1.2 Basic Concepts in Probability Theory

In this section we will introduce some basic concepts in probability theory.

Let us consider a random experiment of which all possible results are included in a non-empty set Ω, usually called the *sample space*. An element $\omega \in \Omega$ is called a *sample point* or *outcome* of the experiment. An *event* of a random experiment is specified as a subset of Ω. Event A is called *true* if an outcome $\omega \in \Omega$ has been chosen with $\omega \in A$. Otherwise A is called *false*. A system $\mathcal{A}$ of events (or, in general, subsets of Ω) is called an *algebra* if

- $\Omega \in \mathcal{A}$,
- if $A, B \in \mathcal{A}$, then $A \cup B \in \mathcal{A}$ and $A \cap B \in \mathcal{A}$, and
- if $A \in \mathcal{A}$, then $\bar{A} \in \mathcal{A}$.

Given an algebra $\mathcal{A}$, a function $\mu : \mathcal{A} \to \mathrm{IR}_+$ is called a *measure* on $\mathcal{A}$ if for every pair of disjoint sets $A, B \in \mathcal{A}$ we have

$$\mu(A \cup B) = \mu(A) + \mu(B) \ .$$

This definition clearly implies that $\mu(\emptyset) = 0$ and that for any set of pairwise disjoint events $A_1, \ldots, A_k \in \mathcal{A}$ we have

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

Furthermore, it implies that for any pair of sets $A, B \in \mathcal{A}$ we have

$$\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B) \ .$$

We say that a function $p : \mathcal{A} \to [0, 1]$ is a *probability measure* if

- p is a measure on $\mathcal{A}$ and
- $p(\Omega) = 1$.

Given a probability measure p, the *probability* of an event A to be true is defined as

$$\Pr[A] = p(A) \ .$$

We say that a triple $(\Omega, \mathcal{A}, p)$ is a *probability space* if $\mathcal{A}$ is an algebra over Ω and p is a probability measure on $\mathcal{A}$.

1.2.1 Events

Starting from a given collection of sets that represent events, we can form new events by means of statements containing the logical connectives "or," "and," and "not," which correspond in the language of set theory to the operations "union," "intersection," and "complement."

If A and B are events, their *union*, denoted by $A \cup B$, is the event consisting of all outcomes realizing either A or B. The *intersection* of A and B, denoted by $A \cap B$, consists of all outcomes realizing both A and B. The *difference* of B and A, denoted by $B \setminus A$, consists of all outcomes that belong to B but not to A. If A is a subset of Ω, its *complement*, denoted by $\bar{A}$, is the set of outcomes in Ω that do not belong to A. That is, $\bar{A} = \Omega \setminus A$.

Two events A and B are called *disjoint* if $A \cap B$ is empty. In probability theory, $\emptyset$ is called the *impossible* event. The set Ω is naturally called the *certain* event.

If Ω is a countable sample space (i.e., its elements can be arranged in a sequence so that the rth element is identifiable for any $r \in \mathbb{N}$), we define the *size* of an event A, denoted by $|A|$, to be the number of outcomes it contains.

1.2.2 The Inclusion-Exclusion Principle

Let $A_1, \ldots, A_n$ be any collection of events. The *inclusion-exclusion principle* stated in Section 1.1.1 implies that

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] = \sum_{k=1}^{n} (-1)^{k+1} \sum_{i_1<i_2<\ldots<i_k} \Pr\left[\bigcap_{j=1}^{k} A_{i_j}\right] . \tag{1.1}$$

For the special case of $n = 2$ we obtain

$$\Pr[A_1 \cup A_2] = \Pr[A_1] + \Pr[A_2] - \Pr[A_1 \cap A_2] .$$

In cases where it is too difficult to evaluate (1.1) exactly, *Bonferroni's inequalities* may be used to find suitable approximations:

- For every odd m,

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] \leq \sum_{k=1}^{m} (-1)^{k+1} \sum_{i_1<i_2<\ldots<i_k} \Pr\left[\bigcap_{j=1}^{k} A_{i_j}\right] .$$

- For every even m,

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] \geq \sum_{k=1}^{m} (-1)^{k+1} \sum_{i_1<i_2<\ldots<i_k} \Pr\left[\bigcap_{j=1}^{k} A_{i_j}\right] .$$

Special cases of these inequalities are *Boole's inequalities*:

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] \leq \sum_{i=1}^{n} \Pr[A_i]$$

and

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] \geq \sum_{i=1}^{n} \Pr[A_i] - \sum_{1\leq i<j\leq n} \Pr[A_i \cap A_j] .$$

We demonstrate the usefulness of the inclusion-exclusion principle by some simple examples.

Example 1: Consider the problem of coloring the nodes of a hypergraph G with the help of 2 colors such that no hyperedge is monochromatic (i.e., no hyperedge only contains nodes of a single color). We would like to identify a class of hypergraphs for which this is always possible. Such a class is given in the following claim.

Claim 1.5 *For every hypergraph G with n hyperedges in which every hyperedge is of size at least $\log n + 2$, there is a 2-coloring of the nodes such that no hyperedge is monochromatic.*

Proof. Consider the random experiment of choosing for each node independently and uniformly at random one of the two possible colors. In this case, the probability that a hyperedge of size k is monochromatic is equal to $2 \cdot 2^{-k} = 2^{-k+1}$. Assume the hyperedges to be numbered from 1 to n and let A_i be the event that hyperedge i is monochromatic. Then, by the inclusion-exclusion principle,

$$\Pr[A_1 \cup \ldots \cup A_n] \leq \sum_{i=1}^{n} \Pr[A_i] \leq \sum_{i=1}^{n} 2^{-(\log n + 1)} = \frac{1}{2} .$$

Hence,

$$\Pr[\bar{A}_1 \cap \ldots \cap \bar{A}_n] = 1 - \Pr[A_1 \cup \ldots \cup A_n] \geq \frac{1}{2} ,$$

and therefore there must exist a 2-coloring such that no edge is monochromatic. □

Example 2: Assume that we have n balls and n bins. Each ball is placed in a bin that is chosen independently and uniformly at random. The goal is to provide an upper bound for the maximum number of balls in a bin. This will be given in the following claim.

Claim 1.6 *With a probability of at least $1/2$, the maximum number of balls in a bin is at most $(1 + o(1)) \log n / \log \log n$.*

Proof. Let $k = \alpha \log n / \log \log n$ for some non-negative α that will be specified later. For every $i \in [n]$, let the event A_i be true if and only if bin i has at least k balls. Furthermore, let E_1 denote the event that there exists a bin with at least k balls. Then, by Boole's inequality,

$$\Pr[E_1] = \Pr[A_1 \cup \ldots \cup A_n] \leq \sum_{i=1}^{n} \Pr[A_i] .$$

For any subset $S \subseteq [n]$, let the event $A_{i,S}$ be true if and only if all balls in S are placed in bin i. Then, again by Boole's inequality,

$$\Pr[A_i] = \Pr\left[\bigcup_{S \subseteq [n],\ |S|=k} A_{i,S}\right] \leq \sum_{S \subseteq [n],\ |S|=k} \Pr[A_{i,S}]$$

for all i. Since $|\{S \subseteq [n],\ |S| = k\}| = \binom{n}{k}$ and $\Pr[A_{i,S}] = 1/n^{|S|}$ for all $S \subseteq [n]$, we obtain

$$\begin{aligned}
\Pr[E_1] &\leq n \cdot \binom{n}{k} \left(\frac{1}{n}\right)^k \\
&\leq n \cdot \left(\frac{\mathrm{e} \cdot n}{k}\right)^k \left(\frac{1}{n}\right)^k = n \cdot \left(\frac{\mathrm{e}}{k}\right)^k \\
&\leq n \cdot 2^{-\alpha \frac{\log n}{\log \log n} \cdot \log\left(\frac{\alpha \log n}{\mathrm{e} \log \log n}\right)} \\
&\leq n \cdot 2^{-\log n - 1} \leq \frac{1}{2}
\end{aligned}$$

if $\alpha = 1 + o(1)$ and n are sufficiently large. □

We note that the bound for the maximum number of balls in Claim 1.6 is essentially best possible, since the probability that the maximum number of balls is at least $(1 - o(1)) \log n / \log\log n$ can be shown to be also at least $1/2$.

Example 3: Again, consider the situation that we have n balls and n bins, and each ball is placed in a bin chosen independently and uniformly at random.

Claim 1.7 *The probability that bin 1 has at least one ball is at least* $1/2$.

Proof. For any $i \in [n]$, let A_i be the event that ball i is placed in bin 1. Then, by Boole's inequality,

$$\begin{aligned}
\Pr[\text{bin 1 has at least one ball}] &= \Pr\left[\bigcup_{i \in [n]} A_i\right] \\
&\geq \sum_{1 \leq i \leq n} \Pr[A_i] - \sum_{1 \leq i < j \leq n} \Pr[A_i \cap A_j] \\
&= \sum_{1 \leq i \leq n} \frac{1}{n} - \sum_{1 \leq i < j \leq n} \frac{1}{n^2} \\
&= 1 - \binom{n}{2} \frac{1}{n^2} \geq 1 - \frac{1}{2} = \frac{1}{2} \, .
\end{aligned}$$

□

Observe that the probability bound in Claim 1.7 is not far away from the exact bound:

$$\Pr[\text{bin 1 has at least one ball}] = 1 - \left(1 - \frac{1}{n}\right)^n \stackrel{n \to \infty}{=} 1 - \frac{1}{\mathrm{e}} \, .$$

1.2.3 Conditional Probability

The *conditional probability* of event B assuming an event A with $\Pr[A] > 0$ is denoted by $\Pr[B \mid A]$ and defined as

$$\Pr[B \mid A] = \frac{\Pr[A \cap B]}{\Pr[A]} \, .$$

From this definition it follows that for all events A and B with $\Pr[A] > 0$,

$$\Pr[A \cap B] = \Pr[A] \cdot \Pr[B \mid A] \, . \tag{1.2}$$

This can be generalized as follows: If $A_1, \ldots, A_n$ are events with $\Pr[A_1 \cap \ldots \cap A_{n-1}] > 0$, then

$$\Pr[A_1 \cap \ldots \cap A_n] = \prod_{i=1}^{n} \Pr[A_i \mid A_1 \cap \ldots \cap A_{i-1}] \, .$$

Suppose that A and B are events with $\Pr[A] > 0$ and $\Pr[B] > 0$. Then, in addition to the equality (1.2), we have

$$\Pr[A \cap B] = \Pr[B] \cdot \Pr[A \mid B] \, . \tag{1.3}$$

From (1.2) and (1.3) we obtain *Bayes's formula*

$$\Pr[A \mid B] = \frac{\Pr[A] \cdot \Pr[B \mid A]}{\Pr[B]} .$$

Two events A and B are called *independent* if and only if

$$\Pr[B \mid A] = \Pr[B] .$$

Note that, due to Bayes's formula, in this case also $\Pr[A \mid B] = \Pr[A]$, that is, the independence property is *symmetric*.

If $\Pr[B \mid A] \neq \Pr[B]$, then A and B are said to be *correlated*. A and B are called

- *negatively correlated* if $\Pr[B \mid A] < \Pr[B]$ and
- *positively correlated* if $\Pr[B \mid A] > \Pr[B]$.

By Bayes's formula, all of these correlation properties are also symmetric.

As an example, any two disjoint events A and B with positive probabilities cannot be independent, since $\Pr[B \mid A] = 0$. However, they are always negatively correlated. Furthermore, they have the property that

$$\Pr[A \cup B] = \Pr[A] + \Pr[B] .$$

To illustrate negative and positive correlation, let us give a very simple example.

Example: Consider the problem of throwing n balls into n bins, where each ball is placed in a bin chosen independently and uniformly at random. Let A be the event that bin 1 has no ball, and let B be the event that some fixed bin $i > 1$ has no ball. For $n = 2$ it holds

$$\Pr[B \mid A] = 0 \leq \frac{1}{4} = \Pr[B] ,$$

and for $n = 3$ it holds

$$\Pr[B \mid A] = \frac{1}{8} \leq \frac{8}{27} = \Pr[B] .$$

Thus, A and B are negatively correlated for these n. (One can observe that this also holds for higher n.)

Now, let A be the event that bin 1 has exactly one ball, and let B be the event that some fixed bin $i > 1$ has exactly one ball. Since the expected number of balls in a bin is 1, one may assume that A does not influence B. However, for $n = 2$ it holds

$$\Pr[B \mid A] = 1 \geq \frac{1}{2} = \Pr[B] ,$$

and for $n = 3$ it holds

$$\Pr[B \mid A] = \frac{1}{2} \geq \frac{4}{9} = \Pr[B] .$$

Thus, A and B are positively correlated for these n. However, one can observe that $\Pr[B \mid A]$ tends to $\Pr[B]$ for $n \to \infty$.

If we define A and B to be events for which a bin has exactly $k > 1$ balls, one can show that A and B are, as in the first case, negatively correlated.

1.2.4 Random Variables

Any numerical function $X = X(\omega)$ defined on a sample space Ω may be called a *random variable*. In this thesis we will only consider real-valued random variables, i.e., functions of the form $X : \Omega \to \mathbb{R}$. It has been found convenient to separate the random variables into three categories: discrete, continuous, and mixed.

A *discrete* random variable is supposed to take only isolated values with nonzero probabilities. The number of values it is allowed to take may be infinite, but it is essential that they are *countable*. That is, it must be possible to arrange its values in a sequence so that the rth number is identifiable for any $r \in \mathbb{N}$. An *(absolutely) continuous* random variable X is one for which $\Pr[X \leq x]$ can be expressed as the integral

$$F_X(x) = \int_{-\infty}^{x} p_X(x)dx$$

of a function $p_X(x)$ commonly called the *probability density* of X. F_X is called the *distribution function* of X. A random variable is called *mixed* if it has both discrete and continuous parts.

In the following, we will only consider discrete random variables. A random variable X is called *non-negative* if $X(\omega) \geq 0$ for all $\omega \in \Omega$. For the special case that X maps elements in Ω to $\{0, 1\}$, X is called a *binary* or *Bernoulli* random variable. A binary random variable X is called an *indicator* of event A (denoted by I_A) if $X(\omega) = 1$ if and only if $\omega \in A$ for all $\omega \in \Omega$.

For any random variable X and any number $x \in \mathbb{R}$, we define $[X = x] = \{\omega \in \Omega : \ X(\omega) = x\}$. Instead of using set operations to express combinations of events associated with random variables, we will use logical expressions in the following, that is,

- instead of $\Pr[[X = x] \cap [Y = y]]$ we write $\Pr[X = x \wedge Y = y]$, and
- instead of $\Pr[[X = x] \cup [Y = y]]$ we write $\Pr[X = x \vee Y = y]$.

For any discrete random variable X and any $k \in \mathbb{R}$, we define

$$\Pr[X \leq k] = \sum_{\ell \leq k} \Pr[X = \ell] \quad \text{and} \quad \Pr[X \geq k] = \sum_{\ell \geq k} \Pr[X = \ell]\ .$$

The function $p_X(k) = \Pr[X = k]$ is called the *probability distribution* of X, and the function $F_X(k) = \Pr[X \leq k]$ is called the *(cumulative) distribution function* of X. Furthermore, the function $G_X(k) = \Pr[X \geq k]$ is called the *survival distribution function* of X. For any two random variables X and Y, X is said to *(stochastically) dominate* Y (denoted by $X \succ Y$) if $G_X(k) \geq G_Y(k)$ for all $k \in \mathbb{R}$.

1.2.5 Expectation

The most important measure used in combination with random variables is the expectation.

Definition 1.8 *Let $(\Omega, \mathcal{A}, p)$ denote an arbitrary probability space and $X : \Omega \to \mathbb{R}$ be an arbitrary discrete function. Then the* expectation *of X is defined as*

$$\mathrm{E}[X] = \sum_{x \in \mathbb{R}} x \cdot \Pr[X = x]\ . \tag{1.4}$$

Assume that the set $S = \{x_1, x_2, \ldots\}$ contains all values X can take, where $x_1 < x_2 < \ldots$. Let the function $\Delta_X : \mathbb{R} \to \mathbb{R}_+$ be defined as

$$\Delta_X(x) = \begin{cases} 0 & : \quad x \notin S \\ x & : \quad x = x_1 \\ x - x_{i-1} & : \quad x = x_i \text{ for some } i \geq 2 \end{cases}$$

That is, $\Delta_X(x)$ provides the distance to the predecessor of x in S. If X is non-negative, it follows from equation (1.4) and the definition of Δ_X that

$$\mathrm{E}[X] = \sum_{x \geq 0} \Delta_X(x) \cdot \Pr[X \geq x] = \sum_{x \geq 0} \Delta_X(x) \cdot G_X(x) \ . \tag{1.5}$$

In the case that X is integer-valued, we can simplify this expression to

$$\mathrm{E}[X] = \sum_{x \in \mathbb{N}} \Pr[X \geq x] = \sum_{x \in \mathbb{N}} G_X(x) \ . \tag{1.6}$$

For binary random variables X it holds that

$$\mathrm{E}[X] = \Pr[X = 1] \ .$$

Basic properties

The following fact lists some basic properties of the expectation.

Fact 1.9 *For arbitrary random variables X and Y and any $c \in \mathbb{R}$ it holds:*

- *If X is non-negative, then* $\mathrm{E}[X] \geq 0$.
- $|\mathrm{E}[X]| \leq \mathrm{E}[|X|]$.
- $\mathrm{E}[c \cdot X] = c \cdot \mathrm{E}[X]$.
- $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$.

Two random variables X and Y are called *independent* if, for all $x, y \in \mathbb{R}$,

$$\Pr[X = x \mid Y = y] = \Pr[X = x] \ .$$

Independent random variables have the following important property.

Claim 1.10 *If X and Y are independent random variables, then* $\mathrm{E}[X \cdot Y] = \mathrm{E}[X] \cdot \mathrm{E}[Y]$.

Proof. For all $x, y \in \mathbb{R}$, let A_x and B_y be the events with

$$A_x = \{\omega \in \Omega : \ X(\omega) = x\} \quad \text{and} \quad B_y = \{\omega \in \Omega : \ Y(\omega) = y\} \ .$$

Since X and Y are independent, $\Pr[A_x \cap B_y] = \Pr[A_x] \cdot \Pr[B_y]$ for all $x, y \in \mathbb{R}$. Recall that, for any event E, the indicator variable I_E is equal to 1 if and only if $\omega \in E$. As $X = \sum_x x \cdot I_{A_x}$ and $Y = \sum_y y \cdot I_{B_y}$, we obtain

$$\begin{aligned}
\mathrm{E}[X \cdot Y] &= \mathrm{E}\left[\left(\sum_{x\in\mathbb{R}} x \cdot I_{A_x}\right) \cdot \left(\sum_{y\in\mathbb{R}} y \cdot I_{B_y}\right)\right] \\
&= \mathrm{E}\left[\sum_{x,y\in\mathbb{R}} x \cdot y \cdot I_{A_x \cap B_y}\right] = \sum_{x,y\in\mathbb{R}} x \cdot y \cdot \Pr[A_x \cap B_y] \\
&= \sum_{x,y\in\mathbb{R}} x \cdot y \cdot \Pr[A_x] \cdot \Pr[B_y] \\
&= \left(\sum_{x\in\mathbb{R}} x \cdot \Pr[A_x]\right) \cdot \left(\sum_{y\in\mathbb{R}} y \cdot \Pr[B_y]\right) = \mathrm{E}[X] \cdot \mathrm{E}[Y] \; .
\end{aligned}$$

□

A set $X_1, \ldots, X_n$ of random variables is called *(mutually) independent* if, for all $i \in [n]$ and $S \subseteq [n] \setminus \{i\}$,

$$\Pr\left[X_i = x_i \mid \bigwedge_{j\in S} X_j = x_j\right] = \Pr[X_i = x_i] \tag{1.7}$$

for all $x_i \in \mathbb{R}$ and $x_j \in \mathbb{R}$ with $j \in S$. $X_1, \ldots, X_n$ are called *k-wise independent* if (1.7) holds for all $S \subseteq [n] \setminus \{i\}$ with $|S| \leq k$.

Modeling Dependence

The *conditional expectation* of a random variable Y with respect to an event A is defined by

$$\mathrm{E}[Y \mid A] = \sum_{y\in\mathbb{R}} y \cdot \Pr[Y = y \mid A] \; . \tag{1.8}$$

If the event A is $X = a$ for some random variable X, this equation defines a function f with

$$f(a) = \mathrm{E}[Y \mid X = a] \; .$$

Thus, $\mathrm{E}[Y \mid X]$ is a random variable, namely the random variable $f(X)$. (Observe that it does *not* hold in general that $\mathrm{E}[X \cdot Y] = \mathrm{E}[X] \cdot \mathrm{E}[Y \mid X]$. A way to study the relationship between $\mathrm{E}[X \cdot Y]$ and $\mathrm{E}[X]$ and $\mathrm{E}[Y]$ will be given in Section 1.2.6.)

A fundamental property of the conditional expectation is that

$$\mathrm{E}[Y] = \mathrm{E}[\mathrm{E}[Y \mid X]]$$

for any two random variables X and Y (which is easy to check). If Y is independent of X, then (1.8) implies that $\mathrm{E}[Y \mid X] = \mathrm{E}[Y]$. An important case for which $\mathrm{E}[Y]$ depends on X is given in the following definition.

Definition 1.11 *A sequence of random variables $X_0, X_1, \ldots$ is called a* martingale *if for all $i \geq 1$,*

$$\mathrm{E}[X_i \mid X_0, X_1, \ldots, X_{i-1}] = X_{i-1} \; .$$

Martingales will be considered in a later section.

Tail estimates

We start with a fact that contains some straightforward probability bounds.

Fact 1.12 *Let X be an arbitrary random variable. Then*

$$\Pr[X < \mathrm{E}[X]] < 1 \quad \textit{and} \quad \Pr[X > \mathrm{E}[X]] < 1 \ .$$

The next result provides a first, simple probability bound that depends on the deviation from the expected value. It has apparently first been used by Chebychev (which is why some authors call it Chebychev inequality [Shi96]), but it is commonly called *Markov inequality*.

Theorem 1.13 (Markov Inequality) *Let X be an arbitrary non-negative random variable. Then, for any $k > 0$,*

$$\Pr[X \geq k] \leq \frac{\mathrm{E}[X]}{k} \ .$$

Proof. Obviously,

$$\mathrm{E}[X] = \sum_{x \geq 0} x \cdot \Pr[X = x] \geq k \cdot \Pr[X \geq k] \ .$$

□

As we will show later, this inequality has a tremendous number of consequences. We will start with some simple examples.

Example 1: A directed graph G is called a *tournament* if for every pair of nodes u, v with $u \neq v$, either (u, v) or (v, u) is an edge in G. A *Hamiltonian path* in a directed graph G is a directed path that visits every node of G exactly once. We will show the following claim.

Claim 1.14 *There exists a tournament of size n with at least $n!/2^{n-1}$ Hamiltonian paths.*

Proof. Let K_n be the complete undirected graph of size n. We generate a random tournament T out of K_n by choosing a direction for each edge independently and uniformly at random. Let π be a permutation of the nodes in T. We define the binary random variable X_π to be 1 if and only if π defines a directed path in T. Clearly, for all $\pi \in S_n$, $\Pr[X_\pi = 1] = 2^{-(n-1)}$. Let X be the random variable counting the number of Hamiltonian paths in T, that is,

$$X = \sum_{\pi \in S_n} X_\pi \ .$$

Since $|S_n| = n!$, we obtain

$$\mathrm{E}[X] = \sum_{\pi \in S_n} \Pr[X_\pi = 1] = n!/2^{n-1} \ .$$

According to Fact 1.12, $\Pr[X < \mathrm{E}[X]] < 1$. Thus, there must exist a tournament with at least $n!/2^{n-1}$ Hamiltonian paths. □