Probability measure

1. **Given space** \( \Omega \) ("abstract space").

2. Need a \( \sigma \)-algebra \( \mathcal{F} \) on \( \Omega \) ("measurable events")
   - \( A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F} \)
   - \( (A_i)_{i \in \mathbb{N}} \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F} \) ("countable unions")
   - \( \emptyset, \Omega \in \mathcal{F} \)
   - countable intersections

3. A measure \( \mu \) on \( (\Omega, \mathcal{F}) \) is a function
   \[ \mu : \mathcal{F} \rightarrow [0, \infty] \]
   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 \( \mu \) on a measurable space \( (\Omega, \mathcal{F}) \) is called a **probability measure** if \( \mu(\Omega) = 1 \).

The elements in \( \mathcal{F} \) are called **events**.

Then \( (\Omega, \mathcal{F}, \mu) \) is called a **probability space**.
Example (1): Throw one die

\[ \mathcal{S} = \{1, 2, \ldots, 6\}, \quad \mathcal{A} = \mathcal{P}(\mathcal{S}) \] (\sigma\text{-algebra generated by the "elementary events" } \{1\}, \{2\}, \ldots, \{6\}).

\( \mathcal{P} \) can be defined uniquely by assigning

\[ \mathcal{P}(\{1\}) = \mathcal{P}(\{2\}) = \ldots = \mathcal{P}(\{6\}) = \frac{1}{6} \]

For example

\[ \mathcal{P}(\{1, 5\}) = \mathcal{P}(\{1\}) + \mathcal{P}(\{5\}) = \frac{1}{3} \]

Throw two dice:

\[ \mathcal{S} = \{1, 2, \ldots, 6\} \times \{1, 2, \ldots, 6\} \]

\[ \mathcal{A} = \mathcal{P}(\mathcal{S}) \]

\[ \mathcal{P}(\{(1, 3)\}) = \frac{1}{36} \]

Example (2): Normal distribution

\( \mathcal{S} = \mathbb{R} \)

\( \mathcal{A} = \sigma\text{-algebra} \)

\[ f_{\mu_1, \sigma_1} : \mathbb{R} \to \mathbb{R}, \quad x \mapsto \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \]

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

\[ \mathcal{P}(A) := \int_A f_{\mu_1, \sigma_1}(x) \, dx \]
Different types of probability measures

**Discrete measure:**

$\mathcal{S} = \{x_1, x_2, \ldots \}$ finite or at most countable.

$\mathcal{A} = 2^\mathcal{S}$

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

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

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

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

$$P(A) = \sum_{\{i | x_i \in A\}} p_i$$

Examples: tossing a coin, distribution on $\mathbb{R}$

**Dirac measure:**

For $x \in \mathbb{R}$, we define the Dirac measure $\delta_x$ 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 $x$. 

\[ \cdot \]

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

**Measures with a density**

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

Consider a function \(f: \mathbb{R}^n \to \mathbb{R}_{\geq 0}\) that is measurable and satisfies \(\int f \, d\lambda = 1\). \((= \int f \, \text{vol} \, dx)\)

Then we define a measure \(\nu\) on \(\mathbb{R}^n\) by setting, for all \(A \in \mathcal{B}^n\),

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

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

Notation: \(\nu = f \cdot \lambda\)

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

Counterexample: \(\delta_0\) Dirac measure
Def. A prob. measure $\nu$ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is called absolutely continuous with respect to another measure $\mu$ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ if every $\mu$-null set is also a $\nu$-null set:

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

Notation: $\nu \ll \mu$

\[ \mu(A) = 0 \implies \int_A d\mu = 0 \]

\[ \nu(A) \]

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

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

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

Theorem (Radon-Nikodym)

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

(See next page)
\( (1) \) \( \nu \) has a density w.r.t \( \mu \).

\( (2) \) \( \nu \) is absolutely continuous w.r.t \( \mu \).

Proof idea

\((1) \Rightarrow (2)\) easy

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

Consider the set \( G \) of all functions \( g \) with the following properties:

- \( g \) is measurable, \( g \geq 0 \)
- \( g \cdot \mu \leq \nu \), that is
  \[ \forall A \in \mathcal{A}(\mathbb{R}): \int g \, d\mu \leq \nu(A). \]

- \( \emptyset \) obviously \( g = 0 \) satisfies \((\star)\), so \( G \) is not empty.
- If \( g, h \) both satisfy \((\star)\), then \( \sup(g, h) \) satisfies \((\star)\).
- Define \( \gamma := \sup_{g \in G} \int g \, d\mu \) and construct a sequence \((g_n)_{n \in \mathbb{N}}\) such that \( \lim \int g_n \, d\mu = \mu \).
- Define "density" \( f := \sup g_n \)
- Now prove: \( f \) does the job.
**Definition**

A probability measure \( \mu \) on \((\Omega, \mathcal{B})\) is called **singular** wrt \( \nu \) if there exists \( A \in \mathcal{B} \) such that

\[
\mu(A) = 0 \quad \text{but} \quad \nu(A^c) > 0.
\]

Notation: \( \mu \perp \nu \).

**Example:** \( \lambda \perp \delta_0 \)

---

**Theorem (Decomposition by Lebesgue)**

\( \mu, \nu \) proto-measures on \((\Omega, \mathcal{B})\). Then there exists a unique decomposition \( \nu = \nu_1 + \nu_2 \) such that

\( \nu_1 \ll \mu \) and \( \nu_2 \perp \mu \).

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

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

---

**Proof**

Let \( \mathcal{M} \) be the set of all null-sets wrt \( \mu \). Let

\[
\alpha := \sup \{ \nu(A) \mid A \in \mathcal{M} \}.
\]

Can construct a countable sequence \((A_n)_{n \in \mathbb{N}}\) such that \( A_n \in \mathcal{M} \), \( \alpha = \sum \nu(A_n) \).
such that $\nu(\mathcal{A}_n) \to \omega$. By countable additivity we get

$$\nu\left(\bigcup_{m \in \mathbb{N}} \mathcal{A}_m\right) = \omega.$$ 

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

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

Don the job.

**Construction of the Cantor set:**

- Start with $C_0 := [0, 1]$  
  "Remove middle part"

- $C_1 := [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$  
  "Remove middle part from all intervals"

- $C_2 = \ldots$

The Cantor set is the limit in this process.
Now construct a probability distribution:

Consider the cdf's of the sets $C_0, C_1, C_2, \ldots$

$C_0 : \quad \begin{array}{c|c}
0 & 1 \\
\hline
& \\
\end{array} \\
\text{uniform on } [0, 1]

$C_1 : \quad \begin{array}{c|c}
0 & 1 \\
\hline
\hline
\frac{1}{3} & 1 \\
\frac{2}{3} & 1 \\
\end{array} \\
\text{uniform on } [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]

$C_2 : \quad \begin{array}{c|c}
0 & 1 \\
\hline
\hline
\frac{1}{3} & 1 \\
\frac{2}{3} & 1 \\
\frac{1}{2} & 1 \\
\frac{5}{6} & 1 \\
\frac{2}{3} & 1 \\
\frac{1}{2} & 1 \\
\frac{5}{6} & 1 \\
1 & 1 \\
\end{array} \\
\text{uniform on } [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]

Take limit. Can prove many strange properties:

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

$\Rightarrow \lambda \perp r$
Cumulative distribution function

Let $P$ be a probability 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 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:
If $(x_n)_n$ a sequence with $x_n \not\to x$ (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 $\mathbb{P}$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that

$$
P(-\infty, x] = F(x).$$
**Random variable**

**Def.** Let \((\Omega, \mathcal{F}, \mathbb{P})\) be a probability space, \((\tilde{\Omega}, \tilde{\mathcal{F}})\) be another measurable space. A mapping: \(X: \Omega \rightarrow \tilde{\Omega}\) is called a random variable if \(X\) is measurable, i.e.

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

**Example: sum of two dice**

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

\(\mathcal{F} = 2^\Omega\)

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

\(X\) "sum of the two values"

\(X: \Omega \rightarrow \{2, \ldots, 12\}, \ (i, j) \mapsto i + j\)

is measurable.
A random variable $X: \Omega \to \tilde{\Omega}$ induces a measure on the target space.

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

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

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

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

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

is a $\sigma$-algebra on $\Omega$ and it is called the $\sigma$-algebra induced by $X$ (it is the smallest $\sigma$-algebra on $\Omega$ that makes $X$ measurable).
Conditional probabilities

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

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

Def. \((\Omega, \mathcal{A}, P)\) probability space, 
if \(A, B \in \mathcal{A}, P(B) > 0\). Then

\[
P(A | B) = \frac{P(A \cap B)}{P(B)}
\]

is called the conditional probability of \(A\) given \(B\).

Theorem. The mapping \(P_B: \mathcal{A} \rightarrow [0,1], A \rightarrow P_A|B\) is a probability measure on \((\Omega, \mathcal{A})\); it is called the conditional distribution of \(P\) with respect to \(B\).
Example:
\[ \Omega = \text{all persons on earth} \]
\[ A = \mathcal{B}(\Omega) \]
\[ \rho = \text{"uniform"} \]

Event \( A : = \text{"person has been vaccinated"} \)
Event \( B : = \text{"person has disease"} \)

\[ p(\text{disease} \mid \text{vaccinated}) \]

\[ p(\text{vaccinated} \mid \text{disease}) \]

Example: two dice
\[ p(\text{"sum is 10"} \mid \text{"first die was 5"}) \]
Bayes formula

Law of total probability: Let $B_1, B_2, \ldots, B_n$ be a disjoint partition of $\Omega$ with $B_i \in \mathcal{A}$ for all $i$, and $A \in \mathcal{A}$. Then

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

Bayes formula:

$$p(B_i | A) = \frac{p(A | B_i) \cdot p(B_i)}{\sum_{i} p(A | B_i) \cdot p(B_i)} = \frac{p(A \cap B_i)}{p(A)}$$

Example: breast cancer screening

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

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

80% of women without breast cancer will receive a positive result as well. ("false positives")

Given that a woman verso receives a positive test result, what is the likelihood that she has breast cancer?
\[ P(\text{cancer} \mid \text{positive}) = \frac{P(\text{positive} \mid \text{cancer}) \cdot P(\text{cancer})}{P(\text{pos.} \mid \text{cancer}) P(\text{cancer}) + P(\text{pos.} \mid \text{not cancer}) \cdot P(\text{not cancer})} \]

\[
\approx \frac{0.9 \cdot 0.01}{0.9 \cdot 0.01 + 0.09 \cdot 0.99} \approx 10 \%
\]
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 \(B \iff P(A|B) = P(A)\)

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 \(A_i, A_j \in I: P(A_i \cap A_j) = P(A_i) \cdot P(A_j)\). This does not imply independence!)

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

\[ A \perp B \]
\[ X \perp Y \]
Expectation (discrete case)

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

**Definition** $(\Omega, \mathcal{A}, \mathbb{P})$ 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

$$E(X) := \sum_{r \in S} r \cdot P(X=r)$$

is called the expectation of $X$.

(sometimes people write $EX$, $E X$, or $E(X)$).

**Examples**

- Toss a coin. $\Omega = \{\text{head, tail}\}$, $\mathcal{A} = 2^\Omega$, $P(\text{head}) = p$, $P(\text{tail}) = 1-p$.

  $0 < p < 1$.

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

  $E(X) = 0 \cdot P(X=0) + 1 \cdot P(X=1) = p$.

- Test error of a classifier.

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

- **Linear:** \( E ( a \cdot X + b \cdot Y ) = a \cdot E(X) + b \cdot E(Y) \), for \( \alpha \in \mathbb{R} \) and \( \beta \in \mathbb{R} \).

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

\[
\sum_{i,j} |x_i \cdot y_j| \cdot P(X=x_i, Y=y_j) = \sum_{i,j} |x_i \cdot y_j| \cdot P(X=x_i) \cdot P(Y=y_j) \\
= \sum_{i,j} |x_i \cdot y_j| \cdot \frac{P(X=x_i)}{\left| \frac{x_i}{y_j} \right|} \\
= \left( \sum_{i} |x_i| \cdot P(X=x_i) \right) \cdot \left( \sum_{j} |y_j| \cdot P(Y=y_j) \right)
\]
Variance, covariance, correlation
(discrete case)

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

Then \(\text{Var}(X) := E((X - E(X))^2)\)

is called the variance of \(X\)

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

is called the standard deviation.

\(\text{Cov}(X, Y) := E((X - E(X)) \cdot (Y - E(Y)))\)

is called the covariance of \(X\) and \(Y\).

\(\rho_{X,Y} := \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y} \in [-1, 1]\)

is called the correlation coefficient.

If \(\text{Cov}(X,Y) = 0\), then \(X\) and \(Y\) are called uncorrelated.

More generally, for \(k \in \mathbb{N}\) we define the formulas
\(E(X^k)\) ("\(k\)-th moment"),
\(E((X - E(X))^k)\) ("\(k\)-th centered moment").
Intuition about covariance

\[ \text{Cov}(X, Y) = \mathbb{E}( (X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y)) ) \]

Positive, large covariance \( g \approx 0.9 \)

Negative covariance, large in absolute values \( g \approx -0.9 \)

\[ \text{Cov} \approx 0 \] (uncorrelated)
Properties

- \( \text{Var} (X) = E(X^2) - (E(X))^2 \)
- \( \text{Cov}(X, Y) = E(XY) - E(X) \cdot E(Y) \)
- \( E(aX + b) = a \cdot E(X) + b \)
- \( \text{Var}(aX + b) = a^2 \text{Var}(X) \)
- \( \text{Cov}(X, Y) = \text{Cov}(Y, X) \)
- \( \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + \text{Cov}(X, Y) \)
- \( X, Y \text{ independent } \iff \text{Cov}(X, Y) = 0 \)
- \( X, Y \text{ independent } \iff \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \)
Expectation and variance in the general setting

\[ \mathbb{L}^k(\Omega, \mathcal{A}, \rho) := \left\{ X: \Omega \to \mathbb{R} \mid X \text{ measurable and} \int_{\Omega} |X|^k \, d\rho < \infty \right\} \]

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

\[
E(X) := \int_{\Omega} X \, d\rho = \int_{\mathbb{R}} x \, d\rho_X(x) \\
\text{(can of density}\ f) \quad \int_{\mathbb{R}} x \, f(x) \, dx
\]

If \(X^k \in \mathbb{L}^k(\Omega, \mathcal{A}, \rho)\) then

\[ E(X^k) = \int_{\mathbb{R}} X^k \, d\rho \] is called the \(k\)-th moment of \(X\).

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

\[
\text{Var}(X) = E((X - E(X))^2) \\
\text{Cov}(X,Y) = E((X - E(X))(Y - E(Y)))
\]
Chebyshev inequality: \( \varepsilon > 0, X \in L^2(\Omega, \mathcal{A}, \mathbb{P}) \). Then:
\[
\mathbb{P}(|X - \mathbb{E}(X)| > \varepsilon) \leq \frac{\text{Var}(X)}{\varepsilon^2}
\]

Markov inequality: \( \varepsilon > 0, f: \mathbb{E}_{0, \infty} \rightarrow \mathbb{E}_{0, \infty}, f \text{ monotonically increasing} \). Then
\[
\mathbb{P}(1_{|Y| > \varepsilon}) \leq \frac{\mathbb{E}(f(1_{|Y|}))}{f(\varepsilon)}
\]
In particular,
\[
\mathbb{P}(1_{|Y| > \varepsilon}) \leq \frac{\mathbb{E}(1_{|Y|})}{\varepsilon}
\]

Cachy-Schwarz inequality: \( X, Y \in L^2(\Omega, \mathcal{A}, \mathbb{P}) \). Then:
\[
\mathbb{E}(|X \cdot Y|^2) \leq \mathbb{E}(X^2) \cdot \mathbb{E}(Y^2)
\]
Examples of probability distributions

**Discrete distributions**

- **Uniform distribution on \([1, \ldots, n] \):** \( P(\{i\}) = \frac{1}{n} \)  

- **Binomial distribution on \([0, \ldots, n]\)  
  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 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  
  \[ f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \]
Normal distribution on $\mathbb{R}$

Density: parameters $\mu$ (mean), $\sigma$ (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

\[ X: \mathcal{X} \to \mathbb{R}^n, \quad X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \mu_i \in \mathbb{E}(X_i), \quad \mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} \]

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

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

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

Prop. \( \Sigma \) is positive definite and symmetric.

Consequence: \( \Sigma \) has real-valued, non-negative eigenvalues.

Contour lines of \( f_{\mu, \Sigma} \)

If \( X_1, \ldots, X_n \) are independent \( \Leftrightarrow \) \( \Sigma = \begin{pmatrix} \delta_{11} & \delta_{12} & \cdots & 0 \\ \delta_{21} & \delta_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \delta_{nn} \end{pmatrix} \)
\[ X \sim N(\mu_1, \Sigma_1), \quad Y \sim N(\mu_2, \Sigma_2), \quad \text{independent, then} \]
\[ X + Y \sim N(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2) \]

**Mixture of Gaussians**

Consider \( \pi_1, \pi_2, \ldots, \pi_k \) with \( 0 \leq \pi_i \leq 1 \) and \( \sum \pi_i = 1 \)

Consider the following density:

\[
f(x) = \sum_{i=1}^{k} \pi_i \cdot f_{\mu_i, \Sigma_i}(x)
\]
Consider in $X_i: \Omega \to \mathbb{R}$, $i \in \mathbb{N}$, $X: \Omega \to \mathbb{R}$, $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space.

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

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

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

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

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

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

Case (1):

$$\lim_{i \to \infty} X_i(\omega) = X(\omega)$$

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

So we get:

...
\[
\{ w \mid X_i(w) \to X(w) \} = \\
\bigcap_{k \in \mathbb{N}} \bigcup_{n \in \mathbb{N}} \bigcap_{n \geq N} \{ w \mid |X_n(w) - X(w)| < \frac{1}{k} \} \in \mathcal{F}
\]

\text{countable union and intersection}

\text{X_n, X are measurable} \implies \text{|X_n - X| is measurable}

\text{so } \{ \ldots \} \in \mathcal{F}

(\text{3}) \quad X_n \Rightarrow X \text{ in } L^p \quad \text{("in the } p \text{-th mean")}: \iff \\
X_n, X \in L^p \text{ and } \|X_i - X\|_p \to 0.

(\text{4}) \quad \text{Let } M^1(\mathbb{R}^n) \text{ be the set of all probability measures on } (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)). \text{ Assume } (\mu_n)_n \subset M^1(\mathbb{R}^n), \mu \in M^1(\mathbb{R}^n).

C_b(\mathbb{R}^n) := \text{space of bounded continuous functions}.

\mu_n \to \mu \text{ weakly } \iff \\
\forall f \in C_b(\mathbb{R}^n) : \int f \, d\mu_n \to \int f \, d\mu

\text{Diagram:}

\mu_n \quad \text{to} \quad \mu
In functional analysis, a sequence \((x_n)_n\) in a Banach space \(B\) converges weakly if for all bounded linear functionals \(f\), we have that \(f(x_n) \to f(x)\). (i.e. for all \(f \in B^*\)).

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

The dual space of \(M^n(\mathbb{R}^n)\) is \(C_0(\mathbb{R}^n)\).

\[
(5) \quad X_i, X : (\Omega, \Delta, \mathbb{P}) \to \mathbb{R}^n. \text{ The sequence } X_n \text{ converges in distribution to } X : \iff \\
\text{ the distributions } P_{X_n} \text{ converge to } P_X \text{ weakly.}
\]

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

\[
\begin{align*}
\text{almost surely} & \quad \iff \quad \text{in } L^1 \\
\text{in probability} & \quad \iff \quad \text{in } L^p (p > 1) \\
\text{in distribution} & \quad \iff \quad \text{in } L^p (p > 1)
\end{align*}
\]
Example (convergence a.s., in prob., but not in $L^1$)

$$X_n : \mathbb{R} \to \mathbb{R}, \quad X_n(x) = \left\{ \begin{array}{ll}
n & \text{for } 0 \leq x \leq \frac{1}{n} \\
0 & \text{otherwise} \end{array} \right.$$  

$$\lim_{n \to \infty} X_n = 0 : \ X_n(x) \to 0.$$  

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"

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

$$f_2 = \mathcal{M}_{[0,\frac{1}{2}]} , \quad f_3 = \mathcal{M}_{[\frac{1}{2},1]}$$

$$f_4 = \mathcal{M}_{[0,\frac{1}{3}]} , \quad f_5 = \mathcal{M}_{[\frac{1}{3},\frac{2}{3}]} , \quad f_6 = \mathcal{M}_{[\frac{2}{3},1]}$$
Example (Conv. in distribution, but not in prob.)

\[ X_n : [0, 1] \to \mathbb{R} \] \( \\xRightarrow{\text{d}} \) \[ \mathcal{D}[0, 1/2] \]

\[ X \sim \mathcal{U}[\frac{1}{2}, 1] \]

Obviously \( X_n \xrightarrow{\text{d}} X \) in prob., but:

\[ p_{X_n} = \frac{1}{2}(\delta_0 + \delta_1) = p_{X_2} = p_{X_3} = \ldots = p_X \]

so \( X_n \xrightarrow{\text{d}} X \) in distribution.
Theorem of Borel–Cantelli

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

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

$$= P(\{w \in \mathcal{S} \mid w \in A_n \text{ for infinitely many } n\})$$

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

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

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

**Proof intuition:**

$$\{ \lim X_n = X \}$$

$$= \{ \forall k : |X_n - X| > \frac{\varepsilon}{k} \text{ at most finitely often} \}$$

$$= \bigcap_{k \in \mathbb{N}} \{ |X_n - X| > \frac{\varepsilon}{k} \text{ at most fin. often} \}$$

$$= \left( \bigcup_{k \in \mathbb{N}} \{ |X_n - X| > \frac{\varepsilon}{k} \text{ inf. often} \} \right)^{\text{complement}}$$
Theorem: Consider a sequence of events \( (A_n) \subseteq \Omega \).

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 \( (A_n) \) are independent, then \( P(A_n \text{ i.o.}) = 1 \).

Application in Learning Theory:

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

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

Borel-Cantelli to prove that

\( P( |X_n - \xi| > \frac{1}{n} \text{ i.o.} ) = 0 \),

Thus \( X_n \rightarrow \xi \text{ a.s.} \).
Limit Theorems: LLN and CLT

**Strong Law of Large Numbers**

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

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

**Remarks:**

- Many versions of this theorem exist, (slightly relaxing iid)
- "Strong law" $\Rightarrow$ convergence a.s.
- "Weak law" $\Rightarrow$ convergence in probability

**Central Limit Theorem**

\( (X_i)_{i=1}^\infty \text{ iid rv with mean } \mu_1 \text{ variance } \sigma_1^2 < \infty. \)

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

\[ Y_n := \frac{S_n - n\mu}{\sqrt{n}\sigma} \]

Then \( Y_n \to Y \) in distribution where \( Y \sim N(0,1) \).
Illustration: $X_i$: coin, head $\in \Lambda$, tail $\in O$

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

Motivation: random projections

Motivation: random projections

\[ \mathbb{R}^d \rightarrow \mathbb{R}^l \]

want to project in \( \mathbb{R}^d \rightarrow \mathbb{R}^l \) "small"

\[ \text{Theorem of Johnson-Lindenstrauss:} \]

Can guarantee (for certain parameters \( \varepsilon, k \))

\[
(n - \varepsilon) \| x_i - x_j \|_{\mathbb{R}^d} \leq \| \pi(x_i) - \pi(x_j) \|_{\mathbb{R}^l} \leq (n + \varepsilon) \| x_i - x_j \|_{\mathbb{R}^d}
\]

Construction / proof steps:

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

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

(b) \( \mathbb{P}(\| \pi(x_i) - \pi(x_j) \|_{\mathbb{R}^l} - \mathbb{E}(\cdots) > \varepsilon) ? \)
**Hoeffding inequality**

**Theorem (Hoeffding):** \( X_1, \ldots, X_n \) rv, independent, assume that \( X_i \in \mathcal{C}(a_i, b_i) \) a.s. for \( i = 1, \ldots, n \).

Let \( S_n = \sum_{i=1}^{n} (X_i - \mathbb{E}(X_i)) \). 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.i.d rv, a \leq X_i \leq b, let \( K \) have the same distribution as \( X_i \). Then:

\[
\frac{1}{n} \sum_{i=0}^{n} X_i \to \mathbb{E}(X) \text{ a.s.}
\]

**Proof.** Hoeffding \( \Rightarrow \)

- \( P \left( \frac{1}{n} \sum X_i - \mathbb{E}(X) > t \right) \leq \exp \left( -\frac{2nt^2}{(b-a)^2} \right) \)
- \( P \left( \frac{1}{n} \sum X_i - \mathbb{E}(X) < -t \right) \)
\[ p \left( \frac{1}{n} \sum (-x_i) - E(x) \geq t \right) \leq \exp \left( -\frac{2nt^2}{(b-a)^2} \right) \]

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:

\[ Z_n := \frac{1}{n} \sum_{i=1}^{n} x_i \]

\[ \sum_{n=0}^{\infty} p (Z_n - E(x) > t) \leq 2 \sum_{n=0}^{\infty} \exp \left( -\frac{2nt^2}{(b-a)^2} \right) \leq \infty \]

\[ \text{Sum} \]

\[ \text{Substitute: } r := \exp \left( -\frac{2nt^2}{(b-a)^2} \right) \in [0, 1] \]

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

\[ \text{Sum} = 2 \sum_{n=0}^{\infty} r^n = 2 \cdot \frac{1}{1-r} < \infty. \]

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 \( \Rightarrow \) need other inequalities.
Bernstein inequality

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

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

**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 \( c_1, \ldots, c_n \) such that

\[
\sup_{x_1, \ldots, x_n} \left| f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, \frac{X_i}{\Lambda}, x_{i+1}, \ldots, x_n) \right| \leq c_i
\]

**Example:** \( f(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i \), and \( a \leq x_i \leq b \) for \( i \), then \( f \) satisfies \( \oplus \) with \( c_i = b - a \).
Theorem (Mc Diarmid) \( X_1, \ldots, X_n \) independent rv, \( X_i \in \mathcal{X}_i \), \( f: X_1 \times \cdots \times X_n \to \mathbb{R} \) function with bounded difference property.

Then, for any \( t > 0 \),

\[
P \left( f(X_1, \ldots, X_n) - E(f(X_1, \ldots, X_n)) \geq t \right) \leq \exp \left( - \frac{2t^2}{\sum_{i=1}^{n} c_i^2} \right)
\]

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

\[
A = \begin{pmatrix}
X_{11} & \cdots & \cdots \\
\vdots & \ddots & \vdots \\
\cdots & \cdots & X_{nn}
\end{pmatrix} \sim \text{draw iid}
\]
Glivenko–Cantelli Theorem

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

\( X_1, \ldots, X_n \sim F, \text{iid} \)

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

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

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

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

Because \( I\{X_i \leq a_0\} \) is a Binomial rv with

\[ p = P(X_i \leq a_0). \]

So it is clear that \( F_n \rightarrow F \) pointwise (i.e. \( \forall a_0 \))

Now let's look at uniform convergence.

Theorem \( X_1, \ldots, X_n \text{ iid random variables with cdf } F. \)

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, \( \sup_{a} | F_n - F | \to 0 \text{ a.s.}, \)

i.e. \( F_n \to F \text{ uniformly a.s.} \)
Proof: Observe: \( P(\sup_{a \in R} | F_u(a) - F(a) | > \varepsilon) \to 0 \) for any fixed \( a_0 \).

Problem: need to look at

\[
P(\sup_{a \in R} | F_u(a) - F(a) | > \varepsilon)\]

difficult because \( R \) is uncountable

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

\[
P(\max_{i=1}^n | u_i | > \varepsilon) = \]

\[
= P(|u_1| > \varepsilon \text{ or } |u_2| > \varepsilon \text{ or } \ldots \text{ or } |u_n| > \varepsilon) = \sum_{i=1}^n P(|u_i| > \varepsilon)
\]

Trick of the proof: cannot sup over \( a \in R \) to something "finite".

How could we achieve this?

\[
| \text{red - green} | 
\leq 2 | \text{green - blue} |
\]
Step 1: Symmetrization by ghost sample
Assume $X_1', \ldots, X_n' \sim \mathcal{F}$ independently ("ghost sample"),
Denote by $F_n'$ the empirical cdf induced by ghost sample
Now it is easy to prove:

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

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

Step 2: Want to split this in two terms

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

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

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

Distribution of $\sigma_i$ is the same as the cdf of the following:

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

Now we can:
\[
2 \mathbb{P} \left( \sup_a \left( E(u) - E_u'(a) \right) \right) \\
= 2 \mathbb{P} \left( \sup_a \left( \frac{1}{n} \sum_{i=1}^{n} \xi_i \Delta \lambda_{x_i \in a} - \Delta \lambda_{x_i \in a} \right) \left| \frac{1}{n} \sum_{i=1}^{n} \xi_i \right| > \frac{\varepsilon}{2} \right) \\
\leq 2 \mathbb{P} \left( \sup_a \left( \frac{1}{n} \sum_{i=1}^{n} \xi_i \Delta \lambda_{x_i \in a} \right) > \frac{\varepsilon}{4} \right) + 2 \mathbb{P} \left( \sup_a \left( \frac{1}{n} \sum_{i=1}^{n} \xi_i \Delta \lambda_{x_i \in a} \right) \left| \frac{1}{n} \sum_{i=1}^{n} \xi_i \right| > \frac{\varepsilon}{4} \right)
\]

\text{Observe:}
\[
\mathbb{P} \left( \left| \mu - \nu \right| > \frac{\varepsilon}{2} \right) \leq \mathbb{P} \left( \left| \mu \right| > \frac{\varepsilon}{4} \right) \text{ or } \mathbb{P} \left( \left| \nu \right| > \frac{\varepsilon}{4} \right)
\]

\[
\leq 4 \cdot \mathbb{P} \left( \sup_a \left( \frac{1}{n} \sum_{i=1}^{n} \xi_i \Delta \lambda_{x_i \in a} \right) > \frac{\varepsilon}{4} \right)
\]

\textbf{Step 3:}

\textbf{Exploit "finiteness structure":}

\textbf{Fix } X_{k_1}, \ldots, X_k \textbf{ (i.e. condition on } X_1, \ldots, X_k) \textbf{.}

\textbf{We look at } \Delta \lambda_{x_i \in a} \textbf{.}

\textbf{The rest } \Delta \lambda_{x_i \in a}, \ldots, \Delta \lambda_{x_k \in a} \textbf{. For fixed } a \textbf{, one can only have } a \textbf{-realizable}

\[
\mathbb{P} \left( \sup_a \left( \frac{1}{n} \sum_{i=1}^{n} \xi_i \Delta \lambda_{x_i \in a} \right) \left| \frac{1}{n} \sum_{i=1}^{n} \xi_i \right| > \frac{\varepsilon}{4} \ \left| X_1, \ldots, X_k \right) \right. 
\]

\[
\leq \left( n+1 \right) \sup_a \mathbb{P} \left( \frac{1}{n} \sum_{i=1}^{n} \xi_i \Delta \lambda_{x_i \in a} \left| \frac{1}{n} \sum_{i=1}^{n} \xi_i \right| > \frac{\varepsilon}{4} \ \left| X_1, \ldots, X_k \right) \right.
\]

\[
\text{use Hoeffding (44)}
\]
Step 4: Apply Hoeffding to (1.5):

$k_\alpha:
\Pr \left( \frac{1}{n} \sum_{i=1}^{n} \epsilon_i < \epsilon \mid \epsilon_1, \ldots, \epsilon_n \right) \leq 2 \exp \left( -\frac{3n}{\varepsilon^2} \right)

Combining everything gives the theorem.
Product space, joint distributions

Consider two measurable spaces \((\Omega_1, \mathcal{A}_1, P_1), (\Omega_2, \mathcal{A}_2, P_2)\).

Define the product space \((\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)\) with
\[
\Omega_1 \times \Omega_2 = \{(w_1, w_2) \mid w_1 \in \Omega_1, w_2 \in \Omega_2\}
\]
\[
\mathcal{A}_1 \otimes \mathcal{A}_2 = \{A_1 \times A_2 \mid A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2\}.
\]

Consider two rvs \(X_1 : (\Omega_1, \mathcal{A}_1, P_1) \to (\Omega_1, \mathcal{A}_1)\),
\(X_2 : (\Omega_1, \mathcal{A}_1, P_1) \to (\Omega_2, \mathcal{A}_2)\).

\(X : (X_1, X_2) : (\Omega_1, \mathcal{A}_1, P_1) \to (\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)\)
\[
(X_1, X_2)(\omega) = (X_1(\omega), X_2(\omega)).
\]

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

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

\[P(X_1, X_2) = P_1 \otimes P_2.\]
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 $(\mathbb{R}_1, \mathcal{B}(\mathbb{R}))$, namely $P_{X_1}$. Similarly for $P_{X_2}$.

Example in the discrete case:

<table>
<thead>
<tr>
<th>$Y$ \ $X$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$\Sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_1$</td>
<td>$p_1$</td>
<td>$p_2$</td>
<td>$p_3$</td>
<td>$p_1 + p_2 + p_3 = P(Y = y_1)$</td>
</tr>
<tr>
<td>$y_2$</td>
<td>$p_4$</td>
<td>$p_5$</td>
<td>$p_6$</td>
<td>$\sum$ marginal distribution wrt $Y$.</td>
</tr>
</tbody>
</table>

Marginal distributions in case of densities

$X, Y : (\mathbb{R}, \mathcal{B}(\mathbb{R}), P) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, $Z := (X, Y)$. 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

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

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

(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 $X = (X_1, X_2)$ with mean

$$\mu = (\mu_1, \mu_2) \in \mathbb{R}^2$$

and cov. $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$.

Then the marginal distribution of $X$ with $X_1$ is again
a normal distribution with mean $\mu_1$ and var $\sigma_1^2$.

\[ x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n. \] Group the variables: \[ x_i^{(k)} \} x \in \mathbb{R}^k \]

Want to look at the marginal of $X$ wrt $x^k$.

\[ p = \begin{pmatrix} \mu^k \\ \mu_n \end{pmatrix} \text{ mean } \quad \tilde{\mu} = \begin{pmatrix} \mu^k \\ \mu_n \end{pmatrix}, \quad \mu^k = \begin{pmatrix} \mu^k_{x^k} \\ \mu_n \end{pmatrix} \]

\[ \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \] \[ \tilde{\Sigma} = \begin{pmatrix} \Sigma_{11} \\ \Sigma_{21} \end{pmatrix} \]

Now the marginal of $X$ wrt $x^k$ is a normal dist. on $\mathbb{R}^k$ with mean $\tilde{\mu}$ and cov. $\tilde{\Sigma}_{11}$. 
**Conditional distributions**

**Direct case:**

Know conditional probabilities: \( P(A | B) \)
defined for events \( A, B \subseteq \Omega \), and \( P(B) > 0 \).

Let \( X, Y : (\Omega, A, \mathbb{P}) \to \mathbb{R} \) be discrete \( X, Y \in \mathbb{R} \) such that
\( P(Y = y) > 0 \). Then we can define the conditional probability measure
\[
P_{X | Y = y} : A \to P(X \in A | Y = y).
\]
This is a probability measure.

For general \( X, Y \) this is surprisingly complicated!

\( \Rightarrow \) "regular conditional probabilities" are specified.

**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

$$\mu = \left( \begin{array}{c} \mu_1 \\ \vdots \\ \mu_n \end{array} \right) \quad \Sigma = \left( \begin{array}{cc} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{array} \right)$$

If $X = \left( \begin{array}{c} X_1 \\ \vdots \\ X_n \end{array} \right) \sim N(\mu, \Sigma)$, then the conditional distribution of $\tilde{X} = \left( \begin{array}{c} X_n \end{array} \right)$ w.r.t $x^u = \left( \begin{array}{c} x_{1:n} \end{array} \right)$ is given by

$$p_{\tilde{X} \mid X^u} \sim N\left( \mu_n + \Sigma_{1n} \Sigma_{22}^{-1} (x^u - \tilde{\mu}), \Sigma_{22} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{12} \right).$$
\textbf{Conditional expectation}

\textbf{Definition (discrete case)} \( X, Y : (\mathcal{X}, \mathcal{A}, \mathbb{P}) \rightarrow \mathbb{R} \)

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

\[
E(Y \mid X = x_i) := \sum_{j=1}^{m} y_j \cdot P(Y = y_j \mid X = x_i)
\]

\textit{well defined}

\textit{Example:} two dice, \( X \) = first one, \( Y \) = second one, independent

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

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

\[
= \sum_{k=1}^{6} (i + k) \cdot \frac{1}{6} = \frac{1}{6} \sum_{k=1}^{6} (i + k) = 4.5
\]

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 \( \mathbb{R} \)-valued random variable:

\[
E(Y \mid X) : (\Omega, \mathcal{A}, \mathbb{P}) \rightarrow (\mathbb{R}, \mathcal{B})
\]

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

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

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

$E(Y|X)$ is only defined a.s.

Now we want to move to the more general case.

Sketch: $X$ continuous rv

$Y$ discrete rv $\sim Y_1, \ldots, Y_5$

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

$$E(X|Y=y_i) = \sum_{i=1}^{5} E(X|Y=y_i) \cdot \mathbb{P}(\omega)$$

But need to make sure that it is measurable w.r.t $\sigma(Y)$. ("the clue")
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 $\mathcal{G}$ be a sub-$\sigma$-algebra of $\mathcal{F}_0$. (Intuitively $\mathcal{F}_0$ will be the $\sigma$-alg. generated by the variable $Y$ we want to condition on).

We now define the **conditional expectation** of $X$ given $\mathcal{G}$
$E(X | \mathcal{G})$ as any random variable $Z$ that satisfies

1. $Z$ is measurable w.r.t $\mathcal{G}$
2. For all $A \in \mathcal{G}$ we have

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

Existence of $E(X | \mathcal{G})$ is not clear a priori, it needs to be proved.

$E(X | Y) := E(X | \sigma(Y))$

Examples (two extreme cases)

- $X = Y$. Then $E(X | Y) = X$ (a.s.)
- $X \perp \perp Y$. $E(X | Y) = E(X)$ (a.s.)
Can of joint densities

$x, z : \mathbb{R} \rightarrow \mathbb{R}$ have a joint density $f(x, z)$.

Let $g : \mathbb{R} \rightarrow \mathbb{R}$ bounded, set $y := g(z)$. Assume we want to compute $E( Y \mid X ) = E(g(Z) \mid X)$.

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

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

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

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

$$E(Y \mid X) = h(x).$$