Preface

This set of notes is an extended, although not entirely complete version of my lectures on undergraduate probability at the Faculty of Mathematics at The Yerevan State University. The entire course comprises around 30 lectures in total and is split into two parts, with part one developing the main concepts primarily in discrete setting and part two presuming knowledge of Lebesgue integral and measure theory. This being said the intersection of concepts in the two parts is brought to a minimum and one is not merely an extension of another.

There are already a number of excellent resources for undergraduate probability courses, and I benefited enormously from many of them, in particular lecture notes of R. Weber, monographs of G. Grimmett & D. Welsh, D. Williams, R. Durrett has been some of the primary sources I was consulting when preparing this course. This notes are by no means a near to complete account for an introductory material to probability but are somewhat customized for the specific course I teach. The notes incorporate various examples where techniques and ideas of probability can be applied in often distant looking scenarios with a hope to give a student a glimpse into breadth and beauty of the subject. All the topics which we discuss are accompanied by exercises that are meant to reinforce the material and with problems (colored in magenta) that might require a bit more effort, but solving them carries a promise to give one a feeling of finding a hidden spring in mountains or encountering an exotic bird in a forest.

I owe my gratitude to students who took the class and to their questions which helped me in shaping and refining the course material and purifying the arguments I had. In particular I wish to thank Hayk Aprikyan, Anna Gyanjumyan, Gagik Melkumyan, Gor Melkumyan, Tigran Chtchyan who kindly pointed out to several inaccuracies, and also to Hayk Tarkhanyan for organizing video recordings for some of the lectures and for his interest. As for the cover image, I wanted to have something that will tell its own story, with no words and titles, and hint toward the direction that there is a beautiful and exciting journey ahead. My thanks are due to Amalia Tucci who accomplished this quest and created the artwork[1] which I now use as a cover to these notes.

In all of its appearance the notes, including this very preface, are under constant development and are being changed, rewritten, extended, contracted, as we go. Should you find a mistake or will have any other comment, please feel free to drop me a line using the contacts at the end of the notes.

Hayk Aleksanyan, Yerevan 2024

[1]Here is a version of interpretation of the symbolism on the cover image. The tree represents knowledge and is associated with various mathematical and algorithmic concepts, including decision trees, Monte Carlo trees, random forests, tree as a data structure, etc., and is, after all, a strong and beautiful tree. The coins falling from the tree symbolize the basic experiment of a coin toss, commonly encountered in probability. The walking man is performing a 2D random walk, a central concept in probability, and part of his route is depicted on the ground. Similarly, the bird's flight represents a 3D random walk, with the point cloud illustrating the locations visited along its path. The floating tape, covered with a sequence of H's (Heads) and T's (Tails), represents a series of independent experiments. Throwing dice, along with coin tossing, is another model of simple random experiments we frequently encounter. Finally, the creeks are a tribute to the law of large numbers where random paths initially divergent, eventually converge toward a common value.

PROBABILITY LECTURE NOTES

HAYK ALEKSANYAN

Contents

Date: September 15, 2024.

1. Events and probabilities

This section introduces the notion of probability and discusses how one may formalize uncertainty in an experiment. We also formulate basic concepts around probability, such as sample space, event space and probability space.

1.1. **Notion of a probability.** The outcomes of many actions cannot be predicted with certainty in advance, simple examples of which being rolling a dice or tossing a coin. Probability theory studies such actions and their consequences.

The mathematical theory of probability began in seventeenth-century France when two renowned French mathematicians B. Pascal and P. Fermat developed an interest in understanding games involving chance. The game they considered is the following:

> Two players A and B play a series of games where the winner of each game gets a point. The player who reaches 10 points first wins the series. When A had 8 points and B had 7 they were forced to stop the game. How should they divide the stake if both A and B are equally skilled?

Today, probability theory is a well-established branch of mathematics that has applications in other areas of mathematics and beyond such as *machine learning*, *mathematical finance*, *data science*, *weather prediction*, *medical trials*, to name a few.

At the heart of the mathematical theory of probability is the idea of an *experiment*. It encapsulates the course of action with an uncertain outcome. The mathematical object modeling such experiment is called a *probability space*, which in broad terms comprises three *building blocks*:

1. the set of all possible outcomes of the experiment,
2. the list of all events which may possibly occur as a consequence of the experiment,
3. an assessment of the likelihoods of these events.

For instance, consider tossing of a (fair) coin as our experiment. Then, the points above become

1. the set of all possible outcomes is the set $\{H, T\}$ for Heads and Tails,
2. the possible events are *no toss*, *toss of H*, *toss of T*, or a *toss of H or T*,
3. both H and T are equally likely to occur.

Later, we will study such examples in greater details and mathematical rigor.

1.2. **Outcomes and events.** We now formalize the above discussion about experiments and their outcomes. Let $\mathcal{E}$ be some experiment whose outcome is not deterministic: for the sake of illustration consider rolling of a six-sided dice. Then the set of all possible outcomes of the experiment, which is usually denoted by Ω and referred to as the **sample space** is

$$\Omega = \{1, 2, 3, 4, 5, 6\},$$

in correspondence with the numbers on the facets of the dice. Questions one may ask about experiment $\mathcal{E}$ can be as follows: "is the outcome equal to 3?", "is it smaller than 5?", "is it a prime number ?", etc. Such questions can be reformulated in terms of the subsets of Ω. For example, the last question is equivalent to asking if the outcome of the experiment lies in $\{2, 3, 5\}$ which is a subset of Ω. Next, we collect all subsets of Ω that are of interest to us into a list. Each member of such list, which we will refer to as an **event**, corresponds to a possible outcome of the experiment. There is a natural link between events and subsets of Ω. Indeed, take two subsets A and B of Ω, then

- $A \cup B$ corresponds to the event "either A or B occurred",
- $A \cap B$ corresponds to the event "both A and B occurred",
- $\Omega \setminus A$ corresponds to the event "A did not occur".

We thus see that set-theoretic concepts applied to subsets of Ω can be put in correspondence with the outcomes of the experiment $\mathcal{E}$.

Given an experiment $\mathcal{E}$, let Ω be the set of all possible outcomes of the experiment and let $\mathcal{F} := \{A_i : i \in I\}$ be the collection of subsets of Ω that are of interest to us (the *events*) in relation to $\mathcal{E}$. In the light of the above correspondence between *set-theoretic operations* and the *outcomes of the experiment*, we need to require some consistency from the set $\mathcal{F}$ which we formalize next.

Definition 1.1. `(Event space)` *The collection $\mathcal{F}$ of a sample space Ω is called an* ***event space*** *if*

$$\mathcal{F} \text{ is not empty}, \tag{1.1a}$$

$$\text{if } A \in \mathcal{F} \text{ then } \Omega \setminus A \in \mathcal{F}, \tag{1.1b}$$

$$\text{if } A_1, A_2, \ldots \in \mathcal{F} \text{ then } \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}. \tag{1.1c}$$

The axioms above simply state that $\mathcal{F}$ is a non-empty set of subsets that is closed under taking complements and countable unions. In measure-theoretic context, the *event space* as defined above is **identical** to the concept of a ***σ-algebra***: an event space is simply a σ-algebra of subsets of Ω. There are several straightforward but important implications from Definition 1.1.

Property 1.1. *Let Ω be a sample space and $\mathcal{F}$ be an event space on Ω. Then*

1. *$\mathcal{F}$ contains $\emptyset$ and Ω,*
2. *$\mathcal{F}$ is closed under operations of finite union,*
3. *$\mathcal{F}$ is closed under finite or countable intersections.*

Proof. For the first item, observe that by (1.1a) there exists $A \subset \Omega$ that lies in $\mathcal{F}$. Then in view of (1.1b) we get $A^c = \Omega \setminus A \in \mathcal{F}$. Now set $A_1 = A$, $A_2 = \Omega \setminus A$, and $A_i = A$ for all $i = 3, 4, \ldots$. Thanks to (1.1c) we have

$$\Omega = A \cup (\Omega \setminus A) = \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}.$$

Since $\Omega \in \mathcal{F}$ we get $\emptyset = \Omega \setminus \Omega \in \mathcal{F}$ due to (1.1b), proving item 1.

To prove that an event space is closed under operations of finite union, take any collection $A_1, \ldots, A_n \in \mathcal{F}$ and define $A_i = \emptyset$ for $i \geq n+1$. We showed above that the empty set lies in $\mathcal{F}$ hence (1.1c) concludes the claim on finite unions.

Finally, to see that an event space is closed under finite or countable intersections we use de Morgan's law which states that for any collection $A_i \in \mathcal{F}$, $i = 1, 2, \ldots$ one has $\Omega \setminus \left(\bigcap_{i=1}^{\infty} A_i \right) = \bigcup_{i=1}^{\infty} (\Omega \setminus A_i)$. The right-hand side of this equality is in $\mathcal{F}$ due to (1.1b) and (1.1c), therefore the left-hand side is also in $\mathcal{F}$ thanks to (1.1b). □

Let us consider a few examples of Ω and an event space $\mathcal{F}$.

Example 1.1.1. *Ω is any non-empty set and $\mathcal{F} = 2^{\Omega}$ - the power set*[1].

[1]For any set Ω the notation 2^{Ω} stands for the set of all subsets of Ω, i.e. $2^{\Omega} = \{A : A \subset \Omega\}$. This set is called the **power set** of Ω.

Example 1.1.2. *Ω is any non-empty set and $\mathcal{F} = \{\emptyset,\ \Omega\}$.*

The two examples above represent the edge-cases: for the given Ω the first is the largest event space and the second is the smallest, both in terms of inclusions.

Remark 1.2. *Informally, we may think about $\mathcal{F}$ as the set of all outcomes of the experiment that we can verify if they have happened or not in a particular realization of the experiment. In other words $\mathcal{F}$ is the set of all questions one may ask about the outcome of the experiment, that can be addressed.*

Example 1.2.1. *Let our experiment be rolling of a fair 6-sided dice. Assume that the outcome of the roll is not directly revealed to us, but one can only know if the roll was even or odd (e.g. we want to build a simplistic random-bit generator). Then $\Omega = \{1, 2, 3, 4, 5, 6\}$ and*

$$\mathcal{F} = \{\emptyset, \{1, 3, 5\}, \{2, 4, 6\}, \{1, 2, 3, 4, 5, 6\}\}.$$

Example 1.2.2. *Ω is any non-empty set, $A \subset \Omega$ and $\mathcal{F} = \{\emptyset, \Omega, A, \Omega \setminus A\}$. If A is a non-empty proper subset of Ω (i.e. $A \neq \Omega$), then $\mathcal{F}$ is the smallest event space containing A.*

Below Ω is a set and $\mathcal{F}$ is an event space as formulated in Definition 1.1.

☐ Exercise 1.2.1. *For any $A, B \in \mathcal{F}$ show that $A \cap B \in \mathcal{F}$ and $A \setminus B \in \mathcal{F}$.*

☐ Exercise 1.2.2. *Given $A_1, ..., A_n \in \mathcal{F}$ and let B be the set of all points in Ω that belong to exactly k of the A_i-s, where $1 \leq k \leq n$ is fixed. Show that $B \in \mathcal{F}$.*

☐ Exercise 1.2.3. *Show that for finite Ω the number of elements in $\mathcal{F}$ is even.*

☐ Exercise 1.2.4. *Construct an example of a sample space Ω and a collection of its subsets $\mathcal{G}$ such that $\mathcal{G}$ is non-empty, is closed under taking complements and finite unions with any number of elements, but is not closed under taking countable unions.*

Note: we showed that being closed under taking finite unions is a consequence of the set of axioms in Definition 1.1. This exercise emphasizes the fact that we cannot recover the case of countable unions from finite unions.

☐ Exercise 1.2.5. *Let Ω be a non-empty set and assume it is partitioned into a union of non-intersecting sets as follows*

$$\Omega = \bigcup_{i \in \mathcal{I}} A_i,$$

*where $\mathcal{I}$ is some index set, $A_i \subset \Omega$ and $A_i \cap A_j = \emptyset$ if $i \neq j$. We will call the sets A_i **atoms**. Consider all subsets of Ω that are a union of some atoms, namely define*

$$\mathcal{A} := \left\{ \bigcup_{i \in I} A_i :\ I \subset \mathcal{I} \right\},$$

where the set of indicies I can be empty in which case the union above is the empty set. Prove that

(a) *$\mathcal{A}$ is an event space (σ-algebra).*

(b) *event spaces described in Example 1.1.1 and Example 1.1.2 are both special cases of the above construction (what are the atoms in each case?).*

→ **Problem 1.2.1.** *Let Ω be an uncountable set. Consider the following subsets of Ω,*

$$\mathcal{A} := \{A \subset \Omega : \text{ if either } A \text{ or } \Omega \setminus A \text{ is at most countable}\}.$$

Prove that

(a) *$\mathcal{A}$ is an event space (σ-algebra),*

(b) *$\mathcal{A}$ cannot be represented in the form of atoms as in Exercise 1.2.5.*

→ **Problem 1.2.2.** *Consider the set of positive integers $\mathbb{N}$. A subset $A \subset \mathbb{N}$ is said to have Cesaro density $\gamma(A)$ if the limit*

$$\lim_{n\to\infty} \frac{|A \cap \{1, 2, ..., n\}|}{n}$$

exists and equals $\gamma(A)$. Show that the set of subsets of $\mathbb{N}$ that have Cesaro density is not closed under intersection.

1.3. **Probability measure.** Getting back to the building blocks of a probability space discussed in subsection 1.1, notice that we have not yet allocated any *quantitative measure* to an event. In this section for each event $A \in \mathcal{F}$ we assign a probability and will write $\mathbb{P}(A)$ for its value. The axiomatic approach that comes next is due to **N. Kolmogorov**.

Definition 1.2. `(Probability measure)` *A mapping $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ is called a **probability measure** on $(\Omega, \mathcal{F})$ if*

1. *$0 \leq \mathbb{P}(A) \leq 1$ for any $A \in \mathcal{F}$,*
2. *$\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$,*
3. *for any collection $A_1, A_2, ... \in \mathcal{F}$ of disjoint events, i.e. $A_i \cap A_j = \emptyset$ for $i \neq j$, we have*

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

The last condition is referred to as **countable additivity** (σ-additivity for short) of the measure. Notice that the probability measure $\mathbb{P}$ is defined only on those subsets[2] of Ω that are in $\mathcal{F}$. Two remarks are in order:

(a) The condition that $\mathbb{P}(\emptyset) = 0$ is redundant and can be inferred from the rest of the axioms. Indeed, take $A_1 = \Omega$ and $A_i = \emptyset$ for $i \geq 2$. In view of condition 3 of Definition 1.2 we have

$$\mathbb{P}(\Omega) = \mathbb{P}(\Omega) + \sum_{i=2}^{\infty} \mathbb{P}(\emptyset),$$

hence $\mathbb{P}(\emptyset) = 0$.

(b) $\mathbb{P}$ is *finitely additive*, i.e. for a collection of disjoint events A_i, $i = 1, 2, ..., n$ we have

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i).$$

This follows from countable additivity taking $A_i = \emptyset$ for all $i = n + 1, ...$.

[2]The structural assumption on $\mathcal{F}$ is instrumental in defining $\mathbb{P}$. A result due to S. Banach and K. Kuratowski (1929) states that under *continuum hypothesis* there is no non-vanishing σ-additive finite measure μ defined for all subsets of $[0, 1]$ such that $\mu(\{x\}) = 0$ for all $x \in [0, 1]$.

(c) The condition that $\mathbb{P}(A) \leq 1$ for any $A \in \mathcal{F}$ is also redundant and follows from the **monotonicity** of the probability measure. Indeed, take any $A, B \in \mathcal{F}$ where $A \subset B$. Then $B = A \cup (B \setminus A)$ and since the two sets in the union are disjoint the finite additivity proved above implies

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A), \tag{1.2}$$

where the inequality is due to non-negativity of $\mathbb{P}$. Since for any $A \in \mathcal{F}$ we have $A \subset \Omega$, using monotonicity of $\mathbb{P}$ we get $\mathbb{P}(A) \leq \mathbb{P}(\Omega) = 1$.

Example 1.2.3. *Let Ω be a non-empty set and $A \subset \Omega$ be a non-empty proper subset. Then any probability measure $\mathbb{P}$ defined on $\{\emptyset, \Omega, A, \Omega \setminus A\}$ has the form*

$$\mathbb{P}(\emptyset) = 0, \ \mathbb{P}(\Omega) = 1, \ \mathbb{P}(A) = p \text{ and } \ \mathbb{P}(\Omega \setminus A) = 1 - p,$$

for some $0 \leq p \leq 1$.

Example 1.2.4. `(Equiprobable sample space)` *This is one of the most recurring scenarios, when the sample space Ω has finite number of elements and all its elements have equal probability of occurrence. Namely, let $\Omega = \{\omega_1, ..., \omega_n\}$ and $\mathcal{F} = 2^\Omega$. For each A let $|A|$ be the cardinality of A (number of elements in A) and define*

$$\mathbb{P}(A) := \frac{|A|}{|\Omega|}. \tag{1.3}$$

Then $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$.

□ Exercise 1.2.6. *Let Ω be at most countable set (finite or countable) and let $p_i \geq 0$ be so that $\sum_i p_i = 1$. Let $\mathcal{F}$ be an event space defined on Ω and for $A \in \mathcal{F}$ set*

$$\mathbb{P}(A) := \sum_{i:\ \omega_i \in A} p_i.$$

Show that $\mathbb{P}$ is a probability measure.

→ Problem 1.2.3. *Construct an example of sample space Ω, an event space $\mathcal{F}$ on Ω and a mapping $\mathbb{P} : \mathcal{F} \to [0, 1]$ such that $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$, $\mathbb{P}$ is finitely additive, i.e. the last requirement of Definition 1.2 is true for finite unions with any number of elements, but $\mathbb{P}$ is not σ-additive.*

Note: we showed that finite additivity follows from σ-additivity as we had in Definition 1.2. This problem shows that we cannot recover the case of countable additivity from finite additivity.

1.4. **Probability space.** We combine the definitions of the event space and probability measure into the following.

Definition 1.3. *A **probability space** is a triplet $(\Omega, \mathcal{F}, \mathbb{P})$ such that*

(a) *Ω is a non-empty set,*

(b) *$\mathcal{F}$ is an event space of subsets of Ω,*

(c) *$\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$.*

A few important properties of a probability space are stated below.

Property 1.3. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Then*

(1) *$\mathbb{P}(A) + \mathbb{P}(\Omega \setminus A) = 1$ for any $A \in \mathcal{F}$,*

(2) *for any $A, B \in \mathcal{F}$ with $A \subset B$ one has $\mathbb{P}(A) \leq \mathbb{P}(B)$,*

(3) *for any $A, B \in \mathcal{F}$ we have $\mathbb{P}(A \cup B) + \mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B)$.*

Proof. For (1) observe that $\Omega = A \cup (\Omega \setminus A)$. The sets in the union are non-intersecting events, hence, in view of finite additivity of $\mathbb{P}$ and item 2 of Definition 1.2 we get (1).

The proof of (2) - the monotonicity of $\mathbb{P}$, is given in (1.2).

For part (3) notice that

$$A = (A \cap B) \cup (A \setminus B),$$
$$B = (A \cap B) \cup (B \setminus A),$$

and

$$A \cup B = (A \setminus B) \cup (B \setminus A) \cup (A \cap B).$$

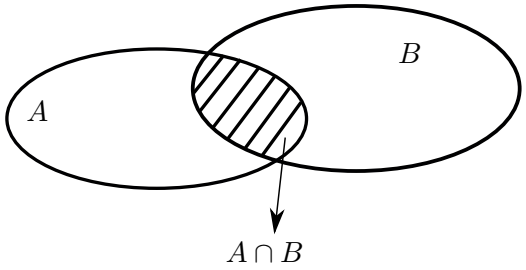

FIGURE 1. In the union $A \cup B$ the subset $A \cap B$ contributes both in $\mathbb{P}(A)$ and $\mathbb{P}(B)$ thus we need to subtract $\mathbb{P}(A \cap B)$ to compensate this doubled contribution.

In all three equalities the left-hand side is represented as a union of disjoint events. Thus, employing the finite additivity of a probability measure we get

$$\mathbb{P}(A) + \mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(A \setminus B) +$$
$$\mathbb{P}(A \cap B) + \mathbb{P}(B \setminus A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cup B),$$

completing the proof. □

Rearranging the equality in (3) of Property 1.3 we obtain

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B), \tag{1.4}$$

which is the **inclusion-exclusion** principle for two sets. Note that when $\mathbb{P}$ is an equiprobable measure as we have in (1.3), then inclusion-exclusion formula takes a form of a combinatorial counting principle. We will return to this equality later on and establish its generalization for the case of n events.

□ EXERCISE 1.3.1. *Use the inclusion-exclusion principle to count the number of integers from* 1 *to* 2021 *that are divisible by* 2 *or* 7.

□ EXERCISE 1.3.2. *In a garden of* 100 *plants,* 80 *have been regularly watered, and* 60 *have received proper sunlight. Use inclusion-exclusion to estimate the number of plants that have received both regular watering and proper sunlight.*

□ EXERCISE 1.3.3. *Consider coin tossing experiment where we toss a coin five times. Describe the probability space for this experiment where*

1. *we are interested in all outcomes of the five tosses* (2^{2^5} *events*),
2. *we are only interested in the total number of heads that occur in five tosses* (2^5 *events*).

2. Combinatorial principles and their relation to probability

In this section we discuss various counting principles with an eye on their probabilistic interpretation.

2.1. **Multiplication principle.** Consider this simple question: *in how many different ways we can form a* 3*-question exam ticket out of* 5 *probability,* 7 *analysis and* 4 *geometry questions, if all* 3 *topics must be presented on the list?* Consider this as a joint experiment consisting of three separate experiments, where the first, second, and third experiments involve selecting a probability question, an analysis question, and a geometry question, respectively. Since we know the number of outcomes for each experiment the answer to the original question becomes $5 \times 7 \times 4$ in view of the **multiplication principle**, also known as the *fundamental rule of counting*, or *rule of product*, that we now define.

The multiplication principle allows one to compute the number of outcomes of a sequence of k experiments $\mathcal{E}_i$, $i = 1, 2, ..., k$, when there are m_1 possible outcomes for the first experiment, m_2 for the second, and so on until m_k scenarios for the last one. Then, the outcome of the joint experiment is an ordered k-tuple and the total number of such tuples equals (see Figure 2)

$$m_1 \cdot m_2 \cdot ... \cdot m_k.$$

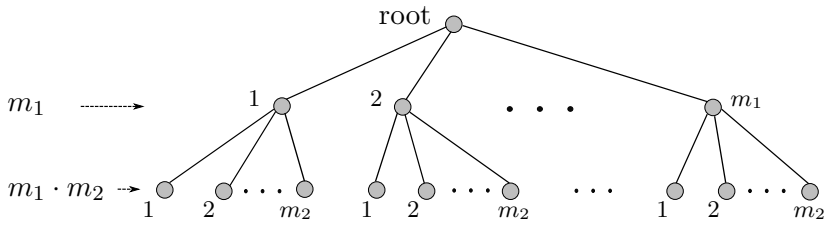

FIGURE 2. It is convenient to think about such k-tuples as leaves of a tree. Namely, the root of the tree is the starting position where no experiment has been made yet, then there are m_1 nodes at the first level representing outcomes for the first step, and then each child on the first level has m_2 children for the second choice, and so on up to the k-th level. The set of all sequences of length k is then the set of all paths starting from the root of the tree and ending at a leaf. The image depicts such tree corresponding to two steps, when there are m_1 outcomes for the first step and m_2 outcomes for the second one. Notice that this tree has $m_1 \cdot m_2$ leaves.

Example 2.0.1. *Consider sequences of length k made of $\{H, T\}$ corresponding to k tosses of a coin. The total number of such sequences is 2^k.*

Example 2.0.2. *The number of words of length k of an alphabet with n letters is n^k. Notice that this is a generalization of the previous examples. Indeed, if we take $\{H, T\}$ as our alphabet, i.e. $n = 2$, and consider words of length k (i.e. sequences made of H or T of length k) we get 2^k such words.*

□ Exercise 2.0.1. *Assume you have* 2 *coats,* 3 *scarves,* 2 *pairs of boots and* 2 *pairs of gloves. How many different outfits you can have if each of the four items in your wardrobe must be used exactly once?*

2.2. **Permutations.** Another fundamental concept we will encounter in combinatorics related to probability theory is the **permutation**.

Definition 2.1. *A permutation of a set S is a bijection from S to itself.*

When the set S is *finite*, any permutation of S is simply an ordered arrangement of its elements. For example, when $S = \{1, 2, 3\}$ the set of all permutations of S is the following:

$$123,\ 132,\ 213,\ 231,\ 312,\ 321$$

In general, we can compute the number of ways integers $1, 2, ..., n$ can be ordered from the multiplication principle. Indeed, there are n choices for the first element, $n-1$ for the second, so on until 1 choice for the last one, hence the total number of permutations becomes

$$n \times (n-1) \times ... \times 2 \times 1 = n!.$$

□ Exercise 2.0.2. *In how many different ways n different hats can be assigned to n different people so that each person gets exactly one hat?*

2.2.1. ***Stirling's formula.*** We will frequently encounter factorials in estimating probabilities. The fundamental tool in dealing with factorials is Stirling's formula, which we recall here:

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n, \text{ as } n \to \infty. \tag{2.1}$$

The notation $\sim$ means that the ratio of the left and right sides converges to 1 as $n \to \infty$.

We will prove the weak version of Stirling's approximation which is still useful and refer to analysis textbooks for the proof of the full version stated above.

Lemma 2.1. `(Toy version of Stirling)` *We have $\log n! \sim n \log n$, as $n \to \infty$.*

Proof. Notice that $\log n! = \sum_{k=1}^{n} \log k$. By monotonicity of $x \mapsto \log x$ we have

$$\log k \leq \int_k^{k+1} \log x\, dx \leq \log(k+1), \text{ with } k = 1, 2, ...,$$

hence $\int_{k-1}^{k} \log x\, dx \leq \log k \leq \int_k^{k+1} \log x\, dx$, where $k = 2, 3, ...$. Summing the inequalities over k and using the fact that $\log 1 = 0$ we get

$$\int_1^n \log x\, dx \leq \log n! \leq \int_1^{n+1} \log x\, dx. \tag{2.2}$$

Using integration by part we get $\int_1^n \log x\, dx = n \log n - n$. Plugging this into (2.2) completes the proof. □

→ **Problem 2.1.1.** *How many trailing 0-s does $n!$ have? For example, there are none for $0 \leq n \leq 4$, there is one for $5 \leq n \leq 9$, there are two for $10 \leq n \leq 14$, etc.*

→ **Problem 2.1.2.** *Prove that the fractional parts*[3] *of* $\log n!$ *are dense in* $[0,1]$. *Namely, show that for any interval* $[a,b] \subset [0,1]$ *there exists* $n \in \mathbb{N}$ *such that* $\{\log n!\} \in [a,b]$.

2.3. **Sampling.** For some experiments, we are given a set $\mathcal{X}$ and need to pick elements from $\mathcal{X}$ according to a specific rule. For instance, consider a deck of cards (the set $\mathcal{X}$), where the experiment consists of choosing three cards from the deck. When choosing a card, we might either put it back into the deck so that it can be chosen again, or set it aside to enforce the uniqueness of the selected items. In this context, we refer to such a process as **sampling** from $\mathcal{X}$, and we will discuss the following sampling schemes from a set:

- ◇ ordered with replacement,
- ◇ ordered without replacement,
- ◇ without order without replacement,
- ◇ without order with replacement.

Here (*without*) *with replacement* refers to the condition of (not) using the same element more than once, and (*without*) *with order* means that the way the elements are arranged (does not) matters. For example $(1,2,3)$ is different from $(2,1,3)$ if order is enforced (i.e. in ordered schemes) otherwise they are the same. The aim of this section is to introduce counting principles that address the four cases listed above. Table 2.3 summarizes the number of possibilities of sampling k times from n objects with the schemes mentioned above. We next proceed to the detailed analysis of these schemes.

	with replacement	without replacement
ordered	n^k	$\frac{n!}{(n-k)!}$
without order	$\frac{(k+n-1)!}{k!(n-1)!}$	$\frac{n!}{k!(n-k)!}$

2.3.1. ***Ordered with replacement.*** This is the easiest case of the four. We have n objects and need to sample k of them regarding the order and allowing replacement. The result of such sampling is a tuple of length k where for each position we have n possible choices. Hence the number of such tuples is n^k in view of the multiplication principle.

Example 2.1.1. *Let us compute the probability of getting at least one* 6 *in* 10 *throws of a fair* 6*-sided dice. Set* $\mathcal{X} = \{1,2,3,4,5,6\}$*, then each outcome of the result of the experiment of throwing the dice* 10 *times can be represented by a tuple of the form* $(a_1,...,a_{10})$ *where* $a_i \in \mathcal{X}$*. Thus the sample space becomes*

$$\Omega = \{(a_1,a_2,...,a_{10}) : a_i \in \mathcal{X} \text{ for } i = 1,2,...,10\},$$

with $\mathcal{F} = 2^\Omega$ *and* $\mathbb{P}$ *defined as the equiprobable measure on* $\mathcal{F}$*, i.e.*

$$\mathbb{P}(E) = \frac{|E|}{|\Omega|}, \text{ for any } E \in \mathcal{F}.$$

[3]For $x \in \mathbb{R}$ let $\lfloor x \rfloor \in \mathbb{Z}$ be the largest integer that $\lfloor x \rfloor \leq x$, i.e. $\lfloor x \rfloor \in \mathbb{Z}$ is the closest integer to x not exceeding it. Then this difference $x - \lfloor x \rfloor$ is called the **fractional part** of x and is denoted by $\{x\}$. By definition we have $\{x\} \in [0,1)$ for any $x \in \mathbb{R}$. For example $\{5.6\} = 0.6$ while $\{8\} = 0$.

Elements of Ω *can equivalently be described as outcomes of* **ordered sampling with replacement** *from* $\mathcal{X}$*, hence* $|\Omega| = 6^{10}$*. We will compute the probability of not getting* 6 *in* 10 *throws, which in our notation is the event*

$$E = \{(a_1, a_2, ..., a_{10}) : a_i \in \mathcal{X} \setminus \{6\} \text{ for } i = 1, 2, ..., 10\}.$$

Similarly to Ω*, the elements of* E *are the results of sampling with replacement and with regard to order from the set* $\mathcal{X} \setminus \{6\}$ *that has* 5 *elements. Hence* $|E| = 5^{10}$*. In particular, the probability of getting at least one* 6 *in* 10 *throws becomes*

$$1 - \mathbb{P}(E) = 1 - \frac{|E|}{|\Omega|} = 1 - \left(\frac{5}{6}\right)^{10} \approx 0.838$$

□ Exercise 2.1.1. *In a similar setup as in Example 2.1.1 compute the probability of getting at least one* 5 *and one* 2 *in* 15 *throws of a fair* 6*-sided dice.*

2.3.2. ***Ordered without replacement.*** So far when discussing permutations we were shuffling n objects and taking all n of them. In certain scenarios we might need to choose only a fixed part of the n elements, say $0 \leq k \leq n$ of them. For example, *consider a code of length* 3 *that is made of digits* 0 *to* 9*. Assume in this code one can use the same digit only once: what is the total number of such codes?*

In this problem we are sampling from the set $D := \{0, 1, ..., 9\}$ and hence $n = 10$. Observe that when $k = n$, then the set of the codes coincides with the set of permutations of D. For $k < n$ we have n items for the first character of the code, $n-1$ for the second, so on until $n-(k-1)$ for the k-th position (last one). Hence, in view of the multiplication principle the total number becomes

$$n \cdot (n-1) \cdot ... \cdot (n-(k-1)) = \frac{n!}{(n-k)!}.$$

This is usually denoted by $P(n, k)$ or A_n^k. The sampling scheme discussed here is also referred to as *permutations of* n *objects taken* k *at a time.*

□ Exercise 2.1.2. *Show that the number of injective mappings from a set* A *with* m *elements into set* B *with* n *elements, where* $m \leq n$*, equals* $\frac{n!}{(n-m)!}$.

2.3.3. ***Birthday problem.*** We now discuss a classical problem on counting, often referred to as a **birthday paradox**, which can be modeled using the scheme of *ordered sampling without replacement.* Here is the statement of the problem.

> What is the **minimum** number of people that should be present in the same room, so that the probability that at least two of them have their birthdays on the same day of the year is at least 1/2?

Since we have 365 days in a year (disregarding the leap year's 1 additional day) it might be tempting to guess that we need at least half of that, 182, as our answer. The correct answer, however, is 23, significantly less than that!

To arrive at the answer, assume there are k people in the room and let $n = 365$ be the number of days in a year. There are n^k arrangements of the birthdays k people can have (ordered sampling with replacement) and we now compute the number of such arrangements that result in all people having birthdays on different days. The latter is clearly a problem about permutations of n taken k at a time as discussed in subsection 2.3.2. Thus, there are A_n^k of them. Hence the probability of all k having birthdays on

different days is

$$\frac{A_n^k}{n^k} = \frac{1}{n^k} n \cdot (n-1) \cdot ... \cdot (n-(k-1)) = \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) = ... = \left(1 - \frac{k-1}{n}\right) =: f(k).$$

The probability that there are at least two people having birthdays on the same day is $1 - f(k)$. Plugging $n = 365$ in the above equality one may estimate that $1 - f(22) \approx 0.47$ and $1 - f(23) \geq 0.5$ thus the answer is $k = 23$.

→ **Problem 2.1.3.** *There are* 360 *different breeds of dogs, each breed is equally likely to be owned by a given person, and each person owns exactly one dog. Every dog owner shares a single image (with permissive license) of their dog on the internet, and you need to collect data for your algorithm that recognizes dog breeds. What is the minimum number of random images you need to pick to ensure that there is at least a* 20% *chance that all breeds are represented in your collection?*

2.3.4. ***Without order without replacement.*** For permutations of n objects taken $0 \leq k \leq n$ at a time, as discussed in the previous section, the order of the elements matter. For example, $(0, 1, 2)$ and $(1, 0, 2)$ are treated as different elements. There are cases when we might need to disregard the order. For example, in how many ways 10 players can form a group of 4 members? In this case the order of the players forming a group does not matter. In general, assume we have n objects and need to choose k of them. Then, in view of the results of Section 2.3.2 we have $\frac{n!}{(n-k)!}$ ways to do this if take into account the order of the elements. Each choice is formed of an ordered tuple of length k which itself can be ordered in $k!$ different ways. Thus, if we disregard the order, then each of the $\frac{n!}{(n-k)!}$ elements will be counted $k!$ times. Hence dividing this number by $k!$ gives

$$\frac{n!}{k!(n-k)!},$$

which is the number of ways k elements can be chosen out of n without order and without replacement. This number is usually denoted by $\binom{n}{k}$, is called a **binomial coefficient** and reads "n choose k". This method of sampling is called *combinations of n objects taken k at a time.*

☐ Exercise 2.1.3. *Show that $\binom{n}{k} = \binom{n}{n-k}$ for any integer $0 \leq k \leq n$.*

☐ Exercise 2.1.4. *Show that for any $1 \leq k \leq n-1$ one has*

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

Here $\binom{n}{0} = \binom{n}{n} = 1$.

☐ Exercise 2.1.5. *Show that the number of subsets $\{1, 2, ..., n\}$ of $0 \leq k \leq n$ elements equals $\binom{n}{k}$. Use this to prove that*

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n.$$

Interpret the result of Exercise 2.1.3 in terms of the subsets of $\{1, 2, ..., n\}$.

→ **Problem 2.1.4.** *Show that for any $n \geq 1$ one has*

$$\sum_{k=0}^{n} \binom{n}{k}^2 = \binom{2n}{n}.$$

Hint: use Exercise 2.1.3 and the first part of Exercise 2.1.5. Compare the two sides of the equation with the number of ways of selecting n-element subsets from a set of 2n elements.

2.3.5. ***Without order with replacement.*** This is the most delicate scenario of counting out of the four cases that we discus here - *combinations of k objects from n with replacement and without order*. The setting of this sampling scheme is as follows: given n distinct objects, in how many ways we can select $k \geq 0$ of them if replacement is allowed (i.e. we can reuse the same element more than once) and order of the elements is disregarded?

Before formulating the main result consider a simple case when $n = 3$ and $k = 2$. Assuming that we are selecting from $\{1, 2, 3\}$, the options we have[4] are

$$11,\ 12,\ 13,\ 22,\ 23,\ 33.$$

Observe that each sample in the list is characterized uniquely by the number of 1-s, 2-s and 3-s it contains (recall that we do not take order of the tuple into account). In the light of this observation we way rewrite the list above as

$$(2,0,0),\ (1,1,0),\ (1,0,1),\ (0,2,0),\ (0,1,1),\ (0,0,2), \tag{2.3}$$

where the i-th index of the 3-tuple indicates the count of i-s used with $i = 1, 2, 3$. We see that for a given 3-tuple in (2.3) the sum of its elements is 2, thus what we do, effectively, is equivalent to representing k as a sum of n non-negative integers. We arrive at the following problem: find the number of integer solutions to $x_1 + x_2 + x_3 = 2$ with $0 \leq x_i \leq k$ for all $i = 1, 2, 3$. This observation settles the general case.

Theorem 2.2. *Given $n \geq 1$ and $k \geq 0$, there are*

$$\binom{n+k-1}{k}$$

ways to select k objects from n with replacement and without order.

Proof. Assume the n objects are enumerated as $1, 2, ..., n$. Then any collection of size k from n with replacement and disregarding the order is uniquely represented by an n-tuple of integers $(x_1, ..., x_n)$ where x_i is the number of times i was used in the collection of size k. This correspondence implies that the count we are interested in equals the number of integer solutions to the following equation

$$x_1 + ... + x_n = k \text{ where } 0 \leq x_i \leq k, \text{ for all } 1 \leq i \leq n. \tag{2.4}$$

Each equation of this form can be put in one-to-one correspondence with a sequence of length $k + n - 1$ composed of symbols $|$ and $+$, where each $x_i > 0$ is replaced by x_i number of $|$ symbols and if $x_i = 0$ it is replaced by an empty symbol. Thus, for example the corresponding sequences for (2.3), with the full sequence of reductions we did, becomes

[4]Notice that there are 3^2 ordered tuples.

11	(2, 0, 0)		$\|\|++$
12	(1, 1, 0)		$\|+\|+$
13	(1, 0, 1)	$x_1+x_2+x_3=2,$	$\|++\|$
22	(0, 2, 0)	$x_i \in \{0,1,2\}$	$+\|\|+$
23	(0, 1, 1)	for all $i=1,2,3.$	$+\|+\|$
33	(0, 0, 2)		$++\|\|$

We see that the number of solutions to (2.4) equals the number of sequences of length $k+n-1$ made of symbols | and + where the number of | symbols equals k and the number of + symbols is $n-1$. In view of section 2.3.4 the number of such sequences equals $\binom{n+k-1}{k}$ which completes the proof. □

□ Exercise 2.2.1. *We have n keys at our disposal and one and only one of them is the right key for the given lock. At each step we randomly pick a key and try to unlock with it. If we fail, we put the key back and select another one at random. What type of sampling is this and what is the number of arrangements when the door gets unlocked for the first time at step k?*

This is sampling with replacement. There will be n^k total ways to reach step k out of which $(n-1)^{k-1}$ are successful.

□ Exercise 2.2.2. *Given integers $n \geq 1$ and $k \geq 1$ compute the number of positive integer solutions to $x_1+...+x_n=k$ where $x_i>0$, $i=1,2,...,n$.*

Hint: reduce to the case of non-negative solutions, which we know how to compute from Theorem 2.2.

2.4. **Binomial and multinomial coefficients.** We have already seen the binomial coefficient $\binom{n}{k}=\frac{n!}{k!(n-k)!}$ in relation to the sampling method discussed above. Here, we will apply combinatorial reasoning to prove the celebrated binomial theorem and then discuss an extension of the binomial coefficient that can be used in sampling schemes involving more than two types of elements, instead of just two.

Theorem 2.3. `(Binomial theorem)` *For any $x,y \in \mathbb{R}$ and any integer $n \geq 0$ we have*

$$(x+y)^n=\sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k}.$$

Proof. There are many proofs of this result, but here we will follow the combinatorial approach. We will prove by induction on n that

$$(x+y)^n=\sum a_1 a_2 \cdot ... \cdot a_n, \tag{2.5}$$

where the sum is over all ordered n-tuples of the form $(a_1,a_2,...,a_n)$ where each a_i is either x or y, i.e. all strings of length n formed of $\{x,y\}$. The case of $n=0,1$ is trivial. The claim is obviously true for $n=2$ since

$$(x+y)^2=xx+xy+yx+yy.$$

Assume that (2.5) if true for $n-1$. To prove it for n notice that each string of length n composed of $\{x,y\}$ is constructed from a string of length $n-1$ adding either x or y at the beginning. Hence writing $(x+y)^n=x(x+y)^{n-1}+y(x+y)^{n-1}$ and applying the inductive hypothesis on $(x+y)^{n-1}$ completes the proof of (2.5). Now each summand of (2.5) can be rearranged into the form $x^k y^{n-k}$ for some $0 \leq k \leq n$. But for a given

$0 \leq k \leq n$ there are exactly $\binom{n}{k}$ summands of the form $x^k y^{n-k}$ since they are formed by choosing k indicies out of $1, 2, ..., n$ for which $a_i = x$ in (2.5) and the rest of a_i-s equal y. Thus, grouping terms of the form $x^k y^{n-k}$ in (2.5) of which we have $\binom{n}{k}$, completes the proof of the theorem. □

We now introduce the **multinomial coefficient**. Consider the following scenario: we have $n \geq 1$ distinct tasks that we need to distribute among k workers, where each worker receives n_i tasks and $n_1 + ... + n_k = n$ with each n_i being a non-negative integer. An equivalent question is: $n \geq 1$ students are forming $k \geq 1$ teams where the i-th team must have exactly n_i members and $n_1 + ... + n_k = n$, how many such teams can be formed? Both questions can be represented as sampling from $\{1, 2, ..., k\}$ with replacement n times where we want to draw each number i exactly n_i times. To compute the number of such possibilities, we will use the multiplication principle. The result of this sampling is an n-tuple of the form $(a_1, ..., a_n)$ where $a_i \in \{1, ..., k\}$. Since the number of 1-s is n_1 out of n total positions there are $\binom{n}{n_1}$ ways to place the 1. When all 1-s are set, we are left with $n - n_1$ free slots out of which n_2 must be 2, hence there are $\binom{n-n_1}{n_2}$ ways to fix the places of 2, so on until $\binom{n-n_1-...-n_{k-1}}{n_k}$ ways to place k. Thus, due to the multiplication principle the total number of outcomes of such sampling becomes

$$\binom{n}{n_1}\binom{n-n_1}{n_2} \cdot ... \cdot \binom{n-n_1-...-n_{k-1}}{n_k}.$$

Using the value of binomial coefficient $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ the last expression simplifies to

$$\frac{n!}{n_1!\, n_2!\, ...\, n_k!} =: \binom{n}{n_1\ n_2\ ...\ n_k}. \tag{2.6}$$

The number in (2.6) is called a **multinomial coefficient**.

Example 2.3.1. *In how many way can we arrange* 5 *red balls,* 3 *blue balls and* 4 *green balls in a line? For this problem* $n = 5 + 3 + 4 = 12$ *and we are asking for*[5]

$$\binom{12}{5\ 3\ 4} = \frac{12!}{5!\ 3!\ 4!} = 27720.$$

A similar argument that we used to prove Theorem 2.3 leads to the following.

Theorem 2.4. *For any* $x_1, x_2, ..., x_k \in \mathbb{R}$ *and any integer* $n \geq 0$ *we have*

$$(x_1 + x_2 + ... + x_k)^n = \sum \binom{n}{n_1\ n_2\ ...\ n_k} x_1^{n_1} x_2^{n_2} \cdot ... \cdot x_k^{n_k},$$

where the sum is over all integers $(n_1, ..., n_k)$ *with* $0 \leq n_i \leq n$ *and* $n_1 + ... + n_k = n$.

□ Exercise 2.4.1. *How many terms does the sum in Theorem 2.4 have?*

□ Exercise 2.4.2. *Adapt the proof of Theorem 2.3 to prove Theorem 2.4.*

[5]An alternative way to think about it is the following. Consider any permutation of $1, ..., 12$, say $(\sigma_1, ..., \sigma_{12})$, then we can use the first 5 elements to position the red balls, i.e. in our notation red balls will be placed at $\sigma_1, ..., \sigma_5$. Similarly, the blue balls will be positioned at $\sigma_6, \sigma_7, \sigma_8$, and the rest will be allocated to green balls. But notice that any permutation of $\sigma_1, ..., \sigma_5$ produces the same positions for the red balls, and the same applies to blue and green balls. Since out of each permutation of 12 elements we can generate 5! permutations that produce the same positions for red balls, 3! for blue and 4! for green, we get $12!/(5!\ 3!\ 4!)$ as the total number of different arrangements.

□ Exercise 2.4.3. *Assume $n \geq 1$ fair 6-sided dices are thrown. What is the probability of getting exactly n_i number of i-s where $i = 1, 2, ..., 6$ and n_i is a non-negative integer with $n_1 + ... + n_6 = n$?*

3. Probabilities of unions of events and counting techniques

In many examples we will deal with unions of events and will be in a position where we need to estimate probabilities of such unions. The first instance we encountered of this case is the third axiom in Definition 1.2 - the countable additivity. Here, we will expand the set of tools we have for handling such cases and establish some basic techniques for working with the probability of unions of events. Specifically, we establish the continuity of the probability measure and prove Boole's inequality, also known as the union bound. We then discuss an application of Boole's inequality in graph theory. The section concludes with the general form of the inclusion-exclusion principle, along with its applications and related inequalities.

3.1. **Continuity of probability measure.**

Theorem 3.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Consider a sequence of monotone events $A_1, A_2, ... \in \mathcal{F}$. If the sequence is increasing, i.e. $A_1 \subset A_2 \subset ...$, then*

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

For decreasing sequence $A_1 \supset A_2 \supset ...$ we have

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

Proof. We consider the case of increasing sequence first. Define a new sequence B_n as follows

$$\begin{aligned} B_1 &:= A_1, \\ B_n &:= A_n \setminus \left(\bigcup_{i=1}^{n-1} A_i\right) = A_n \setminus A_{n-1}. \end{aligned}$$

By definition $\{B_n\}$ is a sequence of disjoint events and

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i \quad \text{and} \quad \bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} B_i \text{ for all } n \geq 1.$$

We thus get

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) \\
&= \sum_{i=1}^{\infty} \mathbb{P}(B_i) \text{ (by } \sigma\text{-addititivity)} \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(B_i) \\
&= \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{i=1}^{n} B_i\right) \text{ (by finite addititivity)} \\
&= \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \\
&= \lim_{n\to\infty} \mathbb{P}(A_n) \text{ (since } A_i \text{ is increasing).}
\end{aligned}$$

To prove the theorem for decreasing events $\{A_i\}$ notice that $\{A_i^c\}$ is increasing, hence from the already proved case we have

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i^c\right) = \lim \mathbb{P}(A_i^c) = 1 - \lim \mathbb{P}(A_i).$$

For the left-hand side, in view of de Morgan's law, we have

$$\bigcup_{i=1}^{\infty} A_i^c = \left(\bigcap_{i=1}^{\infty} A_i\right)^c,$$

which combined with the previous equality settles the case of decreasing sequences. □

□ Exercise 3.1.1. *Construct a counterexample when the continuity of measure does not hold for uncountable number of events, i.e. construct a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*, an uncountable set of indicies* I *with limiting point, say* α*, and a set of events* $A_i \in \mathcal{F}$*, with* $i \in I$ *such that* $A_i \subset A_j$ *if* $i \leq j$ *and* $i, j \in I$ *but*

$$\mathbb{P}\left(\bigcup_{i\in I} A_i\right) \neq \lim_{i\to\alpha} \mathbb{P}(A_i).$$

□ Exercise 3.1.2. *Let* $\mathbb{P}$ *be any probability measure defined on the set of positive integers* $\mathbb{N}$ *with the set of events defined as* $2^{\mathbb{N}}$*.*

(a) *Let* $A_n := \{1, 2, ..., n\}$*. Show that* $\lim_{n\to\infty} \mathbb{P}(A_n) = 1$*.*

(b) *Let* $\sigma : \mathbb{N} \to \mathbb{N}$ *be any permutation of* $\mathbb{N}$ *(i.e. a bijection on* $\mathbb{N}$*), and define* $A_n := \{\sigma(1), ..., \sigma(n)\}$*. Show that* $\lim_{n\to\infty} \mathbb{P}(A_n) = 1$*.*

→ **Problem 3.1.1.** *Prove that the axiom 3 of Definition 1.2, namely the countable additivity of the probability measure, can be replaced by finite additivity and continuity in a sense of Theorem 3.1. More precisely, prove that axioms 1 and 2 of Definition 1.2*

together with the following axiom: for any $A_1, A_2, ... \in \mathcal{F}$ with $A_1 \subset A_2 \subset ...$ one has

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{i\to\infty} \mathbb{P}(A_i),$$

and $\mathbb{P}$ is finitely additive, imply countable additivity as stated in Definition 1.2.

Hint: Take a sequence $\{A_i\}$ of disjoint events and define $B_i := \bigcup_{n=i}^{\infty} A_n$, where $i = 1, 2, ...$ Notice that $\{B_i\}$ is decreasing hence the continuity of the measure applies to it. Then use the finite additivity on the events $A_1, A_2, ..., A_{n-1}, B_n$ to conclude the proof.

3.2. **Boole's inequality.** We now introduce a basic inequality due to Boole, often referred to as a *union bound*, which provides upper bound on probability of the union of events. While the proof is straightforward, nevertheless the inequality is very useful in various scenarios when one needs to estimate that at least one of the several events will happen.

Theorem 3.2. `(Union bound or Boole's inequality)` *For any sequence of events $A_1, A_2, ...$ we have*

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

Proof. Set $B_1 = A_1$ and $B_i = A_i \setminus (\cup_{k=1}^{i-1} A_k)$. Clearly $B_i \subset A_i$ and hence $\mathbb{P}(B_i) \leq \mathbb{P}(A_i)$ for all $i = 1, 2, ...$ in view of the monotonicity of the probability measure. Notice also that $\{B_i\}$ are disjoint, and $\cup_i A_i = \cup_i B_i$, hence

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(B_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

□

Note that for disjoint events we have equality in the union bound above inequality thanks to the σ-additivity of probability measure.

□ Exercise 3.2.1. *Prove that for any events $A_1, ..., A_n$ we have*

$$\mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{n} \mathbb{P}(A_i) - (n-1).$$

Hint: Use de Morgan's law to pass from intersection to union followed by Boole's inequality.

□ Exercise 3.2.2. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $A_1, A_2, ... \in \mathcal{F}$ be so that $\mathbb{P}(A_n) = 1$ for all $n \in \mathbb{N}$. Show that*

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = 1.$$

3.3. **Graph coloring and Ramsey numbers.** In this section we will discuss an application of Boole's inequality in the graph theory. We will first introduce the problem on a well-known puzzle that asks to prove that among any group of 6 people there will be 3 who all know each other or 3 people that none knows another. Now a standard approach to this problem is to model it via a graph, i.e. we represent people as points (verticies) on a plane, 6 of them here, and assume all points are connected via lines between each other (all edges of the graph are present, in this case $(6 \cdot 5)/2 = 15$ in total). Then, if two people know each other we color the line (edge of the graph) connecting them in **red** and if two are unrelated we color the edge in **blue**. Now all edges are colored in red

or blue and the original question asks whether there is a triangle with all vertices being either red or blue (i.e. a monochromatic triangle). We now put this forward in a more general setting, where we will study the existence of certain monochromatic subgraphs in complete graphs.

In the *complete graph*[6] on 6 vertices from the last paragraph given any *red-blue coloring* of its edges[7] we were looking for a subgraph on 3 vertices with all edges being red or a subgraph on 3 vertices with all edges being blue. Now a theorem due to F. Ramsey states that for any positive integers r and s there is an integer N such that any red-blue coloring of edges of a complete graph on N vertices contains a subgraph on r vertices with all edges being blue, or a subgraph on s vertices with all edges being red. The smallest such N is denoted by $R(r,s)$ and is a called a *Ramsey number*.

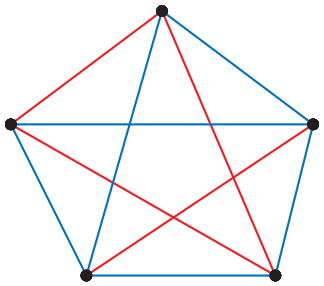

FIGURE 3. A particular red-blue coloring (one of the $2^{\binom{5}{2}} = 1024$ possible variants) of a complete graph on 5 verticies. Notice that we are coloring edges of the graph and not verticies. Observe as well that there is a single monochromatic subgraph with 3 verticies with all edges being red (best seen online).

The exact value of $R(r,s)$ is not known for general r,s except for some special cases. A particularly striking example of applications of probability to combinatorics is a lower bound on $R(s,s)$ due to P. Erdős (1947). The technique that we will use here is called a **probabilistic method** and has far reaching applications beyond the graph coloring problem (see [1] for more).

Proposition 3.3. *If* $\binom{n}{k} \cdot 2^{1-\binom{k}{2}} < 1$ *then* $R(k,k) > n$. *In particular* $R(k,k) > \lfloor 2^{k/2} \rfloor$ *for all* $k \geq 3$.

Proof. Consider a complete graph G on n vertices. By definition it has $\binom{n}{2} = \frac{n(n-1)}{2} =: N$ edges which we will denote by $e_1, ..., e_N$. We now define the experiment of randomly coloring an edge by blue $(+1)$ or red (-1) and describe the probability space corresponding to this experiment. The sample space of all possible colorings of the graph becomes

$$\Omega = \{\vec{c} = (c_1, ..., c_N) : c_i \in \{-1, 1\}\},$$

in particular we have $|\Omega| = 2^N$. We then set $\mathcal{F} = 2^\Omega$ - the power set of Ω, as the event space and use the equiprobable measure 1.2.4 on $\mathcal{F}$, i.e. for any $E \in \mathcal{F}$ we define

$$\mathbb{P}(E) := \frac{|E|}{2^N}, \tag{3.1}$$

In particular for any edge coloring it either red or blue gets equal probability. As discussed in Example (1.2.4) the triplet $(\Omega, \mathcal{F}, \mathbb{P})$ is indeed a probability space.

Take any subgraph on k verticies. Since G is a complete graph, the subgraph has $\binom{k}{2} =: K$ edges. Without loss of generality assume edges of this subgraph are $e_1, ..., e_K$.

[6] A graph is called *complete* if there is an edge between any pair of vertices.

[7] Each edge of a graph is colored either red or blue without any restrictions.

Then, the event that this particular subgraph is monochromatic (i.e. all its edges have the same color) is

$$E := \{(\underbrace{c, ..., c}_{K \text{ times}}, c_{K+1}, ..., c_N) : \ c_i \in \{-1, 1\} \text{ for } K < i \leq N \text{ and } c \in \{-1, 1\}\}.$$

From here and (3.1), we get

$$\mathbb{P}(E) = 2 \cdot 2^{-K} = 2^{1-\binom{k}{2}}.$$

The number of subgraphs on k vertices equals $\binom{n}{k} := M$ and let A_i be the probability that the i-th such subgraph is monochromatic. By Boole's inequality we have

$$\mathbb{P}\left(\bigcup_{i=1}^{M} A_i\right) \leq \sum_{i=1}^{M} \mathbb{P}(A_i) \leq M \cdot 2^{1-\binom{k}{2}} = \binom{n}{k} 2^{1-\binom{k}{2}} < 1.$$

Thus with positive probability none of the events A_i occurs, hence there is a red-blue coloring of G without monochromatic subgraph on k vertices. From this, in view of the definition of the Ramsey number, we get $R(k, k) > n$.

Finally, setting $n = \lfloor 2^{k/2} \rfloor$ we obtain

$$\begin{aligned}\binom{n}{k} 2^{1-\binom{k}{2}} &= \frac{(n-k+1) \cdot ... \cdot n}{k!} \cdot 2^{1-\binom{k}{2}} \leq \frac{n^k}{k!} 2^{1-\binom{k}{2}} \\ &< 2 \frac{2^{\frac{k^2}{2}}}{k!} \frac{2^{\frac{k}{2}}}{2^{\frac{k^2}{2}}} = 2 \frac{2^{\frac{k}{2}}}{k!} < 1 \text{ for } k \geq 3.\end{aligned}$$

□

□ Exercise 3.3.1. *Show that* $2^{\frac{k}{2}+1} < k!$ *for* $k \geq 3$.

□ Exercise 3.3.2. *Prove that among* 6 *people there are* 3 *who all know each other or* 3 *people that no one knows another*[8].

Construct an example of a group of 5 *people such that there is no subgroup of* 3 *people that all know each other or a subgroup of* 3 *people that none knows another. This, coupled with the first claim shows that* $R(3, 3) = 6$.

□ Exercise 3.3.3. *A network of* $n \geq 1$ *wireless devices is distributed randomly in a certain area. If two devices are close to each other, they can connect and exchange messages. Based on prior statistics, you know that two devices will be connected with probability* $0 < p < 1$ *and any configuration of connectivity is equally probable. You pass a message to one of the devices. Prove that the probability that all devices will receive your message is bounded below by* $1 - n(1-p)^{n-1}$.

Hint: Model this as a problem on graphs as we did above and estimate the probability that the graph is connected, or equivalently that there is a vertex that is not connected to any of the rest of vertices using union bound.

→ **Problem 3.3.1.** *Let* $A_1, ..., A_n \subset A$ *be subsets of a non-empty set* A *where* $|A_i| = k$ *for all* $1 \leq i \leq n$. *Assuming* $n < 2^{k-1}$ *show that there is a red-blue coloring of* A *such that each subset* A_i *contains elements of both colors.*

Hint: Model the probability space corresponding to the 2-coloring and use probabilistic method with Boole's inequality.

[8]In terms of the Ramsey numbers we are trying to show that $R(3, 3) \leq 6$. Notice also that Proposition 3.3 cannot be applied here as for $n = 6$ and $k = 3$ the condition $\binom{n}{k} 2^{1-\binom{k}{2}} < 1$ is not satisfied.

→ Problem 3.3.2. *Consider set of binary strings (i.e. sequences of $\{0,1\}$-s) of finite length where no string in the collection is a prefix of another string from the collection (e.g. 00 and 001 cannot both be in the collection as the former is a prefix of the latter). For each integer $i \geq 1$ let N_i be the number of such strings of length i. Show that*

$$\sum_i \frac{N_i}{2^i} \leq 1.$$

Hint: Notice that 2^i is the total number of strings of length i. Then interpret what it means that a shorter string is not a prefix of a longer string.

Note 1: This is called Kraft's inequality for prefix-free codes.

Note 2: There is an interesting realization of the the above inequality on prefix codes generated by binary trees. Consider a binary tree with finite number of nodes. Then prove that

$$\sum_{v \text{ is a leaf}} 2^{-depth(v)} \leq 1,$$

where depth(v) is the length of a path from the root of the tree to the leaf. This claim is related to the main inequality of this problem as follows: each path from root to a leaf can be identified with a string of $\{0,1\}$-s (binary string) where we put 0 if we move to the left child of a node and 1 if we go to the right. Such correspondence between binary strings and the leafs sets the stage for applying the inequality of the problem. Try to fill in the details.

3.4. **Inclusion-exclusion principle and related inequalities.** Here we will study one of the important tools in counting problems, namely the *inclusion-exclusion formula*, which in its basic form was proved in Property 1.3. We will settle the case of $n \geq 2$ events, discuss some of its applications, and then look at related inequalities known as Bonferroni's inequalities.

Proposition 3.4. `(Inclusion-exclusion formula)` *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $A_1, ..., A_n \in \mathcal{F}$. Then*

$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \mathbb{P}(A_i) - \sum_{1\leq i<j\leq n} \mathbb{P}(A_i \cap A_j) + \sum_{1\leq i<j<k\leq n} \mathbb{P}(A_i \cap A_j \cap A_k) + \ ... \\ + (-1)^{n-1}\mathbb{P}(A_1 \cap ... \cap A_n). \tag{3.2}$$

A few remarks are in order. When Ω is a finite set, $\mathcal{F} = 2^\Omega$ and $\mathbb{P}$ is the equiprobable measure (see (1.2.4)), then the inclusion-exclusion principle becomes a counting formula that computes the number of elements in the union of A_i-s. Also note that the terms in the right-hand side of (3.2) are the missing components that turn Boole's inequality into an equality.

A more concise way of writing (3.2), and emphasizing the fact that we are summing over all possible intersections of the events A_i, is the following

$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_{\substack{\mathcal{I}\subset\{1,2,...,n\}\\ \mathcal{I}\neq\emptyset}} (-1)^{|\mathcal{I}|-1}\mathbb{P}\left(\bigcap_{j\in\mathcal{I}} A_j\right). \tag{3.3}$$

Proof of the inclusion-exclusion formula. The proof is by induction on n. The formula for $n = 2$ was proved in (1.4). Next, we represent the union of $n > 2$ elements as the union of the last one with the rest of $n - 1$ sets, and apply the case of two sets

on this representation. We thus get

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \mathbb{P}\left(A_n \cup \left(\bigcup_{i=1}^{n-1} A_i\right)\right) \text{ (applying the case of n=2)}$$

$$= \mathbb{P}(A_n) + \mathbb{P}\left(\bigcup_{i=1}^{n-1} A_i\right) - \mathbb{P}\left(\bigcup_{i=1}^{n-1}(A_i \cap A_n)\right) \text{ (using induction)}$$

$$= \mathbb{P}(A_n) + \sum_{\substack{\mathcal{I}\subset\{1,2,\dots,n-1\}\\ \mathcal{I}\neq\emptyset}} (-1)^{|\mathcal{I}|-1}\mathbb{P}\left(\bigcap_{j\in\mathcal{I}} A_j\right) + \sum_{\substack{\mathcal{I}\subset\{1,2,\dots,n-1\}\\ \mathcal{I}\neq\emptyset\\ \tilde{\mathcal{I}}=\mathcal{I}\cup\{n\}}} (-1)^{|\tilde{\mathcal{I}}|-1}\mathbb{P}\left(\bigcap_{j\in\tilde{\mathcal{I}}} A_j\right).$$

The first sum in the last equality above is the part of (3.2) that does not have A_n in it, and the second one together with $\mathbb{P}(A_n)$ covers all terms with A_n. This concludes the inductive step and completes the proof. □

□ Exercise 3.4.1. *The list of codes contains all strings of length $n \geq 1$ that consist of symbols $\{1, 2, ..., n\}$, e.g. 322, 232, 123, 222 are all valid codes when $n = 3$. If you pick a random code from this list, what is the probability that it will have all numbers $1, 2, ..., n$ in it?*

Hint: compute the total number of such codes via the multiplication principle. Then for each $1 \leq i \leq n$ set A_i to be the codes without i in it. Now represent the probability in question with A_i-s and use inclusion-exclusion.

□ Exercise 3.4.2. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $A_1, ..., A_n \in \mathcal{F}$ be so that $\mathbb{P}(A_i \cap A_j) = 0$ if $i \neq j$ and $\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = 1$. Prove that for any $B \in \mathcal{F}$ we have $\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B \cap A_i)$.*

3.4.1. ***Combinatorial version of the inclusion-exclusion principle.*** A particular corollary of the inclusion-inclusion formula is the *combinatorial version* of it when the underlying probability space is finite and the probability measure is the equiprobable measure. More precisely, take $\Omega = \{\omega_1, ..., \omega_n\}$ where $n \in \mathbb{N}$, $\mathcal{F} = 2^{\Omega}$ and $\mathbb{P}(A) := \frac{|A|}{|\Omega|}$ for $A \in \mathcal{F}$. Then (3.3) for this probability space becomes

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{i=1}^{n} |A_i| - \sum_{1\leq i<j\leq n} |A_i \cap A_j| + \sum_{1\leq i<j<k\leq n} |A_i \cap A_j \cap A_k| + \dots + (-1)^{n-1}|A_1 \cap \dots \cap A_n|, \tag{3.4}$$

which provides a very useful counting principle.

An alternative proof of (3.4) based purely on counting principles is the following. Take any element $\omega \in A_1 \cup \dots \cup A_n$. Then there is an integer $1 \leq k \leq n$ such that ω belongs to exactly k of the sets $\{A_i\}$. Since ω contributes by 1 in the count of the left-hand side (l.h.s.) of (3.4), to complete the proof we need to show that the contribution of ω in the right-hand side (r.h.s.) is also 1. To this end observe that the input of ω in r.h.s. equals

$$k - \binom{k}{2} + \binom{k}{3} + \dots + (-1)^{k-1}\binom{k}{k} \tag{3.5}$$

since there are precisely $\binom{k}{i}$ different non-zero summands in the i-th expression of the the r.h.s. of (3.4) where $1 \leq i \leq k$ and for $i > k$ the intersections with more than i sets do not contain ω in view of our assumption that ω is contained in exactly k sets from the collection $\{A_1, ..., A_n\}$. To complete the proof of (3.4) it is left to verify that (3.5) equals 1. For this notice that the Binomial theorem 2.3 implies

$$0 = (1-1)^k = \sum_{i=0}^{k} \binom{k}{i} (-1)^i = 1 - \sum_{i=1}^{k} \binom{k}{i} (-1)^{i-1}.$$

The last sum in the above expression is precisely the sum we have in (3.5). This settles the combinatorial proof of (3.4).

3.4.2. ***Number of derangements***. As an application of the inclusion-exclusion principle consider the following problem.

> Two identical decks of cards are placed side by side and the first deck is shuffled randomly. One by one we take a card from each of the decks until the decks become empty. What is the probability that at some step we will pick the same card?

Let us model this problem mathematically. Denote the number of cards in a deck by n. Shuffling a deck corresponds to a permutation of $\{1, 2, ..., n\}$. With this in mind define Ω to be the set of all permutations of $\{1, 2, ..., n\}$, set $\mathcal{F} = 2^{\Omega}$ and let $\mathbb{P}$ be the equiprobable measure on Ω. Clearly $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $|\Omega| = n!$. The problem then asks to compute the probability that a random permutation (an element of Ω) will preserve at least one element at its place[9]. To this end define $A_i = \{\sigma \in \Omega : \sigma(i) = i\}$, where $i = 1, 2, ..., n$ (i.e. the permutation σ leaves i unaltered). Then, for any non-empty $\mathcal{I} \subset \{1, 2, ..., n\}$ we have

$$\mathbb{P}\left(\bigcap_{i \in \mathcal{I}} A_i\right) = \frac{(n - |\mathcal{I}|)!}{n!},$$

since we fix $|\mathcal{I}|$ positions and permute the rest $n - |\mathcal{I}|$ arbitrarily. Notice also that there are $\binom{n}{|\mathcal{I}|}$ ways to choose $|\mathcal{I}|$ positions that will be fixed. Hence, by (3.2)

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = 1 - \binom{n}{2}\frac{(n-2)!}{n!} + \binom{n}{3}\frac{(n-3)!}{n!} + \ ... \ + (-1)^{n-1}\frac{1}{n!} = \tag{3.6}$$

$$1 - \frac{1}{2!} + \frac{1}{3!} + \ ... \ + (-1)^{n-1}\frac{1}{n!} \approx 1 - e^{-1},$$

where the last approximation is due to Taylor's expansion of e^{-x}. We conclude that the probability of seeing the same card at some point from the two decks is approximately $1 - e^{-1}$.

□ Exercise 3.4.3. *Assume each of the two decks is shuffled randomly (instead of one being fixed) and then one by one we pick a card from each deck until they are*

[9]A permutation $\sigma : \{1, 2, ..., n\} \to \{1, 2, ..., n\}$ is called a **derangement** if for all $1 \leq i \leq n$ we have $\sigma(i) \neq i$, i.e. the permutation has no fixed points. Notice that we may assume that the first deck of cards corresponds to the identity permutation. Indeed, in any enumeration of all $n!$ permutations if μ and σ are permutations corresponding to the first and second decks then we are interested in the event when $\mu(i) = \sigma(i)$ for some $1 \leq i \leq n$. The latter is equivalent to $i = (\mu^{-1} \circ \sigma)(i)$ for some $1 \leq i \leq n$. Now the mapping $\sigma \mapsto \mu^{-1} \circ \sigma$ is a bijection on Ω and hence will not change the number of derangments. Thus, we may assume that the first deck is the identity permutation, and the second one is chosen randomly.

empty. Model the probability space corresponding to this experiment and show that the probability of getting the same card on each deck at some point is the same as in (3.6).

→ Problem 3.4.1. *Let $A = \{1, 2, ..., k\}$ and $B = \{1, 2, ..., n\}$ where $k, n \in \mathbb{N}$. Use the inclusion-exclusion principle to compute the number of surjective (onto) functions from A to B.*

Hint: notice that the number of all functions from A to B is n^k (sampling with replacement and regarding the order). It is easier to compute the number of functions which are not surjective. To that end for each $i \in B$ define A_i as the set of all functions $f : A \to B$ which do not contain i in their image, i.e. $f(x) \neq i$ for any $x \in A$. Then use inclusion-exclusion principle with A_i-s.

→ Problem 3.4.2. *For $n \in \mathbb{N}$ let $\varphi(n)$ be the Euler's totient function, i.e. $\varphi(n)$ is the number of integers $1 \leq k \leq n$ that are coprime with n. For example, $\varphi(4) = 2$, $\varphi(5) = 4$, $\varphi(11) = 10$. Using the inclusion-exclusion principle prove that*

$$\varphi(n) = n \prod_{\substack{p \text{ is prime} \\ p \mid n}} \left(1 - \frac{1}{p}\right),$$

i.e. the product is over all primes that divide n.

Hint: consider the prime factorization of n, let $n = p_1^{\alpha_1} p_2^{\alpha_2} \cdot ... \cdot p_k^{\alpha_k}$, where p_i is prime and $\alpha_i \in \mathbb{N}$. Then an integer $1 \leq m < n$ is coprime with n if and only if its prime factorization does not contain any of p_i-s. Now define $A_i = \{1 \leq m \leq n : p_i \mid m\}$ where $i = 1, 2, ..., k$ and use the inclusion-exclusion principle.

→ Problem 3.4.3. *Given any map with $n \geq 1$ countries prove that there is a polynomial χ with integer coefficients such that for any $k \geq 0$ the number of ways to color the countries (each country is colored with a single color) so that no neighboring countries are of the same color on the map equals $\chi(k)$.*

Hint: model the problem as coloring of verticies of a graph so that no two adjacent verticies get the same color. Then for an edge e consider the set of all colorings that assign the same color to the endpoints of e. Let $\mathcal{C}_e$ be this set of coloring, then we are interested in the complement of $\cup_e \mathcal{C}_e$. Now use inclusion-exclusion and the combinatorial sampling schemes of the earlier sections.

*Note: the type of coloring mentioned in the problem is called a **proper coloring** and the polynomial discussed above is the **chromatic polynomial** of the graph.*

3.4.3. ***Bonferroni's inequalities***. When we sum the terms appearing in the inclusion-exclusion formula up to a certain intermediate point we lose the exact equality for the probability of a union of events, but interestingly we obtain either an upper or a lower bound for the probability of a union. This we formulate and prove next.

Proposition 3.5. `(Bonferroni's inequalities)` *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $A_1, A_2, ..., A_n \in \mathcal{F}$. For each $1 \leq k \leq n$ define*

$$S_k = \sum_{\substack{\mathcal{I} \subset \{1,2,...,n\} \\ |\mathcal{I}| = k}} \mathbb{P}\left(\bigcap_{i \in \mathcal{I}} A_i\right).$$

Then for $1 \leq k \leq n$ odd we have

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{k} (-1)^{i-1} S_i \tag{3.7}$$

and for $1 \leq k \leq n$ even we have

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{k} (-1)^{i-1} S_i. \tag{3.8}$$

A few clarifying remarks are in order. Notice that S_k is simply the k-th term in the inclusion-exclusion formula disregarding the sign. In particular, the summation in S_k is over all k-element subsets of $\{1, 2, ..., n\}$ and hence there are $\binom{n}{k}$ terms in the sum. Observe also that the case of $k = 1$ in (3.7) corresponds to Boole's inequality for finite number of terms.

Proof of Bonferroni's inequalities. The proof is by induction on n. When $n = 2, 3$ both inequalities in (3.7) and (3.8) follow from inclusion-exclusion formula and non-negativity of a probability measure. We now assume the inequalities hold for $n-1$ and prove the case of n.

Let us first eliminate the edge cases when $k = 1$ or $k = n$. The former is the Boole's inequality, the latter is inclusion-exclusion formula. Thus we will assume that $1 < k < n$ *is odd* and prove (3.7). Using formula (3.2) for two events, namely A_n and $\cup_{i=1}^{n-1} A_i$ we get

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \mathbb{P}(A_n) + \mathbb{P}\left(\bigcup_{i=1}^{n-1} A_i\right) - \mathbb{P}\left(\bigcup_{i=1}^{n-1} (A_i \cap A_n)\right).$$

By the inductive hypothesis we can apply (3.7) on the first union in the above equality and (3.8) with $k-1$ on the second one (note that $k - 1 \geq 1$ since we assumed $k > 1$) arriving at

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &\leq \mathbb{P}(A_n) + \sum_{i=1}^{k} \sum_{\substack{\mathcal{I}\subset\{1,2,\dots,n-1\}\\ |\mathcal{I}|=i}} (-1)^{i-1}\mathbb{P}\left(\bigcap_{j\in\mathcal{I}} A_j\right) \\
&- \sum_{i=1}^{k-1} \sum_{\substack{\mathcal{I}\subset\{1,2,\dots,n-1\}\\ |\mathcal{I}|=i}} (-1)^{i-1}\mathbb{P}\left(A_n \cap \bigcap_{j\in\mathcal{I}} A_j\right) \\
&= \mathbb{P}(A_n) + \sum_{\substack{\mathcal{J}\subset\{1,2,\dots,n-1\}\\ 1\leq|\mathcal{J}|\leq k}} (-1)^{|\mathcal{J}|-1}\mathbb{P}\left(\bigcap_{j\in\mathcal{J}} A_j\right) + \sum_{\substack{\mathcal{J}\subset\{1,2,\dots,n-1\}\\ 1\leq|\mathcal{J}|\leq k-1\\ \tilde{\mathcal{J}}=\mathcal{J}\cup\{n\}}} (-1)^{|\tilde{\mathcal{J}}|-1}\mathbb{P}\left(\bigcap_{j\in\tilde{\mathcal{J}}} A_j\right).
\end{aligned}$$

The first sum in the above inequality covers all terms of (3.7) which do not contain A_n, and the second sum coupled with $\mathbb{P}(A_n)$ cover all terms with A_n. This completes the induction step.

The proof of (3.8) is similar and we leave it as an exercise.

□

4. Conditional probability and Independent events

In this section we introduce the notion of *independence* of events, *conditional probability* and two important results concerning it, namely the *law of total probability* and *Bayes' theorem*.

4.1. **Conditional probability.** Consider an experiment $\mathcal{E}$ and the associated probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Sometimes we might have incomplete information about the outcome of $\mathcal{E}$. For example, assume the experiment is a throw of a fair 6-sided die and we are interested in the event where the number on the die shows 2. Clearly, in the standard equiprobable setting, this event, call it A, has probability $1/6$. Now assume

someone told us that the outcome of the experiment is a prime number. This information does not determine the actual outcome with certainty, nevertheless, it narrows the space of possible values we can see on the die from $\{1,2,3,4,5,6\}$ to $\{2,3,5\}$. This new information affects our calculations of probabilities as from six equiprobable outcomes we had before we now have only three, hence the probability of seeing number 2 becomes $1/3$. In general, for two events A and B, the probability of A, given that B has occurred, might no longer have the value $\mathbb{P}(A)$. A reason for this is that in this case A occurs if and only if $A \cap B$ occurs, which suggests that the adjusted probability of A should be linked with $\mathbb{P}(A \cap B)$. This leads to the following definition.

Definition 4.1. `(Conditional probability)` *Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and two events $A, B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. The **conditional probability** of A given B is denoted by $\mathbb{P}(A \,|\, B)$ and is defined as*

$$\mathbb{P}(A \,|\, B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{4.1}$$

Notice that the proportionality constant $\mathbb{P}(B)$ is chosen so that $\mathbb{P}(B \,|\, B) = 1$, i.e. the probability of B given that B has occurred is 1.

Theorem 4.1. `(Conditional probability induces a new measure)` *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and assume $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. Then the mapping $\mathbb{P}_B : \mathcal{F} \to \mathbb{R}$ defined as $\mathbb{P}_B(A) := \mathbb{P}(A \,|\, B)$ is a probability measure on $(\Omega, \mathcal{F})$, in particular the triple $(\Omega, \mathcal{F}, \mathbb{P}_B)$ is a probability space.*

Proof. To check that $\mathbb{P}_B$ is a probability measure we need to verify that $\mathbb{P}_B$ satisfies all requirements of Definition 1.2.

We clearly have $\mathbb{P}_B(A) \geq 0$ for any $A \in \mathcal{F}$, and the fact that $\mathbb{P}_B(A) \leq 1$ follows from the monotonicity of the probability measure (see (1.2)). The conditions $\mathbb{P}_B(\emptyset) = 0$ and $\mathbb{P}_B(\Omega) = 1$ follow directly from (4.1). The only thing that remains to check is the σ-additivity of $\mathbb{P}_B$. To that end fix a sequence of disjoint events $A_1, A_2, \ldots \in \mathcal{F}$. Then

$$\mathbb{P}_B\left(\bigcup_{i=1}^{\infty} A_i\right) = \frac{1}{\mathbb{P}(B)}\mathbb{P}\left(\bigcup_{i=1}^{\infty}(A_i \cap B)\right) =$$
$$\frac{1}{\mathbb{P}(B)}\sum_{i=1}^{\infty}\mathbb{P}(A_i \cap B) = \sum_{i=1}^{\infty}\mathbb{P}(A_i \,|\, B) = \sum_{i=1}^{\infty}\mathbb{P}_B(A_i),$$

where we used the fact that the events $\{A_i \cap B\}_{i=1}^{\infty}$ are disjoint and that $\mathbb{P}$ is σ-additive. Hence $\mathbb{P}_B$ is a probability measure and the proof is complete. □

4.2. **Law of total probability.** Informally, the conditional probability $\mathbb{P}(A \,|\, B)$ can be interpreted as adjusting the probability of the event A given context of B, where B carries certain piece of information about the random experiment we study. Clearly, if A is the event we are interested in, we can keep it fixed and vary the context B on which conditioning is applied. Now assume we can partition the probability space into disjoint pieces where the behavior of A can be studied more easily. The question that stands then is about possibility to recover the probability of A using pieces of information encoded in each piece of partition. The next result we study answers this question in positive and will be of fundamental importance when studying conditional probabilities.

Definition 4.2. (Partition of a probability space) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We say that the collection $\{B_i : i \in I\} \subset \mathcal{F}$ forms a **partition** of Ω if $B_i \cap B_j = \emptyset$ for $i \neq j$ (the events are disjoint) and $\bigcup_{i\in I} B_i = \Omega$.*

□ Exercise 4.1.1. *Let $\{B_i\}_{i\in I}$ be a partition of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where I is some index set. Assume $\mathbb{P}(B_i) > 0$ for all $i \in I$. Prove that I is at most countable.*

Theorem 4.2. (Partition Theorem or the Law of Total Probability) *Let $\{B_i : i \in I\}$ be a partition of Ω with $\mathbb{P}(B_i) > 0$ for all $i \in I$. Then for any event $A \in \mathcal{F}$ we have*

$$\mathbb{P}(A) = \sum_{i\in I} \mathbb{P}(A \,|\, B_i)\mathbb{P}(B_i).$$

Proof. Using the fact that $\bigcup_i B_i = \Omega$ we get

$$\begin{aligned}
\mathbb{P}(A) &= \mathbb{P}\left(A \cap \left(\bigcup_{i\in I} B_i\right)\right) \\
&= \mathbb{P}\left(\bigcup_{i\in I} (A \cap B_i)\right) \text{ (by } \sigma\text{-additivity of } \mathbb{P}) \\
&= \sum_{i\in I} \mathbb{P}(A \cap B_i) \text{ (by (4.1))} \\
&= \sum_{i\in I} \mathbb{P}(A \,|\, B_i)\mathbb{P}(B_i).
\end{aligned}$$

□

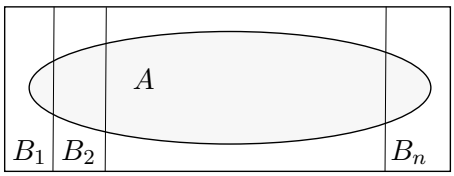

Figure 4. A schematic view of partitioning Ω into n disjoint events $B_1, B_2, ..., B_n$. The law of total probability says that the probability of any $A \in \mathcal{F}$ becomes a weighted sum over its conditional probabilities on B_i weighted by the probabilities of the partition element.

Let us now see this result in action starting with a simple example. Assume a student enters the exam which has a single question from one of these subjects *analysis*, *probability*, *geometry*. The student knows 1/2 of the analysis questions, 2/3 of the probability and 4/5 of the geometry. If all subjects are equally probable to be on the question list and once the subject is selected, all its questions are equally probable, what is the probability that the student will pass the exam?

Let B_1 the list of *analysis*, B_2 - *probability* and B_3 - *geometry* questions. This defines a partition of the probability space. Let also A be the event that the student passes the exam. Since all subjects have equal probability to be on the exam, we have $\mathbb{P}(B_i) = \frac{1}{3}$ for $i = 1, 2, 3$, and from the law of total probability we get

$$\mathbb{P}(A) = \mathbb{P}(A \,|\, B_1)\mathbb{P}(B_1) + \mathbb{P}(A \,|\, B_2)\mathbb{P}(B_2) + \mathbb{P}(A \,|\, B_3)\mathbb{P}(B_3) = \frac{1}{3}\left(\frac{1}{2} + \frac{2}{3} + \frac{4}{5}\right) = \frac{59}{90}.$$

☐ Exercise 4.2.1. *What is the probability space in the above example with exam tickets?*

Here is another example around the partition theorem.

> In a theater with N seats, N people bought tickets. Assume that the first person lost their ticket and chooses a seat uniformly at random from the N seats. The rest of the people then enter one by one. If their assigned seat is empty, they occupy it; otherwise, they choose a seat uniformly at random from the remaining empty seats and sit there. What is the probability that the last person to enter the theater will find their assigned seat still empty?

Let us enumerate the seats by $1, 2, ..., N$ and let seat i be the seat of person i entering the theater hall. Let A_N be the event that the seat of the last person is empty when the first $N-1$ people become seated. Denote $a_N := \mathbb{P}(A_N)$ and for each $1 \leq i \leq N$ define B_i to be the event that the first person takes seat i. In view of the condition of the problem we have $\mathbb{P}(B_i) = \frac{1}{N}$ and clearly $\{B_i\}$ defines a **partition** of the probability space in question. Applying Theorem 4.2 we get

$$\mathbb{P}(A_N) = \sum_{i=1}^{N} \mathbb{P}(A_N \,|\, B_i)\mathbb{P}(B_i) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}(A_N \,|\, B_i). \tag{4.2}$$

Clearly $\mathbb{P}(A_N \,|\, B_1) = 1$ and $\mathbb{P}(A_N \,|\, B_N) = 0$. Now observe that when the first person takes the i-th seat, with $i > 1$, then everyone up to $i-1$ will take their own seat, except possibly for the i-th person, who will choose a seat from the remaining $N - i$ seats uniformly at random. In the light of this observation, if the event B_i occurs, then the i-th person acts as the first person for the same problem but with $N - i$ seats. Hence,

$$a_N = \frac{1}{N} \left(1 + a_2 + ... + a_{N-1}\right).$$

Observe that $a_2 = \frac{1}{2}$. Using induction by N it follows from the last equality that $a_N = \frac{1}{2}$ for all $N = 2, 3, ...$, in particular the probability that the last person's seat will be empty is $1/2$ regardless of the number of seats[10] N.

☐ Exercise 4.2.2. *What is the probability space in the above example with theater seats?*

☐ Exercise 4.2.3. *Write a computer program, in a language of your choice, to simulate the process of seat assignment described above. Ensure that your computations suggest the same outcome as we obtained mathematically.*

☐ Exercise 4.2.4. *In the problem with theater seats assume everyone up to $k-1$-th person takes their seat and it is the k-th person who chooses a seat uniformly at random, where $1 \leq k < N$. What is the probability that the last person will find their seat free?*

[10]Here is an alternative approach to think about this problem, that leverages underlying symmetry in the problem. Notice that people entering the theater have no identity. Hence, when i-th person enters the hall and finds their place occupied, instead of taking a seat uniformly at random from the remaining seats, the i-th person forces the person seated at i-th seat out, takes the i-th seat and then the evicted person takes a seat uniformly at random. Since the identity of the people entering the hall is irrelevant, this step is equivalent to the process described in the formulation of the problem. With this modified rule of taking seats, everyone at steps $2, ..., N-1$ will take their seats. Hence the person mixing their sit will have two choices, either seat on their seat or on the seat of the last person, both with probability $1/2$. This implies that the probability of the last person finding their seat free equals $1/2$.

→ Problem 4.2.1. *Assume in the problem with theater seats two people, the first and the second, mix their seats in the same way as in the original problem. What is the probability that the last person will find their seat unoccupied? How about the first k people mixing their seats?*

Our next example where the law of total probability is used is a well-known problem often referred to as the **secretary problem**.

> The director of a company needs to hire a secretary from a pool of $n \geq 1$ candidates. The candidates are presented for interviews one by one, and after each interview, the director can identify the best candidate among those interviewed so far. The objective is to devise an optimal strategy for the director to choose the best candidate, considering the following conditions:
>
> - once a candidate is rejected, they cannot be reconsidered,
> - once a candidate is selected, the hiring process stops,
> - the total number of candidates n is known at the start and the candidates are arranged in a random order, with each of the $n!$ arrangements equally likely,
> - in any group of candidates that are interviewed the director can determine the best one without ties, specifically the director is able to identify if the current candidate is superior to everyone else interviewed thus far after each interview,
> - if we reached the last candidate in the line we must choose them.
>
> The question is to define a strategy that maximizes the chance of selecting the best candidate under the given assumptions.

Since the order of applicants is equally likely to be any of the $n!$ arrangements, then the probability that the first candidate is the best among all is $1/n$. Thus a trivial strategy of selecting the first candidate will be successful with probability $1/n$. Now assume, we do not pick the first applicant. Then starting from the second we should not select a candidate that is not better that everyone else seen up to that point, simply because that candidate cannot be the best one (recall that we only win if we select the best in the group). This observation naturally leads to the following family of strategies:

> (††) Fix some $1 \leq k \leq n$ and reject all candidates up to $k-1$ inclusive. Then starting from k select the first candidate that is better than everyone else seen in steps up to $k-1$ inclusive. If we reach candidate n, the last one, in this search then n-th applicant must be selected.

Let also E_k be the event that we will choose the best candidate following the strategy (††) outlined above, and let $p_k = \mathbb{P}(E_k)$. For $k = 1$ the strategy means selecting the first applicant, thus $p_1 = 1/n$ (see Exercise 4.2.5). Now assume $k > 1$. For computing p_k we will **partition** the probability space as follows: let A_i be the probability that i-th applicant is the best. Since each alignment is equally likely and there is only one best candidate we get $\mathbb{P}(A_i) = 1/n$ for each $1 \leq i \leq n$ (Exercise 4.2.5). Then

$$\begin{aligned} p_k &= \sum_{i=1}^{n} \mathbb{P}(E_k \,|\, A_i)\mathbb{P}(A_i) \text{ (since the first } k-1 \text{ are rejected)} \\ &= \sum_{i=k}^{n} \mathbb{P}(E_k \,|\, A_i)\frac{1}{n}. \end{aligned}$$

We now proceed to computing $\mathbb{P}(E_k \,|\, A_i)$ where k is fixed and $k \leq i \leq n$ is any. This is the conditional event that we will choose the i-th candidate given that i-th is the best. Since we are following strategy (††) this event can happen if and only if the i-th candidate is the first starting from k-th that is better than everyone else up to $k-1$ inclusive. But this is identical with the event that the best up to $i-1$ is located in $\{1, ..., k-1\}$. The probability of the latter is $(k-1)/(i-1)$ since each arrangement is equally likely (see Exercise 4.2.6). We thus get

$$p_k = \frac{1}{n} \sum_{i=k}^{n} \frac{k-1}{i-1} = \frac{k-1}{n} \sum_{i=k-1}^{n-1} \frac{1}{i}.$$

Now recall that $1 \leq k \leq n$ is a free parameter and we can optimize with respect to k. Considering the inequality $p_k < p_{k+1}$ and doing simple algebraic manipulations indicate that p_k is increasing as long as

$$\sum_{i=k}^{n-1} \frac{1}{k} > 1.$$

Using the asymptotics of the harmonic sum the last expression can be simplified to $\ln n - \ln k > 1$ or equivalently to $k < n/e$ for large n. Thus we will need to reject everyone up to $\approx n/e$ and then pick the first one that is better than what we saw in the first n/e candidates. To argue that strategy (††) with the optimal k is indeed the best we can have, observe that in any strategy where we reject some of the first, say k, candidates there is no point in choosing one that is not better than the best in the first k since that one cannot be the best overall. On the other hand, if we skip the first one after k that is better than everyone else in the first k and choose say the second best then that is equivalent to the initial strategy but with larger k.

□ Exercise 4.2.5. *Let $A = \{a_1, ..., a_n\}$ be any set of distinct numbers. Pick a permutation of n elements uniformly at random from $n!$ permutations and permute A with it. Prove that the probability that the maximum of A appears at index $1 \leq k \leq n$ equals $1/n$.*

□ Exercise 4.2.6. *Let $A = \{a_1, ..., a_n\}$ be any set of numbers where $a_i \neq a_j$ if $i \neq j$. Assume we rearrange A by a permutation sampled uniformly at random from all $n!$ permutations. Prove that for any $1 \leq k \leq i \leq n$ the maximum of the first i elements is in the first k elements with probability k/i.*

4.3. **Independence.** In our study of conditional probabilities, we have seen that new information (context) about the random experiment we are studying can affect the original probability of the event we are interested in. Let us consider another example where additional information is available about the random experiment, but the conditional probability is identical with the original probability of the event.

Take the experiment of tossing two identical coins. Each outcome of this experiment can be described by a pair of values: H (Head) or T (Tails), representing the outcomes of the first and second tosses, respectively. Thus, the sample space becomes $\Omega = \{(H,H), (H,T), (T,H), (T,T)\}$, and we take $\mathcal{F} = 2^{\Omega}$, and as $\mathbb{P}$ we take the equiprobable measure. Now let A be the event that the second toss is H. Clearly $\mathbb{P}(A) = 1/2$ since only two out of four items in Ω satisfy this. Assume we condition A on the event B which is the event that the first toss was H also. We have $\mathbb{P}(B) = 1/2$ and $\mathbb{P}(A \cap B) = 1/4$, hence $\mathbb{P}(A \,|\, B) = 1/2$. Notice that we get $\mathbb{P}(A \,|\, B) = \mathbb{P}(A)$, or $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ after rearrangement. In other words the new information we had from B did not alter the probability of A, in other words the information in A and B are independent in

some sense. This, as we will often encounter throughout this notes, is another typical situation, where the new context does not provide any additional information.

Definition 4.3. (`Independent events`) *Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and two events $A, B \in \mathcal{F}$, call A and B **independent** if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Otherwise A and B are called* dependent*.*

Property 4.3. *If A and B are independent, then so are A^c and B, A and B^c, and A^c and B^c.*

Proof. We start with the first pair, namely A^c and B. Representing B as a union of two disjoint events as

$$B = (A^c \cap B) \cup (A \cap B),$$

and using the finite additivity of $\mathbb{P}$ we get

$$\mathbb{P}(B) = \mathbb{P}(A^c \cap B) + \mathbb{P}(A \cap B) = \mathbb{P}(A^c \cap B) + \mathbb{P}(A)\mathbb{P}(B),$$

where the last equality is due to independence of A and B. Rearranging the last equality leads to

$$\mathbb{P}(A^c \cap B) = \mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(B)(1 - \mathbb{P}(A)) = \mathbb{P}(B)\mathbb{P}(A^c),$$

which is the independence of A^c and B.

Independence of A and B^c follows by symmetry (we can change the places of A and B). The claim for A^c and B^c follows from the independence of A and B^c and the already proved property that we can replace A by its complement. □

□ Exercise 4.3.1. *Prove that if A is independent of itself, then $\mathbb{P}(A) \in \{0, 1\}$.*

Example 4.3.1. *Two fair dice are thrown. Let A_1 be the event that the first one shows an* odd number*, and let A_2 be the event that the second one shows an* even number*.*

Clearly the sample space of this experiment is the following

$$\Omega = \{(i, j) : i, j \in \{1, 2, 3, 4, 5, 6\}\}.$$

Then, $\mathcal{F} = 2^\Omega$ and $\mathbb{P}((i, j)) = 1/36$ for each $1 \le i, j \le 6$. From here we have $\mathbb{P}(A_1) = \mathbb{P}(A_2) = 1/2$ and $\mathbb{P}(A_1 \cap A_2) = \frac{3 \cdot 3}{36} = \frac{1}{4}$. It follows that A_1 and A_2 are independent.

*Let A_3 be the event that the sum of the first and the second throws is 4. Since the only possibilities are $\{(1, 3),\ (2, 2),\ (3, 1)\}$, we have $\mathbb{P}(A_3) = \frac{3}{36} = \frac{1}{12}$. We also have $\mathbb{P}(A_1 \cap A_3) = \frac{2}{36} = \frac{1}{18} \neq \mathbb{P}(A_1)\mathbb{P}(A_3)$ hence A_1 and A_3 are **not** independent.*

Independence is also defined for a family of events.

Definition 4.4. (`Independence of a family of events`) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Then a family of events $\{A_i : i \in I\}$ where $A_i \in \mathcal{F}$ is called **independent** (or **mutually independent**), if for any finite $J \subset I$ we have*

$$\mathbb{P}\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} \mathbb{P}(A_j). \tag{4.3}$$

*The family $\{A_i : i \in I\}$ is called **pairwise independent** if (4.3) holds for any $J \subset I$ with $|J| = 2$.*

Example 4.3.2. (`Pairwise independence does not imply mutual independence`) *It is important to note that the pairwise independence does* ***not*** *necessarily imply (mutual) independence. To see this, consider an experiment of rolling three* 6*-sided dices. For* $1 \leq i \neq j \leq 3$ *let* A_{ij} *be the event that dices* i *and* j *show the same number. The family* $\{A_{12}, A_{13}, A_{23}\}$ *is pairwise independent but is not mutually independent. Indeed,* $\mathbb{P}(A_{ij}) = \frac{1}{6}$ *for all pairs* (i,j)*, and* $\mathbb{P}(A_{12} \cap A_{13}) = \mathbb{P}(A_{12} \cap A_{23}) = \mathbb{P}(A_{13} \cap A_{23}) = \frac{1}{36}$*. It follows that* A_{ij} *are pairwise independent. However,*

$$\mathbb{P}(A_{12} \cap A_{13} \cap A_{23}) = \frac{1}{36} \neq \mathbb{P}(A_{12})\mathbb{P}(A_{13})\mathbb{P}(A_{23})$$

hence the family is ***not*** *mutually independent.*

→ **Problem 4.3.1.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and* $A_1, ..., A_n \in \mathcal{F}$ *be any events, where* $n \geq 2$*. Assume we have*

$$\mathbb{P}(\mathcal{X}_1 \cap \mathcal{X}_2 \cap ... \cap \mathcal{X}_n) = \prod_{i=1}^{n} \mathbb{P}(\mathcal{X}_i),$$

where $\mathcal{X}_i \in \{A_i, A_i^c\}$*, for all* $i = 1, 2, ..., n$*. Prove that the events* $A_1, ..., A_n$ *are mutually independent.*

Hint: observe that

$$\bigcup_{I \subset \{1,...n\}} \left(\left(\bigcap_{i \in I} B_i \right) \cap \left(\bigcap_{i \notin I} B_i^c \right) \right) = \Omega,$$

for any events $B_1, ..., B_n \in \mathcal{F}$*. For example, when* $n = 3$ *we get* $A_1 \cap A_2 = (A_1 \cap A_2 \cap A_3) \cup (A_1 \cap A_2 \cap A_3^c)$*.*

→ **Problem 4.3.2.** *Construct an example of a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *and events* $A_1, ..., A_n$ *with* $n \geq 3$*, for which*

$$\mathbb{P}(A_1 \cap ... \cap A_n) = \prod_{i=1}^{n} \mathbb{P}(A_i) > 0,$$

but no pair A_i, A_j *with* $i \neq j$*, is independent.*

4.3.1. ***Independent experiments***. Informally, independent events model the idea of independent experiments. Let us illustrate this on an example of two probability spaces $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ with $i = 1, 2$, where both Ω_1 and Ω_2 are at most countable[11], and $\mathcal{F}_i = 2^{\Omega_i}$. Set $\Omega := \Omega_1 \times \Omega_2$, $\mathcal{F} := 2^{\Omega}$ and define $\mathbb{P}(\omega_1, \omega_2) := \mathbb{P}_1(\omega_1)\mathbb{P}_2(\omega_2)$ for $(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2$. Then any $A \in \Omega_1$ can be associated with $\tilde{A} := A \times \Omega_2 \subset \Omega$ and any $B \in \Omega_2$ with $\tilde{B} := \Omega_1 \times B \subset \Omega$. We get $\tilde{A} \cap \tilde{B} = A \times B$ and using the definition of $\mathbb{P}$ we obtain

$$\mathbb{P}(\tilde{A} \cap \tilde{B}) = \mathbb{P}(A \times B) = \sum_{(\omega_1, \omega_2) \in A \times B} \mathbb{P}(\omega_1, \omega_2) =$$

$$\sum_{\omega_1 \in A} \sum_{\omega_2 \in B} \mathbb{P}_1(\omega_1)\mathbb{P}_2(\omega_2) = \sum_{\omega_1 \in A} \mathbb{P}_1(\omega_1) \sum_{\omega_2 \in B} \mathbb{P}_2(\omega_2) =$$

$$\mathbb{P}_1(A)\mathbb{P}_2(B) = \mathbb{P}(\tilde{A})\mathbb{P}(\tilde{B}),$$

hence the events $\tilde{A}$ and $\tilde{B}$ are independent.

The argument used above can be easily extended to $n \geq 2$ experiments. Consider $n \geq 2$ experiments $\mathcal{E}_i$ with corresponding probability spaces $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ where $i = 1, 2, ..., n$.

[11]The requirement of Ω_i to be at most countable makes it easier to define a probability measure on their cross product. Later on we will study the concept of **product measure** in detail that will allow us to eliminate the restriction on Ω_i to be discrete.

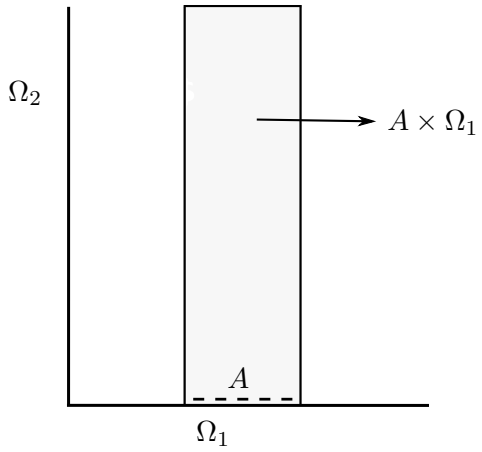

FIGURE 5. A schematic view of the extension of $A \subset \Omega_1$ into $\Omega_1 \times \Omega_2$ as a cylindrical set with base A.

The aim is to build a probability space that will model the *joint experiment* $(\mathcal{E}_1, ..., \mathcal{E}_n)$ as a series of independent experiments. Similar to $n = 2$ define

$$\Omega := \Omega_1 \times ... \times \Omega_n,$$

take $\mathcal{F} := 2^\Omega$ and set

$$\mathbb{P}(\omega_1, ..., \omega_n) := \mathbb{P}_1(\omega_1) \cdot ... \cdot \mathbb{P}_n(\omega_n).$$

It is easy to see that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. The events of the individual experiments $\mathcal{E}_i$ can be associated with events of the larger probability space $(\Omega, \mathcal{F}, \mathbb{P})$ as follows: any $A_i \in \mathcal{F}_i$ is associated with[12]

$$\tilde{A}_i := \Omega_1 \times ... \times \Omega_{i-1} \times A_i \times \Omega_{i+1} \times ... \times \Omega_n \in \mathcal{F},$$

where $i = 1, 2, ..., n$. In view of the definition of $\mathbb{P}$ we have $\mathbb{P}(\tilde{A}_i) = \mathbb{P}_i(A_i)$. The same argument used for the case $n = 2$ shows that (**cylindrical**) events of the form $\tilde{A}_1, ..., \tilde{A}_n$ are independent in a sense of Definition 4.4.

We conclude that finite number of different experiments can be joined together with the construction described above into a single probability space such that the joint experiment is represented as a series of independent experiments.

□ EXERCISE 4.3.2. *Show that the cross product $\mathcal{F}_1 \times \mathcal{F}_2$ of two event spaces is not an event space in general.*

□ EXERCISE 4.3.3. *Fill in the necessary details for the construction for $n \geq 2$ experiments.*

□ EXERCISE 4.3.4. *Get back to Example 4.3.1 and recover the probability space discussed there using the construction with cross products described above. Individual experiments will be two (independent) throws of a dice.*

□ EXERCISE 4.3.5. *Convince yourself that the probability space defined in Proposition 3.3 can be obtained by cross product as above from independent colorings of the edges of the graph. Namely, if we enumerate the edges as $e_1, ..., e_N$ and for each $1 \leq i \leq N$ let*

[12]The event A_i of the single experiment is put in correspondence with the **cylindrical** set $\tilde{A}_i$. Notice that projection $\pi_i : \Omega \to \Omega_i$ defined as $\pi_i(\omega_1, ..., \omega_n) = \omega_i$ maps $\tilde{A}_i$ to A_i.

$\Omega_i = \{-1, 1\}$, $\mathcal{F}_i = 2^{\Omega_i}$ and $\mathbb{P}_i(-1) = \mathbb{P}_i(1) = 1/2$ then the cross product of $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ coincides with the probability space described in the proposition.

→ Problem 4.3.3. *Consider the matrix multiplication operation over the field $\mathbb{Z}_2 = \{0, 1\}$. Namely, let A, B, C all be $n \times n$ matrices with elements from $\mathbb{Z}_2$ and consider the problem of verifying whether $AB = C$. Prove that if $AB \neq C$, then*

$$\mathbb{P}\left(\vec{b} \in \{0,1\}^n : \ (AB)\vec{b} = C\vec{b}\right) \leq \frac{1}{2}.$$

*Here the equality is understood in $\mathbb{Z}_2$, i.e. modulo 2, and the probability measure on $\{0, 1\}^n$ is the equiprobable measure. Once you prove the above inequality, investigate how one can verify with high confidence whether $AB = C$ by doing **independent** tests by choosing a vector $\vec{b} \in \{0, 1\}^n$ uniformly at random and checking $A(B\vec{b}) = C\vec{b}$. Compare the computational complexity of the series of such tests with direct computation of the matrix product and element-wise comparison.*

→ Problem 4.3.4. *For each $n \geq 3$ construct a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a family of events $\mathcal{A} = \{A_1, ..., A_n\} \subset \mathcal{F}$ such that $\mathcal{A}$ is pairwise independent, but any $2 < k \leq n$ events from $\mathcal{A}$ are not mutually independent.*

4.4. **Bayes rule.** The result discussed in this section is of fundamental importance and has far reaching applications beyond probability theory.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $A, B \in \mathcal{F}$ satisfy $\mathbb{P}(A), \mathbb{P}(B) > 0$. Then,

$$\mathbb{P}(B \,|\, A) = \frac{\mathbb{P}(A \,|\, B)\mathbb{P}(B)}{\mathbb{P}(A)} \tag{4.4}$$

which is called the **Bayes rule**. To see why (4.4) is true, observe that from the definition of the conditional probability (4.1) we have

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

and hence (4.4) follows. We next prove a slightly more general form of this rule.

Theorem 4.4. `(Bayes' theorem)` *Let $\{B_i : \ i \in I\}$ be a partition of Ω with $\mathbb{P}(B_i) > 0$ for all $i \in I$ and let $A \in \mathcal{F}$ be such that $\mathbb{P}(A) > 0$. Then*

$$\mathbb{P}(B_j \,|\, A) = \frac{\mathbb{P}(A \,|\, B_j)\mathbb{P}(B_j)}{\sum_{i \in I} \mathbb{P}(A \,|\, B_i)\mathbb{P}(B_i)}.$$

Proof. Rearranging the formula of conditional probability (4.1) we get

$$\mathbb{P}(B_j \,|\, A) = \frac{\mathbb{P}(B_j \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(B_j \cap A)}{\mathbb{P}(B_j)} \frac{\mathbb{P}(B_j)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \,|\, B_j)\mathbb{P}(B_j)}{\mathbb{P}(A)}.$$

Now the claim follows from the law of total probability (Theorem 4.2) applied to the denominator of the last equation. □

We next discuss some examples when Bayes theorem can be utilized.

Example 4.4.1. `(Hit-and-run accident)` *A cab was involved in a hit-and-run accident at night. There are two types of cabs, **Green** (85% of the total cabs) and **Blue** (15% of the total cabs), that operate in the city. A witness identified the cab as Blue. The court tested the reliability of the witness under the circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time.*

What is the probability that the cab involved in the accident was Blue?

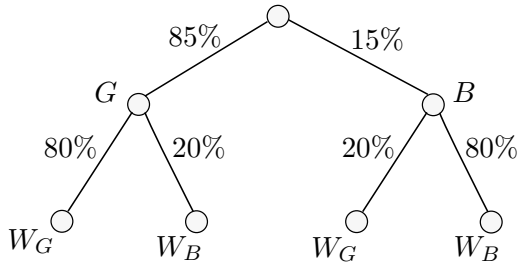

FIGURE 6. It is helpful to describe this problem in schematic way as a tree. The first level represents the partition of the underlying probability space and the last level is the witness' testimony. We then look at the ways of arriving at W_B (the **W**itness claims the car is **B**lue) from the root of the tree. The value of each path from the root to W_B becomes the product of the weights of its edges (see the denominator of (4.5)).

Solution to Example 4.4.1. Let G be the event that the cab involved in the accient, is **G**reen, B be the event that the cab is **B**lue. Let also W_B be the event that the witness testimonies that the cab was blue. We are interested in $\mathbb{P}(B\,|\,W_B)$. Since the events $\{G, B\}$ form a partition of the probability space underlying this question, by Theorem 4.4 we get

$$\mathbb{P}(B\,|\,W_B) = \frac{\mathbb{P}(W_B\,|\,B)\mathbb{P}(B)}{\mathbb{P}(W_B\,|\,B)\mathbb{P}(B) + \mathbb{P}(W_B\,|\,G)\mathbb{P}(G)} = \frac{\frac{80}{100}\frac{15}{100}}{\frac{80}{100}\frac{15}{100} + \frac{20}{100}\frac{85}{100}} \approx 0.413. \tag{4.5}$$

□

The following **terminology** is widely used in Bayesian theory (cf. Theorem 4.4):

- the events of the partition $\{B_i : i \in I\}$ are called the **hypotheses**,
- the set A is called **data**,
- probabilities $\mathbb{P}(B_k)$ are called **prior** probabilities,
- probabilities $\mathbb{P}(B_k\,|\,A)$ are called **posterior** probabilities,
- probabilities $\mathbb{P}(A\,|\,B_k)$ are called **likelihoods**,
- the transition from *priors* to *posteriors* via Bayes theorem is called *Bayesian update*.

Getting back to the example with cabs, notice that without the witness's testimony $\mathbb{P}(B) = \frac{15}{100} = 0.15$ (the *prior* probability). With the data (new information), we can adjust our estimate to $\mathbb{P}(B\,|\,W_B) \approx 0.413$ (the *posterior* probability).

Example 4.4.2. *We have ten coins in our pocket and 9 of them are fair, and one of them is biased with $\mathbb{P}(H) = 0.3$ and $\mathbb{P}(T) = 0.7$ We randomly pick a coin from the pocket, flip it and it lands **H** (head). What is the conditional probability that the selected coin is **biased**?*

Let B be the event that the coin is biased. The prior probability of B is 0.1 since out of 10 coins we have exactly one is biased. Applying Bayes' rule we have

$$\mathbb{P}(B\,|\,H) = \frac{\mathbb{P}(H\,|\,B)\mathbb{P}(B)}{\mathbb{P}(H\,|\,B)\mathbb{P}(B) + \mathbb{P}(H|B^c)\mathbb{P}(B^c)} = \frac{0.3 \cdot 0.1}{0.3 \cdot 0.1 + 0.5 \cdot 0.9} = \frac{3}{48}.$$

Example 4.4.3. *A certain disease occurs in about* 0.05% *of the population. A diagnostic test was developed for it, which correctly detects a person with the disease* 99% *of the time. However, in* 3% *of the cases it will falsely diagnose a healthy patient as having the disease (false positive cases).*

A person is selected at random from the population, and the test indicates that this person has the disease. What are the conditional probabilities that

(a) *the person has the disease?*

(b) *the person does not have the disease?*

Define $A = \{\text{test is positive}\}$ and $B = \{\text{patient has the disease}\}$. We are given

$$\mathbb{P}(B) = 0.0005,\ \mathbb{P}(A\,|\,B) = 0.99 \ \text{ and } \ \mathbb{P}(A\,|\,B^c) = 0.03.$$

We need to compute $\mathbb{P}(B\,|\,A)$ for part (a) and $\mathbb{P}(B^c\,|\,A)$ for part (b). From Bayes formula we have

$$\mathbb{P}(B\,|\,A) = \frac{\mathbb{P}(A\,|\,B)\mathbb{P}(B)}{\mathbb{P}(A\,|\,B)\mathbb{P}(B) + \mathbb{P}(A\,|\,B^c)\mathbb{P}(B^c)} = \frac{0.99 \cdot 0.0005}{0.99 \cdot 0.0005 + 0.03 \cdot 0.9995} \approx 0.016.$$

For part (b) we get $\mathbb{P}(B^c\,|\,A) = 1 - \mathbb{P}(B\,|\,A) \approx 0.984$.

Observe that while the test has high accuracy (99%), being positive on a randomly selected individual only provides 0.016 (1.6%) chance that the person has the disease, which is rather small. This phenomenon is called **base rate fallacy**. It happens when

- there is a large population,
- only a small fraction of the population has the disease,
- there is a large scale testing of mixed population of healthy and sick individuals.

One should take into account that the characteristics of the sampled population are a contributing factor, along with the accuracy of the test itself, in determining the probability of a positive test result. Even for a very accurate test the number of false positives can potentially be large.

□ Exercise 4.4.1. *In Example 4.4.3 let* $0 < p < 1$ *be the fraction of the population that have the disease. Assume that the accuracy of the test is* a% *(i.e. the test gives a positive result on sick individual), and let* b% *be the false positive percentage of the test (i.e. in* b% *of the cases the test gives positive result on healthy individuals).*

A person, randomly selected from the population, gets a positive result on the test. What is the probability that the person actually has the disease? Play with the parameters p, a, b *to see when the results of the test can be reliable.*

4.4.1. ***Naïve Bayes.*** We conclude the section with a toy example of using Bayesian approach for *classification problems*. For the sake of the exposition in this subsection we do away the mathematical rigor and concentrate on describing the general approach in a somewhat hand-waving way. Everything here can be made rigorous and mathematically sound with a slight effort.

A typical setup in classification problem asks to determine the *class label* of the object given a set of *features*. Here is an example.

> Consider the problem of classifying an incoming email as *spam* or *non-spam* (sometimes called *ham*). The data one might use for this problem include the text of the email, subject line, sender's address, etc. All these are features of the email that contribute to it being *spam* or *ham*.

In mathematical terms, assume we have $K \geq 2$ classes denoted by $C_1, C_2, ..., C_K$ and data is represented by a sequence of features $\vec{x} := (x_1, x_2, ..., x_n)$. We then aim to construct a probabilistic model which given the vector $\vec{x}$ will compute the conditional probabilities $\mathbb{P}(C_k \,|\, \vec{x})$, i.e. how likely it is that the object is from class C_k if we received the data $\vec{x}$. In view of the Bayes rule we have

$$\mathbb{P}(C_k \,|\, \vec{x}) = \frac{\mathbb{P}(\vec{x} \,|\, C_k)\mathbb{P}(C_k)}{\mathbb{P}(\vec{x})}. \tag{4.6}$$

Recall that we proved in Theorem 4.1 that conditional probability $\vec{x} \to \mathbb{P}(\vec{x} \,|\, C_k)$ is a probability measure itself. We thus have that $\mathbb{P}_{C_k}(\vec{x})$ for fixed C_k defines a probability measure on the set of features $\vec{x}$. Now the crucial *assumption* in constructing the naïve Bayes classifier is assuming that the features $x_1, ..., x_n$ are conditionally independent[13] given C_k, i.e.

$$\mathbb{P}_{C_k}(\vec{x}) = \prod_{i=1}^{n} \mathbb{P}_{C_k}(x_i).$$

Getting back to (4.6) and using the assumption on conditional independence, we get

$$\mathbb{P}(C_k \,|\, \vec{x}) = \frac{1}{\mathbb{P}(\vec{x})}\mathbb{P}(C_k) \prod_{i=1}^{n} \mathbb{P}_{C_k}(x_i).$$

We now define our classifier: given a data point with features $\vec{x}$ assign it to a class C_k where $1 \leq k \leq K$, so that the probability $\mathbb{P}(C_k \,|\, \vec{x})$ is **maximized**. From the above relation we get[14]

$$\arg\max_{1 \leq k \leq K} \mathbb{P}(C_k \,|\, \vec{x}) = \arg\max_{1 \leq k \leq K} \left(\mathbb{P}(C_k) \prod_{i=1}^{n} \mathbb{P}_{C_k}(x_i) \right).$$

This decision rule is known as **maximum a posteriori** or **MAP**. It picks the hypothesis that is most probable given the data (see the terminology introduced in subsection 4.4). For numerical stability one usually applies log on the right-hand side of the above relation. That will not change the arg max in view of the monotonicity of the logarithm.

→ Problem 4.4.1. *Write a computer program, in a language of your choice, to model a spam, non-spam classifier for emails using Bayesian approach described above. A data to build a model can be found at this address.*

5. Discrete probability distributions

In this section we define *discrete probability distributions*, discuss several important examples, and study the approximation of Poisson distribution by Bernoulli.

5.1. Important examples.

Definition 5.1. *A sequence $\{p_i\}_{i \in I}$, where the index set I is at most countable, is called a **probability distribution** if $0 \leq p_i \leq 1$ for all $i \in I$ and $\sum_{i \in I} p_i = 1$.*

[13] This is an extremely crude assumption, which is not true in most of the practical cases, and this is the reason the classifier is called naïve. Nevertheless, the idea is stil useful and can lead to meaningful results. Getting back to our example with spam classification, we note that there are more efficient ways of classification based on Machine Learning principles.

[14] Given a finite list of numbers $A = (a_1, a_2, ..., a_n)$ by $\arg\max A$ we denote the index $1 \leq k \leq n$ such that $\max A = a_k$. If the maximum in A is not unique then arg max returns any of the indicies of existing maximums.

5.1.1. ***Bernoulli distribution.*** Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ associated with an experiment of tossing a coin. Here $\Omega = \{H, T\}$ and assume $\mathbb{P}(H) = p$ and $\mathbb{P}(T) = 1-p$ where $0 \leq p \leq 1$. The distribution $\{p, 1-p\}$ is called ***Bernoulli distribution*** and is denoted by $B(1, p)$. We will sometimes refer to the probability p as the probability of *success* of the experiment.

For example, in a *clinical trial* a treatment given to a particular patient may succeed or fail. Denoting by $0 \leq p \leq 1$ the probability of a success, we get a Bernoulli distribution. Notice that for each $p \in [0, 1]$ we get a different distribution, but all from the same family of Bernoulli distributions.

5.1.2. ***Bernoulli trials and binomial distribution.*** Suppose that 5% of items produced by a certain machine are defective. Assume also that different items are found to be defective or not *independent* of each other. We are sampling n items at random and inspect them. For a single item we may model the experiment of it being defective by a Bernoulli distribution with probability $5/100 = 0.05$ (i.e. probability of success p is 0.95). When sampling n items, we get a sequence of *Bernoulli trials*, as the experiment of a sampled item being defective follows Bernoulli distribution and the experiments are independent in view of our assumption. This leads to the following definition.

A series of n independent experiments all following Bernoulli distribution $B(1, p)$, is called **Bernoulli trials**. The number of successes in n experiments is an integer from $\{0, 1, ..., n\}$ and

$$p_k := \mathbb{P}(k \text{ successes}) = \binom{n}{k} p^k (1-p)^{n-k}, \text{ where } 0 \leq k \leq n, \tag{5.1}$$

since there are $\binom{n}{k}$ ways to fix k positions in $\{1, 2, ..., n\}$ where the success will happen. The sequence $\{p_k\}_{k=0}^n$ is called **binomial distribution** and is denoted by $B(n, p)$.

A common example of Bernoulli trials is the series of independent experiments with a coin toss (see subsection 4.3.1 for the underlying probability space).

Notice that the sum of the sequence p_k defined in (5.1) equals 1, since there can be only $0, 1, ..., n$ number of successes which provide a partition of the probability space underlying the trials. Another way to see why the sum equals 1 is through Theorem 2.3 (Binomial theorem).

Example 5.0.1. `(Selling extra tickets for a flight)` *An airline sells* 100 *tickets for a certain flight on an airplane that has only* 98 *seats. On the average,* 1% *of passengers do not show up for their flight and assume the passengers choose to show up or not for the flight independently of each other. What is the probability that everyone who appears for the departure will have a seat?*

Solution to Example 5.0.1. We need to understand the distribution of people that will (or will not) show up for the flight. To that end let X be the number of purchasers who do **not** appear for their flight. We will treat the result of a passenger not appearing for the flight a success (for the airplane company). Now, the probability of success is $p = 0.01$ and hence the number of successes in $n = 100$ trials follows a Bernoulli distribution of the form $B(100, 0.01)$ (recall the independence condition in the problem). Thus everyone will have a seat if and only if there are at least 2 successes in $n = 100$ trials. The complement of this event is easier to compute, i.e. the probability of at most 1 success, which equals

$$\mathbb{P}(X = 0) + \mathbb{P}(X = 1) = \binom{100}{0}(1 - 0.01)^{100} + \binom{100}{1}0.01(1 - 0.01)^{99} \approx 0.7357,$$

hence the probability in question becomes $1 - 0.7357 = 0.2643$. □

5.1.3. ***Hypergeometric distribution.*** Consider an urn with n_1 red balls and n_2 black balls. Suppose n balls are drawn **without replacement**, where $n \leq n_1 + n_2$. The probability of drawing exactly k red balls equals

$$p_k := \frac{\binom{n_1}{k}\binom{n_2}{n-k}}{\binom{n_1+n_2}{n}}, \text{ where } \max(0, n-n_2) \leq k \leq \min(n, n_1),$$

because there are $\binom{n_1+n_2}{n}$ ways to choose n balls out of the total n_1+n_2 of which there are $\binom{n_1}{k}$ choices with exactly k red balls and $\binom{n_2}{n-k}$ choices for exactly $n-k$ black balls. The multiplication principle then gives the combined choices of red and black balls.

The sequence $\{p_k\}$ defined above is called *hypergeometric distribution.*

Observe that if we treat drawing of the balls one by one as our experiment, then it is not a Bernoulli trial, since separate experiments are not independent. Indeed, the event of drawing a red ball on the second step depends on the result of drawing on the first step. Hypergeometric distribution is used in scenarios where the total number of possible successes is predetermined.

Summing the probabilities of the hypergeometric distribution we get

$$\sum_{k=0}^{n} \binom{n_1}{k}\binom{n_2}{n-k} = \binom{n_1+n_2}{n}.$$

This is known as *Vandermonde's identity* in combinatorics (cf. Problem 2.1.4).

5.1.4. ***Geometric distribution.*** Consider an infinite sequence of Bernoulli trials with success probability $0 < p \leq 1$. Then the probability of getting the first success on the k-th trial equals $(1-p)^{k-1}p$ where $k = 1, 2, \ldots$. The sequence $\{p(1-p)^k\}_{k=0}^{\infty}$ is called *geometric distribution* with parameter p.

If we consider the coin toss as the single trial, and treat the probability of getting Heads as the success, then the sample space can be modeled as

$$\Omega = \{H,\ TH, ..., T^kH, ...\} \cup \{T^\infty\},$$

where T^kH is the sequence of k Tails followed by a Head, and T^∞ is the sequence of getting no heads.

5.2. **Bernoulli approximation to Poisson.** In this section we introduce Poisson distribution and prove that it is a limit of the binomial distribution in a certain sense. We start with an example.

> In a certain store customers arrive at a constant rate of 4.5 customers per hour on average. The customers' arrivals at different time periods are assumed to be independent. Under this assumption, the store owner wants to find out the number X of actual customers that will arrive during a particular hour of a day.

One approach to this problem is to build an approximate model for the arrivals, taking into account the independence condition, as Bernoulli trials. To build the model, we split the 1 hour period into 3600 seconds and distribute the arrival rate over seconds, getting on average $4.5/3600 = 0.00125$ customers per second. Next, for each second we assume that the arrival of a customer follows a Bernoulli distribution $B(1, p)$ with probability of success being $p = 0.00125$. Thus the number of customers in a given hour follows a binomial distribution of the form $B(3600, 0.00125)$ in view of the independence assumption on arrivals of the customers.

While this model provides a meaningful estimate on the number of customers in a given hour, it imposes a bound of 3600 on the maximum number of customers. To get rid of that bound we split the hour in an even smaller pieces and apply the same model with Bernoulli trials. Namely, fix an integer $n \geq 1$ and split an hour into n equal segments. Let also $\lambda = 4.5$ and $p = \lambda/n$. Then in view of the same analysis we have that the number of customers in an hour follows binomial distribution with parameters $B(n, p)$. The next result provides a way to approximate the values of this distribution.

Theorem 5.1. *Consider a Bernoulli distribution $B(n, p_n)$ and assume $np_n \to \lambda$ as $n \to \infty$ where $\lambda > 0$. Then, for each integer $k \geq 0$ we have*

$$\binom{n}{k} p_n^k (1 - p_n)^{n-k} \to \frac{\lambda^k}{k!} e^{-\lambda}.$$

Proof. By the definition of the binomial coefficient we have

$$\binom{n}{k} p_n^k (1 - p_n)^{n-k} = \frac{n(n-1) \cdot \ldots \cdot (n-k+1)}{k!} p_n^k (1 - p_n)^{n-k} = \frac{1}{k!} \frac{n(n-1) \cdot \ldots \cdot (n-k+1)}{n^k} (np_n)^k \left(1 - \frac{np_n}{n}\right)^{n-k}.$$

The second term in the product above tends to 1 as $n \to \infty$, since it has a fixed number of terms and each term tends to 1. The third one tends to λ^k in view of the condition $np_n \to \lambda$. For the last one recall that

$$\left(1 - \frac{\lambda + o(1)}{n}\right)^n \to e^{-\lambda}, \text{ as } n \to \infty,$$

which follows from the Taylor expansion of the logarithm. Putting these together completes the proof of the theorem. □

The sequence $\{\frac{\lambda^k}{k!} e^{-\lambda}\}_{k=0}^{\infty}$ is called the **Poisson distribution** with parameter λ. This distribution expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known *constant mean rate* and *independently* of the time since the last event. As in the example of arrival of the customers many experiments consist in observing the occurrence times of the random arrivals. More examples include arrivals of calls at a switch-board, occurrence of floods or other natural and man-made disasters, etc. For such cases the family of Poisson distributions is used to model the number of arrivals occurring in a fixed time period. The distribution also serves as an approximation to binomial with very small success probabilities.

6. Discrete random variable

In this section we introduce the concept of *discrete random variable* and its *probability mass function*. We then define *mathematical expectation* of discrete random variables and discuss several basic properties of the expectation.

6.1. **Discrete random variable and its probability mass function.** Often we deal with situations involving some function X defined on a sample space Ω of a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For example, consider the experiment of throwing a fair dice and assume we gamble on the outcome E of this dice in such a way to win 1 if the

outcome is a prime number, lose 1 in case of 4 or 6 and stay even if 1 shows up. Then our profit is a function X defined on Ω where

$$X(\omega) = \begin{cases} 1, & \text{if } \omega \in \{2,3,5\}, \\ -1, & \text{if } \omega \in \{4,6\}, \\ 0, & \text{if } \omega = 1. \end{cases}$$

The mapping $X : \Omega \to \mathbb{R}$ is an example of a *discrete random variable.*

Definition 6.1. (`Discrete random variable`) *Given a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*, a mapping* $X : \Omega \to \mathbb{R}$ *is called a discrete random variable if*

(a) *the image of* X*, i.e. the set*

$$\mathrm{Im}X = \{x \in \mathbb{R} : \ \exists \omega \in \Omega \ s.t. \ X(\omega) = x\},$$

is at most a countable subset of $\mathbb{R}$*,*

(b) $\{\omega \in \Omega : \ X(\omega) = x\} \in \mathcal{F}$ *for any* $x \in \mathbb{R}$*.*

The term *discrete* here refers to the fact that X takes at most countably many values. For the second condition, notice that $\{\omega \in \Omega : \ X(\omega) = x\}$ is the preimage of $\{x\}$ under the mapping X, i.e. $X^{-1}(x)$. We may not be able to predict the actual values of X with certainty, but we would like to measure the probability of X taking a certain value x. To be able to do that we require $X^{-1}(x)$ to be an event.

For a random variable X we will write $\{X = x\}$ for the set $\{\omega \in \Omega : \ X(\omega) = x\}$.

□ Exercise 6.0.1. *Let* X *be a discrete random variable defined on* $(\Omega, \mathcal{F}, \mathbb{P})$*. Prove that for any* $A \subset \mathbb{R}$ *we have* $\{\omega \in \Omega : \ X(\omega) \in A\} \in \mathcal{F}$*.*

Definition 6.2. (`Probability mass function`) *The probability mass function (or pmf) of a discrete random variable* X *defined on* $(\Omega, \mathcal{F}, \mathbb{P})$ *is the function* $p_X : \mathbb{R} \to [0,1]$ *defined as*

$$p_X(x) = \mathbb{P}(X = x), \ x \in \mathbb{R}.$$

Here $p_X(x)$ is simply the probability that X takes the value x, and is sometimes referred to as **discrete density function** too. We have

$$p_X(x) = 0, \ x \notin \mathrm{Im}X,$$

and

$$\begin{aligned} \sum_{x \in \mathrm{Im}X} p_X(x) &= \sum_{x \in \mathrm{Im}X} \mathbb{P}(X = x) \text{ (by } \sigma\text{-additivity of } \mathbb{P}) \\ &= \mathbb{P}\left(\bigcup_{x \in \mathrm{Im}X} \{\omega \in \Omega : \ X(\omega) = x\} \right) = \mathbb{P}(\Omega) = 1. \end{aligned}$$

The last equation is sometimes written as

$$\sum_{x \in \mathbb{R}} p_X(x) = 1,$$

as at most there will be countably many non-zero values in the sum above due to the assumption that X is discrete. The last equation also characterizes discrete probability mass functions as stated in the next theorem.

Theorem 6.1. (Existence of random discrete variables with given pmf) *Let $S = \{s_i : \ i \in I\}$ be a countable set of distinct real numbers and let $\{\pi_i : \ i \in I\}$ be a collection of real numbers satisfying*

$$\pi_i \geq 0 \text{ and } \sum_{i \in I} \pi_i = 1.$$

Then, there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a discrete random variable X defined on it such that the probability mass function of X is given by

$$p_X(s) = \begin{cases} \pi_i, & \text{if } s = s_i \text{ for some } i \in I, \\ 0, & \text{if } s \notin S. \end{cases}$$

Proof. Take $\Omega = S$, $\mathcal{F} = 2^\Omega$ and

$$\mathbb{P}(A) = \sum_{i: \ s_i \in A} \pi_i \quad \text{for } A \in \mathcal{F}.$$

Then $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. Now define $X : \Omega \to \mathbb{R}$ by $X(\omega) = \omega$ for $\omega \in \Omega$. Then the probability mass function (pmf) of X is p_X. □

The power of this theorem is that it allows us to forget about the probability space in many practical scenarios where we have a discrete probability distribution and need to consider a random variable that takes certain values following the given distribution. The theorem ensures that such a random variable exists.

We next show that compositions of discrete random variables with any real-valued functions are still discrete random variables.

Lemma 6.2. (Functions of random variables) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X : \Omega \to \mathbb{R}$ be a discrete random variable, let also $g : \mathbb{R} \to \mathbb{R}$ be any function. Then $Y := g \circ X = g(X)$ is also a discrete random variable on $(\Omega, \mathcal{F}, \mathbb{P})$.*

Proof. First we show that $\mathrm{Im}Y$ is at most countable. Take any $y \in \mathrm{Im}Y$, then there exists $x \in \mathrm{Im}X$ such that $y = g(x)$. It follows that

$$\mathrm{Im}Y \subset \{g(x) : \ x \in \mathrm{Im}X\}.$$

Since $\mathrm{Im}X$ is discrete we get that so is $\mathrm{Im}Y$. Next, we show that for any $y \in \mathrm{Im}Y$ the set $\{\omega \in \Omega : g(X(\omega)) = y\}$ is an event, which will complete the proof of the lemma. Fix any $y \in \mathrm{Im}Y$, then

$$\begin{aligned} \{g(X) = y\} \ & \text{(consider the preimage } g^{-1}(y) = \{x \in \mathbb{R} : g(x) = y\}) \\ &= \{X \in g^{-1}(y)\} \ \ (X \text{ only takes values from } \mathrm{Im}X) \\ &= \{X \in g^{-1}(y) \cap \mathrm{Im}X\} \\ &= \bigcup_{z \in g^{-1}(y) \cap \mathrm{Im}X} \{X = z\} \in \mathcal{F}, \end{aligned}$$

since $z \in g^{-1}(y) \cap \mathrm{Im}X$ is at most countable and $\{X = z\} \in \mathcal{F}$ for any $z \in \mathrm{Im}X$ thanks to the fact that X is a discrete random variable. Thus the last expression is at most a countable union of events under which $\mathcal{F}$ is closed by definition. The proof of the lemma is complete. □

This lemma allows us to conclude that expressions of the form

$$X^2, \ e^X, \ 2023X + \log(1 + |X|),$$

etc, are all discrete random variables given that so is X, where in the examples we took as g functions x^2, 2^x and $2023x + \log(1 + |x|)$ correspondingly.

6.2. **Expectation and variance.** Consider a long series of Bernoulli trials with throwing of a fair dice. As each of the outcomes $1, 2, ..., 6$ is equally likely to appear, the average of observed numbers will be approximately

$$\frac{1}{6}1 + \frac{1}{6}2 + ... + \frac{1}{6}6 = \frac{7}{2},$$

which is the *mean value* of the observable variable - the value of the dice. This notion of mean value is at the heart of the next definition.

Definition 6.3. (Mathematical expectation) *If X is a random variable, the **expectation** of X is denoted by $\mathbb{E}(X)$ and is defined by*

$$\mathbb{E}(X) = \sum_{x \in \mathrm{Im}X} x\mathbb{P}(X = x), \tag{6.1}$$

whenever this sum converges absolutely, i.e. $\sum_{x \in \mathrm{Im}X} |x\mathbb{P}(X = x)| < \infty$.

If X is a discrete random variable for which

$$\sum_{\substack{x \in \mathrm{Im}X \\ x \leq 0}} x\mathbb{P}(X = x) = -\infty \quad \text{and} \quad \sum_{\substack{x \in \mathrm{Im}X \\ x \geq 0}} x\mathbb{P}(X = x) = +\infty,$$

then $\mathbb{E}(X)$ is not defined for such X. If one of the sums above converges but another diverges to infinity, we define $\mathbb{E}(X) = \infty$ with the sign equal to the sign of the diverging sum. For example, if $X \geq 0$ and the sum in (6.1) diverges to $+\infty$, then we define $\mathbb{E}(X) = +\infty$. Equation (6.1) is often written as

$$\mathbb{E}(X) = \sum_x x\mathbb{P}(X = x) = \sum_x x p_X(x),$$

and the expectation of X is often called the *expected value* or *mean value* of X. A few remarks on expectation are in order.

- To reinforce the idea that mathematical expectation models the mean value of a random variable consider a random variable X where

  $$\mathbb{P}(X = x_i) = \frac{1}{n}, \quad \text{with } x_i \in \mathbb{R} \text{ and } x_i \neq x_j \text{ if } i \neq j,\ 1 \leq i, j \leq n.$$

  Then from (6.1) we get

  $$\mathbb{E}X = \frac{x_1 + ... + x_n}{n},$$

  which in this simple case is precisely the arithmetic mean of the values of X.
- Notice that when the X takes only finitely many values, the requirement of absolute convergence of the series in (6.1) becomes redundant. However, when $\mathrm{Im}X$ is infinite (countable though), then the absolute convergence of the series guarantees that its sum is the same for any rearrangement of the series (i.e. irrespective of the summation order).
- The physical analogy of expectation is the idea of "centre of gravity". If masses with weights $\pi_1, \pi_2, ...$ are placed at points $x_1, x_2, ... \in \mathbb{R}$, the position of the centre of gravity is

  $$\frac{\sum \pi_i x_i}{\sum \pi_i} = \sum_i x_i \frac{\pi_i}{\sum\limits_j \pi_j}.$$

  Setting $p_i := \pi_i / \sum_j \pi_j$ gives a probability distribution that amounts for the proportion of the total weight assigned to point x_i.

Bernoulli distribution. Let X be a random variable with probability mass function equal to binomial distribution $B(n,p)$ (thanks to Theorem 6.1 such X exists). Then

$$\mathbb{E}(X) = \sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k} = \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!}p^k(1-p)^{n-k} =$$
$$np\sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!\,(n-1-(k-1))!}p^{k-1}(1-p)^{n-1-(k-1)} =$$
$$np\sum_{k=0}^{n-1}\binom{n-1}{k}p^k(1-p)^{n-1-k} = np.$$

Poisson distribution. Let X be a random variable with Poisson distribution with parameter $\lambda > 0$. Then

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} = \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = \lambda.$$

□ Exercise 6.2.1. *Assume we toss a fair coin until it comes up Heads. Let N be the first time in the trials when we see a Head. Compute $\mathbb{E}(N)$.*

Note: Notice that N takes values $1, 2, ...$ and ∞ and observe that N follows the geometric distribution.

Theorem 6.3. (`Basic properties of expectation`) *Let X be a discrete random variable for which $\mathbb{E}(X)$ is defined. Then*

(a) *if $\mathbb{P}(X \geq 0) = 1$ then $\mathbb{E}(X) \geq 0$,*

(b) *if $\mathbb{P}(X \geq 0) = 1$ and $\mathbb{E}(X) = 0$ then*[15] *$\mathbb{P}(X = 0) = 1$,*

(c) *$\mathbb{E}(aX) = a\mathbb{E}(X)$ for any $a \in \mathbb{R}$,*

(d) *for any random variable Y for which the expectation exists, the expectation of $X + Y$ exists and $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$,*

Notice that (c) and (d) together imply that $\mathbb{E}$ is a **linear operator** on discrete random variables that have expectation, more precisely if X and Y are discrete random variables whose expectations exist and $a, b \in \mathbb{R}$ are any, then

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y). \tag{6.2}$$

Proof of Theorem 6.3. (a) We have that $\mathbb{P}(X \geq 0) = 1$ hence $\mathbb{P}(X < 0) = 0$, in particular for any $x < 0$ we have $\mathbb{P}(X = x) = 0$. Hence, all non-zero terms in (6.1) are positive and we get $\mathbb{E}X \geq 0$.

(b) Since $\mathbb{P}(X \geq 0) = 1$ we get $\mathbb{P}(X < 0) = 0$ and hence

$$0 = \mathbb{E}X = \sum_{\substack{x\in \mathrm{Im}X \\ x>0}} x\mathbb{P}(X = x).$$

Since all terms on the right-hand side are non-negative we get that $\mathbb{P}(X = 0)$ if $x > 0$, in particular $\mathbb{P}(X > 0) = 0$. From here we obtain

$$\mathbb{P}(X = 0) = 1 - \mathbb{P}(X < 0) - \mathbb{P}(X > 0) = 1.$$

[15]Recall from the elementary calculus that for a continuous function $f \in C[a,b]$ if $\int_a^b |f(x)|dx = 0$, then $f \equiv 0$ on $[a,b]$. When f is Lebesgue integrable on (a,b) and $\int_a^b |f|d\mu = 0$ where μ is the Lebesgue measure, then $f = 0$ μ-almost everywhere on $[a,b]$.

(c) First observe that if $a = 0$ then both sides of the equation in (c) become 0. Now assume $a \geq 0$. Then $y \mapsto \frac{1}{a}y$ is a bijection between $\mathrm{Im}(aX)$ and $\mathrm{Im}X$. Using this we get

$$\begin{aligned}\mathbb{E}(aX) &= \sum_{y\in \mathrm{Im}(aX)} y\mathbb{P}(aX = y)\\ &= a\sum_{\frac{1}{a}y\in \mathrm{Im}X} \frac{1}{a}y\mathbb{P}\left(X = \frac{1}{a}y\right) \text{ (setting } x = \frac{1}{a}y \text{ and using that } y \mapsto \frac{1}{a}y \text{ is a bijection)}\\ &= a\sum_{x\in \mathrm{Im}X} x\mathbb{P}\left(X = x\right) = a\mathbb{E}X.\end{aligned}$$

(d) Let us first show that $\mathbb{E}(X + Y)$ exists, for which we need to show that the sum in (6.1) corresponding to $X + Y$ converges absolutely. For $x, y, z \in \mathbb{R}$ we will let $\delta_{x+y=z}$ be the *Kronecker delta symbol*, i.e.

$$\delta_{x+y=z}(x, y, z) = \begin{cases}1, & \text{if } x + y = z,\\ 0, & \text{otherwise.}\end{cases}$$

We have

$$\begin{aligned}&\sum_{z\in \mathrm{Im}(X+Y)} |z|\mathbb{P}(X + Y = z) \quad (\{X = x\}_{x\in \mathrm{Im}X} \text{ is a partition})\\ &\quad= \sum_{z\in \mathrm{Im}(X+Y)} |z| \sum_{x\in \mathrm{Im}X} \mathbb{P}(x + Y = z, X = x) \quad (\{Y = y\}_{y\in \mathrm{Im}Y} \text{ is a partition})\\ &\quad= \sum_{z\in \mathrm{Im}(X+Y)} |z| \sum_{x\in \mathrm{Im}X}\sum_{y\in \mathrm{Im}Y} \mathbb{P}(X = x, Y = y)\delta_{x+y=z}\\ &\quad= \sum_{x\in \mathrm{Im}X}\sum_{y\in \mathrm{Im}Y} \mathbb{P}(X = x, Y = y) \sum_{z\in \mathrm{Im}(X+Y)} |z|\delta_{x+y=z} =: \mathcal{E}\end{aligned}$$

Denote $\Delta_{x,y} := \sum_{z\in \mathrm{Im}(X+Y)} |z|\delta_{x+y=z}$ and observe that the sum over z can have at most one non-zero summand corresponding to the case when $z = x + y$, in particular $\Delta_{x,y}$ is finite. We now claim that

$$\mathbb{P}(X = x, Y = y)\Delta_{x,y} = \mathbb{P}(X = x, Y = y)|x + y|.$$

Indeed, if $\mathbb{P}(X = x, Y = y) = 0$ both sides become 0, otherwise we have $z \in \mathrm{Im}(X + Y)$ and hence[16] $\Delta_{x,y} = |x + y|$. With this getting back to $\mathcal{E}$ we get

$$\begin{aligned}\mathcal{E} &= \sum_{x\in \mathrm{Im}X}\sum_{y\in \mathrm{Im}Y} \mathbb{P}(X = x, Y = y)|x + y|\\ &\leq \sum_{x\in \mathrm{Im}X} |x| \sum_{y\in \mathrm{Im}Y} \mathbb{P}(X = x, Y = y) + \sum_{y\in \mathrm{Im}Y} |y| \sum_{x\in \mathrm{Im}X} \mathbb{P}(X = x, Y = y)\\ &= \sum_{x\in \mathrm{Im}X} |x|\mathbb{P}(X = x) + \sum_{y\in \mathrm{Im}Y} |y|\mathbb{P}(Y = y) < \infty,\end{aligned}$$

[16]Consider $(\Omega, \mathcal{F}, \mathbb{P})$ corresponding to a single fair coin toss experiment and let $X = 1$ for heads and $X = 0$ for tails. Let also $Y = 1 - X$. Then $1 \in \mathrm{Im}X \cap \mathrm{Im}Y$ but $2 = 1 + 1$ is a value that is not from $\mathrm{Im}(X + Y)$.

where the convergence of the last two sums is due to the existence of $\mathbb{E}(X)$ and $\mathbb{E}(Y)$. We thus get that $\mathbb{E}(X+Y)$ exists. The same argument as above shows that $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$. □

□ Exercise 6.3.1. *Construct a discrete random variable X, for which $\mathbb{E}|X| < +\infty$, but $\mathbb{E}(X^2) = +\infty$.*

Note: the example shows that for two discrete random variables X and Y, having the existence of $\mathbb{E}(X)$ and $\mathbb{E}(Y)$ does not imply the existence of $\mathbb{E}(XY)$ in general.

□ Exercise 6.3.2. *Let X and Y be two discrete random variables and assume $\mathbb{P}(|X| \leq |Y|) = 1$. Prove that if $\mathbb{E}Y$ exists then so does $\mathbb{E}X$.*

We next establish a formula allowing to compute expectation for functions of a random variable.

Theorem 6.4. (Law of the subconscious statistician) *Let X be a discrete random variable and $g : \mathbb{R} \to \mathbb{R}$ be any function. Then $\mathbb{E}g(X)$ exists and*

$$\mathbb{E}g(X) = \sum_{x \in \mathrm{Im}X} g(x)\mathbb{P}(X = x), \tag{6.3}$$

whenever the series converge absolutely.

Proof. The proof is based on a similar idea used in part (d) of Theorem 6.3. Notice that $g(X)$ is a random variable in view of the results of subsection **??**. To show that $\mathbb{E}g(X)$ exists, we will show absolute convergence of the series underlying the expectation of $g(X)$.

Denote by $\delta_{g(x)=y}$ the Kronecker delta symbol, i.e. $\delta_{g(x)=y} = 1$ if $g(x) = y$ otherwise it is 0. Then, to estimate the series (6.1) for $g(X)$ becomes

$$\begin{aligned}
&\sum_{y \in \mathrm{Im}g(X)} |y|\mathbb{P}(g(X) = y) \quad (\{X = x\}_{x \in \mathrm{Im}X} \text{ is a partition}) \\
&= \sum_{y \in \mathrm{Im}g(X)} |y| \sum_{x \in \mathrm{Im}X} \mathbb{P}(X = x,\ g(x) = y) \\
&= \sum_{y \in \mathrm{Im}g(X)} \sum_{x \in \mathrm{Im}X} |y|\mathbb{P}(X = x)\delta_{g(x)=y} \\
&= \sum_{x \in \mathrm{Im}X} \mathbb{P}(X = x) \sum_{y \in \mathrm{Im}g(X)} |y|\delta_{g(x)=y} \quad (\text{only } y \text{ s.t. } y = g(x) \text{ remains}) \\
&\sum_{x \in \mathrm{Im}X} |g(x)|\mathbb{P}(X = x) < \infty
\end{aligned}$$

where convergence of the last sum and all rearrangements of the series above are due to the absolute convergence of series in (6.3). The last inequality shows that $\mathbb{E}g(X)$ exists and the same argument without absolute values on values of $g(x)$ confirms (6.3). □

As an application of Theorem 6.4 consider a random variable X whose probability mass function is a Poisson distribution with parameter $\lambda > 0$, i.e. $\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$ for $k = 0, 1, \ldots$. Let also $Y = e^X$. To compute the expectation of Y, we apply Theorem 6.4 with $g(x) = e^x$ getting

$$\mathbb{E}Y = \sum_{k=0}^{\infty} e^k e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(e\lambda)^k}{k!} = e^{\lambda(e-1)}.$$

We next define another important concept related to random variables. It is often necessary to measure to which extent given random variable's values diverge from its mean. One of the ways to quantify such divergence is by measuring the mean square difference of its values from its mean. This brings us to the next definition.

Definition 6.4. (`Variance of a random variable`) *A **variance** of a discrete random variable X, for which $\mathbb{E}(X)$ exists, is denoted by* $\operatorname{var}(X)$ *and is defined as*

$$\operatorname{var}(X) = \mathbb{E}\left(X - \mathbb{E}(X)\right)^2. \tag{6.4}$$

*The square root of the variance is called the **standard deviation**.*

Note that Theorem 6.3 part (a) implies that given $\mathbb{E}X$ exists then so does $\operatorname{var}X$ with a value possibly being $+\infty$.

Property 6.5. (`Basic properties of variance`) *Let X be a discrete random variable and assume $\mathbb{E}(X)$ exists. Then*

(a) $\operatorname{var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$,

(b) $\operatorname{var}(aX) = a^2\operatorname{var}(X)$ *for any* $a \in \mathbb{R}$,

(c) $\operatorname{var}(X+Y) = \operatorname{var}(X) + \operatorname{var}(Y) + 2\left(\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)\right)$, *where Y is a discrete random variable for which* $\operatorname{var}(Y) < \infty$.

(d) *if $\mathbb{E}(X^2)$ is defined then $\mathbb{E}(X)$ is the minimizer of $c \mapsto \mathbb{E}[(X-c)^2]$ with $c \in \mathbb{R}$.*

Proof. (a) Setting $\mu = \mathbb{E}(X)$ and applying Theorem 6.4 with $g(x) = (x-\mu)^2$, we get

$$\begin{aligned}\operatorname{var}(X) = \sum_{x \in \operatorname{Im}X} (x-\mu)^2 \mathbb{P}(X=x) = \\ \sum_{x \in \operatorname{Im}X} (x^2 - 2x\mu + \mu^2)\mathbb{P}(X=x) = \\ \sum_{x \in \operatorname{Im}X} x^2\mathbb{P}(X=x) - 2\mu \sum_{x \in \operatorname{Im}X} x\mathbb{P}(X=x) + \mu^2 \sum_{x \in \operatorname{Im}X} \mathbb{P}(X=x) = \\ \mathbb{E}(X^2) - 2\mu\mathbb{E}(X) + 2\mu^2 = \mathbb{E}(X^2) - (\mathbb{E}(X))^2,\end{aligned}$$

completing the proof of (a)[17].

Formula (b) is a direct consequence of Property 6.3 (c) and definition of the variance given by (6.4).

We now prove (c). Observe that in view of (a) we have $\mathbb{E}X^2 < \infty$. Notice also that

$$|XY| \leq \frac{X^2 + Y^2}{2},$$

and since $\mathbb{E}X^2,\ \mathbb{E}Y^2 < \infty$ we also get that $\mathbb{E}|XY| < \infty$ directly from the definition of the expectation. Using these and part (a) we obtain

$$\begin{aligned}\operatorname{var}(X+Y) &= \mathbb{E}(X+Y)^2 - (\mathbb{E}(X+Y))^2 \quad (\text{linearity of } \mathbb{E} \text{ and } \mathbb{E}|XY|, \mathbb{E}X^2, \mathbb{E}Y^2 < \infty) \\ &= \mathbb{E}(X)^2 + 2\mathbb{E}(XY) + \mathbb{E}(Y)^2 - \left((\mathbb{E}(X))^2 + 2\mathbb{E}(X)\mathbb{E}(Y) + (\mathbb{E}(Y))^2\right) \\ &= \operatorname{var}(X) + \operatorname{var}(Y) + 2\left(\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)\right).\end{aligned}$$

It remains to show (d). For any $c \in \mathbb{R}$ we have

[17]Notice that variance is non-negative by (6.4). This in particular implies that the right-hand side of (a) is non-negative. Independently of the definition of the variance the non-negativity of the right-hand side of (a) is a consequence of **Jensen's inequality** which we will study later.

$$\begin{aligned}(6.5)\quad \mathbb{E}[(X-c)^2] &= \mathbb{E}[(X-\mathbb{E}(X)+\mathbb{E}(X)-c)^2] \text{ (by linearity of } \mathbb{E}) \\ &= \mathbb{E}[(X-\mathbb{E}(X))^2] + 2\mathbb{E}[(X-\mathbb{E}(X))(\mathbb{E}(X)-c)] + \mathbb{E}[(\mathbb{E}(X)-c)^2] \\ &= \mathbb{E}[(X-\mathbb{E}(X))^2] + (\mathbb{E}(X)-c)^2,\end{aligned}$$

where we used the following

$$\begin{aligned}\mathbb{E}[(X-\mathbb{E}(X))(\mathbb{E}(X)-c)] = \mathbb{E}[(X-\mathbb{E}(X))](\mathbb{E}(X)-c) = \\ (\mathbb{E}(X)-\mathbb{E}(X))(\mathbb{E}(X)-c) = 0.\end{aligned}$$

Now (d) follows directly from (6.5) as both terms in the last equation are non-negative.
The proof is now complete. □

We next discuss some examples involving expectation.

Example 6.5.1. `(Student trying to study probability)` *A student decided to start studying probability. Every morning she throws a* 6*-sided dice and if* 1 *shows up she will start her studies the same day, otherwise she will skip that day and repeat her experiment with a dice the next morning. On average, how many days one should wait for her to start studying?*

The waiting time of the student follows a geometric distribution with parameter $1/6$ (see subsection 5.1.4). Using it, one may compute the expected waiting time directly from the definition of expectation. But there is an alternative approach which we consider below that becomes useful in certain scenarios.

Claim 6.6. *Let X be a discrete random variable with pmf (probability mass function) $\mathbb{P}(X=k)=p_k$, where $k=0,1,2,\ldots$ and $\sum_{k=0}^{\infty} p_k = 1$. Then*

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k). \tag{6.6}$$

Proof. Notice that $\mathbb{P}(X=k) = \mathbb{P}(X\geq k) - \mathbb{P}(X \geq k+1)$ for all $k=0,1,2,\ldots$. Using this and (6.1) we have

$$\begin{aligned}\mathbb{E}(X) = \sum_{k=0}^{\infty} k\mathbb{P}(X=k) = \sum_{k=0}^{\infty} k\left(\mathbb{P}(X\geq k) - \mathbb{P}(X\geq k+1)\right) = \\ \mathbb{P}(X\geq 1) + \sum_{k=2}^{\infty} \mathbb{P}(X\geq k)(k+1-k) = \sum_{k=1}^{\infty} \mathbb{P}(X\geq k),\end{aligned}$$

where we used the Abel's summation rule[18] to pass from the first row to the second. □

Solution to Example 6.5.1. Define X to be the first day that 1 shows up when throwing a dice. Clearly $\mathbb{P}(X=1)=\frac{1}{6}$, $\mathbb{P}(X=2)=\frac{5}{6}\frac{1}{6}$ and in general $\mathbb{P}(X=k) =$

[18]Given a sum of the form $\sum_{k=1}^{\infty} a_k(b_k - b_{k+1})$ we can rewrite it in the form

$$\begin{aligned}a_1(b_1-b_2) + a_2(b_2-b_3) + \ldots = a_1b_1 + b_2(a_2-a_1) + b_3(a_3-a_2) + \ldots \\ = a_1b_1 + \sum_{k=2}^{\infty} b_k(a_{k+1}-a_k),\end{aligned}$$

whenever the series converge. This can be considered as the analogue of *integration by parts* rule for the integrals and is often referred to as *Abel's summation by parts*.

$\left(\frac{5}{6}\right)^{k-1}\frac{1}{6}$ for $k \geq 1$, thus X follows a geometric distribution with parameter $p = 1/6$. From here, for each $n \geq 1$ we obtain

$$\mathbb{P}(X \geq n) = \sum_{k=n}^{\infty}\left(\frac{5}{6}\right)^{k-1}\frac{1}{6} = \frac{1}{6}\left(\frac{5}{6}\right)^{n-1}\sum_{k=0}^{\infty}\left(\frac{5}{6}\right)^{k} = \left(\frac{5}{6}\right)^{n-1}$$

Using this we compute $\mathbb{E}(X)$ from (6.6) as follows

$$\mathbb{E}(X) = \sum_{n=1}^{\infty}\mathbb{P}(X \geq n) = \sum_{n=1}^{\infty}\left(\frac{5}{6}\right)^{n-1} = 6.$$

Thus the expected number of days to wait equals 6. □

□ Exercise 6.6.1. *Confirm the answer to the Example 6.5.1 directly from the definition of expectation, by showing that*

$$\sum_{k=0}^{\infty} k\mathbb{P}(X = k) = 6.$$

Example 6.6.1. (Group testing of a disease) *A certain disease in a population can be tested and identified by a blood sample. Assume the size of the population is $n \geq 1$ and each individual has probability $0 < p < 1$ of having the disease independently of others. A straightforward approach in identifying all people carrying the disease is to collect blood samples from each individual and test them. This will require n tests - the size of the population. Is there a procedure that requires smaller number of tests on average and still identifies individuals with the disease?*

Solution to Example 6.6.1. One strategy toward the example above is to split the population into groups of k, for some fixed $k \geq 1$, then mix the samples from individuals of the same group and test the mixed sample. If the test on the mix is negative, then no one in the group has the disease, otherwise we test everyone in the group one by one, using another k tests. Thus, in the worst case we will need $k+1$ tests in total for a single group to know with certainty who has a disease in the group. Let X be the number of tests required for a single group. Then X takes two values: 1 with probability $(1-p)^k$ and $k+1$ with probability $1-(1-p)^k$. Hence

$$\mathbb{E}(X) = (1-p)^k + (k+1)\left(1-(1-p)^k\right).$$

For simplicity assume that k divides n, hence there will be n/k groups. Consequently,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{n/k} X_i\right] \\
&= \frac{n}{k}\left[(1-p)^k + (k+1)\left(1-(1-p)^k\right)\right] \\
&= n\left[1-(1-p)^k + \frac{1}{k}\right].
\end{aligned}$$

Recall that k is a parameter of the testing scheme and we now optimize the value of the expectation with respect to k. To that end, differentiate the right-hand side of the last expression with respect to k and consider the value of k that makes the derivative to become 0. We thus get

$$\frac{1}{k^2} + (1-p)^k\log(1-p) = 0.$$

If p is much smaller than $\frac{1}{k}$, then we can approximate $(1-p)^k$ by 1 and $\log(1-p)$ by $-p$, arriving at $p \approx \frac{1}{k^2}$. Thus $k = \frac{1}{\sqrt{p}}$ and the resulting expectation becomes $n\sqrt{p}$ which is significantly less than n for smaller values of p. □

→ **Problem 6.6.1.** *Assume we toss a fair coin until we get two Heads in a row, and then we stop (i.e. the first time we see the pattern HH, we stop). Let N be the first time when we see two Heads.*

- *Compute the expectation of N.*
- *Do the same computation for the pattern TH (i.e. seeing for the first time a Tail followed immediately by Heads) and observe that the expecting time for the two patters HH and TH is different.*
- *How about getting the pattern HHT for the first time?*

→ **Problem 6.6.2.** *A fair 6-sided dice is thrown. We can either take the number that appears on the dice as our winning (e.g. if its 3 we get 3$ as a winning) and the game stops, or we can ask to throw it again. We are given $n \geq 1$ chances to ask for a new round (this includes the initial round). If we play it up to the last possible throw, then we must take whatever appears on the dice on that throw. What is the optimal strategy of playing this game and what is the expected winning if we play according to that strategy?*

Note: For example, when $n = 1$, there is only 1 throw, which we must accept, and thus our expected winning is $\frac{1}{6}(1+2+...+6) = 3.5$. If $n = 2$ we throw it once and after seeing the result we can decide if we want to keep it or move to the second round in which case we must take whatever appears there.

Hint: The case of $n = 1$ is described above. Start with $n = 2$ and think about what is your expected winning if you move on the first throw.

→ **Problem 6.6.3.** *Let $N > 1$ be a fixed integer. Your are offered to play a game where the referee picks an integer n from 1 to N inclusive and if you guess correctly you will win n, otherwise 0. How much would you be willing to pay to play this game?*

Note: first try to think about the optimal strategy for both the referee and the player. The question asks about the entry fee that will make the game fair both for the player and the organizer, in a sense that the expected gain of the player equals expected loss of the organizer.

6.3. **Conditional expectation.** Recall from section 4.1 that given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and an event $B \in \mathcal{F}$, knowing that B occurred can affect the probabilities of other events. To formalize this idea, we introduced the concept of *conditional probability*. In the same spirit knowing that some $B \in \mathcal{F}$ occurred can affect the values of probability mass function (pmf) of a given random variable X as there we are dealing with events $\{X = x\}$ for some $x \in \mathrm{Im}X$. More precisely, if $B \in \mathcal{F}$ and $\mathbb{P}(B) > 0$, then the events $\mathbb{P}(\{X = x\})$ can be conditioned on B.

Definition 6.5. `(Conditional expectation)` *Let X be a discrete random variable and $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. The conditional expectation of X is denoted by $\mathbb{E}(X \,|\, B)$ and is defined as*

$$\mathbb{E}(X \,|\, B) = \sum_{x \in \mathrm{Im}X} x\mathbb{P}(X = x \,|\, B), \tag{6.7}$$

whenever the sum converges absolutely.

Notice that the only difference of (6.7) from (6.1) is that we are taking conditional probabilities of X taking the value x. The analogue of Theorem 4.2 (law of total probability) is the following.

Theorem 6.7. (Partition theorem for conditional expectation) *Let X be a discrete random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and $\{B_i : i = 1, 2, ...\} \subset \mathcal{F}$ be a finite or countable partition of Ω with $\mathbb{P}(B_i) > 0$ for all $i = 1, 2, ...$. Then*

$$\mathbb{E}(X) = \sum_i \mathbb{E}(X \,|\, B_i)\mathbb{P}(B_i),$$

whenever the sum converges absolutely.

Proof. We have

$$\begin{aligned} &\sum_i \mathbb{E}(X \,|\, B_i)\mathbb{P}(B_i) \text{ (by (6.7))} \\ = &\sum_i \sum_{x \in \mathrm{Im}X} x\mathbb{P}(X = x \,|\, B)\mathbb{P}(B_i) \text{ (by absolute convergence)} \\ = &\sum_{x \in \mathrm{Im}X} x \sum_i \mathbb{P}(X = x \,|\, B)\mathbb{P}(B_i) \text{ (by Theorem 4.2)} \\ &\sum_{x \in \mathrm{Im}X} x\mathbb{P}(X = x) = \mathbb{E}(X). \end{aligned}$$

□

→ **Problem 6.7.1.** *For an integer $n \geq 1$ let $X(n)$ be an integer chosen uniformly at random from $\{0, 1, ..., n-1\}$. If $X(n)$ is 0, we stop, otherwise choose an integer uniformly at random from $\{0, ..., X(n)-1\}$, in other words we consider $X(X(n))$. Let τ be the number of times X was applied before getting 0, i.e. $(\underbrace{X \circ ... \circ X}_{\tau \text{ times}})(n) = 0$. What is the expected value of τ?*

Note: For example, when $X(n) = 0$, i.e. we choose 0 at the first step, then $\tau = 1$. If $n = 10$ and we choose 5 on the first step, then 0 on the second, then $\tau = 2$. Since on each step we decrease the upper bound n be at least 1, then $\tau \leq n$, meaning the process will terminate in at most n steps.

6.4. **Indicator functions.** An important class of random variables are the *indicator functions* of events. For $A \in \mathcal{F}$ define

$$\mathbb{I}_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{otherwise,} \end{cases} \tag{6.8}$$

and call $\mathbb{I}_A$ the **indicator function** of A (sometimes it is called a *characteristic function* of A). It follows from the definition that $\mathbb{I}_A$ is a random variable in a sense of Definition 6.1. Moreover, it enjoys the following nice properties.

Property 6.8. (Basic properties of indicators) *Let $A \in \mathcal{F}$ and $\mathbb{I}_A$ be its indicator function. Then*

1. $\mathbb{E}(\mathbb{I}_A) = \mathbb{P}(A)$,
2. $\mathbb{I}_A = 1 - \mathbb{I}_{A^c}$,
3. $\mathbb{I}_{A \cap B} = \mathbb{I}_A \cdot \mathbb{I}_B$ *for any* $B \in \mathcal{F}$,
4. $\mathbb{I}_{A \cup B} = \mathbb{I}_A + \mathbb{I}_B - \mathbb{I}_{A \cap B}$ *for any* $B \in \mathcal{F}$,
5. $\mathbb{E}(\mathbb{I}_A \cdot \mathbb{I}_B) = \mathbb{E}\mathbb{I}_A \cdot \mathbb{E}\mathbb{I}_B$ *for any* $B \in \mathcal{F}$ *if A and B are independent.*

Proof. All claims follow directly from definitions. □

□ Exercise 6.8.1. *Prove Property 6.8.*

The rest of the subsection is devoted to various applications of indicator functions that are of independent interest. We will start with the *inclusion-exclusion principle*. Notice that applying Property 6.8 (4) and the *linearity of expectation* proved in Theorem 6.3 we get the inclusion-exclusion principle for two sets. We now recover the inclusion-exclusion for n sets using indicator functions.

Inclusion-exclusion principle (Proposition 3.4) via indicator functions. Let $A_1, ..., A_n \in \mathcal{F}$ and set $A = \cup_{i=1}^n A_i$. Observe that

$$(\mathbb{I}_A - \mathbb{I}_{A_1}) \cdot ... \cdot (\mathbb{I}_A - \mathbb{I}_{A_n}) = 0 \text{ on } \Omega. \tag{6.9}$$

Indeed, if $\omega \notin A$ then all factors above are vanishing. If $\omega \in A$ then there exists $1 \leq i \leq n$ such that $\omega \in A_i$, hence $\mathbb{I}_A(\omega) - \mathbb{I}_{A_i}(\omega) = 0$ and we get (6.9). Next, notice that $\mathbb{I}_A \mathbb{I}_{A_i} = \mathbb{I}_{A_i}$ for all $1 \leq i \leq n$. Thus, expanding the product in (6.9) we get[19]

$$\mathbb{I}_A + \sum_{\substack{J \subset \{1,2,...,n\} \\ J \neq \emptyset}} (-1)^{|J|} \prod_{i \in J} \mathbb{I}_{A_i} = 0 \text{ on } \Omega.$$

Applying expectation on both sides of the above equation and using items 1 and 3 of Property 6.8 we get (3.3) and complete the proof. □

Indicator functions become very useful in many counting problems. Below we discuss a few examples of the so-called ***method of indicators***.

Example 6.8.1. `(Kings and queens at a round table)` *There are $n \geq 2$ pairs of kings/queens seated at a round table where kings take the even numbered seats and queens have the odd numbered ones, the point being that kings and queens seat interleaving. On average how many kings will have their queens seated next to them?*

Solution to Example 6.8.1. Enumerate the pairs by $1, 2, ..., n$ and for $1 \leq i \leq n$ let A_i be the event that the queen of the i-th pair seats next to the king of the i-th pair. Let also N be the number of the correctly seated pairs. Observe that N is a random variable taking values from $\{0, 1, ..., n\}$ and we are asked to compute $\mathbb{E}(N)$. It is clear that

$$N = \mathbb{I}_{A_1} + ... + \mathbb{I}_{A_n},$$

and thanks to the linearity of the expectation (Theorem 6.3 parts (c) and (d)) and Property 6.8 part 1 of indicators we get $\mathbb{E}(N) = \sum_{i=1}^n \mathbb{P}(A_i)$. Notice that for each $1 \leq i \leq n$ we have $\mathbb{P}(A_i) = \frac{2}{n}$ since if the position of the king is fixed, then out of n available seats the queen can take only 2 seats that are next to their king. We thus get that $\mathbb{E}(N) = n\frac{2}{n} = 2$ for all $n \geq 2$. □

[19] In (6.9) we have a product of the form $(a_1 - b_1) \cdot ... \cdot (a_n - b_n)$. To expand it we take a single factor from each bracket a_i or $-b_i$ for all $i = 1, 2, ..., n$. Let $I \subset \{1, 2, ..., n\}$ be the set of indicies where we choose a_i and the complement of I will comprise of indicies where $-b_i$ is chosen. Thus, we get a sum of the form

$$\sum_{I \subset \{1,2,...,n\}} \prod_{i \in I} a_i \prod_{i \in \{1,2,...,n\} \setminus I} (-b_i).$$

Applying the expansion above to (6.9) the part with $\prod_{i \in I} \mathbb{I}_A$ becomes $\mathbb{I}_A$ thanks to $\mathbb{I}_A \mathbb{I}_A = \mathbb{I}_A$ and $\mathbb{I}_A \mathbb{I}_{A_i} = \mathbb{I}_{A_k}$ for all $k = 1, 2, ..., n$.

□ Exercise 6.8.2. *A fair coin is tossed $n \geq$ times. What is the expected number of Heads? What is the coin turns Heads with probability $0 \leq p \leq 1$?*

Notice that we are considering Bernoulli trials here, see subsection 5.1.2. Compute the expected number of Heads using two methods: first, directly from the definition of the expectation and Bernoulli trials, second, representing the number of Heads as a sum of indicators of Heads for each toss and using the linearity of expectation (cf. Example 6.8.1).

Example 6.8.2. `(Parallel series of coin flips)` *Assume two coins, one with probability $0 \leq p < 1$ of turning up Heads, another with probability $q > p$ are tossed repeatedly. Show that for any $k \in \mathbb{N}$ the probability that in a series of n independent tosses the first coin will produce at least k heads is less than or equal to the probability that the second one will produce at least k heads.*

Solution to Example 6.8.2. The claim is of course intuitive, since the second coin has a greater chance of landing up Heads. Nevertheless, proving this by counting arguments is not entirely straightforward. We will instead use a method related to indicator functions. For $1 \leq i \leq n$ define

$$X_i = \begin{cases} 1, & \text{if the } i\text{-th toss of the first coin is Head,} \\ 0, & \text{otherwise} \end{cases}$$

which is simply the indicator of the i-th toss of the first coin being Heads. Consider an auxiliary series of coin tosses where the probability of landing Heads equals $\frac{q-p}{1-p}$, and let Z_i be the indicator of the i-th toss turning up Heads for this new series of tosses, with $i = 1, 2, ..., n$. We thus have that $\mathbb{P}(Z = 1) = \frac{q-p}{1-p}$. Also assume that the new series of tosses is independent of the tosses of the first coin (we know from section 4.3.1 that such series of independent experiments exist). Now for the second coin define $Y_i := X_i + (1 - X_i)Z_i$, where $i = 1, 2, ..., n$. Clearly Y takes only two values 0 and 1 and by definition of Z_i we have $\mathbb{P}(Y_i = 1) = q$ and $\mathbb{P}(Y_i = 0) = 1 - q$, hence Y_i has the same distribution as the indicator of the Heads in a coin toss of the second coin. By construction we have $X_i \leq Y_i$ for all $i = 1, 2, ..., n$ and hence

$$\mathbb{P}(X_1 + ... + X_n \geq k) \leq \mathbb{P}(Y_1 + ... + Y_n \geq k)$$

proving the claim[20]. □

6.4.1. ***Cycles in random permutation.*** Let σ be a permutation of $\{1, 2, ..., n\}$. A **cycle** of length m in σ is a sequence $(i_1\ i_2\ ...\ i_{m-1})$ of distinct integers from $\{1, ..., n\}$ where $\sigma(i_k) = \sigma(i_{k+1})$ for all $k = 1, 2, ..., m-2$ and $\sigma(i_{m-1}) = \sigma(i_1)$. A cycle of length 1 is a fixed point of permutation σ, i.e. an integer i such that $\sigma(i) = i$. For example, let $n = 4$ and consider a permutation σ defined as

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 \end{pmatrix}.$$

Then we have two cycles, namely (1 3 4) and (2). It is well-known that any permutation can be decomposed as a product of disjoint cycles. In this section we will be interested in the *average length* of a cycle in a random permutation.

Step 1. The probability that a given element is in a cycle of length m.

[20]The technique used here is an example of a **coupling** method, which is a powerful technique in probability theory and has deep applications that are beyond the scope of this notes.

Fix an element $1 \leq i \leq n$ and assume it is in a cycle of length m. Then there exist distinct integers $i_1, i_2, ..., i_{m-1}$ from $\{1, 2, ..., n\}$ forming the cycle

$$i \mapsto i_1 \mapsto i_2 \mapsto ... \mapsto i_{m-1} \mapsto i.$$

Since the sequence $i_1, ..., i_{m-1}$ is ordered there are $\frac{(n-1)!}{((n-1-(m-1))!}$ ways to choose such a sequence (see subsection 2.3.2). When the cycle is fixed the rest of $n-m$ integers can be permuted arbitrarily, thus there are $\frac{(n-1)!}{(n-m)!}(n-m)!$ permutations out of $n!$ such that i is in a cycle of length m. So the probability that i is in a cycle of length m equals $\frac{1}{n}$.

Step 2. The expected number of cycles of length m.

For each $1 \leq i \leq n$ let $\mathbb{I}_i$ be the indicator of the event that i lies in a cycle of length m. Let also N be the total number of cycles of length m. Then

$$N = \frac{1}{m}\left(\mathbb{I}_1 + \mathbb{I}_2 + ... + \mathbb{I}_n\right),$$

where the factor $\frac{1}{m}$ is due to the fact that each cycle is counted exactly m times in the sum of indicators. We thus get

$$\mathbb{E}(N) = \frac{1}{m}\sum_{i=1}^{n}\mathbb{E}(\mathbb{I}_i) = \frac{1}{m}\sum_{i=1}^{n}\mathbb{P}\left(i \text{ is in a cycle of length } m\right) \overset{\text{(by Step 1)}}{=} \frac{1}{m}.$$

Step 3. Probability of a long cycle.

Clearly for $m > \frac{n}{2}$ there can be at most one cycle of length m. Let p_m be the probability that a randomly chosen permutation has a cycle of length m where $m > \frac{n}{2}$ is fixed. By Step 2 we have $\mathbb{E}(N) = \frac{1}{m}$ where N is the number of cycles of length m. But in this case N takes only values 1 and 0 thus

$$\frac{1}{m} = \mathbb{E}(N) = 1 \cdot \mathbb{P}(N = 1) = p_m.$$

We conclude that for $m > \frac{n}{2}$ the probability that a random permutation has a cycle of length m equals $\frac{1}{m}$. Hence the probability that a randomly chosen permutation has a cycle of length greater than $n/2$ is bounded above by

$$\sum_{m:\ m>\frac{n}{2}}^{n}\frac{1}{m} \approx \ln n - \ln\frac{n}{2} = \ln 2.$$

Problem about names in boxes. There is a peculiar application of the above estimates on cycles in permutation.

> Assume the names of 100 prisoners are placed randomly in boxes with numbers $1, 2, ..., 100$ - one name in each box. One by one, prisoners enter the room where the boxes are placed, with the goal to find their names from the boxes. Each prisoner is allowed to open at most 50 boxes. If all prisoners find their names, they are freed. The prisoner leaves the room with boxes without altering its state. They cannot communicate with each other once the process starts. What strategy can prisoners choose to have a non-trivial chance of being freed?

Here is a possible strategy: prisoners start by forming a line which will not be changed during the process. Starting with the first in the line, the i-th prisoner enters the room and looks into the box with number i. If it contains his name we are done, otherwise it contains the name of a prisoner lined i_1 and then the current prisoner goes and checks the box number i_1. If he finds the name of a prisoner lined i_2 instead he opens the box i_2 and so on. In this process, the prisoner will eventually find his name, numbered i, unless i is in a cycle of length larger than 50 (the maximal number of boxes one can open). As we proved above, the probability that the permutation of names in boxes has a cycle of length more than 50 is bounded above by $\approx \ln 2 \approx 0.693$. Thus the probability that all will find their names is at least $\approx 1 - 0.693 = 0.307$.

□ Exercise 6.8.3. *What are the chances of survival in the above problem if all prisoners choose their* 50 *boxes uniformly at random?*

→ **Problem 6.8.1.** *Is there a better strategy to the above problem with names in boxes? Justify your answer.*

7. Independent random variables

In this section we introduce an important concept of *independence* of discrete random variables. We study several criteria for independence and prove a key property related to the expectation of a product of independent random variables. We also discuss extensions of Theorem 6.4 (the law of subconscious statistician) to the multivariate case.

7.1. Multivariate discrete distribution.

Let X and Y be discrete random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Often it is necessary to consider X and Y as a pair taking values in $\mathbb{R}^2$ rather than individual variables. In analogy with Definition 6.2, here for the pair of random variables we have the following.

Definition 7.1. (`Joint probability mass function`) *Let X and Y be discrete random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Their **joint probability mass function** $p_{X,Y} : \mathbb{R}^2 \to [0,1]$ is defined by*

$$p_{X,Y}(x,y) = \mathbb{P}\left(\{\omega \in \Omega : \ X(\omega) = x \text{ and } Y(\omega) = y\}\right),$$

which is usually abbreviated to $p_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y)$.

Clearly we have

$$p_{X,Y}(x,y) = 0 \text{ if } x \notin \mathrm{Im}X \text{ or } y \notin \mathrm{Im}Y,$$

and

$$\sum_{x \in \mathrm{Im}X} \sum_{y \in \mathrm{Im}Y} p_{X,Y}(x,y) = 1.$$

Notice that the events $\{\{X = x\} : x \in \mathrm{Im}X\}$ and $\{\{Y = y\} : y \in \mathrm{Im}Y\}$ are both partitions of Ω in the usual sense of Definition 4.2. Thus

$$\sum_{x \in \mathrm{Im}X} p_{X,Y}(x,y) = \sum_{x \in \mathrm{Im}X} \mathbb{P}(X = x, Y = y) = \mathbb{P}(Y = y) = p_Y(y), \tag{7.1}$$

similarly

$$\sum_{y \in \mathrm{Im}Y} p_{X,Y}(x,y) = p_X(x). \tag{7.2}$$

The formulas for p_X and p_Y we obtained above indicate that mass function of individual variables X and Y can be recovered from the joint probability mass function of X and

Y. The distributions in (7.1) and (7.2) are called **marginal** mass functions of Y and X respectively.

	y_1	y_2	...	y_m	...
x_1	$p_{X,Y}(x_1,y_1)$	$p_{X,Y}(x_1,y_2)$	...	$p_{X,Y}(x_1,y_m)$	...
x_2	$p_{X,Y}(x_2,y_1)$	$p_{X,Y}(x_2,y_2)$	...	$p_{X,Y}(x_2,y_m)$	...
$\vdots$					
x_n	$p_{X,Y}(x_n,y_m)$	$p_{X,Y}(x_n,y_m)$	...	$p_{X,Y}(x_n,y_m)$	...
$\vdots$	...	...	...	...	...

TABLE 1. One may think about the joint mass function $p_{X,Y}$ as a matrix where (n,m)-th element is the probability of the event $\{\omega\in\Omega:\ X(\omega)=x_n \text{ and } Y(\omega)=y_m\}$. Here $\mathrm{Im}X=\{x_1,x_2,\ ...\}$ and $\mathrm{Im}Y=\{y_1,y_2,\ ...\}$. To recover the marginal distributions we sum over the columns of this matrix for fixed row to get p_X and over the rows of the matrix for a fixed column to recover p_Y.

Given two random variables X and Y on $(\Omega,\mathcal{F},\mathbb{P})$ and a function $g:\mathbb{R}^2\to\mathbb{R}$ define $Z(\omega)=g(X(\omega),Y(\omega))$ where $\omega\in\Omega$. Then, one may easily check that conditions of Definition 6.1 are satisfied for Z, i.e. Z is a discrete random variable. We now formulate and prove the analogue of Theorem 6.4 for Z which will become a very useful tool in computing expectations in multivariate case.

Theorem 7.1. (`Law of the subconscious statistician - 2d case`) *Let X, Y and g be as above. Then $\mathbb{E}g(X,Y)$ exists and*

$$\mathbb{E}g(X,Y)=\sum_{x\in\mathrm{Im}X}\sum_{y\in\mathrm{Im}Y}g(x,y)p_{X,Y}(x,y), \tag{7.3}$$

whenever the sum converges absolutely.

Proof. The proof is similar to that of Theorem 6.4. We first show that $\mathbb{E}g(X,Y)$ exists for which one needs to prove absolute convergence of the following series

$$\begin{aligned}\sum_{z\in\mathrm{Im}g(X,Y)}|z|\mathbb{P}(g(X,Y)=z)\quad&(\{X=x,Y=y\}_{x\in\mathrm{Im}X,y\in\mathrm{Im}Y}\text{ is a partition})\\&=\sum_{z\in\mathrm{Im}g(X,Y)}|z|\sum_{x\in\mathrm{Im}X}\sum_{y\in\mathrm{Im}Y}\mathbb{P}(X=x,Y=y)\delta_{g(x,y)=z}\\&=\sum_{x\in\mathrm{Im}X}\sum_{y\in\mathrm{Im}Y}\mathbb{P}(X=x,Y=y)\sum_{z\in\mathrm{Im}g(X,Y)}|z|\delta_{g(x,y)=z}=:\mathcal{E}\end{aligned}$$

where as before δ stands for the Kronecker symbol, i.e.

$$\delta_{g(x,y)=z}=\begin{cases}1, & \text{if } g(x,y)=z,\\ 0, & \text{otherwise .}\end{cases}$$

Set $\Delta_{x,y}:=\sum_{z\in\mathrm{Im}g(X,Y)}|z|\delta_{x+y=z}$. Clearly there is at most one non-zero summand in this sum corresponding to the case when $z=g(x,y)$. We have in particular that $\Delta_{x,y}$ is finite. Next, we show that

$$\mathbb{P}(X=x,Y=y)\Delta_{x,y}=\mathbb{P}(X=x,Y=y)g(x,y).$$

Indeed, if $\mathbb{P}(X = x, Y = y) = 0$ both sides become 0, otherwise we have $z \in \mathrm{Im}g(X, Y)$ and hence[21] $\Delta_{x,y} = g(x, y)$. Getting back to $\mathcal{E}$ we obtain

$$\mathcal{E} = \sum_{x \in \mathrm{Im}X} \sum_{y \in \mathrm{Im}Y} \mathbb{P}(X = x, Y = y)|g(x, y)|,$$

which converges thanks to the condition of the theorem. It follows that $\mathbb{E}g(X, Y)$ exists and the same argument as above but without absolute values on the values of $g(x, y)$ settles the equality in (7.4). □

We now recover a result proved in Theorem 6.3.

Corollary 7.2. (`Linearity of expectation, yet again`) *$\mathbb{E}$ acts linearly on random variables, namely if X and Y are discrete random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, for which $\mathbb{E}X$ and $\mathbb{E}Y$ both exist, $a, b \in \mathbb{R}$, then $\mathbb{E}(aX + bY)$ exists and we have*

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y).$$

Proof. Take $g(x, y) = ax + by$, then

$$\begin{aligned}
&\sum_{x \in \mathrm{Im}X} \sum_{y \in \mathrm{Im}Y} |g(x, y)| p_{X,Y}(x, y) \\
&\qquad \leq |a| \sum_{x \in \mathrm{Im}X} |x| \sum_{y \in \mathrm{Im}Y} p_{X,Y}(x, y) + |b| \sum_{y \in \mathrm{Im}Y} |y| \sum_{x \in \mathrm{Im}X} p_{X,Y}(x, y) \\
&\qquad = |a| \sum_{x \in \mathrm{Im}X} |x| \mathbb{P}(X = x) + |b| \sum_{y \in \mathrm{Im}Y} |y| \mathbb{P}(Y = y) < \infty,
\end{aligned}$$

where the last two series are convergent due to the existence of $\mathbb{E}X$ and $\mathbb{E}Y$. Now the corollary follows from Theorem 7.1. □

Note, that the result of Theorem 7.1 works similarly for the case of n random variables. More precisely, consider $g : \mathbb{R}^n \to \mathbb{R}$ and let $X_1, X_2, ..., X_n$ be discrete random variables defined on Ω. Then $Z = g(X_1, ..., X_n)$ defines a discrete random variable on Ω and the analogue of (7.4) is true for joint mass function $p_{X_1,...,X_n}$, which is defined in analogy with the case of two variables as follows:

$$p_{X_1,...,X_n}(x_1, ..., x_n) = \mathbb{P}\left(X_1 = x_1, ..., X_n = x_n\right).$$

We formulate the result for n variables.

Theorem 7.3. *Let $X_1, X_2, ..., X_n$ be random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and $g : \mathbb{R}^n \to \mathbb{R}$ be any function. Then $g(X_1, X_2, ..., X_n)$ is a random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and*

$$\mathbb{E}g(X_1, ..., X_n) = \sum_{x_1 \in \mathrm{Im}X_1} ... \sum_{x_n \in \mathrm{Im}X_n} g(x_1, ..., x_n) p_{X_1,...,X_n}(x_1, ..., x_n), \tag{7.4}$$

whenever the sum converges absolutely.

□ Exercise 7.3.1. *Prove Theorem 7.3 following the proof of Theorem 7.1.*

[21] Let g be identically 0 and both X and Y be identically 1. Then $1 \in \mathrm{Im}X \cap \mathrm{Im}Y$ but $2 = 1 + 1$ is a value that is not from $\mathrm{Im}g(X, Y)$.

7.2. **Independence of two random variables.** Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, recall that two events $A, B \in \mathcal{F}$ are called independent, if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. We now extend the concept of independence from events to random variables.

Definition 7.2. (`Independence of discrete random variables`) *Discrete random variables X and Y defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are called independent, if*

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y) \text{ for all } x, y \in \mathbb{R}. \tag{7.5}$$

The next result provides a simple criteria for independence.

Theorem 7.4. *Two discrete random variables X and Y defined on $(\Omega, \mathcal{F}, \mathbb{P})$ are independent if and only if there exist functions $f, g : \mathbb{R} \to \mathbb{R}$ such that*

$$p_{X,Y}(x, y) = f(x)g(y), \quad \forall x, y \in \mathbb{R}. \tag{7.6}$$

Proof. The necessity is due to the definition of independence with $f(x) := \mathbb{P}(X = x)$ and $g(y) := \mathbb{P}(Y = y)$ with $x, y \in \mathbb{R}$. We now prove the sufficiency. Assume $p_{X,Y}(x, y) = f(x)g(y)$ where $x, y \in \mathbb{R}$. We need to check that (7.5) holds. We have

$$\begin{aligned} 1 = \sum_{x \in \mathrm{Im}X} \sum_{y \in \mathrm{Im}Y} p_{X,Y}(x, y) &= \sum_{x \in \mathrm{Im}X} \sum_{y \in \mathrm{Im}Y} f(x)g(y) \\ &= \left(\sum_{x \in \mathrm{Im}X} f(x) \right) \left(\sum_{y \in \mathrm{Im}Y} g(y) \right). \end{aligned} \tag{7.7}$$

We also have

$$p_X(x) = \sum_{y \in \mathrm{Im}Y} p_{X,Y}(x, y) = \left(\sum_{y \in \mathrm{Im}Y} g(y) \right) f(x)$$

and

$$p_Y(y) = \sum_{x \in \mathrm{Im}X} p_{X,Y}(x, y) = \left(\sum_{x \in \mathrm{Im}X} f(x) \right) g(y).$$

Multiplying the last two equalities implies

$$\begin{aligned} p_X(x)p_Y(y) &= f(x)g(y) \left(\sum_{y \in \mathrm{Im}Y} g(y) \right) \left(\sum_{x \in \mathrm{Im}X} f(x) \right) \text{ (by (7.7))} \\ &= f(x)g(y) \text{ (by (7.6))} \\ &= p_{X,Y}(x, y). \end{aligned}$$

The proof is now complete. □

Example 7.4.1. *Consider random variables X and Y that take values from non-negative integers and have the joint mass function*

$$p_{X,Y}(n, m) = \frac{1}{n!\, m!} \lambda^n \mu^m e^{-(\lambda+\mu)}, \ n, m = 0, 1, 2, \dots .$$

Then, Theorem 7.4 implies that X and Y are independent.

An important implication of independence concerns expectation of the product of random variables.

Theorem 7.5. *Let X and Y be discrete random variables that are independent, and assume $\mathbb{E}(X)$ and $\mathbb{E}(Y)$ exist. Then $\mathbb{E}(XY)$ exists and*

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

Proof. We will apply Theorem 7.1 with $g : \mathbb{R}^2 \to \mathbb{R}$ defined as $g(x, y) = xy$, for which we need to show that the following series converge

$$\begin{aligned}\sum_{x\in \mathrm{Im}X}\sum_{y\in \mathrm{Im}Y} |xy|\mathbb{P}(X = x, Y = y) \text{ (by independence)}\\ &= \sum_{x\in \mathrm{Im}X}\sum_{y\in \mathrm{Im}Y} |xy|\mathbb{P}(X = x)\mathbb{P}(Y = y)\\ &= \left(\sum_{x\in \mathrm{Im}X} |x|\mathbb{P}(X = x)\right)\left(\sum_{y\in \mathrm{Im}Y} |y|\mathbb{P}(Y = y)\right) < \infty,\end{aligned}$$

where the last two series converge thanks to the existence of $\mathbb{E}(X)$ and $\mathbb{E}(Y)$. Thus, Theorem 7.1 shows the existence of $\mathbb{E}(XY)$, and also that

$$\mathbb{E}(XY) = \sum_{x\in \mathrm{Im}X}\sum_{y\in \mathrm{Im}Y} xy\mathbb{P}(X = x, Y = y).$$

The last sum equals $\mathbb{E}(X)\mathbb{E}(Y)$ by the same argument as we had above for the existence of $\mathbb{E}(XY)$. The proof is complete. □

Corollary 7.6. *Let X and Y be discrete random variables that are independent. Then*

$$\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y).$$

Proof. This follows directly from Property 6.5 (c) and Theorem 7.5. □

Notice that we have **linearity of expectation** regardless of the independence of the random variables involved. The *linearity of the variance*, however, in not true in general and requires additional conditions. The next example shows that the converse of Theorem 7.5 is not true.

Example 7.6.1. *Let X be a discrete random variable taking values $-1, 0, 1$ all with probability $1/3$ and let $Y = |X|$. Then*

$$\mathbb{P}(X = 0, Y = 1) = 0 \quad \textit{and} \quad \mathbb{P}(X = 0)\mathbb{P}(Y = 1) = \frac{1}{3}\frac{2}{3} \neq 0,$$

hence X and Y are not independent. However, $\mathbb{E}(XY) = 0 = \mathbb{E}(X)\mathbb{E}(Y)$.

The next theorem provides another criteria for independence in terms of expectations.

Theorem 7.7. *Discrete random variables X and Y are independent if and only if for any functions $f, g : \mathbb{R} \to \mathbb{R}$ we have*

$$\mathbb{E}(f(X)g(Y)) = \mathbb{E}(f(X))\mathbb{E}(g(Y)),$$

whenever the expectations on the right-hand side exist.

Proof. The necessity follows by the same argument as we used in Theorem 7.5, one only needs to consider the function $(x, y) \mapsto f(x)g(y)$ instead of $(x, y) \mapsto xy$.

We now prove sufficiency. Take any $a, b \in \mathbb{R}$ and let

$$f(x) = \begin{cases} 1, & x = a,\\ 0, & x \neq a\end{cases} \qquad \text{and} \qquad g(y) = \begin{cases} 1, & y = b,\\ 0, & y \neq b\end{cases}.$$

Then $\mathbb{E}(f(X)g(Y)) = \mathbb{P}(X = a, Y = b)$ and $\mathbb{E}(f(X)) = \mathbb{P}(X = a)$, $\mathbb{E}(g(Y)) = \mathbb{P}(Y = b)$. Combining these with the condition of the theorem gives independence of X and Y. □

7.3. **Independence of n random variables.** We extend the discussion in the previous subsection to n random variables.

Definition 7.3. (Independence of n random variables) *Discrete random variables X_1, X_2, ..., X_n defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are called (jointly) independent, if for any $x_1, x_2, ..., x_n \in \mathbb{R}$ we have*

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i). \tag{7.8}$$

Clearly, both sides of (7.8) become zero unless $x_i \in \mathrm{Im} X_i$ for all $i = 1, 2, ..., n$.

☐ Exercise 7.7.1. *Prove that condition (7.8) for discrete random variables is equivalent to the following: for any non-empty index set $I \subset \{1, 2, ..., n\}$ we have*

$$\mathbb{P}\left(\bigcap_{i \in I} \{X_i = x_i\}\right) = \prod_{i \in I} \mathbb{P}(X_i = x_i)$$

for any $x_i \in \mathbb{R}$.

We now formulate the analogues of the main results proved for the case of two discrete random variables for for n variables.

Theorem 7.8. (cf. Theorem 7.4) *Discrete random variables $X_1, ..., X_n$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ are independent if and only if there exist functions $f_1, ...f_n : \mathbb{R} \to \mathbb{R}$ such that*

$$p_{X_1,...X_n}(x_1, ..., x_n) = f_1(x_1) \cdot ... \cdot f_n(x_n), \quad \forall x_1, ..., x_n \in \mathbb{R}.$$

Theorem 7.9. (cf. Theorem 7.5) *Let $X_1, ..., X_n$ be discrete random variables that are jointly independent, and assume $\mathbb{E}(X_1), ..., \mathbb{E}(X_n)$ exist. Then $\mathbb{E}(X_1 \cdot ... \cdot X_n)$ exists and*

$$\mathbb{E}(X_1 \cdot ... \cdot X_n) = \mathbb{E}(X_1) \cdot ... \cdot \mathbb{E}(X_n).$$

Corollary 7.10. (cf. Corollary 7.6) *Let $X_1, ..., X_n$ be discrete random variables that are pairwise independent. Then*

$$\mathrm{var}(X_1 + ... + X_n) = \mathrm{var}(X_1) + ... + \mathrm{var}(X_n).$$

Theorem 7.11. (cf. Theorem 7.7) *Discrete random variables $X_1, ..., X_n$ are jointly independent if and only if for any functions $f_1, ..., f_n : \mathbb{R} \to \mathbb{R}$ we have*

$$\mathbb{E}(f_1(X_1) \cdot ... \cdot f_n(X_n)) = \mathbb{E}(f_1(X_1)) \cdot ... \cdot \mathbb{E}(f_n(X_n)),$$

whenever the expectations on the right-hand side exist.

Proofs of the formulated results follow along the same lines of their counterparts for two random variables.

☐ Exercise 7.11.1. *Prove Theorems 7.8, 7.9, 7.11, and Corollary 7.10.*

We also include a new result that we will prove for the case of n random variables.

Theorem 7.12. (Independence is preserved under compositions) *Let $X_1, X_2, ..., X_n$ be independent random variables, and let $f_1, f_2, ..., f_n$ be any real-valued functions defined on $\mathbb{R}$. Then $f_1(X_1), f_2(X_2), ..., f_n(X_n)$ are independent.*

Proof. Fix any $y_1, ..., y_n \in \mathbb{R}$, we need to show (see Exercise 7.7.1) that

$$\mathbb{P}\left(f_1(X_1) = y_1, ..., f_n(X_n) = y_n\right) = \prod_{i=1}^{n} \mathbb{P}\left(f_i(X_i) = y_i\right).$$

To this end define a function $g : \mathbb{R}^n \to \mathbb{R}$ as follows

$$g(x_1, ..., x_n) = \begin{cases} 1, & \text{if } f_i(x_i) = y_i, \text{ for all } i = 1, 2, ..., n, \\ 0, & \text{otherwise,} \end{cases}$$

and for $i = 1, 2, ..., n$ define $g_i : \mathbb{R} \to \mathbb{R}$ as

$$g_i(x) = \begin{cases} 1, & \text{if } f_i(x) = y_i, \\ 0, & \text{otherwise.} \end{cases}$$

We have

$$\begin{aligned}
&\mathbb{P}\left(f_1(X_1) = y_1, ..., f_n(X_n) = y_n\right) \text{ (by Property 6.8 part 1)} \\
&= \mathbb{E}\mathbb{I}_{\{f_1(X_1)=y_1,...,f_n(X_n)=y_n\}} \text{ (by definition of } g) \\
&= \mathbb{E}g(X_1, ..., X_n) \text{ (by Theorem 7.3)} \\
&= \sum_{x_1 \in \mathrm{Im}X_1} ... \sum_{x_n \in \mathrm{Im}X_n} g(x_1, ..., x_n) p_{X_1,...,X_n}(x_1, ..., x_n) \\
&= \sum_{\substack{x_1 \in \mathrm{Im}X_1 \\ f(x_1)=y_1}} ... \sum_{\substack{x_n \in \mathrm{Im}X_n \\ f(x_n)=y_n}} \mathbb{P}(X_1 = x_1, ..., X_n = x_n) \text{ (by independence of } X_i) \\
&= \prod_{i=1}^{n} \sum_{\substack{x_i \in \mathrm{Im}X_i \\ f(x_i)=y_i}} \mathbb{P}(X_i = x_i) \text{ (by definition of } g_i) \\
&= \prod_{i=1}^{n} \sum_{x \in \mathrm{Im}X_i} g_i(x)\mathbb{P}(X_i = x) \text{ (by Theorem 6.4)} \\
&= \prod_{i=1}^{n} \mathbb{E}\, g_i(X_i) \text{ (by definition of } g_i) \\
&\prod_{i=1}^{n} \mathbb{E}\mathbb{I}_{f(X_i)=y_i} = \prod_{i=1}^{n} \mathbb{P}(f(X_i) = y_i),
\end{aligned}$$

and the proof is complete. □

The morale of Theorem 7.12 is that given a sequence of independent variables any transformation of a single entry in the sequence will not alter the independence. For example, if X, Y, Z are independent random variables, then so are

$$X^2,\ e^Y,\ (1 + |Z|^{2021})^{-1/2},$$

a fact which might not seem entirely trivial on this particular example.

□ Exercise 7.12.1. `(Independent copies of a discrete random variable)` *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X : \Omega \to \mathbb{R}$ be a discrete random variable. Using the ideas of section 4.3.1 prove that for any $n \in \mathbb{N}$ there exists a probability space $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and discrete random variables $X_1, X_2, ..., X_n$ defined on that space which are independent and all have the same distribution (pmf) as X.*

→ **Problem 7.12.1.** *Let $v_1, v_2, ..., v_n \in \mathbb{R}^d$ be unit vectors, i.e. $||v_i|| = 1$ for all $1 \le i \le n$. Prove that there exists a sequence of signs $\varepsilon_i \in \{-1, 1\}$ such that the vector $\sum_{i=1}^n \varepsilon_i v_i$ has length at most (at least) $\sqrt{n}$.*

Hint: choose the signs uniformly at random and compute the expectation of the length (Euclidean norm) of $\sum_i \varepsilon_i v_i$.

8. Inequalities

In this section we state and prove several important inequalities, and discuss the concept of information entropy.

8.1. **Jensen's inequality.** Recall the definition of a convex function. A function $f : \mathbb{R}^d \to \mathbb{R}$ is called **convex** if for any $x_1, x_2 \in \mathbb{R}^d$ and $0 \le \lambda \le 1$ one has

$$f\left((1-\lambda)x_1 + \lambda x_2\right) \le (1-\lambda)f(x_1) + \lambda f(x_2). \tag{8.1}$$

A function is called **strictly** convex if the inequality (8.1) is strict whenever $x \ne y$ and $0 < \lambda < 1$.

A function is called **concave** if $-f$ is convex.

Examples of convex functions are $f(x) = |x|$, $f(x) = x^2$, $f(x) = e^x$, etc.

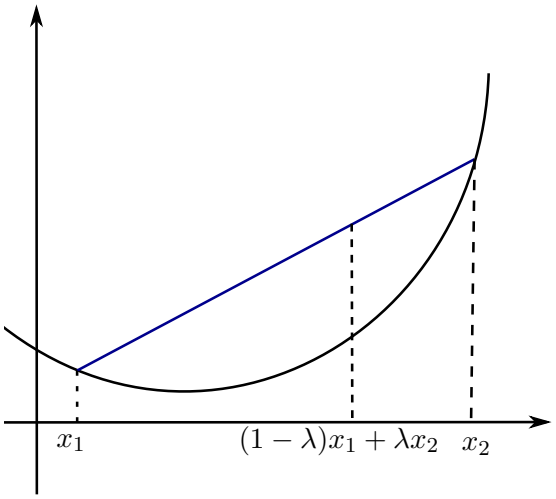

FIGURE 7. An example of a convex function. For any x_1, x_2 the graph of f lies below the chord joining points $(x_1, f(x_1))$ and $(x_2, f(x_2))$ on the graph. More precisely, the value of f at $(1-\lambda)x_1 + \lambda x_2$, for any $\lambda \in [0, 1]$, is bounded above by the value of the chord at the same point.

Theorem 8.1. (`Jensen's inequality`) *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Then for any $x_1, ..., x_n \in \mathbb{R}^d$ and any probability distribution $p_1, ..., p_n$ i.e. $0 \le p_i \le 1$ where $i = 1, 2, ..., n$ and $p_1 + ... + p_n = 1$, one has*

$$f(p_1 x_1 + ... + p_n x_n) \le p_1 f(x_1) + ... + p_n f(x_n). \tag{8.2}$$

Proof. The proof is by induction on the number of points. When $n = 2$ the inequality (8.2) coincides with the definition of the convex function (8.1). Now assume (8.2) holds

for $n-1$ points. To prove the case of n points notice first that the inequality is trivially true if $p_1 = 1$. Otherwise, if $p_1 < 1$, then $q := p_2 + ... + p_n > 0$, and hence

$$\begin{aligned} f(p_1x_1 + ... + p_nx_n) &= f\left(p_1x_1 + q\left(\frac{p_2}{q}x_2 + ... + \frac{p_n}{q}x_n\right)\right) \text{ (by (8.1))} \\ &\leq p_1f(x_1) + qf\left(\frac{p_2}{q}x_2 + ... + \frac{p_n}{q}x_n\right) \text{ (by induction)} \\ &\leq p_1f(x_1) + q\left(\frac{p_2}{q}f(x_2) + ... + \frac{p_n}{q}f(x_n)\right) = \\ &\leq p_1f(x_1) + p_2f(x_2) + ... + p_nf(x_n), \end{aligned}$$

completing the inductive step and the proof of the inequality. □

An important corollary of the Jensen's inequality is the following.

Theorem 8.2. (`Jensen's inequality for expectation`) *Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function, and let X be a discrete random variable. If $\mathbb{E}X$ and $\mathbb{E}f(X)$ both exist, then*

$$f(\mathbb{E}X) \leq \mathbb{E}f(X). \tag{8.3}$$

Proof. We split the proof in two parts, namely when X takes only finitely many values, and the general case of discrete random variables.

Assume first that X takes finitely many values with the pmf $\mathbb{P}(X = x_i) = p_i$ where $0 \leq p_i \leq 1$ for $i = 1, 2, ..., n$ and $p_1 + ... + p_n = 1$. Using the definition of expectation we get

$$\begin{aligned} f(\mathbb{E}X) = f\left(\sum_{i=1}^{n} x_ip_i\right) &\text{ (by Jensen's inequality)} \\ &\leq \sum_{i=1}^{n} p_if(x_i) = \mathbb{E}f(X), \end{aligned}$$

where the last equality is due to (6.4). This settles the case of random variables with finitely many values.

To prove the general case assume X has distribution $\mathbb{P}(X = x_i) = p_i$ with $0 \leq p_i \leq 1$ and $\sum\limits_{i=1}^{\infty} p_i = 1$. We will use the fact that convex functions are continuous. In view of continuity of f at $\mathbb{E}X$ for any $\varepsilon > 0$ there exists $N(\varepsilon) \in \mathbb{N}$ such that for all $N \geq N(\varepsilon)$ we have

$$f(\mathbb{E}X) \leq \varepsilon + f\left(\sum_{i=1}^{N} x_i\tilde{p_i}\right), \tag{8.4}$$

where $\tilde{p_i} = \frac{p_i}{\sum\limits_{i=1}^{N} p_i}$. Indeed, by definition $\mathbb{E}X = \sum\limits_{i=1}^{\infty} x_ip_i$ and the series is absolutely convergent. Hence $\sum\limits_{i=1}^{N} x_i\tilde{p_i} = \frac{1}{\sum\limits_{i=1}^{N} p_i} \sum\limits_{i=1}^{N} x_ip_i \to \mathbb{E}X$ as $N \to \infty$ since $\{p_i\}$ is a probability distribution and sums to 1. Getting back to (8.4) we use the fact that $\{\tilde{p_i}\}_{i=1}^{N}$ is a finite distribution and apply Jensen's inequality on the right-hand-side arriving at

$$f(\mathbb{E}X) \leq \varepsilon + \sum_{i=1}^{N} f(x_i)\tilde{p_i} = \varepsilon + \frac{1}{\sum\limits_{i=1}^{N} p_i} \sum_{i=1}^{N} f(x_i)p_i.$$

Now both sums in the above inequality are convergent. Passing to limit as $N \to \infty$ and using Theorem (6.4) we get

$$f(\mathbb{E}X) \leq \varepsilon + \mathbb{E}f(X).$$

Since $\varepsilon > 0$ is arbitrary the last inequality completes the proof of the general case of discrete random variable. □

Geometric proof of Jensen's inequality. We also include a geometric proof of inequality (8.3). A point $v \in \mathbb{R}$ is called a **sub-derivative** of f at $x_0 \in \mathbb{R}$ if $f(x) \geq f(x_0) + v(x - x_0)$ for all $x \in \mathbb{R}$. The set of all sub-derivatives at x_0 is denoted by $\partial f(x_0)$ and is called the **sub-differential**[22] of f at x_0. It is a well-known fact that for convex functions the sub-differential at any point is a non-empty closed convex set.

Let v by any sub-derivative of f at $\mathbb{E}X$. By definition we have

$$f(x) \geq f(\mathbb{E}X) + v(x - \mathbb{E}X) \text{ for all } x \in \mathbb{R}.$$

Applying the inequality above for $x = x_i$, multiplying both sides by p_i and summing over $i = 1, 2, ...$ we get

$$\sum_{i=1}^{\infty} f(x_i)p_i \geq f(\mathbb{E}X) + v \cdot \left(\sum_{i=1}^{\infty} p_i x_i - \mathbb{E}X \right).$$

The left-hand side of this inequality equals $\mathbb{E}f(X)$ by Theorem 6.4, and the right-hand side is precisely $f(\mathbb{E}X)$ as the sum in the brackets equals $\mathbb{E}X$. The proof of inequality (8.3) is complete. □

For example, applying Jensen's inequality for specific convex functions f we get

$$e^{\mathbb{E}X} \leq \mathbb{E}e^X \quad \text{by taking } f(x) = e^x$$

and

$$(\mathbb{E}X)^{2n} \leq \mathbb{E}X^{2n} \quad \text{by taking } f(x) = x^{2n} \text{ where } n \in \mathbb{N},$$

whenever the expectations above exist.

8.2. **Arithmetic mean - Geometric mean inequality.**

Theorem 8.3. `(AM-GM inequality)` *Let $x_1, x_2, ..., x_n \in \mathbb{R}$ be a set of non-negative numbers. Then*

$$(x_1 x_2 \cdot ... \cdot x_n)^{1/n} \leq \frac{x_1 + x_2 + ... + x_n}{n}. \tag{8.5}$$

Proof. We may assume that $x_i > 0$ for all $1 \leq i \leq n$ as otherwise the inequality is trivially true. Since $x \mapsto \log x$ is increasing in x proving (8.5) is equivalent to proving the inequality with log applied to both sides which reads

$$\sum_{i=1}^{n} \frac{1}{n} \log x_i \leq \log \left(\sum_{i=1}^{n} \frac{1}{n} x_i \right).$$

Now the above inequality is a direct consequence of the Jensen's inequality and convexity of $f(x) = -\log(x)$, $x > 0$. The proof is complete. □

[22] In geometric terms, the elements of the sub-differential are precisely the slopes for which the graph of f lies above the line through $(x_0, f(x_0))$ with the given slope. The sub-differential is defined for functions $f : \mathbb{R}^d \to \mathbb{R}$ as well where $v \in \mathbb{R}^d$ is a sub-derivative if $f(x) \geq f(x_0) + \langle v, x - x_0 \rangle$ for all $x \in \mathbb{R}^d$, where $\langle \cdot, \cdot \rangle$ is the standard inner product of $\mathbb{R}^d$. Moreover, a convex function f is differentiable at x_0 if and only if the sub-differential is a single element set; in which case the only sub-derivative is the gradient of f at x_0.

8.3. **Cauchy-Schwarz inequality.**

Theorem 8.4. (`Cauchy-Schwarz inequality`) *Let X and Y be two discrete random variables such that $\mathbb{E}(X^2), \mathbb{E}(Y^2) < \infty$. Then*

$$(\mathbb{E}(XY))^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2). \tag{8.6}$$

The inequality becomes equality if and only if X and Y are linearly dependent with probability one, i.e. there is some $t \in \mathbb{R}$ such that $\mathbb{P}(X = tY) = 1$.

Proof. Since $2|XY| \leq X^2 + Y^2$ we have that XY has finite expectation. Define $W_t := tX + Y$, where $t \in \mathbb{R}$. Then

$$0 \leq \mathbb{E}(W_t^2) = t^2\mathbb{E}(X^2) + 2t\,\mathbb{E}(XY) + \mathbb{E}(Y^2).$$

Notice that we have a quadratic polynomial in t which is always non-negative. Hence its discriminant must be non-positive, and we get that

$$4\left(\mathbb{E}(XY)\right)^2 - 4\mathbb{E}(X^2)\mathbb{E}(Y^2) \leq 0,$$

which is equivalent to (8.6).

To see when (8.6) becomes an equality, observe that $\mathbb{E}(W_t^2) = 0$ for some $t \in \mathbb{R}$ if and only if the discriminant of the above quadratic polynomial becomes 0 which is precisely the case of equality in (8.6). But if $\mathbb{E}(W_t)^2 = 0$, then $\mathbb{P}(W_t = 0) = 1$ which is exactly the case of linear dependence between X and Y. The proof is complete. □

□ Exercise 8.4.1. *Prove that $(\mathbb{E}|XY|)^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$ for two discrete random variables X and Y.*

8.4. **Covariance and correlation.** In this section we define *covariance* and *correlation* as measure of joint variability of two random variables.

Definition 8.1. (`Covariance`) *Given two random variables X and Y the covariance between them is defined as*

$$\operatorname{cov}(X, Y) = \mathbb{E}\left((X - \mathbb{E}X)(Y - \mathbb{E}Y)\right), \tag{8.7}$$

whenever the expectations on the right-hand side exist.

Note that the formula for covariance can be simplified to

$$\operatorname{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

in view of the linearity of expectation. Hence the formula for the variance of sums of random variables proved in Property 6.5 (c), in terms of the covariance, becomes

$$\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y) + 2\operatorname{cov}(X, Y).$$

Similarly, for the variance of the sum of n random variables $X_1, X_2, ..., X_n$ we have

$$\operatorname{var}(X_1 + ... + X_n) = \operatorname{var}(X_1) + ... + \operatorname{var}(X_n) + 2\sum_{1 \leq i < j \leq n} \operatorname{cov}(X_i, X_j). \tag{8.8}$$

□ Exercise 8.4.2. *Verify (8.8) following the proof of Property 6.5 (c).*

We next state some basic properties of covariance, proofs of which follow directly from definitions.

Property 8.5. (`Basic properties of covariance`) *Let X and Y be discrete random variables and $a \in \mathbb{R}$ be a constant. Whenever the covariances defined below exist, we have*

1. $\operatorname{cov}(X, a) = 0$,
2. $\operatorname{cov}(X + a, Y) = \operatorname{cov}(X, Y)$,
3. $\operatorname{cov}(aX, Y) = a \operatorname{cov}(X, Y)$,
4. $\operatorname{cov}(X, Y) = \operatorname{cov}(Y, X)$
5. $\operatorname{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$,
6. $\operatorname{cov}(X, Y) = 0$ *if* X *and* Y *are independent,*
7. $\operatorname{cov}(X + Z, Y) = \operatorname{cov}(X, Y) + \operatorname{cov}(Z, Y)$,
8. $\operatorname{cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \operatorname{cov}(X_i, Y_j)$ *for any discrete random variables* X_i, Y_j, *and constants* $a_i, b_j \in \mathbb{R}$, *whenever the covariances exist*[23].

☐ Exercise 8.5.1. *Prove Property 8.5.*

☐ Exercise 8.5.2. *What is* $\operatorname{cov}(X, X)$ *for a given random variable* X*?*

Example 8.5.1. (Zero covariance does not imply independence) *The aim of this example is to show that the converse of Property 8.5 (6) is not true in general. To this end, consider a random variable* X *with probability mass function defined as*

$$\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = \frac{1}{4} \text{ and } \mathbb{P}(X = 0) = \frac{1}{2}.$$

Also, define a random variable Y *as*

$$Y = \begin{cases} 1, & \text{if } X = 0, \\ 0, & \text{if } X \neq 0. \end{cases}$$

Clearly $\mathbb{E}(X) = 0$ *and* $XY = 0$, *hence* $\operatorname{cov}(X, Y) = 0$. *However,*

$$\mathbb{P}(X = 0, Y = 1) = \frac{1}{2} \neq \mathbb{P}(X = 0)\mathbb{P}(Y = 1),$$

implying that X *and* Y *are not independent.*

☐ Exercise 8.5.3. *Construct a discrete random variable* X *such that* $\mathbb{E}X = 0$, $0 < \mathbb{E}|X|^3 < \infty$ *and* $\operatorname{cov}(X, X^2) = 0$.

Note: This exercise gives another example of dependent random variables with vanishing covariance. As an example of such X *one may take a random variable whose probability mass function (pmf) is symmetric around zero and choose the values of the* X *and pmf so that to guarantee* $0 < \mathbb{E}|X|^3 < \infty$ *(by Theorem 6.1 there exists a random variable with given pmf.)*

☐ Exercise 8.5.4. *A fair coin is tossed 3 times. Let* X *be the number of Heads in the first two tosses, and let* Y *be the number of Heads in the last two tosses (note the overlap). Compute* $\operatorname{cov}(X, Y)$.

Definition 8.2. (Correlation) *For two random variables* X *and* Y *such that* $\operatorname{var}(X)$, $\operatorname{var}(Y) > 0$, *the correlation (or correlation coefficient) between them is defined as*

$$\operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}. \tag{8.9}$$

[23] This is the **bilinearity** of the covariance.

Theorem 8.6. *For any two random variables X an Y for which the correlation is defined, we have*

$$|\mathrm{corr}(X,Y)| \leq 1. \tag{8.10}$$

Moreover, $\mathrm{corr}(X,Y) = \epsilon$ *with* $\epsilon \in \{-1,1\}$ *if and only if* $Y = aX + b$ *for some* $a, b \in \mathbb{R}$ *where* $a \neq 0$ *and* $\frac{a}{|a|} = \epsilon$.

Proof. Applying Cauchy-Schwarz inequality (Theorem 8.4) on formula (8.7) of covariance we get

$$|\mathrm{cov}(X,Y)|^2 \leq \left(\mathbb{E}(X - \mathbb{E}X)^2\right)\left(\mathbb{E}(Y - \mathbb{E}Y)^2\right) = \mathrm{var}(X)\mathrm{var}(Y).$$

This inequality, combined with (8.9) completes the proof of (8.10).

We now establish the second part of the Theorem. If $Y = aX + b$, where $a \neq 0$, then $\mathrm{corr}(X,Y) = \frac{a}{|a|}$ by a direct computation (or using Property 8.5). To see the other side of the equivalence, notice that from the proof of (8.10) we have that $|\mathrm{corr}(X,Y)| = 1$ if and only if there is an equality in the Cauchy-Schwarz inequality for $X - \mathbb{E}X$ and $Y - \mathbb{E}Y$. Hence, by Theorem 8.4 we obtain that $X = aY + b$ and the sign of a must coincide with the sign of $\mathrm{corr}(X,Y)$. □

Remark 8.7. *From Theorem 8.6 we see that the correlation is the "normalized" version of the covariance taking values in* $[-1,1]$. *The second part of the Theorem hints toward the direction that covariance (correlation) measures the extent to which variables X and Y are related linearly. For instance, if* $\mathrm{corr}(X,Y) = \pm 1$, *then there is a linear dependence between X and Y.*

8.5. **Information entropy.** How can one quantify the *level of surprise* or *information* that a certain event bears? For example, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $A \in \mathcal{F}$ be an event with probability $0 \leq p \leq 1$; how surprising is A? We will measure the level of the surprise as a function of p, and denote that function by I (here I stands for "information"). We now outline a few requirements on I that convey the intuitive meaning of surprise, namely

- if $\mathbb{P}(A) = 1$, then the event A is certain to happen and hence there is no surprise of it happening, thus we require $I(1) = 0$,
- the more unlikely the event is (i.e. if p is small), the more surprising its occurrence is, hence $p \mapsto I(p)$ is required to be decreasing in p,
- for two independent events A and B the occurrence of $A \cap B$ should sum up the surprise (or information content) coming from A and B, thus we require $I(\mathbb{P}(A \cap B)) = I(\mathbb{P}(A)) + I(\mathbb{P}(B))$.

It turns out these conditions on I characterize it.

Theorem 8.8. `(Characterization of information function)` Let $I : (0,1] \to \mathbb{R}_+$ be a function satisfying

$$I(pq) = I(p) + I(q), \tag{8.11}$$

for any $p, q \in (0,1]$. If I is continuous then $I(p) = c \log p$ for some $c \leq 0$.

Proof. Make a change of variables in I by setting $p = e^{-x}$ where $p \in (0,1]$. Let $f(x) = I(e^{-x})$ for $x \in [0,\infty)$. Clearly (8.11) transforms to[24]

$$f(x+y) = f(x) + f(y) \text{ for } x, y \in \mathbb{R}_+. \tag{8.12}$$

[24]Equation 8.12 is called Cauchy's functional equation.

For any $p, q \in \mathbb{N}$ we have $\frac{p}{q} = \underbrace{\frac{1}{q} + ... + \frac{1}{q}}_{p \text{ times}}$, hence applying (8.12) we get $f\left(\frac{p}{q}\right) = pf\left(\frac{1}{q}\right)$. Similarly, writing $1 = \underbrace{\frac{1}{q} + ... + \frac{1}{q}}_{q \text{ times}}$ from (8.12) we arrive at $f(1) = qf\left(\frac{1}{q}\right)$, and hence

$$f\left(\frac{p}{q}\right) = \frac{p}{q} f(1) \text{ for all } p, q \in \mathbb{N}. \tag{8.13}$$

Now the continuity of f coupled with (8.13) implies that $f(x) = xf(1)$ for all $x \geq 0$. Getting back to relation between f and I we see that for $x \in (0, 1]$ we have

$$I(e^{-\log \frac{1}{x}}) = I(x) = f\left(\log \frac{1}{x}\right) = \log \frac{1}{x} I(e^{-1}) = -c \log x,$$

where $c = I(e^{-1})$ is a constant. The proof is complete. □

Proof of Theorem 8.8 when I is C^2. We include another proof that the information function is a multiple of logarithm, when I has continuous second derivative.

Using the fact that I is twice differentiable we take the derivative in (8.11) with respect to p arriving at

$$qI'(pq) = I'(p).$$

Differentiating both sides of the last equality with respect to q we get

$$I'(pq) + pqI''(pq) = 0.$$

With a change of variable $z := pq$, and setting $g(z) = I'(z)$, the last equation becomes $g(z) + zg'(z) = 0$. Thus $\frac{dg}{dz} = -\frac{g}{z}$ which is a differential equation with *separable variables*. Solving it we obtain $g(z) = c\frac{1}{z}$ for some constant $c > 0$.

Recall that $g(z) = I'(z)$ hence from the formula of g we see that $I(z) = c \log z + c_1$ where c and c_1 are constants. To determine the range of these constants recall that I must satisfy (8.11). In that formula take $q = 1$ which leads to $I(p) = I(p) + I(1)$ hence $I(1) = 0$. Hence, setting $z = 1$ we get $0 = I(1) = c \log 1 + c_1$, consequently $c_1 = 0$. This leads to $I(z) = c \log z$ and the proof is complete. □

→ **Problem 8.8.1.** *Let I be as in Theorem 8.8. Prove that the requirement of continuity of I can be relaxed to measurability of I. More precisely, show that if I is a Lebesgue measurable function satisfying (8.11), then $I(p) = c \log p$ for some $c \leq 0$.*

→ **Problem 8.8.2.** *Using the concept of Hamel basis and considering $\mathbb{R}$ as a vector space over rationals $\mathbb{Q}$, show that there exists a non-linear solution to (8.12). By the previous problem, such solutions cannot be measurable. Thus, the condition of measurability cannot be eliminated in the characterization of the information function I.*

We choose the constant c in the definition of the information function to get $I(p) = -\log_2 p$ where $p \in (0, 1]$. Now, if X is a random variable with pmf p_X given by $p_X(x_i) = \mathbb{P}(X = x_i) = p_i$, with $i = 1, 2, ..., n$, then the average level of information

(surprise) obtained on learning X equals[25]

$$H(X) := \mathbb{E}I(p_X) = -\sum_{i=1}^{n} p_i \log_2 p_i.$$

The value $H(X)$ is called **information entropy** (Shannon entropy) of X, which among other things measures the level of uncertainty in X. In this light it is plausible to expect that the *highest level of uncertainty* is obtained for the *uniform distribution*. Indeed, by concavity of $x \mapsto \log x$ on $x > 0$, and Jensen's inequality we have

$$H(X) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i} \leq \log_2 \left(\sum_{i=1}^{n} p_i \frac{1}{p_i} \right) = \log_2 n.$$

But $\log_2 n$ is the entropy of the uniform distribution on n elements and the Jensen's inequality is strict (for strictly convex functions) unless $p_i = const$. Hence the uniform distribution has the maximal entropy.

→ Problem 8.8.3. *There are* 12 *coins of identical shape and size, except one of them is fake. It is known that the fake coin has different weight than a real coin, but it is not known whether the fake is heavier or lighter than a real one. You have a balancing weight scale, and can weight any two subsets of the* 12 *coins with it. When placing two sets of coins on the two sides of the scale (either set of coins can be empty) there are three possible outcomes: the left and right sides have equal weights, the left side is heavier, the right side is heavier.*

Using the scale, find the fake coin and whether it is lighter or heavier in as few weightings as possible.

Hint: If we enumerate the coins from $1, 2, ..., 12$*, then there are* 24 *scenarios in total. Namely, the fake coin is coin number* i *for some* $1 \leq i \leq 12$ (12 *possibilities) and it is either lighter or heavier than a real one* (2 *possibilities). Think about a weighting step as a process of gaining information about these scenarios and when organize a weighting step think about splitting the aforementioned possibilities as evenly as possible.*

Note: If you can do $W \geq 1$ *weightings, what is the maximal number of coins, of which one is fake, that can be checked as in the original problem?*

9. Weak Law of Large Numbers

In this section we discuss Markov's and Chebyshev's inequalities. We then prove a version of Weak Law of Large Numbers (WLLN) and as an example, conclude with probabilistic proofs of few results, including Weierstrass approximation theorem, where ideas from WLLN can be used.

9.1. Markov and Chebyshev.

Theorem 9.1. (`Markov inequality`) *Let X be a discrete random variable and let $a > 0$ be fixed. Then*

$$\mathbb{P}(|X| \geq a) \leq \frac{1}{a}\mathbb{E}|X|.$$

Proof. We will assume that $\mathbb{E}|X| < \infty$, as otherwise the inequality is trivial. By the linearity of expectation and Property 6.8 (1) of indicators we have

$$\begin{aligned}\mathbb{E}(|X|) = \mathbb{E}(|X|\mathbb{I}_{|X|\geq a}) + \mathbb{E}(|X|\mathbb{I}_{|X|<a}) \geq \\ \mathbb{E}(|X|\mathbb{I}_{|X|\geq a}) \geq a\mathbb{E}(\mathbb{I}_{|X|\geq a}) = a\mathbb{P}(|X| \geq a).\end{aligned}$$

[25] In the formula for $H(X)$ we use the convention that $0 \cdot \log 0 = 0$ based on the fact that $p \log p \to 0$ when $p \to 0+$.

Rearranging the last inequality completes the proof. □

Theorem 9.2. (Chebyshev inequality) *Let X be a discrete random variable and let $\varepsilon > 0$ be fixed. Then*

$$\mathbb{P}(|X| \geq \varepsilon) \leq \frac{1}{\varepsilon^2}\mathbb{E}|X|^2.$$

Proof. In view of the monotonicity of $x \mapsto x^2$ we have

$$\mathbb{P}(|X| \geq \varepsilon) = \mathbb{P}(|X|^2 \geq \varepsilon^2) \leq \frac{1}{\varepsilon^2}\mathbb{E}|X|^2,$$

where the last inequality follows by Markov's inequality. □

Corollary 9.3. *Let X and $\varepsilon > 0$ be as in Theorem 9.2. Assume $\mathbb{E}X$ exists, then*

$$\mathbb{P}(|X - \mathbb{E}X| \geq \varepsilon) \leq \frac{1}{\varepsilon^2}\mathrm{var}(X). \tag{9.1}$$

Proof. The proof follows by applying Chebishev's inequality to $X - \mathbb{E}X$. □

9.2. **WLLN - easy version.**

Theorem 9.4. (Weak Law of Large Numbers - easy version) *Let $X_1, X_2, ..., X_n, ...$ be a sequence of independent and identically distributed (i.i.d.) discrete random variables with mean μ and variance σ^2. Define*

$$S_n = X_1 + ... + X_n, \quad n = 1, 2, ... \ .$$

Then, for any $\varepsilon > 0$ we have

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \to 0, \ \text{as } n \to \infty.$$

Proof. Notice that $\mathbb{E}(S_n) = n\mu$, hence applying inequality (9.1) we get

$$\begin{aligned}\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) &\leq \frac{1}{\varepsilon^2}\mathrm{var}\left(\frac{S_n}{n}\right) = \frac{1}{\varepsilon^2}\frac{1}{n^2}\mathrm{var}(S_n) \text{ (by independence)} \\ &= \frac{1}{\varepsilon^2}\frac{1}{n^2}n\sigma^2 = \frac{1}{\varepsilon^2}\frac{1}{n}\sigma^2 \to 0, \ n \to \infty,\end{aligned}$$

and the proof is complete. □

The assertion of the WLLN holds under different forms and conditions. An example demonstrating this is provided in the following exercise.

□ Exercise 9.4.1. *Let $X_1, X_2, ...$ be any sequence of identically distributed random variables (not necessarily independent) with finite variance. Assume there exists a constant $0 \leq \alpha < 1$ such that $|\mathrm{cov}(X_i, X_j)| \leq \alpha^{|i-j|}$ for any $i \neq j$. Prove that the claim of WLLN holds for the sequence $\{X_i\}$.*

Hint: Use Chebyshev's inequality. Then use (8.8) and apply the power bound on covariances.

Note: The power decay on covariance can be replaced by a more general condition. Namely, prove that if $|\mathrm{cov}(X_i, X_j)| \to 0$ when $|i - j| \to +\infty$ then the convergence in WLLN holds.

The next exercise is a prelude to the probabilistic proof of the Weierstrass approximation theorem. It can be useful to go through the exercise first before moving on.

□ Exercise 9.4.2. *Consider a sequence of Bernoulli trials, i.e. a sequence of random variables $X_1, X_2, ..., X_n, ...$ that are independent and*

$$\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = 0) = \frac{1}{2}, \quad \text{for all } i = 1, 2, ... \ .$$

Prove that for any $f \in C[0,1]$ one has

$$\mathbb{E}f\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) \to f\left(\frac{1}{2}\right), \ n \to \infty.$$

9.3. **Probabilistic proofs.** In this section we will prove several results, which are not necessarily probabilistic in nature, but can be proved using the ideas around the weak law of large numbers (WLLN).

9.3.1. **Weierstrass approximation theorem**. We will recover a classical result due to Weierstrass on approximation of continuous functions with polynomials with ideas used in this section.

Theorem 9.5. (`Weierstrass approximation theorem`) *Let $f \in C[0,1]$. Then, for any $\varepsilon > 0$ there exists a polynomial $p(x)$ such that $|f(x) - p(x)| < \varepsilon$ for all $x \in [0,1]$.*

Proof. Continuous function on a closed interval is uniformly continuous, hence for any $m \in \mathbb{N}$ there exists $\delta_m > 0$ such that

$$|f(x) - f(y)| < \frac{1}{m}, \quad \text{if } x, y \ \in [0,1] \text{ and } |x - y| < \delta_m. \tag{9.2}$$

For $n \in \mathbb{N}$ and $0 \leq k \leq n$ consider the following polynomial (called Bernstein polynomial)

$$p_{n,k}(x) = \binom{n}{k} x^k (1-x)^{n-k}, \qquad x \in [0,1],$$

and define

$$p_n(x) := \sum_{k=0}^{n} f\left(\frac{k}{n}\right) p_{n,k}(x).$$

Clearly p_n is a polynomial for any $n \in \mathbb{N}$. The aim is to show that the sequence p_n approximates f. Consider a sequence of Bernoulli trials $X_1, X_2, ..., X_n$, where $X_i \sim B(1,x)$, i.e. $\mathbb{P}(X_i = 1) = x$ and $\mathbb{P}(X_i = 0) = 1 - x$, and the random variables are independent. Define $S_n := X_1 + ... + X_n$ and notice that

$$\mathbb{E}f\left(\frac{S_n}{n}\right) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) \mathbb{P}(S_n = k) = p_n(x), \tag{9.3}$$

where the first equality is due to Theorem 6.3 and the second one follows by definition of Bernoulli trials[26]. For $m \in \mathbb{N}$ define

$$A_{m,n,x} := \left\{\omega \in \Omega : \ \left| f\left(\frac{S_n(\omega)}{n}\right) - f(x)\right| \geq \frac{1}{m}\right\}.$$

[26]The intuition behind the argument is the following. By WLLN we know that S_n/n concentrates around x. Then, the continuity of f implies that $f(S_n/n)$ should concentrate around $f(x)$ (cf. Exercise 9.4.2). But $\mathbb{E}f(S_n/n)$ is a polynomial, and we get a candidate polynomial that can approximate f.

Let also $\overline{S}_n := \frac{S_n}{n}$, then using (9.3) we get

$$
\begin{aligned}
(9.4)\quad |p_n(x) - f(x)| &= |\mathbb{E}f(\overline{S}_n) - f(x)| \le \mathbb{E}|f(\overline{S}_n) - f(x)| \\
&= \mathbb{E}\left(|f(\overline{S}_n) - f(x)|\mathbb{I}_{A^c_{m,n,x}}\right) + \mathbb{E}\left(|f(\overline{S}_n) - f(x)|\mathbb{I}_{A_{m,n,x}}\right) \\
&\le \frac{1}{m} + 2\left(\max_{x\in[0,1]}|f(x)|\right)\mathbb{P}(A_{m,n,x}).
\end{aligned}
$$

To estimate $\mathbb{P}(A_{m,n,x})$, observe that by (9.2) we have

$$A_{m,n,x} \subset \{\omega \in \Omega :\ |\overline{S}_n(\omega) - x| \ge \delta_m\},$$

hence by monotonicity of probability measure we get

$$
\begin{aligned}
\mathbb{P}(A_{m,n,x}) &\le \mathbb{P}(\{\omega \in \Omega :\ |\overline{S}_n(\omega) - x| \ge \delta_m\}) \text{ (by Corollary (9.3))} \\
&\le \frac{1}{\delta_m^2}\mathrm{var}(\overline{S}_n) = \frac{1}{\delta_m^2}\frac{1}{n^2}nx(1-x) \le \frac{1}{\delta_m^2}\frac{1}{n}.
\end{aligned}
$$

Combining this estimate with (9.4) we obtain[27]

$$|p_n(x) - f(x)| \le \frac{1}{m} + 2\left(\max_{x\in[0,1]}|f(x)|\right)\frac{1}{\delta_m^2}\frac{1}{n}. \tag{9.5}$$

Now, for a given $\varepsilon > 0$ we fix $m \in \mathbb{N}$ large enough so that $\frac{1}{m} < \frac{\varepsilon}{2}$ and then take $n \in \mathbb{N}$ so that the second summand on the right-hand side of (9.5) is less than $\frac{\varepsilon}{2}$, completing the proof. □

9.3.2. **Polling a large population**. The example studied here is borrowed from [11], see Section 1.9 of [11] for more details. Assume we have a finite set, call it $\mathcal{X}$, and are interested in some particular property of the elements of $\mathcal{X}$, call this property $\mathcal{P}$. Suppose we have a way of determining for any point of $\mathcal{X}$ whether it has the property $\mathcal{P}$ or not. If the set $\mathcal{X}$ is very large, it might be impractical to check all its elements in order to understand if they have the property we are looking for or not. In this light, how many elements should we inspect from $\mathcal{X}$ in order to have a reasonable guess on the proportion of elements in $\mathcal{X}$ that have property $\mathcal{P}$?

As an example of the above situation consider a country-wide election (here $\mathcal{X}$ is the set of voters) and the question one is interested is whether a voter will vote for the given candidate or not (YES / NO question: the property $\mathcal{P}$). As another example, assume we have a large number of the same product (the set $\mathcal{X}$) all produced in the same conditions, and want to check whether a particular sample will have a defect or not (the property $\mathcal{P}$). Both situations can be modeled by the above setup. Provided that the property we are interested in is *uniformly distributed* among the elements of $\mathcal{X}$, a random sampling of a small number of elements gives an efficient way of estimating the proportion of elements in $\mathcal{X}$ with property $\mathcal{P}$. We now formulate and prove this as a theorem.

Theorem 9.6. *Let $\mathcal{X}$ be a finite non-empty set and let $A \subset \mathcal{X}$ be fixed with $p = \frac{|A|}{|\mathcal{X}|}$, i.e. the portion of A in X equals p. Let $X_1, ..., X_n$ be independent and all uniformly distributed on $\mathcal{X}$. Set*

$$\bar{p} := \frac{|\{1 \le i \le n :\ X_i \in A\}|}{n}.$$

[27]Notice that by the WLLN we have that $\mathbb{P}(A_{m,n,x}) \to 0$ as $n \to \infty$ for any fixed m and x. While we keep m fixed, we need uniformity with respect to $x \in [0,1]$, which is why we used Corollary (9.3) instead of directly applying Theorem 9.4.

Then

$$\mathbb{P}(|\bar{p}-p|<\varepsilon)\geq 1-\frac{1}{4n^2\varepsilon} \quad \textit{for any } \varepsilon>0.$$

Proof. Let $\mathbb{I}_i$ be the indicator of the event $\{X_i\in A\}$, where $i=1,2,...,n$. As the proportion of A in $\mathcal{X}$ is p and each X_i is uniformly distributed in $\mathcal{X}$, we have $\mathbb{P}(\mathbb{I}_i=1)=p$ and hence $\mathbb{E}\mathbb{I}_i=p$ and $\mathrm{var}\mathbb{I}_i=p(1-p)\leq\frac{1}{4}$. Notice that

$$\bar{p}=\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_i$$

which, together with the linearity of expectation implies $\mathbb{E}(\bar{p})=p$. Now using Corollary 9.3 of the Chebishev's inequality we get[28]

$$\begin{aligned}\mathbb{P}(|\bar{p}-p|\geq\varepsilon)&\leq\frac{1}{\varepsilon^2}\mathrm{var}(\bar{p}) \text{ (using independence of } X_i)\\ &=\frac{1}{\varepsilon^2}\frac{1}{n^2}n\mathrm{var}(\mathbb{I}_i)\leq\frac{1}{\varepsilon^2}\frac{1}{4n},\end{aligned}$$

completing the proof. □

9.3.3. **Caratheodory's theorem on convex hulls**. Here we discuss applications of ideas around WLLN to convex geometry. Given a set of points $\mathcal{X}:=\{\vec{x}_1,\vec{x}_2,...,\vec{x}_n\}\subset\mathbb{R}^d$ ($d\geq 1$) the smallest convex set containing $\mathcal{X}$ is called the **convex hull** of $\mathcal{X}$, which we will denote here by $\mathrm{co}(\mathcal{X})$. It is easy to see that

$$\mathrm{co}(\mathcal{X})=\left\{\sum_{k=1}^{n}\alpha_k\vec{x}_k:\ \text{where } \alpha_k\geq 0 \text{ and } \sum_{k=1}^{n}\alpha_k=1\right\}. \tag{9.6}$$

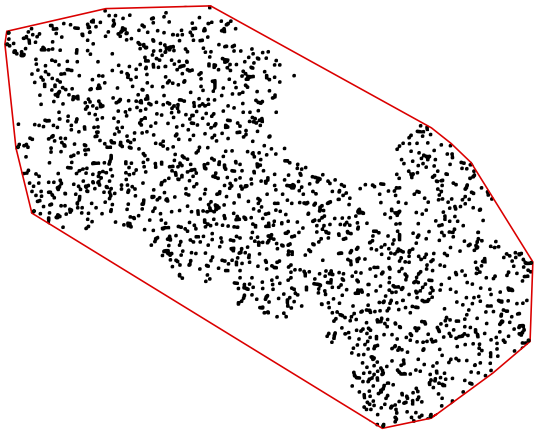

FIGURE 8. A random set of points (black dots on the figure) on $\mathbb{R}^2$ and their convex hull - the region bounded by the red-colored lines.

□ EXERCISE 9.6.1. *Prove that for a set of points $\mathcal{X}=\{\vec{x}_1,...,\vec{x}_n\}$ the convex hull is given by (9.6), i.e. the set defined in the right-hand side of (9.6) contains all points of $\mathcal{X}$, is convex, and that any convex set $\mathcal{C}$ containing $\mathcal{X}$ must also contain points defined in the right-hand side of (9.6).*

[28]By WLLN we know that $\mathbb{P}(|\bar{p}-p|>\varepsilon)\to 0$ as $n\to\infty$. The theorem here quantifies the rate of this convergence with respect to n and ε.

A well-known theorem of C. Caratheodory from convex geometry states that if a point x lies in $\mathrm{co}(\mathcal{X})$ then x can be written as a convex combination (just as above, i.e. a linear combination with non-negative coefficients that sum to 1) of at most $d+1$ points of $\mathcal{X}$. Here we study an approximate version of this result.

Theorem 9.7. *Let $\mathcal{X} := \{\vec{x}_1, ..., \vec{x}_n\} \subset \mathbb{R}^d$ be any, then for any $z \in \mathrm{co}(\mathcal{X})$ and any $k \in \mathbb{N}$ there exists $\vec{z}_1, ..., \vec{z}_k \in \mathcal{X}$ such that*

$$\left|\left|\vec{z} - \frac{1}{k}\sum_{k=1}^{n} \vec{z}_k\right|\right|_2 \leq \frac{\mathrm{diam}(\mathrm{co}(\mathcal{X}))}{\sqrt{k}},$$

where $\mathrm{diam}(\mathrm{co}(\mathcal{X}))$ is the diameter of the set $\mathrm{co}(\mathcal{X})$, i.e. the largest distance between any two points in the set.

Proof. After translating the coordinate system or $\mathbb{R}^d$ if necessary, we may assume without loss of generality that $\vec{0} \in \mathcal{X}$. This, in particular, implies the estimate

$$||\vec{x}||_2 = ||\vec{x} - \vec{0}||_2 \leq \mathrm{diam}(\mathrm{co}(\mathcal{X})). \tag{9.7}$$

Since $\vec{z} \in \mathrm{co}(\mathcal{X})$ by definition of the convex hull there exists $\alpha_1, ..., \alpha_n \geq 0$ and $\sum_{k=1}^{n} \alpha_k = 1$ such that $\vec{z} = \sum_{k=1}^{n} \alpha_k \vec{x}_k$. Let $\vec{Z}$ be a random variable such that

$$\mathbb{P}(\vec{Z} = \vec{x}_k) = \alpha_k, \ k = 1, 2, ..., n.$$

Observe that[29] $\mathbb{E}\vec{Z} = \vec{z}$. Now consider a sequence of independent discrete random variables $\vec{Z}_1, \vec{Z}_2, ...$, where all have the same distribution[30] as Z. We then have

$$\begin{aligned}
\mathbb{E}\left|\left|\vec{x} - \frac{1}{k}\sum_{i=1}^{k} \vec{Z}_i\right|\right|_2^2 &= \frac{1}{k^2}\mathbb{E}\left|\left|\sum_{i=1}^{k}\left(\vec{x} - \vec{Z}_i\right)\right|\right|_2^2 \\
&= \frac{1}{k^2}\sum_{i=1}^{k}\mathbb{E}\left|\left|\vec{x} - \vec{Z}_i\right|\right|_2^2 + \frac{1}{k^2}\sum_{\substack{1\leq i,j\leq k \\ i\neq j}} \mathbb{E}\langle \vec{x} - \vec{Z}_i, \vec{x} - \vec{Z}_j\rangle \\
&= \frac{1}{k^2}\sum_{i=1}^{k}\left(\mathbb{E}||\vec{Z}_i||_2^2 - ||\vec{x}||_2^2\right) \ \text{(by (9.7))} \\
&\leq \frac{1}{k^2}\sum_{i=1}^{k}\mathbb{E}||\vec{Z}_i||_2^2 \\
&\leq \frac{1}{k}\mathrm{diam}(\mathrm{co}(\mathcal{X}))^2.
\end{aligned}$$

From the last bound it follows that there exists a realization of $\{Z_i\}$-s such that satisfies the inequality claimed in the formulation of the theorem. The proof is complete. □

[29] Both sides of the equation are vectors and the expectation as well as the equality are both understood coordinate-wise. Observe that at this point we can conclude from WLLN that the averages of $\vec{Z}_i$-s converge to x in a weak sense. The estimates of this theorem quantify this convergence.

[30] We only need k (finite number of) copies of Z. The existence of these random variables follows from Exercise 7.12.1

10. Simple symmetric random walk

In this section we glimpse into the world of random walks. We introduce the simple one-dimensional random walk on integer lattice $\mathbb{Z} = \{..., -2, -1, 0, 1, 2, ...\}$, prove the *reflection principle* and the *arcsin law* for it.

10.1. **Random walks on $\mathbb{Z}$ and the reflection principle.** Consider a particle *walking* on the integer lattice $\mathbb{Z}$ one step at a time either to the left or to the right of its current position, deciding the direction of the next move by a coin toss experiment. The next definition formalizes this process.

Definition 10.1. (`Simple random walk`) *Let $X_1, X_2, ..., X_n, ...$ be a sequence of independent and identically distributed (i.i.d.) random variables where $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = -1) = 1 - p$, for some $0 \leq p \leq 1$. Define a new sequence $\{S_n\}$ where*

$$S_0 = 0, \quad S_n = X_1 + X_2 + ... + X_n, \text{ for } n \geq 1. \tag{10.1}$$

*The sequence $\{S_n\}$ is called a (simple) **random walk** on $\mathbb{Z}$. When $p = 1/2$, the random walk is called **symmetric**.*

Throughout this section we will assume that the walk is **symmetric**, i.e. $p = 1/2$ in Definition 10.1.

Random walks serve as a fundamental model for stochastic activities arising in various scenarios, such as fluctuating stock values, a path of a molecule traveling in a liquid, etc. Along with a setting of the integer lattice $\mathbb{Z}$ described above there are a number of variations. For example, a random walk can be defined on the d-dimensional lattice $\mathbb{Z}^d$, with $d \geq 2$, where at each step the particle walks by selecting one of its $2d$ lattice neighbors. There are versions of a walk, where the particle can choose to stay at its current position with a certain probability (the so-called *lazy* walks). Other spaces where a version of the random walk is defined and studied include graphs, Riemannian manifolds, finitely generated groups, etc. There is a vast theory developed for random walks in various settings, with deep and beautiful mathematical results. The aim of this section is to introduce the concept of random walks in the simplest form and prove with very basic techniques some important properties of simple symmetric random walks.

10.1.1. ***Counting paths.*** We can think about the sequence $\{S_n\}$ as a polygonal line joining points $(n, S_n) \to (n + 1, S_{n+1})$ (see Figure 9). Each such **path** corresponds to a particular realization of the random walk which can be identified with an outcome of the joint experiment with $X_1, ..., X_n$. Thus the number of **all paths** starting at $(0, 0)$ and moving for n time-steps equals 2^n.

Observe that a walk starting from $(0, 0)$ and moving for n steps cannot terminate on any integer from $[-n, n]$. For example, when $n = 2$, none of the four paths of the walk ends at 1. In the light of this observation let us compute the number of paths from $(0, 0)$ to (n, x) for a given point (n, x). Assume the walk does a steps up and b steps down, i.e. the number of $1 \leq i \leq n$ for which $X_i = 1$ equals a and for $X_i = -1$ equals b. Since we do n steps in total and arrive at x from 0, we must have

$$a + b = n \text{ and } a - b = x.$$

Solving this system we obtain that

$$a = \frac{n + x}{2} \text{ and } b = \frac{n - x}{2}.$$

Let $N_{(0,0)\mapsto(n,x)}$ be the number of paths from $(0, 0)$ to (n, x). If the context is clear we will drop the $(0, 0)$ from the subscript and write $N_{(n,x)}$ for the number of paths from

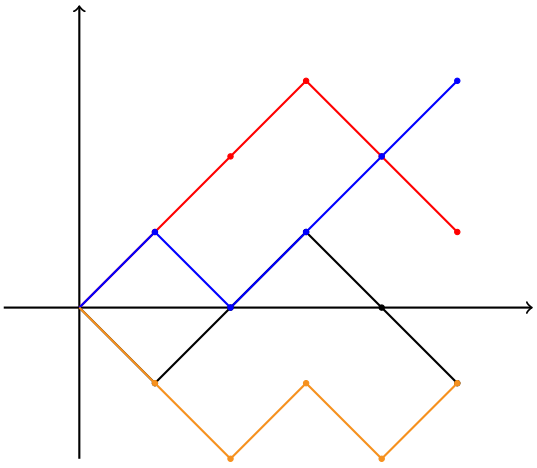

FIGURE 9. Each of the four colors corresponds to a particular realization of a random walk starting from $(0,0)$ and moving for 5 steps. An outcome of the joint experiment with $(X_1,...,X_n)$, as defined in (10.1), can be put in bijective correspondence with a path similar to those depicted here. In the case of 5 time-steps there are 2^5 different paths.

$(0,0)$ to (n,x). We will think about n as *time*, and about x as the *location* of the walk at time n. It is now clear that if $n+x$ is not an even number, then there are no paths ending at x in n steps, otherwise the number of such paths equals

$$N_{(0,0)\mapsto(n,x)} = \begin{cases} 0, & \text{if } n+x \text{ is odd,} \\ \binom{n}{a}, & \text{otherwise, for } a = \frac{n+x}{2}. \end{cases} \tag{10.2}$$

Let us also note that we can shift the time and location of the starting position of the path. Namely, for any $t \in \mathbb{Z}$ and any $x_0 \in \mathbb{Z}$ we have

$$N_{(0,0)\mapsto(n,x)} = N_{(t,0)\mapsto(n+t,x)} = N_{(t,x_0)\mapsto(n+t,x+x_0)}. \tag{10.3}$$

By $N^0_{(0,0)\mapsto(n,x)}$ denote the number of paths from $(0,0)$ to (n,x) that become 0 for some $1 \le i \le n$. The paths that start and end at positive integers but become zero at some intermediate point are of special interest. The next result links the number of paths that hit zero at some intermediate point with the number of all paths.

Theorem 10.1. `(Reflection principle for random walks)` *Let $x,y \ge 1$ be integers. Then*

$$N^0_{(0,x)\mapsto(n,y)} = N_{(0,-x)\mapsto(n,y)},$$

i.e. the number of paths from $(0,x)$ to (n,y) that become 0 at some intermediate point $1 \le i \le n-1$ equals the number of all paths from $(0,-x)$ to (n,y).

Proof. Let A be the set of all paths from $(0,x)$ to (n,y) that become 0 at some point $1 \le i \le n-1$, and let B be the set of all paths from $(0,-x)$ to (n,y). We need to show that $|A| = |B|$. Let $(0,s_0) \to (1,s_1) \to ... \to (n,s_n)$ be any path from A. By definition we have $s_0 = x$, $s_n = y$ and there exists $1 \le i \le n-1$ such that $s_i = 0$. Define $k = \min\{1 \le i \le n-1 : \ s_i = 0\}$, i.e. the first time the path hits 0. Now define a new path as follows (see Figure 10):

$$\tilde{s}_i = \begin{cases} -s_i, & \text{for } 0 \le i \le k, \\ s_i, & \text{for } k < i \le n. \end{cases} \tag{10.4}$$

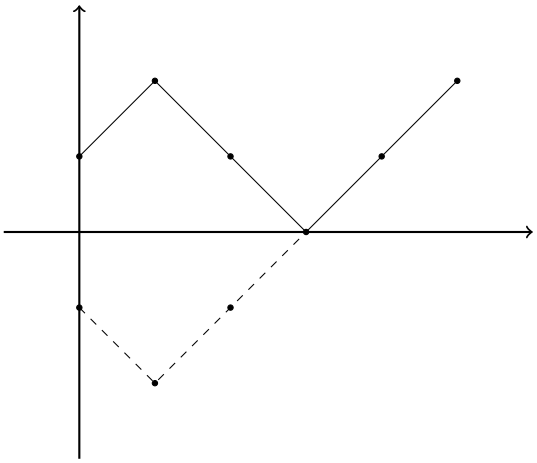

FIGURE 10. One of the paths from $(0,1)$ to $(5,2)$ that hits 0 in an intermediate point (at step 3 in this example). Part of the path that is reflected with respect to the x axis is depicted by dashed lines. This reflection generates a path from $(0,-1)$ to $(5,2)$.

Clearly we have a path from $(0,-x)$ to (n,y). Such reflection defines a one-to-one (injective) mapping from A to B implying that $|A| \leq |B|$.

To show $|A| \geq |B|$ take any path from B. Since $x, y > 0$ the path must intersect 0 before reaching y. Let $1 \leq i \leq n-1$ be the first time the path becomes 0. Then the reflected path given by (10.4) defines a one-to-one correspondence between B and A, hence $|B| \leq |A|$ and the proof is complete. □

Here is a peculiar application of the reflection principle.

Theorem 10.2. (`Ballot theorem`) *Assume in a certain election candidate $\mathcal{A}$ gets α votes and candidate $\mathcal{B}$ gets β votes where $\alpha > \beta > 0$. Then, in counting of votes the probability that $\mathcal{A}$ will always be ahead of $\mathcal{B}$ equals $\frac{\alpha-\beta}{\alpha+\beta}$.*

Notice that each realization of a counting process of votes is a path that leads from $(0,0)$ to $(\alpha+\beta, \alpha-\beta)$. The counting scenarios where candidate $\mathcal{A}$ is always ahead of candidate $\mathcal{B}$ correspond to the paths that never cross 0.

Proof. Define $S_0 = 0$ and for $n \geq 1$ set $S_n = X_1 + ... + X_n$, where

$$X_i = \begin{cases} 1, & \text{if the } i\text{-th vote was for } \mathcal{A}, \\ -1, & \text{if the } i\text{-th vote was for } \mathcal{B}. \end{cases}$$

Set $T := \alpha + \beta$ and $D := \alpha - \beta$ representing the total number of votes and the difference of votes for $\mathcal{A}$ and $\mathcal{B}$. By definition $\{S_n\}$ represents the counting process that starts at 0 and ends at D. We are interested in the number of paths that do not intersect 0, i.e. when $S_n > 0$ for all $n = 1, 2, ..., T$. The latter is equal to the number of paths from $(1,1)$ to (T, D) that do not cross 0. Shifting the time axis by 1 we need to count paths from $(0,1)$ to $(T-1, D)$ that stay positive. By Theorem 10.1 (reflection

principle) the number of such paths equals

$$\begin{aligned} &N_{(0,1)\mapsto(T-1,\,D)} - N_{(0,-1)\mapsto(T-1,\,D)} \\ &= N_{(0,0)\mapsto(T-1,\,D-1)} - N_{(0,0)\mapsto(T-1,\,D+1)} \quad \text{(by (10.2))} \\ &= \binom{\alpha+\beta-1}{\alpha-1} - \binom{\alpha+\beta-1}{\alpha} = \frac{(\alpha+\beta-1)!}{(\alpha-1)!\beta!} - \frac{(\alpha+\beta-1)!}{\alpha!(\beta-1)!} \\ &= \frac{\alpha-\beta}{\alpha+\beta}\binom{\alpha+\beta}{\alpha} = \frac{\alpha-\beta}{\alpha+\beta} N_{(0,0)\mapsto(\alpha+\beta,\,\alpha-\beta)}. \end{aligned}$$

We proved that the number of paths where $\mathcal{A}$ always stays ahead of $\mathcal{B}$ equals $(\alpha-\beta)/(\alpha+\beta)$ of the total, hence the proof is complete. □

10.2. **Arcsin law.** Recall that we assume the walk is *symmetric*. In this section we will study distributions of *first* and *last* visits to 0 of a walk starting from the origin.

Lemma 10.3. *Let S_n be as in (10.1). Then*

$$\mathbb{P}(S_1 \neq 0, S_2 \neq 0, ..., S_{2n} \neq 0) = \mathbb{P}(S_{2n} = 0).$$

Proof. We will show that the numbers of paths on both sides of the equality in question are equal. This will complete the proof in view of the fact that the walk is symmetric and all paths have the same probability, .

The walk starts from 0 hence in $2n$ steps it can reach at most $2n$, thus

$$\mathbb{P}(S_1 > 0, S_2 > 0, ..., S_{2n} > 0) = \sum_{k=1}^{n} \mathbb{P}(S_1 > 0, ..., S_{2n-1} > 0, S_{2n} = 2k).$$

For each $1 \leq k \leq n$ the number of paths with $S_1 > 0, ..., S_{2n-1} > 0, S_{2n} = 2k$ equals

$$\begin{aligned} &N_{(1,1)\mapsto(2n,2k)} - N^0_{(1,1)\mapsto(2n,2k)} \text{ (shifting time)} \\ &= N_{(0,1)\mapsto(2n-1,2k)} - N^0_{(0,1)\mapsto(2n-1,2k)} \text{ (by reflection principle, Theorem 10.1)} \\ &= N_{(0,1)\mapsto(2n-1,2k)} - N_{(0,-1)\mapsto(2n-1,2k)} \text{ (shifting location)} \\ &= N_{(0,0)\mapsto(2n-1,2k-1)} - N_{(0,0)\mapsto(2n-1,2k+1)}. \end{aligned}$$

Thus the number of paths with $S_1 > 0, S_2 > 0, ..., S_{2n} = 2k$ for some $1 \leq k \leq n$ equals

$$\begin{aligned} &\sum_{k=1}^{n} \left(N_{(0,0)\mapsto(2n-1,2k-1)} - N_{(0,0)\mapsto(2n-1,2k+1)} \right) \\ &= N_{(0,0)\mapsto(2n-1,1)} - N_{(0,0)\mapsto(2n-1,2n+1)} = \binom{2n-1}{n}, \end{aligned}$$

where we used (10.2) for the last equality. In view of the symmetry the number of paths that are always negative equals the number of paths that are always positive. Hence the number of paths such that $S_1 \neq 0, S_2 \neq 0, ..., S_{2n} \neq 0$ becomes

$$2\binom{2n-1}{n} = \binom{2n}{n} = N_{(0,0)\to(2n,0)}, \tag{10.5}$$

where the last equality follows by (10.2). We see that the number of paths that never hit 0 equals the number of paths that end at 0 at time $2n$. This settles the lemma as noted at the beginning of the proof. □

An important corollary of the last lemma is an estimate on probability of the **first visit to zero**. Let $T = \inf\{m \geq 1 : \ S_m = 0\}$, i.e. T is the first time a symmetric random walk started at 0 will revisit it. Then, by Lemma 10.3 we have

$$\mathbb{P}(T > 2n) = \mathbb{P}(S_{2n} = 0) = \frac{1}{2^{2n}}\binom{2n}{n} \sim \pi^{-1/2} n^{-1/2}, \text{ as } n \to \infty, \tag{10.6}$$

where we used (10.5) and Stirling's formula (2.1).

We now consider the **last visits to zero**. Let $L_{2n} := \sup\{m \leq 2n : \ S_m = 0\}$, and set $u_{2m} := \mathbb{P}(S_{2m} = 0)$.

Lemma 10.4. *For each $1 \leq k \leq n$ we have $\mathbb{P}(L_{2n} = 2k) = u_{2k}u_{2n-2k}$.*

Proof. Notice that after hitting 0 at $2k$ the walk does not visit 0 until step $2n$. Hence

$$\begin{aligned}
\mathbb{P}(L_{2n} = 2k) &= \mathbb{P}(S_{2k} = 0, \ S_{2k+1} \neq 0, ..., S_{2n} \neq 0) \\
&= \mathbb{P}(S_{2k} = 0)\mathbb{P}(S_{2k+1} \neq 0, ..., S_{2n} \neq 0 \,|\, S_{2k} = 0) \\
&= u_{2k}\mathbb{P}(S_1 \neq 0, ..., S_{2n-2k} \neq 0) \text{ (using Lemma 10.3)} \\
&= u_{2k}u_{2n-2k},
\end{aligned}$$

where the passage from the second to the third row is due to the fact that a walk conditioned to be 0 at step $2k$ has the same distribution as the walk starting at 0. The lemma is proved. □

Theorem 10.5. `(Arcsin law for the last visit to 0)` *For any $0 < a < b < 1$ we have*

$$\mathbb{P}\left(a < \frac{L_{2n}}{2n} < b\right) \to \frac{1}{\pi}\int_a^b \frac{dx}{\sqrt{x(1-x)}}, \text{ as } n \to \infty.$$

Remark 10.6. *To see why this law is called **arcsin**, set $\sqrt{x} = y$ in the integral, which leads to*

$$\int_a^b \frac{dx}{\sqrt{x(1-x)}} = 2\int_{\sqrt{a}}^{\sqrt{b}} \frac{dy}{\sqrt{1-y^2}} = 2(\arcsin\sqrt{b} - \arcsin\sqrt{a}).$$

Since $\arcsin\frac{1}{\sqrt{2}} = \frac{\pi}{4}$ *we get*

$$\mathbb{P}\left(\frac{L_{2n}}{2n} < \frac{1}{2}\right) \to \frac{1}{2} \text{ as } n \to \infty. \tag{10.7}$$

Here is an interesting interpretation of this law in gambling terms. Assume two players with initial capital of 0 gamble on fair coin toss. It follows from (10.7) that in a long series of tosses one of the players will be ahead the entire second half of the series with probability close to 1/2.

Proof of Theorem 10.5. By Lemma 10.4 and (10.6) we have

$$\mathbb{P}(L_{2n} = 2k) = \frac{1}{\pi}\frac{1}{\sqrt{k(n-k)}}\beta_{n,k}, \tag{10.8}$$

where $\beta_{n,k} \to 1$ as both $k, n \to \infty$. For each $n \in \mathbb{N}$ let $a \leq a_n < b_n \leq b$ be chosen so that $2na_n$ is the smallest even integer larger than equal to $2na$, and $2nb_n$ is the largest

integer smaller than equal to $2nb$. Clearly $a_n \to a$ and $b_n \to b$ as $n \to \infty$. It follows that

$$\mathbb{P}\left(a < \frac{L_{2n}}{2n} < b\right) = \sum_{k=na_n}^{nb_n} \mathbb{P}(L_{2n} = 2k) = \sum_{k=na_n}^{nb_n} \frac{1}{\pi} \frac{1}{\sqrt{k(n-k)}} \beta_{n,k}. \tag{10.9}$$

where the last equality is due to (10.8). Let $f(x) = \frac{1}{\sqrt{x(1-x)}}$, where $0 < x < 1$. Writing the integral sum for f on $[a, b]$ for partition

$$\left\{ a_n + \frac{k}{n} : \ k = 0, 1, ..., (b_n - a_n)n \right\}$$

we get

$$\frac{1}{n} \sum_{k=0}^{(b_n-a_n)n} f\left(a_n + \frac{k}{n}\right) = \frac{1}{n} \sum_{k=0}^{(b_n-a_n)n} \frac{1}{\sqrt{\frac{na_n+k}{n}\left(1 - \frac{na_n+k}{n}\right)}} =$$

$$\sum_{k=na_n}^{nb_n} \frac{1}{\sqrt{k(n-k)}} \to \int_a^b f(x)dx, \text{ as } n \to \infty.$$

Combining this with (10.9) and taking into account that $\beta_{n,k} \to 1$ the proof is complete. □

11. Introduction to geometric probabilities: motivating the need to formalize randomness

In this section we study several classical examples on geometric probabilities and motivate the need for formalizing the intuitive meaning of a random experiment. In the following example we do not formalize the notion of probability, but rather recourse to its intuitive meaning.

11.0.1. ***Bertrand's paradox.*** Consider the following problem.

> Given a circle of radius 2 and a concentric circle of radius 1, take a random chord of the larger circle. What is the probability that the chord will intersect the smaller concentric circle?

This problem (along with certain variations of it) is often referred to as *Bertrand's paradox*. We now discuss several approaches to it.

Approach 1. The chord is uniquely determined by its midpoint, unless the chord is passing through the origin. This is due to the fact that the line through the origin and the midpoint must be perpendicular to the chord as the height of a triangle formed by the chord and the two radii of the circle. Hence, in order for the chord to intersect the inner circle we need its midpoint to be in the concentric circle of radius 1. In view of this the probability of intersection becomes

$$\frac{\pi 1^2}{\pi 2^2} = \frac{1}{4}.$$

Approach 2. By symmetry we may assume that the chords are all vertical. Indeed, whether the chord intersects the inner circle or not is invariant under rotation of the plane. Hence we may rotate the plane so that the chord becomes vertical. Then, a

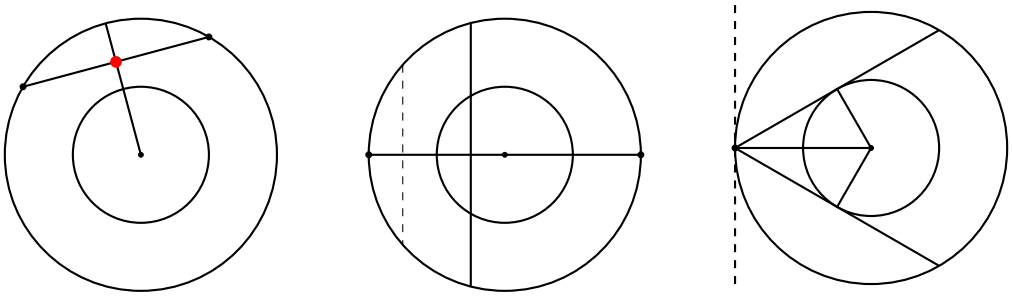

FIGURE 11. A schematic diagram of the problem and the three approaches discussed here related to Bertrand's paradox. Both the smaller and the larger circles are centered at $(0,0)$, with radii being 1 and 2 correspondingly. Each of the graphs (best seen online) in the order from left to right corresponds to the approach 1, 2 and 3.

vertical chord intersects the inner circle if and only if it passes through the segment $[-1,1]$. Hence the probability of intersection becomes

$$\frac{\text{length of } [-1,1]}{\text{length of } [-2,2]} = \frac{1}{2}.$$

Approach 3. By symmetry, rotating the plane if necessary, we may assume that the chord passes through $(-2,0)$. Let $\theta \in (-\pi/2, \pi/2)$ be the angle that the chord makes with the $\vec{x}$-axis. It is easy to see that the chord intersects the inner circle if and only if $\theta \in (-\pi/6, \pi/6)$. Hence the probability of intersection becomes

$$\frac{\text{length of } (-\pi/6, \pi/6)}{\text{length of } (-\pi/2, \pi/2)} = \frac{1}{3}.$$

□ EXERCISE 11.0.1. *Describe the probability space according to Definition 1.3 in all three approaches described above.*

11.0.2. ***Back to probability spaces.*** Doing logical steps in the approaches described for Bertrand's problem we arrived at three different answer. There is no error in each approach per se, the apparent contradiction, however, arises from the fact that all three approaches, while seemingly the same at the first glance, utilize different models for the same experiment. More precisely, the probabilistic model for the experiment with random chord is different in each of the three approaches described above. The morale of the example is that one can get ambiguous results if the random experiment under consideration is not formally defined, even in simple situations as we had above. We now describe some examples of probability spaces, i.e. triples $(\Omega, \mathcal{F}, \mathbb{P})$ that satisfy Definition 1.3.

Example 1. Take $\Omega = \{\omega_1, ..., \omega_n\}$ and let $\{p_i\}_{i=1}^n$ be a probability distribution, i.e. $0 \le p_i \le 1$ for $i = 1, ..., n$ and $\sum_{i=1}^n p_i = 1$. Then take $\mathcal{F} = 2^\Omega$ and define $\mathbb{P} : \mathcal{F} \to [0,1]$ as

$$\mathbb{P}(A) = \sum_{i:\ \omega_i \in A} p_i \quad \text{for any } A \subset \Omega.$$

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

Example 2. Take $\Omega = \mathbb{R}$, let $\mathcal{F}$ be the Borel σ-algebra on $\mathbb{R}$. Let also $f : \mathbb{R} \to \mathbb{R}_+$ be a Lebesgue integrable non-negative function such that $\int_{\mathbb{R}} f d\mu = 1$ where μ is the

Lebesgue measure. Define $\mathbb{P} : \mathcal{F} \to [0,1]$ as

$$\mathbb{P}(A) = \int_A f d\mu \quad \text{for any } A \in \mathcal{F}. \tag{11.1}$$

Thanks to the choice of f and the standard properties of the Lebesgue measure we get that $\mathbb{P}$ defined by (11.1) is indeed a probability measure. Thus the triple $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. Some particular choices of f, described next, are of special interest.

Example 2a. Take $f(x) = \mathbb{I}_{[0,1]}(x)$ - the indicator function of $[0,1]$, i.e. $f(x) = 1$ for $x \in [0,1]$ and $f(x) = 0$ for $x \notin [0,1]$. Then $\mathbb{P}$ defined by (11.1) for this particular choice of f is precisely the Lebesgue measure restricted to Borel subsets of $[0,1]$.

Example 2b. Take

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}.$$

Claim 11.1. *Let f be as above. Then $\int_{\mathbb{R}} f(x)dx = 1$.*

Proof. Define $I := \int_{\mathbb{R}} e^{-\frac{x^2}{2}} dx$. We need to show that $I^2 = 2\pi$. We have

$$\begin{aligned} I^2 &= \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-\frac{x^2+y^2}{2}} dxdy \quad \text{(passing to polar coordinates)} \\ &= \int_0^{\infty} \int_0^{2\pi} e^{-\frac{r^2}{2}} r d\varphi dr = \pi \int_0^{\infty} e^{-\frac{r^2}{2}} dr^2 = \pi \int_0^{\infty} e^{-\frac{t}{2}} dt = 2\pi, \end{aligned}$$

and the proof is complete. □

We thus get that f has integral equal to 1 on $\mathbb{R}$ and hence defines a probability measure by (11.1).

Example 3. Fix any $x_0 \in \mathbb{R}$ and for $A \subset \mathbb{R}$ define

$$\mathbb{P}(A) = \begin{cases} 1, & \text{if } x_0 \in A, \\ 0, & \text{if } x_0 \notin A. \end{cases}$$

With such $\mathbb{P}$ the triple $(\mathbb{R}, 2^{\mathbb{R}}, \mathbb{P})$ is a probability space.

11.0.3. ***Buffon's needle problem.*** Here we study another classical problem in geometric probability.

> The plane $\mathbb{R}^2$ is partitioned into strips by parallel lines that are 2 units away from each other. A segment of length 1 is thrown on the plane. What is the probability that this segment intersects one of the parallel lines?

Let $h \geq 0$ be the distance of the midpoint of the chord to its closest line, and let $\theta \in [0, \pi]$ be the angle the chord makes with the positive direction of the $\vec{x}$ axis. Let also x be the length of the hypothenus of the triangle outlined in the Figure 12. We see that $\frac{h}{x} = \sin\theta$. Since the length of the segment is 1 we get that the segment intersects its closest line if and only if

$$x = \frac{h}{\sin\theta} \leq \frac{1}{2}. \tag{11.2}$$

We now formalize the discussion above. Define $\Omega = [0,1] \times [0, \pi)$ and let $\mathcal{B}$ be the Borel σ-algebra on Ω. For each $A \in \mathcal{B}$ define

$$\mathbb{P}(A) = \frac{1}{\pi}\mu(A), \quad \text{where } \mu \text{ is the Lebesgue measure.}$$

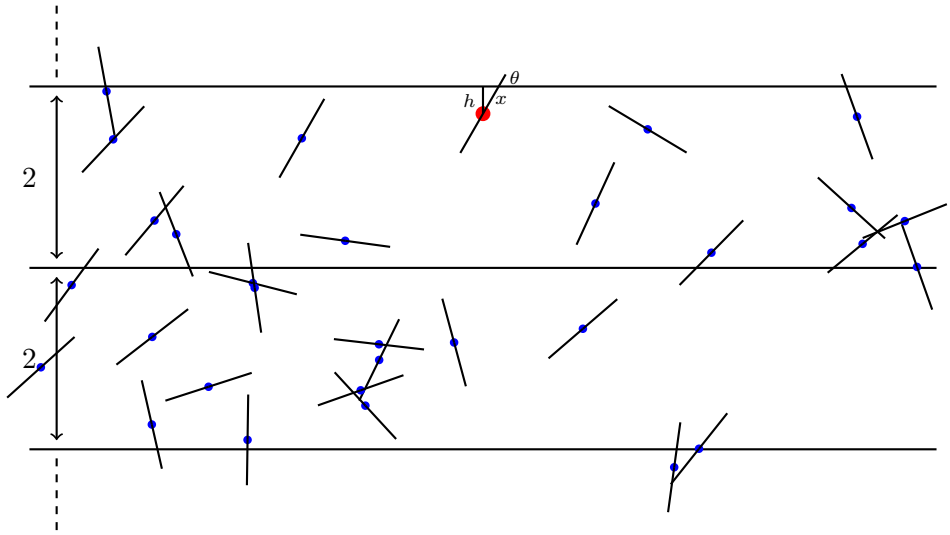

FIGURE 12. A schematic diagram (best seen online) of the Buffon's needle problem. The plane is separated by parallel horizontal lines 2 units apart from each other, and the experiment consists of throwing a random segment of length 1 on the plane. Each of the segments seen on the graph is obtained as a result of this experiment. The experiment here is modeled as a random sampling of a distance of the midpoint of the segment (the blue points on the graph) from its nearest line, followed by random sampling (independently of the midpoint selection) of the angle that the segment makes with the positive direction of $\vec{x}$ axis. The handpicked segment with red dot as its midpoint incorporates the necessary parameters. Namely, h is the distance from the midpoint to the closest line and x is the hypothenus of the right-angle triangle whose one vertex is the red midpoint and one edge equals h. Also, θ is the angle that the segment makes with the horizontal $\vec{x}$ axis.

We have that $(\Omega, \mathcal{B}, \mathbb{P})$ is a probability space. By (11.2) the probability of intersection becomes

$$\mathbb{P}(\text{intersection}) = \frac{1}{\pi}\int_0^1\int_0^\pi \mathbb{I}_{\left\{(h,\theta)\in\Omega:\ \frac{h}{\sin\theta}\le\frac{1}{2}\right\}} d\theta dh$$
$$= \frac{1}{\pi}\int_0^\pi\int_0^{(\sin\theta)/2} dhd\theta = \frac{1}{\pi}\int_0^\pi \frac{1}{2}\sin\theta d\theta = \frac{1}{\pi}.$$

□ EXERCISE 11.1.1. *Compute the probability in the Buffon's needle problem when the lines are $2d > 0$ apart from each other instead of 2, and the segment thrown onto the plane has length $0 < a < d$.*

12. RANDOM VARIABLE AND ITS LAW

In this section we introduce random variables, its law and discuss some basic properties.

12.1. **Random variable and associated σ-algebras.** In Section 6 we defined random variables in the discrete setting. Here we will work with mappings defined on Ω with values that are not necessarily discrete.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X : \Omega \to \mathbb{R}$ be any mapping. For any $A \subset \mathbb{R}$ we set

$$X^{-1}(A) := \{\omega \in \Omega : \ X(\omega) \in A\},$$

and call $X^{-1}(A)$ the **preimage** of A under X. As a shorthand for $X^{-1}(A)$ we will often write $\{X \in A\}$. In what follows we prove structural results concerning preimages.

Claim 12.1. (`Sets with preimages in a` σ`-algebra form a` σ`-algebra`) *Let Ω be a non-empty set, $X : \Omega \to \mathbb{R}$ be any mapping and let $\mathcal{G}$ be a σ-algebra of subsets of Ω. Then the set*

$$\mathcal{A} = \{A \subset \mathbb{R} : \ X^{-1}(A) \in \mathcal{G}\}$$

is a σ-algebra on $\mathbb{R}$.

Proof. We need to check that $\mathcal{A}$ satisfies conditions of Definition 1.1. Clearly $\mathcal{A}$ is non-empty. Now take any $A \in \mathcal{A}$, we need to show that $A^c = \mathbb{R} \setminus A \in \mathcal{A}$. To this end observe that

$$\begin{aligned} X^{-1}(\mathbb{R} \setminus A) &= \{\omega \in \Omega : \ X(\omega) \notin A\} \\ &= \Omega \setminus \{\omega \in \Omega : \ X(\omega) \in A\} = \Omega \setminus X^{-1}(A). \end{aligned}$$

Since $\mathcal{G}$ is a σ-algebra and $X^{-1}(A) \in \mathcal{G}$ by the choice of A we get that $\Omega \setminus X^{-1}(A) \in \mathcal{G}$ and hence $\mathbb{R} \setminus A \in \mathcal{A}$. It is left to show that $\mathcal{A}$ is closed under countable unions. Take $A_1, A_2, ... \in \mathcal{A}$, then

$$\begin{aligned} X^{-1}\left(\bigcup_i A_i\right) &= \left\{\omega \in \Omega : \ X(\omega) \in \bigcup_i A_i\right\} \\ &= \bigcup_i \{\omega \in \Omega : \ X(\omega) \in A_i\} = \bigcup_i X^{-1}(A_i). \end{aligned}$$

By the choice of A_i we have that $X^{-1}(A_i) \in \mathcal{G}$ and since $\mathcal{G}$ is closed under countable unions we get the same property for $\mathcal{A}$. The proof is complete. □

Claim 12.2. (`Preimage of a` σ`-algebra is a` σ`-algebra`) *Let Ω be a non-empty set and $X : \Omega \to \mathbb{R}$ be any mapping. Let also $\mathcal{G}$ be a σ-algebra of subsets of $\mathbb{R}$. Then*

$$\{X^{-1}(A) : \ A \in \mathcal{G}\}$$

is a σ-algebra in Ω.

Proof. One needs to check the requirements of Definition 1.1, which follow in a straightforward manner, similar to the proof of Claim 12.1, using the fact that $\mathcal{G}$ is a σ-algebra. □

□ Exercise 12.2.1. *Complete the details in the proof of Claim 12.2.*

An important σ-algebra associated with a mapping $X : \Omega \to \mathbb{R}$ is defined as follows

$$\sigma(X) := \{X^{-1}(B) : \ B \in \mathcal{B}(\mathbb{R})\}, \tag{12.1}$$

where $\mathcal{B}(\mathbb{R})$ is the Borel σ-algebra of $\mathbb{R}$. By Claim 12.2 we know that $\sigma(X)$ is indeed a σ-algebra. It is called a **σ-algebra generated by X**. Notice that we have imposed no restriction on X so far. The mappings X defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, for which $\sigma(X)$ becomes a sub-σ-algebra of $\mathcal{F}$ are of special interest, as we define next.

Definition 12.1. (Random variable) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ be any mapping. Then X is called a **random variable** if $X^{-1}(B) \in \mathcal{F}$ for any $B \in \mathcal{B}(\mathbb{R})$ where $\mathcal{B}(\mathbb{R})$ is the Borel σ-algebra on $\mathbb{R}$.*

A few remarks are in order.

- Definition 12.1 requires $\sigma(X) \subset \mathcal{F}$. By Claim 12.2 we know that $\sigma(X)$ is indeed a σ-algebra in Ω. But when working with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the event space $\mathcal{F}$ is all the information we have that we can measure (compute probabilities) and requiring $\sigma(X) \subset \mathcal{F}$ means that X does not carry an information which we cannot measure.
- To stress the fact that preimage of B under X is required to be from $\mathcal{F}$ we might say that X is *$\mathcal{F}$-measurable*. This distinction will be useful when considering sub-σ-algebras of $\mathcal{F}$.
- In the light of the previous two remarks and Claim 12.2, one can think about $\sigma(X)$ as the **smallest σ-algebra** with respect to which X is measurable. For example, let $\Omega = \{1, 2, 3, 4, 5, 6\}$ be the usual sample space from the experiment of rolling a 6-sided dice. Consider the mapping $X : \Omega \to \mathbb{R}$ defined as
$$X(\omega) = \begin{cases} 1, & \omega = 1, \\ 0, & \omega \neq 1. \end{cases}$$
Then $\sigma(X) = \{\emptyset, \Omega, \{1\}, \{2, 3, 4, 5, 6\}\}$, and thus the event space $\mathcal{F}$ must be at least as large as $\sigma(X)$ to make X a random variable. As another example on the same sample space Ω, consider $X(\omega) = \omega$ for all $\omega \in \Omega$. Then $\sigma(X) = 2^{\Omega}$ and hence the only event space that makes X a random variable is the set of all subsets of Ω.
- Instead of real-valued mappings we can consider mappings into $\mathbb{R}^d$ with $d > 1$. In such case we say that X is a *random vector* (instead of a *variable*) to stress the fact that the values it takes are vectors rather than real numbers.

Simple yet very important examples of random variables are indicator functions of events. Namely, given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ take any $A \in \mathcal{F}$ and consider its indicator function, i.e.
$$\mathbb{I}_A(\omega) = \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A. \end{cases}$$
Then $\mathbb{I}_A$ is a random variable in a sense of Definition 12.1.

□ Exercise 12.2.2. (Consistency with the discrete case) *Assume $X : \Omega \to \mathbb{R}$ takes at most countably many values (i.e. X is discrete). Show that for such X Definitions 12.1 and 6.1 are equivalent.*

By Definition 12.1 to check that a given mapping is a random variable we need to assert that the preimage of any Borel subset of $\mathbb{R}$ is an event in Ω. The next result allows one to concentrate only on a subset of Borel sets with a nicer structure.

Lemma 12.3. (Measurability of sublevel sets is enough) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ be any mapping. Then X is a random variable if and only if $\{\omega \in \Omega : \ X(\omega) \leq x\} \in \mathcal{F}$ for any $x \in \mathbb{R}$.*

Proof. Notice that $\{\omega \in \Omega : \ X(\omega) \leq x\} = X^{-1}((-\infty, x])$ and $(-\infty, x] \in \mathcal{B}(\mathbb{R})$, hence if X is a random variable then $\{X \leq x\} \in \mathcal{F}$ for any $x \in \mathbb{R}$. We now prove the other direction of the lemma.

By Claim 12.1 we have that

$$\mathcal{A} := \{A \subset \mathbb{R} : \ X^{-1}(A) \in \mathcal{F}\}$$

is a σ-algebra in $\mathbb{R}$. The condition of the lemma states that $(-\infty, x] \in \mathcal{A}$ for any $x \in \mathbb{R}$. Hence $\mathcal{A}$, being a σ-algebra itself, contains the smallest σ-algebra containing the sets of the form $\{(-\infty, x] : \ x \in \mathbb{R}\}$. But the latter is the Borel σ-algebra of $\mathbb{R}$, hence for any $B \in \mathcal{B}(\mathbb{R})$ we get that $B \in \mathcal{A}$. By definition of $\mathcal{A}$ this is equivalent to $X^{-1}(B) \in \mathcal{F}$. Hence X is a random variable and the proof is complete. □

The following remarks provide more context to Lemma 12.3.

Remark 12.4. (`Sublevel sets and other generators of Borel σ-algebra`) *The sets $\{X \leq x\}$ considered in Lemma 12.3 are called* **sublevel sets** *of X. What we used in fact is that the collection of sets $\{(-\infty, x], \ x \in \mathbb{R}\}$ generate $\mathcal{B}(\mathbb{R})$ - the Borel σ-algebra of $\mathbb{R}$. The same proof works for any family of sets that generate $\mathcal{B}(\mathbb{R})$, for example we can restrict $x \in \mathbb{R}$ to rationals only, or consider sets of the form $\{X \geq x\}$, or sets of the form $\{a \leq X < b\}$ with $a < b$ real numbers, etc.*

Remark 12.5. (`Measurability with respect to a sub-σ-algebra of` $\mathcal{F}$) *We may sometimes need to consider mappings with preimages in a sub-σ-algebra of $\mathcal{F}$. For such scenarios, assume $\mathcal{G} \subset \mathcal{F}$ is σ-algebra and for some mapping $X : \Omega \to \mathbb{R}$ we have $\{X \leq x\} \in \mathcal{G}$. Then the same proof as we had for the lemma above implies that $X^{-1}(B) \in \mathcal{G}$ for any Borel set $B \in \mathcal{B}(\mathbb{R})$. Similarly to Remark 12.4 the sublevel sets $\{X \leq x\}$ can be replaced by preimages of any collection of sets that generate Borel σ-algebra of $\mathbb{R}$.*

→ **Problem 12.5.1.** *Prove that the minimal σ-algebra containing all sets in $\mathcal{X} := \{(-\infty, x], \ x \in \mathbb{R}\}$ is the Borel σ-algebra of $\mathbb{R}$.*

In the spirit of Lemma 12.3 we have the following.

Lemma 12.6. (`Preimages of generators of a σ-algebra`) *Let Ω be a non-empty set and $X : \Omega \to \mathbb{R}$ be any mapping. Assume $\mathcal{G}$ is a σ-algebra of subsets of $\mathbb{R}$ and let $G_0 \subset \mathcal{G}$ be a collection of sets that generate $\mathcal{G}$, i.e. the smallest σ-algebra containing all sets of G_0 is $\mathcal{G}$. Then $\mathcal{X} := \{X^{-1}(A) : \ A \in G_0\}$ generates $\{X^{-1}(A) : \ A \in \mathcal{G}\}$.*

Proof. By Claim 12.2 and the given assumption that $\mathcal{G}$ is a σ-algebra we have that $\{X^{-1}(A) : \ A \in \mathcal{G}\}$ is a σ-algebra in Ω. Let $\sigma(\mathcal{X})$ be the σ-algebra generated by $\mathcal{X}$, i.e. the smallest σ-algebra in Ω that contains all sets of $\mathcal{X}$. Since $G_0 \subset \mathcal{G}$, we get that $\sigma(\mathcal{X}) \subset \{X^{-1}(A) : A \in \mathcal{G}\}$. Hence there exists a collection of sets G_1 such that $G_0 \subset G_1 \subset \mathcal{G}$ and $\sigma(\mathcal{X}) = \{X^{-1}(A) : A \in G_1\}$. Since $\sigma(\mathcal{X})$ is a σ-algebra, by Claim 12.1 we get that G_1 is also a σ-algebra. Since G_0 generates $\mathcal{G}$ we thus get that $G_1 = \mathcal{G}$ and hence $\sigma(\mathcal{X}) = \{X^{-1}(A) : A \in \mathcal{G}\}$. The proof is now complete. □

An important corollary of the above lemma is that the collection of sublevel sets $\{\{X \leq x\} : x \in \mathbb{R}\}$ of a mapping X generates $\{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}$, where $\mathcal{B}(\mathbb{R})$ is the Borel σ-algebra of $\mathbb{R}$.

In the spirit of Lemma 12.6 here we formulate and prove a similar result for the sequence of random variables. We will need such results later on.

Lemma 12.7. (Preimages of generators of a σ-algebra for mappings) *Let Ω be a non-empty set and $X_n : \Omega \to \mathbb{R}$ be any set of mappings where $n \in \mathcal{I}$ and the index set $\mathcal{I}$ is at most countable. Assume $\mathcal{G}$ is a σ-algebra of subsets of $\mathbb{R}$ and for each $n \in \mathcal{I}$ let $G_n \subset \mathcal{G}$ be a collection of sets that generate $\mathcal{G}$, i.e. the smallest σ-algebra containing all sets of G_n is $\mathcal{G}$. Then*[31]

$$\sigma\left(\bigcup_{n\in\mathcal{J}}\{X_n^{-1}(A) : \ A \in G_n\}\right) = \sigma\left(\bigcup_{n\in\mathcal{J}}\{X_n^{-1}(A) : \ A \in \mathcal{G}\}\right). \tag{12.2}$$

Proof. Clearly the left hand side is included in the right-hand side as $G_n \subset \mathcal{G}$. To see the other side of the inclusion, notice that the left-hand side contains the σ-algebra $\{X_n^{-1}(A) : \ A \in \mathcal{G}\}$ by Lemma 12.6. Hence it contains the union of these σ-algebras. But the minimal σ-algebra containing these union is the one on the right hand side of (12.2). This completes the proof. □

12.2. **Law and distribution of a random variable.** We saw in the previous section that there are σ-algebras which are naturally associated with random variables. Here we will work with *probability measures* and *distributions* that are generated by random variables. Recall that by $\mathcal{B}(\mathbb{R})$ we denote the Borel σ-algebra of $\mathbb{R}$.

Lemma 12.8. (Law of a random variable is a probability measure on $(\mathbb{R}, \mathcal{B})$) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X : \Omega \to \mathbb{R}$ be a random variable. Define $\mathbb{P}_X : \mathcal{B}(\mathbb{R}) \to \mathbb{R}$ as follows*

$$\mathbb{P}_X(B) := \mathbb{P}\left(\{\omega \in \Omega : \ X(\omega) \in B\}\right) \ \text{where } B \in \mathcal{B}(\mathbb{R}). \tag{12.3}$$

Then $\mathbb{P}_X$ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Proof. We need to verify conditions of Definition 1.2 for $\mathbb{P}_X$. The first two conditions of the definition are trivially satisfied. To verify the σ-additivity of $\mathbb{P}_X$ fix a collection $A_i \in \mathcal{B}(\mathbb{R})$, $i = 1, 2, ...$ of Borel sets such that $A_i \cap A_j = \emptyset$ for $i \neq j$. Then

$$\begin{aligned}
\mathbb{P}_X\left(\bigcup_i A_i\right) &= \mathbb{P}\left(\{\omega \in \Omega : \ X(\omega) \in \bigcup_i A_i\}\right) \\
&= \mathbb{P}\left(\bigcup_i \{\omega \in \Omega : \ X(\omega) \in A_i\}\right) \ \text{(since } X^{-1}(A_i) \text{ are disjoint)} \\
&= \sum_i \mathbb{P}\left(\{\omega \in \Omega : \ X(\omega) \in A_i\}\right) = \sum_i \mathbb{P}_X(A_i).
\end{aligned}$$

The proof is now complete. □

Some remarks on notation are in order.

- The measure $\mathbb{P}_X$ defined by (12.3) is called the **probability law** or the **distribution** of a random variable X. Notice also that $\mathbb{P}_X(A) = \mathbb{P}(X^{-1}(A))$.
- We will use a shorthand notation and for $A \in \mathcal{B}(\mathbb{R})$ will write $\mathbb{P}(X \in A)$ for the right-hand side of (12.3).

[31] Here, as usual, for a collection of sets $\mathcal{A}$ by $\sigma(\mathcal{A})$ we denote the smallest σ-algebra containing all sets in $\mathcal{A}$. More precisely, $\sigma(\mathcal{A})$ is a σ-algebra containing all sets in $\mathcal{A}$, and if $\mathcal{H}$ is any σ-algebra that contains all set of $\mathcal{A}$, then $\mathcal{H}$ must contain all elements of $\sigma(\mathcal{A})$ as well. Thus left and right sides of (12.2) are understood in this way.

In the spirit of Lemma 12.3 we now concentrate on sets of the form $\{X \leq x\}$ (sublevel sets).

Definition 12.2. (`Cumulative distribution function or CDF`) *Let $X : \Omega \to \mathbb{R}$ be a random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$. The function*

$$F(x) := \mathbb{P}(X \leq x) \text{ where } x \in \mathbb{R}, \tag{12.4}$$

is called the ***cumulative distribution function*** *(**CDF**) of X.*

Notice that we do not explicitly require the set $\{\omega \in \Omega : \ X(\omega) \leq x\} = X^{-1}((-\infty, x])$ to be from $\mathcal{F}$ so that $\mathbb{P}$ can be applied to it. It follows automatically from the fact that $(-\infty, x]$ is a Borel subset of $\mathbb{R}$ for any $x \in \mathbb{R}$ and that X is a random variable.

Theorem 12.9. (`Basic properties of CDF`) *Assume F is a cumulative distribution function of some random variable X as defined by (12.4). Then F satisfies the following*

(a) *F is non-decreasing on $\mathbb{R}$,*

(b) $\lim_{x \to +\infty} F(x) = 1$ *and* $\lim_{x \to -\infty} F(x) = 0$,

(c) *F is right-continuous, i.e.* $\lim_{y \to x+} F(y) = F(x)$ *for any $x \in \mathbb{R}$,*

(d) $\lim_{y \to x-} F(y) = \mathbb{P}(X < x)$ *for any $x \in \mathbb{R}$,*

(e) $\mathbb{P}(X = x) = F(x) - F(x-)$, *where $F(x-)$ is the limit of F at x from the left as defined in (d).*

Proof. (a). For any $x < y$ we have

$$\{\omega \in \Omega : \ X(\omega) \leq x\} \subset \{\omega \in \Omega : \ X(\omega) \leq y\},$$

hence the monotonicity of F follows directly from the monotonicity of the probability measure (see (1.2)).

(b). Take any sequence $x_n \uparrow +\infty$ and let $A_n := \{X \leq x_n\}$. Clearly $A_n \subset A_{n+1}$ for any $n = 1, 2, ...$ and $\bigcup_n A_n = \Omega$. Using Theorem 3.1 on continuity of the probability measure we get

$$1 = \mathbb{P}(\Omega) = \mathbb{P}\left(\bigcup_n A_n\right) = \lim_{n \to \infty} \mathbb{P}(A_n) = \lim_{n \to \infty} F(x_n).$$

To pass to any sequence x_n, without assuming it is increasing, notice that for any $x_n \to +\infty$, using the boundedness of F from above, we may pass to a subsequence, if necessary, to ensure that the limit $F(x_n)$ exists. Furthermore, since $x_n \to +\infty$ we may pass to a subsequence of x_n which is increasing, for which we know that the limit of F is 1. Hence for any sequence $\{x_n\}$ such that $x_n \to +\infty$, the limit of $F(x_n)$ exists and equals 1. Since we have the existence of the limit for any sequence converging to $+\infty$, we get that $\lim_{x \to +\infty} F(x) = 1$.

The case of the limit at $-\infty$ follows from a similar argument.

(c). Fix $x \in \mathbb{R}$ and let $\{y_n\}$ be a sequence of real numbers such that $y_n > x$ and $y_n \downarrow x$, i.e. y_n is decreasing and converges to x. Define $A_n := \{X \leq y_n\}$, then

$$\{X \leq x\} = \bigcap_{n=1}^{\infty} A_n.$$

Since y_n is decreasing, we get that $A_n \supset A_{n+1}$ for $n = 1, 2, ...$, hence using the continuity of probability measure (Theorem 3.1) we obtain

$$F(x) = \mathbb{P}(X \leq x) = \lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} F(y_n).$$

We get the convergence for decreasing sequences y_n. The requirement of y_n being decreasing can be removed by a similar approach as we had above in the proof of (b) and is left as an exercise. Thus the proof of (c) is complete.

(d). The proof here follows a similar argument as we had for (c). Fix $x \in \mathbb{R}$ and let $\{x_n\}$ be an increasing sequence converging to x. Then $\{X_n \leq x_n\}$ is an increasing sequence of events such that $\bigcup_{n=1}^{\infty} \{X_n \leq x_n\} = \{X < x\}$. By Theorem 3.1 (continuity of the probability measure) we have that

$$\mathbb{P}(X < x) = \lim_{n\to\infty} \mathbb{P}\left(\{X_n \leq x_n\}\right) = \lim_{n\to\infty} F(x_n).$$

This proves the claim of (d) for increasing sequences. Removing the restriction on $\{x_n\}$ being increasing follows a similar argument as we had above in (b). Thus, the proof is complete.

(e). For any $x \in \mathbb{R}$ we have $\{X \leq x\} = \{X = x\} \cup \{X < x\}$. Since the events in the union do not intersect, the equality in (e) follows from additivity of probability measure and the equality we proved in (d).

The proof of the theorem is now complete. □

Our next aim is the characterization of the set of all cumulative distribution functions (CDFs). As we show next, the first three conditions of Theorem 12.9 are necessary and sufficient for a function F to be a CDF of a random variable.

Theorem 12.10. `(Characterization of CDFs)` *Let $F : \mathbb{R} \to \mathbb{R}$ be a function that satisfies the following conditions*

(a) *F is non-decreasing on $\mathbb{R}$,*

(b) $\lim_{x\to+\infty} F(x) = 1$ *and* $\lim_{x\to-\infty} F(x) = 0$,

(c) *F is right-continuous, i.e.* $\lim_{y\to x+} F(y) = F(x)$ *for any $x \in \mathbb{R}$.*

Then, there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable $X : \Omega \to \mathbb{R}$ such that the CDF of X is F.

Proof. The proof is constructive. Take $\Omega = (0, 1)$, $\mathcal{F} = \mathcal{B}$ - the Borel σ-algebra on $(0, 1)$, and let $\mathbb{P}$ be the Lebesgue measure on $(0, 1)$. We have that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. Define[32]

$$X(\omega) := \sup\{y \in \mathbb{R} : \ F(y) < \omega\}, \quad \omega \in \Omega. \tag{12.5}$$

Thanks to conditions (a) and (b) we have that $0 \leq F(x) \leq 1$ for all $x \in \mathbb{R}$, and hence X takes real-values for all[33] $\omega \in \Omega$. We claim that $(\Omega, \mathcal{F}, \mathbb{P})$ and X satisfy the

[32]The intuition behind (12.5) is the following. Assume F is continuous and strictly increasing. Then it has an inverse $F^{-1} : (0, 1) \to \mathbb{R}$. It is easy to see that the function $X(\omega) = F^{-1}(\omega)$ where $\omega \in \Omega$ satisfies requirements of the Theorem. In general, we do not have invertibility of F, but can use the right-hand side of (12.5) as a substitute, in a certain sense, of the inverse of F.

[33]Notice that if $\lim_{x\to+\infty} F(x) = a < 1$, then $X(\omega) = +\infty$ for $a < \omega < 1$. Similarly, if $\lim_{x\to-\infty} F(x) = a > 0$, then $X(\omega) = -\infty$ for $0 < \omega < a$. Thus, conditions $\lim_{x\to-\infty} F(x) \leq 0$ and $\lim_{x\to+\infty} F(x) \geq 1$ guarantee that X defined by (12.5) is real-valued. The actual values of the limits of F at infinities are used elsewhere in the proof related to (12.6).

requirements of the Theorem. To prove that, it is enough to show that

$$\{\omega \in \Omega : \ X(\omega) \leq x\} = \{\omega : \ \omega \leq F(x)\}, \ \forall x \in \mathbb{R}. \tag{12.6}$$

Indeed, once we show (12.6) we immediately get that X is a random variable thanks to Lemma 12.3. Moreover, applying $\mathbb{P}$ on both sides of (12.6) and taking into account that in our construction $\mathbb{P}$ is the Lebesgue measure on $(0,1)$ and $0 \leq F \leq 1$ in view of (a) and (b), we get $\mathbb{P}(X \leq x) = F(x)$ which means that F is the CDF of X. We now proceed to the proof of (12.6).

Fix $x \in \mathbb{R}$ and let $\omega \in \Omega$ satisfy $X(\omega) \leq x$. We claim that $F(X(\omega)) \geq \omega$. Indeed, assume for contradiction that $F(X(\omega)) < \omega$. Using condition (c) - the right continuity of F, we have

$$F(X(\omega)) = \lim_{n\to\infty} F\left(X(\omega) + \frac{1}{n}\right) < \omega,$$

hence $F\left(X(\omega) + \frac{1}{n}\right) < \omega$ for $n \in \mathbb{N}$ large enough. The latter contradicts the fact that $X(\omega)$ is the supremum of $y \in \mathbb{R}$ with $F(y) < \omega$ and we conclude that $F(X(\omega)) \geq \omega$. Next, using the condition (a) - the monotonicity of F and the choice of x we get

$$\omega \leq F(X(\omega)) \leq F(x),$$

hence $\omega \in \{\omega : \ \omega \leq F(x)\}$, thus the left-hand side of (12.6) is included in the right-hand side. To see the other side of the inclusion, take ω such that $\omega \leq F(x)$. We will show that $X(\omega) \leq x$ which will complete the proof of the opposite inclusion in (12.6) and the theorem. Take any $y \in \mathbb{R}$ for which $F(y) < \omega$. Hence

$$F(y) < \omega \leq F(x),$$

and in view of the monotonicity of F we get $y \leq x$. Since $X(\omega)$ is the supremum over such y, we obtain $X(\omega) \leq x$, completing the proof of the theorem. □

Cumulative distribution functions (CDFs) give rise to a certain weak equivalence between random variables.

Definition 12.3. `(Equal in distribution)` *Given two random variables X and Y defined on $(\Omega, \mathcal{F}, \mathbb{P})$, we say that X and Y are **equal in distribution** and write $X \overset{d}{=} Y$, if*

$$\mathbb{P}(X \leq x) = \mathbb{P}(Y \leq x) \quad \text{for all } x \in \mathbb{R}.$$

Of special interest are random variables whose CDF is *absolutely continuous*, in which case the CDF becomes an integral of its derivative.

Definition 12.4. `(Probability density)` *Let F be the CDF of some random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$. If there exists a Lebesgue integrable function $\rho : \mathbb{R} \to \mathbb{R}$ such that*

$$F(x) = \int_{-\infty}^{x} \rho(t)dt, \quad \text{for all } x \in \mathbb{R},$$

*then ρ is called the **probability density** of X.*

Since F is non-decreasing, and its limit at $+\infty$ equals 1, we get that the probability density function, whenever it exists, is **non-negative** almost everywhere on $\mathbb{R}$, and its **integral over $\mathbb{R}$ equals** 1.

Example 1. (Uniform distribution) We say that a random variable X has a *uniform distribution* on $(0,1)$ and write $X \sim \mathrm{U}(0,1)$, if it has a density function ρ equal to

$$\rho(x) = \begin{cases} 1, & \text{if } x \in (0,1), \\ 0, & \text{if } x \notin (0,1). \end{cases}$$

Then, the CDF of X becomes

$$F(x) = \begin{cases} 0, & \text{if } x \leq 0, \\ x, & \text{if } x \in (0,1), \\ 1, & \text{if } x \geq 1. \end{cases}$$

In general, given an interval (a,b) we say that X is uniformly distributed in (a,b) if its density function equals $\frac{1}{b-a}\mathbb{I}_{[a,b]}$, where $\mathbb{I}$ is the characteristic function.

Example 2. (Exponential distribution) A random variable X is said to have exponential distribution with parameter $\lambda > 0$, if its density function equals

$$\rho(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{if } x < 0. \end{cases}$$

Integrating f we get that the CDF in this case equals

$$F(x) = \begin{cases} 0, & \text{if } x \leq 0, \\ 1 - e^{-x}, & \text{if } x > 0. \end{cases}$$

Example 3. (Standard normal or Gaussian distribution) This is arguably the most important probability distribution. We say that X has standard normal distribution and write $X \sim N(0,1)$, if its density function equals

$$\rho(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \qquad x \in \mathbb{R}.$$

There is no closed form for the CDF in this case.

Example 4. (Cantor distribution) Not all probability distributions have density. Here is an example. Let C be the Cantor set on $[0,1]$. Recall that C has the form

$$\mathrm{C} = [0,1] \setminus \bigcup_{n=0}^{\infty} \bigcup_{k=1}^{2^n} \Delta_{n,k},$$

where $\Delta_{n,k}$ is the k-th open interval removed from $[0,1]$ on step n. In particular, $\Delta_{0,1} = (\frac{1}{3}, \frac{2}{3})$, $\Delta_{1,1} = (\frac{1}{9}, \frac{2}{9})$ and $\Delta_{1,2} = (\frac{7}{9}, \frac{8}{9})$. Define a function $F : [0,1] \to [0,1]$ as follows. Set $F(0) = 0$, $F(1) = 1$ and $F(x) = 1/2$ on $\Delta_{0,1}$ - the interval that is dropped on the first step of constructing C. Next, define F on $\Delta_{1,1}$ and $\Delta_{1,2}$ as a constant equal to the average of the values of F at the neighboring points to each interval, where F is already defined. In particular, we will have $F = \frac{1}{4}$ on $\Delta_{1,1}$ and $F = \frac{3}{4}$ on $\Delta_{1,2}$. With this process we will have F defined on the complement of C. Then, for each $x \in \mathrm{C}$ we define

$$F(x) := \sup_{\substack{y < x \\ y \in [0,1] \setminus \mathrm{C}}} F(y).$$

Extending F to $(-\infty, 0]$ as 0 and to $[1, +\infty)$ as 1, we get a function satisfying all requirements of Theorem 12.10, hence F is a CDF of some random variable. However,

F has no density function. Indeed, if there is a function ρ such that $F(x) = \int_{-\infty}^{x} \rho(t)dt$, then $F'(x) = \rho(x)$ almost everywhere. However, by definition of F we have $F' = 0$ for a.e. on $\mathbb{R}$, which will imply that $\rho = 0$ a.e. on $\mathbb{R}$, thus $F = 0$ everywhere, which is a contradiction.

Example 5. (Discrete measures) A probability measure $\mathbb{P}$ is called *discrete*, if there exists an element S of $\mathcal{F}$ which is countable and such that $\mathbb{P}(\Omega \setminus S) = 0$. For example, consider a CDF given by

$$F(x) = \begin{cases} 1, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

As another example, let $q_1, q_2, ...$ be the set of rational numbers in some enumeration, and let $\alpha_i > 0$ be such that $\sum_i \alpha_i = 1$. Now define

$$F(x) = \sum_{i=1}^{\infty} \alpha_i \mathbb{I}_{[q_i, +\infty)}(x), \qquad x \in \mathbb{R}.$$

13. Borel-Cantelli lemmas

In a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ let $E_i \in \mathcal{F}$, $i = 1, 2, ...$, be a sequence of events. Consider a scenario where E_i describes the outcome of some experiment: for example, we are tossing a coin, and E_i is the event that the i-th toss is heads (H). A natural question in this setting is to understand if *infinitely many* of these experiments result in success. In our example with a coin toss, one may be interested in the probability of getting heads *infinitely many times*. A different but related question is whether all experiments are *eventually successful*, which in terms of our example with coin tosses means if we get only heads (H) starting form some point in the sequence. We now formally define the events corresponding to the discussion above as follow

$$\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} E_m =: \limsup E_n =: \{E_n \text{ i.o.}\}, \tag{13.1}$$

$$\bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} E_m =: \liminf E_n =: \{E_n \text{ ev.}\}, \tag{13.2}$$

where **i.o.** in (13.1) and **ev.** in (13.2) stand for **infinitely often** and **eventually**. We have the following properties:

- $\limsup E_n,\ \liminf E_n \in \mathcal{F}$, thanks to the fact that $\mathcal{F}$ is a σ-algebra,
- $\liminf E_n \subset \limsup E_n$ by a direct inspection, indeed by de Morgan's law

$$(\limsup E_n)^c = \left(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} E_m \right)^c = \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} E_m^c = \liminf E_m^c, \tag{13.3}$$

Lemma 13.1. `(Borel-Cantelli's first lemma)` *Assume* $\sum_{n=1}^{\infty} \mathbb{P}(E_n) < \infty$. *Then* $\mathbb{P}(\limsup E_n) = 0$.

Proof. We have $\limsup E_n = \bigcap_{n=1}^{\infty} A_n$, where $A_n = \bigcup_{m=n}^{\infty} E_m$. Clearly, $A_n \supset A_{n+1}$ for $n = 1, 2, ...$, i.e. the sequence of events $\{A_n\}$ is decreasing. Using Theorem 3.1 (the continuity of measure) we get

$$\mathbb{P}(\limsup E_n) = \lim_{n\to\infty} \mathbb{P}(A_n) \leq \lim_{n\to\infty} \sum_{m=n}^{\infty} \mathbb{P}(E_m),$$

where we used Boole's inequality (Theorem 3.2) to bound $\mathbb{P}(A_n)$ from above. The last limit is 0 as a tail of convergent series. The proof is now complete. □

Lemma 13.2. (Borel-Cantelli's second lemma) *Let $\{E_n\}$ be a sequence of independent events and assume that $\sum_{n=1}^{\infty} \mathbb{P}(E_n) = \infty$. Then $\mathbb{P}(\limsup E_n) = 1$.*

Proof. By (13.3) we have

$$(\limsup E_n)^c = \bigcup_{n=1}^{\infty} A_n,$$

where $A_n := \bigcap_{m=n}^{\infty} E_m^c$ is a sequence of increasing events. Thus,

$$\begin{aligned}
&\mathbb{P}\left(\bigcap_{m=n}^{\infty} E_m^c\right) \quad \text{(by Theorem 3.1)} \\
&= \lim_{r\to\infty} \mathbb{P}\left(\bigcap_{m=n}^{r} E_m^c\right) \quad \text{(by independence)} \\
&= \lim_{r\to\infty} \prod_{m=n}^{r} (1 - \mathbb{P}(E_m)) \quad \text{(by inequality } 1 - x \leq e^{-x}) \\
&= \lim_{r\to\infty} \prod_{m=n}^{r} e^{-\mathbb{P}(E_m)} = \lim_{r\to\infty} e^{-\sum_{m=n}^{r} \mathbb{P}(E_m)} = 0,
\end{aligned}$$

where the last limit is 0 in view of the divergence of the series $\sum_n \mathbb{P}(E_n)$. We thus get that $\mathbb{P}(A_n) = 0$ hence, by Boole's inequality (Theorem 3.2) we have

$$\mathbb{P}\left((\limsup E_n)^c\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n) = 0,$$

and hence $\mathbb{P}(\limsup E_n) = 1$. The proof is complete. □

Notice that we cannot eliminate the **independence condition**. Indeed, let $E \in \mathcal{F}$ be such that $0 < \mathbb{P}(E) < 1$. Take $E_n = E$ for all $n \geq 1$. Then $\sum_n \mathbb{P}(E_n) = \infty$ but $\mathbb{P}(E_n \text{ i. o.}) = \mathbb{P}(E) < 1$.

A more non-trivial example on independence condition is the following. Take a random variable $X \sim \mathrm{U}[0, 1]$ and let $E_n := \{X \leq 1/n\}$, $n \in \mathbb{N}$. Then $\mathbb{P}(E_n) = 1/n$ hence the sum $\sum_n \mathbb{P}(E_n) = \infty$ (divergence of harmonic series). However, $\mathbb{P}(E_n \text{ i. o.}) = \{X = 0\}$ which is an event of probability 0.

There are, however, results on relaxing the independence condition in Borel-Cantelli's second lemma, but they are out of scope of these notes.

While simple in their formulations and proofs, the Borel-Cantelli lemmas are very helpful on many occasions. We will now see them in action on a nice example which is commonly referred to as **Monkey typing Shakespeare**.

> Assume a monkey types letters A-Z of the English alphabet uniformly at random. What is the probability that in this process the monkey will produce all works of Shakespeare[34] put together?

To address the question above we need to formulate it as a mathematical problem. Let $\mathcal{A} = \{a_1, a_2, ..., a_n\}$ be a finite set of distinct symbols (alphabet). To model the typing process described above consider a sequence of independent random variables $X_1, X_2, ...$ all with uniform distribution on $\mathcal{A}$, i.e.

$$\mathbb{P}(X_i = a) = \frac{1}{|\mathcal{A}|} \text{ for any } a \in \mathcal{A} \text{ and all } i = 1, 2, ...$$

Fix $\vec{w} = (c_1, c_2, ..., c_k)$, a sequence (word) made of symbols of $\mathcal{A}$. In a process of typing, getting the word $\vec{w}$ straight away means that $(X_1, ..., X_k) = \vec{w}$.

Lemma 13.3. *Let $\mathcal{A}$, $\vec{w}$ and the sequence $\{X_i\}$ be as above. For $n \in \mathbb{N}$ define*

$$A_n = \{\omega : \big(X_n(\omega), ..., X_{n+|\vec{w}|-1}(\omega)\big) = \vec{w}\},$$

i.e. the event of getting the word $\vec{w}$ starting from step n. Then

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

Proof. Recall that k is the length of the word $\vec{w}$. Thanks to the independence of the sequence $\{X_i\}$, we get that the events

$$A_1, A_{k+1}, ..., A_{ik+1}, ..., \text{ where } i = 1, 2, ...,$$

are independent. Again, in view of independence of $\{X_i\}$ we have

$$\mathbb{P}(A_n) = \prod_{i=1}^{k} \mathbb{P}(X_{n+i-1} = c_i) = \left(\frac{1}{|\mathcal{A}|}\right)^k.$$

From here we get that $\sum_i \mathbb{P}(A_{ik+1}) = \infty$ and hence Borel-Cantelli's second lemma (see 13.2) settles the the proof. □

14. Sequence of random variables

In this section we study convergence of a sequence of random variables. We introduce different types (modes) of convergence and study relations between them.

Definition 14.1. `(Convergence of random variables)` *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, X_n, $n = 1, 2, ...$ be a sequence of random variables and let X be another random variable all defined on this probability space. We say that*

- *X_n converges to X **almost surely**, and write $X_n \to X$ a.s., if*

$$\mathbb{P}\left(\{\omega \in \Omega : \ X_n(\omega) \to X(\omega)\}\right) = 1. \tag{14.1}$$

[34] Of course, for the sake of the example we consider texts that are capitalized and any symbol that is not a letter of the English alphabet is dropped.

- *X_n converges to X* ***in probability****, and write $X_n \xrightarrow{P} X$ if for any $\varepsilon > 0$*
$$\mathbb{P}(|X_n - X| \geq \varepsilon) \to 0 \text{ as } n \to \infty, \tag{14.2}$$

- *X_n converges to X* ***in distribution****, and write $X_n \xrightarrow{d} X$ if*
$$F_n(x) \to F(x) \text{ as } n \to \infty, \text{ for every } x \in \mathbb{R} \text{ where } F \text{ is continuous}, \tag{14.3}$$
where F_n is the CDF of X_n and F is the CDF of X.

The notions of *almost sure convergence* and the *convergence in probability* are equivalent to the notions of *convergence almost everywhere* and *convergence in measure* (*weak convergence*) correspondingly for ordinary functions. Here we will establish several connections between these different modes of convergence. Let us also emphasize that the types of convergence discussed above are not exhaustive of course: there are many other ways of defining convergence of random variables (for example in L^p spaces which we will encounter later). Here we study arguably the most basic forms of convergence.

Before moving forward let us clarify why in the definition of *convergence in distribution* we consider only points of continuity for F.

Example 14.0.1. `(Only points of continuity for convergence in distribution)` *Consider a sequence of random variables $\{X_n\}_{n=1}^{\infty}$ defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where each $X_n \sim U(0, 1/n)$, i.e. X_n is uniformly distributed on $(0, 1/n)$, $n = 1, 2, \dots$. Denote the CDF of X_n by F_n. By definition of the uniform distribution we have*

$$F_n(x) = \begin{cases} 0, & x \leq 0, \\ nx, & x \in \left(0, \frac{1}{n}\right), \\ 1, & x \geq \frac{1}{n}. \end{cases}$$

Clearly

$$F_n(x) \to \begin{cases} 0, & x \leq 0, \\ 1, & x > 0, \end{cases}$$

and $X_n \xrightarrow{P} 0$, however, the point-wise limit of F_n, as we see above, is not a cumulative distribution function of any random variable (the limit is not right-continuous at 0). Nevertheless, if F is the CDF of a random variable X with $\mathbb{P}(X = 0) = 1$ (the limit of X_n in probability), we still have have that $\lim_{n\to\infty} F_n(x) = F(x)$ at points $x \in \mathbb{R}$ where F is continuous.

Lemma 14.1. (Convergence **almost surely** implies convergence **in probability**) *Keeping the notation of Definition 14.1, if $X_n \to X$ a.s. then $X_n \xrightarrow{P} X$.*

Proof. Fix any $\varepsilon > 0$ and let $\omega \in \Omega$ be such that $X_n(\omega) \to X(\omega)$ as $n \to \infty$. For such $\omega \in \Omega$ there exists $N = N(\omega) \in \mathbb{N}$ such that $|X_n(\omega) - X(\omega)| < \varepsilon$ for all $n \geq N(\omega)$, or equivalently

$$\omega \in \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \{|X_n(\omega) - X(\omega)| < \varepsilon\}.$$

Since $X_n(\omega) \to X(\omega)$ with probability 1, it follows that

$$\mathbb{P}\left(\bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \{|X_n(\omega) - X(\omega)| < \varepsilon\} \right) = 1, \quad \text{for any } \varepsilon > 0.$$

Taking the complements and using de Morgan's rule we get that

$$\mathbb{P}\left(\bigcap_{N=1}^{\infty}\bigcup_{n=N}^{\infty}\{|X_n(\omega)-X(\omega)|\geq\varepsilon\}\right)=0.$$

For $N\in\mathbb{N}$ set $A_N:=\bigcup_{n=N}^{\infty}\{|X_n(\omega)-X(\omega)|\geq\varepsilon\}$. Clearly we have $A_N\supset A_{N+1}$, i.e. the events $\{A_N\}$ are decreasing, hence by the continuity of measure (Theorem 3.1) and the last equality we get $\mathbb{P}(A_N)\to 0$. Since $\{|X_N(\omega)-X(\omega)|\geq\varepsilon\}\subset A_N$ it follows that

$$\mathbb{P}\left(\{|X_N(\omega)-X(\omega)|\geq\varepsilon\}\right)\leq\mathbb{P}(A_N)\to 0,$$

as $N\to\infty$ completing the proof. □

Example 14.1.1. `(Convergence in probability does not imply convergence almost surely)` *Consider a probability space $(\Omega,\mathcal{F},\mathbb{P})$ where $\Omega=(0,1)$, $\mathcal{F}$ is the Borel σ-algebra on Ω and $\mathbb{P}$ is the Lebesgue measure on Ω. For $k=0,1,2,...$ and $0\leq i\leq 2^k-1$ consider the interval $\Delta_k^i:=\left(\frac{i}{2^k},\frac{i+1}{2^k}\right)$. Each $n\in\mathbb{N}$ can be represented uniquely as $n=2^k+i$ with $0\leq i\leq 2^k-1$. Given such representation define*

$$X_n:=X_k^i:=\begin{cases}1, & x\in\Delta_k^i,\\ 0, & x\notin\Delta_k^i.\end{cases}$$

Then $X_n\xrightarrow{P}0$ but X_n does not converge to 0 almost surely, as for all $\omega\in\Omega$ that are not of the form $i/2^k$ there are infinitely many $n\in\mathbb{N}$ for which $X_n(\omega)=1$. Indeed, for each k the intervals Δ_k^i with $i=0,1,...,2^k-1$ cover Ω, hence also the point ω, therefore the index $n=2^k+i$ for which Δ_n contains ω will have $X_n(\omega)=1$.

Lemma 14.2. (Convergence **in probability** implies convergence **in distribution**) *Keeping the notation of Definition 14.1, if $X_n\xrightarrow{P}X$ then $X_n\xrightarrow{d}X$.*

Proof. Let F be the CDF of X and let F_n be the CDF of X_n for $n=1,2...$. Then, for any $\varepsilon>0$ and any $x\in\mathbb{R}$ we have

$$\begin{aligned}F_n(x)&=\mathbb{P}(X_n\leq x)\\&=\mathbb{P}(X_n\leq x,X\leq x+\varepsilon)+\mathbb{P}(X_n\leq x,X>x+\varepsilon)\\&\leq\mathbb{P}(X\leq x+\varepsilon)+\mathbb{P}\left(|X_n-X|\geq\varepsilon\right)\\&=F(x+\varepsilon)+\mathbb{P}\left(|X_n-X|\geq\varepsilon\right).\end{aligned}$$

Passing to the limit in the last inequality and using the fact that $X_n\xrightarrow{P}X$ we get

$$\limsup_{n\to\infty}F_n(x)\leq F(x+\varepsilon). \tag{14.4}$$

We now establish a bound from below.

$$\begin{aligned}F(x-\varepsilon)&=\mathbb{P}(X\leq x-\varepsilon)\\&=\mathbb{P}(X\leq x-\varepsilon,X_n\leq x)+\mathbb{P}(X\leq x-\varepsilon,X_n>x)\\&\leq\mathbb{P}(X_n\leq x)+\mathbb{P}\left(|X_n-X|\geq\varepsilon\right)\\&=F_n(x)+\mathbb{P}\left(|X_n-X|\geq\varepsilon\right).\end{aligned}$$

We again pass to the limit in the last inequality and using the fact that X_n converges to X in probability, arrive at

$$F(x-\varepsilon)\leq\liminf_{n\to\infty}F_n(x). \tag{14.5}$$

Combining (14.4) and (14.5) leads to

$$F(x-\varepsilon) \leq \liminf_{n\to\infty} F_n(x) \leq \limsup_{n\to\infty} F_n(x) \leq F(x+\varepsilon).$$

Now if x is a point of continuity for F, we may pass to the limit in the last inequality as $\varepsilon \to 0$. Since both sides converge to the same limit which equals $F(x)$ thanks to continuity, it follows that the limit $\lim_{n\to\infty} F_n(x)$ exists and is equal to $F(x)$. The proof is complete. □

Example 14.2.1. `(Convergence in distribution does not imply convergence in probability)` *Consider a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *where* $\Omega = \{0,1\}$, $\mathcal{F} = 2^{\Omega}$ *and* $\mathbb{P}(0) = \mathbb{P}(1) = 1/2$. *Define* $X_n : \Omega \to \mathbb{R}$ *as* $X_n(0) = 0$, $X_n(1) = 1$ *for all* $n \in \mathbb{N}$. *Then* $X_n \xrightarrow{d} Y$ *where* $Y(0) = 1$ *and* $Y(1) = 0$, *but* $|X_n - Y| = 1$ *on* Ω *thus* X_n *does not converge to* Y *in probability.*

Lemma 14.3. (**Almost surely** convergent subsequence from **convergence in probability**) *Assume* $X_n \xrightarrow{P} X$. *Then, there exists an increasing sequence* $\{n_k\}$ *such that* $X_{n_k} \to X$ *almost surely.*

Proof. Since $X_n \to X$ in probability, by (14.2) for any $j \in \mathbb{N}$ there exists $k_j \in \mathbb{N}$ large enough such that

$$\mathbb{P}\left(|X_{k_j} - X| \geq \frac{1}{j}\right) \leq \frac{1}{j^2}.$$

Set $A_j := \{|X_{k_j} - X| \geq 1/j\}$. We have $\sum_{j=1}^{\infty} \mathbb{P}(A_j) < \infty$, hence by Borel-Cantelli's first lemma 13.1 we get $\mathbb{P}(\limsup A_j) = 0$. From here and (13.3) we have

$$\mathbb{P}(\liminf A_j^c) = 1,$$

hence with probability 1 for $\omega \in \Omega$ there exists $J = J(\omega) \in \mathbb{N}$ such that $\omega \in \bigcap_{j=J(\omega)}^{\infty} A_j^c$, meaning that

$$|X_{k_j}(\omega) - X(\omega)| < \frac{1}{j}, \text{ for all } j \geq J(\omega).$$

The latter means that $\lim_{j\to\infty} X_{k_j}(\omega) = X(\omega)$. The proof of the lemma is complete. □

15. Independence of random variables

This section is devoted to the concept of *independence of random variables*. Recall that earlier in the notes in Section 7 we defined and studied independence for discrete random variables. Here, we will proceed in the general case, without assuming random variables being *discrete*. This setup will also cover the case of discrete random variables.

15.1. **Basic definitions and independence via π-systems.** Here we provide basic definitions of independence and prove a useful equivalent criteria for independence.

Definition 15.1. `(Independence of σ-algebras)` *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and let* $\mathcal{G}, \mathcal{H} \subset \mathcal{F}$ *be two sub-*σ*-algebras of* $\mathcal{F}$. *We say that* $\mathcal{G}$ *and* $\mathcal{H}$ *are independent, if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B), \quad \text{for any } A \in \mathcal{G} \text{ and } B \in \mathcal{H}.$$

Similarly, sub-σ*-algebras* $\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_n \subset \mathcal{F}$, $n \geq 2$, *are called independent if*

$$\mathbb{P}(A_1 \cap A_2 \cap ... \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2) \cdot ... \cdot \mathbb{P}(A_n),$$

for any events $A_i \in \mathcal{G}_i$, where $i = 1, 2, ..., n$.

For a mapping $X : \Omega \to \mathbb{R}$ in (12.1) we defined $\sigma(X) = \{X^{-1}(B) : \ B \in \mathcal{B}(\mathbb{R})\}$ and proved that $\sigma(X)$ is a σ-algebra itself. Independence between random variables will be defined via their generated σ-algebras.

Definition 15.2. (`Independence of random variables`) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X_1, X_2, ..., X_n : \Omega \to \mathbb{R}$ be random variables. Then, these random variables are called **independent** if the σ-algebras generated by them, namely $\sigma(X_1), \sigma(X_2), ..., \sigma(X_n)$ are independent in the sense of Definition 15.1.*

□ Exercise 15.0.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X, Y : \Omega \to \mathbb{R}$ be **discrete** random variables. Prove that Definition 15.2 is equivalent to Definition 7.2.*

Assume we have random variables X and Y that are independent. It should be intuitively clear that X^2 and e^Y (say) are also independent since we did not change the source of information in the new random variables. Our next lemma settles this question formally. Similarly to Theorem 7.12 for discrete random variables, here as well we have that independence is preserved under compositions with a large class of functions.

Lemma 15.1. (`Composition of Borel functions with independent random variables preserves independence`) *Let $X_1, ..., X_n$ $(n \geq 2)$ be independent random variables, and let $g_1, ..., g_n : \mathbb{R} \to \mathbb{R}$ be Borel(-measurable) functions, i.e. $g_i^{-1}(B) \in \mathcal{B}$ for any $B \in \mathcal{B}(\mathbb{R})$ for all $i = 1, 2, ..., n$. Then $g_1(X_1), ..., g_n(X_n)$ are also independent.*

Proof. By definition of independence we need to show the independence of σ-algebras $\sigma(g_1(X_1)), ..., \sigma(g_n(X_n))$. To this end take $B_i \in \mathcal{B}(\mathbb{R})$ for $i = 1, 2, ..., n$. Then

$$\begin{aligned}
&\mathbb{P}\left(g_1(X_1) \in B_1, ..., g_n(X_n) \in B_n\right) \\
=&\mathbb{P}\left(X_1 \in g_1^{-1}(B_1), ..., X_n \in g_n^{-1}(B_n)\right) \text{ (since } g_i \text{ is Borel and } X_i\text{-s are independent)} \\
=&\mathbb{P}\left(X_1 \in g_1^{-1}(B_1)\right) \cdot ... \cdot \mathbb{P}\left(X_n \in g_n^{-1}(B_n)\right) \\
=&\mathbb{P}\left(g_1(X_1) \in B_1\right) \cdot ... \cdot \mathbb{P}\left(g_n(X_n) \in B_n\right),
\end{aligned}$$

which competes the proof. □

□ Exercise 15.1.1. *Prove that any continuous function $f : \mathbb{R}^n \to \mathbb{R}$ $(n \geq 1)$ is Borel, i.e. for any Borel set $B \in \mathcal{B}(\mathbb{R})$ the preimage $f^{-1}(B)$ is Borel in $\mathbb{R}^n$.*

□ Exercise 15.1.2. *Let $X_1, X_2, ..., X_n$ $(n \geq 2)$ be independent and let $g : \mathbb{R}^{n-1} \to \mathbb{R}$ be a Borel function, i.e. for any Borel set $B \in \mathcal{B}(\mathbb{R})$ the preimage $g^{-1}(B)$ is a Borel set in $\mathbb{R}^{n-1}$. Prove that $g(X_1, ..., X_{n-1})$ and X_n are also independent.*

We will next establish a criteria for independence based on *sublevel sets* of random variables. To do this we need some results from measure theory. Let S be a non-empty set and let $I \subset 2^S$ be a collection of subsets of S. We say that S is a **π-system** if it is closed under **finite intersections**, i.e. if $A, B \in S$ then $A \cap B \in S$. Examples of π-systems include any σ-algebra, the collection of sets $\{(-\infty, x] : x \in \mathbb{R}\}$, the collection of sets $\{(-\infty, a] \times (-\infty, b] : \ a, b \in \mathbb{R}\}$, any topology, the family of sets that can be represented as a Cartesian product, namely $\{A_1 \times ... \times A_n : \ A_i \subset \mathbb{R}, \ i = 1, ..., n\}$, etc. The morale here is that σ-algebras might be hard to work with but π-systems usually have an easier structure.

Lemma 15.2. (Uniqueness lemma for a π-system generating σ-algebra) *Given a non-empty set S and a π-system I in S, let $\mathcal{G} := \sigma(I)$ be the smallest σ-algebra containing I (σ-algebra generated by I). Assume μ_1 and μ_2 are two measures on $(S, \mathcal{G})$ such that $\mu_1(S) = \mu_2(S) < \infty$ and $\mu_1 = \mu_2$ on I, i.e. $\mu_1(A) = \mu_2(A)$ for any $A \in I$. Then $\mu_1 = \mu_2$ on $\mathcal{G}$.*

For the proof of this lemma see [9, Chapter A1] or [2, Section 2.1.1]. It uses another measure-theoretical construct called λ-*systems* (or *d-systems*) and is based on Dynkin's $\pi - \lambda$ theorem.

The core idea of Lemma 15.2, which we will use here, is that probability measures which coincide on a certain π-system will coincide on the minimal σ-algebra containing that system.

Lemma 15.3. (Independent π-systems generate independent σ-algebras) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let I and J be two π-systems of sets in $\mathcal{F}$. Let also $\mathcal{G} = \sigma(I)$ and $\mathcal{H} = \sigma(J)$ be the minimal σ-algebras generated by I and J correspondingly. If I and J are independent, then so are $\mathcal{G}$ and $\mathcal{H}$.*

Proof. Fix any $A_0 \in I$ and define a measure μ_1 by

$$\mu_1(B) := \mathbb{P}(A_0 \cap B), \quad B \in \mathcal{H}.$$

Similarly define another measure μ_2 by

$$\mu_2(B) := \mathbb{P}(A_0)\mathbb{P}(B), \quad B \in \mathcal{H}.$$

Clearly $\mu_1(\Omega) = \mu_2(\Omega) = \mathbb{P}(A_0) < \infty$. Also, by assumption I and J are independent, hence $\mu_1(B) = \mu_2(B)$ if $B \in J$. Thus we can apply Lemma 15.2 which implies that $\mu_1 = \mu_2$ on $\mathcal{H}$. More explicitly this means

$$\mathbb{P}(A_0 \cap B) = \mathbb{P}(A_0)\mathbb{P}(B) \text{ for any } A_0 \in I \text{ and } B \in \mathcal{H}.$$

It is only left to show that I can be extended to $\sigma(I) = \mathcal{G}$ in the last equality. To see that fix some $B_0 \in \mathcal{H}$ and define new measures μ_1 and μ_2 by

$$\mu_1(A) = \mathbb{P}(A \cap B_0), \ \mu_2(A) = \mathbb{P}(A)\mathbb{P}(B_0), \quad A \in \mathcal{G}.$$

Again, μ_1 and μ_2 coincide on I and $\mu_1(\Omega) = \mu_2(\Omega) = \mathbb{P}(B_0) < \infty$ hence by Lemma 15.2 they coincide on $\sigma(I) = \mathcal{G}$. The proof is now complete. □

An important corollary of the last lemma is the following.

Corollary 15.4. (Independence of random variables via factoring of CDFs) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X, Y : \Omega \to \mathbb{R}$ be two random variables. Then X and Y are independent if and only if*

$$\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x)\mathbb{P}(Y \leq y) \quad \text{for all } x, y \in \mathbb{R}.$$

Proof. We only need to prove the sufficiency of the claim as the necessity follows by definition and the fact that $\{X \leq x\}$ is a preimage of a Borel set $(-\infty, x]$.

Define $I := \{\{X \leq x\} : \ x \in \mathbb{R}\}$. Clearly I is a π-system in $(\Omega, \mathcal{F})$. Also, since $\{X \leq x\} = X^{-1}((-\infty, x])$ and the collection of sets $\{(-\infty, x] : \ x \in \mathbb{R}\}$ generates Borel σ-algebra $\mathcal{B}(\mathbb{R})$, then by Claim 12.6 the minimal σ-algebra containing I coincides with $\sigma(X)$. Similarly, $J = \{\{Y \leq y\} : \ y \in \mathbb{R}\}$ is a π-system that generates $\sigma(Y)$. The assumption of the corollary implies that I and J are independent. Hence by Lemma 15.3 we get that $\sigma(X)$ and $\sigma(Y)$ are independent thus so are X and Y by definition. This completes the proof. □

Lemma 15.5. (*$n \geq 2$ independent π-systems generate independent σ-algebras*) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $I_1, I_2, ..., I_n$ be π-systems of sets in $\mathcal{F}$ such that $\Omega \in I_k$ for all $k = 1, 2, ..., n$ and*

$$\mathbb{P}(A_1 \cap A_2 \cap ... \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2) \cdot ... \cdot \mathbb{P}(A_n), \tag{15.1}$$

for any $A_1 \in I_1, A_2 \in I_2, ..., A_n \in I_n$. For each $k = 1, 2, ..., n$ let $\mathcal{G}_k := \sigma(I_k)$ be the σ-algebra generated by I_k. Then $\mathcal{G}_1, ..., \mathcal{G}_n$ are independent.

Proof. We need to prove that each I_k used in (15.1) can be replaced by $\mathcal{G}_k$, i.e. by the σ-algebra they generate. To achieve this we will show that for any list of π-systems, say $J_1, J_2, ..., J_n$, that satisfy (15.1) and each contain Ω, the first item in the list, in this case J_1, can be replaced by the minimal σ-algebra containing it. This will complete the proof of the lemma since we can repeatedly apply this argument to the following lists of π-systems (each with n items):

$$[I_1, I_2, ..., I_n] \mapsto [I_2, I_3, ..., I_n, \mathcal{G}_1] \mapsto [I_3, I_4, ..., I_n, \mathcal{G}_1, \mathcal{G}_2] \mapsto ... \mapsto [I_n, \mathcal{G}_1, ..., \mathcal{G}_{n-1}],$$

each time replacing the first item by the minimal σ-algebra it generates and cyclically moving it to the end of the list.

We now prove that I_1 can be extended to $\mathcal{G}_1$ which will complete the proof as described above. Fix $A_2 \in I_2, ..., A_n \in I_n$, and for $A \in \mathcal{G}_1$ consider set functions

$$\mu(A) := \mathbb{P}(A \cap A_2 \cap ... \cap A_n),$$

and

$$\nu(A) := \mathbb{P}(A)\mathbb{P}(A_2) \cdot ... \cdot \mathbb{P}(A_n).$$

It is easy to see that both μ and ν are measures on $\mathcal{G}_1$. Thanks to (15.1) we have that μ and ν coincide on I_1. Now this, coupled with $\mu(\Omega) = \nu(\Omega)$ which is due to $\Omega \in I_1$ and (15.1), allows us to apply Lemma 15.2 (uniqueness) and conclude that $\mu = \nu$ on $\mathcal{G}_1$. The proof of the lemma is now complete. □

15.2. **Joint and marginal distributions.** Here we generalize the concept of CDF (cumulative distribution function) and probability density defined for a single random variable to $n \geq 2$ random variables and introduce a criteria for independence based on CDFs and probability densities of a finite collection of random variables.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X_1, X_2, ..., X_n : \Omega \to \mathbb{R}$ be random variables defined on it. Similarly to the case of a single random variable here as well we can define a CDF (cumulative distribution function) for the random vector $\mathbf{X} := (X_1, ..., X_n)$. Indeed, for $x_1, ..., x_n \in \mathbb{R}$ set

$$F_{\mathbf{X}}(\mathbf{x}) := F(x_1, ..., x_n) := \mathbb{P}\left(X_1 \leq x_1, ..., X_n \leq x_n\right), \tag{15.2}$$

and call F the **joint CDF** of $X_1, ..., X_n$. Throughout the section we will be using the boldface notation $\mathbf{X}$ and $\mathbf{x}$ to indicate that we are dealing with multidimensional entities instead.

We can easily recover the CDFs of single random variables X_i from the joint CDF. Just as in the case of $n = 1$ here too a similar argument as we had in the proof of Theorem 12.9, implies that

$$\lim_{\mathbf{x} \to +\infty} F_{\mathbf{X}}(\mathbf{x}) = 1,$$

where by $\mathbf{x} \to +\infty$ we mean that all coordinates of $\mathbf{x} = (x_1, ..., x_n)$ tend to $+\infty$. Moreover, for any index set $I \subset \{1, 2, ..., n\}$ that is non-empty and does not coincide

with $\{1, ..., n\}$ we can prove, again similarly to the case of $n = 1$, that

$$\lim_{\substack{x_i \to +\infty \\ i \in I}} F(x_1, ..., x_n) = \mathbb{P}\left(\bigcap_{j \in \{1,...,n\} \setminus I} \{X_j \leq x_j\} \right), \tag{15.3}$$

i.e. when passing to the limit in (15.2) with respect to some collection of coordinates, then that collection of random variables becomes excluded from the intersection in the right-hand side of (15.2). For example, if as an index set we take $I = \{2, ..., n\}$ then the limit in (15.3) recovers the CDF of X_1. In general, the right-hand side of (15.3) which cover the CDFs for subcollection of $X_1, ..., X_n$ (including single X_i) is called (joint) **marginal CDF**. What is important here to note is that the value of such marginal CDFs can be recovered from the joint CDF of $\mathbf{X}$ defined in (15.2).

□ Exercise 15.5.1. *Let F be the joint CDF of $X_1, ..., X_n$ as defined above. Using the approach of Theorem 12.9 show that*

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

where $\boldsymbol{x} \to \pm\infty$ means that all coordinates of $\boldsymbol{x} = (x_1, ..., x_n) \in \mathbb{R}^n$ converge to $\pm\infty$.

Similarly to what we had above for two random variables, factoring of the joint CDF into a product of individual CDFs is equivalent to independence of $n \geq 2$ random variables as we prove next.

Lemma 15.6. (`Independence of` n`-random variables via factoring CDFs`) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X_1, X_2, ..., X_n : \Omega \to \mathbb{R}$ be random variables. Then $X_1, X_2, ..., X_n$ are independent if and only if*

$$\mathbb{P}(X_1 \leq x_1, ..., X_n \leq x_n) = \mathbb{P}(X_1 \leq x_1) \cdot ... \cdot \mathbb{P}(X_n \leq x_n)$$

for all $x_1, ..., x_n \in \mathbb{R}$.

Proof. We only need to prove the sufficiency of the claim as the necessity follows by definition and the fact that $\{X \leq x\}$ is a preimage of a Borel set $(-\infty, x]$.

For each $k = 1, 2, ..., n$ define $I_k := \{\{X_k \leq x\} : \ x \in \mathbb{R} \cup \{+\infty\}\}$. Clearly I_k is a π-system in $(\Omega, \mathcal{F})$ and $\Omega \in I_k$. We also have that $I_1, ..., I_n$ are independent thanks to the assumption of the lemma and (15.3) to cover the cases of $\Omega \in I_k$ which effectively means taking $x_k \to +\infty$. Now in view of $\{X_k \leq x\} = X_k^{-1}((-\infty, x])$ and the fact that the collection of sets $\{(-\infty, x] : \ x \in \mathbb{R}\}$ generates Borel σ-algebra $\mathcal{B}(\mathbb{R})$, we get by Claim 12.6 that the minimal σ-algebra containing I_k coincides with $\sigma(X_k)$. With this, Lemma 15.5 implies that $\sigma(X_1), \sigma(X_2), ..., \sigma(X_n)$ are independent and the proof is complete. □

As we can see from the proof of Lemma 15.6 above, the particular form of the sublevel sets $\{X \leq x\}$ plays no role. We only use the fact that the sets of the form $(-\infty, x]$ generate Borel σ-algebra of $\mathbb{R}$ and that sublevel sets form a π-system in Ω. We use this observation to formulate the next result, which is a generalization of the last lemma beyond sublevel sets. The proof of the next result, however, is the same as for the case of sublevel sets and will be left as an exercise.

Lemma 15.7. (`Independence of` n`-random variables via factoring CDFs, the case of general` π`-systems`) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X_1, ..., X_n : \Omega \to \mathbb{R}$ be random variables. Let also $\mathcal{A}_1, ..., \mathcal{A}_n$ be π-systems consisting of Borel subsets*

of $\mathbb{R}$ *such that* $\mathcal{A}_k$ *generates* $\mathcal{B}(\mathbb{R})$ *and* $\mathbb{R} \in \mathcal{A}_k$ *for each* $k = 1, ..., n$. *Then* $X_1, ..., X_n$ *are independent if and only if*

$$\mathbb{P}(X_1 \in A_1, ..., X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdot ... \cdot \mathbb{P}(X_n \in A_n),$$

for any $A_1 \in \mathcal{A}_1, ..., A_n \in \mathcal{A}_n$.

□ Exercise 15.7.1. *Prove Lemma 15.7 following the proof of Lemma 15.6.*

We now develop a criteria for independence based on joint probability densities. Given random variables $X_1, ..., X_n$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, assume there exists a function $\rho : \mathbb{R}^n \to \mathbb{R}_+$ such that for any Borel subset $B \in \mathcal{B}(\mathbb{R}^n)$ we have

$$\mathbb{P}(\mathbf{X} \in B) = \int_B \rho(\mathbf{x}) d\mathbf{x}. \tag{15.4}$$

Then such ρ is called the **joint probability density** of $\mathbf{X} := (X_1, ..., X_n)$. As in the case with joint CDF here as well we can easily recover the densities for particular random variables X_i. Indeed, consider X_1 and let $A \in \mathcal{B}(\mathbb{R})$ be any Borel set. Then $A \times \mathbb{R}^{n-1}$ is a Borel set in $\mathbb{R}^n$ and hence

$$\mathbb{P}(X_1 \in A) = \mathbb{P}(\mathbf{X} \in A \times \mathbb{R}^{n-1}) = \int_{A \times \mathbb{R}^{n-1}} \rho(\mathbf{x}) d\mathbf{x} = \int_A \rho_1(x_1) dx_1,$$

where

$$\rho_1(x_1) := \int_{\mathbb{R}^{n-1}} \rho(x_1, x_2, ..., x_n) dx_2 ... dx_n, \ x_1 \in \mathbb{R},$$

and is called the **marginal density**.

Lemma 15.8. (`Independence of` n `random variables via factorization of the joint probability distribution`) *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and let* $X_1, X_2, ..., X_n : \Omega \to \mathbb{R}$ *be random variables with a joint density* $\rho : \mathbb{R}^n \to \mathbb{R}_+$. *Then these random variables are independent if and only if* ρ *can be factored into*

$$\rho(\boldsymbol{x}) = \rho_1(x_1)\rho_2(x_2) \cdot ... \cdot \rho_n(x_n), \ \textit{almost everywhere for } \boldsymbol{x} = (x_1, ..., x_n) \in \mathbb{R}^n,$$

where ρ_i *are the marginal densities as defined above.*

Proof. We start with the necessity of the claim. Assume the random variables are independent and consider probability measures μ and ν on Borel subsets of $\mathbb{R}^n$ defined as

$$\mu(A) := \int_A \rho(\mathbf{x}) d\mathbf{x}$$

and

$$\nu(A) := \int_A \rho_1(x_1) \cdot ... \cdot \rho_n(x_n) d\mathbf{x}.$$

By definition of independence we get that μ and ν coincide on the sets A of the form $A_1 \times ... \times A_n$ where $A_i \in \mathcal{B}(\mathbb{R})$. Now the family of sets $\{A_1 \times ... \times A_n : \ A_i \in \mathcal{B}(\mathbb{R})\}$ is a π-system and measures μ and ν coincide on that π-system. From here and Lemma 15.2 (uniqueness) we conclude that μ and ν also coincide on the minimal σ-algebra of this π-system, which in this case is the Borel σ-algebra of $\mathbb{R}^n$.

We can now take $A = \{\mathbf{x} \in \mathbb{R}^n : \rho(x_1, ..., x_n) - \rho_1(x_1) \cdot ... \cdot \rho_n(x_n) > 0\}$ and let $\mathcal{L}$ be the Lebesgue measure in $\mathbb{R}^n$. The aim is to show that $\mathcal{L}(A) = 0$. The set A might

not be Borel, but it is Lebesgue measurable and using the regularity of the Lebesgue measure[35] we have that

$$\mathcal{L} = \sup\{\mathcal{L}(K) : \ K \subset A \text{ and } K \text{ is compact}\}.$$

Compact sets being closed are Borel, hence we can take any compact subset of A, say K and for that subset we get

$$\mu(K) = \nu(K),$$

in particular

$$0 \leq \int\limits_K (\rho(\mathbf{x}) - \rho_1(x_1) \cdot ... \cdot \rho_n(x_n))\, d\mathbf{x} = 0.$$

But the function is positive on K and has 0 integral, hence $\mathcal{L}(K) = 0$. Since $K \subset A$ was arbitrary from here and the regularity of $\mathcal{L}$ we conclude that $\mathcal{L}(A) = 0$. The same works for the set where $\rho - \rho_1 \cdot ... \cdot \rho_n < 0$. It now follows that $\rho = \rho_1 \cdot ... \cdot \rho_n$ almost everywhere on $\mathbb{R}^n$ which is what we needed to prove.

It is left to show the sufficiency of the claim. Assume we have $\rho = \rho_1 \cdot ... \cdot \rho_n$ almost everywhere on $\mathbb{R}^n$. Take any Borel sets $A_1, ..., A_n \in \mathcal{B}(\mathbb{R})$. Then

$$\begin{aligned}
\mathbb{P}(X_1 \in A_1, ..., X_n \in A_n) &= \int\limits_{A_1 \times ... \times A_n} \rho(\mathbf{x}) d\mathbf{x} \\
&= \int\limits_{A_1 \times ... \times A_n} \rho_1(x_1) \cdot ... \cdot \rho_n(x_n) d\mathbf{x} \\
&= \int\limits_{A_1} \rho_1(x_1) dx_1 \cdot ... \cdot \int\limits_{A_1} \rho_n(x_n) dx_n \\
&= \mathbb{P}(X_1 \in A_1) \cdot ... \cdot \mathbb{P}(X_n \in A_n).
\end{aligned}$$

This shows the independence of $X_1, ..., X_n$ and completes the completes the proof of the lemma. □

It is worth mentioning that there is an alternative way of showing the necessity of the last lemma without resorting to π-systems and Lemma 15.2 on uniqueness. Indeed, observe that in view of independence we have that

$$\int\limits_{\mathcal{C}} (\rho - \rho_1 \cdot ... \cdot \rho_n)\, d\mathbf{x} = 0$$

[35]The regularity of Lebesgue measure $\mathcal{L}$ refers to the property of approximating a set by more regular sets, more precisely, if $A \subset \mathbb{R}^n$ is Lebesgue measureable then

$$\mathcal{L}(A) = \inf\{\mathcal{L}(G) : A \subset G \text{ and } G \subset \mathbb{R}^n \text{ is open}\},$$

and

$$\mathcal{L}(A) = \sup\{\mathcal{L}(K) : K \subset A \text{ and } K \subset \mathbb{R}^n \text{ is compact}\}.$$

In fact the approximation by open sets holds for any set A with the difference being that the Lebesgue measure of A is replaced by the outer measure. For more details, see, for example, *Evans, L.C., Gariepy, R.F., Measure Theory and Fine Properties of Functions (Studies in Advanced Mathematics) 1st Edition, CRC Press, (1991).*

for any cube $\mathcal{C}$ in $\mathbb{R}^n$. In view of *Lebesgue's differentiation theorem*[36] for any integrable function $f \in L^1(\mathbb{R}^n)$ we have

$$\frac{1}{\mathcal{L}(\mathcal{C}_n)} \int\limits_{\mathcal{C}_n} f(\mathbf{y}) d\mathbf{y} \to f(\mathbf{x}) \text{ for almost every } \mathbf{x} \in \mathbb{R}^n,$$

where $\mathcal{C}_n$ is a family of cubes centered at $\mathbf{x}$ with diameter converging to 0. This theorem applied to $\rho - \rho_1 \cdot ... \cdot \rho_n$ directly implies that the function is 0 almost everywhere in $\mathbb{R}^n$.

15.3. **Construction of a sequence of independent random variables with given CDFs.** We will often encounter situations were we are dealing with a sequence of independent random variables. For example, in a series of independent coin tosses or dice rolls, or when sampling uniformly from say the interval $(0, 1)$ independently. We will prove in this section that such setups are mathematically well-posed by proving a general result in this regard.

Definition 15.3. `(Independence of a sequence of random variables)` *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X_1, X_2, ..., X_n, ...$ be sequences of random variables on it. We say that the sequence $\{X_n\}$ is **independent** if for any **finite set of indicies** $\mathcal{J} \subset \mathbb{N}$ the random variables $\{X_i : \ i \in \mathcal{J}\}$ are independent in a sense of Definition 15.2.*

□ Exercise 15.8.1. *Keeping the notation of Definition 15.3, a sequence $\{X_n\}$ is independent if and only if for any $N \geq 2$ the random variables $X_1, ..., X_N$ are independent.*

Theorem 15.9. `(Existence of independent sequence of random variables with given CDFs)` *Let $F_1, F_2, ..., F_n, ...$ be any sequence of CDFs. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a sequence of independent random variables $X_1, X_2, ...$ such that the CDF of each X_i is F_i for all $i = 1, 2, ...$.*

The theorem above is the main result of this section. The proof we will present here is constructive and will be based on a series of auxiliary results. First we show the existence of independent sequence of Bernoulli 0 - 1 variables (a particular case of the theorem when $\mathbb{P}(X_i = 0) = \mathbb{P}(X_i = 1) = 1/2$ for all $i \in \mathbb{N}$) using binary expansion of reals in $(0, 1)$. Then, using a sequence of Bernoulli variables construct a random variable uniformly distributed on $(0, 1)$, and then a sequence of independent uniformly distributed random variables on $(0, 1)$. After this a slight adjustment in the proof of Theorem 12.10 on characterization of CDFs leads to the proof of Theorem 15.9 in its full generality. In what follows we execute this plan.

Recall that the each $x \in [0, 1]$ can be represented as a sum of series of the form

$$x = \sum_{n=1}^{\infty} \frac{b_n}{2^n}, \tag{15.5}$$

where $b_n \in \{0, 1\}$. The expansion (15.5) is called **binary expansion** of x and we will refer to the coefficient b_n as the n-th **bit** of x owing the name to information theory. Unless x is from the set

$$\mathcal{D}_2 := \left\{ \frac{i}{2^k} : \ i = 0, 1, ..., 2^k, \ k = 1, 2, ... \right\}, \tag{15.6}$$

[36] See, for example, *Folland, G. B., Real analysis: modern techniques and their applications (1999)*, for a precise statement and proof.

which is called the set of **dyadic rationals**, the expansion (15.5) is *unique*. To see why expansion of points of $\mathcal{D}_2$ might not be unique, observe that both sequences of bits (0111...) and (1000...) sum to 1/2 using (15.5).

□ Exercise 15.9.1. (Existence of dyadic expansion) *Prove the existence of the binary expansion, namely show that for each $x \in [0,1]$ there exists a sequence $x_n \in \{0,1\}$, $n = 1, 2, ...$ such that (15.5) holds.*

Hint: split the interval $[0,1]$ into two equal parts at $1/2$. Then each $x \in [0,1]$ is either in the left side or the right side. If its in the left take the first bit as 0, if in the right, then take the first bit as 1. For $1/2$, which can be considered both in the right or left take any. We get that $|x - 1/2| \leq 1/2$. Now apply the same process for $1 - 2x$ if x was in the left segment or to $2x - 1$ otherwise.

□ Exercise 15.9.2. (Uniqueness of dyadic expansion) *Prove that if $x \in [0,1] \setminus \mathcal{D}_2$, i.e. x is not a dyadic rational, then the expansion (15.5) is unique.*

□ Exercise 15.9.3. (Existence of b-adic expansion) *Prove the existence of b-ary expansion for any $x \in [0,1]$, i.e. for a fixed integer $b \geq 2$ (which will be called **base**) show that there exist a sequence $\{x_n\}_{n=1}^{\infty}$ where $x_n \in \{0, 1, ..., b-1\}$ such that*

$$x = \sum_{n=1}^{\infty} \frac{x_n}{b^n}.$$

In analogy with dyadic rationals (15.6) define b-adic rationals, i.e. rational points in base b, and show that if x is not a b-ary rational, then its base-b expansion is unique.

Lemma 15.10. (Structure of points with given first n bits) *Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where $\Omega = (0,1)$, $\mathcal{F}$ is the standard Borel σ-algebra on $(0,1)$ and $\mathbb{P}$ is the Lebesgue measure on $(0,1)$. Define a sequence of mappings $X_n : \Omega \to \mathbb{R}$ where each X_n is the n-th bit of the binary representation of ω. Namely, if the binary representation of $\omega \in \Omega$ is the following*

$$\omega = \sum_{n=1}^{\infty} \frac{b_n}{2^n},$$

then we define $X_n(\omega) = b_n$. Then for any sequence $\varepsilon_1, ..., \varepsilon_n$ with each $\varepsilon_i \in \{0,1\}$ we have

$$\{\omega \in \Omega \setminus \mathcal{D}_2 : X_1(\omega) = \varepsilon_1, ..., X_n(\omega) = \varepsilon_n\} = \left(w, w + \frac{1}{2^n}\right) \setminus \mathcal{D}_2, \tag{15.7}$$

where $\mathcal{D}_2$ is the set of dyadic rationals defined in (15.6) and

$$w = \frac{\varepsilon_1}{2} + ... + \frac{\varepsilon_n}{2^n}.$$

Proof. For any $\omega \in \Omega$ from the left-hand side of (15.7) the first n bits are fixed and thus ω must be of the form

$$\omega = \frac{\varepsilon_1}{2} + ... + \frac{\varepsilon_n}{2^n} + \sum_{k=n+1}^{\infty} \frac{b_k}{2^k},$$

where $b_k \in \{0,1\}$ for all $k \geq n+1$. The sum above is bounded above by 2^{-n} (when all bits are 1) and is bounded below by 0 (all bits are 0). The bounds are strict as we assumed ω is not a dyadic rational and hence its dyadic representation is not eventually periodic. From this we get that

$$\frac{\varepsilon_1}{2} + ... + \frac{\varepsilon_n}{2^n} < \omega < \frac{\varepsilon_1}{2} + ... + \frac{\varepsilon_n}{2^n} + \frac{1}{2^n},$$

or equivalently that ω is from the right-hand side of (15.7).

To see the other way of inclusion, take any point $w < x < w + 2^{-n}$ which is not a dyadic rational, and let $x = \sum_{k=1}^{\infty} \frac{b_k}{2^k}$ be the binary representation of x. Then, we must have

$$\frac{\varepsilon_1}{2} + ... + \frac{\varepsilon_n}{2^n} < \frac{b_1}{2} + ... + \frac{b_n}{2^n} + \sum_{k=n+1}^{\infty} \frac{b_k}{2^k} < \frac{\varepsilon_1}{2} + ... + \frac{\varepsilon_n}{2^n} + \frac{1}{2^n}.$$

Multiplying both sides of the above inequality by 2^n we get

$$2^{n-1}\varepsilon_1 + ... + \varepsilon_n < 2^{n-1}b_1 + ... + b_n + 2^n \sum_{k=n+1}^{\infty} \frac{b_k}{2^k} < 2^{n-1}\varepsilon_1 + ... + \varepsilon_n + 1.$$

The left and right bounds of the last inequality are integers whose difference is 1. Hence the integer part of the middle term must coincide with the left-most side of the above inequality. Hence we get the relation

$$2^{n-1}\varepsilon_1 + ... + \varepsilon_n = 2^{n-1}b_1 + ... + b_n.$$

The latter implies that $b_n = \varepsilon_n$ since the rest of the terms are divisible by 2 and $|b_n - \varepsilon_n| \leq 1$. Canceling these terms and dividing the rest by 2 by induction we obtain that $\varepsilon_i = b_i$ for all $i = 1, 2, ..., n$. We conclude that the dyadic expansion of the any point from the right-hand side of (15.7) must have its first n bits identical with ε_i-s. This completes the proof the lemma. □

An important corollary of the structural result we just proved is that the binary bits of numbers of $(0, 1)$ form a sequence of independent Bernoulli random variables.

Lemma 15.11. (The n-th bit is Bernoulli, independent of other bits) *Let X_n be defined in Lemma 15.10. Then for each $n \in \mathbb{N}$ we have that X_n is a random variable with*

$$\mathbb{P}(X_n = 0) = \mathbb{P}(X_n = 1) = \frac{1}{2}$$

and the sequence $\{X_n\}$ is independent.

Proof. To see that X_n is indeed a random variable notice that by Lemma 15.10 for each $\varepsilon_n \in \{0, 1\}$ we have

$$\begin{aligned}
&\{\omega \in \Omega \setminus \mathcal{D}_2 : X_n(\omega) = \varepsilon_n\} \\
&= \bigcup_{(\varepsilon_1,...,\varepsilon_{n-1}) \in \{0,1\}^{n-1}} \{\omega \in \Omega \setminus \mathcal{D}_2 : X_1(\omega) = \varepsilon_1, ..., X_{n-1}(\omega) = \varepsilon_{n-1}, X_n(\omega) = \varepsilon_n\} \\
&= \bigcup_{(\varepsilon_1,...,\varepsilon_{n-1}) \in \{0,1\}^{n-1}} \left(\frac{\varepsilon_1}{2} + ... + \frac{\varepsilon_n}{2^n}, \quad \frac{\varepsilon_1}{2} + ... + \frac{\varepsilon_n}{2^n} + \frac{1}{2^n} \right) \setminus \mathcal{D}_2.
\end{aligned}$$

Clearly the right-hand side of the above equality is from $\mathcal{F}$, and since $\mathcal{D}_2$ is countable, hence of probability 0, we get that X_n is a random variable. Next, observe that each interval in the union above has length 2^{-n}, all intervals are non-intersecting and there are 2^{n-1} of them. Hence $\mathbb{P}(X_n = \varepsilon_n) = 1/2$. This proves the first part of the lemma.

To establish the independence of $\{X_n\}$, again using Lemma 15.10 for any $\varepsilon_1, ..., \varepsilon_n \in \{0, 1\}$ we have

$$\mathbb{P}(\{X_1 = \varepsilon_1, ..., X_{n-1} = \varepsilon_{n-1}, X_n = \varepsilon_n\}) = 2^{-n}.$$

But for each $1 \leq i \leq n$ we proved that $\mathbb{P}(X_i = \varepsilon_i) = 1/2$. Therefore

$$\mathbb{P}(\{X_1 = \varepsilon_1, ..., X_{n-1} = \varepsilon_{n-1}, X_n = \varepsilon_n\}) = \prod_{i=1}^{n} \mathbb{P}(X_i = \varepsilon_i)$$

which shows the independence of the sequence and completes the proof of the lemma. □

Lemma 15.12. *Let X_n be a sequence of independent Bernoulli random variables with $\mathbb{P}(X_n = 0) = \mathbb{P}(X_n = 1) = 1/2$ for all $n \in \mathbb{N}$. Define*

$$Y(\omega) = \sum_{n=1}^{\infty} \frac{X_n(\omega)}{2^n}.$$

Then Y is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ with uniform distribution on $(0, 1)$.

Proof. Since Y is a pointwise limit of random variables (partial sums of the series) it follows that Y is a random variable. We now prove that $Y \sim U(0, 1)$. Since X_n takes values 0 and 1, we have $\mathbb{P}(0 \leq Y \leq 1) = 1$. To show that Y has uniform distribution on $(0, 1)$ we are left to prove that the CDF of Y satisfies

$$\mathbb{P}(Y \leq x) = x \quad \text{for all } x \in (0, 1). \tag{15.8}$$

Observe that it is enough to show the above equality only for dyadic rationals $\mathcal{D}_2$. Indeed $\mathcal{D}_2$ is dense in $[0, 1]$ and the CDF is right-continuous, thus we can approximate any point of $[0, 1]$ by dyadic rationals from the right and pass to the limit using right-continuity of cumulative distribution function of Y and (15.8) for x restricted to $\mathcal{D}_2$. Thus the proof will be complete once we show (15.8) for $x \in \mathcal{D}_2$.

Take any point of the form $k/2^m \in [0, 1]$, where $0 \leq k \leq 2^m$ and $m \in \mathbb{N}$. Then

$$\mathbb{P}\left(Y \leq \frac{k}{2^m}\right) = \mathbb{P}\left(2^{m-1}X_1 + ... + X_m + \tilde{Y} \leq k\right), \tag{15.9}$$

where

$$\tilde{Y}(\omega) := \sum_{n=m+1}^{\infty} \frac{X_n(\omega)}{2^{n-m}}.$$

Since $\mathbb{P}(X_n \in \{0, 1\}) = 1$ it follows that the term $2^{m-1}X_1 + ... + X_m$ is an integer (with probability 1). Also notice that $\mathbb{P}(\tilde{Y} \leq 1) = 1$ thanks to the fact that X_n is bounded above by 1. We also have

$$\mathbb{P}(\tilde{Y} \leq 0) \leq \mathbb{P}(X_{m+1} = 0, ..., X_{m+i} = 0) = 2^{-i} \to 0, \text{ as } i \to \infty,$$

where we used the independence of X_n. It follows that[37] $\mathbb{P}(\tilde{Y} \in [0,1]) = 1$ and getting back to (15.9) we obtain

$$\begin{aligned}\mathbb{P}\left(Y \le \frac{k}{2^m}\right) &= \mathbb{P}\left(2^{m-1}X_1 + ... + X_m = k,\ \tilde{Y} \le 0\right)\\ &+ \mathbb{P}\left(2^{m-1}X_1 + ... + X_m \le k-1,\ \tilde{Y} \in [0,1]\right)\\ &= \mathbb{P}\left(2^{m-1}X_1 + ... + X_m \le k-1\right) \text{ (the left-hand side is an integer)}\\ &= \sum_{i=0}^{k-1} \mathbb{P}(2^{m-1}X_1 + ... + X_m = i) \text{ (by Lemma 15.10)}\\ &= \frac{k}{2^m},\end{aligned}$$

completing the proof[38]. □

Lemma 15.13. (Independent sequence of uniformly distributed random variables) *There exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a sequence $U_1, U_2, ...$ of random variables defined on it such that each U_k is uniformly distributed on the interval $(0,1)$ and the sequence $\{U_k\}$ is independent.*

Proof. We take $(\Omega, \mathcal{F}, \mathbb{P})$ as in Lemma 15.10, i.e. $\Omega = (0,1)$ with the usual Lebesgue measure defined on Borel sets. Thanks to Lemma 15.11 there exists a sequence of independent Bernoulli random variables $X_n : \Omega \to \{0,1\}$ with $\mathbb{P}(X_n = 0) = \mathbb{P}(X_n = 1) = 1/2$ for all $n \in \mathbb{N}$. Partition the set of positive integers into countable number of non-intersecting countable subsets, more precisely write $\mathbb{N} = \mathbb{N}_1 \cup \mathbb{N}_2 \cup ...$ where each $\mathbb{N}_i$ is countable and $\mathbb{N}_i \cap \mathbb{N}_j = \emptyset$ where $i \ne j$ (see Figure 13 for a particular scheme of such partition, an exact form of $\mathbb{N}_i$ is not relevant here).

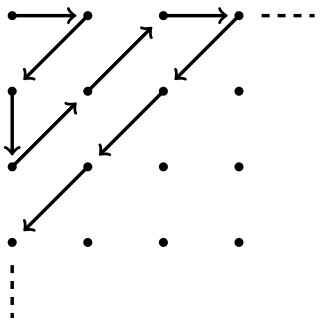

FIGURE 13. An example of a scheme that can be used to partition the set of positive integers $\mathbb{N}$ into countable number of countable sets. Starting from top-left corner fill the numbers $1, 2, 3, 4, ...$ following the arrows. For example, the dots in the first row will get values $1, 2, 6, 7, ...$, the second row will start with $3, 5, 8, ...$. Clearly all elements of $\mathbb{N}$ will be covered in this way and each row of the table is a countable subset of $\mathbb{N}$ that does not intersect with other rows.

[37] In fact, $\tilde{Y}$ has the same distribution as Y as the lemma eventually proves.

[38] For the passage to the third line we are using the fact that if A and B are any events and $\mathbb{P}(B) = 1$ then $\mathbb{P}(A \cap B) = \mathbb{P}(A)$ and similarly if $\mathbb{P}(B) = 0$ then $\mathbb{P}(A \cap B) = 0$ (prove these). In our case B is either $\{\tilde{Y} \in [0,1]\}$ (en event of probability 1) or $\{\tilde{Y} \le 0\}$ (an event of probability 0).

For $k=1,2,...$ consider the sum

$$U_k(\omega):=\sum_{n=1}^{\infty}\frac{X_{\mathcal{J}(k,n)}}{2^n},$$

where $\mathcal{J}(k,n)$ is the n-th element of $\mathbb{N}_k$. By Lemma 15.12 we have $U_k\sim U(0,1)$. It is left to show that the sequence $\{U_k\}$ is independent. To this end fix an integer $n\geq 2$ and take any dyadic rationals

$$\frac{i_1}{2^{p_1}},\frac{i_2}{2^{p_2}},...,\frac{i_n}{2^{p_n}}\in\mathcal{D}_2.$$

Using a similar argument as we had in Lemma 15.9 with computing the CDF at the dyadic rationals, here as well we have

$$\begin{aligned}
&\mathbb{P}\left(U_1\leq\frac{i_1}{2^{p_1}},\ U_2\leq\frac{i_2}{2^{p_2}},\ ...\ ,U_n\leq\frac{i_n}{2^{p_n}}\right)\\
&=\mathbb{P}(2^{p_1-1}X_{\mathcal{J}(1,1)}+...+X_{\mathcal{J}(1,p_1)}\leq i_1-1,\\
&\qquad 2^{p_2-1}X_{\mathcal{J}(2,1)}+...+X_{\mathcal{J}(2,p_2)}\leq i_2-1,\\
&\qquad\qquad ...\\
&\qquad 2^{p_n-1}X_{\mathcal{J}(n,1)}+...+X_{\mathcal{J}(n,p_n)}\leq i_n-1)\\
&\qquad\qquad\text{(by independence of }\{X_k\})\\
&=\mathbb{P}(2^{p_1-1}X_{\mathcal{J}(1,1)}+...+X_{\mathcal{J}(1,p_1)}\leq i_1-1)\cdot...\cdot\mathbb{P}(2^{p_n-1}X_{\mathcal{J}(n,1)}+...+X_{\mathcal{J}(n,p_n)}\leq i_n-1)\\
&=\mathbb{P}\left(U_1\leq\frac{i_1}{2^{p_1}}\right)\cdot...\cdot\mathbb{P}\left(U_n\leq\frac{i_n}{2^{p_n}}\right).
\end{aligned}$$

Consider the set of Borel subsets of $\mathbb{R}$ defined as

$$\mathcal{A}:=\{(-\infty,x]:\ x\in(-\infty,0]\cup\mathcal{D}_2\cup[1,+\infty]\}.$$

Then $\mathcal{A}$ is a π-system that generates $\mathcal{B}(\mathbb{R})$. Since the sublevel sets of U_k-s restricted to dyadic rationals are independent and also $\mathbb{P}(U_k\in(0,1))=1$, Lemma 15.7 applied for $U_1,...,U_n$ and π-systems $\mathcal{A},...,\mathcal{A}$ where we have n copies of $\mathcal{A}$, implies the independence of $U_1,U_2,...,U_n$. The proof of the lemma is complete. □

Proof of Theorem 15.9. As before take $(\Omega,\mathcal{F},\mathbb{P})$ the interval $(0,1)$ with standard Lebesgue measure defined on Borel subsets. By Lemma 15.13 there exists a sequence of independent random variables $U_n\sim U(0,1)$ on $(\Omega,\mathcal{F},\mathbb{P})$. As in the proof of Theorem 12.10 on characterization of CDFs define (cf. (12.5))

$$X_n(\omega):=\sup\{y\in\mathbb{R}:\ F_n(y)<U_n(\omega)\},\quad \omega\in\Omega.$$

Repeating the same argument as we had for (12.5) leads to $\mathbb{P}(X_n<\infty)=1$ and

$$\{\omega\in\Omega:\ X_n(\omega)\leq x\}=\{\omega\in\Omega:\ U_n(\omega)\leq F_n(x)\},\ x\in\mathbb{R}.$$

The above relation coupled with Lemma 12.3 on measurability of sublevel sets implies that X_n is a random variable. Moreover, since $U_n\sim U(0,1)$ we get

$$\mathbb{P}(X_n\leq x)=\mathbb{P}\left(U_n\leq F_n(x)\right)=F_n(x),\ x\in\mathbb{R},$$

which means that the CDF of X_n is F_n. It is left to show that the sequence $\{X_n\}$ is independent. To this end fix $n\geq 2$ and take any $x_1,x_2,...,x_n\in\mathbb{R}\cup\{+\infty\}$. Again, repeating the same argument as we had for (12.5) we get

$$\begin{aligned}
&\{\omega\in\Omega:\ X_1(\omega)\leq x_1,...,X_n(\omega)\leq x_n\}\\
&=\{\omega\in\Omega:\ U_1(\omega)\leq F_1(x_1),...,U_n(\omega)\leq F_n(x_n)\}.
\end{aligned}$$

Using the independence of $\{U_n\}$, from the last expression we obtain

$$\begin{aligned}&\mathbb{P}\left(X_1 \leq x_1, ..., X_n \leq x_n\right)\\ &= \mathbb{P}\left(U_1 \leq F_1(x_1), ..., U_n \leq F_n(x_n)\right)\\ &= \mathbb{P}\left(U_1 \leq F_1(x_1)\right) \cdot ... \cdot \mathbb{P}\left(U_n \leq F_n(x_n)\right)\\ &= \mathbb{P}(X_1 \leq x_1) \cdot ... \cdot \mathbb{P}(X_n \leq x_n).\end{aligned}$$

It follows that the π-systems $I_k := \{\{X_k \leq x\} : \ x \in \mathbb{R} \cup \{+\infty\}\}$, $k = 1, 2, ..., n$ are independent. Now Lemma 15.6 implies that $X_1, X_2, ..., X_n$ are independent. Since $n \geq 2$ was arbitrary the proof is complete. □

□ Exercise 15.13.1. *Fill the details in the proof of Theorem 15.9 on applying the arguments used for proving (12.5) in the proof of the theorem.*

16. Tail events and Kolmogorov's zero-one law

In this section we study Kolmogorov's zero-one law, which is a powerful tool in the analysis of a large family of events called **tail events**. Let us start with an example: consider the experiment of throwing a fair 6-sided dice infinitely many times. Assume we are interested in the event of seeing the number 5 infinitely many times in the sequence of throws[39]. An interesting observation about this event is that ignoring or changing finitely many of the outcomes of the throws has no influence on the validity of the event. This is of course not an isolated example, but is rather tip of the iceberg of a larger phenomenon as we will see shortly.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X_1, X_2, ..., X_n, ...$ be random variables defined on it. For an index set $\mathcal{I} \subset \mathbb{N}$ we write $\sigma(X_n : \ n \in \mathcal{I})$ for the smallest σ-algebra generated by the random variables $\{X_n : \ n \in \mathcal{I}\}$. More precisely

$$\sigma(X_n : \ n \in \mathcal{I}) := \sigma\left(\bigcup_{n \in \mathcal{J}} \{X_n^{-1}(B) : \ B \in \mathcal{B}(\mathbb{R})\}\right),$$

where as usual $\mathcal{B}(\mathbb{R})$ is the Borel σ-algebra of $\mathbb{R}$. With this notation in mind, for each $n \in \mathbb{N}$ define

$$\mathcal{T}_n := \sigma\left(X_{n+1}, X_{n+2}, ...\right) \quad \text{and} \quad \mathcal{T} := \bigcap_{n=1}^{\infty} \mathcal{T}_n,$$

Here $\mathcal{T}$ is called the **tail** σ-algebra of the sequence $\{X_n\}_{n=1}^{\infty}$, and the elements of $\mathcal{T}$ are called *tail events*. Remarkably, many naturally occurring events are tail events for some sequence of random variables. Here are a few examples of tail events. Set

$$\begin{aligned}E_1 &:= \{\omega \in \Omega : \ \exists \lim_{n \to \infty} X_n(\omega)\},\\ E_2 &:= \{\omega \in \Omega : \ \sum_{n=1}^{\infty} X_n(\omega) < \infty\},\\ E_3 &:= \left\{\omega \in \Omega : \ \lim_{n \to \infty} \frac{X_1 + ... + X_n}{n} = \frac{1}{2}\right\}.\end{aligned}$$

[39] Using the result of section 15.3 we can rigorously define a probability space for this experiment and show that "getting infinitely many 5s" is indeed an event.

All events $E_1, E_2,\ E_3$ are tail events, i.e. elements of $\mathcal{T}$. Let us illustrate on the example of E_1 how that can be proved formally. The existence of a limit of a sequence is equivalent to the sequence being Cauchy (fundamental). Thus, $\omega \in E_1$ if and only if

$$\forall \varepsilon > 0\ \exists N = N(\varepsilon) \text{ s.t. } \forall n \geq N(\varepsilon),\ \ \forall m \in \mathbb{N} \ \text{ we have } \ |X_{n+m}(\omega) - X_n(\omega)| < \varepsilon.$$

Now writing the above in the language of sets and replacing (equivalently) $\varepsilon > 0$ by a small number of the form $1/k$ we get

$$E_1 = \bigcap_{k=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \bigcap_{m=1}^{\infty} \left\{ \omega \in \Omega :\ |X_{n+m}(\omega) - X_n(\omega)| < \frac{1}{k} \right\}.$$

First of all, from here we conclude that E_1 is indeed an event, i.e. an element of $\mathcal{F}$, as each X_n is a random variable. To see that $E_1 \in \mathcal{T}_n$ for any $n \in \mathbb{N}$, denote by $A_{N,k}$ the set in the definition of E_1 containing the last two intersections, i.e. $E_1 = \bigcap_k \bigcup_N A_{N,k}$. By construction, if $\omega \in A_{N,k}$ then $\omega \in A_{N_1,k}$ for any $N_1 \geq N$ because of the set intersection in the definition of $A_{N,k}$. Hence the lower limit of $N = 1$ can be replaced by $N = N_1$ for any $N_1 \geq 1$ implying that $E_1 \in \mathcal{T}_n$ for any $n \in \mathbb{N}$ and thereby $E_1 \in \mathcal{T}$.

Informally speaking, tail events are the events whose behavior does not change by changing any finite number of elements from the sequence $\{X_n\}$. For example, to see if the sequence X_n has a limit or not, we can study the sequence with any offset, i.e. instead of starting from the first element of the sequence, we can skip any *finite* number of elements from $\{X_n\}$: the existence and the value of the limit will not change with this.

□ Exercise 16.0.1. *Prove that E_2 and E_3 are elements of $\mathcal{T}$.*

Theorem 16.1. (`Kolmogorov's 0-1 law`) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X_n : \Omega \to \mathbb{R}$ be a sequence of independent random variables. Let also $\mathcal{T}$ be the tail σ-algebra for the sequence $\{X_n\}$. Then*

(a) *$\mathcal{T}$ is trivial, i.e. for any $A \in \mathcal{T}$ we have $\mathbb{P}(A) \in \{0, 1\}$,*

(b) *For any $\xi : \Omega \to \mathbb{R} \cup \{\pm\infty\}$ which is $\mathcal{T}$-measurable, there exists a constant $c \in \mathbb{R} \cup \{\pm\infty\}$ such that*

$$\mathbb{P}(\xi = c) = 1,$$

i.e. ξ is deterministic.

Proof. We start with part (a) and will prove that $\mathcal{T}$ is independent of itself, i.e. for any $A, B \in \mathcal{T}$ one has $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Once we get this, we can take $A = B$ and obtain $\mathbb{P}(A) = \mathbb{P}(A)^2$ implying that $\mathbb{P}(A)$ is either 0 or 1. We now turn to the proof of independence of $\mathcal{T}$ from itself, which will be split into several steps.

Step 1. Define $\mathcal{X}_n = \sigma(X_1, X_2, ..., X_n)$, i.e. the σ-algebra generated by random variables $X_1, ..., X_n$. We claim that $\mathcal{X}_n$ is independent of $\mathcal{T}_n$. This should be intuitively clear as the indicies of random variables in $\mathcal{X}_n$ and $\mathcal{T}_n$ do not intersect. To prove the independence rigorously, consider the collection of sets

$$\mathcal{A} := \{X_1 \leq x_1, ..., X_n \leq x_n :\ x_i \in (-\infty, +\infty],\ 1 \leq i \leq n\}.$$

Clearly $\mathcal{A}$ is a π-system and by Lemma 12.7 generates $\mathcal{X}_n$. The latter is due to the fact that allowing $x_i = +\infty$ implies that $\mathcal{A}$ contains events of the $\{X_k \leq x_k\}$ for any $1 \leq k \leq n$ and any $x_k \in \mathbb{R}$, hence Lemma 12.7 can be applied. Similarly, the collection of sets

$$\mathcal{B} := \{X_{n+1} \leq x_{n+1}, ..., X_{n+k} \leq x_{n+k} :\ x_i \in (-\infty, +\infty],\ k = 1, 2, ...\}$$

is a π-system and by Lemma 12.7 generates $\mathcal{T}_n$. By independence of the sequence $\{X_n\}$ the π-systems $\mathcal{A}$ and $\mathcal{B}$ are independent. Hence by Lemma 15.3 on independence of σ-algebras generated by independent π-systems we get that $\mathcal{X}_n$ and $\mathcal{T}_n$ are independent.

Step 2. We claim that $\mathcal{X}_n$ and $\mathcal{T}$ are independent for each $n \in \mathbb{N}$. Indeed, $\mathcal{T} \subset \mathcal{T}_n$ by construction and by Step 1 $\mathcal{X}_n$ is independent of $\mathcal{T}_n$. Hence $\mathcal{X}_n$ is also independent from any subset of $\mathcal{T}_n$, in particular from $\mathcal{T}$.

Step 3. Here we show that $\sigma(X_n : \ n \in \mathbb{N})$ is independent of $\mathcal{T}$. Indeed, consider the collection of sets

$$\mathcal{X}_\infty := \bigcup_{n=1}^{\infty} \mathcal{X}_n.$$

Since $\mathcal{X}_1 \subset \mathcal{X}_2 \subset ...$ and each $\mathcal{X}_k$ is a σ-algebra, the collection of sets $\mathcal{X}_\infty$ defines a π-system. This π-system is independent of $\mathcal{T}$ by Step 2, thus applying Lemma 15.3 to $\mathcal{X}_\infty$ and $\mathcal{T}$ (which itself is a π-system) we get that the σ-algebras $\sigma(\mathcal{X}_\infty)$ and $\mathcal{T}$ are independent. But thanks to Lemma 12.7 we have that $\sigma(\mathcal{X}_\infty) = \sigma(X_n : \ n \in \mathbb{N})$ hence the proof of Step 3 is complete.

Step 4. From Step 3 we have that $\mathcal{T}$ is independent of $\sigma(X_n : \ n \in \mathbb{N})$. By construction $\mathcal{T} \subset \sigma(X_n : \ n \in \mathbb{N})$ and hence $\mathcal{T}$ is independent of $\mathcal{T}$ as well. This completes the proof of part (a) of the theorem.

We now prove part **(b)**. For any $x \in \mathbb{R}$ thanks to part (a) we have $\mathbb{P}(\xi \leq x) \in \{0, 1\}$. Define

$$c_* := \sup\{x \in \mathbb{R} : \ \mathbb{P}(\xi \leq x) = 0\}.$$

If $\{x \in \mathbb{R} : \ \mathbb{P}(\xi \leq x) = 0\} = \emptyset$ then $\mathbb{P}(\xi \leq x) = 1$ for any $x \in \mathbb{R}$ and hence $\mathbb{P}(\xi = -\infty) = 1$, and if $c_* = +\infty$ then $\mathbb{P}(\xi \leq x) = 0$ for any $x \in \mathbb{R}$ implying that $\mathbb{P}(\xi = +\infty) = 1$. Hence we are left with the case when $c_* \in \mathbb{R}$. We write

$$\{\xi \leq c_*\} = \bigcap_{n=1}^{\infty} \left\{\xi \leq c_* + \frac{1}{n}\right\},$$

and by definition of c_* we have $\mathbb{P}(\xi \leq c_* + 1/n) = 1$. Hence the continuity of measure implies $\mathbb{P}(\xi \leq c_*) = 1$. Similarly,

$$\{\xi < c_*\} = \bigcup_{n=1}^{\infty} \left\{\xi < c_* - \frac{1}{n}\right\},$$

and from the definition of c_* we have $\mathbb{P}(\xi < c_* - 1/n) = 0$. Hence by the continuity of measure we obtain $\mathbb{P}(\xi < c_*) = 0$. Putting everything together we get

$$\mathbb{P}(\xi = c) = \mathbb{P}(\xi \leq c_*) - \mathbb{P}(\xi < c_*) = 1 - 0 = 1.$$

The proof of part (b) and the theorem is now complete. □

□ Exercise 16.1.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\xi : \Omega \to \mathbb{R} \cup \{\pm\infty\}$ be a random variable. Assume for each $x \in \mathbb{R}$ we have $\mathbb{P}(\xi \leq x) = 1$. Prove that $\mathbb{P}(\xi = -\infty) = 1$.*

We now discuss a few scenarios where Kolmogorov's 0-1 law can be used.

Random power series. Consider a power series

$$f(z) = \sum_{n=0}^{\infty} a_n z^n, \quad z \in \mathbb{C},$$

where the coefficients $\{a_n\} \subset \mathbb{C}$ do not depend on z. It is well-known that the radius of convergence of the series is

$$R := \limsup_{n\to\infty} \frac{1}{\sqrt[n]{|a_n|}},$$

i.e. the series above converges for $|z| < R$ and diverges for $|z| > R$. Take a sequence of independent random variables $\{X_n\}$ defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and consider a series

$$F(z,\omega) := \sum_{n=0}^{\infty} X_n(\omega) z^n, \ z \in \mathbb{C}.$$

Remarkably, the radius of convergence of this series, which depends on ω, is constant almost surely. Indeed,

$$R(\omega) := \limsup_{n\to\infty} \frac{1}{\sqrt[n]{|X_n(\omega)|}},$$

hence R is $\mathcal{T}$-measurable where $\mathcal{T}$ is the tail σ-algebra of the sequence $\{X_n\}$. Now, Kolmogorov's 0-1 theorem implies the existence of a constant $c \in [0, +\infty]$ such that $\mathbb{P}(R = c) = 1$.

Bond percolation on $\mathbb{Z}^2$. Let $\mathbb{Z}^2$ be the integer lattice of $\mathbb{R}^2$, i.e. points on the plane with integer coordinates. Consider a graph $G = (V, E)$ where the set of verticies are the points of $\mathbb{Z}^2$ and the set of edges are edges (*bonds*) between neighboring points (each point has 4 neighbors). Now let $e_1, e_2, ...$ be any enumeration of these edges. Consider a sequence of independent Bernoulli random variables X_n where $\mathbb{P}(X_n = 0) = 1 - p$ and $\mathbb{P}(X_n = 1) = p$ and $p \in [0, 1]$ is fixed. Define a (probabilistic) graph $G_p := (V, E_p)$ where $V = \mathbb{Z}^2$ and E_p consists of all edges e_n for which $X_n = 1$. We can think of this as tossing a coin for each edge independently of others and including the edge in the graph if and only if the coin comes up Heads (see Figure 14).

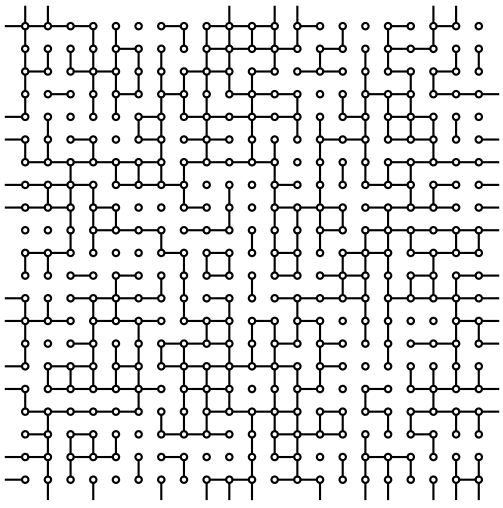

FIGURE 14. Part of $\mathbb{Z}^2$ with random edges. The small circles with empty interiors represent points of the lattice $\mathbb{Z}^2$. Each point has 4 neighbors and connects to a particular neighbor independently from others with probability p which in the case is set to 0.5. Here we see a particular realization of such experiment for a small subset of the lattice.

An important question about the graph G_p is whether it has an *infinite connected component* or not. This problem is called *bond percolation*. It is easy to see that the

existence of an infinite connected component is a tail event for the sequence of Bernoulli variables $\{X_n\}$ defined above. Hence for any $p \in [0,1]$ the probability that G_p has an infinite connected component is either 0 or 1. This remarkable results follows directly from Kolmogorov's theorem[40].

→ **Problem 16.1.1.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X_n : \Omega \to \mathbb{R}$ be a sequence of independent random variables. Assume $X_n \to X$ almost surely. Prove that X is deterministic, i.e. there exists a constant $c \in \mathbb{R} \cup \{\pm\infty\}$ such that $\mathbb{P}(X = c) = 1$.*

Show that the convergence almost surely can be replaced by convergence in probability.

Hint: It is enough to show that X is $\mathcal{T}$-measurable, where $\mathcal{T}$ is the tail σ-algebra of $\{X_n\}$ since then Kolmogorov's theorem part (b) can be applied. To show that X is measurable with respect to the tail σ-algebra consider the sublevel sets $\{X \le x\}$ and use Remark 12.5. For the second part of the problem recall the relation between two modes of convergence of a sequence of random variables.

17. Mathematical expectation as Lebesgue integral

The aim of this section is to define *mathematical expectation* of a random variable. Recall that the discrete case was covered in section 6.2, and here we study the general case, which will also cover the discrete random variables as a particular case. The expectation will be defined as a Lebesgue integral with respect to the probability measure. We will briefly recall the construction of integral and several important results that will be in use throughout the text. Not all results will be given with proofs as a large part of the material in this section is a refreshment of the theory of Lebesgue integral. At the end of the section we will prove some useful results of practical value for computing the expectation.

17.1. **Integration.** Here we briefly recall the construction of Lebesgue integral on a general measure space. Consider a measure space $(S, \mathcal{F}, \mu)$ where S is any non-empty set, $\mathcal{F}$ is a σ-algebra of subsets of S and μ is a measure defined on $\mathcal{F}$, i.e. μ satisfies all requirements of Definition 1.2 except that its values are not necessarily bounded above by 1. Given a measurable function $f : S \to \mathbb{R} \cup \{\pm\infty\}$, i.e. $f^{-1}(B) \in \mathcal{F}$ for any $B \in \mathcal{B}(\mathbb{R})$, the goal is to define an integral $f \mapsto \int f d\mu$ for f. We will follow the construction of Lebesgue integral that is split into the following steps:

- start with the *indicator functions*,
- extend to the sums of indicators with non-negative coefficients, these will be called *simple functions*,
- use approximation argument involving simple functions to define the integral for any *non-negative function*,
- for the general case, split the given measurable function into a sum of non-negative and non-positive parts and apply the construction for non-negative functions to each component.

What follows next is an execution of this plan.

Step 1. Indicators. For $A \in \mathcal{F}$ let $\mathbb{I}_A$ be the indicator of A, i.e.

$$\mathbb{I}_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \in A^c. \end{cases}$$

[40]To find for which $p \in [0,1]$ the probability of existence of an infinite cluster is actually 1 is a very deep result that is beyond the scopes of these notes; see H. Kesten, "The Critical Probability of Bond Percolation on the Square Lattice Equals 1/2", Commun. Math. Phys. 74, 41-59 (1980).

Define $\int \mathbb{I}_A d\mu := \mu(A) \leq \infty$ and call this the integral of $\mathbb{I}_A$.

Step 2. Simple functions. Call a function $f : S \to \mathbb{R} \cup \{+\infty\}$ **simple**, if there exists $A_1, A_2, ..., A_n \in \mathcal{F}$ and scalars $a_k \in [0, +\infty]$, $1 \leq k \leq n$ such that

$$f = \sum_{k=1}^{n} a_k \mathbb{I}_{A_k}. \tag{17.1}$$

Note that we do *not* assume A_k-s to be non-intersecting. If f is a simple function, define its integral as follows:

$$\int f d\mu := \sum_{k=1}^{n} a_k \mu(A_k) \leq \infty, \tag{17.2}$$

with the convention that $0 \cdot \infty = \infty \cdot 0 = 0$.

Clearly, if f is simple, then the representation (17.2) is not unique in general, just think about writing $\mathbb{I}_A = \mathbb{I}_{A\cap B} + \mathbb{I}_{A\cap B^c}$ for any $A, B \in \mathcal{F}$. Nevertheless, the value of (17.2) does not depend on the representation (17.1), as Problem 17.0.1 suggests.

→ Problem 17.0.1. *Let $f : S \to \mathbb{R} \cup \{+\infty\}$ be a simple function and assume we have two equal representations for it, namely*

$$f = \sum_{i=1}^{n} a_i \mathbb{I}_{A_i} = \sum_{j=1}^{m} b_j \mathbb{I}_{B_j},$$

where equality above at all points of X and $A_i, B_j \in \mathcal{F}$ and $a_i, b_j \in [0, +\infty]$ are scalars. Prove that

$$\sum_{i=1}^{n} a_i \mu(A_i) = \sum_{j=1}^{m} b_j \mu(B_j),$$

i.e. the value of the integral of a simple function, defined in (17.2), does not depend on its representation.

Hint: First show that for each representation (17.2) there is an equivalent (equal) representation where all sets are pairwise disjoint. To show this for each $x \in X$ consider the maximal set of indicies $I \subset \{1, 2, ..., n\}$ such that $x \in \bigcap_{i\in I} A_i$ then pass to representations over such intersections where the coefficient will be $\sum_{i\in I} a_i$. Once we may assume that $A_i \cap A_j = \emptyset$, $i \neq j$, partition $\mathbb{I}_{A_i} = \sum_{j=1}^{m} \mathbb{I}_{A_i\cap B_j}$ (here we may assume that $\bigcup_{j=1}^{m} B_j = X$ by adding $X \setminus \bigcup_{j=1}^{m} B_j$ to the representation of f with 0 coefficient) and similarly for B_j, and then replace these partitions of indicators in the corresponding representations of f. In this new representation compare the terms $a_i \mathbb{I}_{A_i\cap B_j}$, $b_j \mathbb{I}_{A_i\cap B_j}$ and their contribution in the integral of f.

We now collect several properties of the integral for simple functions together.

Proposition 17.1. `(Basic properties for simple functions)` *Let f and g be simple functions, and $c \in [0, +\infty]$ be any constant. Simple functions and their integral have the following properties:*

(a) *The integral (17.2) is well-defined, i.e. its value does not depend on the representation (17.1),*
(b) *if $\mu(f \neq g) = 0$ then $\int f d\mu = \int g d\mu$,*
(c) *$f + g$ and cf are also simple functions,*
(d) *if $f \leq g$ then $\int f d\mu \leq \int g d\mu$,*
(e) *$\min(f, g)$ and $\max(f, g)$ are also simple.*

We will use these properties throughout the text. An approach to the first item of the proposition is outlined in Problem 17.0.1, the rest is left as an exercise.

□ Exercise 17.1.1. *Prove Proposition 17.1.*

Step 3. Non-negative measurable functions. Let $f : X \to \mathbb{R} \cup \{+\infty\}$ be measurable and assume $f \geq 0$. Define

$$\int f d\mu := \sup_{\substack{h \leq f \\ h \text{ is simple}}} \int h d\mu. \tag{17.3}$$

If f is simple then the definition (17.3) coincides with the definition (17.2) in view of the monotonicity property in Proposition 17.1 (d).

Lemma 17.2. (`Non-negative function with 0 integral is 0 a.e.`) *Let $f \geq 0$ be a measurable function and $\int f d\mu = 0$, then $\mu\{f > 0\} = 0$, i.e. $f = 0$ almost everywhere (a.e.).*

Proof. For each $n \in \mathbb{N}$ we have

$$f \geq \frac{1}{n} \mathbb{I}_{\{f > 1/n\}},$$

and the right-hand side of the inequality is a simple function. Hence using the definition of integral (17.3) we get

$$0 = \int f d\mu \geq \int \frac{1}{n} \mathbb{I}_{\{f > 1/n\}} d\mu = \frac{1}{n} \mu\{f > 1/n\} \geq 0,$$

from which we conclude that $\mu\{f > 1/n\} = 0$ for each $n \in \mathbb{N}$. Now observe that

$$\{f > 0\} = \bigcup_{n=1}^{\infty} \left\{ f > \frac{1}{n} \right\},$$

and each set in the union is of measure 0, hence so is the union (one can either apply sub-additivity of measure (Boole's inequality), or in this case use continuity of measure observing that the sets $\{f > 1/n\}$ are nested). The proof is complete. □

Theorem 17.3. (`Monotone convergence for integrals (MON)`) *Let $f_n \geq 0$ be a sequence of measurable functions such that $f_n \uparrow f$ everywhere on S. Then*

$$\int f_n d\mu \uparrow \int f d\mu.$$

As a shorthand we will refer to this theorem as `MON`. This is a well-known and very useful theorem related to Lebesgue integral, the proof of which can be found in standard textbooks (see [2] or [9] for example). Soon we will see that the condition of everywhere convergence of f_n can be relaxed to μ-almost everywhere.

Sometimes we might need an explicit example of simple functions that monotonically converge to a given non-negative function. To construct such example, for each $k \in \mathbb{N}$ define a sequence of functions $\alpha^{(k)} : [0, +\infty] \to [0, +\infty)$ as follows

$$\alpha^{(k)}(x) := \begin{cases} 0, & x = 0, \\ \frac{i-1}{2^k}, & \frac{i-1}{2^k} < x \leq \frac{i}{2^k} \leq k, \ i \in \mathbb{N} \\ k, & x > k. \end{cases} \tag{17.4}$$

By definition each $\alpha^{(k)}$ is a simple function and it is easy to verify that $\alpha^{(k)}(x) \leq \alpha^{(k+1)}(x)$ for any fixed $x \geq 0$, i.e. the sequence $\alpha^{(k)}(x)$ is non-decreasing for any $x \geq 0$.

Now for a measurable function $f : S \to [0, +\infty]$ define $f^{(k)} := (\alpha^{(k)} \circ f)(x) = \alpha^{(k)}(f(x))$, $x \in S$. Then $f^{(k)} \uparrow f$ and $f^{(k)}$ is a simple function for each $k \in \mathbb{N}$. By `MON` we also get

$$\int f^{(k)} d\mu \uparrow \int f d\mu$$

We now discuss how to relax the everywhere convergence requirement in `MON`.

Claim 17.4. *If $f, g \geq 0$ are measurable and $\mu(f \neq g) = 0$ then $\int f d\mu = \int g d\mu$.*

Proof. Consider the sequences $f^{(k)} := \alpha^{(k)} \circ f$ and $g^{(k)} := \alpha^{(k)} \circ g$, where $k \in \mathbb{N}$. Both are sequences of simple functions. Since $\mu\{f \neq g\} = 0$ we have $\mu\{f^{(k)} \neq g^{(k)}\} = 0$ by construction for each $k \in \mathbb{N}$ and hence in view of Proposition 17.1

$$\int f^{(k)} d\mu = \int g^{(k)} d\mu, \quad k \in \mathbb{N}.$$

Since $f^{(k)} \uparrow f$ and $g^{(k)} \uparrow g$, by `MON` we can pass to the limit in last equality and obtain $\int f d\mu = \int g d\mu$. This completes the proof. □

Claim 17.5. `(MON with almost sure convergence)` *If $0 \leq f_n \uparrow f$ everywhere on S except possibly some set $S_0 \subset S$ with $\mu(S_0) = 0$, then*

$$\int f_n d\mu \uparrow \int f d\mu.$$

Proof. Define $g_n := f_n \mathbb{I}_{S \setminus S_0}$ and $g := f \mathbb{I}_{S \setminus S_0}$. By construction $g_n \uparrow g$ everywhere, therefore we can apply `MON` as stated in Theorem 17.3 and obtain $\int g_n d\mu \uparrow \int g d\mu$. Since $\mu(S_0) = 0$ by Claim 17.4 we have $\int g_n d\mu = \int f_n d\mu$ and $\int g d\mu = \int f d\mu$. This, coupled with the result we obtained from `MON` completes the proof. □

From here on when referring to `MON` we will assume that the everywhere convergence as stated in Theorem 17.3 can be relaxed to almost everywhere convergence.

Claim 17.6. `(Linearity of integral)` Let $f, g \geq 0$ be measurable functions on $(S, \mathcal{F}, \mu)$, and let $a, b \geq 0$ be scalars. Then

$$\int (af + bg) d\mu = a \int f d\mu + b \int g d\mu.$$

□ Exercise 17.6.1. *Prove Claim 17.6 using approximation by monotone sequence of simple functions and using the* `MON` *theorem.*

Step 4. Measurable functions. For a measurable function f we split it into non-negative and non-positive components by writing $f = f^+ - f^-$ where

$$f^+ = \max(0, f) \quad \text{and} \quad f^- = \max(0, -f). \tag{17.5}$$

We now define $L^1(S, \mathcal{F}, \mu) =: L^1$ as the space of all measurable functions in $(S, \mathcal{F}, \mu)$ for which

$$\int \left(|f^+| + |f^-|\right) d\mu < \infty,$$

and define the integral of f as

$$\int f d\mu := \int f^+ d\mu - \int f^- d\mu.$$

This step completes the construction of Lebesgue integral.

We conclude this subsection with another theorem that we will need in the text. is the dominated convergence theorem (DOM).

Theorem 17.7. (Dominated convergence theorem (DOM)) *Let f_n, f be measurable function on $(S, \mathcal{F}, \mu)$ and $f_n \to f$ on S almost everywhere. Assume also that $|f_n| \leq g$ almost everywhere for some $g \in L^1$. Then $f_n \to f$ in L^1 i.e. $f \in L^1$ and*

$$\int |f_n - f| d\mu \to 0.$$

We will refer to this theorem as DOM. An important special case of the theorem, that can occur frequently, is when the sequence f_n is bounded by some constant.

17.2. **Probabilistic setup.** We now formulate the concepts and main results introduced in the previous section in a probabilistic language. As a measurable space we will take a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and for $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ will define its **expectation** as the (Lebesgue) integral of X with respect to the measure $\mathbb{P}$ as

$$\mathbb{E}X := \int_\Omega X(\omega) dP(\omega). \tag{17.6}$$

One can think about $\mathbb{E}X$, intuitively, as a *mean value* of X (discrete random variables illustrate this idea more evidently, cf. Definition 6.3). Next, we reformulate the main results discussed in the previous section for Lebesgue integral in the probabilistic setting now.

(MON) If X_n is a sequence of non-negative random variables that almost surely monotonically converges to a random variable X, i.e. $0 \leq X_n(\omega) \uparrow X(\omega)$ a.s., then $\mathbb{E}X_n \uparrow \mathbb{E}X$.

(DOM) Let X_n be a sequence of random variables such that $X_n \to X$ almost surely. If X_n is dominated by a random variable Y, i.e. $|X_n| \leq Y$ a.s. for all $n \in \mathbb{N}$, and if $\mathbb{E}Y < \infty$, then $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathbb{E}|X_n - X| \to 0$ as $n \to \infty$.

(bounded DOM) Let X_n be a sequence of random variables such that $X_n \to X$ a.s. If X_n is bounded, i.e. $|X_n| \leq K$ a.s. for all $n \in \mathbb{N}$ where K is some constant, then $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathbb{E}|X_n - X| \to 0$ as $n \to \infty$.

For an exponent $p \geq 1$ we define the usual $L^p(\Omega, \mathcal{F}, \mathbb{P}) =: L^p$ spaces as a space of all random variables X such that $|X|^p \in L^1$. As an L^p-norm of a random variable we set

$$||X||_p := (\mathbb{E}|X|^p)^{1/p}.$$

Here as well we have all the usual properties of L^p-norms, including the following

- (Hölder inequality) For random variables X, Y defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and exponents $1 \leq p, q \leq \infty$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$ one has

$$\mathbb{E}|XY| \leq (\mathbb{E}|X|^p)^{1/p} (\mathbb{E}|X|^q)^{1/q}.$$

- (Minkowski inequality) For random variables X, Y defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and an exponent $1 \leq p \leq \infty$ one has

$$||X + Y||_p \leq ||X||_p + ||Y||_p.$$

Minkowski's inequality is the triangle inequality of the L^p-norms, and from Hölder's inequality one can easily get the monotonicity of the L^p-norms, i.e. $||X||_p \leq ||X||_q$ if $1 \leq p \leq q \leq \infty$.

For elements of L^2 we also define the **variance** by

$$\mathrm{var}X := \mathbb{E}(X - \mathbb{E}X)^2,$$

which measures the mean quadratic deviation of a random variable from its mean. From the definition of the variance we have

$$\mathrm{var}X = \mathbb{E}X^2 - (\mathbb{E}X)^2 .$$

17.3. **Computation of expectation.** Here we discuss ways how one can compute expectation from a more practical viewpoint. The idea is that working directly with (17.6) might not be easy but in many scenarios we can transfer the integration from an abstract space such as $(\Omega, \mathcal{F}, \mathbb{P})$ to the real numbers where carrying out computations should be lot easier. The next result generalizes Theorem 6.4 (the law of subconscious statistician) to the case of non-discrete random variables.

Theorem 17.8. `(Change of variables formula for expectation)` Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X : \Omega \to \mathbb{R}$ be a random variable. Assume $h : \mathbb{R} \to \mathbb{R}$ is a Borel function for which $\mathbb{E}|h(X)| < \infty$. Then

$$\mathbb{E}h(X) = \int_{\mathbb{R}} h(x) d\mathbb{P}_X(x), \tag{17.7}$$

where $\mathbb{P}_X$ is the law (distribution) of X.

A few clarifying remarks before we move on to the proof.

- Here $h(X) = h \circ X$ is the superposition of h and X. Such situations arise regularly, e.g. when computing expectations of the form $\mathbb{E}X^2$ ($h(x) = x^2$), $\mathbb{E}e^X$ ($h(x) = e^x$), etc.
- Saying that h is a Borel function we mean that $h^{-1}(B) \in \mathcal{B}(\mathbb{R})$ for any $B \in \mathcal{B}(\mathbb{R})$, i.e. the preimages of Borel sets under h are also Borel. This implies, in particular, that $h \circ X$ is also a random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$.
- Recall also, that $\mathbb{P}_X$ is a probability measure defined on the Borel subsets of $\mathbb{R}$ by $\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B))$ where $X^{-1}(B) = \{\omega \in \Omega : \ X(\omega) \in B\}$ is the preimage of B under X as usual. Thus the integral in (17.7) is understood in the framework of the previous section: we have a measure space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$ and construct integral as in the previous section.

Proof of Theorem 17.8. Similar to the construction of the integral, here as well we split the proof into several steps starting from the indicator functions.

Step 1. Indicator functions. Let $h = \mathbb{I}_B$ where $B \in \mathcal{B}(\mathbb{R})$ is a given Borel set. Then

$$(h \circ X)(\omega) = \mathbb{I}_{X^{-1}(B)}(\omega), \quad \omega \in \Omega.$$

From here and the definition of integral for indicators we get

$$\mathbb{E}h(X) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}_X(B).$$

Similarly

$$\int_{\mathbb{R}} h(x) d\mathbb{P}_X(x) = \int_B 1 d\mathbb{P}_X(x) = \mathbb{P}_X(B),$$

which proves (17.7) for the case when h is an indicator of a Borel set.

Step 2. Simple functions. Let h be a simple function, i.e. $h = \sum_{k=1}^{m} a_k \mathbb{I}_{B_k}$ where $a_k \geq 0$ and $B_k \in \mathcal{B}(\mathbb{R})$ for all $k = 1, 2, ..., m$. We have

$$(h \circ X)(\omega) = \sum_{k=1}^{m} a_k \mathbb{I}_{X^{-1}(B_k)}.$$

Since the right-hand side of the above equality is a simple function, by definition of integral for simple functions we get

$$\mathbb{E}h(X) = \sum_{k=1}^{m} a_k \mathbb{P}(X^{-1}(B_k)).$$

In a similar vein we use the definition of integral but now for h on the measure space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$ arriving at

$$\int_{\mathbb{R}} h(x) d\mathbb{P}_X(x) = \sum_{k=1}^{m} a_k \mathbb{P}_X(B_k).$$

Since $\mathbb{P}_X(B_k) = \mathbb{P}(X^{-1}(B_k))$ by definition of $\mathbb{P}_X$, we get (17.7) for simple functions too.

Step 3. Non-negative functions. Now assume $h \geq 0$ is any Borel function. Let $\alpha^{(k)}$ be defined as in (17.4) and consider the sequence $h^{(k)} := \alpha^{(k)} \circ h$, $k \in \mathbb{N}$. By construction $h^{(k)}$ is simple and $h^{(k)} \uparrow h$. From here we get

$$\begin{aligned}
&\mathbb{E}(h \circ X) \text{ (by MON)} \\
&\quad = \lim_{k \to \infty} \mathbb{E}(h^{(k)} \circ X) \text{ (by the case of simple functions in Step 2)} \\
&\quad = \lim_{k \to \infty} \int_{\mathbb{R}} h^{(k)}(x) d\mathbb{P}_X(x) \text{ (by MON)} \\
&\quad = \lim_{k \to \infty} \int_{\mathbb{R}} h(x) d\mathbb{P}_X(x).
\end{aligned}$$

Step 4. The general case. We now complete the proof for general Borel function $h : \mathbb{R} \to \mathbb{R}$ that satisfy $\mathbb{E}|h \circ X| < \infty$. For such h write $h = h^+ - h^-$ where $h^+ = \max(h, 0)$ and $h^- = \max(-h, 0)$. Both h^+ and h^- are non-negative Borel functions and in view of the condition $\mathbb{E}|h \circ X| < \infty$ and definition of integral we have $\mathbb{E}h^+(X), \mathbb{E}h^-(X) < \infty$. Hence, applying the Step 3 we obtain

$$\begin{aligned}
&\mathbb{E}h(X) \text{ (by definition of integral)} \\
&\quad = \mathbb{E}h^+(X) - \mathbb{E}h^-(X) \text{ (by Step 3)} \\
&\quad = \int_{\mathbb{R}} h^+(x) d\mathbb{P}_X(x) - \int_{\mathbb{R}} h^-(x) d\mathbb{P}_X(x) \text{ (by definition of integral)} \\
&\quad = \int_{\mathbb{R}} h(x) d\mathbb{P}_X(x).
\end{aligned}$$

The proof is now complete. □

The formula (17.7) carries an important practical value by allowing to transfer the integral from an abstract space such as $(\Omega, \mathcal{F}, \mathbb{P})$ to the real line. However, we still have the law of X in the integral, which might not be easy to work with directly. In what

follows we establish other equivalent forms of (17.7) which can be useful for computations. The idea is to use other but equivalent representation of the measure induced by the law of X that we use in (17.7). To do that we will use Caratheodory's celebrated extension theorem from measure theory. We now start with some preliminaries.

Definition 17.1. `(Ring of sets)` *Let S be a non-empty set. A non-empty family of subsets $\mathcal{R}$ of S is called a* ***ring****, if it is closed under finite unions and relative complements, namely*

1. *if $A, B \in \mathcal{R}$ then $A \cup B \in \mathcal{R}$,*
2. *if $A, B \in \mathcal{R}$ then $A \setminus B \in \mathcal{R}$.*

Notice that from the definition of the ring we have that $\emptyset \in \mathcal{R}$ in view of the second condition and $\mathcal{R}$ being non-empty. We also have that a ring is closed under finite intersections, since $A \cap B = A \setminus (A \setminus B)$.

Clearly each σ-algebra is also a ring. The collection of any finite unions of half-open intervals $(a, b]$, where $a < b$, together with the empty set is an example of a ring which is not a σ-algebra.

Theorem 17.9. `(Caratheodory's extension)` *Let $\mathcal{R}$ be a ring and let $\mu : \mathcal{R} \to [0, +\infty]$ be a pre-measure on $\mathcal{R}$, i.e. whenever*[41] *$A \in \mathcal{R}$ and $A = \bigcup_n A_n$ where $A_n \in \mathcal{R}$ and $A_n \cap A_m = \emptyset$ for any $n \neq m$, we have $\mu(A) = \sum_n \mu(A_n)$. Then*

(1) *μ can be extended to $\sigma(\mathcal{R})$,*

(2) *if μ is σ-finite, i.e. there is a representation of S in the form of $S = \bigcup_n A_n$ with $A_n \in \mathcal{R}$ and $\mu(A_n) < \infty$, then the extension of μ is unique.*

This very important theorem in measure theory (see [2] or [9] for a proof) allows to define a measure on a relatively simple constructs such as a ring, and then extend it to more robust structures which σ-algebras represent. We will use this result here to find an equivalent forms of (17.7) that are more amenable to direct computations. In doing that we will first recall the definition of *Lebesgue-Stieltjes measure*.

Let F be the CDF of X and consider the ring $\mathcal{R}$ of semi-open intervals of $\mathbb{R}$ i.e. $\mathcal{R}$ consists of the empty set and all finite unions of intervals of the form $(a, b]$ with $a < b$. We now define $\mu : \mathcal{R} \to [0, \infty)$ using the function F.

For an interval $(a, b] \in \mathcal{R}$ where $a < b$ set

$$\mu_F((a, b]) := F(b) - F(a).$$

Since F is non-decreasing we have that the values of μ_F are non-negative. One can also show that μ_F is a pre-measure on the ring of semi-open intervals. Indeed, assume

$$(a, b] = \bigcup_{n=1}^{\infty} (a_n, b_n],$$

where $a < b$, $a_n < b_n$ and the intervals $(a_n, b_n]$ are non-intersecting. We may assume, without loss of generality, that the intervals $(a_n, b_n]$ are arranged so that

$$b_1 = b,\ b_2 = a_1,\ b_3 = a_2, ...,$$

[41] Notice that the only difference of pre-measure and a proper measure is that the underlying set $\mathcal{R}$ on which μ is defined is not necessarily a σ-algebra. Nevertheless, whenever it is possible to decompose a set in $\mathcal{R}$ into disjoint union of at most countable number of sets from $\mathcal{R}$ we must have σ-additivity of μ in such cases.

i.e. the left endpoint of the n-th interval coincides with the right endpoint of the $n+1$-th interval. With such arrangement of intervals we have

$$\begin{aligned}
&\sum_{n=1}^{\infty}\mu_F((a_n,b_n]) \text{ (by definition of } \mu)\\
&\quad=\sum_{n=1}^{\infty}(F(b_n)-F(a_n)) \text{ (by the arrangement of the intervals)}\\
&\quad=\lim_{n\to\infty}(F(b)-F(a_n)) \text{ (by right-continuity of } F)\\
&\quad=F(b)-F(a).
\end{aligned}$$

This shows that μ_F is indeed a pre-measure on $\mathcal{R}$. It is also easy to see that μ_F is σ-finite on $\mathbb{R}$, hence it has a unique extension to the Borel subsets of $\mathbb{R}$ thanks to Caratheodory's extension theorem 17.9. This extension is called *Lebesgue-Stieltjes measure*.

Getting back to the law of X defined by $\mathbb{P}_X$ we see that the Lebesgue-Stieltjes measure μ_F and $\mathbb{P}_X$ coincide on the ring $\mathcal{R}$, since by definition of $\mathbb{P}_X$ we have

$$\mathbb{P}_X((a,b])=\mathbb{P}(a<X\le b)=F(b)-F(a)=\mu_F\left((a,b]\right).$$

It follows that the Lebesuge-Stieltjes measure coincides with $\mathbb{P}_X$ and hence formula (17.7) can be rewritten in the form

$$\mathbb{E}h(X)=\int_{\mathbb{R}}h(x)dF(x).$$

In the same way, if the distribution of X has a density, i.e. there is a non-negative integrable function $\rho:\mathbb{R}\to\mathbb{R}_+$ such that

$$F(x)=\int_{-\infty}^{x}\rho(t)dt,\quad x\in\mathbb{R},$$

then the pre-measure $\mu_F\left((a,b]\right):=\int_a^b\rho(x)dx$ defined on the ring of half-open intervals $\mathcal{R}$, coincides with the law of X (as well as with μ_F) on $\mathcal{R}$ and hence extends to the same measure $\mathbb{P}_X$. Thus we get the following representations

$$\mathbb{E}h(X)=\int_{\mathbb{R}}h(x)d\mathbb{P}_X(x)=\int_{\mathbb{R}}h(x)dF(x)=\int_{\mathbb{R}}h(x)\rho(x)dx, \tag{17.8}$$

where the last representation is valid if X has a density.

□ Exercise 17.9.1. *Let $a<b$ be fixed and assume $(a,b]=\bigcup_{n=1}^{\infty}(a_n,b_n]$ where all intervals $(a_n,b_n]$ are disjoint. Prove that there is a rearrangement of intervals $\{(a_n,b_n]\}_{n=1}^{\infty}$, i.e. a bijection $\sigma:\mathbb{N}\to\mathbb{N}$, such that*

$$b_{\sigma(1)}=b, a_{\sigma(1)}=b_{\sigma(2)}, a_{\sigma(2)}=b_{\sigma(3)},...,$$

i.e. we can arrange the intervals in some order so that the left endpoint of the n-th interval coincides with the right endpoint of the $n+1$-th, and the first one coincides with b. In particular, show that in this arrangement $a_{\sigma(n)}\downarrow a$.

17.4. **Expectation and independence.** Recall that expectation acts linearly on a sum of two random variables. For a product of random variables, we saw in the discrete case (Theorem 7.5) that expectation is multiplicative for *independent* random variables. Here we prove this property in the general case.

Theorem 17.10. (`Expectation for product of independent random variables`) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X, Y : \Omega \to \mathbb{R}$ be random variables that are both in L^1. If X and Y are independent, then $\mathbb{E}(XY) = \mathbb{E}X\mathbb{E}Y$.*

Proof. Keeping in mind the construction of integral and decomposing X and Y into non-negative parts as $X = X^+ - X^-$ and $Y = Y^+ - Y^-$ (see (17.5)) we can reduce the problem to the case of non-negative random variables. Thus, we will assume without loss of generality that $X, Y \geq 0$.

Let $\alpha^{(k)}$ be defined as in (17.4). Then $X^{(k)} := \alpha^{(k)} \circ X$ and $Y^{(k)} := \alpha^{(k)} \circ Y$ are simple random variables monotonically converging to X and Y correspondingly. We also have

$$X^{(k)} = \sum_{i=1}^{m} a_i \mathbb{I}_{A_i} \text{ and } Y^{(k)} = \sum_{j=1}^{n} b_j \mathbb{I}_{B_j},$$

where $A_i \in \sigma(X)$ and $B_j \in \sigma(Y)$ by definition of $\alpha^{(k)}$. We thus get

$$\begin{aligned}
\mathbb{E}(X^{(k)}Y^{(k)}) &= \mathbb{E}\left(\sum_{i,j} a_i b_j \mathbb{I}_{A_i}\mathbb{I}_{B_j}\right) \\
&= \sum_{i,j} a_i b_j \mathbb{E}(\mathbb{I}_{A_i}\mathbb{I}_{B_j}) \\
&= \sum_{i,j} a_i b_j \mathbb{E}(\mathbb{I}_{A_i \cap B_j}) \\
&= \sum_{i,j} a_i b_j \mathbb{P}(A_i \cap B_j) \text{ (by independence of } X \text{ and } Y) \\
&= \sum_{i,j} a_i b_j \mathbb{P}(A_i)\mathbb{P}(B_j) \\
&= \left(\sum_i a_i \mathbb{P}(A_i)\right)\left(\sum_j b_j \mathbb{P}(B_j)\right) \\
&= \mathbb{E}X^{(k)}\mathbb{E}Y^{(k)}.
\end{aligned}$$

Now using `MON` the right-hand side of the last expression converges to $\mathbb{E}X\mathbb{E}Y$ which is finite by the condition of the theorem. On the other hand for the left-hand side we have $X^{(k)}Y^{(k)} \uparrow XY$, hence $\mathbb{E}X^{(k)}\mathbb{E}Y^{(k)} \uparrow \mathbb{E}(XY)$, again thanks to `MON`. Comparing the limits of two sides completes the proof. □

The same argument works for $n \geq 2$ independent random variables too. We formulate the result and will leave the proof as an exercise.

Theorem 17.11. (`Expectation for product of` n `independent random variables`) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X_1, X_2, ..., X_n : \Omega \to \mathbb{R}$ be independent random variables that are all in L^1. Then $X_1 \cdot ... \cdot X_n$ is also in L^1 and*

$$\mathbb{E}(X_1 X_2 \cdot ... \cdot X_n) = \mathbb{E}X_1 \mathbb{E}X_2 \cdot ... \cdot \mathbb{E}X_n.$$

□ Exercise 17.11.1. *Prove Theorem 17.11 using*

(a) *the arguments we had in the proof of Theorem 17.10,*

(b) *induction and the results of Exercise 15.1.2.*

□ Exercise 17.11.2. *Construct an example of two random variables X, Y defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that both are L^1 but their product is not.*

→ **Problem 17.11.1.** *A particle starts at the origin of $\mathbb{R}^2$ and at each time step chooses a direction uniformly at random and independently from previous moves, and moves a unit distance in that direction (i.e. if the particle is at point $M \in \mathbb{R}^2$ then for its next move it chooses a point from a uniform distribution on a sphere of radius 1 centered at M and moves there). What is the probability that the particle will be in a disc of radius 1 centered at the origin after moving for $n \geq 1$ steps?*

18. Strong Law of Large Numbers

This section is devoted to the *strong law of large numbers* (**SLLN** for short), one of the most important limit theorems in probability theory. Here we prove SLLN for a large class of random variables and study various scenarios in which the law or ideas related to it can be applied.

18.1. **SLLN.** Assume we toss a fair coin for a certain number of times, say $n \geq 1$. We then record the number of times we saw **H** (Heads) or **T** (Tails) during these tosses. How are we supposed to see the implications of the *fairness* of the coin? The first guess might be that we will have an approximately equal number of **H** and **T**. This thought, however, can easily be seen to be incorrect. Indeed, in $2n$ tosses with a fair coin the probability that we will get **H** n times, in view of independence of the tosses, equals

$$\frac{1}{2^{2n}}\binom{2n}{n} \sim \frac{1}{\sqrt{\pi}}\frac{1}{\sqrt{n}}, \text{ as } n \to \infty,$$

where the asymptotics follows easily from Stirling's formula (2.1). We see that there is a very small chance to get equal number of *Heads* and *Tails* in a long series of tosses. One can also show that the chances to have the absolute difference of the numbers of **H** and **T** within a given fixed bounds will also be small.

A proper relaxation of the setting above will be to follow the asymptotic frequency of seeing **H** or **T**. Namely, instead of comparing the absolute number of times we see these events, in a series of n tosses we compute the number of times that *Heads* appear and divide that number by n - the total number of experiments. With this approach, as we will see soon, the fairness of the coin is duly reflected as the limit of such averages converges to $1/2$ (see Figure 15 for a numerical simulation).

There are many results in probability theory that study such asymptotic behavior of a sequence of random variables, usually going under the name **limit theorems**. One such example that we have already encountered in this notes, is the *weak law of large numbers* (WLLN) as discussed in section 9. We will now add another very important result from this class of *limit theorems*.

Theorem 18.1. `(Strong Law of Large Numbers (SLLN), easy version)` *Consider a sequence of independent identically distributed (i.i.d.) random variables $X_1, ..., X_n, ...$ where $\mathbb{E}X_i = \mu$ and $\mathbb{E}X_i^4 \leq K$ for all $i \in \mathbb{N}$ with some constant K. Denote $S_n := X_1 + ... + X_n$, $n \in \mathbb{N}$, then*

$$\frac{S_n}{n} \to \mu \text{ a.s. as } n \to \infty.$$

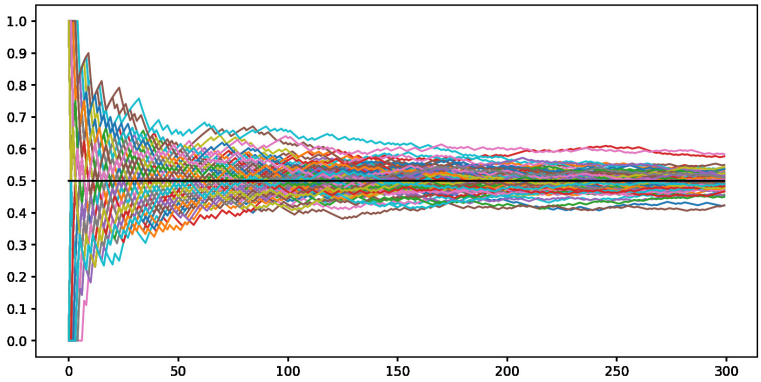

FIGURE 15. Numerical simulation of 100 experiments of tossing a fair coin 300 times in a raw. Each colored graph is a particular realization of the coin tossing experiment, one of the 100 experiments in total. For each such graph, the value at discrete times $t = 1, 2, ..., 300$ is the average number of Heads seen up to that time. Notice, that these averages while fluctuating at the earlier times cluster around the value of 0.5 (the black line on the figure) as predicted by the SLLN.

A few remarks are in order before we move on to the proof of this result.

- `Strong vs Weak law:` We know from Lemma 14.1 that if a sequence of random variables converges almost surely then it also converges in probability, in other words convergence almost surely is a stronger form of convergence than convergence in probability. Thus, whenever the strong law is applicable so is the weak law, hence the naming.
- `On the bound of fourth moments:` The bound $\mathbb{E}X_i^4 \leq K$ that we have in the formulation is not necessary for the conclusion of the theorem to hold. The estimate is for technical reasons, to facilitate a shorter streamlined proof of the result. It does, nevertheless, contain a large class of cases where the current formulation will suffice. In general, one only needs integrability of the random variables, i.e. it suffices to have $\mathbb{E}|X_i| < \infty$. See [2] for more details.

Proof of Theorem 18.1. Without loss of generality we may assume that $\mu = 0$. Indeed, set $Y_i := X_i - \mu$. Then $\mathbb{E}Y_i = 0$ and using Minkowski's inequality (triangle inequality for L^p norms) we get[42] $\mathbb{E}Y_i^4 \leq (||X_i||_4 + ||\mu||_4)^4 \leq \left(K^{1/4} + |\mu|\right)^4 =: K_1$. Thus $\{Y_i\}$ satisfies the requirements of theorem and

$$\frac{Y_1 + ... + Y_n}{n} = \frac{S_n}{n} - \mu.$$

Given this, we will assume that $\mu = 0$. Next, we estimate $\mathbb{E}S_n^4$. We have

$$S_n^4 = \sum_{1 \leq i,j,k,m \leq n} X_i X_j X_k X_m.$$

To study the general term in this sum we consider few cases. Namely

[42]Note, that a simpler inequality, which is

$$(a+b)^p \leq (2\max(a,b))^p \leq 2^p(a^p + b^p),$$

for any $a, b \geq 0$ and $p \geq 0$, will suffice here.

- `all indicies` i,j,k,m `are different:` Then, by independence of the random variables and the assumption that $\mu = 0$ we get
$$\mathbb{E}\left(X_iX_jX_kX_m\right) = \mathbb{E}X_i\mathbb{E}X_j\mathbb{E}X_k\mathbb{E}X_m = 0$$
where we applied Theorem 17.11 to compute the expectation of the product.
- `one index out of` i,j,k,m `is different from the other three:` This case is similar to the previous one. Indeed, assume $i \neq j,k,m$ then
$$\mathbb{E}\left(X_iX_jX_kX_m\right) = \mathbb{E}X_i\mathbb{E}\left(X_jX_kX_m\right) = 0.$$
- `all four indicies are the same:` This leads to terms of the form $\mathbb{E}X_i^4$ for all $i = 1,2,...,n$.
- `there are only two different indicies out of four:` To count the possible configurations here we will simply count the number of ways i can be paired with any of the j,k,m, where by pairing we mean that the two indicies in the pair will have the same value. Since in this case exactly two out of the four indicies are different, fixing the pair for i will automatically define the second pair. Thus we have three different options, namely $i = j \neq k = m$, $i = k \neq j = m$ and $i = m \neq j = k$. Overall the terms we get here are the following
$$3\sum_{1\leq i\neq j\leq n}\mathbb{E}X_i^2\mathbb{E}X_j^2,$$
where in view of the random variables being identically distributed all three arrangements of the pairings above lead to the same sum.

Putting the above cases together we get

$$\begin{aligned}\mathbb{E}S_n^4 &= \sum_{i=1}^n \mathbb{E}X_i^4 + 3\sum_{1\leq i\neq j\leq n}\mathbb{E}X_i^2\mathbb{E}X_j^2\\ &\leq nK + 3n(n-1)K \leq 3n^2K,\end{aligned}$$

where we used the estimate $\mathbb{E}X_i^2 \leq \left(\mathbb{E}X_i^4\right)^{1/2}$ which is due to Hölder's inequality. We thus obtain that

$$\mathbb{E}\left(\frac{S_n}{n}\right)^4 \leq 3K\frac{1}{n^2}. \tag{18.1}$$

Now consider the series

$$Z(\omega) := \sum_{n=1}^{\infty}\left(\frac{S_n(\omega)}{n}\right)^4,\ \omega\in\Omega,$$

and let $Z_N(\omega)$ be its N-th partial sum. Since the general term of the series is non-negative we have $Z_N(\omega) \uparrow Z(\omega)$ for any $\omega \in \Omega$, hence by `MON` we get $\mathbb{E}Z_N \uparrow \mathbb{E}Z$. We have finite number of terms in Z_n and can compute

$$\mathbb{E}Z_N = \sum_{n=1}^N \mathbb{E}\left(\frac{S_n}{n}\right)^4 \uparrow \sum_{n=1}^{\infty}\mathbb{E}\left(\frac{S_n}{n}\right)^4 < \infty,$$

where the convergence of the last series is due to the estimate (18.1). Since we already saw that $\mathbb{E}Z_N \uparrow \mathbb{E}Z$ it follows that

$$\mathbb{E}Z = \sum_{n=1}^{\infty}\mathbb{E}\left(\frac{S_n}{n}\right)^4 < \infty.$$

As $\mathbb{E}Z < \infty$ it follows that Z is finite almost surely in view of the definition of the integral as can be seen from (17.3). The latter implies that the series in the definition of Z must converge almost surely, and hence the general term of the series must converge to 0 whenever the series converges. Thus, for any $\omega \in \Omega$ for which $Z(\omega) < \infty$ we get that

$$\frac{S_n(\omega)}{n} \to 0, \text{ as } n \to \infty.$$

Since $Z(\omega) < \infty$ almost surely, the proof of the theorem is complete.

□

□ EXERCISE 18.1.1. *Let $X_1, X_2, ..., X_n, ...$ be a sequence of independent random variables all distributed as $U(0,1)$. Let $k \geq 1$ be fixed. Prove that the sequence*

$$Z_n := \frac{X_1^k + ... + X_n^k}{X_1 + ... + X_n}$$

converges almost surely and compute the limit of Z_n in L^1.

Hint: Use the SLLN for the first part after adjusting the ratio. For the second part consider using `DOM`.

→ **Problem 18.1.1.** *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of i.i.d. random variables and denote $S_n := X_1 + ... + X_n$, $n = 1, 2, ...$. Assume that for some random variable Y one has*

$$\frac{S_n}{n} \to Y \text{ almost surely as } n \to \infty.$$

Prove that $\mathbb{E}|X_1| < \infty$.

Note: This is in some sense the inverse to SLLN.

Hint: First show that $X_n/n \to 0$ almost surely. Using this, conclude from Borel-Cantelli's second lemma that $\sum_{n\geq 1} \mathbb{P}(|X_n| \geq n) < \infty$. Then establish a bound from above on $\mathbb{E}|X| \leq 1 + \sum_{n\geq 1} \mathbb{P}(|X| \geq n)$ for any random variable X.

18.2. **Applications of SLLN.** Strong law of large numbers is a result of fundamental importance in probability theory and in many applications. We now discuss few scenarios where SLLN or ideas around it can be applied.

Computing areas and integrals. Here we discuss on simple examples how the strong law of large numbers (SLLN) can have practical implications in computing integrals. We start with the question of approximating the value of π, which we will formulate as a problem of estimating an area of a certain 2-dimensional region that relates to π.

Let Δ be the region bounded by the arc of a unit circle from points $(0,1)$ to $(1,0)$ and coordinate axes (see Figure 16). Clearly, the area of Δ equals $\pi/4$. To approximate this value we take a sequence of independent random vectors $\mathbf{X}_1, \mathbf{X}_2, ...$ that are all uniformly distributed on the square[43] $[0,1] \times [0,1]$. Now define $Y_n := \mathbb{I}_{\mathbf{X}_n \in \Delta}$ for $n \in \mathbb{N}$. Since the surface area of Δ is $\pi/4$ we get $\mathbb{P}(Y_n = 1) = \pi/4$. Also, since $\mathbf{X}_i$-s are independent we also have that Y_n-s are also independent thanks to Lemma 15.1. Then SLLN implies that almost surely

$$\frac{Y_1 + ... + Y_n}{n} \to \frac{\pi}{4} \quad \text{as } n \to \infty.$$

Now the practical implication of the above is the following. Sample random points from $[0,1] \times [0,1]$ independent of each other and all uniformly distributed on the unit

[43]The sequences made of each coordinates of $\{\mathbf{X}_i\}$, in this case two such sequences, are both sequences of independent random variables uniformly distributed on $[0,1]$.

square. Then compute the number of such points that fall into Δ. This number divided by n - the total number of sampled points will approximate $\pi/4$ if n is large enough.

An important caveat that one needs to keep in mind, is that SLLN only gives almost sure convergence, which does not eliminate the case that a single realization of the random experiment, as we had above, might end up in the non-convergence set, which has zero probability, but is not necessarily empty. This, however, is usually disregarded from a practical perspective. The scheme for obtaining numerical values of certain quantities via repeated random sampling that we discussed above is called the **Monte Carlo method**, one of the variety of sampling techniques.

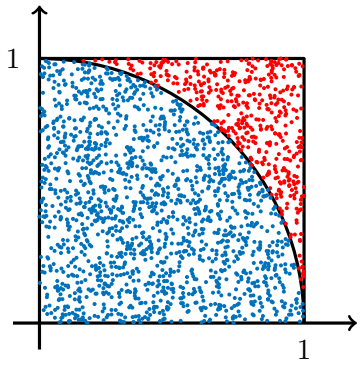

FIGURE 16. Approximation of the value of π (best seen online). All points seen on this figure are sampled uniformly and independently from $[0,1] \times [0,1]$. Points that fall in Δ - the region bounded by the arc and the coordinate axes are colored in blue, and those that are not are colored in red. The number of blue points divided by the total number of points, i.e. the average number of times a point appears in the arc approximate $\pi/4$ - the area of Δ. In this example the number of sampled points is 2023 out of which 1583 fell into Δ, i.e. are blue points. This leads to $1583/2023 \approx 0.7825$ average frequency of blue points, i.e. the approximate value of π becomes $4 * 0.7825 = 3.13$ which is close to the actual value of π.

As another example, we will estimate the numerical value of the integral $I := \int_0^1 e^{x^2} dx$, which is an example of an integral that cannot be computed via elementary functions. The approach is similar to the previous example: consider a sequence $X_1, X_2, ...$ of independent random variables that are all uniformly distributed on the interval $(0,1)$. Let $f(x) = e^{x^2}$ and define $Y_i := f(X_i)$. Due to Lemma 15.1 the sequence of random variables $\{Y_i\}$ is also independent. Moreover, the distribution of each Y_i is the same and by (17.8) we have

$$\mathbb{E}Y_i = \int_0^1 e^{x^2} dx = I.$$

Now the SLLN implies that almost surely we have

$$\frac{Y_1 + ... + Y_n}{n} \to \int_0^1 e^{x^2} dx, \text{ as } n \to \infty.$$

To compute the limit numerically, as in the previous example, we can sample points from $[0,1]$ uniformly and independently of each other, and use the numerical value of the above fraction as an approximation to the value of I.

□ Exercise 18.1.2. *Check that in both examples above the expectation of the fourth powers of random variables (the fourth moments) are uniformly bounded. This is the condition under which we proved Theorem 18.1.*

Normal numbers. Here we present another application of SLLN this time in number theory related to *normal numbers*, a concept introduced by É. Borel.

Fix an integer $b \geq 2$ and for a number $x \in [0,1]$ let

$$x = \sum_{n=1}^{\infty} \frac{x_n}{b^n}$$

be its expansion in base b (recall the dyadic expansion from section 15.3 corresponding to the case of $b = 2$) where $x_n \in \{0, 1, ..., b-1\}$. The number x is called **simply normal in base** b if for any $d \in \{0, 1, ..., b-1\}$ we have

$$\frac{|\{1 \leq i \leq n : x_i = d\}|}{n} \to \frac{1}{b} \text{ as } n \to \infty,$$

in other words the asymptotic frequency of occurrence of each digit d in the expansion of x equals $1/b$. Intuitively, the definition says that asymptotically no digit appears more than any other one. For example, any rational number in base 10 is not simply normal, or any point of the Cantor set on $[0,1]$ is not normal in base 3 (prove these two assertions). One can extend this definition to any finite pattern of symbols instead of just one character. Namely, let $\mathbf{w} = (w_1, ..., w_k)$ be any sequence of digits where $w_i \in \{0, 1, ..., b-1\}$ for all $1 \leq i \leq k$. Then the number x is called **normal in base** b if

$$\frac{|\{1 \leq i \leq n : (x_i, ..., x_{i+k-1}) = \mathbf{w}\}|}{n} \to \frac{1}{b^k} \text{ as } n \to \infty,$$

meaning that the asymptotic frequency of seeing $\mathbf{w}$ as a substring in the sequence of digits of the expansion of x equals $1/b^k$ for any $\mathbf{w}$. Again, intuitively the definition asserts that the asymptotic frequency of the substring $\mathbf{w}$ is determined by its length only and is not affected by the actual pattern of $\mathbf{w}$. Finally, the number x is called **normal** if it is normal in any integer base $b \geq 2$.

Remarkably, one can show that almost any number in $[0,1]$ is normal[44]. This is known as Borel's theorem. We now prove this result using SLLN.

Theorem 18.2. `(Almost all numbers are normal)` *The Lebesgue measure of the set* $\{x \in [0,1] : \ x \text{ is normal}\}$ *is* 1.

Proof. We will first show that for a fixed base $b \geq 2$ the measure of simply normal numbers is 1. This of course is a weaker statement than the normality of the numbers, but will illustrate the main ideas used in proving the full statement.

Fix an integer $b \geq 2$ as a base for expansion and consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where $\Omega = [0,1]$, $\mathcal{F}$ is the Borel σ-algebra of $[0,1]$ and $\mathbb{P}$ is the Lebesgue measure on $[0,1]$. Then, for a point $\omega \in \Omega$ consider its expansion in base b, namely

$$\omega = \sum_{n=1}^{\infty} \frac{X_n(\omega)}{b^n},$$

[44]While showing that almost all numbers are normal is relatively straightforward as we see with Theorem 18.2, providing a concrete example of a normal number is notoriously hard. It is not known, as of this writing, if the constants π, e, or $\sqrt{2}$ are normal or not. A well-known normal number is the so-called **Chaitin's constant** that represents the halting probability of a universal machine, informally speaking it is the probability that a randomly constructed program will halt. See *V. Becher, S. Figueira, An example of a computable absolutely normal number, Theoretical Computer Science, Volume 270, Issues 1-2, 6 (2002), pp. 947-958* for an overview and more details.

where $X_n \in \{0, 1, ..., b-1\}$. Similarly to Lemma 15.11, here as well we have that $\{X_n\}$ forms an independent and identically distributed (i.i.d.) sequence. Hence the sequence $Y_i := \mathbb{I}_{X_i=d}$ where $d \in \{0, 1, ..., b-1\}$ is fixed, is also an i.i.d. sequence in view of Lemma 15.1, moreover, we have $\mathbb{P}(Y_i = 1) = 1/b$ and $\mathbb{E}Y_i = 1/b$. Now applying SLLN on $\{Y_i\}$ we get

$$\mathbb{P}\left(\omega \in \Omega : \ \frac{Y_1(\omega) + ... + Y_n(\omega)}{n} \to \frac{1}{b}\right) = 1.$$

Recall that $\mathbb{P}$ is the Lebesgue measure on $[0, 1]$ and the sum $Y_1(\omega) + ... + Y_n(\omega)$ equals the number of times digit d occurred in the first n terms of the b-adic expansion of ω. Hence from the last relation we obtain that the Lebesgue measure of numbers from $[0, 1]$ that are *simply normal* in base b equals 1.

We now show the stronger claim, that the set of numbers from $[0, 1]$ that are *normal* in base b is 1. Fix a finite sequence of symbols $\mathbf{w} = (w_1, ..., w_k)$ where $k \geq 1$ and $w_i \in \{0, 1, ..., b-1\}$. Keeping the notation for $\{X_n\}$ as above define

$$Y_i := \mathbb{I}_{(X_i,...,X_{i+k-1})=\mathbf{w}}, \ i = 1, 2, ...,$$

i.e. $Y_i = 1$ if we see the sequence $\mathbf{w}$ starting from position i. Due to the independent of $\{X_n\}$ we have $\mathbb{E}Y_i = 1/b^k$, however, unless $k = 1$, i.e. $\mathbf{w}$ has length one, the sequence $\{Y_i\}$ is not independent and hence we cannot apply SLLN directly. To overcome this obstacle, we will split $\{Y_i\}$ into groups of sequences that are independent. Observe that the subsequence $Y_1, Y_{k+1}, ..., Y_{mk+1}, ...$ of $\{Y_i\}$ is independent since the all X_i-s that are present in these Y_i-s are different from each other. Thus, we can group Y_i-s that are k-index apart from each other. More precisely, set $Y_0 = 0$ and observe that

$$Y_1 + ... + Y_n = \sum_{r=0}^{k-1} \sum_{m=0}^{\lfloor \frac{n-r}{k} \rfloor} Y_{km+r},$$

where $\lfloor x \rfloor$ denotes the integer part of x. In other words, indicies with common remainder when divided by m are grouped. By construction, for each $0 \leq r \leq k-1$ the sequence $\{Y_{km+r}\}_{m=0}^{\infty}$ is independent. Notice also that as $n \to \infty$ we have

$$\frac{1}{n}\left\lfloor \frac{n-r}{k} \right\rfloor \to \frac{1}{k},$$

hence from the SLLN we get with probability one that

$$\frac{Y_1 + ... + Y_n}{n} = \sum_{r=0}^{k-1} \underbrace{\frac{1}{n}\left\lfloor \frac{n-r}{k} \right\rfloor}_{\to \frac{1}{k}} \underbrace{\frac{1}{\lfloor \frac{n-r}{k} \rfloor} \sum_{m=0}^{\lfloor \frac{n-r}{k} \rfloor} Y_{km+r}}_{\to \frac{1}{b^k} \text{ by SLLN}} \to \frac{1}{b^k}.$$

This proves that the probability of a number being normal in base b is one. Since the number of bases is countable we get that the probability of a number being normal in any base is also one as the countable intersection of probability 1 events is also of probability 1. This completes the proof of the theorem. □

Geometric implications of SLLN. Consider a sequence of independent random variables $X_1, ..., X_n$ all uniformly distributed on the interval $(-1, 1)$. Clearly $\mathbb{E}X_i^2 = 1/3$, and $\mathbb{E}X_i^4 = 1/5$ for all $i = 1, ..., n$. Hence the requirements of Theorem 18.1 (SLLN)

hold for the sequence $\{X_i^2\}$ and since almost sure convergence implies convergence in probability (see Lemma 14.1) we also have the implication of the WLLN, namely

$$\mathbb{P}\left(\left|\frac{X_1^2+...+X_n^2}{n}-\frac{1}{3}\right|\leq\frac{\delta}{3}\right)\to 1,\ \text{as } n\to\infty,$$

for any fixed $\delta>0$. Regrouping the terms in the inequality we obtain

$$\mathbb{P}\left(\frac{n}{3}(1-\delta)\leq X_1^2+...+X_n^2\leq\frac{n}{3}(1+\delta)\right)\to 1.$$

Now observe that this two-sided inequality defines an annulus in $\mathbb{R}^n$ and the geometric interpretation of the converge in probability above is the following: the volume of the cube

$$\mathcal{C}:=\underbrace{[-1,1]\times...\times[-1,1]}_{n\ times}=:[-1,1]^n$$

is almost entirely concentrated in the annulus with radii $(1+\delta)n/3$ and $(1-\delta)n/3$. This, somewhat counterintuitive implication is due to effects that are typical to higher-dimensional spaces and is a particular manifestation of a general phenomenon called **concentration of measure**. A few remarks are in order.

- Notice, that in the example above we tacitly moved from the underlying probability measure to the Lebesgue measure when discussing volumes, as if they were equivalent. To formally justify this consider the probability space $(\Omega,\mathcal{F},\mathbb{P})$ on which $X_1,...,X_n$ are defined, and let $\mathcal{L}_n$ be the normalized (i.e. with total volume 1) Lebesgue measure on the cube $\mathcal{C}$. Now the joint law of $(X_1,...,X_n)$ coincides with $\mathcal{L}_n$ on a π-system of n-dimensional rectangles in $\mathcal{C}$ thanks to the independence of $\{X_i\}$ and each X_i being uniformly distributed on $(-1,1)$. It is left to use Lemma 15.2 on unique continuation from π-system to minimal σ-algebra.
- Another example that shows how volumes are concentrated in higher-dimensional spaces is the following. Consider the unit cube $[0,1]^n$ and let $\delta>0$ be a small fixed number. Then, the volume of $[\delta,1-\delta]^n$ equals $(1-2\delta)^n$ which tends to 0 as $n\to\infty$. But the volume of the unit cube is 1, hence this means that the entire volume of the cube is concentrated near its boundaries if the dimension of the space is large. A probabilistic interpretation of the above can be the following: take independent random variables $X_1,...,X_n$ all with distribution $U(0,1)$. Now the vector $\mathbf{X}:=(X_1,...,X_n)\in(0,1)^n$ is uniformly distributed in the unit cube. In order that $\mathbf{X}$ stays at least δ away from the boundary of the cube we need $\delta<X_i<1-\delta$ for all coordinates. For each coordinate the probability of this event is $1-2\delta$, and for all coordinates, in view of the independence, the probability of staying away from the boundary becomes $(1-2\delta)^n$ which converges to 0 as $n\to\infty$. Thus, sampling a point uniformly at random from a cube will most likely be close to the boundary of the cube.

□ Exercise 18.2.1. *Convince yourself that you can carry out the necessary steps discussed in the first remark above using the machinery of π-systems to interpret the probability as Lebesgue measure in that example.*

□ Exercise 18.2.2. *Consider a high-dimensional cube $[0,1]^n$, with $n\geq 2$. Let* H *be the hyperplane passing through the center of the cube, i.e. point $(1/2,...,1/2)\in\mathbb{R}^n$ and being orthogonal to the main diagonal of the cube, i.e. the vector $(1,...,1)\in\mathbb{R}^n$. Take $\delta>0$ and consider a strip of width δ around* H. *Show that if n is large then the*

entire volume of the cube is concentrated in that strip, i.e. the volume of points of the cube that are not from that strip converge to 0 *as* $n \to \infty$.

Hint: Model this problem in probabilistic framework, using the ideas discussed in this subsection, so that SLLN can be applied.

→ **Problem 18.2.1.** *Let* $X_1, X_2, ..., X_n$ *be independent all with* $N(0,1)$ *normal distribution. Consider the random vector*

$$\boldsymbol{Z} := \left(\frac{X_1}{||\boldsymbol{X}||_2}, ..., \frac{X_n}{||\boldsymbol{X}||_2} \right),$$

where $\boldsymbol{X} = (X_1, ..., X_n)$ *and* $||\boldsymbol{X}||_2 = \left(X_1^2 + ... + X_n^2\right)^{1/2}$. *Clearly the values of* $\boldsymbol{Z}$ *are points on the unit sphere* $\mathbb{S}^{n-1}$ *of* $\mathbb{R}^n$. *Prove that* $\boldsymbol{Z} \sim U(\mathbb{S}^{n-1})$, *i.e. that* $\boldsymbol{Z}$ *is uniformly distributed on the unit sphere of* $\mathbb{R}^n$. *More precisely, for any Borel set* $B \subset \mathbb{S}^{n-1}$ *prove that* $\mathbb{P}(\boldsymbol{Z} \in B) = \sigma(B)$, *where* σ *is the normalized Lebesgue measure (surface measure) on* $\mathbb{S}^{n-1}$, *and* $\mathbb{P}$ *is the probability measure of the space where all* X_i*-s are defined.*

Note: the scheme above provides a way to sample points from the unit sphere uniformly at random. Notice that while it is clear how the uniform sampling from the unit cube $[0,1]^n$ *works, sampling uniformly from a curved surface, such as the sphere, is not trivial.*

Hint: Show that the distribution (law) of $\boldsymbol{Z}$ *is rotation-invariant, i.e. for any orthogonal* $n \times n$ *matrix* O *prove that* $\mathbb{P}(\boldsymbol{Z} \in B) = \mathbb{P}(\boldsymbol{Z} \in OB)$. *Then use the Haar's theorem on uniqueness of the translation-invariant measure to conclude that the law of* $\boldsymbol{Z}$ *must coincide with the normalized Lebesgue measure on the sphere as it is the only measure with surface area equal to* 1 *that is invariant under rotations.*

→ **Problem 18.2.2.** *Let* $f \in C[0,1]$. *Compute the limit*

$$\lim_{n\to\infty} \underbrace{\int_0^1 ... \int_0^1}_{n \text{ times}} f(\sqrt[n]{x_1 \cdot ... \cdot x_n}) dx_1 ... dx_n.$$

Hint: Consider a sequence of random variables $X_1, ..., X_n, ...$ *that are all uniformly distributed on the interval* $(0,1)$. *Then observe that SLLN can be used to compute the limit of* $\log \sqrt[n]{X_1 \cdot ... \cdot X_n}$ *as* $n \to \infty$. *Then see how the integral above can be computed as an expectation involving the random variables* $\{X_i\}$.

19. Concentration Inequalities

The main theme of this section is **concentration inequalities**, which in probability theory provide bounds on deviation of a random variable from a certain value, such as its mathematical expectation. For example, the law of large numbers, as we saw earlier, states that the sums of independent random variables center around their mean with high probability. Concentration inequalities help to quantitatively assess such clustering effects. We will revisit the classical inequalities of Markov and Chebishev proved earlier in a new light, and study some applications of them. We will then add a new inequality due to Chernoff with applications to some problems in higher-dimensional Euclidean spaces.

19.1. **Deviation from the mean.** In this section we will study some very basic forms of estimating divergence of a random variable from its mathematical expectation, or equivalently the concentration of a random variable around its mean. We start by recalling the inequalities we proved section 9.1 for discrete random variables.

▶ `(Markov)` Let X be a random variable. Then for any $t > 0$

$$\mathbb{P}(|X| \geq t) \leq \frac{1}{t}\mathbb{E}|X|. \tag{19.1}$$

▶ `(Chebishev)` Let X be a random variable. Then for any $p > 0$ and any $t > 0$

$$\mathbb{P}(|X| \geq t) \leq \frac{1}{t^p}\mathbb{E}|X|^p. \tag{19.2}$$

▶ `(Chebishev for variance)` Let X be a random variable. Then for any $t > 0$

$$\mathbb{P}(|X - \mathbb{E}X| \geq t) \leq \frac{1}{t^2}\mathrm{var}X. \tag{19.3}$$

The proofs of the inequalities above are the same as in the case of discrete random variables and will be omitted. Inequality (19.2) is new, its proof, however, is the same as the proof of Chebishev's inequality we had in Theorem 9.2. Observe that inequality (19.3) from the list above provides a way of measuring the concentration of a random variable around its mean. Although in a very basic form, the inequality is still quite useful and we will now discuss some applications of it.

Theorem 19.1. `(`L^2`-version of WLLN)` *Let $X_1, X_2, ...$ be a sequence of random variables that are pairwise uncorrelated. Assume they all have the same finite mean μ and have variance uniformly bounded above, i.e. there exists a constant $C > 0$ such that $\mathrm{var}X_k \leq C$ for all $k \in \mathbb{N}$. Then*

$$\frac{X_1 + ... + X_n}{n} \to \mu \text{ in } L^2.$$

Proof. We have

$$\mathbb{E}\left|\frac{S_n}{n} - \mu\right|^2 = \mathrm{var}\left(\frac{S_n}{n}\right) \text{(thanks to } \mathrm{cov}(X_i, X_j) = 0 \text{ , } i \neq j)$$
$$= \frac{1}{n^2}\mathrm{var}S_n = \frac{1}{n^2}\left(\mathrm{var}X_1 + ... + \mathrm{var}X_n\right) \leq \frac{1}{n}C \to 0.$$

□

Observe that in view of Chebishev's inequality (19.3) the convergence in Theorem 19.1 also holds in probability, thus we get another version of WLLN.

□ Exercise 19.1.1. *Let X_n be a sequence of random variables that converge to a random variable X in L^p, i.e. $\mathbb{E}|X_n - X|^p \to 0$ where $p > 0$. Using Chebishev's inequality show that $X_n \xrightarrow{P} X$, i.e. L^p convergence implies convergence in probability.*

□ Exercise 19.1.2. *Let $p > 0$ be fixed. Construct a sequence of random variables X_n such that $X_n \xrightarrow{P} X$ for some random variable X but $\mathbb{E}|X_n - X|^p$ does not converge to 0, i.e. convergence in probability does not imply convergence in L^p.*

For our next example, imagine we are running a service that processes user search queries, similar to Google or other search engines. We receive an enormous volume of queries (jobs), and we have a fixed number of computing resources (workers) available to handle these tasks. Our primary objective as a service provider is to ensure quick response times for incoming queries. This raises the question of how to efficiently distribute these jobs among our workers to prevent any single worker from becoming overwhelmed and to maintain system stability. To achieve this, system architects typically devise a job distribution policy. The component responsible for implementing such policies sits between the jobs and workers and is called a **load balancer** (refer to Figure 17). Its role is to efficiently and evenly distribute incoming job traffic among the available workers.

The scenario described above represents a simplified real-life architecture that includes a load balancer component. Now, you might wonder how probability comes into play. As a model example, we will explore a straightforward policy of randomly selecting a worker and assigning a job to it uniformly. Our goal is to analyze the workload of each worker and assess the degree of balance achieved by this randomized job assignment algorithm. We now start with the mathematical model of the above.

> Assume we have $n \geq 1$ bins (workers) and want to place some n identical balls (jobs) in them. Assume the placement (job to worker assignment) policy is uniform, i.e. we choose a bin uniformly at random and place the ball in it. What is the average number of balls in each bin (average workload of workers)?

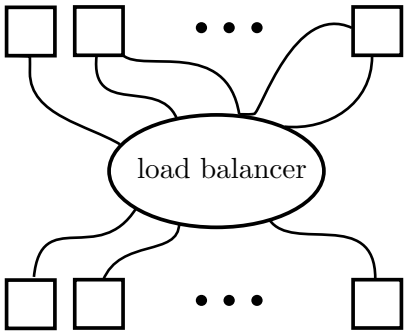

FIGURE 17. A schematic view of a **load balancer**. The bins at the bottom of the image represent tasks (jobs), while the bins at the top represent computing resources (workers). The ellipse-shaped object in between, which is the load balancer, is the module responsible for assigning a job to a worker based on some policy. The lines going in and out of the balancer represent job assignments. The primary task of the load balancer is to efficiently and evenly distribute incoming tasks among the workers, hence its name. In this example, the policy of the balancer is as follows: it selects a worker uniformly at random and assigns the current job to it. The question at hand is to understand how balanced this policy is.

We will first study the average number of balls that fell into bin with number 1. Enumerate the identical balls as $1, 2, ..., n$ and let $\mathbb{I}_i$ be the indicator of the event that the ball number i lands in bin with number 1. Thus

$$\mathbb{I}_i = \begin{cases} 1, & \text{if ball } i \text{ is in bin 1,} \\ 0, & \text{otherwise.} \end{cases}$$

Since the choice of the bins is uniformly at random we get

$$\mathbb{P}(\mathbb{I}_i = 1) = \frac{1}{n}, \quad \mathbb{E}\mathbb{I}_i = \frac{1}{n} \quad \text{and} \quad \text{var}\mathbb{I}_i = \frac{1}{n} - \frac{1}{n^2} \leq \frac{1}{n}. \tag{19.4}$$

The total number of balls in bin 1 becomes

$$X_1 := \sum_{i=1}^{n} \mathbb{I}_i$$

and hence from (19.4) we get

$$\mathbb{E}X_1 = 1 \text{ and } \text{var}X_1 \leq 1.$$

We thus get that the average load, in a sense of mathematical expectation, is 1 (one job per worker on average). The question is then how much the actual load deviates from

the mean or equivalently how strongly is the number of jobs concentrated around their mean. For this we will use Chebishev's inequality (19.3) which implies

$$\mathbb{P}(|X_1 - 1| \geq \sqrt{2n}) \leq \frac{1}{2n},$$

and hence

$$\mathbb{P}(X_1 \geq 1 + \sqrt{2n}) \leq \frac{1}{2n}.$$

Similar to X_1 let X_k be the number of balls that fall into bin number k, with $k = 1, 2, ..., n$. We get

$$\begin{aligned}
&\mathbb{P}\left(\text{there is a bin with load } \geq 1 + \sqrt{2n}\right) \\
&= \mathbb{P}\left(\bigcup_{k=1}^{n} \{X_k \geq 1 + \sqrt{2n}\}\right) \text{(by union bound 3.2)} \\
&\leq \sum_{k=1}^{n} \mathbb{P}\left(X_k \geq 1 + \sqrt{2n}\right) \\
&\leq n\frac{1}{2n} = \frac{1}{2}.
\end{aligned}$$

From the above we conclude that there is at least 50% chance that all bins are well-balanced. For example, take $n = 2000000$ for which $\sqrt{2n} = 2000$. In plain words, this shows that for two million jobs with the random job assignment strategy the probability that there is a worker with over 2000 jobs assigned to it is less than 0.5. In fact more precise bounds on concentration of X_k around its mean can be proved, see Problem 19.2.1.

□ Exercise 19.1.3. *Write a computer program to simulate the random job assignment policy with varying parameters of jobs and workers. How does the bound predicted by Chebishev's inequality compare with the actual load you see in simulations?*

19.2. **Chernoff bound and almost orthogonal vectors in higher-dimensional spaces.** Here we study a stronger form of concentration for a certain family of sums of random variables and discuss some applications of these bounds to higher-dimensional Euclidean spaces.

Theorem 19.2. (`Chernoff bound`) *Let $X_1, ..., X_n$ be a sequence of independent random variables that take value of 0 or 1, i.e. $\mathbb{P}(X_k \in \{0, 1\}) = 1$. Assume $\mathbb{P}(X_k = 1) = p_k$ where $0 \leq p_k \leq 1$ for $k = 1, ..., n$ and set $X = \sum_{k=1}^{n} X_k$. Then for any $\delta > 0$ we have*

$$\mathbb{P}\left(X \geq (1 + \delta)\mu\right) \leq \left[\frac{e^{\delta}}{(1 + \delta)^{1+\delta}}\right]^{\mu},$$

where $\mu := \mathbb{E}X = \sum_{k=1}^{n} p_k$.

Proof. Let $t > 0$ be fixed. Since the random variables $\{X_k\}$ are independent, then so are the random variables $\{e^{tX_k}\}$ due to Lemma 15.1. Hence by Theorem 17.11 we have

$$\begin{aligned}\mathbb{E}e^{tX} &= \mathbb{E}\left(e^{tX_1} \cdot \ldots \cdot e^{tX_n}\right) \\ &= \prod_{i=1}^{n} \mathbb{E}e^{tX_i} = \prod_{i=1}^{n}(1 - p_i + p_i e^t) \\ &= \prod_{i=1}^{n}(1 + p_i(e^t - 1)) \quad (\text{using } 1 + x \leq e^x) \\ &\leq \prod_{i=1}^{n} e^{p_i(e^t-1)} = e^{\mu(e^t-1)}.\end{aligned}$$

From here we get

$$\begin{aligned}\mathbb{P}(X \geq &(1+\delta)\mu) \\ &= \mathbb{P}(e^{tX} \geq e^{t(1+\delta)\mu}) \quad (\text{by Markov (19.1)}) \\ &\leq e^{-t(1+\delta)\mu}\,\mathbb{E}e^{tX} \leq e^{-t(1+\delta)\mu} e^{\mu(e^t-1)} \\ &= e^{\mu(e^t-1-t(1+\delta))}.\end{aligned}$$

Notice that the left-hand side of the last inequality is independent of $t > 0$. Thus we can optimize the right-hand side of the inequality with respect to the parameter $t > 0$ so that it attains its minimum value. Doing this for the term $t \mapsto e^t - 1 - t(1+\delta)$ we get that the optimal value of t satisfies $e^t = 1 + \delta$. Plugging this into the inequality leads to the bound claimed in the theorem. The proof is now complete. □

We next add some remarks to show how Chernoff bounds relate to concentration phenomenon discussed earlier.

- Taking log on the right-hand side (r.h.s.) of the inequality of Theorem 19.2 we get

$$\mu \log \frac{e^\delta}{(1+\delta)^{1+\delta}} = \mu\left[\delta - (1+\delta)\log(1+\delta)\right].$$

  Now for $0 < \delta < 1$ we have

$$\log(1+\delta) \geq \delta - \frac{\delta^2}{2}.$$

  The inequality above can be proved easily by studying the minimum of the difference of the left- and right-hand sides of the inequality as a function of δ. Notice also that the right-hand side of the inequality are the first two terms in Taylor expansion of $\log(1+\delta)$ around 0. We now plug this into the above expression, perform the algebraic simplifications and use the monotonicity of the log to arrive at the following

$$\left[\frac{e^\delta}{(1+\delta)^{1+\delta}}\right]^\mu \leq e^{-\mu\frac{\delta^2}{4}}, \tag{19.5}$$

  when $0 < \delta < 1/2$. It is important to note that the range of δ when this estimate holds does not depend on μ. The simplified exponential bound we obtained in (19.5) will be helpful in applications that we discuss later on.

- For $t > 0$ the probabilities $\mathbb{P}(X \geq t)$ are called the **tail** of a random variable. Chernoff bound provides an exponential decay on the tail of a sum of independent $\{0,1\}$-random variables (Bernoulli variables). Also, to see how the estimate in Theorem 19.2 relates to concentration discussed in this section, we combine the results of Exercise 19.2.2 with the Chernoff bound to estimate $\mathbb{P}(|X - \mathbb{E}X| \geq \delta \mathbb{E}X)$. The simplified bound of (19.5) implies that such deviations of X from $\mathbb{E}X$ have very small probability, or in other words the random variable concentrates strongly around its mean.
- The function $M(t) := \mathbb{E}e^{tX}$ with $t \in \mathbb{R}$, that we used in the proof of Chernoff bound, is called the **moment generating function** of X. The name is motivated by the fact that the formal power series

$$M(t) = \sum_{n=0}^{\infty} \frac{\mathbb{E}X^n}{n!} t^n$$

contain all moments (expectations of the integer powers of X, i.e. $n \mapsto \mathbb{E}X^n$, $n \in \mathbb{N}$) of the random variable X. Thus if the series of M around 0 converges we can recover the $\mathbb{E}X^n$ from the coefficients of the series (see Exercise 19.2.3).

☐ Exercise 19.2.1. *Obtain precise estimates that lead to (19.5).*

☐ Exercise 19.2.2. *Use a similar argument as we had in the proof of Theorem 19.2 to show that*

$$\mathbb{P}\left(X \leq (1-\delta)\mu\right) \leq \left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu}.$$

This estimate coupled with the estimate proved in the theorem shows that the deviation of X from its mean μ is exponentially slow.

☐ Exercise 19.2.3. *A random variable X is distributed as $N(\mu, \sigma^2)$. Compute its moment generating function $M(t) = \mathbb{E}e^{tX}$ and show that*

$$M(t) = e^{t\mu + \frac{1}{2}\sigma^2 t^2}, \quad t \in \mathbb{R}.$$

From this formula find the values of $\mathbb{E}X^n$ of all moments of X using the Taylor expansion of Me^{tX} and comparing the coefficients of t^n.

☐ Exercise 19.2.4. *Let $X_1, ..., X_n, ...$ be a sequence of independent Bernoulli random variables with probability of success p, i.e. $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = 0) = 1 - p$ for any $i \in \mathbb{N}$. Assume $p > 1/2$ and set $S_n := X_1 + ... + X_n$, $n \in \mathbb{N}$. Using Chernoff bound show that*

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} > \frac{1}{2}\right) = 1.$$

Note: The claim above follows directly from SLLN. However, it can be proved with Chernoff also, using Borel-Cantelli lemma on the way.

→ **Problem 19.2.1.** *Prove using Chernoff bounds that in the load balancing problem discussed earlier the load on each worker with randomized job assignment strategy with n jobs and n workers is bounded above by $O\left(\frac{\log n}{\log \log n}\right)$ when n is large with high probability. This bound is substantially better than the one we obtained with Chebishev's inequality.*

Next, we study a peculiar application of Chernoff bound to higher-dimensional Euclidean spaces. It is a standard fact that in $\mathbb{R}^d$ with $d \geq 2$ the maximal number of

pairwise orthogonal vectors is precisely d. Now, assume instead of requiring the vectors being orthogonal, or equivalently their dot product being 0, we allow them to be "almost" orthogonal, in a sense that their dot product is very small while the vectors have length 1. Somewhat unexpectedly, with such relaxation of the orthogonality there can be exponentially many (depending on dimension of the space) vectors that are pairwise almost orthogonal. We now show this using Chernoff bound.

Let $d \geq 2$ be fixed, and consider a sequence of independent and identically distributed (i.i.d.) variables $X_1, ..., X_d, Y_1, ..., Y_d$ where

$$\mathbb{P}\left(X_i = \pm\frac{1}{\sqrt{d}}\right) = \mathbb{P}\left(Y_i = \pm\frac{1}{\sqrt{d}}\right) = \frac{1}{2},\ i = 1, 2, ..., d. \tag{19.6}$$

Now define a random vector $\mathbf{X} := (X_1, ..., X_d) \in \mathbb{R}^d$. Clearly $\mathbf{X}$ is a unit vector with probability 1. Let also $\mathbf{Y} = (Y_1, ..., Y_d)$. For the scalar product of these two vectors we get

$$\langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{k=1}^{d} X_k Y_k =: \sum_{k=1}^{d} Z_k,$$

where in view of independence of X_i and Y_i we have

$$\mathbb{P}\left(Z_i = \pm\frac{1}{d}\right) = \frac{1}{2},\ i = 1, ..., d.$$

Now set

$$W_i := \left(Z_i + \frac{1}{d}\right)\frac{d}{2}.$$

Then clearly W_i is a Bernoulli random variable with

$$W_i \in \{0, 1\} \text{ and } \mathbb{E}W_i = \frac{1}{2}.$$

It now follows from Chernoff's inequality, in particular the simplified bound (19.5) that for $0 < \delta < 1/2$ we get

$$\mathbb{P}\left(\sum_{i=1}^{d} W_i > (1+\delta)\frac{d}{2}\right) \leq e^{-\frac{\delta^2}{2}\frac{d}{2}}.$$

On the other hand we have the following series of equivalences

$$\begin{aligned}
&\sum_{i=1}^{d} W_i > (1+\delta)\frac{d}{2} \\
&\qquad \Leftrightarrow \quad \left(\sum_{i=1}^{d}\left(Z_i + \frac{1}{d}\right)\right)\frac{d}{2} > (1+\delta)\frac{d}{2} \\
&\qquad \Leftrightarrow \quad \sum_{i=1}^{d} Z_i + 1 > 1 + \delta \\
&\qquad \Leftrightarrow \quad \sum_{i=1}^{d} Z_i > \delta.
\end{aligned}$$

It follows that

$$\mathbb{P}\left(\langle \mathbf{X}, \mathbf{Y} \rangle > \delta\right) = \mathbb{P}\left(\sum_{i=1}^{d} Z_i > \delta\right) \leq e^{-\frac{\delta^2}{2}\frac{d}{2}}. \tag{19.7}$$

To pass from a single pair to many pairs we use union-bound argument. As above let $\mathbf{X}_1, ..., \mathbf{X}_n$ be a sequence of random vectors in $\mathbb{R}^d$ where their coordinates are all independent and follow the distribution (19.6). For each pair $i \neq j$ we have the estimate (19.7). Now with the union-bound we get

$$\begin{aligned}&\mathbb{P}\left(\text{there exists a pair } i \neq j \text{ s.t. } \langle \mathbf{X}_i, \mathbf{X}_j \rangle > \delta\right)\\ &\leq \sum_{1 \leq i < j \leq n} \mathbb{P}\left(\langle \mathbf{X}_i, \mathbf{X}_j \rangle > \delta\right)\\ &\leq \frac{n(n-1)}{2} e^{-\frac{\delta^2}{2}\frac{d}{2}}.\end{aligned}$$

If take $n \in \mathbb{N}$ so that

$$\frac{n(n-1)}{2} e^{-\frac{\delta^2}{2}\frac{d}{2}} < 1 \tag{19.8}$$

then the probability that there is a pair of vectors with scalar product larger than δ is smaller than one, hence its complement, when all pairs of vectors have dot products bounded above by δ has positive probability. To achieve (19.8) we can have $n \sim e^{\frac{\delta^2}{4}\frac{d}{2}}$, thus exponentially many with respect to the dimension of the space. Similar bounds hold for estimating $\mathbb{P}\left(\langle \mathbf{X}, \mathbf{Y} \rangle < -\delta\right)$, hence we conclude that there are exponentially many unit vectors that are almost orthogonal in a sense that their dot product is very small.

□ Exercise 19.2.5. *Complete the details on estimating the probabilities of the form* $\mathbb{P}\left(\langle \boldsymbol{X}, \boldsymbol{Y} \rangle < -\delta\right)$ *in the above example.*

□ Exercise 19.2.6. *Show that similar to the almost orthogonality of vectors discussed above, given the dimension* n *of the Euclidean space* $\mathbb{R}^n$ *is large one can find exponentially many, with respect to the dimension, unit vectors that are on almost equal distance from each other.*

References

[1] N. Alon, J. Spencer, The probabilistic method, 2016

[2] **R. Durrett**, Probability theory and examples. 4th edition, Cambridge University Press, 2010

[3] M. H. DeGroot, S. J. Mark, Probability and statistics. 4th edition, Pearson Education, 2014

[4] L. C. Evans, An introduction to stochastic differential equations, (vii) 1 - 149, 2013

[5] **G. Grimmett and D. Welsh**, Probability an Introduction (2nd edition); Oxford University Press, 2014, pp. 1-281; http://www.math.nagoya-u.ac.jp/~richard/teaching/f2017/GW_2014.pdf

[6] D. MacKay, Information theory, inference and learning algorithms. 4th edition, Cambridge University Press, 2014.

[7] S. Shreve, Stochastic Calculus for Finance II. Continuous-Time Models. Springer Finance Textbook, 2004.

[8] **R. Weber**, Lecture notes, 2014, http://www.statslab.cam.ac.uk/~rrw1/prob/prob-weber.pdf

[9] **D. Williams**, Probability with martingales, Cambridge Mathematical Textbooks, (xiv)1-251 (2008)

[10] G. Shafer, V. Vovk, The origins and legacy of Kolmogorov's Grundbegriffe, https://arxiv.org/pdf/1802.06071.pdf, 2018

[11] T. Tao, Poincaré's Legacies, Part I (pages from year two of a mathematical blog), American Mathematical Society, 2009

Email address: hayk.aleksanyan@gmail.com

Web: https://www.haykaleksanyan.com