Probability measure

- Given space Ω ("abstract space").
- Need a σ-algebra $\mathcal{A}$ on Ω ("measurable events")
  - $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$
  - $(A_i)_{i \in \mathbb{N}} \subset \mathcal{A} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$ ("countable unions")
  - $\emptyset, \Omega \in \mathcal{A}$
  - countable intersections

- A measure μ on $(\Omega, \mathcal{A})$ is a function

  $\mu : \mathcal{A} \to [0, \infty]$

  that is countably additive: If $(A_i)_{i \in \mathbb{N}}$ is a sequence of pairwise disjoint sets, then

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$

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.

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

Example (1): Throw one die

$\Omega = \{1, 2, \dots 6\}$, $\mathcal{A} = \mathcal{P}(\Omega)$ (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:

first die, second

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

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

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

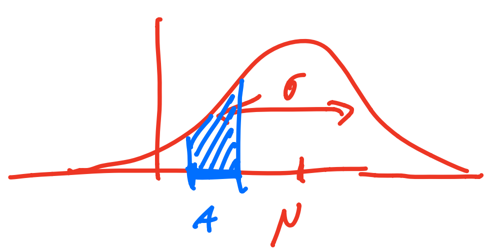
Example (2) Normal distribution

$$\Omega = \mathbb{R}$$

$$\mathcal{A} = \text{Borel-}\sigma\text{-algebra}$$

$$f_{\mu,\sigma} : \mathbb{R} \to \mathbb{R},$$

$$x \longmapsto \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



$$P : \mathcal{A} \to [0, 1]$$

$$P(A) := \int_A f_{\mu,\sigma}(x)\, dx$$

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 $0 \le p_i \le 1$, $\sum_i p_i = 1$.

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

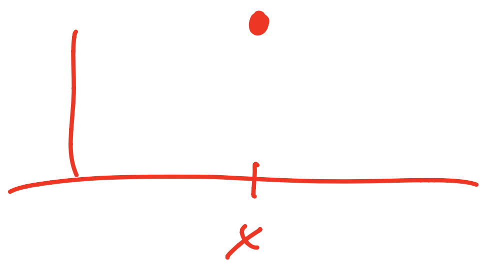
$$P(A) = \sum_{\{i \mid x_i \in A\}} p_i \, .$$

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

Dirac measures:

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.

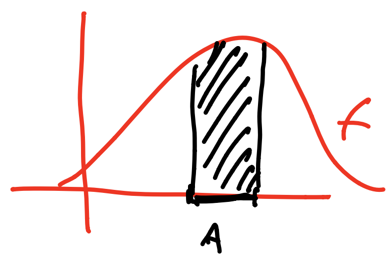
A discrete measure on $\mathbb{R}$ can be written as a sum of Dirac measures. For example, throwing a die can be described as $\frac{1}{6}(\delta_1 + \delta_2 + \dots + \delta_6)$

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: $\nu = f \cdot \lambda$

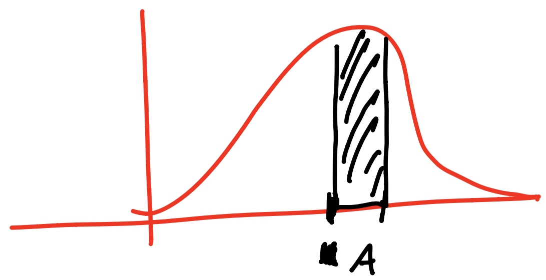
Question? Can we describe every prob measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ in terms of a density? Answer: no!

Counterexample: δ_0 Dirac measure

Def A prob. measure ν on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is called <u>absolutely continuous</u> 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:

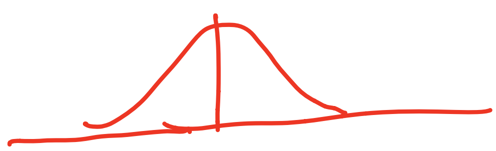
$$\forall B \in \mathcal{B}(\mathbb{R}^n): \quad \mu(B) = 0 \implies \nu(B) = 0.$$

Notation: $\nu \ll \mu$



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

Example: $N(0,1) \ll \lambda$



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

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

Theorem (Radon-Nikodym)

Consider two prob. 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.

<u>Proof idea</u>

(1) $\Rightarrow$ (2) easy

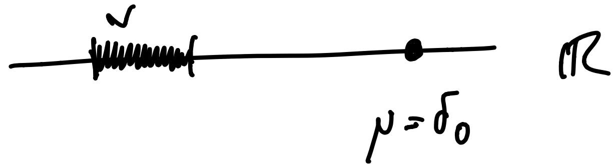
(2) $\Rightarrow$ (1) We need to construct a density!

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. ∎

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: $\mu \perp \nu$.

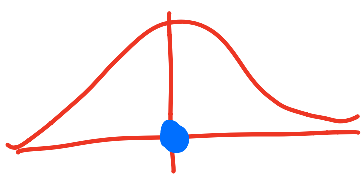


Example: $\lambda \perp \delta_0$

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



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

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

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

Does the job. $\blacksquare$

<u>Cantor-distribution: non-trivial distribution that is singular wrt λ</u>.

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"

- $C_2 =$

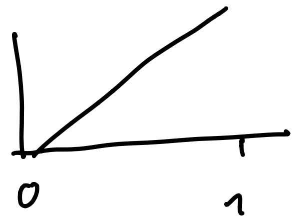
$\vdots$

The Cantor set is the limit in this process.

Now construct a probability distribution:

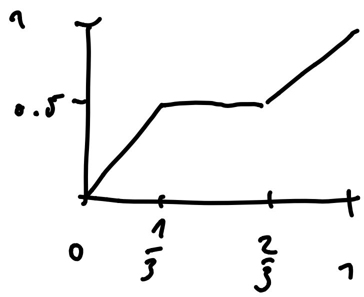
Consider the cdfs of the sets $C_0, C_1, C_2 \dots$

C_0 :



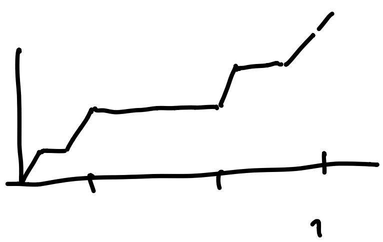
uniform on $[0, 1]$

C_1 :



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

C_2 :



$\vdots$

Take limit. $\to \mu$ Can prove many strange properties:

- Cantor-set is compact, non-empty, empty interior.
- The cdf of "μ" is continuous. μ is a prob. measure.
- But: $\lambda(C) = 0$.

$\Rightarrow \lambda \perp \mu$

Cumulative distribution function

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

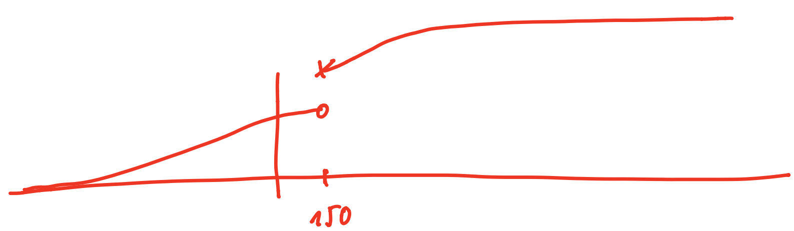
Define the function $F: \mathbb{R} \to \mathbb{R}$, $x \mapsto P(]-\infty, x])$.

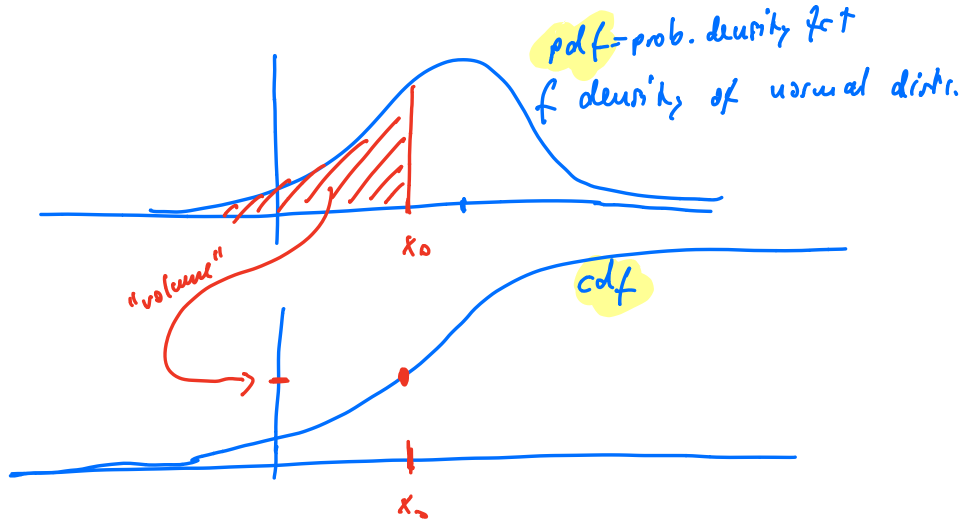
We say that F is a cumulative distribution function (cdf), that is it satisfies the following ~~prob~~ properties:

(i) F is monotonically increasing,

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

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





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

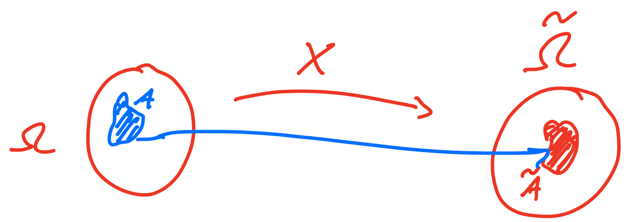
Then there exist a unique prob. measure P on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$

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

Random variable

Def 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 <u>random variable</u> if X is measurable, i.e.

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



people in the world

height

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

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

$P(A) = 0.1$

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"

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

is measurable.

Def A random variable $X : \Omega \to \tilde{\Omega}$ 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.

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

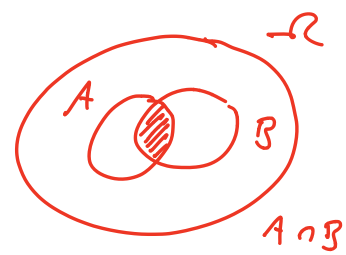
$$\sigma(X) := \{ X^{-1}(\tilde{A}) \mid \tilde{A} \in \tilde{\mathcal{A}} \}$$

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

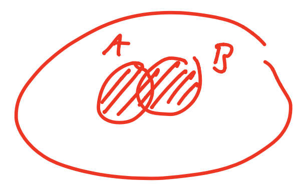
(it is the smallest σ-algebra on Ω that makes X measurable).

Conditional probabilities

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



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



Def $(\Omega, \mathcal{A}, P)$ 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.

Theorem 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.

Example: Ω = all persons on earth

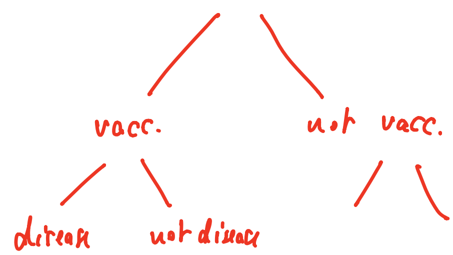
$\mathcal{A} = \mathcal{P}(\Omega)$

P = "uniform"

Event A := "person has been vaccinated"

B := "person has disease"

$P($ disease $|$ vaccinated $)$



$P($ vaccinated $|$ disease $)$

Example: two dice

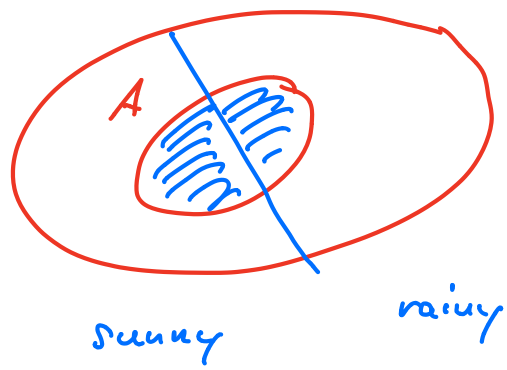
$P($ "sum is 10" $|$ "first dice was 5" $)$

Bayes formula

corrected

Law of total probability: Let $B_1, B_2, \dots, B_k$ be a disjoint partition of Ω with $B_i \in \mathcal{A}$ for all i, and $A \in \mathcal{A}$. 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:

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

Example: breast cancer screening

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

90% of women with breast cancer will be test 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

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

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

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

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!)

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:

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

Notation for independence:

$$A \perp\!\!\!\perp B$$

$$X \perp\!\!\!\perp Y$$

Expectation (discrete case)

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

Definition $(\Omega, \mathcal{A}, P)$ prob. space, $S \subset \mathbb{R}$ at most countable, $X : \Omega \to S$ random variable.

If $\sum_{r \in S} |r| \cdot P(X = r) < \infty$, 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 $\mathbb{E}(X)$).

Examples

- Toss a coin. $\Omega = \{\text{head}, \text{tail}\}$, $\mathcal{A} = \mathcal{P}(\Omega)$, $P(\text{head}) = p$, $P(\text{tail}) = 1 - p$.

  $0 < p < 1$.

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

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

- Test error of a classifier.

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

Important properties :

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

- X, Y independent $\Rightarrow E(X \cdot Y) = E(X) \cdot E(Y)$

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

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

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

Variance, covariance, correlation
(discrete case)

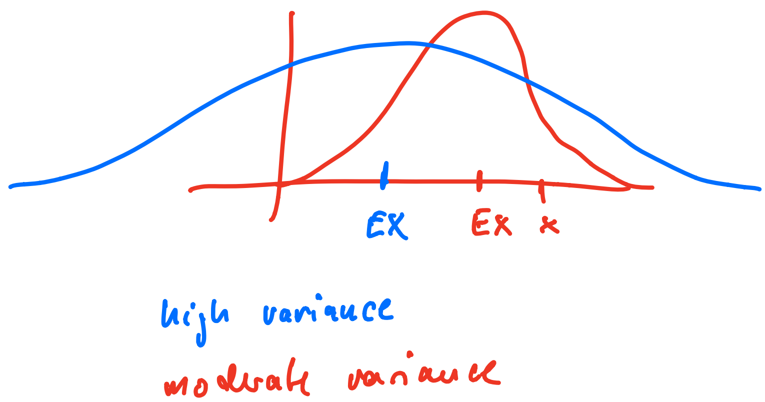
<u>Def</u> $X, Y : (\Omega, \mathcal{A}, P) \to \mathbb{R}$ discrete rvs with
$E(X^2) < \infty$, $E(Y^2) < \infty$.

Then $$Var(X) := E\left((X - E(X))^2 \right)$$

is called the variance of X

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

is called the standard deviation.



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

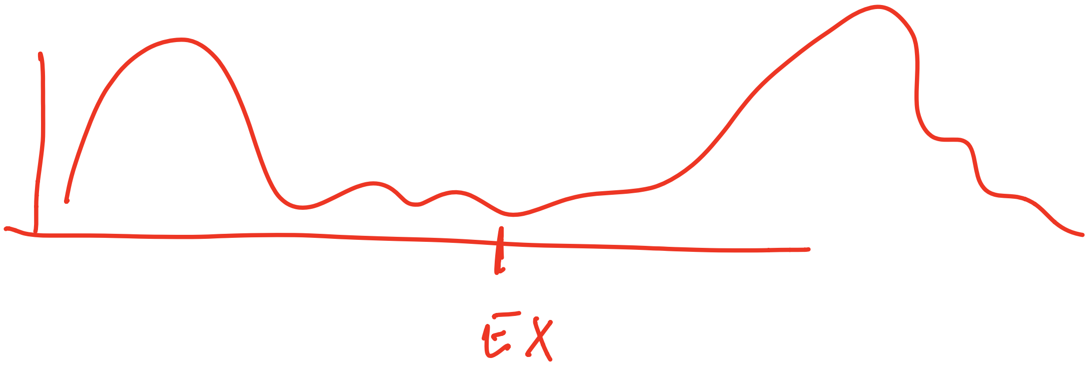
$$\rho_{XY} := \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$ $\in [-1, 1]$ is called the correlation coefficient.

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

More generally, for $k \in \mathbb{N}$ we define the terms

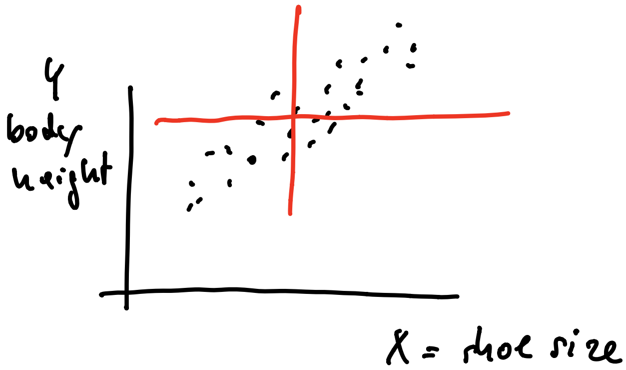
$E(X^k)$ ("k-th moment"),

$E\left((X - E(X))^k \right)$ ("kth centered moment")



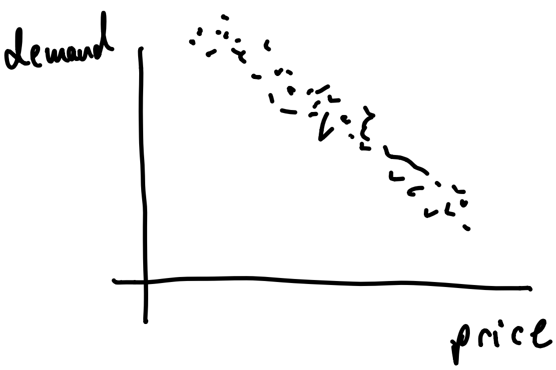
Intuition about covariance

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



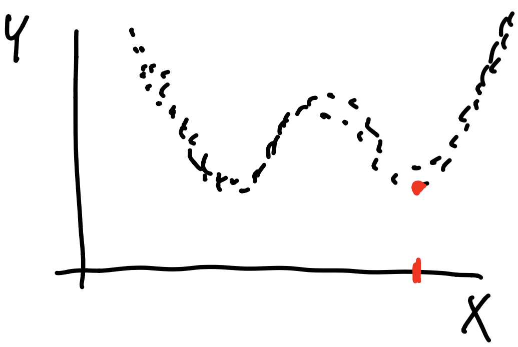
positive, large covariance

$\rho \approx 0.9$



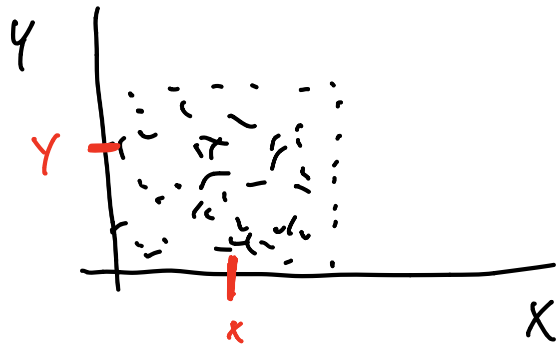
negative cov,
large in absolute values

$\rho \approx -0.9$



$\mathrm{Cov} \approx 0$
(uncorrelated).

⚠ uncorrelated $\not\Rightarrow$ independence!
$\Leftarrow$



independence

Properties

- $Var(X) = E(X^2) - (E(X))^2$
- $Cov(X,Y) = E(X \cdot Y) - E(X) \cdot E(Y)$
- $E(aX + b) = a \cdot E(X) + b$
- $Var(a \cdot X + b) = a^2 \, Var(X)$
- $Cov(X,Y) = Cov(Y,X)$
- $Var(X+Y) = Var(X) + Var(Y) + Cov(X,Y)$
- X, Y independent $\Rightarrow$ $Cov(X,Y) = 0$
  $\not\Leftarrow$
- X, Y independent $\Rightarrow$ $Var(X+Y) = Var(X) + Var(Y)$.

Expectation and variance in the general setting

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

$(\Omega, \mathcal{A}, P)$ prob. space, $X : \Omega \to \mathbb{R}$ with distribution $P_X = X(P)$, $X \in L^1(\Omega, \mathcal{A}, P)$. The expectation of X is then defined as

$$E(X) := \int_\Omega X \, dP = \int_\mathbb{R} x \, dP_X(x)$$

(case of density f: $\int_\mathbb{R} x \, f(x) \, dx$)

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.

If $X \in L^2(\Omega, \mathcal{A}, P)$ we define

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

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

Markov and Chebyshev inequalities

Cauchy-Schwartz-inequality

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

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

Markov inequality:

$\varepsilon > 0$, $f : [0, \infty[\to [0, \infty[$, f monotonically increasing. Then

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

In particular,

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

Chebyshev inequality:

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

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

Examples of probability distributions

Discrete distributions

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

- Binomial distribution on $\{0, \dots, n\}$

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

  $X :=$ # heads

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

- Poisson distribution on $\mathbb{N}$

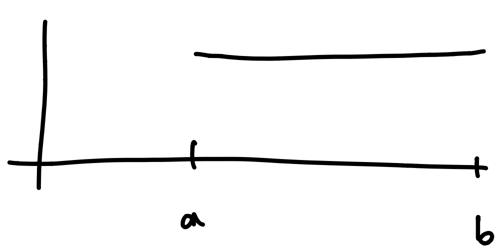
  Parameter $\lambda > 0$

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

Intuition: number of incoming calls at a hotline.

Continuous distributions

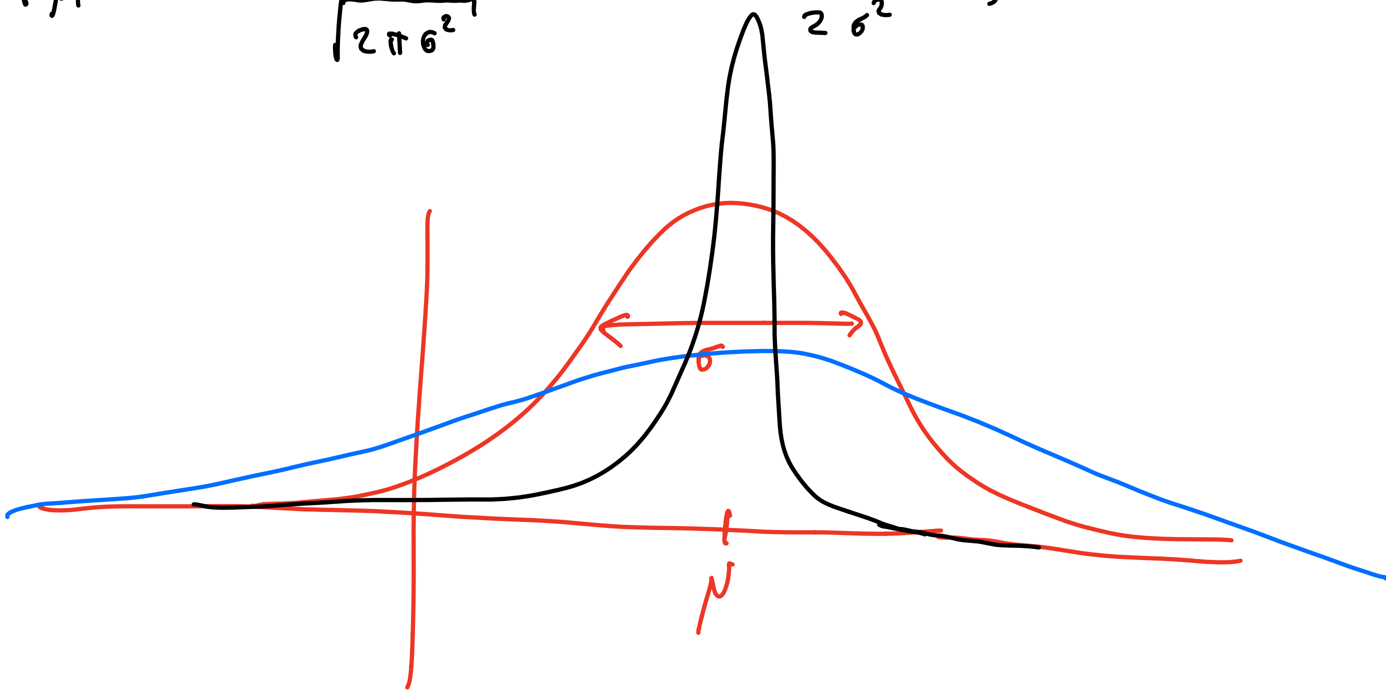
Uniform distribution on $[a, b]$: constant density



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: $N(\mu, \sigma^2)$

Some first properties:

- $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, X, Y independent.

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

Normal distribution in higher dimension

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

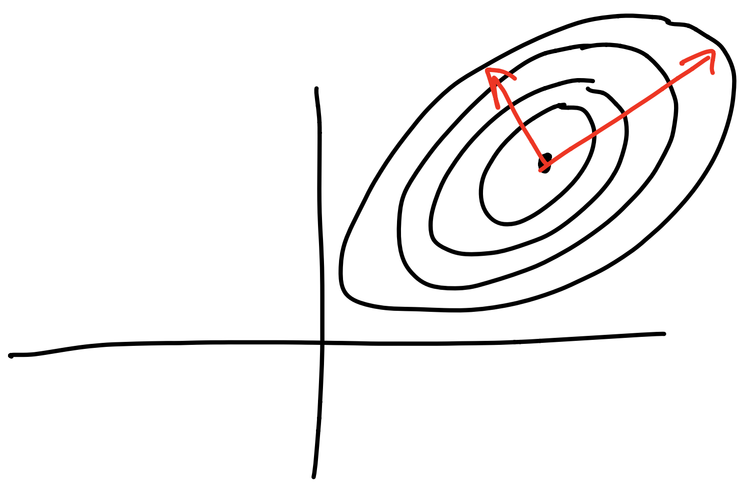
$\Sigma \in \mathbb{R}^{n \times n}$ with $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$, called covariance matrix.

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

Notation: $\mathcal{N}(\mu, \Sigma)$

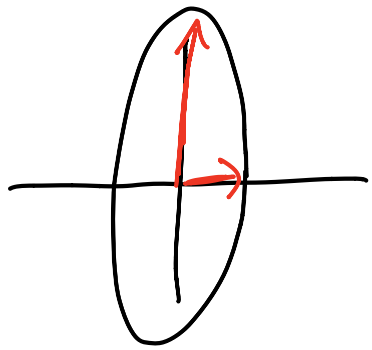
<u>Prop</u> Σ is psd and symmetric.

Consequence: Σ has real-valued, non-negative eig's.



contour lines of $f_{\mu,\Sigma}$

- $\cancel{X}$ $X_1, \ldots, X_n$ are independent $\iff$ $\Sigma = \begin{pmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_n^2 \end{pmatrix}$



- $X \sim N(\mu_1, \Sigma_1)$, $Y \sim N(\mu_2, \Sigma_2)$, independent, then

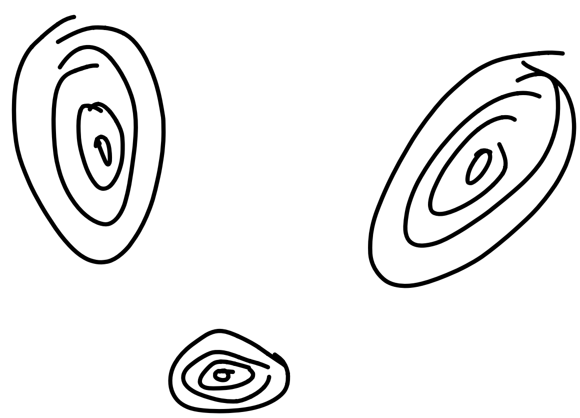
$$X + Y \sim N(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2)$$

Mixture of Gaussians

Consider $\pi_1, \pi_2, \dots, \pi_k$ with $0 \leq \pi_i \leq 1$ and $\sum \pi_i = 1$

Consider the following density:

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



Convergence of random variables

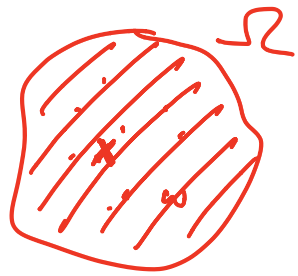
Consider rv $X_i : \Omega \to \mathbb{R}$, $i \in \mathbb{N}$, $X : \Omega \to \mathbb{R}$,

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

(1) $(X_i)_{i \in \mathbb{N}}$ 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.



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

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

Let us check that these definitions make sense. We need to prove that the events in (1) and (2) are in fact in ~~the~~ $\mathcal{A}$.

Case (1):

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

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

So we get:

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

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

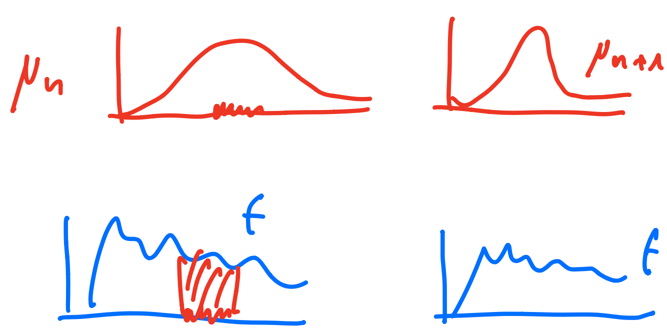
(3) $X_n \to X$ <u>in L^p ("in the p-th mean")</u> $:\Leftrightarrow$

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

(4) 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.

$\mu_n \to \mu$ <u>weakly</u> $:\Leftrightarrow$

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



Excursion:

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

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 $C_b(\mathbb{R}^n)$.

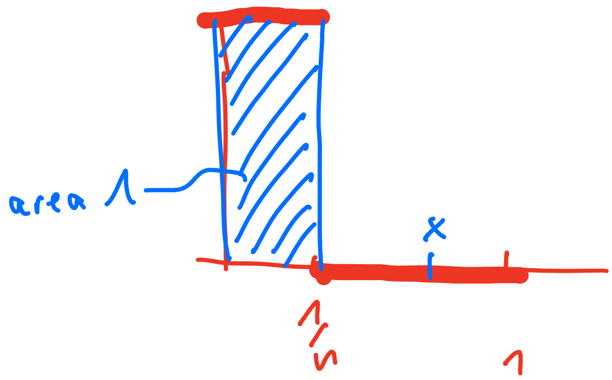
(5) $X_i, X : (\Omega, \mathcal{A}, P) \to \mathbb{R}^n$. The sequence X_n converges in distribution to X $:\Longleftrightarrow$ the distributions P_{X_n} converge to P_X weakly.

We have the following implications (but none of the missing implications is true 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)

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



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

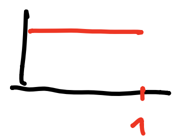
Can formally see: a.s., in prob.

But: no convergence in L^1.

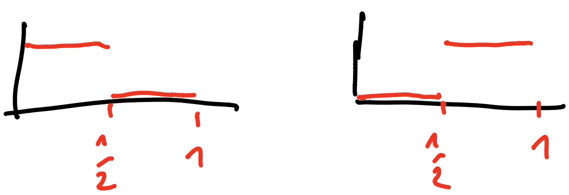
Example (convergence in prob., in L^1, but not a.s.)

"sliding blocks"

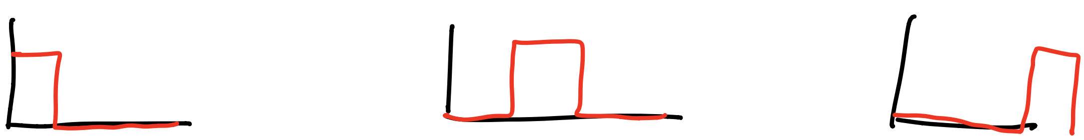
$f_1 = \mathbb{1}_{[0,1]}$



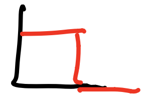
$f_2 = \mathbb{1}_{[0,1/2]}, \quad f_3 = \mathbb{1}_{[1/2,1]}$

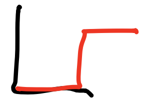


$f_4 = \mathbb{1}_{[0,1/3]}, \quad f_5 = \mathbb{1}_{[1/3,\frac{2}{3}]}, \quad f_6 = \mathbb{1}_{[\frac{2}{3},1]}$



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

- $X_n : [0, 1] \to \mathbb{R}$, $X_1 = X_2 = \dots = \mathbb{1}_{[0, \frac{1}{2}[}$ 

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

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

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

$$P(A_n \text{ infinitely often}) := P(A_n \text{ i.o.})$$
$$= P(\{\omega \in \Omega \mid \omega \in A_n \text{ for infinitely many } n\})$$

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

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

<u>Proof intuition</u> :

$$\{\lim X_n = X\}$$
$$= \{\forall k : |X_n - X| > \tfrac{1}{k} \text{ at most finitely often}\}$$
$$= \bigcap_{k \in \mathbb{N}} \{|X_n - X| > \tfrac{1}{k} \text{ at most fin. often}\}$$
$$\left(\bigcup_{k \in \mathbb{N}} \{|X_n - X| > \tfrac{1}{k} \text{ inf. often}\}\right)^{\text{complement}}$$

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

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.

Limit theorems: LLN and CLT

Strong law of large numbers

$X_n : (\Omega, \mathcal{A}, P) \to \mathbb{R}$ iid (identically distributed and independent). Assume the mean $\mu := E(X_1) < \infty$, and $\mathrm{Var}(X_1) =: \sigma^2 < \infty$. Then:

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

Remarks:

- Many versions of this theorem exist. (slightly relaxing iid)
- "Strong law" $\Leftrightarrow$ convergence a.s.

  "Weak law" $\Leftrightarrow$ convergence in probability

Central Limit 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

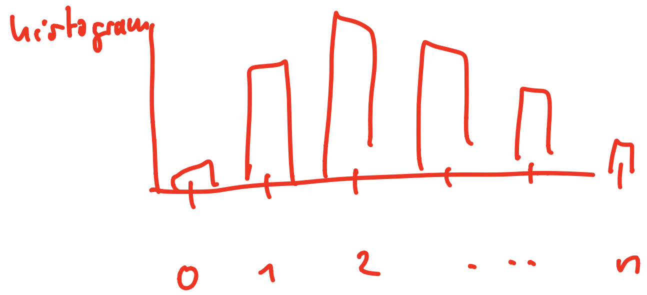
$$Y_n := \frac{S_n - n \cdot \mu}{\sqrt{n}\,\sigma}$$

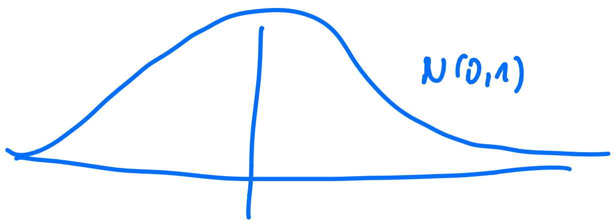
(which has mean 0 and standard dev. 1).

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

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

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



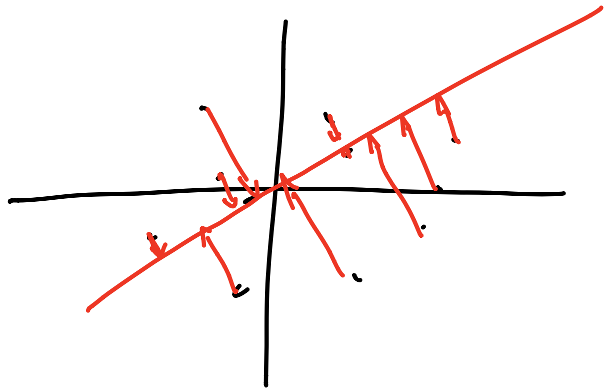


Concentration inequalities

Motivation: random projections

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

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



Theorem of Johnson - Lindenstrauss:

Can guarantee (for certain parameters ε, k)

$$(1-\varepsilon)\,\|x_i - x_j\|_{\mathbb{R}^d} \leq \|\pi(x_i) - \pi(x_j)\|_{\mathbb{R}^\ell}$$

$$\leq (1+\varepsilon)\,\|x_i - x_j\|_{\mathbb{R}^d}$$

Construction / proof steps:

(1) Assume you know $\|x_i - x_j\|_{\mathbb{R}^d} = 1$.

Compute $E(\|\pi(x_i) - \pi(x_j)\|_{\mathbb{R}^\ell})$, "easy".

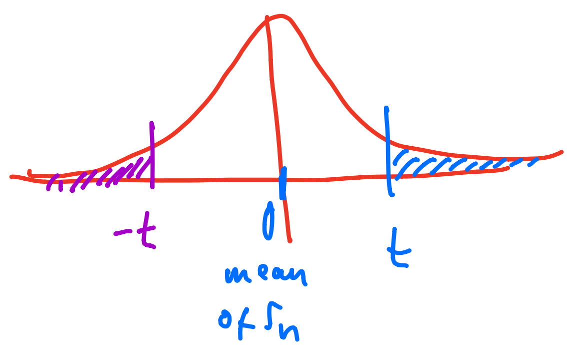
(2) $P(\,|\,\|\pi(x_i) - \pi(x_j)\| - E(\dots)\,| > t)$?

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: SLLN

Prop $(X_i)_{i \in \mathbb{N}}$ 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.

Proof 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:

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

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

(*) 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. ∎

Remark: 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 $t > 0$,

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

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

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

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

Theorem (McDiarmid) $X_1, \dots, X_n$ 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.

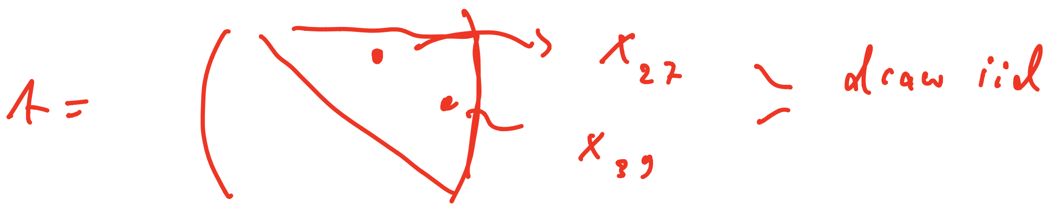
Then, for any $t > 0$,

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

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

Applications:

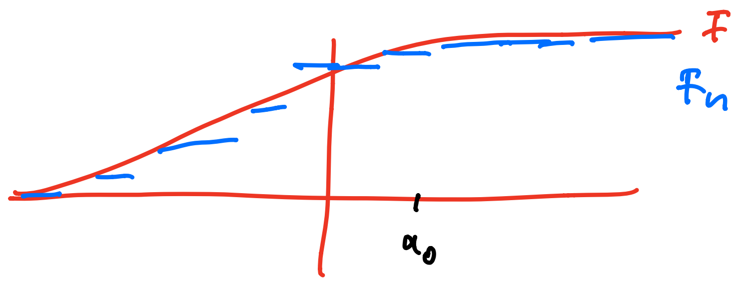
- stability in ML
- standard theoretical CS, randomized algorithms
- largest eigenvalue of a random symmetric matrix



Glivenko - Cantelli Theorem

F cdf : $F(a) = P(X \le a)$

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



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

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

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

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

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

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

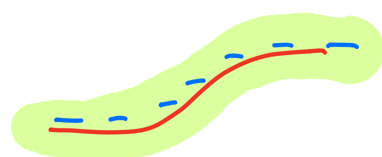
So it is clear that $F_n \xrightarrow{a.s.} F$ pointwise (i.e. $\forall a_0$)

Now let's look at uniform convergence.

__Theorem__ $X_1, \dots, X_n$ iid random variables with cdf F.

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

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



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

In particular, $\sup |F_n - F| \longrightarrow 0$ 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:

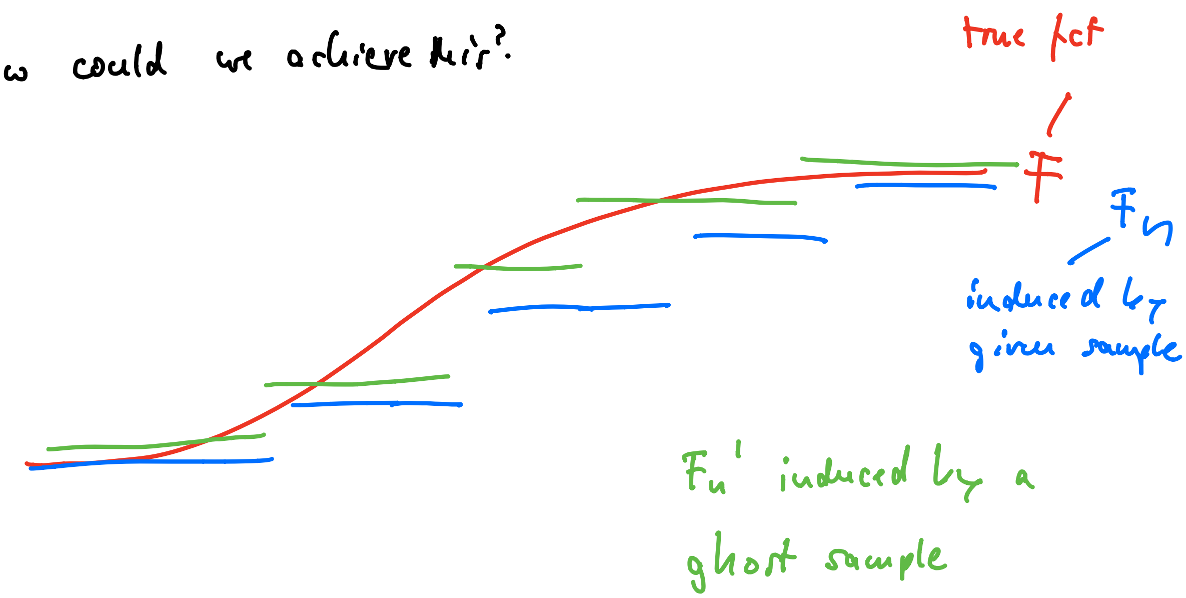
$$P\left(\max_{i=1\dots n} |U_i| > \varepsilon\right) =$$

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

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

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

How could we achieve this?



$|\text{red} - \text{green}|$
$\shortparallel$
$|\text{red} - \text{blue}| \leq 2\,|\text{green} - \text{blue}|$

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:

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

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

Step 2 : Want to split this in two terms

$$| F_n(a) - F_n'(a) | = \Big| \overbrace{\frac{1}{n} \sum_{i=1}^{n} \left(1\!\!1_{\{X_i \leq a\}} - 1\!\!1_{\{X_i' \leq a\}} \right)}^{\circledast} \Big|$$

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

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

Distribution of $\circledast$ is the same as the distr. of the following:

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

Now we have:

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

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

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

Observe:

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

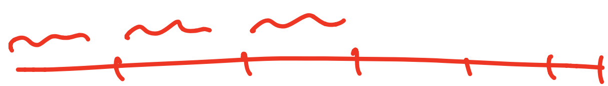
right side is necessary for left side

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

Step 3 Exploit "finite structure":

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

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



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

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

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

<u>Step 4</u> Hoeffding to (*) (apply)

$\forall a$:

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

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

Combining everything gives the theorem.

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
$$X_1 : (\Omega, \mathcal{A}, P) \to (\Omega_1, \mathcal{A}_1)$$
$$X_2 : (\Omega, \mathcal{A}, P) \to (\Omega_2, \mathcal{A}_2).$$

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

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: $(\Omega_1, \mathcal{A}_1, P_1)$, $(\Omega_2, \mathcal{A}_2, P_2)$ 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

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

$\boxed{\begin{array}{l}\underline{\text{Theorem}} \quad \text{Two rvs } X_1, X_2 \text{ are independent if and only if} \\ \text{their joint distribution coincides with the product distribution:} \\ \qquad P_{(X_1, X_2)} = P_1 \otimes P_2 .\end{array}}$

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

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

Mixed cases

For example, consider X a continuous rv with density and Y a discrete rv.

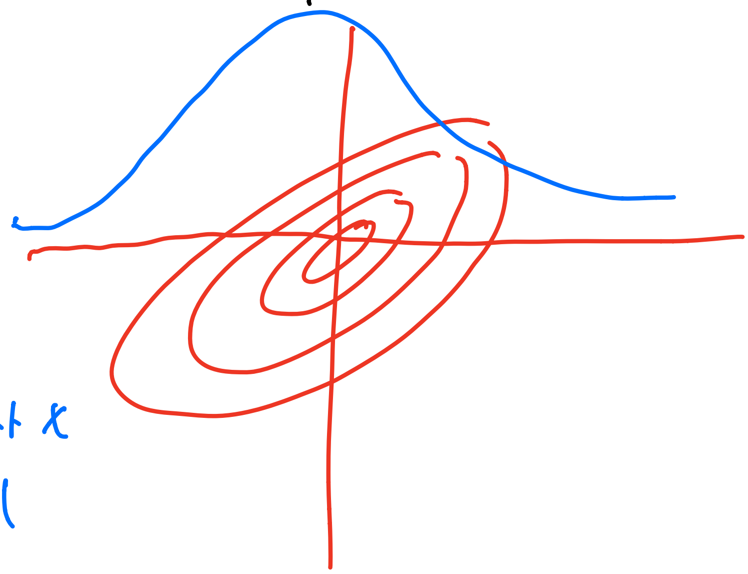
Say, X = income $\in \mathbb{R}$
Y = "yes" or "no", discrete

Special case: marginals of multivariate normal distributions

2 dim 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 w.r.t. 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, $\tilde{\mu} := \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}$, $\mu^{\#} = \begin{pmatrix} \mu_{k+1} \\ \vdots \\ \mu_n \end{pmatrix}$

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

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

Discrete case:

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 $P_{X|Y=y} : A \mapsto P(X \in A \mid Y = y)$.

This is a probability measure.

For general rv this is surprisingly complicated!

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

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

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

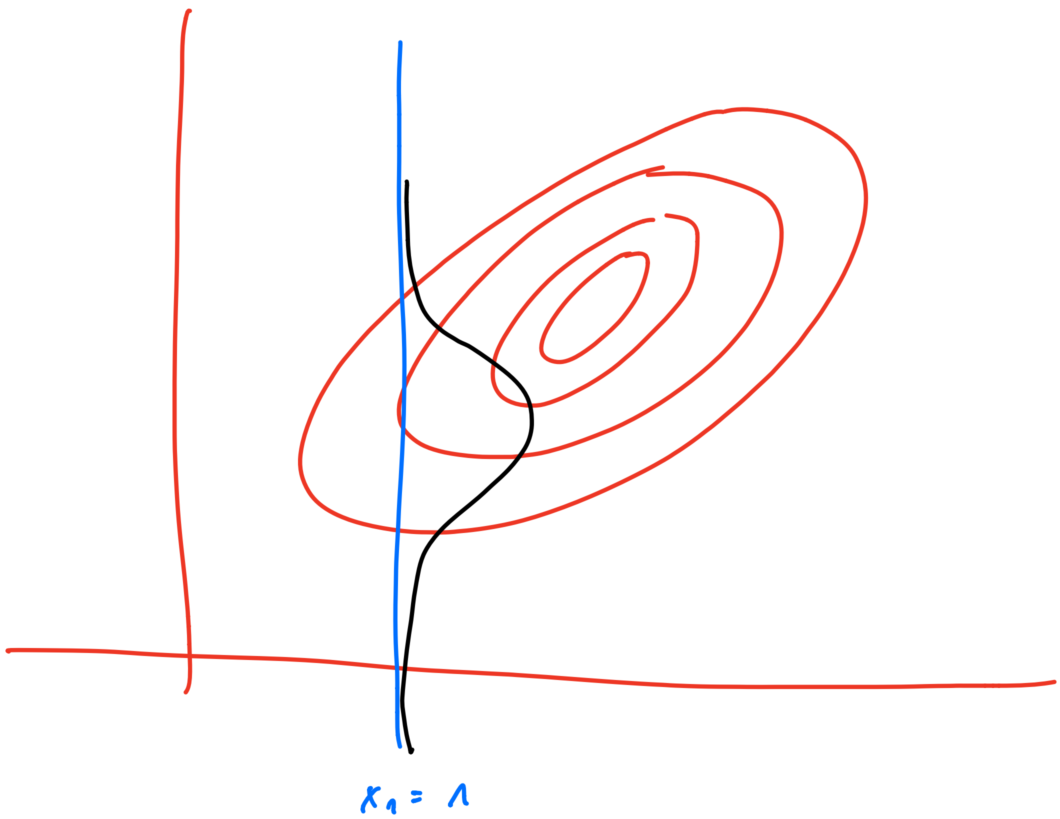
is then also a density on $\mathbb{R}$, called the conditional density of X given $Y = y$.

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 expectation

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.

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

Example: 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:

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

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

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

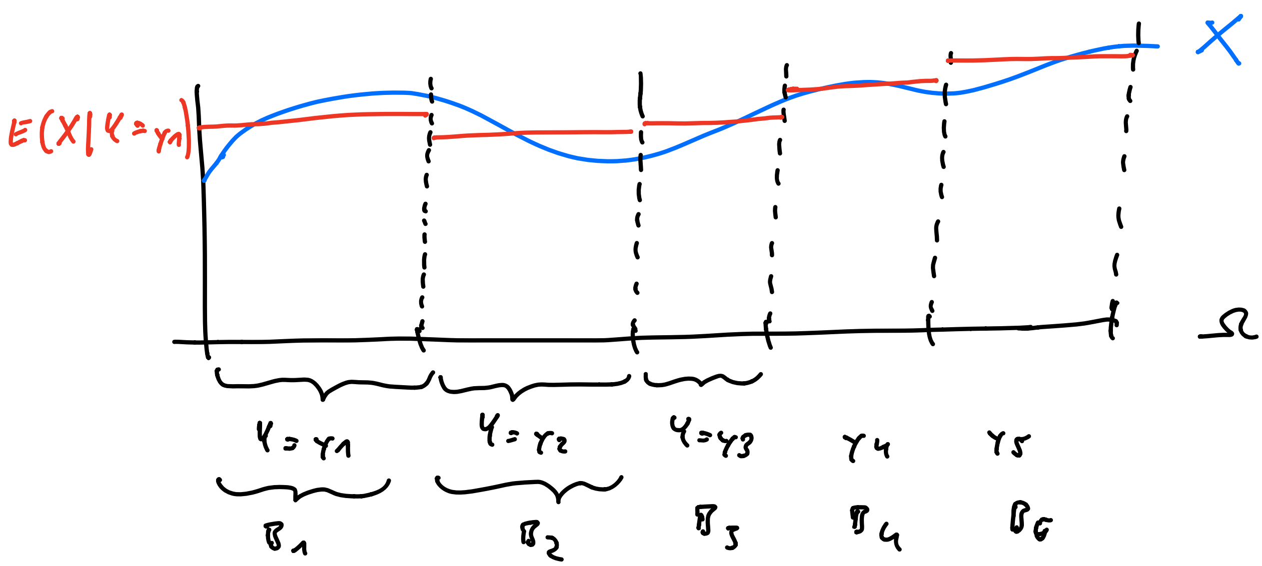
⚠ $E(Y|X)$ is only defined a.s.

Now we want to move to the more general case.

Sketch: X continuous rv

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

Want to look at $E(X|Y)$



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

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

Def (Conditional expectation on L_1)

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

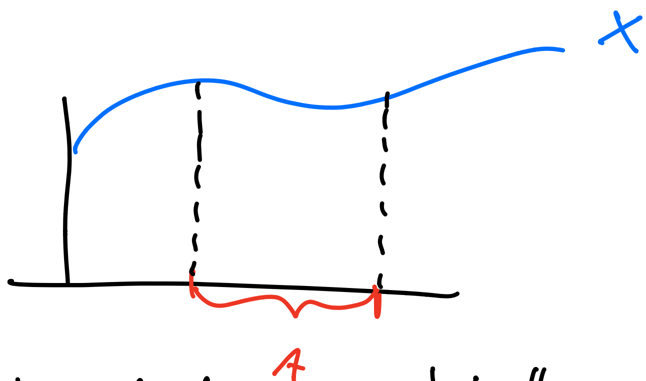
Let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{F}_0$. (Intuition: $\mathcal{F}_0$ will be the σ-alg. generated by the variable Y we want to condition on).

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 $E(X \mid \mathcal{F})$ is not clear a priori, it needs to be proved.

- $E(X \mid Y) := E(X \mid \sigma(Y))$

Examples (extreme cases)

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

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

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

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