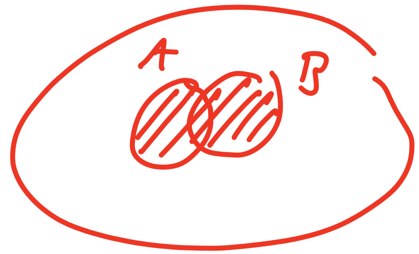
Unions and intersections

Notation:

$$P(A \cap B) = P(\text{"}A \text{ and } B\text{"})$$



$$P(A \cup B) = P(\text{"}A \text{ or } B\text{"})$$



Conditional probability

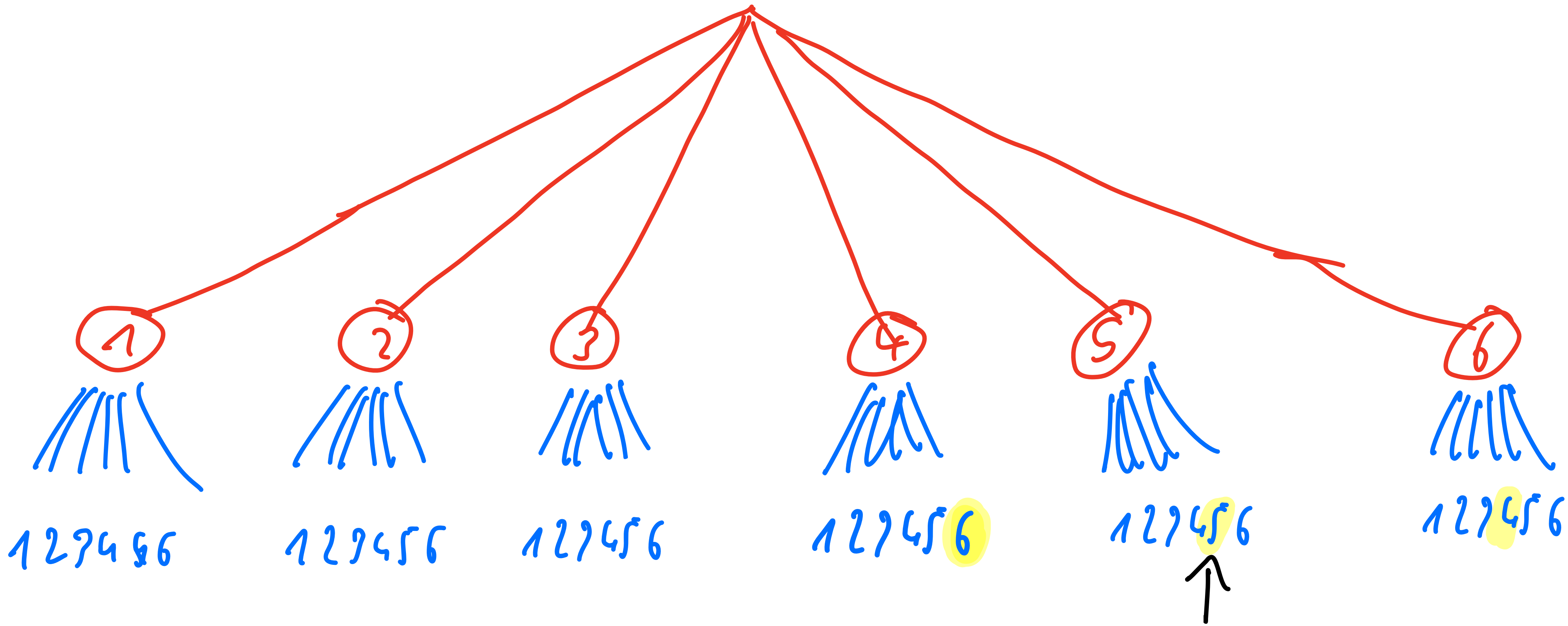
<u>Def</u> $(\Omega, \mathcal{A}, P)$ probability space, $A, B \in \mathcal{A}$, $P(B) > 0$. Then

$$P(A \mid B) := \frac{P(A \cap B)}{P(B)}$$ is called the

conditional probability of A given B.

Example with two dice

$$P(\text{"sum is 10"} \mid \text{"first die is 5"})$$



$$P(\text{"sum is 10"} \ \underline{\text{and}} \ \text{"first die is 5"}) = \frac{1}{36}$$

$$P(\text{"sum is 10"} \mid \text{"first die is 5"})$$

$$= \frac{P(\text{"sum is 10" and "first die is 5"})}{P(\text{"first die is 5"})} = \frac{1/36}{1/6} = \frac{1}{6}$$

Makes sense: when we know the first die is 5, then the second one has to be 5 as well to achieve sum = 10, which happens with 1/6 prob.

Note: of course the "ordering" matters

Ω = "all persons on earth" (a finite set)

$\mathcal{A} = \mathcal{P}(\Omega)$

P = "uniform"

Event A := "person has been vaccinated"

B := "person has disease"

two different conditional distributions:

$$\left.\begin{aligned} &P(\text{disease} \mid \text{vaccinated}) \\ &P(\text{vaccinated} \mid \text{disease}) \end{aligned}\right\} \text{not the same!!!}$$

Conditional distribution (positive condition)

Theorem Let $B \in \mathcal{A}$ with $P(B) > 0$.

The mapping $P_B : \mathcal{A} \to [0,1]$, $A \mapsto P(A \mid B)$ is a probability measure on $(\Omega, \mathcal{A})$, it is called the conditional distribution of P with respect to B.

B is fixed, $P(A \mid B)$ as a fct. of A

Regular conditional distribution

In the definition of the cond. distribution we require $P(B) > 0$.

But:

In ML we often want to condition on events with probability 0:

$$P(Y = 1 \mid \underbrace{X = 0.1}_{\text{training pt. } X \text{ drawn from normal distribution}})$$

Under certain assumptions, the resulting "conditional distribution" exists and well defined (needs heavy maths...). In ML we can typically take these assumptions for granted... see later.

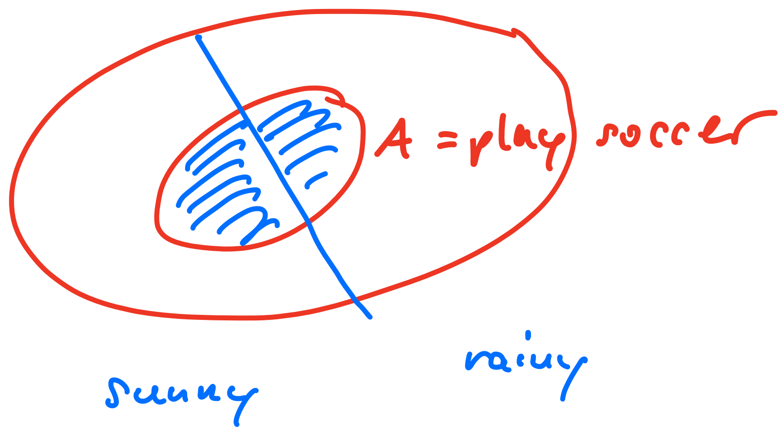
Law of total probability and Bayes formula

Law of total probability

Let $B_1, B_2, \ldots, B_k$ be a disjoint partition of Ω

with $B_i \in \mathcal{A}$, $P(B_i) > 0$ for all i, and $A \in \mathcal{A}$. Then

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i) \cdot P(B_i) = \sum_{i=1}^{n} P(A \cap B_i)$$



Bayes formula

Let $B_1, B_2, \dots, B_k$ be a disjoint partition of Ω with $B_i \in \mathcal{A}$ for all i, and $A \in \mathcal{A}$ with $P(A) \neq 0$.

$$P(B_i \mid A) = \frac{P(A \cap B_i)}{P(A)} = \frac{P(A \mid B_i) \cdot P(B_i)}{\sum_i P(A \mid B_i) \cdot P(B_i)}$$

Example: breast cancer screening

Assume 1% of all women above 40 have breast cancer.

90% of women with breast cancer will be tested positive. ("true positives")

8% of women without breast cancer will receive a positive result as well ("false positives")

Given that a woman receives a positive test result, what is the likelihood that she has breast cancer?

$$P(\text{cancer} \mid \text{positive}) = \frac{P(\text{positive} \mid \text{cancer}) \cdot P(\text{cancer})}{P(\text{pos.} \mid \text{cancer})\, P(\text{cancer}) + P(\text{pos} \mid \text{not cancer}) \cdot P(\text{not cancer})}$$

$$= \frac{0.9 \cdot 0.01}{0.9 \cdot 0.01 + 0.08 \cdot 0.99} \approx 10\,\%$$

Independence

Two independent events

<u>Def</u> Consider a probability space $(\Omega, \mathcal{A}, P)$. Two events $A, B \in \mathcal{A}$ are called independent if

$$P(A \cap B) = P(A) \cdot P(B)$$

Notation: $A \perp\!\!\!\perp B$

Observation: A is independent of $B \iff P(A \mid B) = P(A)$

Many independent events

<u>Def</u> A family of events $(A_i)_{i \in I}$ is called independent if for all finite subsets $J \subset I$ we have

$$P\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} P(A_i).$$

(Family is called pairwise independent if $\forall i,j \in I$:
$P(A_i \cap A_j) = P(A_i) \cdot P(A_j)$. This does not imply independence!)

Independent random variables

<u>Def</u>:

Two random variables $X : \Omega \to \Omega_1$, $Y : \Omega \to \Omega_2$ are called independent if their induced σ-algebras $\sigma(X)$, $\sigma(Y)$ are independent:

$\forall A \in \sigma(X), \ B \in \sigma(Y): \quad P(A \cap B) = P(A) \cdot P(B)$.

Notation: $X \perp\!\!\!\perp Y$

Expectation of independent product

<u>Prop</u>: Let $X, Y : (\Omega, \mathcal{A}, P) \to \mathbb{R}$ be two random variables. Then:

$$X, Y \text{ independent} \Rightarrow E(X \cdot Y) = E(X) \cdot E(Y)$$

<u>Proof sketch</u>: (discrete case)

$$\sum_{ij} x_i y_j \quad P(X = x_i, Y = y_j) \overset{\text{ind.}}{=}$$

$$= \sum_{ij} \underbrace{x_i y_j}_{x_i \cdot y_j} P(X = x_i) \cdot P(Y = y_j)$$

$$= \underbrace{\left(\sum_i x_i P(X = x_i) \right)}_{< \infty} \underbrace{\left(\sum_j y_j P(Y = y_j) \right)}_{< \infty}$$

Variance and covariance

Variance & standard deviation

<u>Def</u> $X, Y : (\Omega, \mathcal{A}, P) \to \mathbb{R}$ rvs with $E(X^2) < \infty$, $E(Y^2) < \infty$.

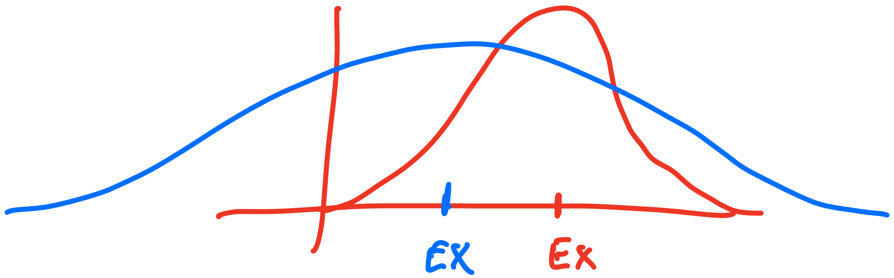
Then

$$\mathrm{Var}(X) := E\big((X - E(X))^2\big)$$

is called the variance of X

and $\sqrt{\mathrm{Var}(X)} =: \sigma_X$

is called the standard deviation.



high variance

moderate variance

Covariance & Correlation

Def:

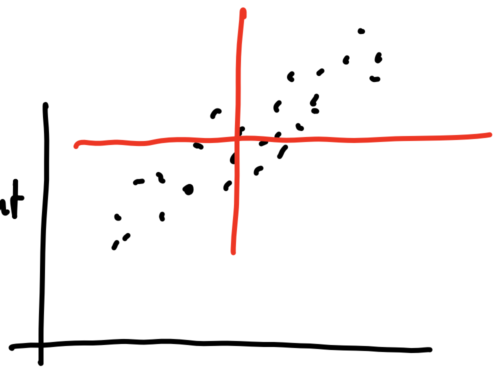
$Cov(X,Y) := E((X-E(X))\cdot(Y-E(Y)))$ is called the covariance of X and Y.

$$\rho_{XY} := \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} \in [-1,1]$$ is called the correlation coefficient.

If $Cov(X,Y)=0$, then X and Y are called uncorrelated.

Intuition about covariance

$$\mathrm{Cov}(X, Y) = E\Big((X - E(X)) \cdot (Y - E(Y)) \Big)$$



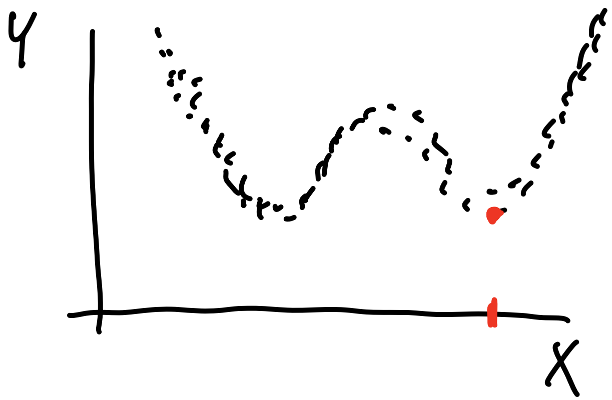
positive, large covariance

$\rho \approx 0.9$

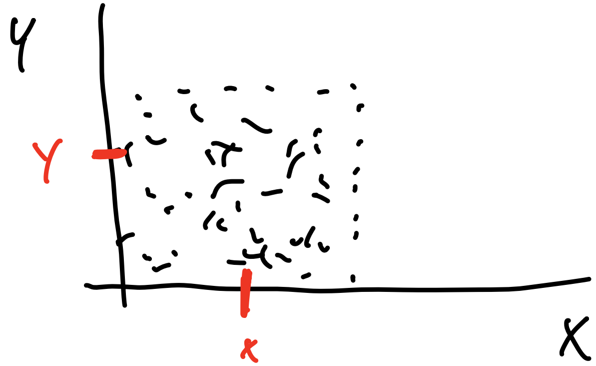
negative cov, large in absolute values

$\rho \approx -0.9$

Intuition about cov



Cov $\approx$ 0
(uncorrelated).
not independent



independence

uncorrelated $\not\Rightarrow$ independence!

$\Leftarrow$

Properties of var and cov

- $Var(X) = E(X^2) - (E(X))^2$
- $Cov(X,Y) = E(X \cdot Y) - E(X) \cdot E(Y)$
- $E(aX + b) = a \cdot E(X) + b$
- $Var(a \cdot X + b) = a^2 Var(X)$
- $Cov(X,Y) = Cov(Y,X)$
- $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$
- X, Y independent $\Rightarrow$ $Cov(X,Y) = 0$

  $\not\Leftarrow$
- X, Y independent $\Rightarrow$ $Var(X+Y) = Var(X) + Var(Y)$.

k-th moment

$$L^k(\Omega, \mathcal{A}, P) := \{ X : \Omega \to \mathbb{R} \mid X \text{ measurable and } \int_\Omega |X^k| \, dP < \infty \}$$

if $X^k \in L^1(\Omega, \mathcal{A}, P)$ then

$E(X^k) = \int X^k \, dP$ is called the k-th moment of X. and

$E((X - E(X))^k)$ the kth centered moment.

Markov and Chebyshev inequality

Cauchy-Schwartz inequality

Proposition: $X, Y \in L^2(\Omega, \mathcal{A}, P)$. Then:

$$E(X \cdot Y)^2 \leq E(X^2) \cdot E(Y^2)$$

Markov inequality

<u>Prop</u>: $\varepsilon > 0$, $f : [0,\infty[\to [0,\infty[$, f monotonically increasing. then

$$P(|Y| > \varepsilon) \le \frac{E(f(|Y|))}{f(\varepsilon)}$$

In particular,

$$P(|Y| > \varepsilon) \le \frac{E(|Y|)}{\varepsilon}$$

Proof

TO Do add on slides

Chebyshev inequality

<u>Proposition</u>: $\varepsilon > 0$, $X \in L^2(\Omega, \mathcal{A}, P)$. Then:

$$\underbrace{P(|X - E(X)| > \varepsilon)}_{\text{key quantity in learning theory!}} \leq \frac{Var(X)}{\varepsilon^2}$$

Proof

To Do: add on slides

SKIPPED

Different types of probability measures

Discrete measure

$\Omega = \{x_1, x_2, \dots\}$ finite or at most countable.

$\mathcal{A} = \mathcal{P}(\Omega)$

We define a probability measure $P: \mathcal{A} \to [0,1]$ by assigning probabilities to the "elementary events":

$$P(\{x_i\}) =: p_i$$

with $\quad 0 \leq p_i \leq 1, \quad \sum_i p_i = 1.$

For $A \in \mathcal{A}$ we assign

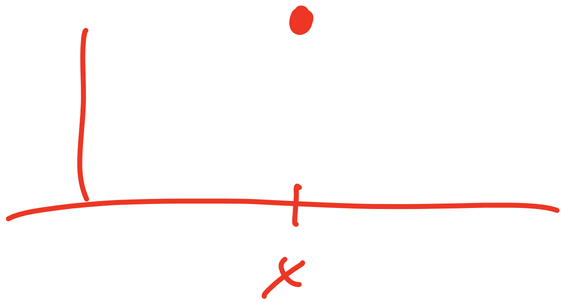
$$P(A) = \sum_{\{i \mid x_i \in A\}} p_i \; .$$

Examples: toss a coin; distribution on $\mathbb{Q}$

Dirac measure

For $x \in \mathbb{R}$, we define the Dirac measure δ_x on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ by setting

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$



Sometimes this is called a point mass at a point x.

A discrete measure on $\mathbb{R}$ can be written as a sum of Dirac measures. For example, throwing a die can be described as

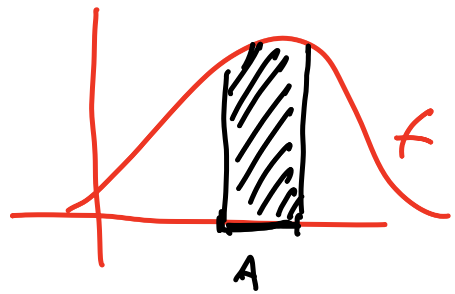
$$\frac{1}{6}\left(\delta_1 + \delta_2 + \dots + \delta_6\right)$$

Measures with a density

Consider $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ and the Lebesgue measure λ.

Consider a function $f: \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ that is measurable

and satisfies $\int f \, d\lambda = 1$. $\left(= \int f(x) \, dx \right)$



Then we define a measure ν on $\mathbb{R}^n$

by setting, for all $A \in \mathcal{A}$,

$$\nu(A) := \int_A f(x) \, dx .$$

ν is the probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ with density f.

Notation: $\nu = f \cdot \lambda$

Question? Can we describe every prob measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$

in terms of a density? Answer: no!

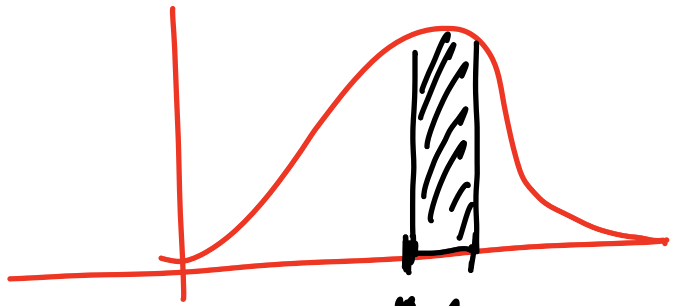
Counterexample: δ_0 Dirac measure

absolutely continuous measures

Def A prob. measure ν on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is called absolutely continuous with respect to another measure μ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ if every μ-null set is also a ν-null set:

$$\forall B \in \mathcal{B}(\mathbb{R}^n): \quad \mu(B) = 0 \Rightarrow \nu(B) = 0.$$

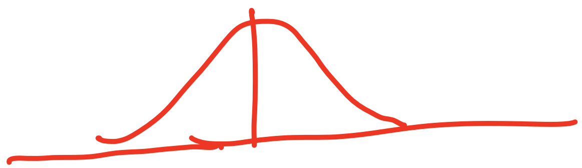
Notation: $\nu \ll \mu$



$$\mu(A) = 0 \Rightarrow \underbrace{\int_A f \, d\mu}_{\nu(A)} = 0$$

Examples

- Example: $\mathcal{N}(0,1) \ll \lambda$



- Example: $\delta_0 \not\ll \lambda$ because

$$\lambda(\{0\}) = 0 \quad \text{but} \quad \delta_0(\{0\}) = 1.$$

Theorem of Radon-Nikodym

Theorem (Radon-Nikodym)

Consider two proto measures ν, μ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Then the following two statements are equivalent:

(see next page)

(1) ν has a density wrt μ.

(2) ν is absolutely continuous wrt μ.

Proof idea

(1) $\Rightarrow$ (2) easy

(2) $\Rightarrow$ (1) We need to construct a density!

Consider the set $\mathcal{G}$ of all functions g with the following properties:

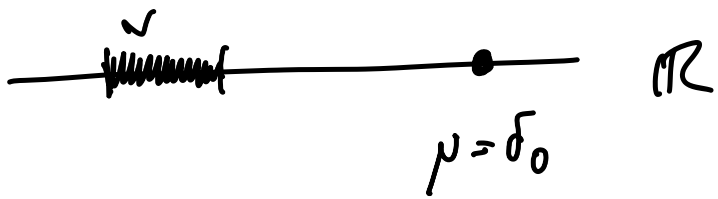
$$(*) \begin{cases} \bullet \ g \text{ is measurable}, \ g \geq 0 \\ \bullet \ g \cdot \mu \leq \nu, \text{ that is} \\ \qquad \forall A \in \mathcal{B}(\mathbb{R}^n): \int_A g \, d\mu \leq \nu(A). \end{cases}$$

- Observe: $g \equiv 0$ satisfies $(*)$, so $\mathcal{G}$ is not empty.
- If g, h both satisfy $(*)$, then $\sup(g, h)$ satisfies $(*)$.
- Define $\gamma := \sup_{g \in \mathcal{G}} \int g \, d\mu$ and construct a

sequence $(g_n)_{n \in \mathbb{N}}$ such that $\lim \int g_n \, d\mu = \gamma$.

- Define "density" $f := \sup g_n$
- Now prove: f does the job. $\blacksquare$

Def μ, ν measures on $(\Omega, \mathcal{A})$. ν is called singular wrt μ if there exists $A \in \mathcal{A}$ such that $\mu(A) = 0$ but $\nu(A^c) = 0$. Notation: $\mu \perp \nu$.



Example: $\lambda \perp \delta_0$

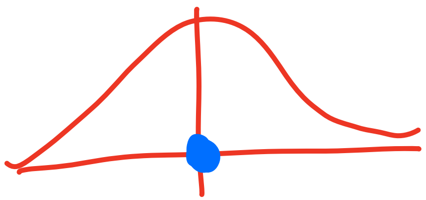
Lebesgue decomposition

Theorem (Decomposition by Lebesgue)

μ, ν prob. measures on $(\Omega, \mathcal{A})$. Then there exists a unique decomposition $\nu = \nu_1 + \nu_2$ such that

$\nu_1 \ll \mu$ and $\nu_2 \perp \mu$.

Example: $\nu = \frac{1}{2}\left(N(0,1) + \delta_0\right)$



$\nu = \nu_1 + \nu_2$ where $\nu_1 = \frac{1}{2} N(0,1)$, $\nu_2 = \frac{1}{2}\delta_0$.

Proof

Proof Let $\mathcal{N}_\mu$ be the set of all null-sets wrt μ. $\subset \mathcal{A}$.

$\alpha := \sup\{\nu(A) \mid A \in \mathcal{N}_\mu\}$

Can construct a countable sequence $(A_n)_{n\in\mathbb{N}}$, $A_n \in \mathcal{N}_\mu$,

such that $\nu(A_n) \nearrow \alpha$. By countable additivity

we get $\nu\Big(\underbrace{\bigcup_{n\in\mathbb{N}} A_n}_{=:N}\Big) = \alpha$.

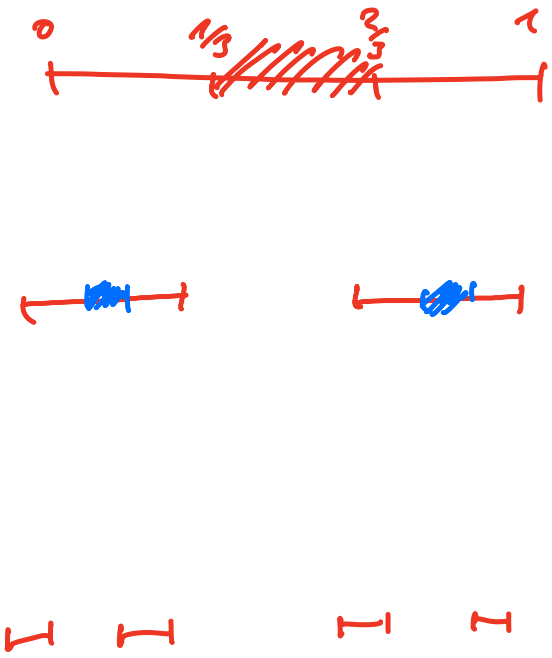
Define $\nu_1 : A \mapsto \nu(A \cap N^c)$

$\nu_2 : A \mapsto \nu(A \cap N)$

Does the job. $\blacksquare$

Cantor - distribution: non-trivial distribution that is singular wrt λ

① Construct the Cantor set:



- Start with $C_0 := [0,1]$

  "Remove middle part"

- $C_1 := [0, 1/3] \cup [2/3, 1]$.

  "Remove middle parts from all intervals"

- $C_2 =$

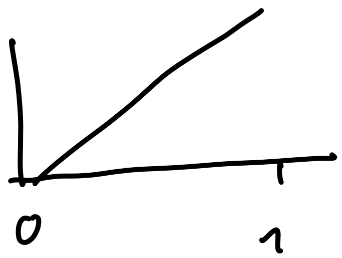
  ⋮

The Cantor set is the limit in this process. It is compact, non-empty, empty interior.

② Now construct a **probability distribution:**

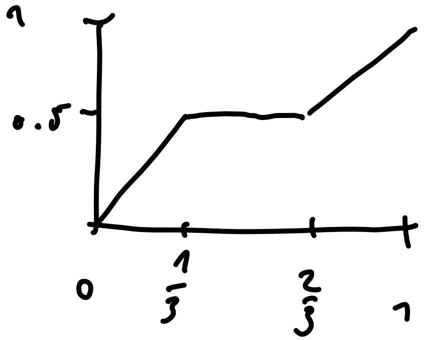
Consider the cdfs of the sets $C_0, C_1, C_2 \ldots$

C_0 :



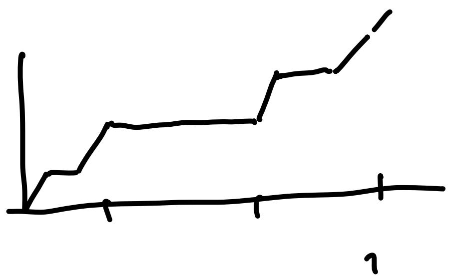
uniform on $[0, 1]$

C_1 :



uniform on $[0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$

C_2 :



... take limit and call the resulting measure μ.

③ Properties:

- The cdf of "μ" is continuous.
- μ is a prob. measure.
- $\lambda(C) = 0$.

Lebesgue measure

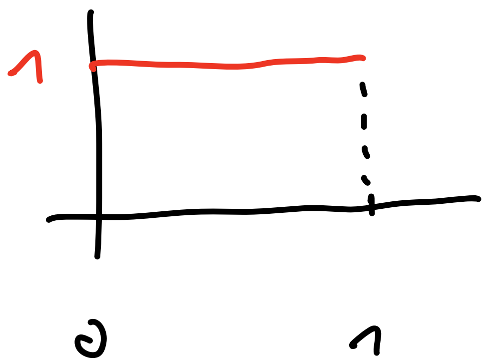
$\Rightarrow \lambda \perp \mu$

Important examples of probability distributions

Uniform distribution

Uniform distr. on $\{1, \dots, n\}$: $P(\{i\}) = \frac{1}{n}$

Uniform distribution on $[0,1]$: distribution with constant density:



Uniform distr. on $[0,1]^d$ $\rightsquigarrow$ independence!

Binomial distribution

- <u>Binomial distribution</u> on $\{0, \dots, n\}$

  Toss a coin n times, independently, each time with probability p of observing head. Denote head $= 1$, tail 0,

  $X :=$ # heads

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

Poisson distribution

Poisson distribution on $\mathbb{N}$:

Parameter $\lambda > 0$

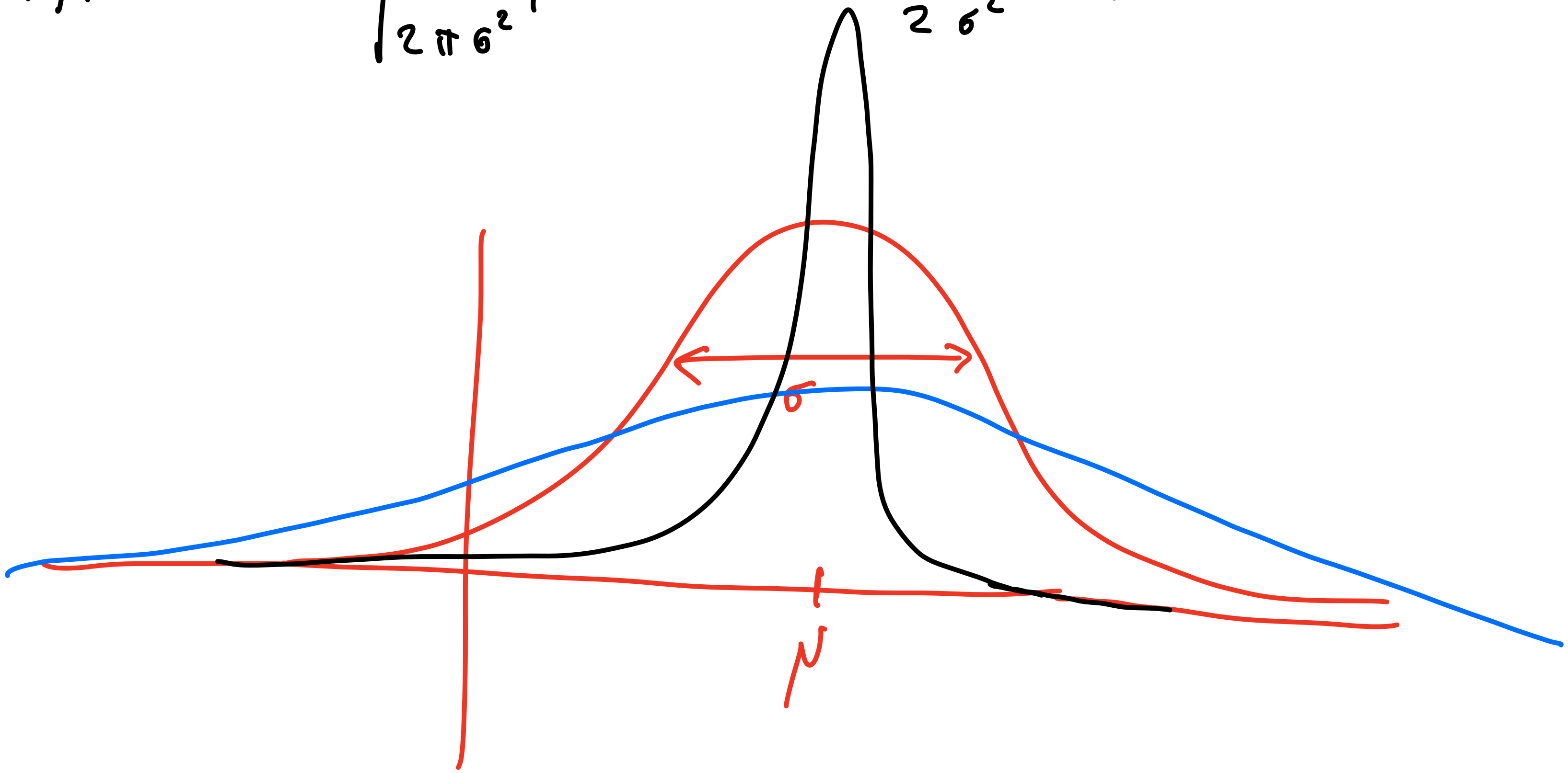
$$P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Intuition: number of incoming calls at a hotline.

Normal distribution on $\mathbb{R}$

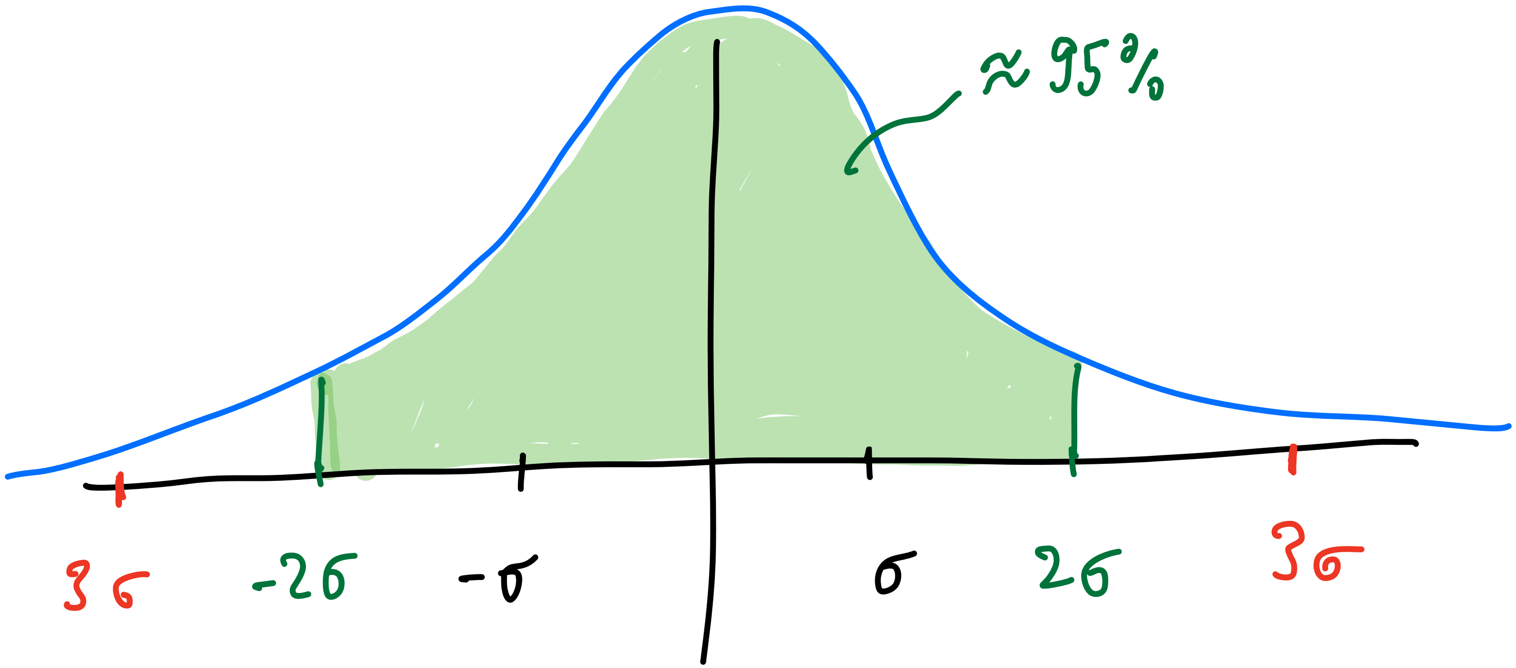
Density: parameter μ (mean), σ (std. deviation)

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



Notation: $\mathcal{N}(\mu, \sigma^2)$

A rule of thumb



The area under the normal densities:

$[-\sigma, \sigma] \rightsquigarrow \approx 70\%$

$[-2\sigma, 2\sigma] \rightsquigarrow \approx 95\%$

$[-3\sigma, 3\sigma] \rightsquigarrow \approx 99\%$

Sum of independent normals is normal

<u>Prop</u>: $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, $X \perp Y$.

Then $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

"<u>Proof</u>": It is elementary to see that the mean of $X+Y$ is $\mu_1 + \mu_2$ and the variance is $\sigma_1^2 + \sigma^2$.

However, it is not trivial to prove that the resulting distribution is again normal, the standard approach requires convolutions and characteristic fcts. Skipped.

Multivariate normal distribution

$X : \Omega \to \mathbb{R}^n$, $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$, $\mu_i \in E(X_i)$, $\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}$

$\Sigma \in \mathbb{R}^{n \times n}$ with $\Sigma_{ij} = Cov(X_i, X_j)$, called covariance matrix.

$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{n/2} (\det \Sigma)^{1/2}} \exp\left(- \frac{1}{2} (x-\mu)^t \Sigma^{-1} (x-\mu)\right)$$

Notation: $N(\mu, \Sigma)$

The covariance matrix is psd

Prop Let C be the covariance matrix of any set of random variables, i.e. $\sum_{i,j=1}^{n} = \mathrm{Cov}(X_i, X_j)$. Then C is symmetric and positive semi-definite.

Proof Symmetry is clear because $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$.

psd need to prove that $\forall a_1, \dots, a_n \in \mathbb{R}$,

$$a^t C a = \sum_{i,j=1}^{n} a_i a_j C_{ij} \geq 0 :$$

$$a^t C a = \sum_{i,j=1}^{n} a_i a_j C_{ij} \overset{\text{def}}{=} \sum_{i,j=1}^{n} a_i a_j E\big((X_i-\mu_i)(X_j-\mu_j)\big)$$

linearity of expectation

$$= E\left(\sum_{i,j=1}^{n} a_i a_j (X_i-\mu_i)(X_j-\mu_j)\right)$$

$$= E\left(\underbrace{\left(\sum_{i=1}^{n} a_i (X_i-\mu_i)\right)^2}_{\geq 0}\right) \geq 0$$

$\blacksquare$

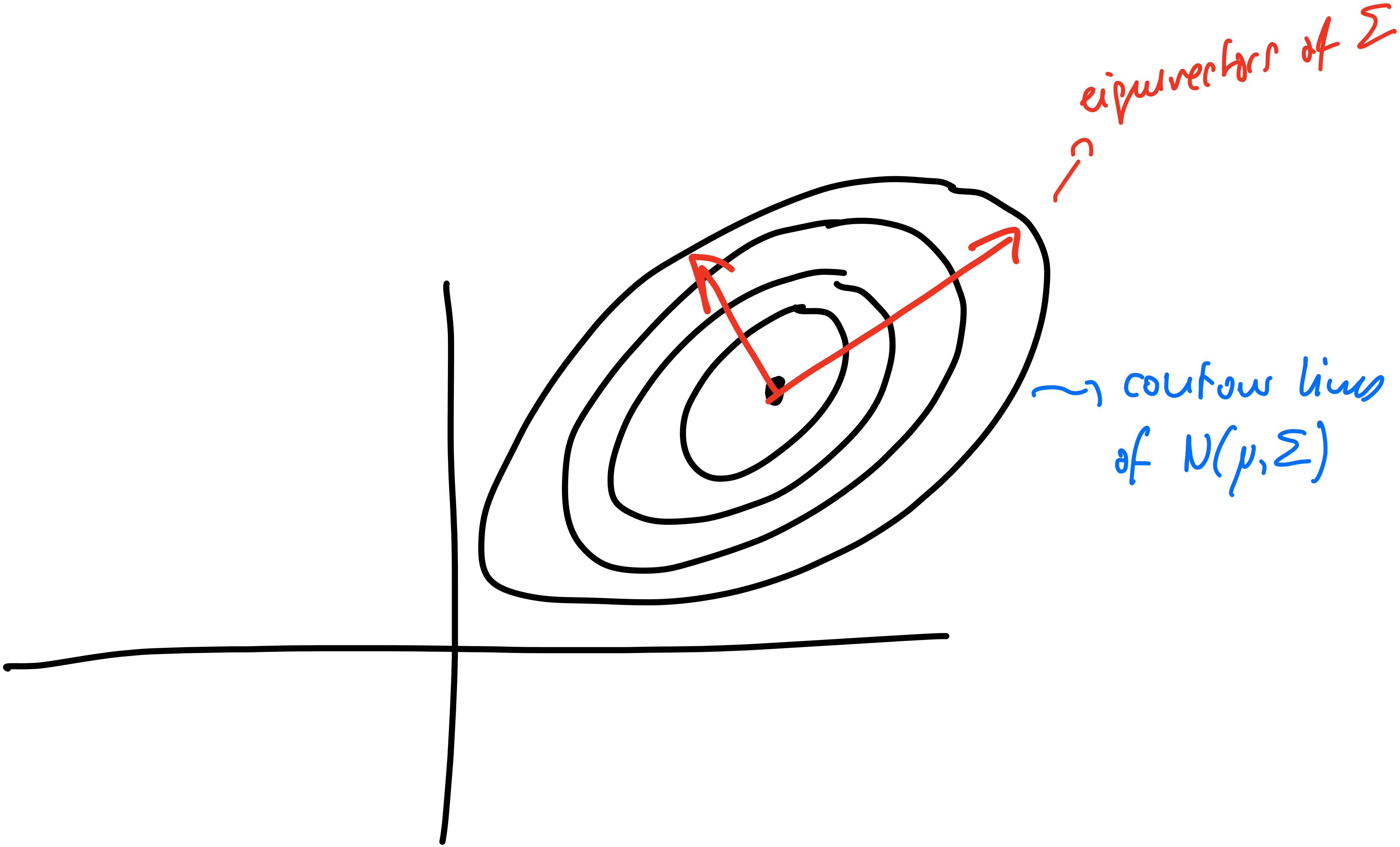
PCA & Contour lines

Because Σ is symmetric and psd, it has real-valued eigenvalues ≥ 0.

The contour lines of a multivariate normal are ellipses:

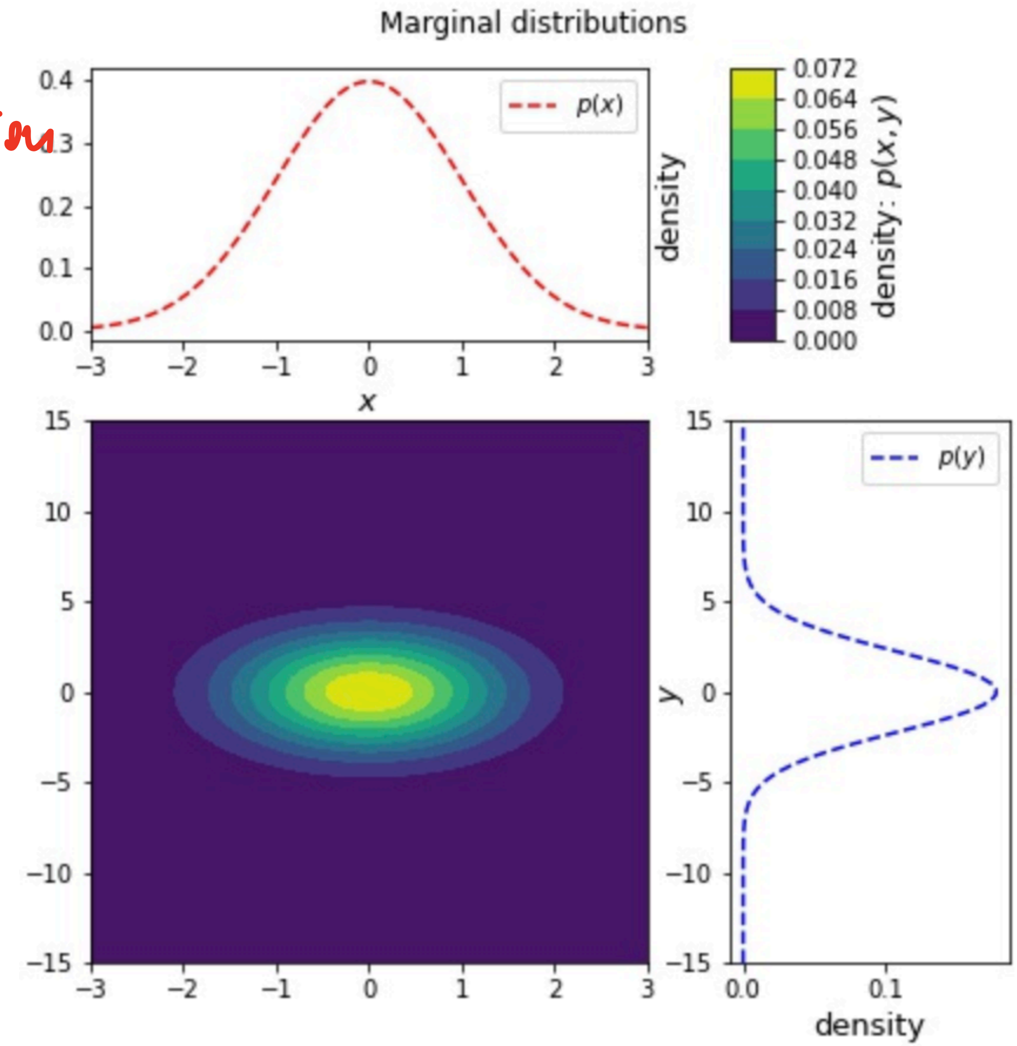
Contour line : set $\{ x \mid f_{\mu,\sigma}(x) = c \}$

$$f_{\mu,\sigma}(x) = c \iff \underbrace{(x-\mu)^t \Sigma^{-1} (x-\mu) = \tilde{c}}_{\text{ellipse equation}}$$


eigenvectors of Σ
contour lines of $N(\mu, \Sigma)$

Marginal distributions of Gaussians are Gaussians

Marginal distribution of x

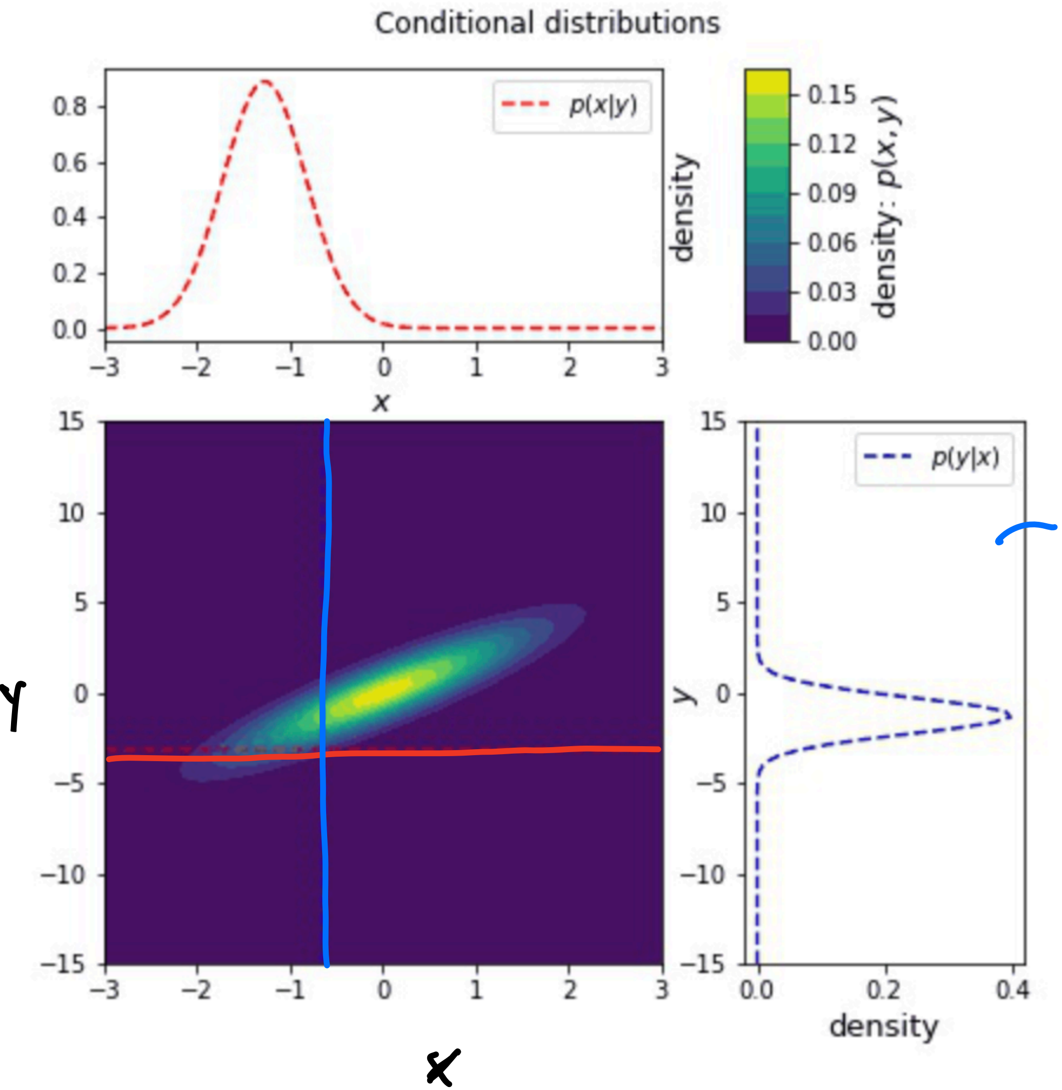


Marginal distribution of Y

Figure credit: Nils Lehmann

Conditional distributions of Gaussians are Gaussians

conditional
$p(x \mid y = -4)$



conditional
$p(y \mid x = -0.7)$

Figure credit: Nils Lehmann

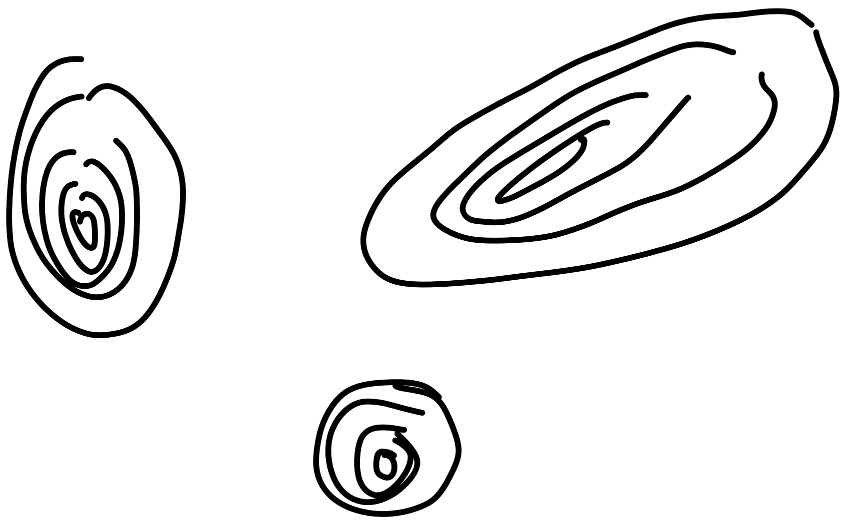
Mixture of Gaussians

Consider "mixing weights" $a_1, \dots, a_k \in [0,1]$ s.t. $\sum_{i=1}^{k} a_i = 1$.

Fix $\mu_1, \dots, \mu_k$ and $\Sigma_1, \dots, \Sigma_k$. Define the new density

$$f(x) = \sum_{i=1}^{k} a_i \, f_{\mu_i, \Sigma_i}(x)$$

This is called a mixture of Gaussians.



Examples of distributions with "infinite" expectation

Cauchy distribution on $\mathbb{R}$ with density $f(x) = \left(\pi r \left(1 + \left(\frac{x - x_0}{r}\right)^2\right)\right)^{-1}$

Power law distributions on $\mathbb{R}$: a family of distributions that satisfies

$$P(X > x) = c \cdot x^{-\alpha} \leftarrow \text{parameters}$$

If $\alpha \leq 3$, no variance exists.

If $\alpha \leq 2$, no mean exists.

ML keyword: preferential attachment model for social networks

Heavy-tailed vs sub-Gaussian distributions

"Tail behavior" of a Gaussian:

$$P(|X| > t) = 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$$

Heavy-tailed distributions are the ones where $P(|X|>t)$ is larger than for a gaussian.

Sub-Gaussian ... smaller

Sub-Gaussian rvs are particularly popular in learning theory because they exhibit strong concentration (as see later).

Convergence of random variables

Almost sure convergence

Def
Consider rv $X_i : \Omega \to \mathbb{R}$, $i \in \mathbb{N}$, $X : \Omega \to \mathbb{R}$,
$(\Omega, \mathcal{A}, P)$ a probability space.

$(X_i)_{i \in \mathbb{N}}$ converges to X almost surely $:\Leftrightarrow$

$$P\left(\left\{ \omega \in \Omega \mid \lim_{i \to \infty} X_i(\omega) = X(\omega) \right\} \right) = 1$$

Notation: $X_i \to X$ a.s.

⚠ Let us check that this definition is well-defined: We need to prove that the event $\left\{ \omega \in \Omega \mid \lim_{i \to \infty} X_i(\omega) = X(\omega) \right\}$ is an element in $\mathcal{A}$:

Well-defined

Prop $\{ \omega \in \Omega \mid \lim_{i\to\infty} X_i(\omega) = X(\omega) \} \in \mathcal{A}$

$\varepsilon = \frac{1}{k}$

Proof sketch $\lim X_i(\omega) = X(\omega)$

$\Leftrightarrow \forall k \in \mathbb{N}\ \exists N \in \mathbb{N}\ \forall n > N: \ |X_n(\omega) - X(\omega)| < \frac{1}{k}$

So we get:

$$\{\omega \mid X_i(\omega) \to X(\omega)\} =$$

$$= \underbrace{\bigcap_{k\in\mathbb{N}} \bigcup_{N\in\mathbb{N}} \bigcap_{n\geq N}}_{\text{countable unions and intersections}} \underbrace{\left\{\omega \,\middle|\, |X_n(\omega) - X(\omega)| < \frac{1}{k}\right\}}_{\substack{X_n,\, X \text{ are measurable} \Rightarrow \\ |X_n - X| \text{ is measurable} \\ \text{so } \{\dots\} \in \mathcal{A}}} \in \mathcal{A}$$

$\square$

Convergence in probability

Def $(X_i)_{i \in \mathbb{N}}$ converges to X in probability $:\Longleftrightarrow$

$$\forall \varepsilon > 0 \quad P\left(\{\omega \in \Omega \mid |X_i(\omega) - X(\omega)| > \varepsilon\}\right) \longrightarrow 0$$

Stronger or weaker than "almost surely"?

Convergence in L_p

<u>Def</u> $X_n \to X$ in L^p ("in the p-th mean") $:\Leftrightarrow$

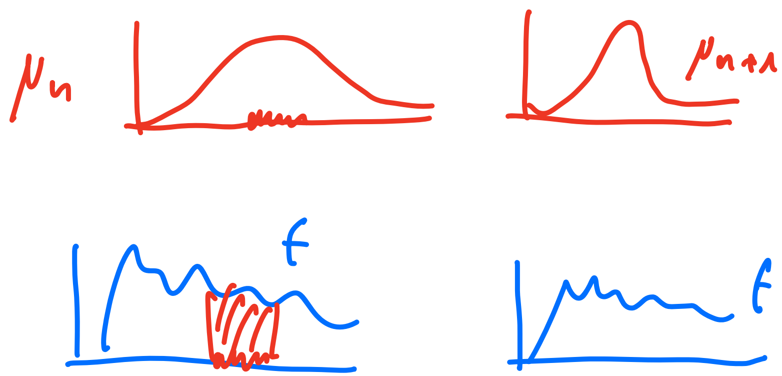
$X_n, X \in L^p$ and $\|X_i - X\|_p \to 0$.

Weak convergence

Def Let $M^1(\mathbb{R}^n)$ be the set of all probability measures on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Assume $(\mu_n)_n \subset M^1(\mathbb{R}^n)$, $\mu \in M^1(\mathbb{R}^n)$.
$C_b(\mathbb{R}^n) :=$ space of bounded continuous functions.

$\mu_n \to \mu$ weakly $:\Leftrightarrow$

$$\forall f \in C_b(\mathbb{R}^n): \int f \, d\mu_n \longrightarrow \int f \, d\mu$$



⚠ weak convergence is defined for measures, not for rvs.

(Excursion: weak convergence in functional analysis)

Excursion:

In functional analysis, a sequence $(x_n)_n$ in a Banach space B converges weakly if for all bounded lin. functionals f, we have that $f(x_n) \to f(x)$. (i.e. for all $f \in B'$).

Space $M^1(\mathbb{R}^n)$ itself is not a Banach space, but $\subset M(\mathbb{R}^n)$, space of all bounded measures.

The dual space of $M(\mathbb{R}^n)$ is $C_b(\mathbb{R}^n)$.

Thus the weak convergence defined above coincides with the notion of weak conv. on $M(\mathbb{R}^n)$.

Convergence in distribution

Def $X_i, X : (\Omega, \mathcal{A}, P) \to \mathbb{R}^n$. The sequence X_n converges in distribution to X :$\Leftrightarrow$ the distributions P_{X_n} converge to P_X weakly.

Relationship between notions of convergence

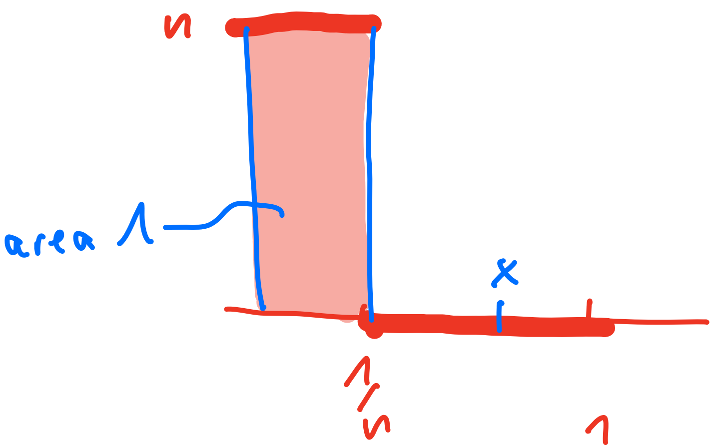
We have the following implications, and none of the "missing directions" holds in general:

$$\begin{tikzcd}
\text{almost surely} \arrow[dr, Rightarrow] & & \text{in } L^1 \arrow[dl, Rightarrow] & \text{in } L^p \ (p>1) \arrow[l, Rightarrow] \\
& \text{in probability} \arrow[d, Rightarrow] & & \\
& \text{in distribution} & &
\end{tikzcd}$$

Example (convergence a.s., in prob., but not in L^1)

Consider $[0,1]$ with the uniform distribution.

Define $X_n : \mathbb{R} \to \mathbb{R}$, $X_n(\omega) = \begin{cases} n & \text{for } 0 \le x \le \frac{1}{n} \\ 0 & \text{otherwise} \end{cases}$



$\forall x > 0: \ X_n(x) \to 0.$

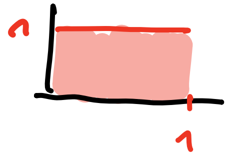
Can formally see: converges a.s., in prob.

But: no convergence in L^1.

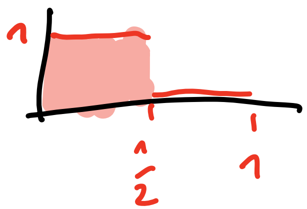
Example (convergence in prob., in L^1, but not a.s.)

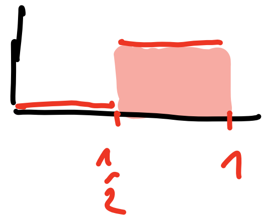
"sliding blocks of smaller and smaller volume":

$f_1 = \mathbb{1}_{[0,1]}$

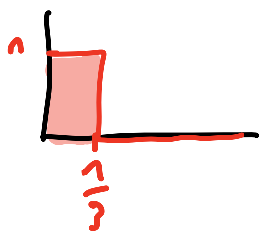


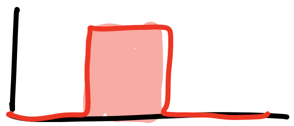
$f_2 = \mathbb{1}_{[0,1/2]}, \quad f_3 = \mathbb{1}_{[1/2,1]}$

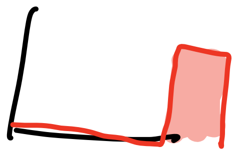




$f_4 = \mathbb{1}_{[0,1/3]}, \quad f_5 = \mathbb{1}_{[1/3,\frac{2}{3}]}, \quad f_6 = \mathbb{1}_{[2/3,1]}$

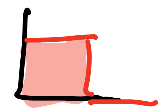




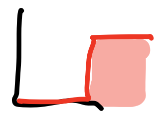


Example (Conv. in distribution, but not in prob.)

- $X_n : [0, 1] \to \mathbb{R}$, $\forall i: \ X_i(\omega) = \mathbb{1}_{[0,\frac{1}{2}[} = \begin{cases} 1 & 0 \leq \omega \leq \frac{1}{2} \\ 0 & \frac{1}{2} < \omega \leq 1 \end{cases}$



- $X = \mathbb{1}_{[\frac{1}{2}, 1]}$



Obviously $X_n \not\to X$ in prob. , but:

$$P_{X_1} = \frac{1}{2}\left(\delta_0 + \delta_1\right) = P_{X_2} = P_{X_3} = \dots = P_X$$

so $X_n \to X$ in distribution.

Theorem of Borel-Cantelli

Definition "A_n infinitely often"

$(\Omega, \mathcal{A}, P)$ prob. space, $(A_n)_n$ sequence of events in $\mathcal{A}$.

$$P(A_n \text{ infinitely often}) := P(A_n \text{ i.o.})$$

$$= P(\{\omega \in \Omega \mid \omega \in A_n \text{ for infinitely many } n\})$$

Almost sure convergence $\Leftrightarrow$ "$> \varepsilon$ infinitely often"

Proposition : X_n, X r.v. on $(\Omega, \mathcal{A}, P)$.

$$X_n \to X \text{ a.s.} \iff$$

$$\forall \varepsilon > 0 : P\Big(\{ |X_n - X| > \varepsilon \} \text{ inf. often} \Big) = 0$$

Proof intuition

$$\{ \lim x_n = x \}$$

$$= \{ \forall k : |x_n - x| > \tfrac{1}{k} \text{ at most finitely often} \}$$

$$= \bigcap_{k \in \mathbb{N}} \{ |x_n - x| > \tfrac{1}{k} \text{ at most fin. often} \}$$

$$\left(\bigcup_{k \in \mathbb{N}} \{ |x_n - x| > \tfrac{1}{k} \text{ inf. often} \} \right)^{\text{complement}}$$

Borel - Cantelli

Theorem: Consider a sequence of events $(A_n)_n \subset \mathcal{A}$.

(1) If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(A_n \text{ i.o.}) = 0$.

(2) If $\sum_{n=1}^{\infty} P(A_n) = \infty$, and if $(A_n)_n$ are independent,

then $P(A_n \text{ i.o.}) = 1$.

Proof

To Do Prof. ess.

Application in learning theory

Assume that $P(|X_n - X| > \frac{1}{n}) < \delta_n$, and

assume that $\sum_{n=1}^{\infty} \delta_n < \infty$. Then you can use

Borel - Cantelli to prove that

$$P(|X_n - X| > \tfrac{1}{n} \text{ i.o.}) = 0,$$

thus $X_n \to X$ a.s.

The law of large numbers
(LLN)

iid

... is an abbreviation for

"independent and identically distributed"

A sequence $(X_n)_{n\in\mathbb{N}}$ of rvs is iid if they are independent and if they all follow the same distribution.

Being iid is one of the standard assumptions in learning theory.

Strong vs weak

"Strong law" $\Leftrightarrow$ convergence a.s.

"Weak law" $\Leftrightarrow$ convergence in probability

Weak law of large numbers

Theorem: Let $(X_i)_{i \in \mathbb{N}}$ iid random variables with $\mathrm{Var}(X_i) =: c < \infty$ and $E(X_i) =: \mu$. Denote $S_n := \frac{1}{n} \sum_{i=1}^{n} X_i$. Then:

$$P(|S_n - \mu| > \varepsilon) \leq \frac{c}{n\varepsilon^2}$$

In particular, $\frac{1}{n} \sum_{i=1}^{n} X_i - \bar{X}$ converges to 0 in probability.

(Observe: iid $\Rightarrow$ all X_i have the same mean and var).

Proof

Wlog assume $\mu = 0$ (otherwise consider the centered rvs $X_i - \mu$).

The weak law then follows directly from the Chebyshev inequality:

$$P(|S_n - \mu| > \varepsilon) \overset{\mu=0}{=} P(|S_n| > \varepsilon) \leq \frac{E(S_n^2)}{\varepsilon^2}$$

Now exploit that $E(S_n^2) = E((\frac{1}{n}\sum X_i)^2) = \frac{1}{n^2} E(\sum_{ij} X_i X_j) \overset{\text{ind.}}{=}$

$$= \frac{1}{n^2}\sum_i \underbrace{E(X_i^2)}_{= Var(X_i)} + \frac{1}{n^2}\sum_{i \neq j} \underbrace{E(X_i)E(X_j)}_{=0} = \frac{1}{n^2} \cdot n \cdot \underbrace{Var(X_i)}_{= c}$$

Plugging this in above immediately gives the desired result. ∎

Strong law of large numbers

<u>Theorem:</u> Let $(X_i)_{i \in \mathbb{N}}$ be iid rvs. Assume that the mean $\mu := E(X_1) < \infty$, and $Var(X_1) =: \sigma^2 < \infty$. Then:

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mu \quad \text{a.s. and in } L^2 .$$

Strong law of large numbers

<u>Theorem:</u> Let $(X_i)_{i \in \mathbb{N}}$ be iid random variables with $\mathrm{Var}(X_i) < c$ and $\mathbb{E}(X_i) = \mu$. Then

$$\frac{1}{n} \sum_{i=1}^{n} X_i \longrightarrow \mu \quad \text{almost surely.}$$

Proof

We prove the theorem under a slightly stronger condition: $E(X_i^4) < \infty$.

Without loss of generality we assume that $\mu = 0$ (otherwise we replace X_i by $X_i - \mu$).

To simplify notation, introduce $S_n := \sum_{i=1}^{n} X_i$.

General idea:

- want to apply Borel-Cantelli to the events of the form

$$\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} - \mu \right| > \delta_n \right\}$$

- Need to find events s.th. $P(\{ \dots > \delta_n \}) \leq \varepsilon_n$

  and $\sum_{n=1}^{\infty} \varepsilon_n < \infty$

- For these individual events we are going to use the Markov inequality.

Proof

Proof Step 1: Under the given assumptions, there exists a constant $K < \infty$ such that $E(S_n^4) \leq K n^2$:

Proof:

$$E(S_n^4) = E\left(\left(\sum_{i=1}^{n} X_i\right)^4\right)$$

$= \dots$

- multiply out
- exploit that all X_i have the same distribution
- and that $E(X_i) = 0$

$$= n \cdot E(X_1^4) + 3n(n-1)\, E(Z_1^2 Z_2^2)$$

$$\leq K \cdot n^2 \quad \text{for some constant } K$$

Proof step 2 Markov inequality for function $f(x) = x^4$:

For all $\mu > 0$,

$$P\left(\tfrac{1}{n} |S_n| > n^{-\mu} \right) \le \frac{E\left(\left(\tfrac{1}{n}|S_n|\right)^4\right)}{n^{-4\mu}} \le \frac{K n^2}{n^{-4\mu}} = K \cdot n^{-2+4\mu}$$

Proof step 3: Borel-Cantelli

Fix some $\mu \in \left]0, \tfrac{1}{4}\right[$ and define the events

$$A_n = \left\{ \tfrac{1}{n} |S_n| \ge n^{-\mu} \right\}.$$

Then $\sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} K\, n^{-2+4\mu} < \infty$. $\quad -2+4\mu < -1$

Borel-Cantelli $\Rightarrow$ $P(A_n \text{ i.o.}) = 0$. Thus "high deviation" can only happen at most finitely often $\Rightarrow$ convergence a.s. $\blacksquare$

Many more versions exist

There exist many, many versions of the LLN, under all kinds of assumptions.

Let us try to get some intuition for what is really needed:

The conditions in this theorem

<u>Independence</u>: The theorem does <u>not</u> hold if we allow for arbitrary dependencies.

Example: $X_1 \sim Ber(\frac{1}{2})$ ("fair coin")

$X_2 = X_3 = \dots =: X_1$ (all other coins get the same result as the first one).

Then: X_i are identically distributed with mean $\mu = \frac{1}{2}$

<u>But</u>: $S_n = \begin{cases} 0 & X_1 = 0 \\ 1 & X_1 = 1 \end{cases}$, hence

$$S_n \to \begin{cases} 0 \\ 1 \end{cases} \quad \text{but not to } E(X_1) = \frac{1}{2}$$

One can weaken independence "a bit" (stationary sequences, martingale differences, ...)

The conditions in this theorem

Id. distributed : This condition is not really necessary.

For example, a popular version of the theorem states that the variance of all the X_i needs to be bounded by the same constant: $\forall i \in \mathbb{N}: \operatorname{Var}(X_i) \leq c$

Bounded variance : In case where the X_is are identically distributed, we do not need a bounded variance.

Bounded expectation : is obviously needed, otherwise we would not even be able to state the result....

Machine learning question

Consider a training set $(x_i, y_i)_{i=1\dots n}$ drawn iid $\sim P$.
Consider a loss function ℓ and a classifier f that has been trained on this data.

The training error is defined as

$$R_n(f) = \frac{1}{n} \sum \ell(f(x_i), y_i).$$

The test error is defined as

$$R(f) = E\big(\ell(f(x_i), y_i)\big).$$

Does the LLN now state that $R_n \to R$?!

the central limit theorem

(CLT)

Central Limit Theorem

Theorem:

$(X_i)_{i \in \mathbb{N}}$ iid rv with mean μ, variance $\sigma^2 < \infty$.

Consider the rv $S_n := \sum_{i=1}^{n} X_i$. We normalize it to

$$Y_n := \frac{S_n - n \cdot \mu}{\sqrt{n}\,\sigma}$$

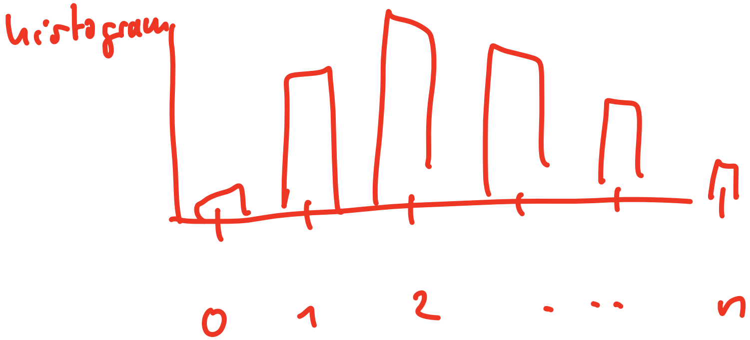
(which has mean 0 and standard dev. 1).

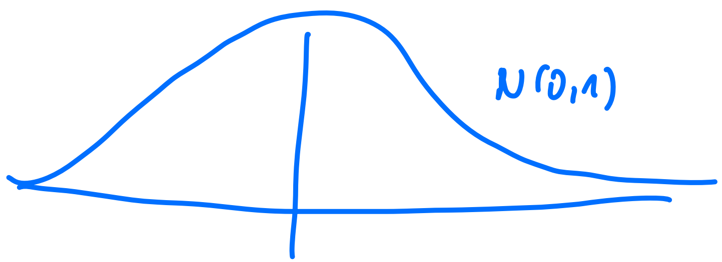
Then $Y_n \to Y$ in distribution where $Y \sim N(0,1)$.

Illustration

Illustration: X_i coin, head $\hat{=} 1$, tail $\hat{=} 0$

$$S_n = \sum X_i \in [0, n]$$





Many more versions

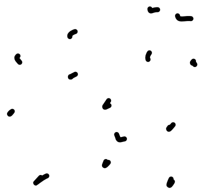
Also here, many more versions ...

- Independence is really important
- We don't need identical distributions at all
- Bounded variance helps but can be weakened. In the end, we need to establish conditions that assert that each individual rv "is negligible in the limit" and does not dominate the resulting sum.

Concentration inequalities

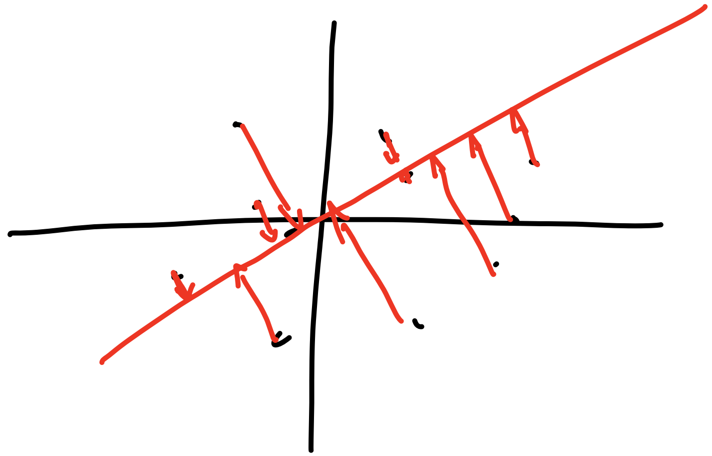
Concentration inequalities

Motivation: random projections



$\mathbb{R}^d$, d large

want to project in $\mathbb{R}^\ell$, ℓ "small"



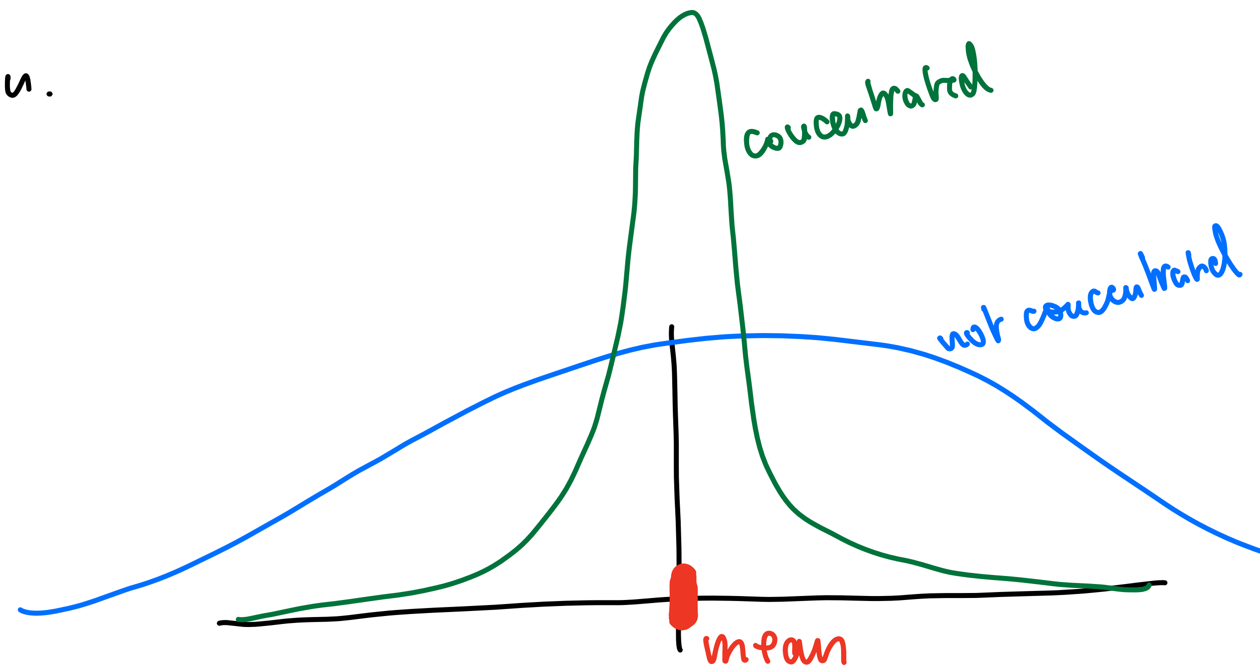
First motivation

In ML, we very often replace individual statistics by their expectation, eg in the training error.

The LLN says that the mean converges to the expected value.

But the speed of convergence can be really slow, and also the

Concentration inequalities tell us that with high probability, an event is very close to its mean.



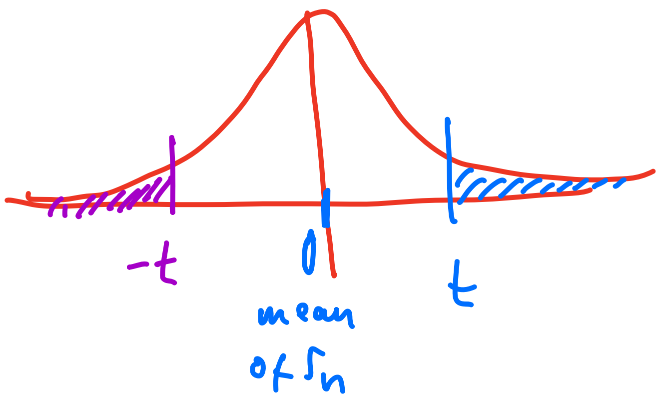
Hoeffding inequality

Theorem (Hoeffding): $X_1, \dots, X_n$ rv $: (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B})$, independent,

assume that $X_i \in [a_i, b_i]$ a.s. for $i = 1, \dots, n$.

Let $S_n := \sum_{i=1}^{n} (X_i - E(X_i))$. Then for any $t > 0$,

$$P(S_n \geq t) \leq \exp\left(- \frac{2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2}\right).$$



Application of Hoeffding: LLN

Consider the following variant of the LLN:

$(X_i)_{i \in \mathbb{N}}$ iid rv, $a \leq X_i \leq b$, let X have the same distribution as the X_i.

Then: $\frac{1}{n} \sum_{i=1}^{n} X_i \to E(X)$ a.s.

Note that we do not make any ass. on $E(X_i^4)$ as in our earlier proof of the LLN. We now use Hoeffding to prove it:

Proof of the LLN with Hoeffding

Step 1: Hoeffding $\Rightarrow$

- $P\left(\frac{1}{n} \sum X_i - E(X) > t \right) \leq \exp\left(- \frac{2nt^2}{(b-a)^2}\right)$

- $P\left(\frac{1}{n} \sum X_i - E(X) < -t \right)$

$$= P\left(\frac{1}{n} \sum (-X_i) - E(-X) > t \right) \leq \exp\left(- \frac{2nt^2}{(b-a)^2}\right)$$

Combined we get

$$P\left(\left| \frac{1}{n} \sum X_i - E(X) \right| > t \right) \leq 2 \exp\left(- \frac{2nt^2}{(b-a)^2}\right).$$

Now want to apply Borel-Cantelli to get a.s. convergence:

$$Z_n := \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$\sum_{n=0}^{\infty} P(Z_n - E(X) > t) \leq 2 \cdot \underbrace{\sum_{n=0}^{\infty} \exp\left(-\frac{2nt^2}{(b-a)^2}\right)}_{=:\ \text{sum}} \overset{!}{\underset{(*)}{\leq}} \infty$$

$(*)$ Substitute: $r := \exp\left(-\frac{2t^2}{(b-a)^2}\right) \in [0,1[$

Observe: $\exp\left(-\frac{2nt^2}{(b-a)^2}\right) = r^n$

$$\text{sum} = 2\sum_{n=0}^{\infty} r^n = 2 \cdot \frac{1}{1-r} < \infty.$$

Now Borel-Cantelli gives almost sure convergence. $\blacksquare$

Hoeffding is tight without further assumptions

Hoeffding is tight (cannot be improved without further assumptions). For fair coin tosses it is tight.

But: not tight if coin is biased $\rightsquigarrow$ need other inequalities

Bernstein inequality

Theorem (Bernstein): $X_1, \dots, X_n$ independent with 0 mean, $|X_i| < 1$ a.s. Let $\sigma^2 := \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(X_i)$. Then for all $t > 0$,

$$P\left(\frac{1}{n}\sum_{i=1}^n X_i > t\right) \leq \exp\left(-\frac{nt^2}{2(\sigma^2 + t/3)}\right)$$

Concentration inequality for functions with bounded differences

Consider a function $f: \mathbb{R}^n \to \mathbb{R}$ (or more generally, $f: \mathcal{X}^n \to \mathbb{R}$ for some "arbitrary" space $\mathcal{X}$).

We say that f has the bounded differences property if there exist constants $c_1, \dots, c_n$ such that

$$(*) \quad \sup_{\substack{x_1, \dots, x_n \in \mathcal{X} \\ \tilde{x}_i \in \mathcal{X}}} \left| f(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_{i-1}, \tilde{x}_i, x_{i+1}, \dots, x_n) \right| \le c_i$$

Example: $f(x_1, \dots, x_n) = \sum_{i=1}^n x_i$, and $a \le x_i \le b \ \forall i$, then f satisfies $(*)$ with $c_i = b - a$.

Theorem of McDiarmid

<u>Theorem</u>: $X_1, \dots, X_n$ independent rv, $X_i \in \mathcal{X}_i$,

$f: \mathcal{X}_1 \times \dots \times \mathcal{X}_n \to \mathbb{R}$ function with bounded difference property.

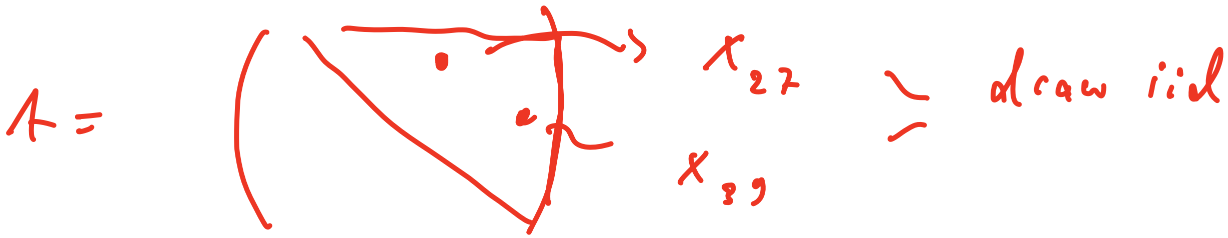
Then, for any $t > 0$,

$$P\left(f(X_1, \dots, X_n) - E(f(X_1, \dots, X_n)) \geq t \right)$$

$$\leq \exp\left(- \frac{2t^2}{\sum_{i=1}^{n} c_i^2} \right)$$

Applications:

- stability in ML
- standard theoretical CS, randomized algorithms, eg Johnson-Lindenstrauss
- largest eigenvalue of a random symmetric matrix



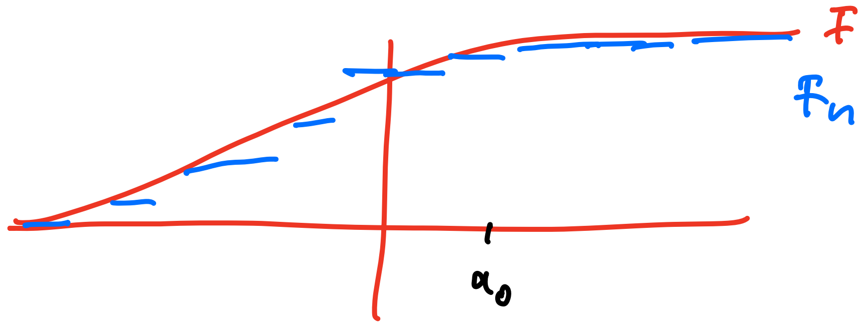
Skipped

Glivenko-Cantelli theorem

Glivenko-Cantelli Theorem

F cdf : $\quad F(a) = P(X \leq a)$

$X_1, \dots, X_n \sim F$, iid



$F_n : \mathbb{R} \to [0,1]$

$F_n(a) := \frac{1}{n} \sum_{i=1}^{n} 1\!\!1_{\{X_i \leq a\}}$

Now fix one particular $a_0 \in \mathbb{R}$.

$F_n(a_0) \longrightarrow F(a_0)$ by the law of large numbers.

Because $1\!\!1_{\{X_i \leq a_0\}}$ is a Binomial rv with

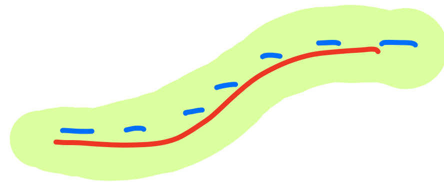
$p = P(X_i \leq a_0)$.

So it is clear that $F_n \xrightarrow{a.s.} F$ pointwise (i.e. $\forall a_0$)

Now let's look at uniform convergence.

<u>Theorem</u> $X_1, \dots, X_n$ iid random variables with cdf F.

Let F_n be the empirical cdf induced by the sample. Then:



$$P\left(\sup_{a \in \mathbb{R}} | F_n(a) - F(a) | > \varepsilon \right) \leq$$

$$\leq 8 \cdot (n+1) \cdot \exp\left(- \frac{n \varepsilon^2}{32}\right) .$$

In particular, $\sup | F_n - F | \to 0$ a.s.,

i.e. $F_n \to F$ uniformly a.s.

__Proof__ Observe: LLN $\Rightarrow$ $P(|F_n(a_0) - F(a_0)| > \varepsilon) \to 0$
for any fixed a_0.

Problem: need to look at

$$P\left(\sup_{a \in \mathbb{R}} |F_n(a) - F(a)| > \varepsilon\right)$$

difficult because $\mathbb{R}$ is uncountable

If we take a supremum over a finite set, it is easier:

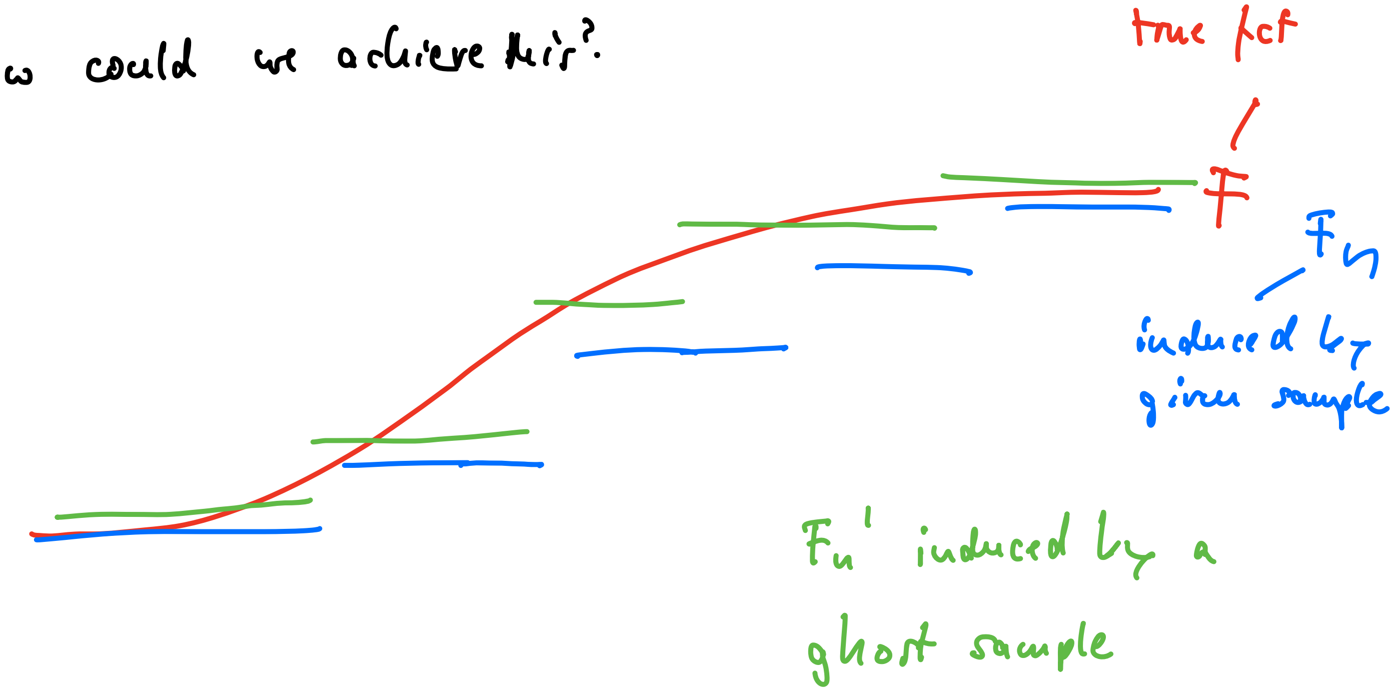
$$P\left(\max_{i=1\ldots n} |U_i| > \varepsilon\right) =$$

$$= P(|U_1| > \varepsilon \text{ or } |U_2| > \varepsilon \text{ or} \ldots \text{ or } |U_n| > \varepsilon)$$

$$\leq \sum_{i=1}^{n} P(|U_i| > \varepsilon)$$

Trick of the proof: convert $\sup\limits_{a \in \mathbb{R}}$ to something "finite".

How could we achieve this?



|red − green|

|red − blue| $\leq$ 2 |green − blue|

Step 1 : Symmetrization by ghost sample

Assume $X_1', \dots, X_n' \sim F$ independently ("ghost sample"),

Denote by F_n' the empirical cdf induced by ghost sample

Now it is easy to prove:

$$P\left(\sup_a \left| \underline{F_n(a)} - \underline{F(a)} \right| > \varepsilon \right)$$

$$\leq 2\, P\left(\sup_a \left| \underline{F_n(a)} - \underline{F_n'(a)} \right| > \frac{\varepsilon}{2} \right)$$

Step 2: Want to split this in two terms

$$\left| F_n(a) - F_n'(a) \right| = \Bigg| \overbrace{\frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{1}_{\{X_i \leq a\}} - \mathbb{1}_{\{X_i' < a\}} \right)}^{(*)} \Bigg|$$

Introduce Rademacher random variables $\sigma_1, \dots, \sigma_n$:

$$\sigma_i(\{-1\}) = \sigma_i(\{1\}) = 1/2.$$

Distribution of $(*)$ is the same as the distr. of the following:

$$\left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left(\mathbb{1}_{\{X_i \leq a\}} - \mathbb{1}_{\{X_i' \leq a\}} \right) \right| = (**)$$

Now we have:

$$2P\left(\sup_a |F_n(a) - F_n'(a)| > \frac{\varepsilon}{2}\right)$$

$$= 2P\left(\sup_a \left| \frac{1}{n} \sum \sigma_i \left(\mathbb{1}_{x_i \le a} - \mathbb{1}_{x_i' \le a} \right) \right| > \frac{\varepsilon}{2}\right)$$

$$\le 2P\left(\sup_a \underbrace{\left| \frac{1}{n} \sum \sigma_i \mathbb{1}_{x_i \le a} \right|}_{u} > \frac{\varepsilon}{4}\right) + 2P\left(\sup_a \underbrace{\left| \frac{1}{n} \sum \sigma_i \mathbb{1}_{x_i' \le a} \right|}_{v} > \frac{\varepsilon}{4}\right)$$

Observe:

$$P\left(|u - v| > \frac{\varepsilon}{2}\right) \le P\left(|u| > \frac{\varepsilon}{4} \text{ or } |v| > \frac{\varepsilon}{4}\right)$$

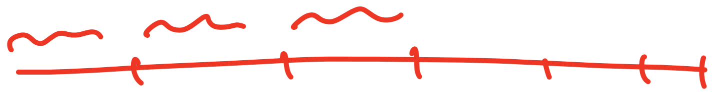
right side is necessary for left side

$$= 4 \cdot P\left(\sup_a \left| \frac{1}{n} \sum \sigma_i \mathbb{1}_{\{x_i \le a\}} \right| > \frac{\varepsilon}{4}\right)$$

Step 3 Exploit "finite structure":

Fix $X_1, \dots, X_n$ ($\equiv$ conditioning on $X_1, \dots, X_n$)

We look at $\mathbb{1}_{X_i \le a}$



The rvs $\mathbb{1}_{X_1 \le a}, \dots, \mathbb{1}_{\{X_n \le a\}}$ for fixed a can only have $n+1$ realizations

$$P\left(\sup_a \frac{1}{n} \left| \sum \sigma_i \mathbb{1}_{X_i \le a} \right| > \frac{\varepsilon}{4} \;\Big|\; X_1 \dots X_n \right) \le$$

$$\le (n+1) \sup_a \underbrace{P\left(\frac{1}{n} \left| \sum \sigma_i \mathbb{1}_{\{X_i \le a\}} \right| > \frac{\varepsilon}{4} \;\Big|\; X_1 \dots X_n \right)}_{\text{use Hoeffding (\#)}}$$

Step 4 Hoeffding apply to (*)

$\forall a$:

$$P\left(\frac{1}{n}\left|\sum \sigma_i \, \mathbb{1}_{x_i \le a}\right| > \frac{\varepsilon}{4} \;\middle|\; X_1 \dots X_n \right)$$

$$\le 2 \exp\left(- \frac{n \varepsilon^2}{32}\right)$$

Combining everything gives the theorem.

Product space

joint distributions

Product space, joint distributions

Consider two measurable spaces $(\Omega_1, \mathcal{A}_1)$, $(\Omega_2, \mathcal{A}_2)$.

Define the **product space** $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$ with

$$\Omega_1 \times \Omega_2 = \{ (\omega_1, \omega_2) \mid \omega_1 \in \Omega_1, \omega_2 \in \Omega_2 \}$$

$$\mathcal{A}_1 \otimes \mathcal{A}_2 = \{ A_1 \times A_2 \mid A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2 \}.$$

Consider two rvs $\quad X_1 : (\Omega, \mathcal{A}, P) \to (\Omega_1, \mathcal{A}_1)$

$\quad X_2 : (\Omega, \mathcal{A}, P) \to (\Omega_2, \mathcal{A}_2)$.

$X := (X_1, X_2) : (\Omega, \mathcal{A}, P) \to (\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$

$$(X_1, X_2)(\omega) = (X_1(\omega), X_2(\omega)).$$

The distribution $P_{(X_1, X_2)}$ on $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$ is called

the **joint distribution** of X_1 and X_2.

Example in ML: (X, Y) where X is the input data, Y is the label

Product measure

Product measure: $(\Omega_1, \mathcal{A}_1, P_1)$, $(\Omega_2, \mathcal{A}_2, P_2)$ two prob. spaces. We define the product measure $P_1 \otimes P_2$ on the product space $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$ as

$$(P_1 \otimes P_2)(A_1 \times A_2) := P_1(A_1) \cdot P_2(A_2).$$

Product ~ independence

Theorem Two rvs X_1, X_2 are independent if and only if their joint distribution coincides with the product distribution:

$$P_{(X_1,X_2)} = P_1 \otimes P_2 .$$

Marginal distribution

Marginal distribution

Consider the joint distribution $P_{(X_1,X_2)}$ of two rvs $X := (X_1, X_2)$. The marginal distribution of X wrt X_1 is the original distribution of X_1 on $(\Omega_1, \mathcal{A}_1)$, namely P_{X_1}. Similarly for P_{X_2}.

Example in the discrete case:

$Y \setminus X$	x_1	x_2	x_3	Σ
y_1	p_1	p_2	p_3	$p_1+p_2+p_3 = P(Y=y_1)$
y_2	p_4	p_5	p_6	$p_4+p_5+p_6 = P(Y=y_2)$
	$p_1+p_4 = P(X=x_1)$	...		

$\hat{=}$ marginal distribution wrt Y.

marginal wrt X

Marginal distributions in case of densities

$X, Y : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, $Z := (X, Y)$. Assume that the joint distribution of Z has a density f on $\mathbb{R}^2$. Then the following statements hold:

(1) Both X and Y have densities on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ given by

joint density

sum over Y

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\, dx$$

(2) X and Y are independent iff

$$f(x,y) = f_X(x) \cdot f_Y(y) \quad \text{a.s.}$$

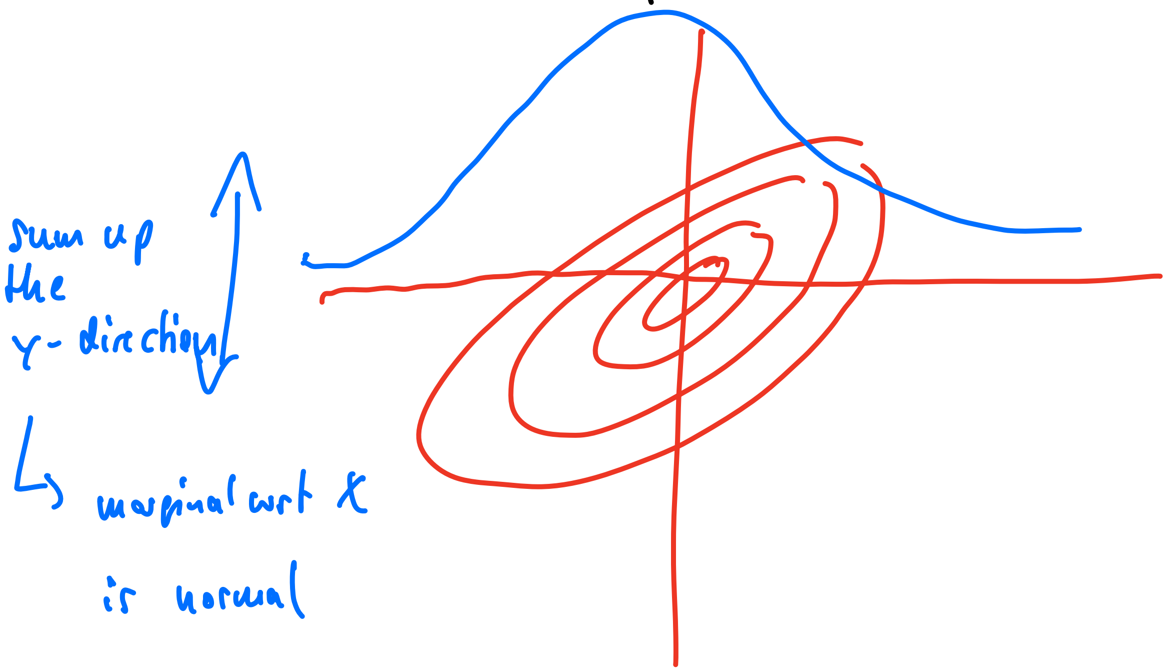
Special case: marginals of multivariate normal distributions

<u>2 dim</u> Consider a 2-dim normal rv $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ with mean

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \in \mathbb{R}^2 \quad \text{and} \quad \text{cov. } \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}.$$

Then the marginal distribution of X wrt X_1 is again

a normal distribution with mean μ_1 and var σ_1^2.



n-dim

$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$. Group the variables: $\left.\begin{matrix} x_1 \\ \vdots \\ x_k \end{matrix}\right\} \tilde{X} \in \mathbb{R}^k$

$\left.\begin{matrix} x_{k+1} \\ \vdots \\ x_n \end{matrix}\right\} X^\# \in \mathbb{R}^{n-k}$

Want to look at the marginal of X wrt $\tilde{X}$.

$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}$ mean, $\tilde{\mu} := \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}$, $\mu^\# = \begin{pmatrix} \mu_{k+1} \\ \vdots \\ \mu_n \end{pmatrix}$

$$\underset{\mathbb{R}^{n\times n}}{\underset{\uparrow}{\Sigma}} = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right) \begin{matrix} \}\, k \\ \\ \end{matrix}$$

$$\underbrace{\qquad}_{k}$$

Now the marginal of X wrt $\tilde{X}$ is a normal distr. on $\mathbb{R}^k$ with mean $\tilde{\mu}$ and cov. Σ_{11}.

Conditional distributions

Conditional distributions in the discrete case

<u>Discrete case:</u>

Know conditional probabilities: $P(A \mid B)$

defined for events $A, B \in \mathcal{A}$, and $P(B) > 0$.

Let $X, Y : (\Omega, \mathcal{A}, P) \to \mathbb{R}$ be discrete rv, $y \in \mathbb{R}$ such that $P(Y = y) > 0$. Then we can define the conditional probability measure $P_{X|Y=y} : A \longmapsto P(X \in A \mid Y = y)$.

This is a probability measure.

Conditional distributions in case of densities

Assume $Z := (X, Y)$ has a joint density $f: \mathbb{R}^2 \to \mathbb{R}$, and marginal densities $f_X, f_Y: \mathbb{R} \to \mathbb{R}$. Then the function

$$f_{X|Y=y}(x) := \frac{f(x,y)}{f_Y(y)}$$

is then also a density on $\mathbb{R}$, called the conditional density of X given $Y=y$.

General case (not discrete, no density)

For general rv this is surprisingly complicated!

$\rightsquigarrow$ "regular conditional probabilities" $\rightsquigarrow$ skipped

Example: normal distribution

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} \begin{matrix} \} \, \tilde{\mu} \\ \} \, \mu^\# \end{matrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

If $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \sim N(\mu, \Sigma)$, then the conditional distribution of $\tilde{X} = \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix}$ wrt $x^\# = \begin{pmatrix} x_{k+1} \\ \vdots \\ x_n \end{pmatrix}$ is given by

$$p_{\tilde{X} \mid x^\#} \sim N\Big(\mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x^\# - \tilde{\mu}), \; \Sigma_{22} - \Sigma_{12}^t \Sigma_{11}^{-1} \Sigma_{12} \Big).$$

Conditional densities

To Do add

Conditional expectations

Conditional expectation in the discrete case (cond. on event)

Def (discrete case) $X, Y : (\Omega, \mathcal{A}, P) \to \mathbb{R}$

assume X takes finitely (countably) many values $x_1, \dots, x_n \in \mathbb{R}$, Y takes finitely (countably) many values $y_1 \dots y_m \in \mathbb{R}$, always with a positive probability.

Then we define the conditional expectations

$$E(Y \mid X = x_i) := \sum_{j=1}^{m} y_j \underbrace{P(Y = y_j \mid X = x_i)}_{\text{well defined}}$$

Example

Example: two dice, X = first one, Y = second one, independent

$$E(\text{sum} \mid X = 1) = \sum_{i=1}^{12} i \cdot P(\text{sum} = i \mid X = 1)$$

$$= \sum_{k=1}^{6} (1+k) \cdot P(Y = k \mid X = 1)$$

$$= \sum_{k=1}^{6} (1+k)\, P(Y = k) = \sum_{k=1}^{6} (1+k) \cdot \frac{1}{6} = 4.5$$

So far we defined $E(Y \mid X = x_i)$, but often we want to consider the "function" $E(Y \mid X)(\omega)$. This is a rv:

$E(Y \mid X) : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B})$. Leads to the following:

Cond. expectation wrt a rv

Def (discrete case) X, Y as before. Then the conditional expectation is defined as follows:

$$E(Y|X) := f(X) \quad \text{with}$$

$$f(x) = \begin{cases} E(Y|X=x) & \text{if } P(X=x) > 0 \\ \text{arbitrary, say } 0 & \text{otherwise} \end{cases}$$

⚠ $E(Y|X)$ is only defined a.s.

General case?

Problem: $P(X=x)$ might be 0 all of the time...

Case of densities is still straightforward:

Case of joint densities

$X, Z : \Omega \to \mathbb{R}$ have a joint density $f(x,z)$.

Let $g : \mathbb{R} \to \mathbb{R}$ bounded, set $Y := g(Z)$. Assume we want to compute $E(Y|X) = E(\underbrace{g(Z)}_{Y}|X)$.

Recall X has density $f_X(x) = \int f(x,z)\,dz$.

The conditional density of Z given $X=x$ is

$$f_{X=x}(z) = \frac{f(x,z)}{f_X(x)} \qquad (\text{if } f_X(x) \neq 0)$$

Now consider $h(x) := \int \underbrace{g(z)}_{Y} f_{X=x}(z)\,dz$, now define

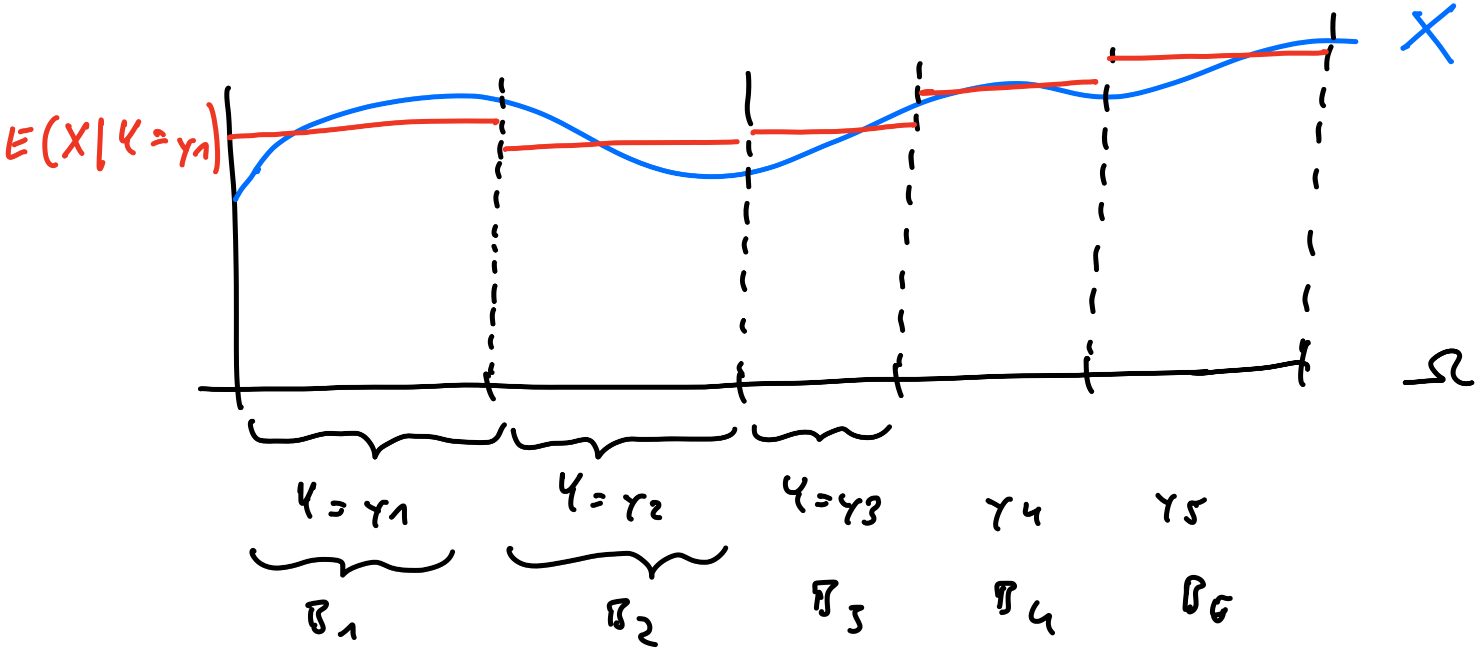
$E(Y|X) = h(X)$.

Idea for the more general case

Consider: X continuous rv

Y discrete rv $\rightsquigarrow y_1, \dots, y_5$

Want to look at $E(X \mid Y)$



Want to "define" $E(X \mid Y) := \sum_{i=1}^{5} E(X \mid Y = y_i) \cdot \underbrace{\mathbb{1}_{B_i}(\omega)}_{rv}$

But need to make sure that it is measurable wrt $\sigma(Y)$. ("technic")

Cond. expectation on L_1

<u>Def</u> (Conditional expectation on L_1)

Consider rv $X : (\Omega, \mathcal{F}_0, P) \to \mathbb{R}$, $X \in L_1(\Omega, \mathcal{F}_0, P)$.

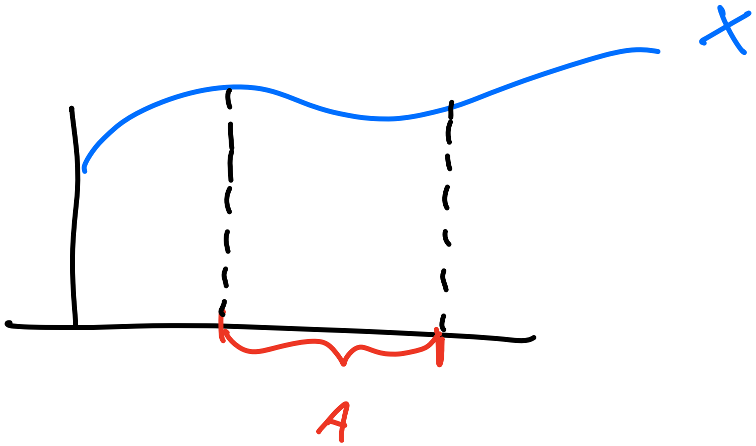
Let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{F}_0$. (Intuition: $\mathcal{F}_0$ will be the σ-alg. generated by the variable Y we want to condition on).

We now define the cond. exp. of X given $\mathcal{F}$ $E(X \mid \mathcal{F})$ as any random variable Z that satisfies

(1) Z is measurable wrt $\mathcal{F}$

(2) For all $A \in \mathcal{F}$ we have

$$\int_A X \, dP = \int_A Z \, dP .$$



- Existence of $E(X|\mathcal{F})$ is not clear a priori, it needs to be proved.
- $E(X|Y) := E(X|\sigma(Y))$

<u>Examples</u>

- $X = Y$. Then $E(X|Y) = X$ (a.s.)
- $X \perp\!\!\!\perp Y$. $E(X|Y) = E(X)$ (a.s.)