

## Part IV

# Probability theory

Recommended text book:

Jacod / Protter: Probability essentials.

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

Shiryaev: Probability

5 - Algebra



# Motivation

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

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

Ideally, would like to assign a "measure" to each subset of  $X$ .

Then the measure might be a mapping

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

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

## $\sigma$ -Algebra

Def  $X$  non-empty set. A  $\sigma$ -algebra on  $X$  is a non-empty collection  $\mathcal{F}$  of subsets of  $X$  such that

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

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

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

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

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

## Examples

- Trivial  $\sigma$ -algebra: Given  $X$ , we can define two (pretty useless)  $\sigma$ -algebras:

- $\mathcal{F}_1 = \{\emptyset, X\}$

- $\mathcal{F}_2 = \mathcal{P}(X) = \{A \mid A \subset X\}$ .

$$A \subset B \Leftrightarrow$$

$$a \in A \Rightarrow a \in B$$

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

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

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

## Borel- $\sigma$ -algebra

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

$\mathbb{Q}$  - Algebra

# Motivation

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

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

Ideally, would like to assign a "measure" to each subset of  $X$ .

Then the measure might be a mapping

$$m: \mathcal{P}(X) \rightarrow \mathbb{R}^+$$

power set, all subsets

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

## $\sigma$ -Algebra

Def  $X$  non-empty set. A  $\sigma$ -algebra on  $X$  is a non-empty collection  $\mathcal{F}$  of subsets of  $X$  such that

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

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

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

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

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

## Examples

- Trivial  $\sigma$ -algebra: Given  $X$ , we can define two (pretty useless)  $\sigma$ -algebras:

- $\mathcal{F}_1 = \{\emptyset, X\}$

- $\mathcal{F}_2 = \mathcal{P}(X) = \{A \mid A \subset X\}$ .

$$A \subset B \Leftrightarrow$$

$$a \in A \Rightarrow a \in B$$

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

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

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



## Borel- $\sigma$ -algebra

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

Measures
----------

# Measurable space

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

# Measure

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

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

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

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

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

# Measure space

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

# Example: Discrete measure

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

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

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

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

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

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

# Lebesgue measure on $\mathbb{R}$

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

To each such interval assign the volume

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

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

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

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

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

Can extend this consistently to the whole  $\sigma$ -algebra

$\leadsto$  multi-dim Lebesgue measure  $\lambda_d$



# A funny measure on $\mathbb{R}$

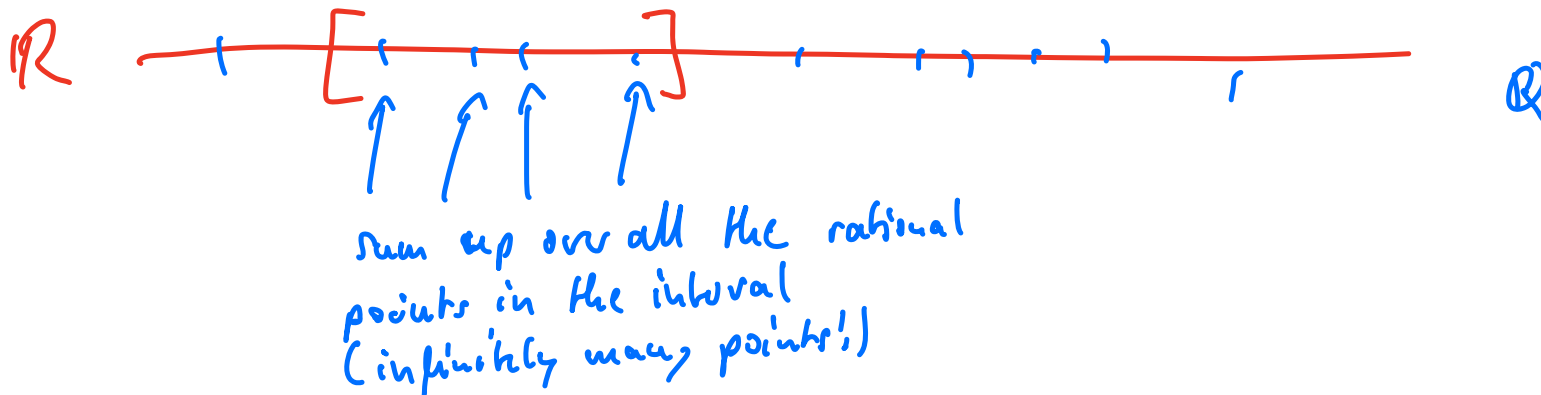
Let  $\mathcal{F}$  be the Borel- $\sigma$ -Algebra on  $\mathbb{R}$ . Want to define a measure that just assigns mass to rational numbers.

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

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

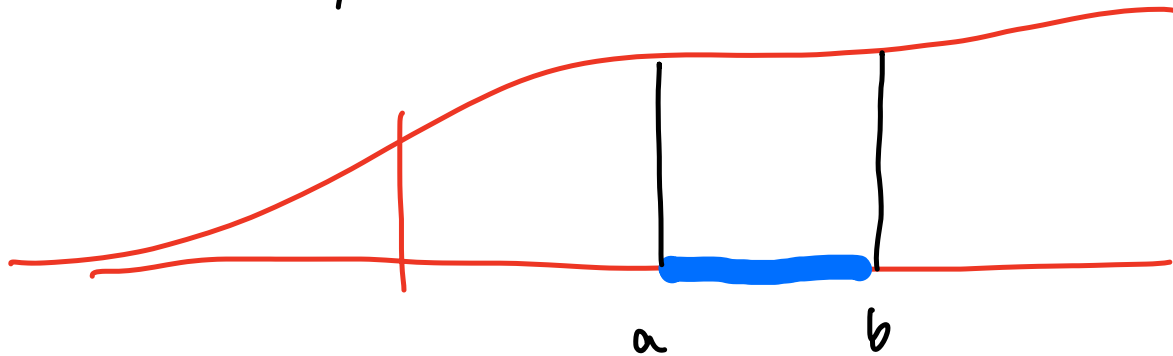
$$\mu(\{q_i\}) = m_i$$

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



## More general measures on $\mathbb{R}$

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



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

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

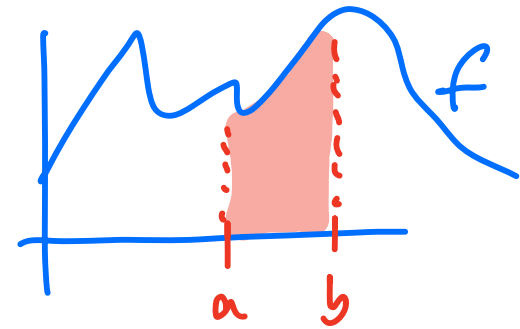
Can prove that one can extend this "measure" to the whole  $\sigma$ -algebra, making it a well-defined measure on  $\mathbb{R}$ .

# Measure with a density

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

For intervals  $[a, b]$  define

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



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

Then  $\gamma$  is the measure with density  $f$  with respect to the Lebesgue measure.

More keywords: Radon-Nikodym theorem

## Carathéodory extension theorem

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

Mathematical basis for this approach is the Carathéodory theorem.

(skipped here)

# Measures with continuous and discrete parts

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

Example: If  $\lambda$  is the Lebesgue measure on  $[0, 1]$  with Borel- $\sigma$ -algebra. Let  $\delta_0$  be the discrete measure that assigns mass 1 to the set  $\{0\}$ .

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

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

$\uparrow$   
Dirac measure

More keywords: Lebesgue decomposition theorem of measures

# Null sets

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

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

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

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

# Measurable mappings

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

A mapping  $f: X \rightarrow \mathcal{W}$  is called **measurable** if  
 $\forall A \in \mathcal{A}, f^{-1}(A) \in \mathcal{F}$  ("pre-images of measurable sets  
are measurable").

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

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

Measures
----------



# Measurable space

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

# Measure

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

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

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

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

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

# Measure space

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

# Example: Discrete measure

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

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

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

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

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

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

# Lebesgue measure on $\mathbb{R}$

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

To each such interval assign the volume

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

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

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

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

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

Can extend this consistently to the whole  $\sigma$ -algebra

$\leadsto$  multi-dim Lebesgue measure  $\lambda_d$

# A funny measure on $\mathbb{R}$

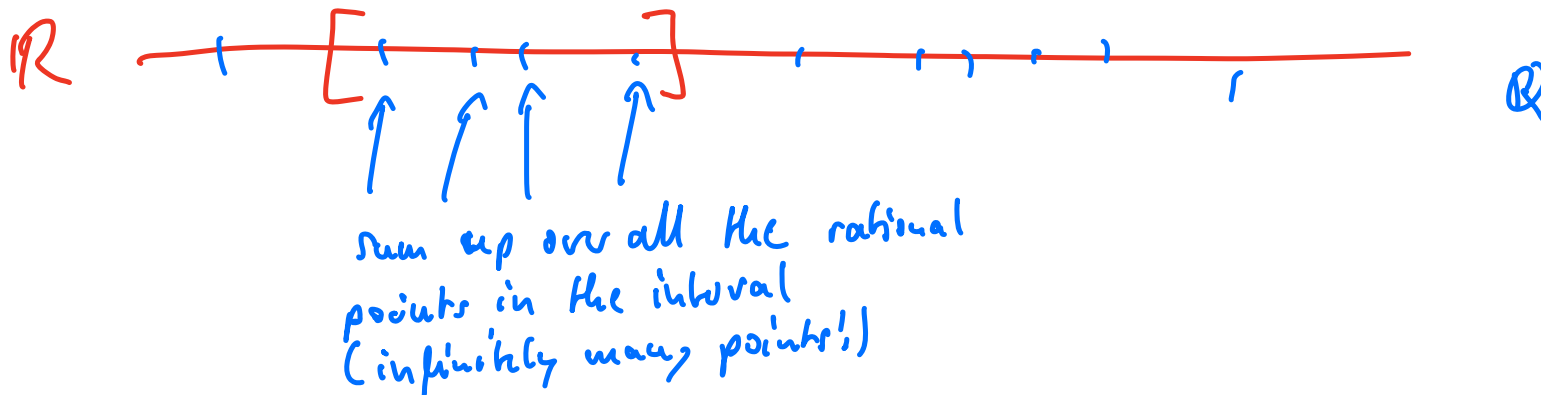
Let  $\mathcal{F}$  be the Borel- $\sigma$ -Algebra on  $\mathbb{R}$ . Want to define a measure that just assigns mass to rational numbers.

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

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

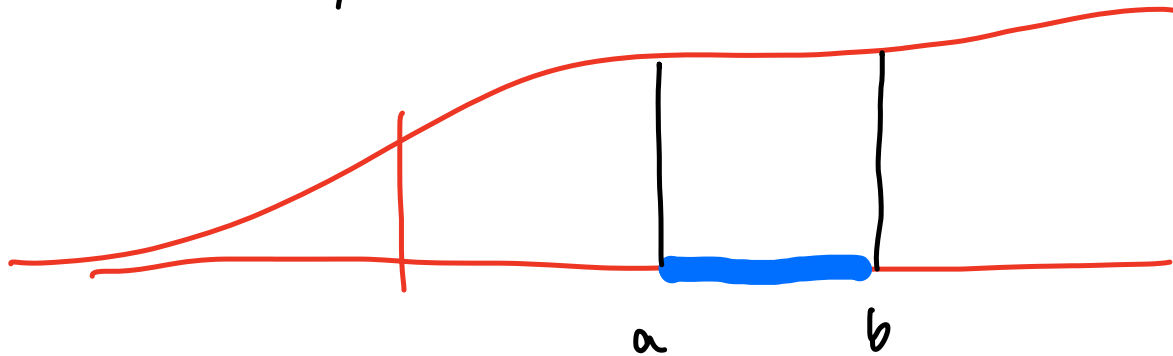
$$\mu(\{q_i\}) = m_i$$

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



## More general measures on $\mathbb{R}$

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



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

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

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

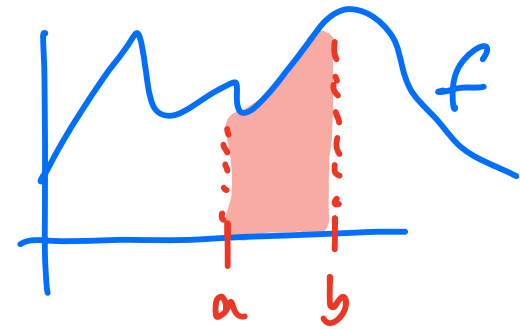


# Measure with a density

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

For intervals  $[a, b]$  define

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



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

Then  $\gamma$  is the measure with density  $f$  with respect to the Lebesgue measure.

More keywords: Radon-Nikodym theorem

## Carathéodory extension theorem

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

Mathematical basis for this approach is the Carathéodory theorem.

(skipped here)

# Measures with continuous and discrete parts

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

Example: If  $\lambda$  is the Lebesgue measure on  $[0, 1]$  with Borel- $\sigma$ -algebra. Let  $\delta_0$  be the discrete measure that assigns mass 1 to the set  $\{0\}$ .

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

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

↑  
Dirac measure

More keywords: Lebesgue decomposition theorem of measures

# Null sets

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

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

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

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

# Measurable mappings

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

A mapping  $f: X \rightarrow \mathcal{W}$  is called **measurable** if  
 $\forall A \in \mathcal{A}, f^{-1}(A) \in \mathcal{F}$  ("pre-images of measurable sets  
are measurable").

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

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

To Do:

add slides for

- integrating wrt measure  $\int f d\mu$
- hopf formula

Probability measure

Probability measure  $\mathcal{A} = \mathcal{A}_1, \mathcal{A}_2, \dots \subset \Omega$   
 $\Omega \in \mathcal{A}$   
 $P: \mathcal{A} \rightarrow \mathbb{R}$

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

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

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



# Example: throw a die

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

$P$  can be defined uniquely by assigning

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

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

Throw two dice:

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

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

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

$(i, j)$

first die second  
 $= \{(1, 1), (1, 2), (1, 3), \dots\}$

# Example: normal distribution

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

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

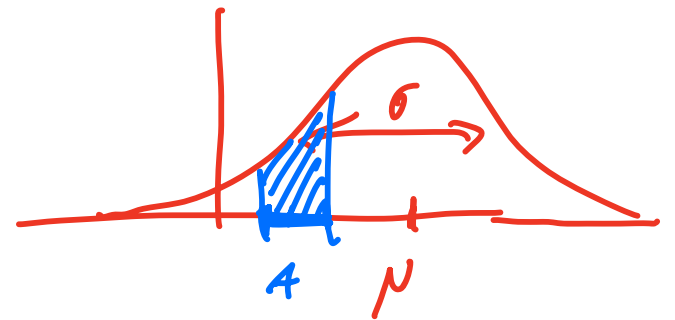
$$f_{\mu, \sigma} : \mathbb{R} \rightarrow \mathbb{R},$$

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

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

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

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



More examples

... see later...

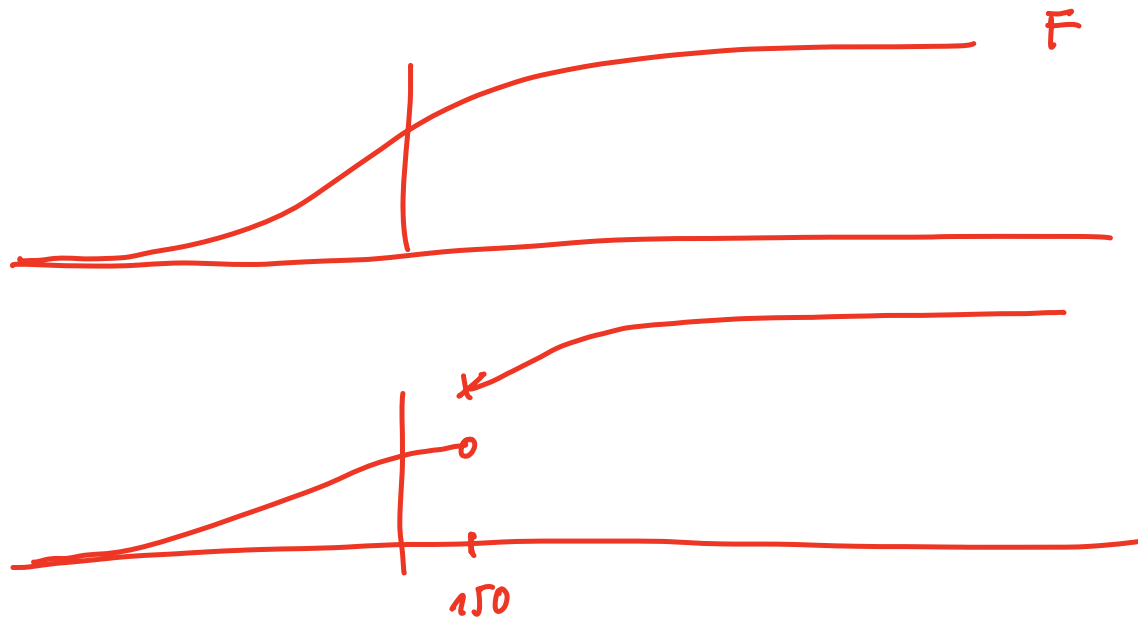
Cumulative distribution  
function (cdf)

# Cumulative distribution function

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

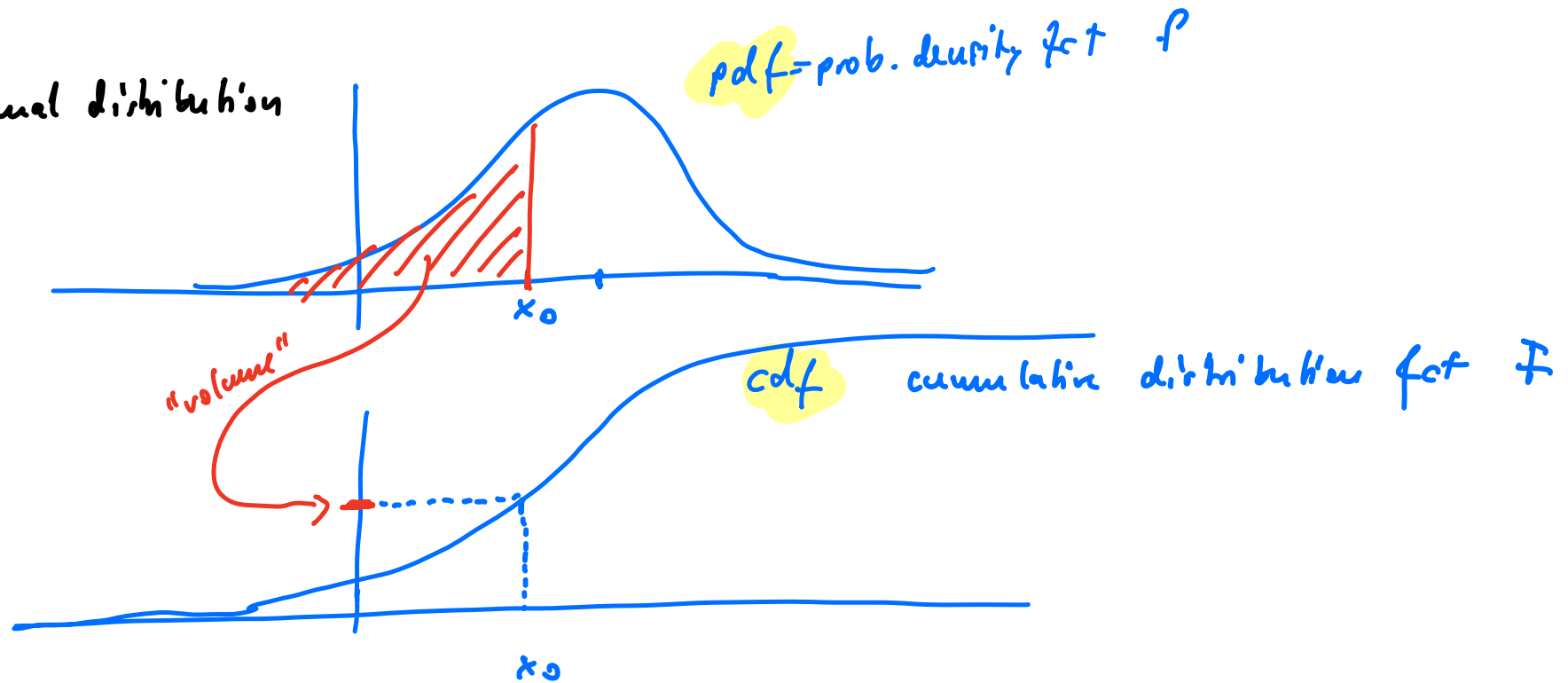
Define the function  $F: \mathbb{R} \rightarrow \mathbb{R}$ ,  $x \mapsto P([-\infty, x])$ .

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



# Cdf vs pdf

Example: normal distribution



# Properties of the cdf

A cdf satisfies the following properties:

(i)  $F$  is monotonically increasing with

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

(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  $x_n \rightarrow x$ ) then also

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

The other way round? Is it true that a function that satisfies

(i) and (ii) is always the cdf of a prob. measure?  $\leadsto$  Yes:

"Cdf" gives rise to a unique prob. measure

Proposition:

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

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

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

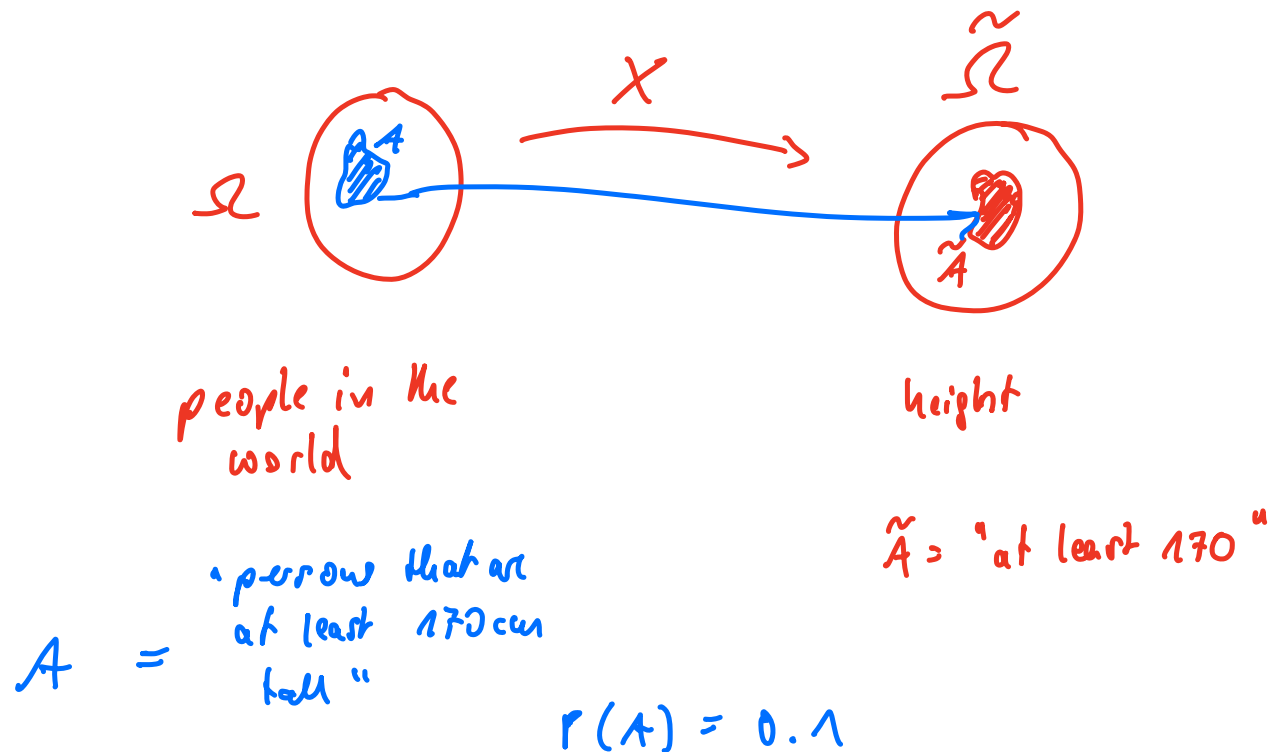


Random variable

# Random variable

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

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



# Example

Example: sum of two dice

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

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

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

$X$  "sum of the two values"

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

is measurable.

$\tilde{\Omega}$

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

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

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

# Distribution of a r.v.

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

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

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

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

# Induced $\sigma$ -algebra

$$(\Omega, \mathcal{A}, P), (\tilde{\Omega}, \tilde{\mathcal{A}}) \quad X: \Omega \rightarrow \tilde{\Omega}$$

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

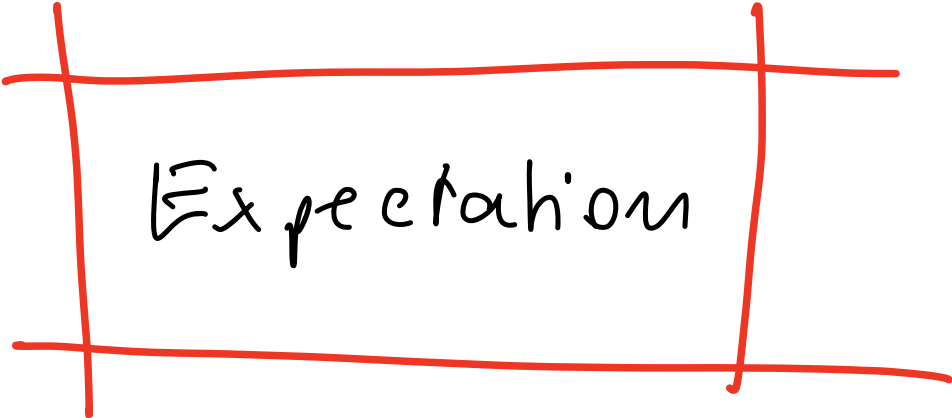
$$\sigma(X) := \{ X^{-1}(\tilde{A}) \mid \tilde{A} \in \tilde{\mathcal{A}} \}$$

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).

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



Expectation

## Expectation (finite case)

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

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

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

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

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

### Examples

- Toss a coin.  $\Omega = \{\text{head}, \text{tail}\}$ ,  $X = \mathcal{P}(\Omega)$ ,  $P(\text{head}) = p$ ,  
 $P(\text{tail}) = 1-p$ ,  $0 \leq p \leq 1$ .

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

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

In ML:

- Train error: expected error wrt empirical distribution which assigns prob  $\frac{1}{n}$  to all the  $n$  training points.
- Test error of a classifier: expected error wrt to the (unknown) underlying data distribution



# Expectation is linear

(discrete)

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

$$E\left(\overset{rv}{a \cdot X} + \overset{rv}{b \cdot Y}\right) = a \cdot E(X) + b \cdot E(Y)$$

$\downarrow$   $\downarrow$   
 $\in \mathbb{R}$   $\in \mathbb{R}$

Proof sketch  
(discrete case)

$$\begin{aligned} E(aX + bY) &= \sum_{x, y} (ax + by) P(X=x, Y=y) = \\ &= a \sum_{x, y} x \cdot P(X=x, Y=y) + b \sum_{x, y} y \cdot P(X=x, Y=y) = \\ &= a \left( \sum_x x \cdot \overbrace{\sum_y P(X=x, Y=y)}^{P(X=x)} \right) + b (\dots) \\ &= a \sum_x x \cdot P(X=x) + b \sum_y y \cdot P(Y=y) = a \cdot E(X) + b \cdot E(Y). \end{aligned}$$

law of  
total prob.



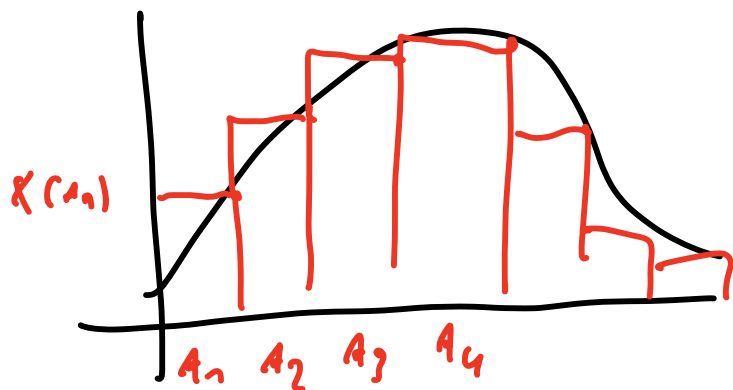
Expectation of a  
general rv

# Motivation

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

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

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



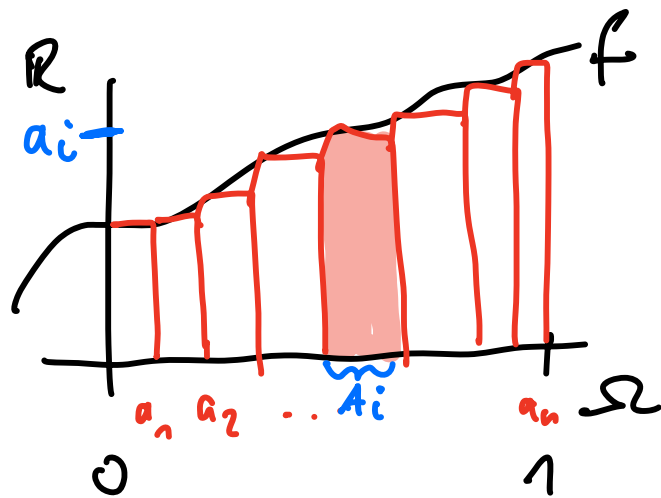
$$E(X) \approx \sum_i X(A_i) \cdot P(A_i)$$

$$\rightarrow \int x p(x) dx$$

# Motivation

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

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

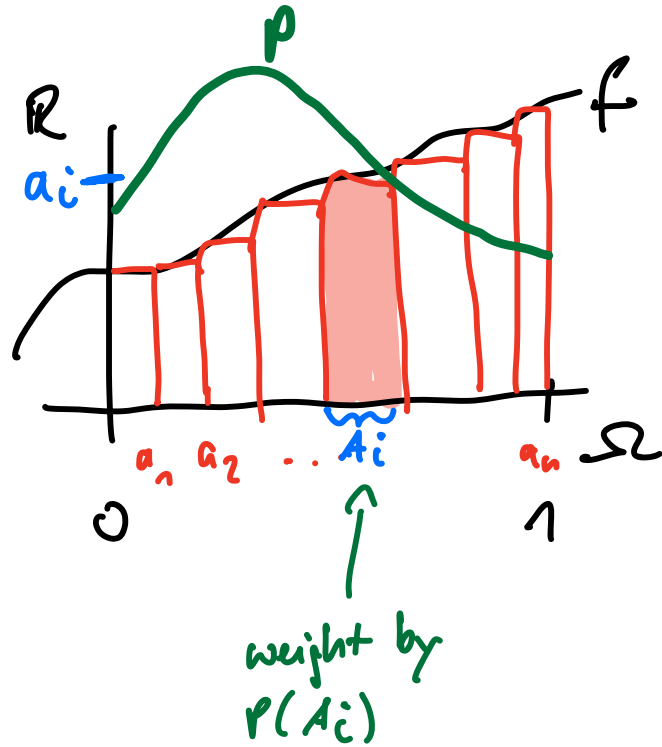


Intuitively, we partition the space  $\Omega = \mathbb{R}$  into sub  $A_1, A_2, \dots, A_n$  and compute the approximate integral:

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

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

② Now consider a case of  $\Omega$  with a prob. distribution with density  $p$ :



We now want to pursue a similar approach. But instead of using the length of  $A_i$  as volume, we now want to use the probability of  $A_i$  as notion of volume:

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

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

to 
$$\int_0^1 f(x) \underline{p(x)} dx$$

weight by the prob. density

③ But what do we do if we don't have a density  $p$ ? Or a more general input space  $\Omega$ ?

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

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

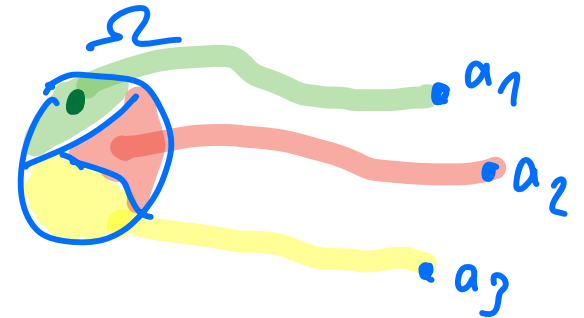
Instead of starting with a partition of the input space  $\Omega$  and refining it (which might be difficult because  $\Omega$  does not have any "structure" beyond a  $\sigma$ -algebra), we use a partition of the target space  $\mathbb{R}$ :

# Construction of an integral : simple rv

Step 1: Assume the random variable  $X: \Omega \rightarrow \mathbb{R}$  only takes finitely many values:  $X \in \{a_1, \dots, a_k\}$ , that is

There exist some sets  $A_1, \dots, A_k$  such that

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



Note that because  $X$  is a rv, the sets  $A_i$  need to be  $\mathcal{F}$ .

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

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

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





## Construction of an integral: non-negative rv

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

$$"E(X)" := \sup \{ E(Y) \mid Y \text{ simple rv with } \underbrace{0 \leq Y \leq X}_{Y \leq X \Leftrightarrow}$$

$$\forall \omega \in \Omega: Y(\omega) \leq X(\omega)$$

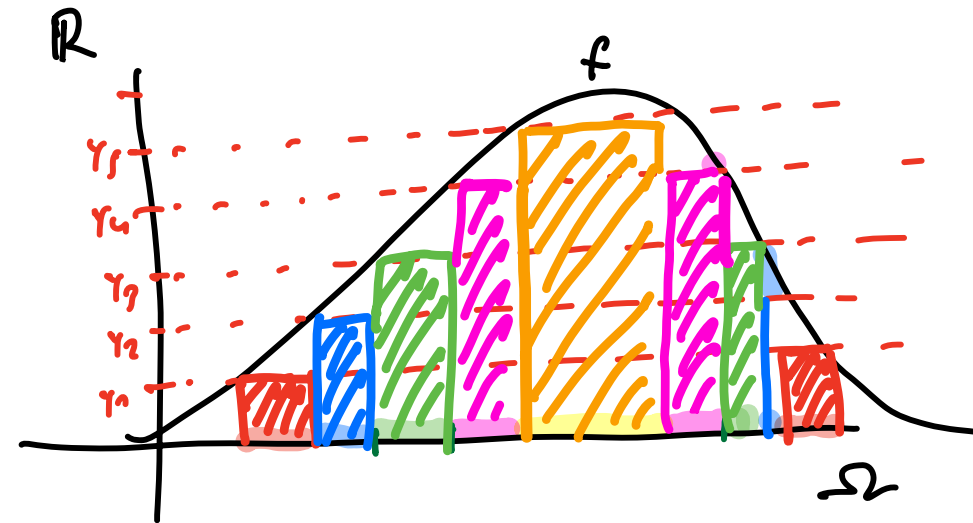
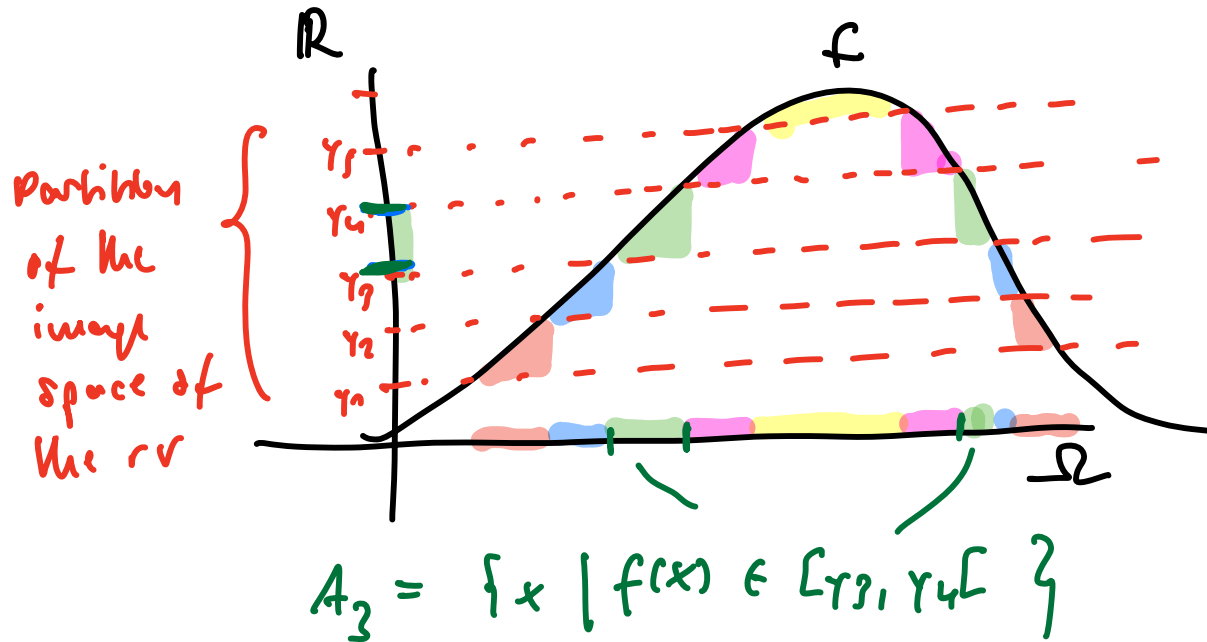
If the supremum exists and is finite, we call the resulting number the expectation  $E(X)$  of the non-neg. rv  $X$ .

Intuitively:

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

# Illustration of Step 2

Consider the case of a r.v  $f: \Omega \rightarrow \mathbb{R}$

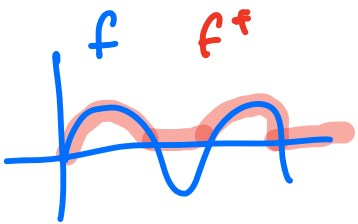


approximation of  $f$   
based on the partition  
induced by the output space

## Construction of an integral: general case

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

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



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

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

Then we define

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

$$\text{Notation: } E(X) = \int_{\Omega} X(\omega) dP(\omega) = \int X dP$$

# Important properties of the expectation

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

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

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

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

- Any bounded r.v. possesses an expectation:

- $X$  bounded  $\Leftrightarrow \forall \omega \in \Omega : |X(\omega)| \leq c$  for some  $c \in \mathbb{R}$

- "possess an expectation": have a well-defined, finite  $E(X)$

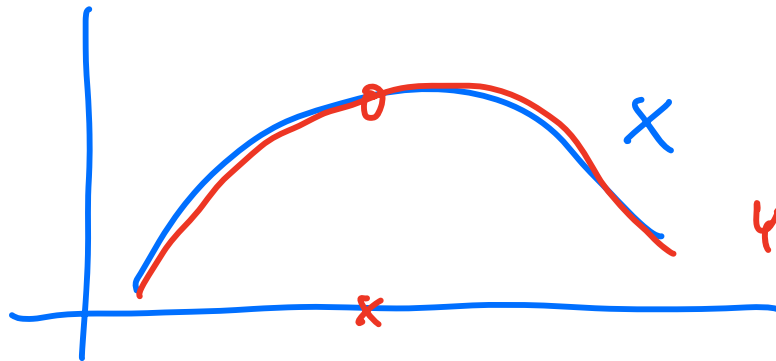
$$\cdot \quad X = Y \text{ a.s.} \quad \Rightarrow \quad E(X) = E(Y)$$

$$\text{Recall: } X = Y \text{ a.s.} \quad \Leftrightarrow \quad \exists N \in \mathcal{A} \text{ with } P(N) = 0$$

such that  $\forall \omega \in \Omega \setminus N$ :

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

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

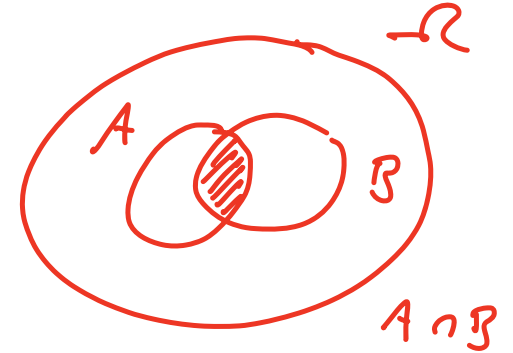


Conditional probability

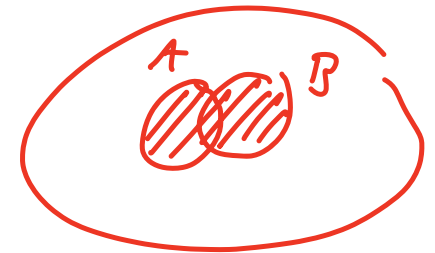
# Union and intersections

Notation:

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



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





# Conditional probability

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

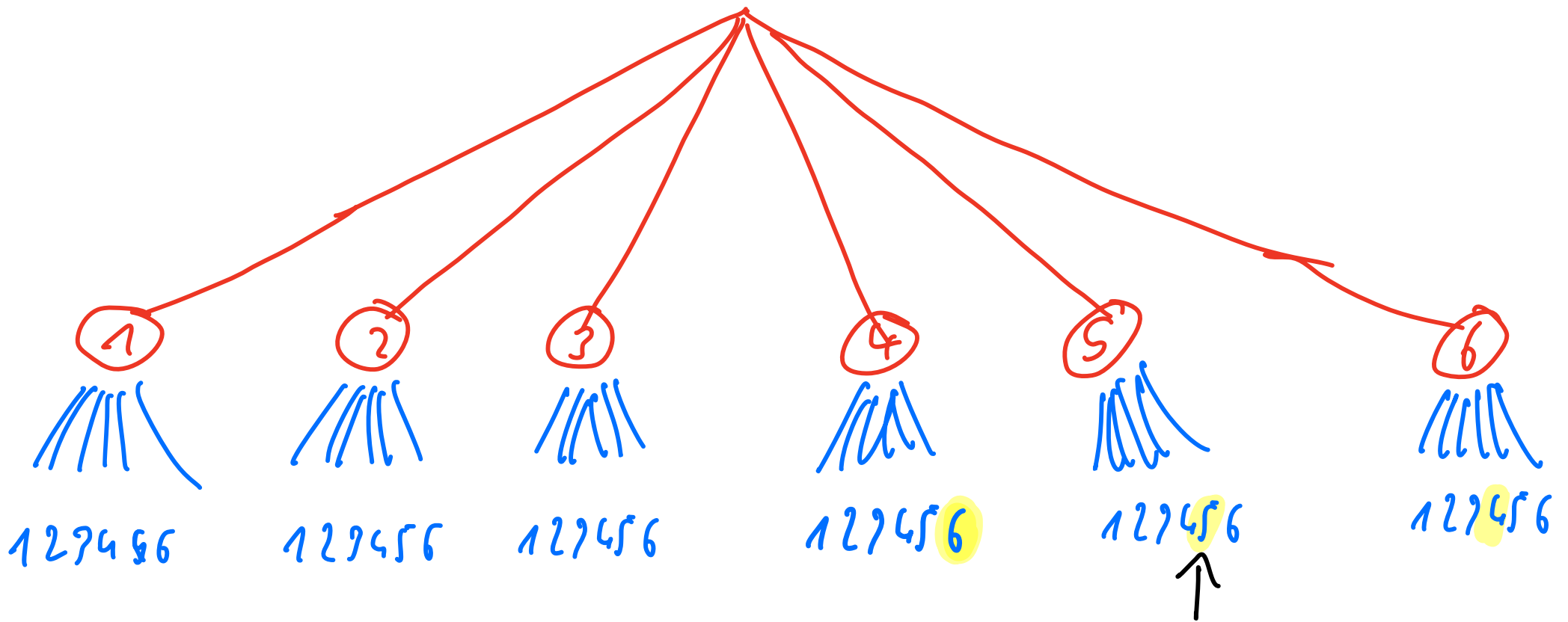
$$P(A | B) := \frac{P(A \cap B)}{P(B)}$$

is called the

conditional probability of A given B.

## Example with two dice

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



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

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

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

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

Note: of course the "ordering" matters

$\Omega$  = "all persons on earth" (a finite set)

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

$P$  = "uniform"

Event  $A :=$  "person has been vaccinated"

$B :=$  "person has disease"

i.e. different conditional distributions:

$P(\text{disease} \mid \text{vaccinated})$

$P(\text{vaccinated} \mid \text{disease})$

} not the same!!!

# Conditional distribution (positive condition)

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

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

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

# Regular conditional distribution

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

But:

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

$$P(Y=1 \mid \underbrace{X=0.1})$$

training pt.  $X$  drawn from normal distribution

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

Law of total probability and

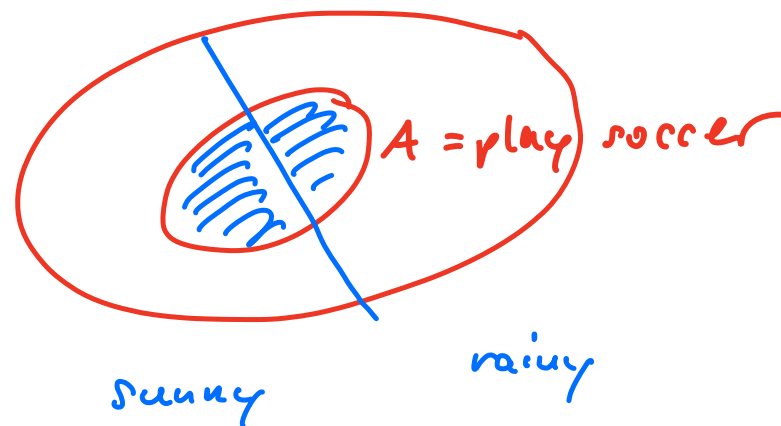
Bayes formula

# Law of total probability

Let  $B_1, B_2, \dots, B_k$  be a disjoint partition of  $\Omega$

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

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





# Bayes formula

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

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

# Example: breast cancer screening

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

90% of women with breast cancer will be 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?

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

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

Independence

# Two independent events

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

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

Notation:  $A \perp B$

Observation:  $A$  is independent of  $B \Leftrightarrow P(A|B) = P(A)$

# Many independent events

Def

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

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

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

# Independent random variables

Def:

Two random variables  $X: \Omega \rightarrow \Omega_1$ ,  $Y: \Omega \rightarrow \Omega_2$   
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:  $X \perp Y$

## Expectation of independent product

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

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

Proof sketch:  
(discrete case)

$$\sum_{i,j} x_i y_j P(X=x_i, Y=y_j) \stackrel{\text{ind.}}{=} \sum_{i,j} \underbrace{x_i y_j}_{x_i \cdot y_j} P(X=x_i) \cdot P(Y=y_j)$$

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

Variance and covariance



# Variance & standard deviation

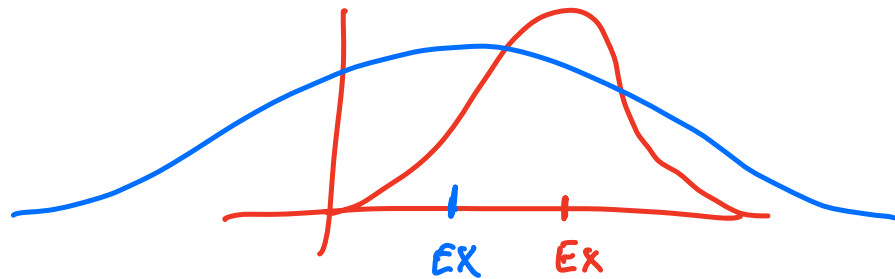
Def  $X, Y: (\Omega, \mathcal{F}, P) \rightarrow \mathbb{R}$  rvs with  
 $E(X^2) < \infty$ ,  $E(Y^2) < \infty$ .

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.



high variance

moderate variance

# Covariance & Correlation

Def :

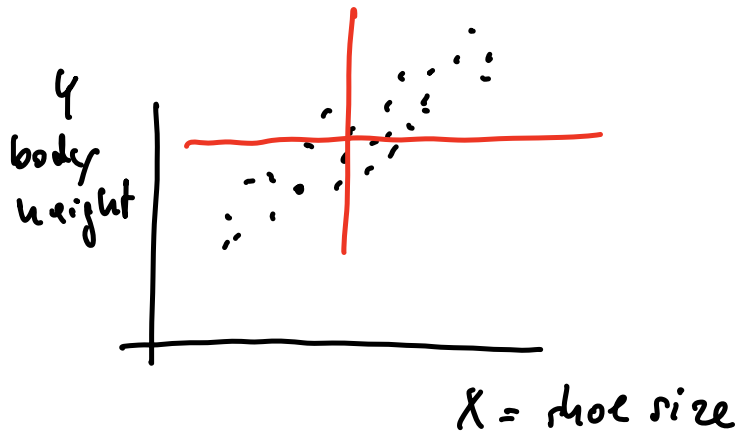
$\text{Cov}(X, Y) := E((X - E(X)) \cdot (Y - E(Y)))$  is called the covariance of  $X$  and  $Y$ .

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

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

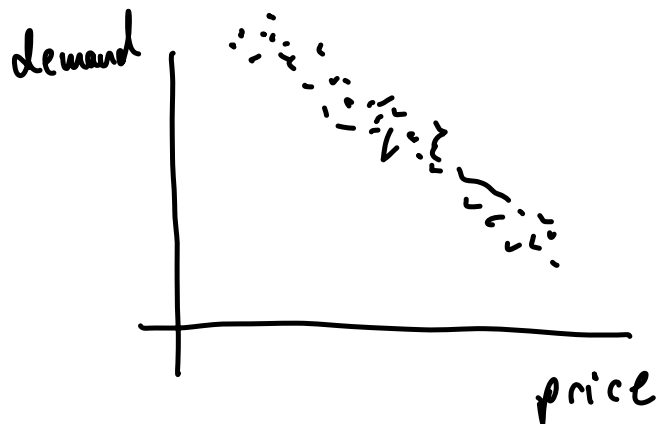
## Intuition about covariance

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



positive, large covariance

$$\rho \approx 0.9$$

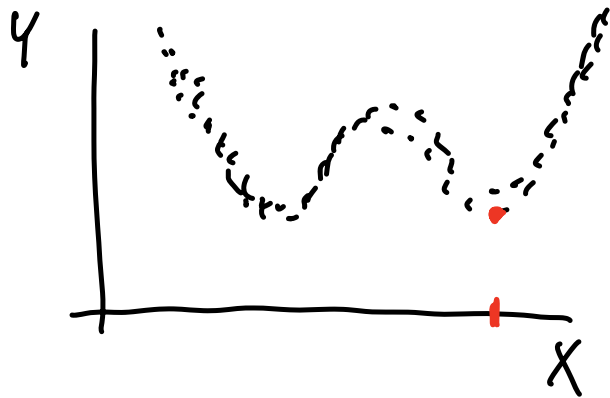


negative cov,

large in absolute values

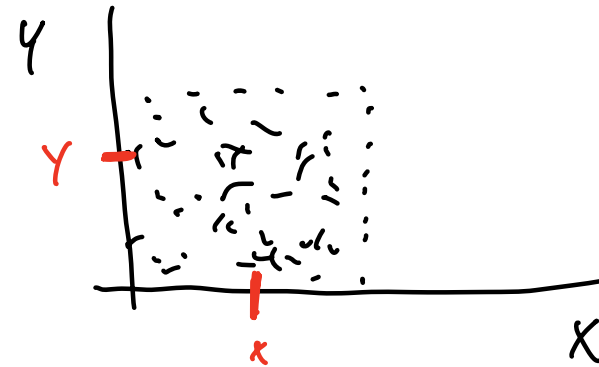
$$\rho \approx -0.9$$

# Intuition about cov



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

not independent!



independence



uncorrelated  $\nRightarrow$  independence!



# Properties of var and cov

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

## $k$ -th moment

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

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

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

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

Markov and Chebyshev  
inequality

# Cauchy-Schwarz inequality

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

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



# Markov inequality

Prop:  $\varepsilon > 0$ ,  $f: [0, \infty[ \rightarrow [0, \infty[$ ,  
 $f$  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

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

$$P(|X - E(X)| > \varepsilon) \leq \frac{\text{Var}(X)}{\varepsilon^2}$$

key quantity in learning theory!

SKIPPED

Different types of probability  
measures

# Discrete measure

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

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

We define a probability measure  $P: \mathcal{A} \rightarrow [0, 1]$  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

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

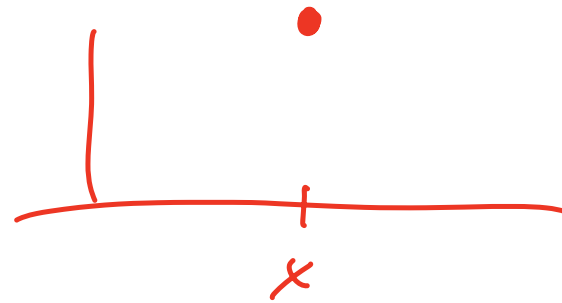
Examples: toss a coin; distribution on  $\mathbb{Q}$

# Dirac measure

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



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

$$\text{as } \frac{1}{6} (\delta_1 + \delta_2 + \dots + \delta_6)$$

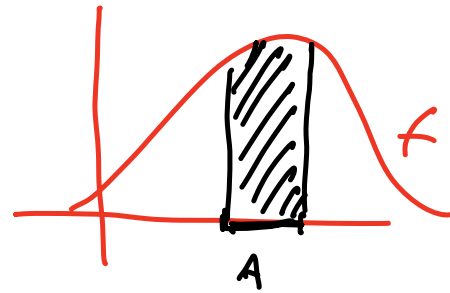
# Measures with a density

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

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

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

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



$\nu$  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:  $\delta_0$  Dirac measure

# absolutely continuous measures

Def A prob. measure  $\gamma$  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  $\gamma$ -null set:

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

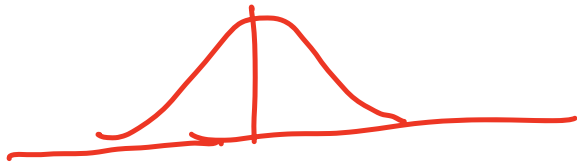
Notation:  $\gamma \ll \mu$



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

# Examples

- 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 of Radon - Nikodym

## Theorem (Radon - Nikodym)

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

(see next page)

(1)  $\nu$  has a density wrt  $\mu$ .

(2)  $\nu$  is absolutely continuous wrt  $\mu$ .

## Proof idea

(1)  $\Rightarrow$  (2) easy

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

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

$$(*) \left\{ \begin{array}{l} \bullet g \text{ is measurable, } g \geq 0 \\ \bullet g \cdot \mu \leq \nu, \text{ that is} \end{array} \right.$$

$$\forall A \in \mathcal{P}(\mathbb{R}^n): \int_A g \, d\mu \leq \nu(A).$$

• Observe:  $g \equiv 0$  satisfies  $(*)$ , so  $\mathcal{G}$  is not empty.

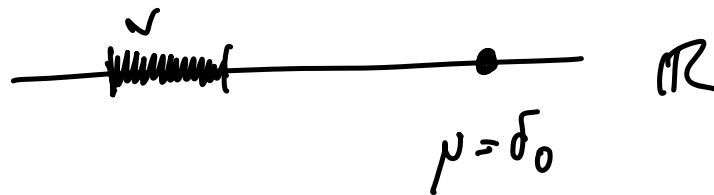
• If  $g, h$  both satisfy  $(*)$ , then  $\sup(g, h)$  satisfies  $(*)$ .

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



Example:  $\lambda \perp \delta_0$

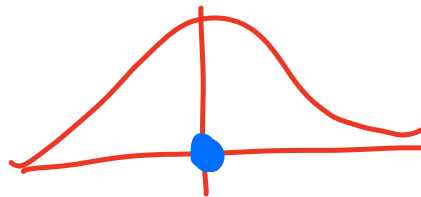
# Lebesgue decomposition

## Theorem (Decomposition by Lebesgue)

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

$$\nu_1 \ll \mu \text{ and } \nu_2 \perp \mu.$$

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



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

# Proof

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

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

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

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

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

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

Don't see job.



Cantor - distribution: non-trivial distribution that is singular w.r.t  $\lambda$

①

Construct the Cantor set:

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

"Remove middle part"



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

"Remove middle parts from all intervals"



- $C_2 =$



$\vdots$

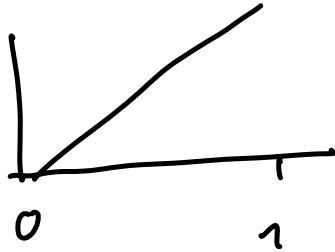
The Cantor set is the limit in this process. It is compact, non-empty, empty interior.

②

Now construct a probability distribution:

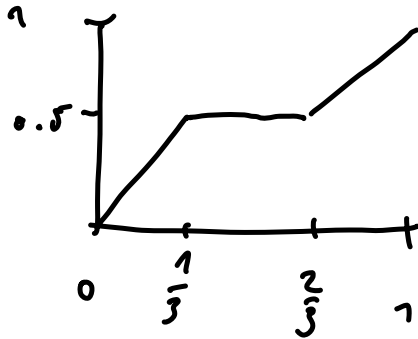
Consider the cdf of the sets  $C_0, C_1, C_2 \dots$

$C_0$ :



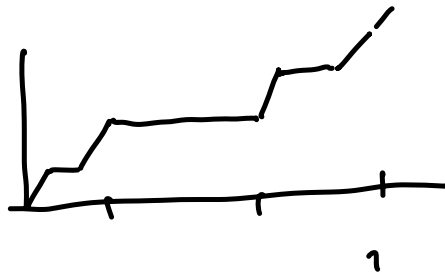
uniform on  $[1/2, 1]$

$C_1$ :



uniform on  $[0, 1/3] \cup [2/3, 1]$

$C_2$ :



... take limit and  
call the resulting measure  $\mu$ .

③ Proposition:

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

$\Rightarrow \lambda \perp \nu$

Lebesgue  
measure

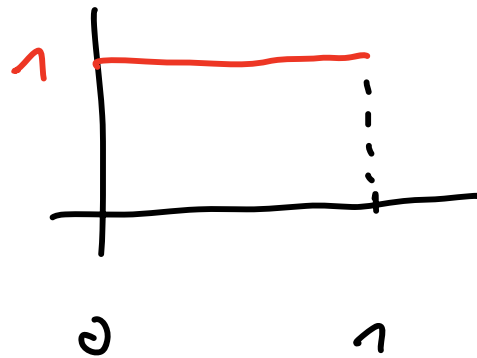


important examples of  
probability distributions

# Uniform distribution

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

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



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

# Binomial distribution

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

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

$X := \# \text{ heads}$

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

# Poisson distribution

Poisson distribution on  $\mathbb{N}$ :

Parameter  $\lambda > 0$

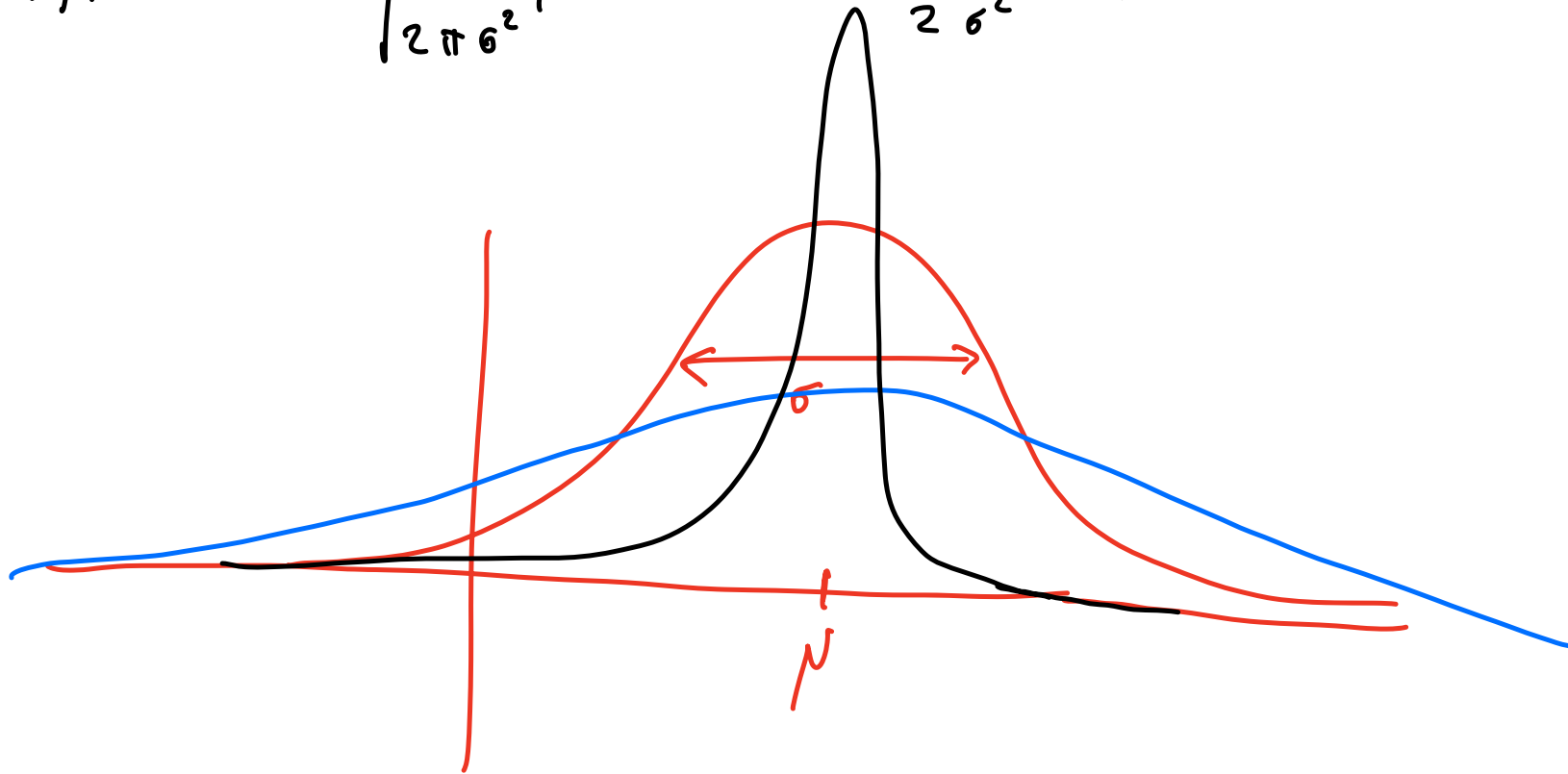
$$P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Intuition: number of incoming calls at a hotline.

# Normal distribution on $\mathbb{R}$

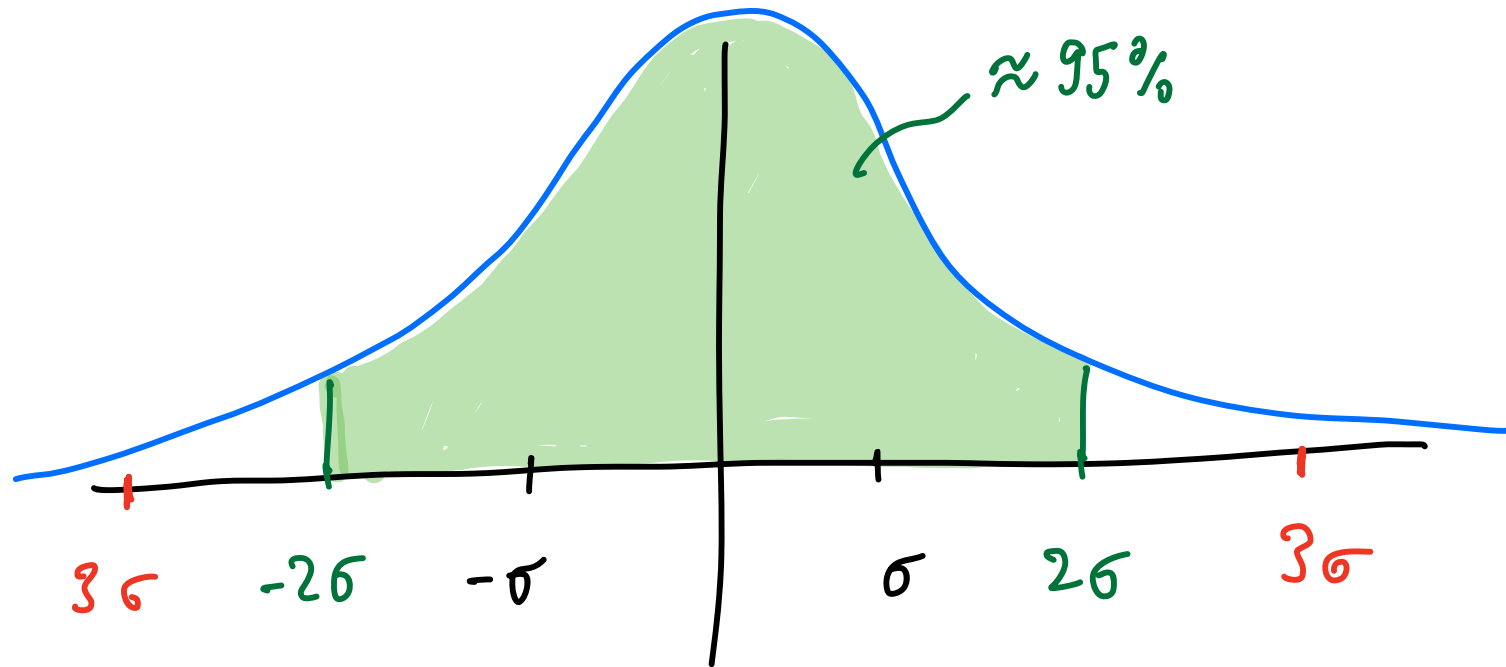
Density: parameter  $\mu$  (mean),  $\sigma$  (std. deviation)

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



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

## A rule of thumb



The area under the normal distribution:

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

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

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

## Sum of independent normals is normal

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

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

"Proof": It is elementary to see that the mean of  $X + Y$  is  $\mu_1 + \mu_2$  and the Variance is  $\sigma_1^2 + \sigma_2^2$ .

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

# Multivariate normal distribution

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

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

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

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



## The covariance matrix is prd

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

Proof Symmetry is clear because  $\text{Cov}(x_i, x_j) = \text{Cov}(x_j, x_i)$ .

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

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

$$\begin{aligned}
 a^t C a &= \sum_{i,j=1}^n a_i a_j C_{ij} \stackrel{\text{def}}{=} \sum_{i,j=1}^n a_i a_j E((X_i - \mu_i)(X_j - \mu_j)) \\
 &= E \left( \sum_{i,j=1}^n a_i a_j (X_i - \mu_i)(X_j - \mu_j) \right) \\
 &= E \left( \underbrace{\left( \sum_{i=1}^n a_i (X_i - \mu_i) \right)^2}_{\geq 0} \right) \geq 0
 \end{aligned}$$

linearity of expectation



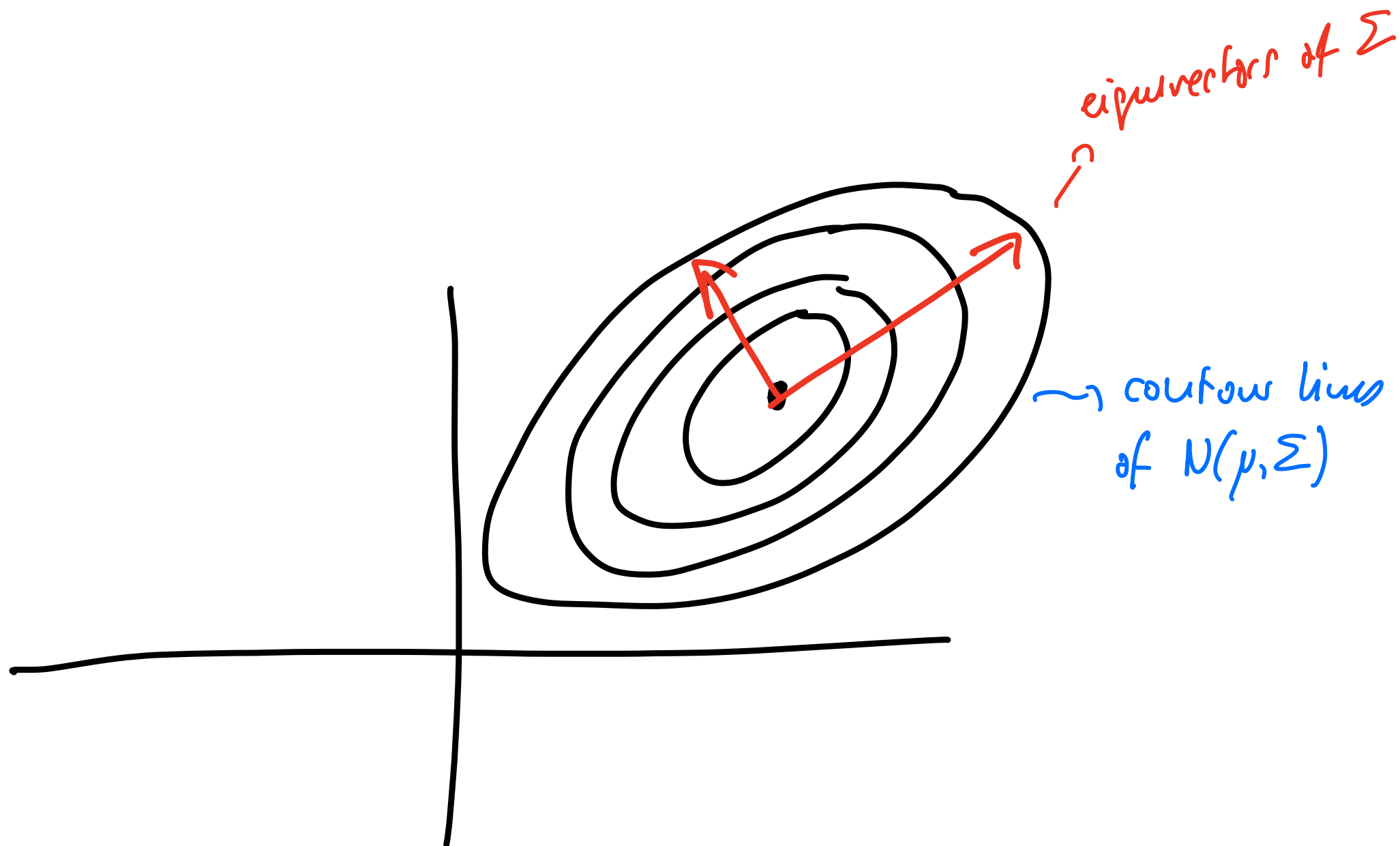
# PCA & Contour lines

Because  $\Sigma$  is symmetric and posd, it has real-valued eigenvalues  $\geq 0$ .

The contour lines of a multivariate normal are ellipses:

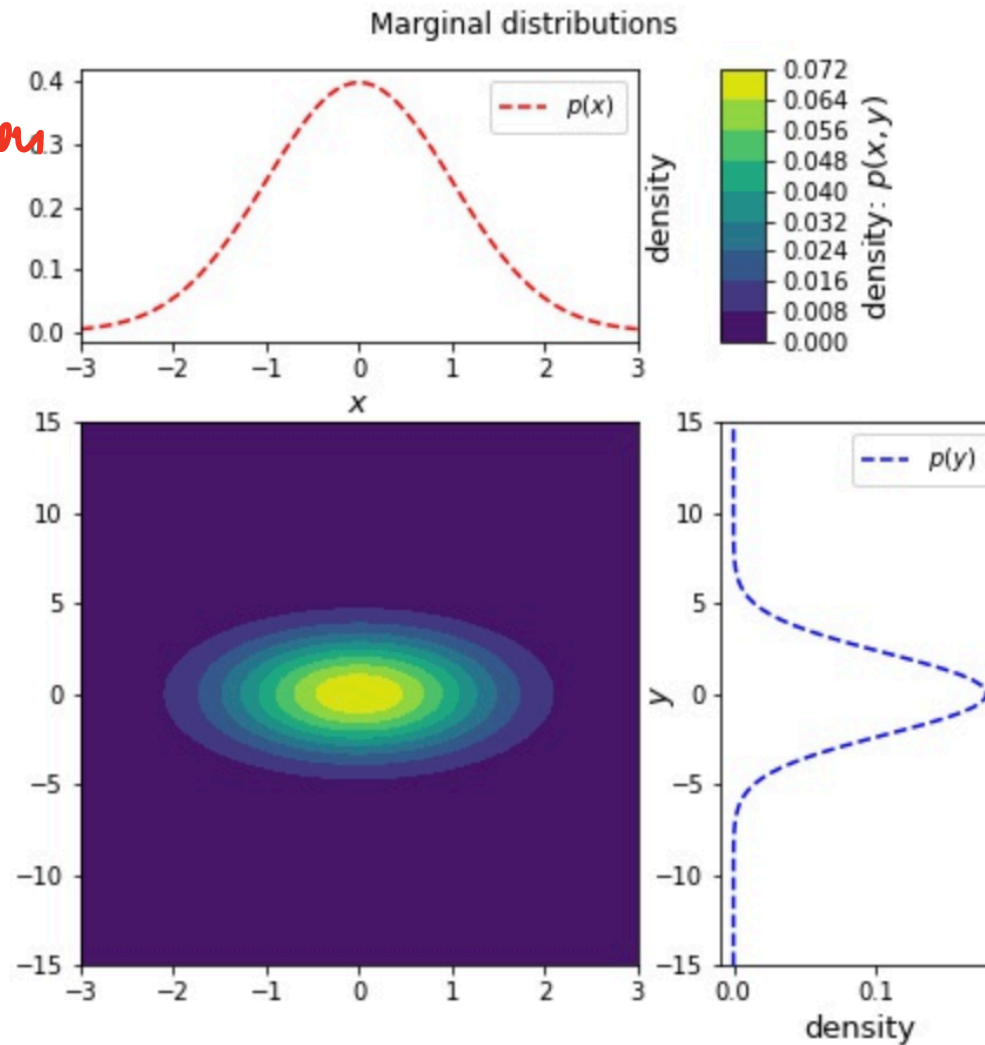
Contour line : set  $\{x \mid f_{\mu, \sigma}(x) = c\}$

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



# Marginal distributions of Gaussian are Gaussian

Marginal distribution  
of  $x$

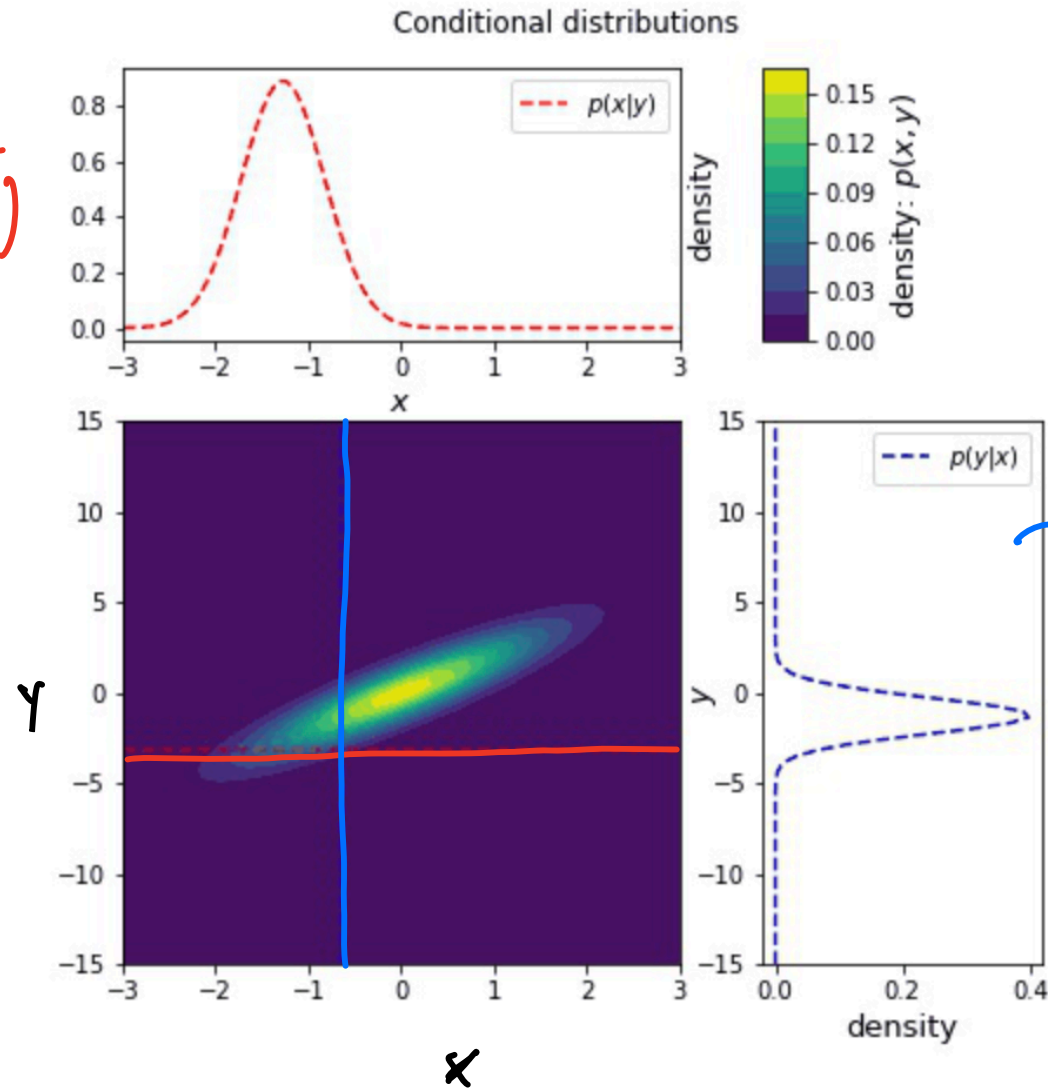


Marginal distribution  
of  $y$

Figure credit: Nils Lehmann

# Conditional distributions of Gaussian are Gaussian

conditional  
 $p(x | y = -4)$



conditional  
 $p(y | x = -0.7)$

# Mixture of Gaussians

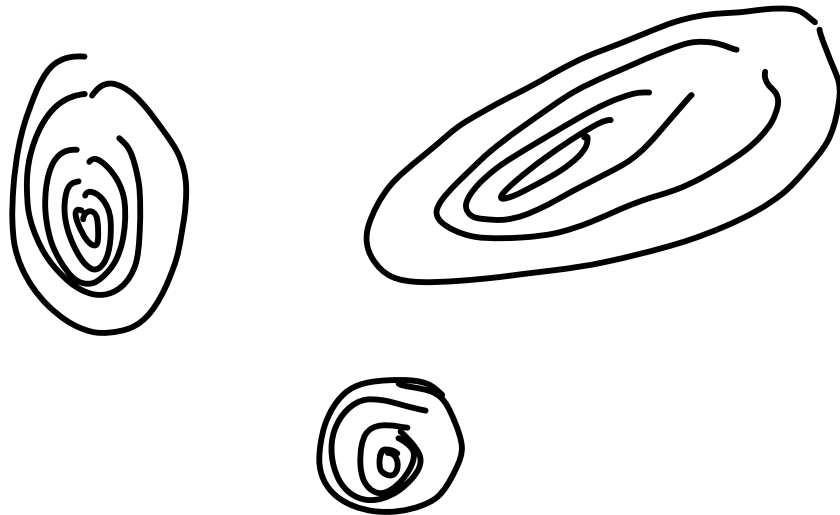
Consider "mixing weights"  $a_1, \dots, a_k \in [0, 1]$  s.t.  $\sum_{i=1}^k a_i = 1$ .

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

density

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

This is called a mixture of Gaussians.



## Examples of distributions with "infinite" expectation

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

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

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

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

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

ML keyword: preferential attachment model for social networks



# Heavy-tailed vs sub-Gaussian distributions

"Tail behavior" of a Gaussian:

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

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

Sub-Gaussian

...

smaller

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

Convergence of  
random variables

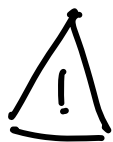
# Almost sure convergence

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

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

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

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



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

## Well-defined

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

Proof  
sketch

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

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

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

So we get:

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

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

countable unions  
and intersections

$X_n, X$  are measurable  $\Rightarrow$   
 $|X_n - X|$  is measurable

$\therefore \{\dots\} \in \mathcal{A}$

□

# Convergence in probability

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

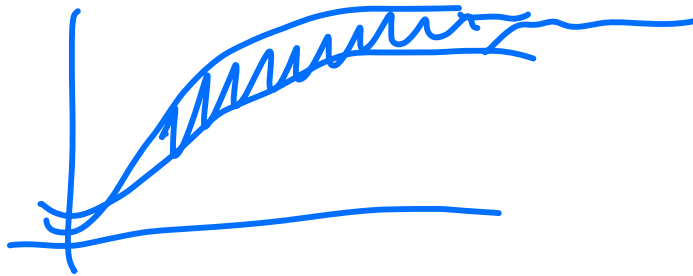
$$\forall \varepsilon > 0 \quad \mathbb{P}(\{\omega \in \Omega \mid |X_i(\omega) - X(\omega)| > \varepsilon\}) \longrightarrow 0$$

# Convergence in $L_p$

Def  $X_n \rightarrow X$  in  $L^p$  ("in the  $p$ -th mean")  $\Leftrightarrow$   
 $x_n, x \in L^p$  and  $\|x_n - x\|_p \rightarrow 0$ .

$$\|x_n - x\|_p^p = \int |x_n - x|^p d\mu(x) \rightarrow 0$$

$$\|z\|_p^p = \int |z|^p d\mu$$



# Weak convergence

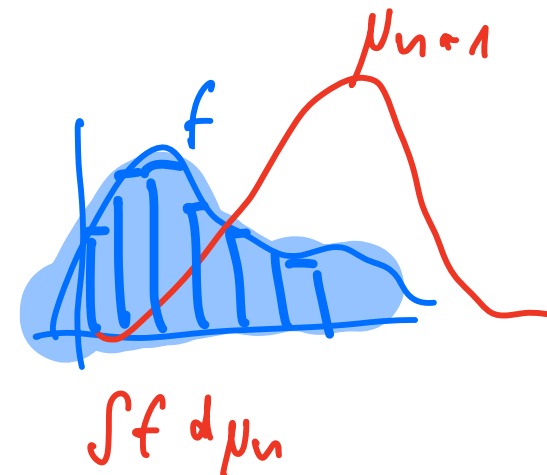
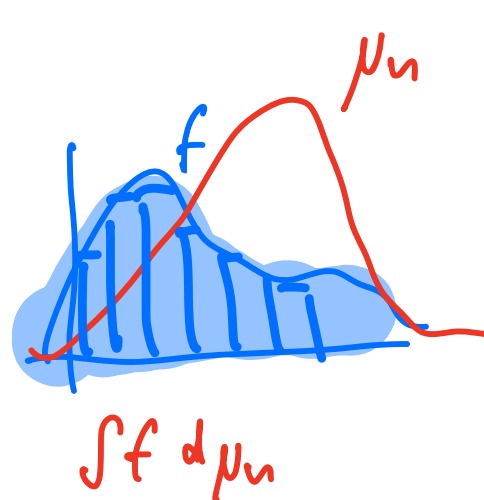
Def

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

$$\mu_n \rightarrow \mu \text{ weakly} : \Leftrightarrow$$

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

⚠ weak convergence is defined for measures, not for r.v.s.



## (Excursion: weak convergence in functional analysis)

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

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

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

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



# Convergence in distribution

Def

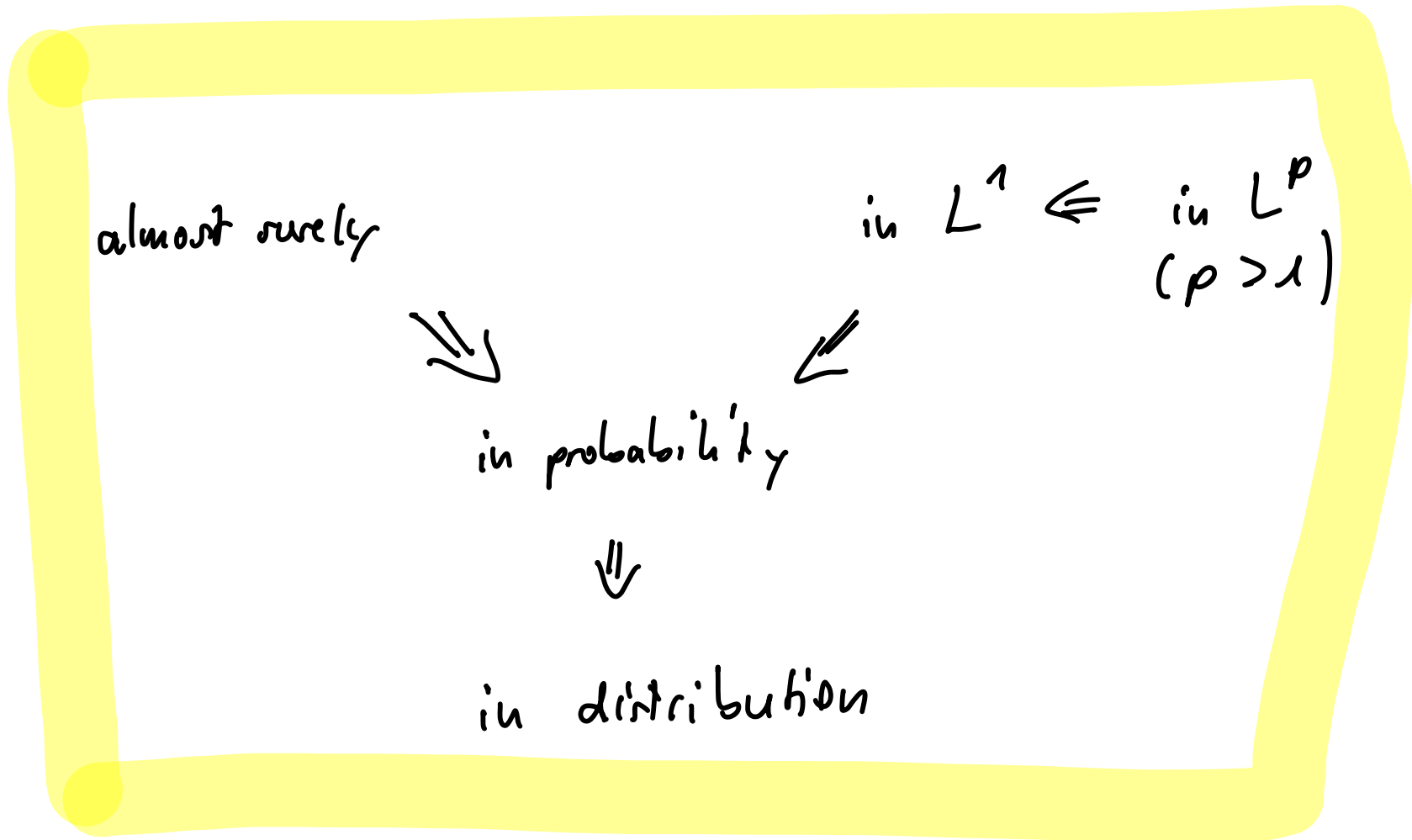
$X_i, X : (\Omega, \mathcal{A}, P) \rightarrow \mathbb{R}^n$ . The sequence  $X_n$

converges in distribution to  $X : \Leftrightarrow$

the distributions  $P_{X_n}$  converge to  $P_X$  weakly.

# Relationship between notions of convergence

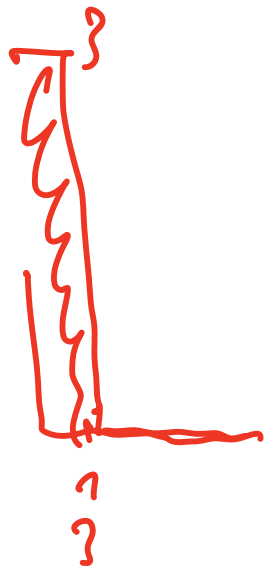
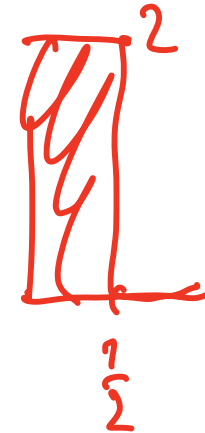
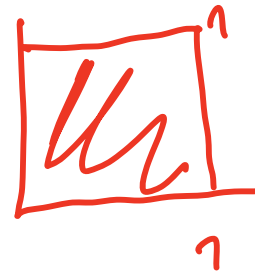
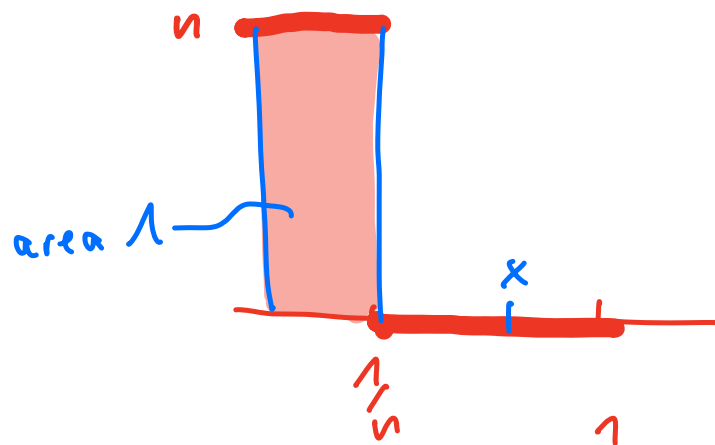
We have the following implications, and none of the "missing directions" hold in general:



## Example (converges a.s., in prob., but not in $L^1$ )

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

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



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

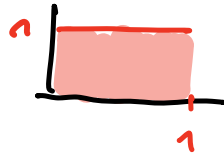
Can formally see: converges a.s., in prob.

But: no convergence in  $L^1$ .

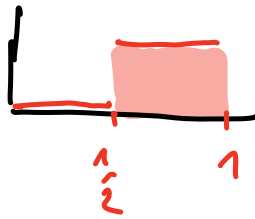
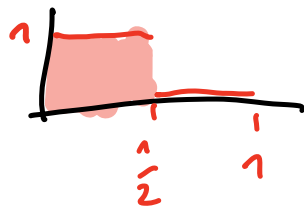
Example (convergence in prob., in  $L^1$ , but not a.s.)

"sliding blocks of smaller and smaller volume":

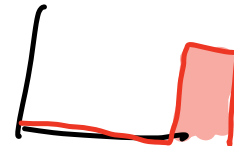
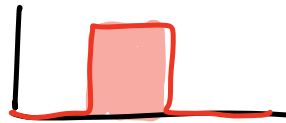
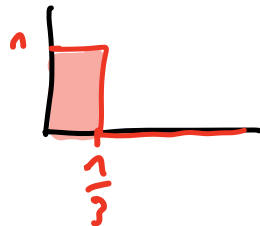
$$f_1 = \mathbb{1}_{[0,1]}$$



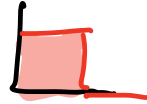
$$f_2 = \mathbb{1}_{[0,1/2]} , \quad f_3 = \mathbb{1}_{[1/2,1]}$$




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



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



•  $X_n : [0, 1] \rightarrow \mathbb{R}$ ,  $\forall i: X_i(\omega) = 1 \mathbb{1}_{[0, \frac{1}{2}[}$  

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

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

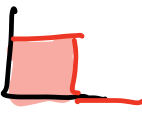
$$\underbrace{P_{X_n}}_{\text{green circle}} = \frac{1}{2}(\delta_0 + \delta_1) = P_{X_2} = P_{X_3} = \dots = P_X$$


so  $X_n \rightarrow X$  in distribution.

$\Omega = \mathbb{R}$ ,  $\delta_0$  is the measure on  $\mathbb{R}$  at 0 s. H.   
$$P(A) = \begin{cases} 1 & \text{if } 0 \in A \\ 0 & \text{otherwise} \end{cases}$$
 

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

Space  $[0,1]$  with uniform distribution,

•  $X_i : [0,1] \rightarrow \{0,1\}, \quad X_i(\omega) = 1_{[0, \frac{1}{2}[}$    
(same distribution for all the  $X_i$  !)

•  $X = 1_{[\frac{1}{2}, 1]}$  

Obviously  $X_n \not\rightarrow X$  in prob., but

$$P_{X_1} = \text{Ber}\left(\frac{1}{2}\right) = P_{X_2} = P_{X_3} = \dots = P_X$$

Bernoulli-distr.  
on  $\{0,1\}$

so  $X_n \rightarrow X$  in distribution!

Theorem of  
Bord-Coullet

## Definition " $A_n$ infinitely often "

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

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

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



Almost sure convergence ( $\Rightarrow$  " $> \varepsilon$  infinitely often")

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

$$X_n \rightarrow X \text{ a.s.}$$

$\Leftrightarrow$

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

$\downarrow$   
"wrt  $n$ "

# Proof intuition

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

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

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

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

$$\text{Thus } P(\textcircled{*}) = 1 \quad \Leftrightarrow \quad P(\textcircled{**}) = 0$$

□

# Borel-Cantelli

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

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

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

## Application in learning theory

$\Rightarrow$  conv. in probability

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

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

thus  $X_n \rightarrow x$  a.s. See assignments.

The law of large numbers  
(LLN)

iid

... is an abbreviation for

"independent and identically distributed"

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

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

# Strong vs weak

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

"Weak law"  $\Leftrightarrow$  convergence in probability

# Weak law of large numbers

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

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

In particular,  $S_n - \mu \rightarrow 0$  in probability.

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



# Proof

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

The weak law then follows directly from the Chebyshev inequality:

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

Now exploit that  $E(S_n^2) = E\left(\left(\frac{1}{n} \sum x_i\right)^2\right) = \frac{1}{n^2} E\left(\sum_{i,j} x_i x_j\right) \stackrel{\text{ind.}}{=}$

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

Plugging this in above immediately gives the desired result. ◻

## Strong law of large numbers (vanilla version)

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

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

## Proof

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

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

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

General idea:

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

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

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

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

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

# Proof

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

Proof :

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

$$= \dots$$

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

$$= n \cdot E(X_1^4) + 3n(n-1) E(X_1^2 X_2^2)$$

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

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

Recall Markov inequality:  $f$  non. increasing, then

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

L

For all  $\nu > 0$ ,

$$P\left(\frac{1}{n} |S_n| > n^{-\nu}\right) \leq \frac{E\left(\left(\frac{1}{n} |S_n|\right)^4\right)}{n^{-4\nu}} \stackrel{\text{step 1}}{\leq} \frac{K n^2}{n^{-4\nu}} = K \cdot n^{-2+4\nu}$$

Proof thp 3 : Borel-Cantelli

Fix some  $p \in ]0, \frac{1}{4}[$  and define the events

$$A_n = \left\{ \frac{1}{n} |S_n| \geq n^{-p} \right\}.$$

$$\text{Then } \sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} K n^{-2+4p} < -1 \quad \leftarrow \text{because} < \infty.$$

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



Many more versions of the LLN exist

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

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

## The conditions in this theorem

Independence: The theorem does not hold if we allow for arbitrary dependencies.

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

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

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

But:  $S_n = \begin{cases} 0 & X_1 = 0 \\ n & X_1 = 1 \end{cases}$ , hence

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

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



## The conditions in this theorem

Id. distributed : This condition is not really necessary.

same  
mean

For example, a popular version of the theorem states that the variance of all the  $X_i$  needs to be bounded by the same constant:

$$\forall i \in \mathbb{N}: \text{Var}(X_i) \leq C$$

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

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

# Machine learning question

Consider a training set  $(x_i, y_i)_{i=1, \dots, n}$  drawn iid  $\sim P$ .

Consider a loss function  $\ell$  and a classifier  $f_n$  that has been trained on this data.

The training error is defined as

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

The test error is defined as

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

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

No! Why?

the central limit theorem  
(CLT)

# Central Limit Theorem (vanilla version)

Theorem:

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

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

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

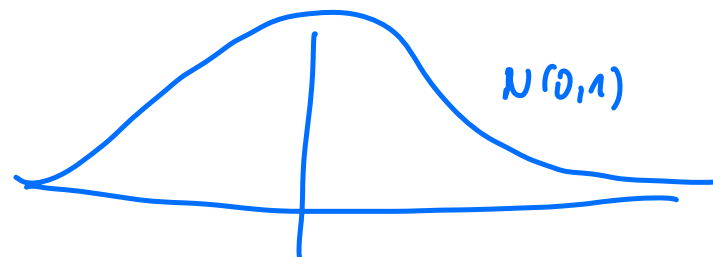
(which has mean 0 and standard dev. 1).

Then  $Y_n \rightarrow Y$  in distribution where  $Y \sim \mathcal{N}(0, 1)$ .

# Illustration

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

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



# Which conditions are needed?

Also here, many more versions ...

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

# CLT in d dimensions

Theorem: Let  $(x_i)_{i \geq 1}$  be iid rvs with values in  $\mathbb{R}^d$ .

let  $\mu$  be the (d-dim) mean and  $C$  the cov. matrix  
between the d variables. Then

$$\lim_{n \rightarrow \infty} \frac{S_n - \mu \cdot n}{\sqrt{n}} = Z$$

where  $Z$  is a d-dim rv with distribution  $N(0, C)$ .

# Discussion

Really impressive!

- We don't need any assumptions on the exact distribution of the  $X_i$
- Whenever something is a sum of many ind. var., it is going to look like a normal distribution
- Eg, measurement errors
- CLT is "the basis" for statistical testing!



Concentration inequalities

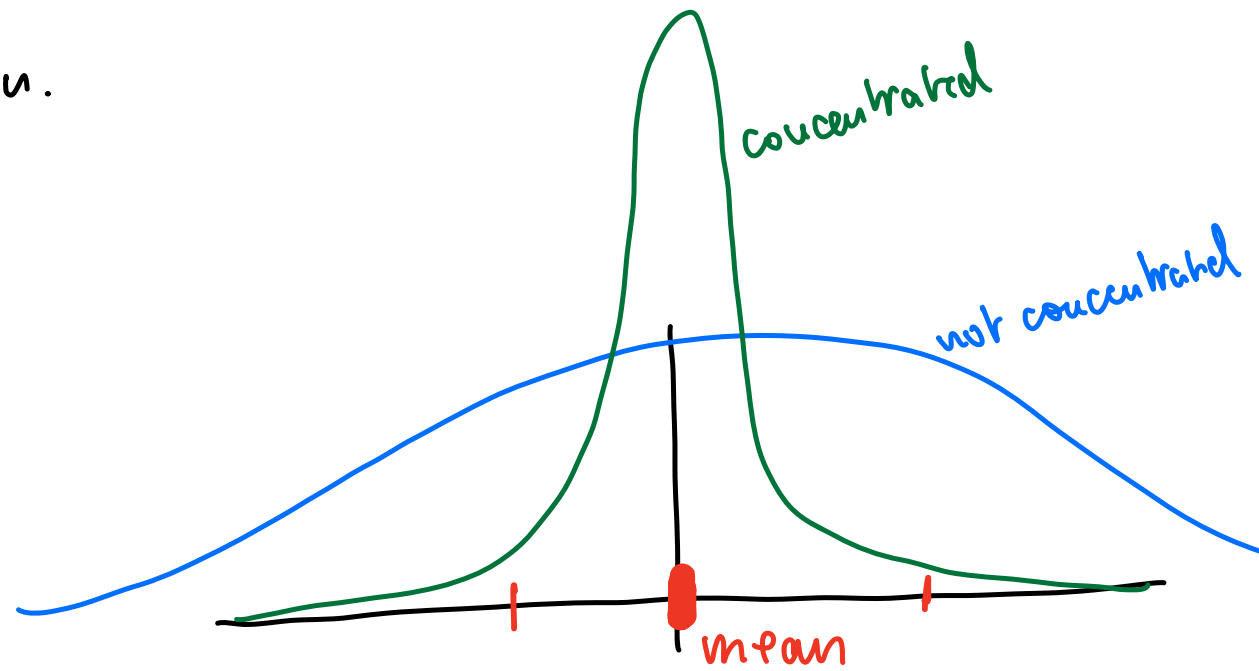
## First motivation

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

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

But the speed of convergence can be really slow.

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



## Hoeffding inequality

Theorem (Hoeffding):  $x_1, \dots, x_n$  r.v., independent,

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

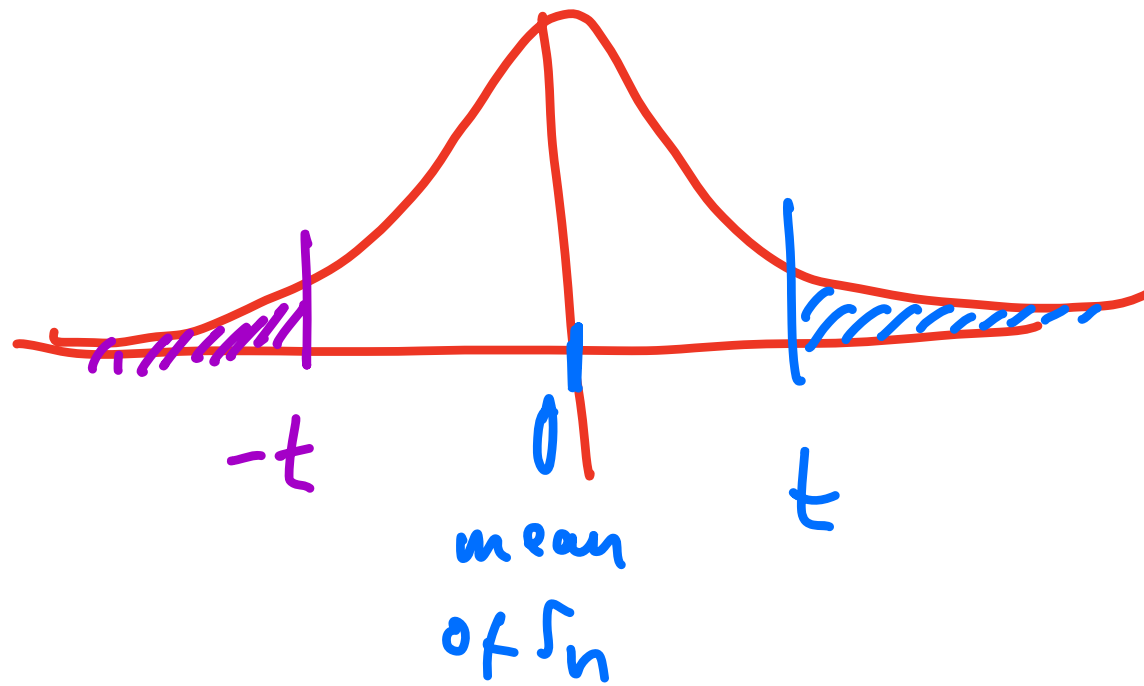
Let  $S_n := \sum_{i=1}^n (x_i - \mathbb{E}(x_i))$ . Then for any  $t > 0$ ,

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

Corollary: If the  $x_1, \dots, x_n$  are iid with  $x_i \in [a, b]$ , then

$$P\left(\frac{1}{n} S_n > t\right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right).$$

# Illustration



## Application of Hoeffding: LLN

Consider the following variant of the LLN:

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

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

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

# Proof of the LLN with Hoeffding

Step 1: Hoeffding  $\Rightarrow$

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

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

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

Combined we get

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

Step 2: Now want to apply Boel-Cauchli to get a.s. convergence of

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

event  $A_n$  from Boel-Cauchli

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

⊛ Subst'ute:  $r := \exp\left(-\frac{Z t^2}{(b-a)^2}\right)$ ; observe that  $r \in ]0, 1[$  if  $t > 0$ .

$$\text{observe: } \exp\left(-\frac{Z_n t^2}{(b-a)^2}\right) = r^n$$

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

Now Boel-Cauchli gives almost sure convergence,  $\square$

# Hoeffding for ML training error?

$(X_1, Y_1), \dots, (X_n, Y_n)$  iid training points

$f$  arbitrary, fixed function,  $\ell$  loss function.

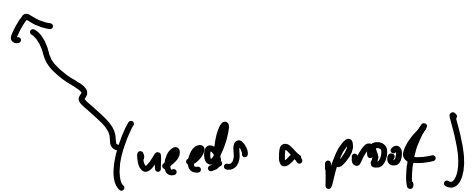
$$R_n(f) = \frac{1}{n} \sum_{i=1}^n \underbrace{\ell(f(x_i), y_i)}_{\text{loss of predicting } f(x_i)} \quad \text{empirical risk}$$

$$R(f) = E(\ell(f(x), y)) \quad \text{true risk}$$

Do we have  $R_n(f) \rightarrow R(f)$  as  $n \rightarrow \infty$ ?



# Hoeffding for ML training error?

(1) If  $f$  is independent of the training pts, then the terms  $\ell(f(x_1), y_1), \dots, \ell(f(x_n), y_n)$  are independent and we can apply Hoeffding to assert that  $R_n(f) \rightarrow R(f)$ . 

(2) But in ML, we typically deal with a function  $f_n$  that has been chosen by an algorithm that looks at all training pts! So the  $\ell(f_n(x_1), y_1), \dots, \ell(f_n(x_n), y_n)$  are not independent and we cannot simply use Hoeffding to assert that  $R_n(f_n) \rightarrow R(f)$ .  
→ see lecture on statistical ML next time!!!

Hoeffding is tight without further assumptions

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

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

## Bernstein inequality

Hoeffding inequality is tight for example for fair coin flips, but not tight in other cases. Might want to use

this instead:

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

for all  $t > 0$ ,

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

## Concentration inequality for functions with bounded differences

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

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

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

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

# Theorem of McDiarmid

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

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

Then, for any  $t > 0$ ,

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

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

## Cool observation

Consider again the ML scenario with iid training pts  $(x_i, y_i)$ , a loss  $\ell$  and a function  $f_n$  that has been chosen based on the training pts. Consider again the term

$$\frac{1}{n} \sum_{i=1}^n \ell(f_n(x_i), y_i) =: g(x_1, \dots, x_n)$$

If we can establish a bounded difference statement for  $g$ , then we can apply McDiarmid and get concentration !!:

Observe that  $g$  involves the construction  $f_n$  as well:

$$\underbrace{x_1, \dots, x_n \xrightarrow{\text{learning algorithm}} f_n \longrightarrow \frac{1}{n} \sum_{i=1}^n \ell(f_n(x_i), y_i)}_{g(x_1, \dots, x_n)}$$

$\leadsto$  Algorithmic stability

# Applications of concentration inequalities

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

Skipped

Glivenko-Cantelli Theorem



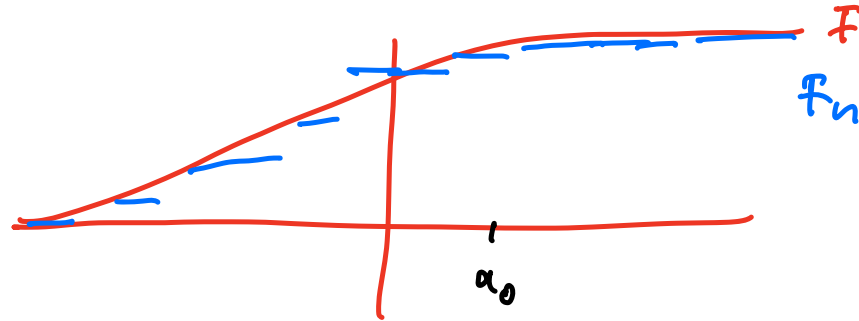
# Glivenko - Cantelli Theorem

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

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

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

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



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

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

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

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

So it is clear that  $F_n \rightarrow F$  <sup>a.p.</sup> pointwise (i.e.  $\forall a_0$ )

Now let's look at uniform convergence.

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

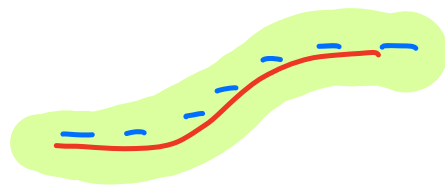
Let  $F_n$  be the empirical cdf induced by the sample. Then:

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

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

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

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



Proof

Observe:  $LLN \Rightarrow P(|F_n(a_0) - F(a_0)| > \varepsilon) \rightarrow 0$   
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:

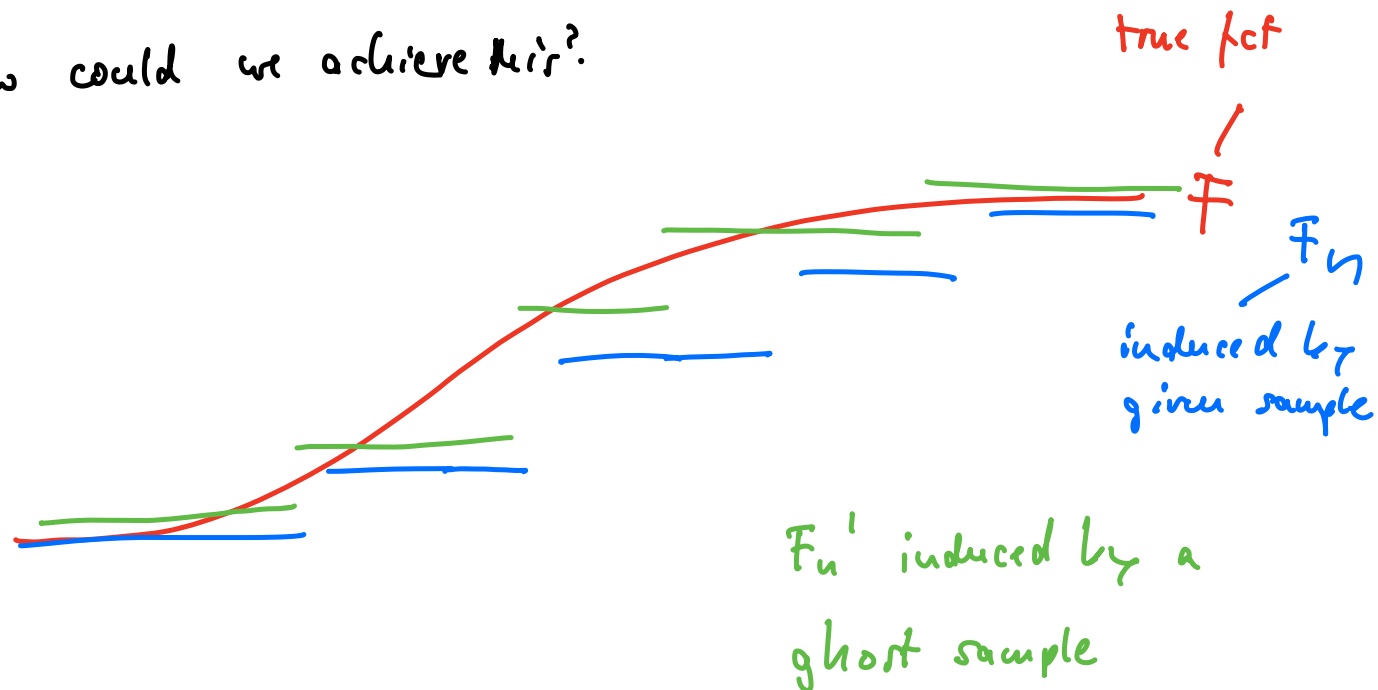
$$P\left(\max_{i=1 \dots n} |U_i| > \varepsilon\right) =$$

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

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

Trick of the proof: cannot  $\sup_{a \in \mathcal{R}}$  to something "finite".

How could we achieve this?



$|red - green|$

$$|red - blue| \leq 2 |green - blue|$$

Step 1 : Symmetrization by ghost sample

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

Denote by  $F_n'$  the empirical cdf induced by ghost sample

Now it is easy to prove:

$$P\left(\sup_a |\underbrace{F_n(a)}_{\text{blue}} - \underbrace{F(a)}_{\text{red}}| > \varepsilon\right)$$

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

Step 2 : Want to split this in two terms

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

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

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

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

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

Now we have:

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

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

$$\leq 2P\left(\sup_a \left| \underbrace{\frac{1}{n} \sum \sigma_i 1_{X_i \leq a}}_u \right| > \frac{\varepsilon}{4}\right) + 2P\left(\sup_a \left| \underbrace{\frac{1}{n} \sum \sigma_i 1_{X_i' \leq a}}_v \right| > \frac{\varepsilon}{4}\right)$$

Observe:

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

↑ right side is necessary for left side

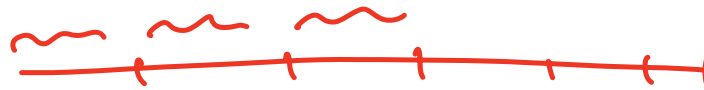
$$= 4 \cdot P\left(\sup_a \left| \frac{1}{n} \sum \sigma_i 1_{\{X_i \leq a\}} \right| > \frac{\varepsilon}{4}\right)$$

Step 3

Exploit "finite structure":

Fix  $x_1, \dots, x_n$  ( $\equiv$  condition on  $x_1, \dots, x_n$ )

We look at  $\mathbb{1}_{x_i \leq a}$



The var  $\mathbb{1}_{x_1 \leq a}, \dots, \mathbb{1}_{x_n \leq a}$  for fixed  $a$  run on by

have  $n+1$  realizations

$$P\left(\sup_a \frac{1}{n} \left| \sum \sigma_i \mathbb{1}_{x_i \leq a} \right| > \frac{\varepsilon}{4} \mid x_1, \dots, x_n\right) \leq$$

$$\leq (n+1) \sup_a \underbrace{P\left(\frac{1}{n} \left| \sum \sigma_i \mathbb{1}_{\{x_i \leq a\}} \right| > \frac{\varepsilon}{4} \mid x_1, \dots, x_n\right)}$$

use Hoeffding (#)



Step 4      apply  
Hoeffding to  $(\pm 1)$

Let:

$$P\left(\frac{1}{n} \left| \sum \sigma_i \right| > \frac{\varepsilon}{4} \mid x_1 \dots x_n\right)$$

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

Combining everything gives the theorem.

Product space  
joint distributions

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

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

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

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

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

Example in ML:  $(x, y)$  where  $x$  is the input data,  $y$  is the label

# Product measure

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

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

## Product $\sim$ independence

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

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

Marginal distribution

# Marginal distribution

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



Example in the discrete case:

$Y \backslash X$	$x_1$	$x_2$	$x_3$	$\Sigma$
$Y_1$	$p_1$	$p_2$	$p_3$	$p_1 + p_2 + p_3 = P(Y = Y_1)$
$Y_2$	$p_4$	$p_5$	$p_6$	$p_4 + p_5 + p_6 = P(Y = Y_2)$
	$p_1 + p_4$ $= P(X = x_1)$	$p_2 + p_5$ $= P(X = x_2)$	$p_3 + p_6$ $= P(X = x_3)$	

marginal wrt  $X$

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

## Marginal distributions in case of densities

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

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

*joint density* (pointing to  $f(x, y)$ )  
*sum over  $Y$*  (pointing to  $dy$ )

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

(2)  $X$  and  $Y$  are independent iff

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

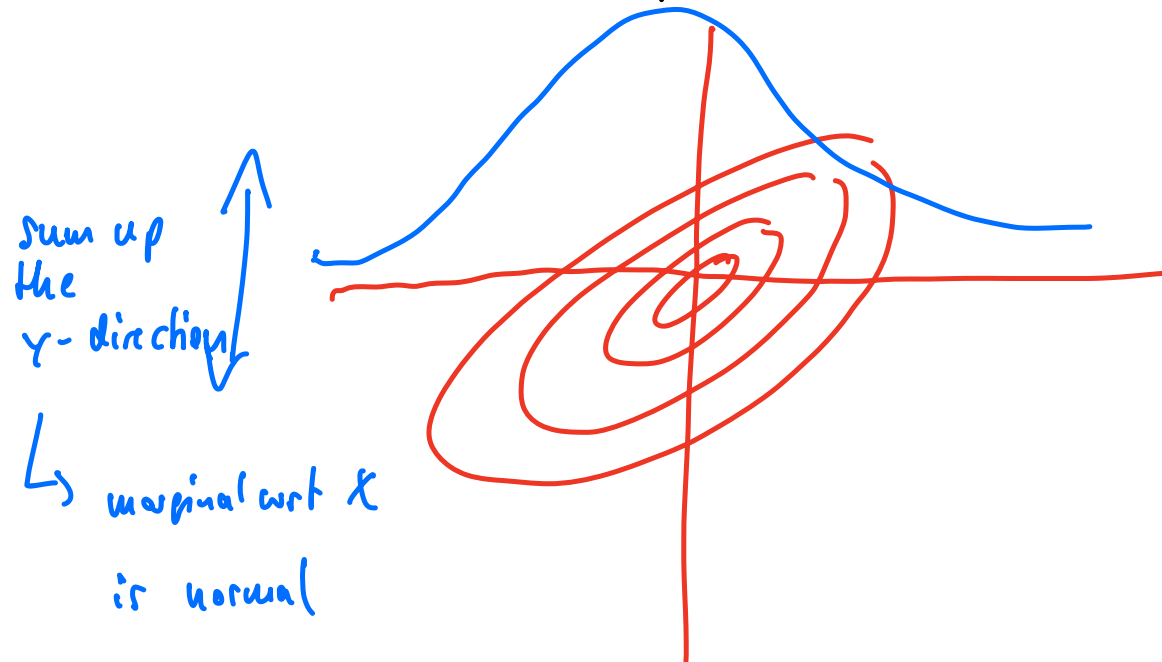
## Special case: marginals of multivariate normal distribution

2 dim Consider a 2-dim normal rv  $X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$  with mean

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

Then the marginal distribution of  $X$  wrt  $x_1$  is again

a normal distribution with mean  $\mu_1$  and var  $\sigma_1^2$ .



n-dim

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n.$$

Group the variables:

$$\begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix} \} \tilde{X} \in \mathbb{R}^k$$

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

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

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

$$\Sigma = \left( \begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array} \right) \} k$$

$k$

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

Conditional distributions

# Conditional distributions in the discrete case

Discrete case:

Know conditional probabilities:  $P(A | B)$

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

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

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

This is a probability measure.

## Conditional distributions in case of densities

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

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

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

General case (not discrete, no density)

For general rv this is surprisingly complicated!

$\leadsto$  "regular conditional probabilities"  $\leadsto$  skipped



Example: normal distributions

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

If  $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \sim N(\mu, \Sigma)$ , then the conditional distributions  
of  $\tilde{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$  w.r.t  $x^\# = \begin{pmatrix} x_{n+1} \\ \vdots \\ x_n \end{pmatrix}$  is given by

$$p_{\tilde{x} | x^\#} \sim N \left( \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x^\# - \tilde{\mu}), \right. \\ \left. \Sigma_{22} - \Sigma_{12}^\top \Sigma_{11}^{-1} \Sigma_{12} \right).$$

Conditional expectations

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

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

assume  $X$  takes finitely (countably) many values

$x_1, \dots, x_n \in \mathbb{R}$ ,  $Y$  takes finitely (countably) many values

$y_1, \dots, y_m \in \mathbb{R}$ , always with a positive probability.

Then we define the conditional expectations

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

## Example

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

$$\begin{aligned} E(\text{sum} \mid X=1) &= \sum_{i=1}^{12} i \cdot P(\text{sum} = i \mid X=1) \\ &= \sum_{k=1}^6 (1+k) \cdot P(Y=k \mid X=1) \\ &= \sum_{k=1}^6 (1+k) P(Y=k) = \sum_{k=1}^6 (1+k) \cdot \frac{1}{6} = 4.5 \end{aligned}$$

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

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

## Cond. expectation wrt a rv

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

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

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

⚠  $E(Y|X)$  is only defined a.s.

## General case?

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

Case of densities is still straight forward:

## Case of joint densities

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

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

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

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

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

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

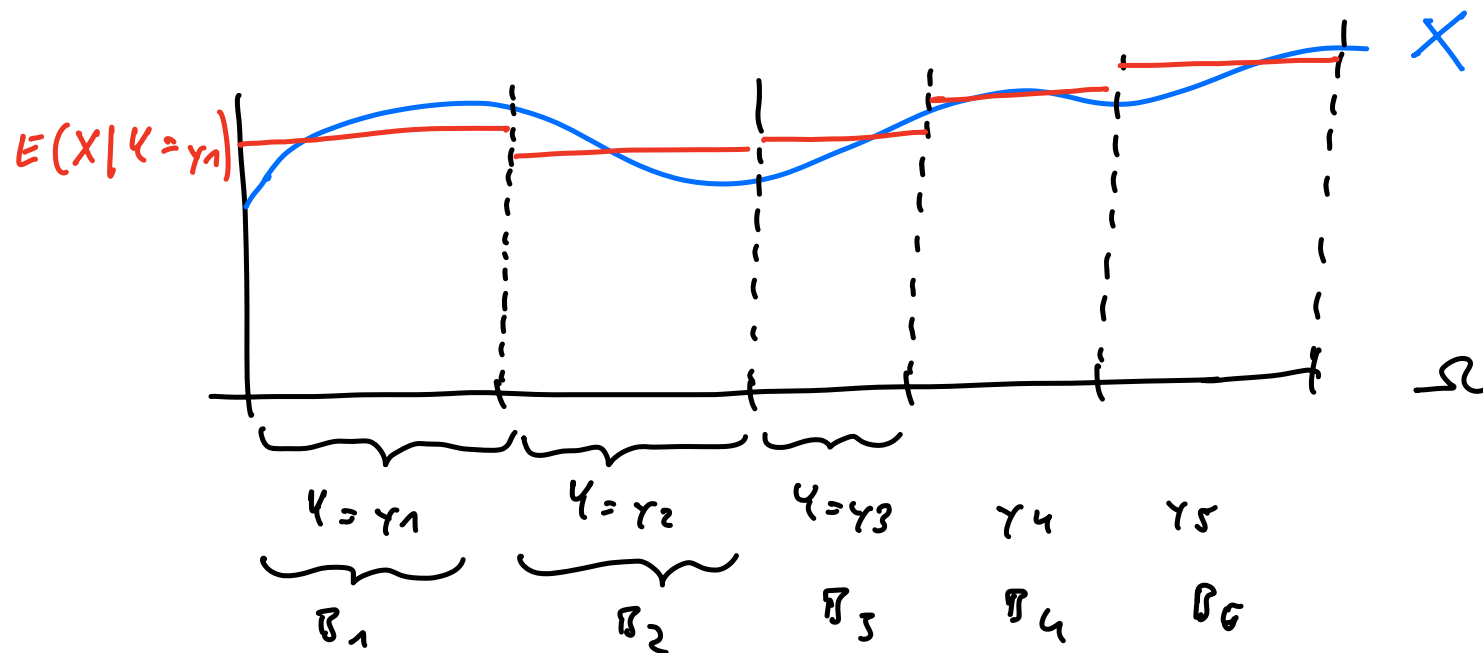
$$E(Y|X) = h(X).$$

# Idea for the more general case

Consider:  $X$  continuous rv

$Y$  discrete rv  $\leadsto Y_1, \dots, Y_5$

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



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

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



# Cond. expectation on $L_1$

Def (Conditional expectation on  $L_1$ )

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

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

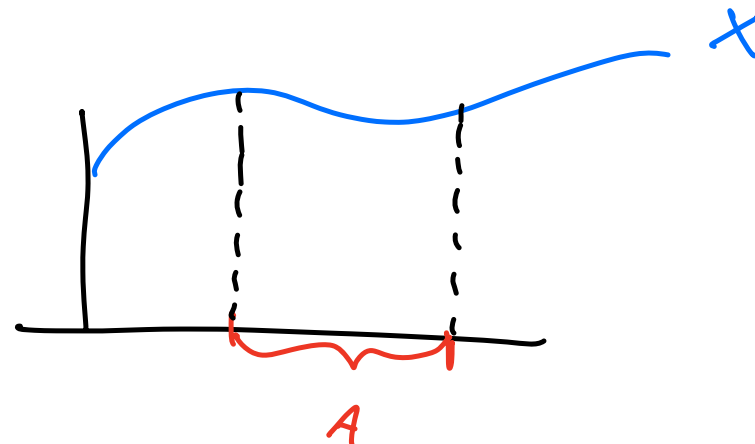
We now define the cond. exp. of  $X$  given  $\mathcal{F}$

$E(X | \mathcal{F})$  as any random variable  $Z$  that satisfies

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

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

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



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

### Examples

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