Recap on probability theory

Felipe Uribe

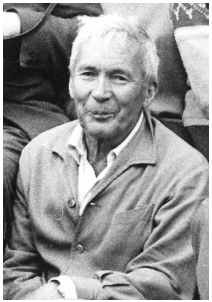
Computational Engineering
School of Engineering Sciences
Lappeenranta-Lahti University of Technology (LUT)

Special Course on Inverse Problems
Lappeenranta, FI — January-February, 2024

1) Probability recap: introduction

Probability theory: basic timeline

- Pierre–Simon Laplace (1814) - *Théorie analytique des probabilités*.
- Henri Poincaré (1896) - *Calcul des probabilités*.
- Andrey N. Kolmogorov (1933) - axiomatic foundation of probability theory.
- Harold Jeffreys (1939) - revival of the Bayesian view of probability.
- Richard T. Cox (1946) - laws of "logical" probability theory.



§ 1. Axioms[1]

Let $\mathfrak{S}$ be a collection of elements $\xi, \eta, \zeta, \ldots$, which we shall call *elementary events*, and $\mathfrak{F}$ a set of subsets of E; the elements of the set $\mathfrak{F}$ will be called *random events*.

I. $\mathfrak{F}$ *is a field*[2] *of sets.*
II. $\mathfrak{F}$ *contains the set* E.
III. *To each set* A *in* $\mathfrak{F}$ *is assigned a non-negative real number* $\mathrm{P}(A)$. *This number* $\mathrm{P}(A)$ *is called the probability of the event* A.
IV. $\mathrm{P}(E)$ *equals* 1.
V. *If* A *and* B *have no element in common, then*

$$\mathrm{P}(A+B) = \mathrm{P}(A) + \mathrm{P}(B)$$

A system of sets, $\mathfrak{F}$, together with a definite assignment of numbers $\mathrm{P}(A)$, satisfying Axioms I-V, is called a *field of probability.*

Our system of Axioms I-V is *consistent.* This is proved by the following example. Let E consist of the single element ξ and let $\mathfrak{F}$ consist of E and the null set 0. $\mathrm{P}(E)$ is then set equal to 1 and $\mathrm{P}(0)$ equals 0.

Why probability theory?

- Bertrand Russell, 1929 Lecture: "Probability is the most important concept in modern science, especially as *nobody* has the slightest notion what it means".

Why probability theory?

- Bertrand Russell, 1929 Lecture: "Probability is the most important concept in modern science, especially as *nobody* has the slightest notion what it means".
- Probability theory: mathematical theory in charge of the analysis and modeling of random phenomena.
  - A random phenomena or "experiment" is one that despite being performed under the same determined conditions produces different results $\Longrightarrow$ coin toss, earthquakes.
  - This is opposite to deterministic phenomena, whose results are always unique and predictable $\Longrightarrow$ speed of light, intrinsic material parameters.

Why probability theory?

- Bertrand Russell, 1929 Lecture: "Probability is the most important concept in modern science, especially as *nobody* has the slightest notion what it means".
- Probability theory: mathematical theory in charge of the analysis and modeling of random phenomena.
  - A random phenomena or "experiment" is one that despite being performed under the same determined conditions produces different results $\Longrightarrow$ coin toss, earthquakes.
  - This is opposite to deterministic phenomena, whose results are always unique and predictable $\Longrightarrow$ speed of light, intrinsic material parameters.
- Formal probability theory works with sets in a given space. Because of this we need to review some relevant results from set theory.

2) Probability recap: probability spaces

Elements of set theory: intro[1]

- A set is a collection of elements. A space Ω is the collection of all elements under consideration.
- A point or atomic set is a set containing a single element, e.g., $\{5\}$. The entire space Ω itself is always a valid set, as is the empty set or null set $\emptyset$, which contains no elements at all.
- Sets are often defined implicitly via an inclusion criterion. These sets are denoted with the *set builder notation*, e.g., $\mathbb{R}^+ = \{x \in \mathbb{R} \mid x > 0\}$.
- There are three natural operations between sets: *complement, union and intersection*.

[1] M. Betancourt. *Probability Theory (For Scientists and Engineers)*. `https://betanalpha.github.io/assets/case_studies/probability_theory.html`. 2021.

Elements of set theory: natural operations

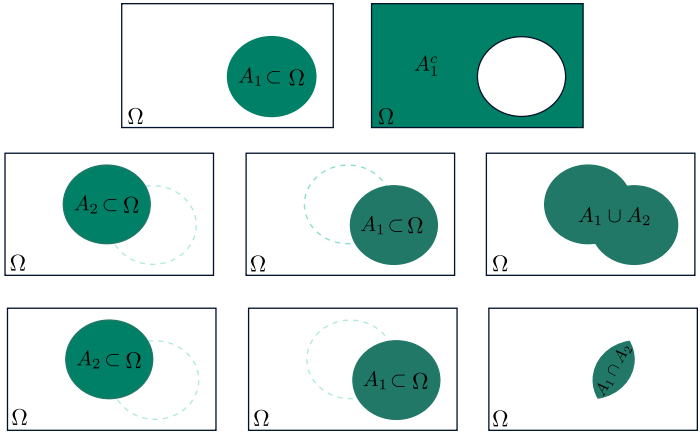


Figure: Elementary set operations: complement, union and intersection (by rows).

Elements of set theory: σ-algebras[2]

- The collection of all sets in a space Ω, is called the power set $\mathcal{P}(\Omega)$. The power set is *massive*; it contains the empty set, the entire space, all of the atomic sets, and more.
- Moreover, even if the space Ω is well-behaved (e.g., the real numbers), $\mathcal{P}(\Omega)$ can contain some mathematically "pathological" elements (namely non-measurable sets). See this nice **video** for a proof.
- We have to define a restriction: *σ-algebras are the patch that fixes the math*.

[2] M. Betancourt. *Probability Theory (For Scientists and Engineers)*. `https://betanalpha.github.io/assets/case_studies/probability_theory.html`. 2021.

Elements of set theory: σ-algebras[3]

- The collection of all sets in a space Ω, is called the power set $\mathcal{P}(\Omega)$. The power set is *massive*; it contains the empty set, the entire space, all of the atomic sets, and more.
- Moreover, even if the space Ω is well-behaved (e.g., the real numbers), $\mathcal{P}(\Omega)$ can contain some mathematically "pathological" elements (namely non-measurable sets). See this nice **video** for a proof.
- We have to define a restriction: *σ-algebras are the patch that fixes the math*.

σ-algebra

Consider a restricted collection of sets in Ω that is *closed under the three natural set operations*. Further, it is closed under a *countable* number of unions/intersections. Such collection is called a **σ-algebra** over the space Ω.

[3] M. Betancourt. *Probability Theory (For Scientists and Engineers)*. `https://betanalpha.github.io/assets/case_studies/probability_theory.html`. 2021.

Another important concept: measures

- Measures provide a mathematical abstraction for common notions like mass, distance/length, area, volume, probability of events. Hence, they are directly related to Lebesgue integration.

- Recall that Ω might contain subsets that are so strange[4] that it is impossible to define a geometrically reasonable notion of measure for them. Hence, *σ-algebras serve as the domains for measures*.

[4] see, e.g., the Banach–Tarski paradox.

Another important concept: measures

- Measures provide a mathematical abstraction for common notions like mass, distance/length, area, volume, probability of events. Hence, they are directly related to integration.
- Recall that Ω might contain subsets that are so strange that it is impossible to define a geometrically reasonable notion of measure for them. Hence, *σ-algebras serve as the domains for measures*.

Measures ("soft definition")

Let Ω be a set and $\mathcal{F}$ a σ-algebra over Ω. A function ν from $\mathcal{F}$ to the extended real line is called a **measure**, if it satisfies the following properties: (i) non-negativity, (ii) $\nu(\emptyset) = 0$, (iii) countable additivity (σ-additive).

What is a probability space?

We define a probability space [3] as the triple $(\Omega, \mathcal{F}, \mathbb{P})$, where

- the sample space Ω is a non-empty set containing the outcomes of a random experiment (elementary events),

What is a probability space?

We define a probability space [3] as the triple $(\Omega, \mathcal{F}, \mathbb{P})$, where

- the sample space Ω is a non-empty set containing the outcomes of a random experiment (elementary events),
- the σ-algebra $\mathcal{F}$ is a collection of subsets of Ω, satisfying: (i) $\varnothing \in \mathcal{F}$ and $\Omega \in \mathcal{F}$, (ii) $A \in \mathcal{F}$ implies that $A^c \in \mathcal{F}$, and (iii) $A_1, A_2, ... \in \mathcal{F}$ implies that $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$. The elements of $\mathcal{F}$ are called *events* (or measurable sets), and

What is a probability space?

We define a probability space [3] as the triple $(\Omega, \mathcal{F}, \mathbb{P})$, where

- the sample space Ω is a non-empty set containing the outcomes of a random experiment (elementary events),
- the σ-algebra $\mathcal{F}$ is a collection of subsets of Ω, satisfying: (i) $\varnothing \in \mathcal{F}$ and $\Omega \in \mathcal{F}$, (ii) $A \in \mathcal{F}$ implies that $A^c \in \mathcal{F}$, and (iii) $A_1, A_2, ... \in \mathcal{F}$ implies that $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$. The elements of $\mathcal{F}$ are called *events* (or measurable sets), and
- the probability measure $\mathbb{P}$ is a mapping $\mathbb{P} : \mathcal{F} \to [0, 1]$, such that (i) $\mathbb{P}$ is real and non-negative, (ii) $\mathbb{P}$ is σ-additive, i.e., $\mathbb{P}[\cup_{i=1}^{\infty} A_i] = \sum_{i=1}^{\infty} \mathbb{P}[A_i]$, for mutually disjoint events A_i (consistent allocation), and (iii) $\mathbb{P}[\varnothing] = 0$ and $\mathbb{P}[\Omega] = 1$ (lossless allocation).

Probability space: example I

- $\Omega = \{\text{head}, \text{tail}\}$, $\Omega = \{1, 2, 3, 4, 5, 6\}$, and $\Omega = [a, b] \subset [0, \infty)$ are sample spaces for the experiments of tossing a coin, rolling a dice and measuring daily rainfall.

Probability space: example I

- $\Omega = \{\text{head}, \text{tail}\}$, $\Omega = \{1, 2, 3, 4, 5, 6\}$, and $\Omega = [a, b] \subset [0, \infty)$ are sample spaces for the experiments of tossing a coin, rolling a dice and measuring daily rainfall.
- The σ-algebra associated with the game in which one wins 10 and looses 5 for outcomes of a die rolling experiment in $\{1, 2\}$ and $\{3, 4, 5, 6\}$ is $\mathcal{F} = \{\emptyset, \{1, 2\}, \{3, 4, 5, 6\}, \Omega\}$.

Probability space: example I

- $\Omega = \{\text{head}, \text{tail}\}$, $\Omega = \{1, 2, 3, 4, 5, 6\}$, and $\Omega = [a, b] \subset [0, \infty)$ are sample spaces for the experiments of tossing a coin, rolling a dice and measuring daily rainfall.
- The σ-algebra associated with the game in which one wins 10 and looses 5 for outcomes of a die rolling experiment in $\{1, 2\}$ and $\{3, 4, 5, 6\}$ is $\mathcal{F} = \{\emptyset, \{1, 2\}, \{3, 4, 5, 6\}, \Omega\}$.
- Finally, for the probability measure, we would map each event to the number of outcomes in that event divided by 6. Hence, $\{1, 2\}$ would be mapped to 2/6=1/3, and $\{3, 4, 5, 6\}$ would be mapped to 4/6=2/3. [5]

[5] Point/counting measures are used in the discrete space context; basically they count the number of elements in the set one is measuring.

Probability space: example II

- $\Omega = \{0, 1\}$, a binary sample space.
- In this case, the valid σ-algebra is the entire power set[6] consisting of the empty set $A_1 = \emptyset$, the atomic sets $A_2 = 0$ and $A_3 = 1$, and the entire space $A_4 = \Omega$.
- Finally, what probabilities can we assign to these sets? The axioms require that $\mathbb{P}[A_4] = 1$, and the complement rule then requires that $\mathbb{P}[A_1] = 0$. We are free to assign any probability to one of the atomic sets, so we can take $\mathbb{P}[A_3] = p$ in which case the complement rule requires that $\mathbb{P}[A_2] = 1 - p$.

[6] If we restrict ourselves to countable sets, then we can take $\mathcal{F} = \mathcal{P}(\Omega)$ and we won't have any problems because for countable Ω, $\mathcal{P}(\Omega)$ consists only of measurable sets.

Probability space: example III

- $\Omega = \{0, 1, 2, \ldots\}$ a sample space consisting of non-negative integers.
- $\mathcal{F} = \{\text{all subsets of } \Omega\}$ (power set of Ω).
- We can define the probability measure, for any $A \in \mathcal{F}$

$$\mathbb{P}[A] = \exp(-5) \sum_{k \in A} \frac{5^k}{k!}. \tag{1}$$

Probability space: example III

- $\Omega = \{0, 1, 2, \ldots\}$ a sample space consisting of non-negative integers.
- $\mathcal{F} = \{\text{all subsets of } \Omega\}$ (power set of Ω).
- We can define the probability measure, for any $A \in \mathcal{F}$

$$\mathbb{P}[A] = \exp(-5) \sum_{k \in A} \frac{5^k}{k!}. \tag{2}$$

This probability triple represents a **Poisson distribution** with rate parameter 5.

Probability space: example IV

- $\Omega = [0, 1]$ a sample space consisting of real numbers in that interval.
- $\mathcal{F} = \{\text{all intervals contained in } \Omega\}$, Borel sets on [0,1].
- We can define the probability measure, for any $I \in \mathcal{F}$, simply as the length of the interval

$$\mathbb{P}[A] = \text{length}(I). \tag{3}$$

Probability space: example IV

- $\Omega = [0, 1]$ a sample space consisting of real numbers in that interval.
- $\mathcal{F} = \{\text{all intervals contained in } \Omega\}$, Borel sets on [0,1].
- We can define the probability measure, for any $I \in \mathcal{F}$, simply as the length of the interval

$$\mathbb{P}[A] = \text{length}(I). \tag{4}$$

This probability triple represents a **uniform distribution**[7] on [0,1], or Lebesgue measure on [0,1]. More generally, the Lebesgue measure on Borel sets in $\mathbb{R}^d$, is given by

$$\lambda((a_1, b_1] \times \cdots \times (a_d, b_d]) = (b_1 - a_1) \cdots (b_d - a_d) \quad \forall\ a_i < b_i. \tag{5}$$

[7] This is a very simplified way of writing it, the proper construction requires more involved concepts such as, semialgebras, the extension theorem, inner and outer measures, among others (the interested student can see, e.g., [4, Sec. 2.4]).

Summary I

- As a rule of thumb, if the sample space Ω is finite or countable, then $\mathcal{F} = \mathcal{P}(\Omega)$, the collection of all subsets of the sample space. If Ω is a Borel subset of the Euclidean space $\mathbb{R}^n$, then $\mathcal{F} = \mathcal{B}(\Omega)$, the Borel sets in $\mathbb{R}^n$ intersected with Ω.
- A probability distribution defined by Kolmogorov's axioms is completely specified by the probability triple.
- We will see that probability is simply a positive and conserved quantity that we want to distribute across a given space. The probability distribution defines a mathematically self-consistent allocation of this conserved across Ω.

Basic definition

Probability theory is simply the study of an object, a probability distribution/measure that assigns values between 0 and 1 to sets (and the transformations of that object).

3) Probability recap: random variables

Random variable as measurable transformations

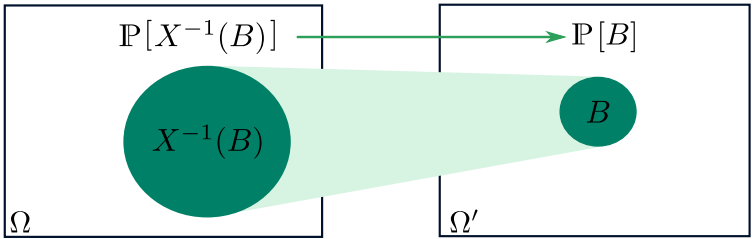


Figure: In measure-theoretic terms, the random variable X 'pushes-forward' the measure $\mathbb{P}$ on Ω to a measure $\mathbb{P}_X$ on Ω'. In this general setting, we see that a random variable defines a new random experiment with Ω' as the new set of outcomes and $\mathcal{F}'$ as the new collection of events.

Random variables

- Once we have defined a probability distribution on Ω, and a well-behaved collection of subsets $\mathcal{F}$, then we can consider how the probability distribution transforms when Ω transforms.
- Recall that Ω is the set of all possible outcomes of some random experiment. A random variable assigns a numerical value to each of these outcomes.

Random variable

Given a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, a (real-valued) **random variable** is a function $X : \Omega \to \mathbb{R}$, such that $X^{-1}(B) \in \mathcal{F}$, for every $B \in \mathbb{R}$ (i.e., X is a measurable function).

Random variables: example

Consider the experiment of rolling a fair dice:

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_6\}, \quad \text{subindex is the number of dots in the face of the dice} \tag{6a}$$

$$\mathcal{F} = \mathcal{P}(\Omega) \quad \text{all possible subsets of } \Omega \tag{6b}$$

$$\mathbb{P}[\omega_i] = 1/6 \quad \text{counting measure.} \tag{6c}$$

Define a RV $X : \Omega \to \mathbb{R}$ by $X(\omega_n) = 0$, if n is even and $X(\omega_n) = 1$, if n is odd. Note that since the outcomes are random X takes the values 0 or 1, randomly.

Random variables: example

Consider the experiment of rolling a fair dice:

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_6\}, \quad \text{subindex is the number of dots in the face of the dice} \tag{8a}$$

$$\mathcal{F} = \mathcal{P}(\Omega) \quad \text{all possible subsets of } \Omega \tag{8b}$$

$$\mathbb{P}[\omega_i] = 1/6 \quad \text{counting measure.} \tag{8c}$$

Define a RV $X : \Omega \to \mathbb{R}$ by $X(\omega_n) = 0$, if n is even and $X(\omega_n) = 1$, if n is odd. Note that since the outcomes are random X takes the values 0 or 1, randomly. In this case, we have

$$\mathbb{P}[X = 0] = \mathbb{P}[\omega_2, \omega_4, \omega_6] = 1/2 \tag{9a}$$

$$\mathbb{P}[X = 1] = \mathbb{P}[\omega_1, \omega_3, \omega_5] = 1/2 \tag{9b}$$

since X takes the value of 1 with probability p and the value of 0 with probability $1 - p$ ($p = 1/2$), it is an example of a *Bernoulli RV*, $X \sim \text{Bern}(1/2)$.

4) Probability recap: probability distributions

Probability distribution

- We want to compute probabilities of events associated to a random variable X, defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Instead of considering a particular value of X, we describe the distribution of the values it takes.

Probability distribution

- We want to compute probabilities of events associated to a random variable X, defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Instead of considering a particular value of X, we describe the *distribution* of the values it takes.

- Such probabilities are specified by the distribution/law/measure of X, which is defined as

$$\mathbb{P}_X(B) := \mathbb{P} \circ X^{-1}(B) = \mathbb{P}[X \in B], \quad B \in \mathcal{B}(\mathbb{R}), \tag{12}$$

  where $\mathcal{B}(\mathbb{R})$ is the collection of Borel sets on $\mathbb{R}$, and the resulting $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$ is valid probability space.

- The distribution/law/measure of X can be studied using its cumulative distribution function (CDF). The CDF is defined as

$$F_X(x) := \mathbb{P}[X \leq x] = \mathbb{P}[\omega : X(\omega) \leq x] \quad \text{for any } x \in \mathbb{R}. \tag{13}$$

Probability distribution functions

- The following are the main properties of the CDF:
  - increasing, $a \leq b \implies F_X(a) \leq F_X(b)$.
  - right-continuous, $\lim_{x \to x_0^+} F_X(x) = F_X(x_0)$.
  - satisfies, $\lim_{x \to -\infty} F_X(x) = 0, \lim_{x \to \infty} F_X(x) = 1$.

Probability distribution functions

- The following are the main properties of the CDF:
  - increasing, $a \leq b \implies F_X(a) \leq F_X(b)$.
  - right-continuous, $\lim_{x \to x_0^+} F_X(x) = F_X(x_0)$.
  - satisfies, $\lim_{x \to -\infty} F_X(x) = 0, \lim_{x \to \infty} F_X(x) = 1$.

- Note that F_X and $\mathbb{P}$ "correspond to each other". By definition of the CDF, the probability that X lies in the semi-closed interval $(a, b]$, where $a < b$, is[8]

$$\mathbb{P}[a < X \leq b] = F_X(b) - F_X(a) = \mathbb{P}[(a, b]], \tag{14}$$

  this is because the intervals $(a, b]$ are in particular Borel sets on $\mathbb{R}$.

[8] R. B. Ash and C. Doléans-Dade. *Probability and measure theory*. 2nd ed. Harcourt/Academic Press, 2000.

Probability density functions

- Mass/density: "*how much more likely is that the random variable would be close to one sample compared to others*"
  - For discrete (or simple): X takes on only a finite number of different values $x_1, x_2, \ldots$, then we can define its probability mass function (PMF), $\pi_X(x) = \mathbb{P}[X = x]$. In this case, the CDF and the PMF are connected by the relation

$$F_X(x) = \mathbb{P}[X \leq x] = \sum_{x_i \leq x} \mathbb{P}[X = x_i] = \sum_{x_i \leq x} \pi_X(x_i). \tag{15}$$

  - For continuous: If F_X is absolutely continuous, then we can define its probability density function (PDF), which is a Lebesgue-integrable function $\pi_X(x)$ such that

$$F_X(x) = \mathbb{P}[X \leq x] = \int_{-\infty}^{x} \pi_X(t)\,\mathrm{d}t; \tag{16}$$

    the PDF is equal to the derivative of the CDF almost everywhere ($\pi(x) = \mathrm{d}F(x)/\mathrm{d}x$).

Other probability functions

- Quantile function (inverse CDF): If the CDF is strictly increasing and continuous then $F^{-1}(p), p \in [0,1]$, is the unique real number x, such that $F(x) = p$. More descriptively, *a p-quantile is the point where we have accumulated p probability*.
  - The most common quantile is 0.5, or the median, which quantifies a sense of where a probability distribution concentrates (other than the mean).
  - Tail quantiles, such as 0.05 and 0.95, quantify a sense of the spread of a probability distribution (other than the variance).
- Tail distribution (complementary CDF): $\bar{F}_X(x) = \mathbb{P}[X > x] = 1 - F_X(x)$. This is used when computing the probabilities of rare/tail events are required.

5) Probability recap: main statistics

Expected values

- In general, if X is a random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then the expected value[9] of X, is defined as the integral

$$\mathbb{E}[X] = \int_{\Omega} X(\omega)\, \mathrm{d}\mathbb{P}(\omega) = \int_{\mathbb{R}} x\, \mathrm{d}\mathbb{P}_X(x). \tag{17}$$

- In words, the expected value of X with respect to $\mathbb{P}$ on Ω is equal to the expected value of its realized value with respect to the distribution/measure $\mathbb{P}_X$ on $\mathbb{R}$.

  - For discrete:

$$\mathbb{E}[X] = \sum_{i=1}^{k} x_i\, \mathbb{P}[X = x_i]\,. \tag{18}$$

  - For continuous:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, \pi_X(x)\, \mathrm{d}x. \tag{19}$$

[9] Back in 1814, Laplace used to call it: *mathematical hope*.

Expected values

- Oftentimes, we use the notation $\mu_X = \mathbb{E}[X]$. Sometimes, we call the expected value of X, with respect to the Lebesgue measure probability triple, its Lebesgue integral.
- Consider the random variables X, Y and scalars a, b, then the following statements about the expectation operator hold[10]:
  - The expectation is linear, $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$.
  - The expectation is order-preserving, if $X \leq Y$, then $\mathbb{E}[X] \leq \mathbb{E}[Y]$.
  - The expectation follows the generalized triangle inequality, $|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$.
  - If X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$.
- In general, the kth moment of X is given by

$$\mathbb{E}\left[X^k\right] = \int_{-\infty}^{\infty} x^k \, \pi_X(x) \, \mathrm{d}x. \tag{20}$$

[10] J. S. Rosenthal. *A first look at rigorous probability theory*. 2nd ed. World Scientific Publishing Company, 2006.

Variance

- Besides the expected value, the second central moment or variance of the random variable X, denoted $\mathbb{V}[X]$, is also important. This function gives information about the variation of X and it is defined as $\mathbb{V}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mu_X^2$.
- Some properties include[11]:
  - $0 \leq \mathbb{V}[X] \leq \mathbb{E}\left[X^2\right]$,
  - $\mathbb{V}[aX + bY] = a^2\mathbb{V}[X] + b^2\mathbb{V}[Y] + 2ab\mathbb{C}\text{ov}[X, Y]$, where $\mathbb{C}\text{ov}[X, Y]$ is the *covariance* between X and Y.
  - The positive square root $\sigma_X = \sqrt{\mathbb{V}[X]}$ is called the standard deviation of X.
- The mean and the variance do not give, in general, enough information to completely specify the distribution of a random variable. However, they may provide useful bounds.

[11] J. S. Rosenthal. *A first look at rigorous probability theory*. 2nd ed. World Scientific Publishing Company, 2006.

Median and MAD

- When the first two moments of a random variable do not exist, location and scale characteristics of its distribution can be summarized using for example, function of the median.
- If X is a random variable, then a median of its distribution is any value med such that $\mathbb{P}[X \geq \text{med}] \geq 1/2$. This value is not necessarily unique.
- If the distribution of X has an unique med, then the *median absolute deviation* is defined as:

$$\text{MAD}(X) = \text{Median}(|X - \text{med}|). \tag{21}$$

The story so far...

- Formal probability theory is simply the study of (i) probability measures that distribute a finite and conserved quantity across a space, (ii) the expectation values that such a distribution induces, (iii) and how the distribution behaves under transformations of the underlying space.
- Still, those concepts have so far been presented in the abstract without any concrete examples to provide context. Why? Because, unfortunately, probability distributions cannot be explicitly defined in most problems!

The story so far...

- Formal probability theory is simply the study of (i) probability measures that distribute a finite and conserved quantity across a space, (ii) the expectation values that such a distribution induces, (iii) and how the distribution behaves under transformations of the underlying space.
- Still, those concepts have so far been presented in the abstract without any concrete examples to provide context. Why? Because, unfortunately, probability distributions cannot be explicitly defined in most problems!
- Fortunately, most probabilistic systems admit representations that faithfully recover all probabilities and expectation values on demand and hence completely specify a given probability distribution.
- In particular, density representations exploit the structure of Ω to fully characterize a probability distribution with special functions.

6) Probability recap: joint distributions

Conditional probability and independence

- A common problem is the computation of the probability of an event, given that another event B has occurred. This leads to the definition of conditional probability:

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}. \tag{22}$$

- This can be generalized to the product rule of probability (or chain rule), which states that for any sequence of events $A_1, \ldots, A_n$,

$$\mathbb{P}[A_1 \cdots A_n] = \mathbb{P}[A_1]\,\mathbb{P}[A_2 \mid A_1]\,\mathbb{P}[A_3 \mid A_1 A_2] \cdots \mathbb{P}[A_n \mid A_1 \cdots A_{n-1}]\,. \tag{23}$$

- Informally, events or random variables are independent, if they do not affect each other's probabilities, hence

$$\mathbb{P}[A_1 \cdots A_n] = \mathbb{P}[A_1]\,\mathbb{P}[A_2] \cdots \mathbb{P}[A_n]\,, \quad \text{holds for arbitrary permutations too.} \tag{24}$$

Conditional probability and independence

- Suppose that $B_1, \ldots, B_n$ is a disjoint partition of Ω and their union is Ω. By the sum rule $\mathbb{P}[A] = \sum_{i=1}^{n} \mathbb{P}[A \mid B_i]$, and hence, by the definition of conditional probability we obtain the total probability theorem:

$$\mathbb{P}[A] = \sum_{i=1}^{n} \mathbb{P}[A \mid B_i]\, \mathbb{P}[B_i]\,. \tag{25}$$

- Combining eq. (25) with the definition of conditional probability, we obtain Bayes' theorem

$$\mathbb{P}[B_j \mid A] = \frac{\mathbb{P}[A \mid B_j]\, \mathbb{P}[B_j]}{\sum_{i=1}^{n} \mathbb{P}[A \mid B_i]\, \mathbb{P}[B_i]}. \tag{26}$$

Example

- Let A be the event of a positive diagnostic of a rare disease, and B be the event that a person gets a rare disease (i.e., is sick), which has probability $\mathbb{P}[B] = 0.001$. The probability that a positive diagnostic is correct given the person is sick, is $\mathbb{P}[A \mid B] = 0.99$. However, there exists a false-positive probability of $\mathbb{P}[A \mid \neg B] = 0.05$.

  Therefore, given a positive test result, what is the probability that the person is actually sick?

Example

- Let A be the event of a positive diagnostic of a rare disease, and B be the event that a person gets a rare disease (i.e., is sick), which has probability $\mathbb{P}[B] = 0.001$. The probability that a positive diagnostic is correct given the person is sick, is $\mathbb{P}[A \mid B] = 0.99$. However, there exists a false-positive probability of $\mathbb{P}[A \mid \neg B] = 0.05$.

  Therefore, given a positive test result, what is the probability that the person is actually sick?

$$\mathbb{P}[B \mid A] = \frac{\mathbb{P}[A \mid B]\,\mathbb{P}[B]}{\mathbb{P}[A]} = \frac{0.09 \cdot 0.001}{\mathbb{P}[A \mid B]\,\mathbb{P}[B] + \mathbb{P}[A \mid \neg B]\,\mathbb{P}[\neg B]} = 0.19$$

Joint distributions

- Consider a d-dimensional *random vector* on a probability space is the function $\boldsymbol{X} : \Omega \to \mathbb{R}^d$. Random vectors can be regarded as d-tuples of random variables, i.e., $\boldsymbol{X} = [X_1, \ldots, X_d]$, where each X_i if the projection of $\boldsymbol{X}$ onto the i-th coordinate space.

- The joint CDF of the random vector $\boldsymbol{X}$ is defined analogously from as

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \mathbb{P}[\omega : X_i(\omega) \leq x_i,\ i = 1, \ldots, d] = \mathbb{P}[X_1 \leq x_1 \cap X_2 \leq x_2 \cap \cdots \cap X_d \leq x_d]\,,$$

where $\boldsymbol{x} = [x_1, x_2, \ldots, x_d] \in \mathbb{R}^d$. The associated joint PDF can be obtained as

$$\pi_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{\partial^d F_{\boldsymbol{X}}(x_1, x_2, \ldots, x_d)}{\partial x_1 \partial x_2 \cdots \partial x_d}. \tag{27}$$

Marginal and conditional distributions

- The marginal probability distributions, give the probabilities for any of the individual variables without reference to the values of the other ones. The conditional probability distributions, which give the probabilities for any grouping of the random variables, conditional on particular values of the remaining ones.

- The conditional PDF between two random variables, say X_j given the occurrence of a value x_i of X_i, is given by

$$\pi_{X_j|X_i}(x_j \mid x_i) = \frac{\pi_{X_iX_j}(x_i, x_j)}{\pi_{X_i}(x_i)} \qquad \text{for any } i, j = 1, \ldots, d,$$

  where each $\pi_{X_i}(x_i)$ represents the marginal PDF of the random variable X_i, which is computed by integration (marginalization) of the joint $\pi_{X_iX_j}(x_i, x_j)$ with respect to X_j.

Marginal and conditional distributions

The conditional PDF $\pi_{X_j|X_i}(x_j \mid x_i)$ can be interpreted as a normalized profile function. For a fixed $x_i = c$, the function $\pi_{X_iX_j}(c, x_j)$ is a *profile* of the joint PDF, since it equals the intersection of the surface $\pi_{X_iX_j}(x_i, x_j)$ by the plane $x_i = c$.

Marginal and conditional distributions

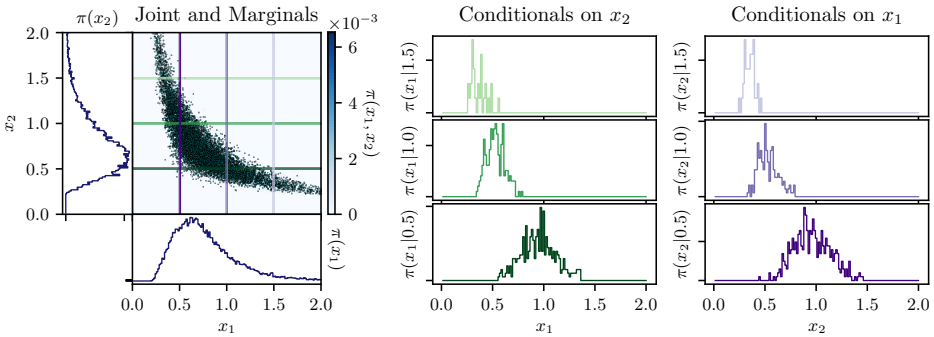


Figure: Lines along which the profiles are taken are shown. The marginals are depicted by histograms at each side of the 2D plot. Three conditionals of x_1 given $x_2 = [0.5, 1.0, 1.5]$ are shown in green. The associated conditionals for x_2 given $x_1 = [0.5, 1.0, 1.5]$ are shown on purple.

The multivariate Gaussian distribution

- A $d \times 1$ random vector $\boldsymbol{X}$ is said to have a Gaussian distribution, if $\boldsymbol{a}^{\mathsf{T}}\boldsymbol{X}$ is a Gaussian random variable for any $\boldsymbol{a} \in \mathbb{R}^d$.

- Multivariate Gaussian distributions are parameterized by their mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The multivariate Gaussian PDF is given by

$$\pi_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{28}$$

- Not every multivariate Gaussian has a density (when $\boldsymbol{\Sigma}$ is singular). These type of distributions still have application in different problems and they are oftentimes called *intrinsic Gaussian distributions*.

7) Probability recap: final aspects

Transformations: linear

- Let $\boldsymbol{x}^{\mathsf{T}}$ a column vector in $\mathbb{R}^d$, and $\mathbf{A} \in \mathbb{R}^{d \times m}$ a matrix. The map $\boldsymbol{x} \rightarrow \boldsymbol{y}$, with $\boldsymbol{y} = \mathbf{A}\boldsymbol{x}$ is called a *linear transformation*. Now, consider a random vector $\boldsymbol{X} = [X_1, \ldots, X_d]^{\mathsf{T}}$, and let $\boldsymbol{Y} = \mathbf{A}\boldsymbol{X}$, then $\boldsymbol{Y}$ is a random vector in $\mathbb{R}^m$. If we know the distribution of $\boldsymbol{X}$, we can derive that of $\boldsymbol{Y}$.

- If $\boldsymbol{X}$ has a mean vector $\boldsymbol{\mu}_X$ and covariance matrix $\boldsymbol{\Sigma}_{XX}$, those of the random vector $\boldsymbol{Y} = \mathbf{A}\boldsymbol{X}$ are given by:

$$\boldsymbol{\mu}_Y = \mathbf{A}\boldsymbol{\mu}_X \qquad \boldsymbol{\Sigma}_{YY} = \mathbf{A}\boldsymbol{\Sigma}_{XX}\mathbf{A}^{\mathsf{T}}. \tag{29}$$

- More generally, if $\mathbf{A}$ is invertible. We have that:

$$\pi_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{|\det(\mathbf{A}^{-1})|}\pi_{\boldsymbol{X}}(\mathbf{A}^{-1}\boldsymbol{y}). \tag{30}$$

Transformations: general

- In many cases, we encounter problems given by functions of random variables. For instance, consider the random vector $\boldsymbol{Y} = f(\boldsymbol{X})$. Suppose than f is invertible and, both the function and its inverse, are differentiable (i.e., f is a diffeomorphism).

- Then, the density function of $\boldsymbol{Y}$ can be obtained as

$$\pi_{\boldsymbol{Y}}(\boldsymbol{y}) = \pi_{\boldsymbol{X}}(f^{-1}(\boldsymbol{y}))\left|\mathbf{J}_{\boldsymbol{y}}(f^{-1})\right|, \tag{31}$$

  where $\left|\mathbf{J}_{\boldsymbol{y}}(f^{-1})\right|$ is the absolute value of the determinant of the Jacobian matrix $\mathbf{J}_{\boldsymbol{y}}(f^{-1})$ at $\boldsymbol{y}$ of the transformation f^{-1}. Here, $\mathbf{J}_{\boldsymbol{y}}(f^{-1})$ has elements $J_{i,j} = \partial f_i^{-1}/\partial y_j$.

- Popular transformations involve a process called *standardization* (also called *whitening* in case of a Gaussian), where the distribution of the random variable is transformed to an analogous distribution but with zero mean and unit variance.

Some useful inequalities

- If X is a non-negative random variable, then for all $x > 0$, we have the Markov's inequality

$$\mathbb{P}[X \geq x] \leq \frac{\mu_X}{x}. \tag{32}$$

(knowing that $\mathbb{E}[\mathbb{1}_A] = \mathbb{P}[A]$ and using the trick $1 = \mathbb{1}_{X \geq x} + \mathbb{1}_{X < x}$).

- If we also know the variance of X, we can give a tighter bound. Namely, for any X, we have the Chebyshev's inequality

$$\mathbb{P}[|X - \mu_X| \geq x] \leq \frac{\sigma_X^2}{x^2}. \tag{33}$$

- We use these two inequalities to show the laws of large numbers, which area pivotal concepts in the Monte Carlo method.

Good references on rigorous probability theory...

Highly recommended:

Joel A. Tropp (2023) - Probability Theory & Computational Mathematics. https://tropp.caltech.edu/notes/Tro23-Probability-Theory-LN.pdf

J. S. Rosenthal (2006) - A first look at rigorous probability theory. 2nd ed. World Scientific Publishing Company.

References

[1] R. B. Ash et al. *Probability and measure theory*. 2nd ed. Harcourt/Academic Press, 2000.

[2] M. Betancourt. *Probability Theory (For Scientists and Engineers)*. `https://betanalpha.github.io/assets/case_studies/probability_theory.html`. 2021.

[3] A. N. Kolmogorov. *Foundations of the theory of probability*. Chelsea Publishing Company, 1956.

[4] J. S. Rosenthal. *A first look at rigorous probability theory*. 2nd ed. World Scientific Publishing Company, 2006.