Chapter 1

Random Variables

1.1. Elementary Examples

We will start with some elementary examples of probability. The most well-known example is that of a fair coin: if flipped, the probability of getting a head or tail both equal to 1/2. If we perform n independent tosses, then the probability of obtaining n heads is equal to $1/2^n$: among the 2^n equally possible outcomes only one gives the result that we look for. More generally, let $S_n = X_1 + X_2 + \cdots + X_n$, where

$$X_j = \begin{cases} 1, & \text{if the result of the } n\text{th trial is a head,} \\ 0, & \text{if the result of the } n\text{th trial is a tail.} \end{cases}$$

Then the probability that we get k heads out of n tosses is equal to

$$\text{Prob}(S_n = k) = \frac{1}{2^n} \binom{n}{k}.$$

Applying Stirling's formula

$$n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n, \quad n \to \infty,$$

we can calculate, for example, the asymptotic probability of obtaining heads exactly half of the time:

$$\text{Prob}(S_{2n} = n) = \frac{1}{2^{2n}} \binom{2n}{n} = \frac{1}{2^{2n}} \frac{(2n)!}{(n!)^2} \sim \frac{1}{\sqrt{\pi n}} \to 0,$$

as $n \to \infty$.

On the other hand, since we have a fair coin, we do expect to obtain heads roughly half of the time; i.e.,

$$\frac{S_{2n}}{2n} \approx \frac{1}{2},$$

for large n. Such a statement is indeed true and is embodied in the law of large numbers that we will discuss in the next chapter. For the moment let us simply observe that while the probability that S_{2n} equals n goes to zero as $n \to \infty$, the probability that S_{2n} is close to n goes to 1 as $n \to \infty$. More precisely, for any $\epsilon > 0$,

$$\text{Prob}\left(\left|\frac{S_{2n}}{2n} - \frac{1}{2}\right| > \epsilon\right) \to 0,$$

as $n \to \infty$. This can be seen as follows. Noting that the distribution $\text{Prob}\{S_{2n} = k\}$ is unimodal and symmetric around the state $k = n$, we have

$$\begin{aligned}
\text{Prob}\left(\left|\frac{S_{2n}}{2n} - \frac{1}{2}\right| > \epsilon\right) &\le 2 \cdot \frac{1}{2^{2n}} \sum_{k > n+2n\epsilon} \frac{(2n)!}{k!(2n-k)!} \\
&\le 2(n - 2n\epsilon) \cdot \frac{1}{2^{2n}} \frac{(2n)!}{\lceil n + 2n\epsilon \rceil! \lfloor n - 2n\epsilon \rfloor!} \\
&\sim \frac{2\sqrt{1-2\epsilon}}{\sqrt{\pi(1+2\epsilon)}} \cdot \frac{\sqrt{n}}{(1-2\epsilon)^{n(1-2\epsilon)}(1+2\epsilon)^{n(1+2\epsilon)}} \to 0
\end{aligned}$$

for sufficiently small ϵ and $n \gg 1$, where $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ are the ceil and floor functions, respectively, defined by $\lceil x \rceil = m+1$ and $\lfloor x \rfloor = m$ if $x \in [m, m+1)$ for $m \in \mathbb{Z}$. This is the *weak law of large numbers* for this particular example.

In the example of a fair coin, the number of outcomes in an experiment is finite. In contrast, the second class of examples involves a continuous set of possible outcomes. Consider the orientation of a unit vector $\boldsymbol{\tau}$. Denote by $\mathbb{S}^2$ the unit sphere in $\mathbb{R}^3$. Define $\rho(\boldsymbol{n})$, $\boldsymbol{n} \in \mathbb{S}^2$, as the orientation distribution density; i.e., for $A \subset \mathbb{S}^2$,

$$\text{Prob}(\boldsymbol{\tau} \in A) = \int_A \rho(\boldsymbol{n}) dS,$$

where dS is the surface area element on $\mathbb{S}^2$. If $\boldsymbol{\tau}$ does not have a preferred orientation, i.e., it has equal probability of pointing at any direction, then

$$\rho(\boldsymbol{n}) = \frac{1}{4\pi}.$$

In this case, we say that $\boldsymbol{\tau}$ is isotropic. On the other hand, if $\boldsymbol{\tau}$ does have a preferred orientation, say $\boldsymbol{n}_0$, then we expect $\rho(\boldsymbol{n})$ to be peaked at $\boldsymbol{n}_0$.

1.2. Probability Space

It is useful to put these intuitive notions of probability on a firm mathematical basis, as was done by Kolmogorov. For this purpose, we need the notion of *probability space*, often written as a triplet $(\Omega, \mathcal{F}, \mathbb{P})$, defined as follows.

Definition 1.1 (Sample space)**.** The sample space Ω is the set of all possible outcomes. Each element $\omega \in \Omega$ is called a sample point.

Definition 1.2 (σ-algebra)**.** A σ-algebra (or σ-field) $\mathcal{F}$ is a collection of subsets of Ω that satisfies the following conditions:

(i) $\Omega \in \mathcal{F}$;

(ii) if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$, where $A^c = \Omega \backslash A$ is the complement of A in Ω;

(iii) if $A_1, A_2, \ldots \in \mathcal{F}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

Each set A in $\mathcal{F}$ is called an event. Let $\mathcal{B}$ be a collection of subsets of Ω. We denote by $\sigma(\mathcal{B})$ the smallest σ-algebra generated by the sets in $\mathcal{B}$, i.e., the smallest σ-algebra that contains $\mathcal{B}$. The pair $(\Omega, \mathcal{F})$ with the above properties is called a *measurable space*.

Definition 1.3 (Probability measure)**.** The probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a set function defined on $\mathcal{F}$ which satisfies

(a) $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$;

(b) if $A_1, A_2, \ldots \in \mathcal{F}$ are pairwise disjoint, i.e., $A_i \cap A_j = \emptyset$ if $i \neq j$, then

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n). \tag{1.1}$$

(1.1) is called countable additivity or σ-additivity.

Example 1.4 (Fair coin)**.** The probability space for the outcome of one trial can be defined as follows. The sample space $\Omega = \{H, T\}$ where H and T represent head and tail, respectively. The σ-algebra

$$\mathcal{F} = \text{ all subsets of } \Omega = \{\emptyset, \{H\}, \{T\}, \Omega\}$$

and

$$\mathbb{P}(\emptyset) = 0, \qquad \mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = \frac{1}{2}, \qquad \mathbb{P}(\Omega) = 1.$$

For n independent tosses, we can take $\Omega = \{H, T\}^n$, $\mathcal{F} =$ all subsets of Ω, and

$$\mathbb{P}(A) = \frac{1}{2^n}|A|,$$

where $|A|$ is the cardinality of the set A.

Example 1.5 (Uniform orientation distribution on $\mathbb{S}^2$)**.** In this case, the sample space $\Omega = \mathbb{S}^2$. Let $\mathcal{B}$ be the set of all open sets of $\mathbb{S}^2$, defined as the intersection of any open set $B \subset \mathbb{R}^3$ and $\mathbb{S}^2$. Then we can take the σ-algebra to be $\mathcal{F} = \sigma(\mathcal{B})$ and

$$\mathbb{P}(U) = \frac{\text{surface area}(U)}{4\pi} \quad \text{for any } U \in \mathcal{F}.$$

Within this framework, the standard rules of set theory are used to answer probability questions. For instance, if both $A, B \in \mathcal{F}$, the probalility that both A and B occurs is given by $\mathbb{P}(A \cap B)$, the probability that either A or B occurs is given by $\mathbb{P}(A \cup B)$, the probability that A but not B occurs is given by $\mathbb{P}(A \backslash B)$, etc. It is easy to check that

$$A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B). \tag{1.2}$$

This is because $B = A \cup (B \backslash A)$; therefore, $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \backslash A) \geq \mathbb{P}(A)$. We also have

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B) \tag{1.3}$$

since $A = A \cup (B \cap A^c)$, and $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c) \leq \mathbb{P}(A) + \mathbb{P}(B)$. This last inequality is known as *Boole's inequality*.

1.3. Conditional Probability

Let $A, B \in \mathcal{F}$ and assume that $\mathbb{P}(B) \neq 0$. Then the *conditional probability* of A given B is defined as

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{1.4}$$

This is the proportion of events that both A and B occur given that B occurs. For instance, the probability to obtain two tails in two tosses of a fair coin is $1/4$, but the conditional probability to obtain two tails is $1/2$ given that the first toss is a tail, and it is zero given that the first toss is a head.

Since $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$ by definition, we also have

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

and so on. It is straightforward to obtain

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\mathbb{P}(B)} \tag{1.5}$$

from the definition of conditional probability. This is called *Bayes's rule*.

Proposition 1.6 (Bayes's theorem)**.** *If $A_1, A_2, \ldots$ are disjoint sets such that $\bigcup_{j=1}^{\infty} A_j = \Omega$, then we have*

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(A_j)\mathbb{P}(B|A_j)}{\sum_{n=1}^{\infty}\mathbb{P}(A_n)\mathbb{P}(B|A_n)} \quad \text{for any } j \in \mathbb{N}. \tag{1.6}$$

This is useful in Bayesian statistics where A_j corresponds to the hypothesis and $\mathbb{P}(A_j)$ is the prior probability of the hypothesis A_j. The conditional probability $\mathbb{P}(A_j|B)$ is the posterior probability of A_j given that the event B occurs.

1.4. Discrete Distributions

If the elements in Ω are finite or enumerable, say, $\Omega = \{\omega_1, \omega_2, \ldots\}$, we have a situation of discrete probability space and discrete distribution. In this case, let $X(\omega_j) = x_j$ and

$$p_j = \mathbb{P}(X = x_j), \qquad j = 0, 1, \ldots.$$

Of course, we have to have

$$0 \le p_j \le 1, \qquad \sum_j p_j = 1.$$

Given a function f of X, its *expectation* is given by

$$\mathbb{E}f(X) = \sum_j f(x_j)p_j \tag{1.7}$$

if the sum is well-defined. In particular, the pth moment of the distribution is defined as

$$m_p = \sum_j x_j^p p_j.$$

When $p = 1$, it is called the *mean* of the random variable and is also denoted by $\mathrm{mean}(X)$. Another important quantity is its *variance*, defined as

$$\mathrm{Var}(X) = m_2 - m_1^2 = \sum_j (x_j - m_1)^2 p_j. \tag{1.8}$$

Example 1.7 (Bernoulli distribution)**.** The Bernoulli distribution has the form

$$\mathbb{P}(X = j) = \begin{cases} p, & j = 1, \\ q, & j = 0, \end{cases}$$

$p + q = 1$ and $p, q \ge 0$. When $p = q = 1/2$, it corresponds to the toss of a fair coin. The mean and variance can be calculated directly:

$$\mathbb{E}X = p, \quad \mathrm{Var}(X) = pq.$$

Example 1.8 (Binomial distribution $B(n,p)$)**.** The binomial distribution $B(n,p)$ has the form

$$\mathbb{P}(X=k)=\binom{n}{k}p^k q^{n-k},\quad k=0,1,\dots,n. \tag{1.9}$$

It is straightforward to obtain

$$\mathbb{E}X=np,\quad \mathrm{Var}(X)=npq.$$

The binomial distribution $B(n,p)$ can be obtained from the sum of n independent Bernoulli trials (Exercise 1.1).

Example 1.9 (Poisson distribution $\mathcal{P}(\lambda)$)**.** The Poisson distribution $\mathcal{P}(\lambda)$ has the form

$$\mathbb{P}(X=k)=\frac{\lambda^k}{k!}e^{-\lambda},\quad k=0,1,\dots. \tag{1.10}$$

The mean and variance are, respectively,

$$\mathbb{E}X=\mathrm{Var}(X)=\lambda.$$

This is often used to model the number of events during a time interval or the number of points that fall in a given set.

The Poisson distribution $\mathcal{P}(\lambda)$ may be viewed as the limit of the binomial distribution in the sense that

$$C_n^k p^k q^{n-k}\longrightarrow\frac{\lambda^k}{k!}e^{-\lambda}\ \text{ as } n\to\infty, p\to 0, np=\lambda.$$

The proof is simple and is left as an exercise.

1.5. Continuous Distributions

Consider now the general case when Ω is not necessarily enumerable. Let us begin with the definition of a *random variable*. Denote by $\mathcal{R}$ the Borel σ-algebra on $\mathbb{R}$, the smallest σ-algebra containing all open sets.

Definition 1.10. A random variable X is an $\mathcal{F}$-measurable real-valued function $X:\Omega\to\mathbb{R}$; i.e., for any $B\in\mathcal{R}$, $X^{-1}(B)\in\mathcal{F}$.

Definition 1.11. The *distribution* of the random variable X is a probability measure μ on $\mathbb{R}$, defined for any set $B\in\mathcal{R}$ by

$$\mu(B)=\mathbb{P}(X\in B)=\mathbb{P}\circ X^{-1}(B). \tag{1.11}$$

In particular, we define the *distribution function* $F(x)=\mathbb{P}(X\le x)$ when $B=(-\infty,x]$.

If there exists an integrable function $\rho(x)$ such that

$$\mu(B) = \int_B \rho(x)dx \tag{1.12}$$

for any $B \in \mathcal{R}$, then ρ is called the *probability density function* (PDF) of X. Here $\rho(x) = d\mu/dm$ is the Radon-Nikodym derivative of $\mu(dx)$ with respect to the Lebesgue measure $m(dx)$ if $\mu(dx)$ is absolutely continuous with respect to $m(dx)$; i.e., for any set $B \in \mathcal{R}$, if $m(B) = 0$, then $\mu(B) = 0$ (see also Section C of the appendix) [**Bil79**]. In this case, we write $\mu \ll m$.

Definition 1.12. The *expectation* of a random variable X is defined as

$$\mathbb{E}X = \int_\Omega X(\omega)\mathbb{P}(d\omega) = \int_{\mathbb{R}} x\mu(dx) \tag{1.13}$$

if the integrals are well-defined.

The *variance* of X is defined as

$$\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2. \tag{1.14}$$

For two random variables X and Y, we can define their *covariance* as

$$\mathrm{Cov}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y). \tag{1.15}$$

X and Y are called *uncorrelated* if $\mathrm{Cov}(X, Y) = 0$.

All of the above definitions can be extended to the vectorial case in which $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)^T \in \mathbb{R}^d$ is a *random vector* and each component X_k is a random variable. In this case, the *covariance matrix* of $\boldsymbol{X}$ is defined as

$$\mathrm{Cov}(\boldsymbol{X}) = \mathbb{E}(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})^T. \tag{1.16}$$

Definition 1.13. For any $p \geq 1$, the space $L^p(\Omega)$ (or L^p_ω) consists of random variables whose pth-order moment is finite:

$$L^p(\Omega) = \{\boldsymbol{X}(\omega) : \mathbb{E}|\boldsymbol{X}|^p < \infty\}. \tag{1.17}$$

For $\boldsymbol{X} \in L^p(\Omega)$, let

$$\|\boldsymbol{X}\|_p = (\mathbb{E}|\boldsymbol{X}|^p)^{1/p}, \quad p \geq 1. \tag{1.18}$$

Theorem 1.14.

(i) *Minkowski inequality.*

$$\|\boldsymbol{X} + \boldsymbol{Y}\|_p \leq \|\boldsymbol{X}\|_p + \|\boldsymbol{Y}\|_p, \qquad p \geq 1,\ \boldsymbol{X}, \boldsymbol{Y} \in L^p(\Omega)$$

(ii) *Hölder inequality.*

$$\mathbb{E}|(\boldsymbol{X}, \boldsymbol{Y})| \leq \|\boldsymbol{X}\|_p \|\boldsymbol{Y}\|_q, \quad p > 1,\ 1/p + 1/q = 1,\ \boldsymbol{X} \in L^p(\Omega),\ \boldsymbol{Y} \in L^q(\Omega),$$

where $(\boldsymbol{X}, \boldsymbol{Y})$ *denotes the standard scalar product in* $\mathbb{R}^d$.

(iii) *Schwartz inequality.*

$$\mathbb{E}|(\boldsymbol{X},\boldsymbol{Y})| \le \|\boldsymbol{X}\|_2\|\boldsymbol{Y}\|_2.$$

Obviously Schwartz inequality is a special case of Hölder inequality when $p = q = 2$.

The proof of these inequalities can be found, for example, in Chapter 2 of [**Shi96**].

It also follows that $\|\cdot\|_p$ is a norm. One can further prove that $L^p(\Omega)$ is a Banach space and $L^2(\Omega)$ is a Hilbert space with inner product

$$(\boldsymbol{X},\boldsymbol{Y})_{L^2_\omega} = \mathbb{E}(\boldsymbol{X},\boldsymbol{Y}). \tag{1.19}$$

Lemma 1.15 (Chebyshev's inequality)**.** *Let $\boldsymbol{X}$ be a random variable such that $\mathbb{E}|\boldsymbol{X}|^p < \infty$ for some $p > 0$. Then*

$$\mathbb{P}\{|\boldsymbol{X}| \ge \lambda\} \le \frac{1}{\lambda^p}\mathbb{E}|\boldsymbol{X}|^p, \tag{1.20}$$

for any positive constant λ.

Proof. For any $\lambda > 0$,

$$\mathbb{E}|\boldsymbol{X}|^p = \int_{\mathbb{R}^d} |\boldsymbol{x}|^p \mu(d\boldsymbol{x}) \ge \int_{|\boldsymbol{x}|\ge\lambda} |\boldsymbol{x}|^p \mu(d\boldsymbol{x}) \ge \lambda^p \int_{|\boldsymbol{x}|\ge\lambda} \mu(d\boldsymbol{x}) = \lambda^p \mathbb{P}(|\boldsymbol{X}| \ge \lambda).$$

□

It is straightforward to generalize the above estimate to any nonnegative increasing function $f(x)$, which gives $\mathbb{P}(|\boldsymbol{X}| \ge \lambda) \le \mathbb{E}f(|\boldsymbol{X}|)/f(\lambda)$ if $f(\lambda) > 0$.

Lemma 1.16 (Jensen's inequality)**.** *Let $\boldsymbol{X}$ be a random variable such that $\mathbb{E}|\boldsymbol{X}| < \infty$ and $\phi : \mathbb{R} \to \mathbb{R}$ is a convex function such that $\mathbb{E}|\phi(\boldsymbol{X})| < \infty$. Then*

$$\mathbb{E}\phi(\boldsymbol{X}) \ge \phi(\mathbb{E}\boldsymbol{X}). \tag{1.21}$$

This follows directly from the definition of convex functions. Readers can also refer to [**Chu01**] for the details.

Below we list some typical continuous distributions.

Example 1.17 (Uniform distribution)**.** The uniform distribution on a domain B (in $\mathbb{R}^d$) is defined by the probability density function:

$$\rho(x) = \begin{cases} \frac{1}{\mathrm{vol}(B)}, & \text{if } \boldsymbol{x} \in B, \\ 0, & \text{otherwise.} \end{cases}$$

In one dimension if $B = [0, 1]$ (denoted as $\mathcal{U}[0, 1]$ later), this reduces to

$$\rho(x) = \begin{cases} 1, & \text{if } x \in [0, 1], \\ 0, & \text{otherwise.} \end{cases}$$

For the uniform distribution on $[0, 1]$, we have

$$\mathbb{E}X = \frac{1}{2}, \quad \mathrm{Var}(X) = \frac{1}{12}.$$

Example 1.18 (Exponential distribution)**.** The exponential distribution $\mathcal{E}(\lambda)$ is defined by the probability density function:

$$\rho(x) = \begin{cases} 0, & \text{if } x < 0, \\ \lambda e^{-\lambda x}, & \text{if } x \geq 0. \end{cases}$$

The mean and variance of $E(\lambda)$ are

$$\mathbb{E}X = \frac{1}{\lambda}, \quad \mathrm{Var}(X) = \frac{1}{\lambda^2}. \tag{1.22}$$

As an example, the waiting time of a Poisson process with rate λ is exponentially distributed with parameter λ.

Example 1.19 (Normal distribution)**.** The one-dimensional normal distribution (also called Gaussian distribution) $N(\mu, \sigma^2)$ is defined by the probability density function:

$$\rho(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) \tag{1.23}$$

with mean μ and variance σ^2.

If $\mathbf{\Sigma}$ is an $n \times n$ symmetric positive definite matrix and $\boldsymbol{\mu}$ is a vector in $\mathbb{R}^n$, we can also define the n-dimensional normal distribution $N(\boldsymbol{\mu}, \mathbf{\Sigma})$ through the density

$$\rho(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}(\det \mathbf{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right). \tag{1.24}$$

In this case, we have

$$\mathbb{E}\boldsymbol{X} = \boldsymbol{\mu}, \quad \mathrm{Cov}(\boldsymbol{X}) = \mathbf{\Sigma}.$$

The normal distribution is the most important probability distribution. It is also called the Gaussian distribution. Random variables with normal distribution are also called Gaussian random variables. In the case of degeneracy, i.e., the covariance matrix $\mathbf{\Sigma}$ is not invertible, which corresponds to the case that some components are in the subspace spanned by other components, we need to define the Gaussian distribution via characteristic functions (see Section 1.9).

Example 1.20 (Gibbs distribution)**.** In equilibrium statistical mechanics, we are concerned with a probability distribution π over a state space S. In the case of an n-particle system with continuous states, we have $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_n) \in S = \mathbb{R}^{6n}$, where $\boldsymbol{x}_k$ and $\boldsymbol{p}_k$ are the position and momentum of the kth particle, respectively. The PDF $\pi(\boldsymbol{x})$, called the *Gibbs distribution*, has a specific form:

$$\pi(\boldsymbol{x}) = \frac{1}{Z} e^{-\beta H(\boldsymbol{x})}, \qquad \boldsymbol{x} \in \mathbb{R}^{6n},\ \beta = (k_B T)^{-1}, \tag{1.25}$$

where H is the energy of the considered system, T is the absolute temperature, k_B is the Boltzmann constant, and

$$Z = \int_{\mathbb{R}^{6n}} e^{-\beta H(\boldsymbol{x})} d\boldsymbol{x} \tag{1.26}$$

is called the partition function. Let $f(\boldsymbol{x})$ be a function defined on the configuration space S. Then its thermodynamic average $\langle f \rangle$, i.e., the expectation of f, is given by

$$\mathbb{E} f = \langle f \rangle = \int_{\mathbb{R}^{6n}} f(\boldsymbol{x}) \pi(\boldsymbol{x}) d\boldsymbol{x}. \tag{1.27}$$

A similar definition holds for the discrete space setting by replacing the integral with summation.

Another important notion is the distribution function of a random variable X:

$$F(x) = \mathbb{P}(X < x). \tag{1.28}$$

One can easily see that if the distribution of X is absolutely continuous with respect to the Lebesgue measure, then the density ρ and the distribution function F of X are related by

$$\rho(x) = \frac{d}{dx} F(x). \tag{1.29}$$

1.6. Independence

We now come to one of the most distinctive notions in probability theory, the notion of independence. Let us start by defining the independence of events. Two events $A, B \in \mathcal{F}$ are *independent* if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Definition 1.21. Two random variables X and Y are said to be independent if for any two Borel sets A and B, $X^{-1}(A)$ and $Y^{-1}(B)$ are independent; i.e.,

$$\mathbb{P}\left(X^{-1}(A) \cap Y^{-1}(B)\right) = \mathbb{P}\left(X^{-1}(A)\right) \mathbb{P}\left(Y^{-1}(B)\right). \tag{1.30}$$

The joint distribution of the two random variables X and Y is defined to be the distribution of the random vector (X, Y). Let μ_1 and μ_2 be the probability distribution of X and Y, respectively, and let μ be their joint distribution. If X and Y are independent, then for any two Borel sets A and B, we have

$$\mu(A \times B) = \mu_1(A)\mu_2(B). \tag{1.31}$$

Consequently, we have

$$\mu = \mu_1\mu_2; \tag{1.32}$$

i.e., the joint distribution of two independent random variables is the product distribution. If both μ_1 and μ_2 are absolutely continuous, with densities p_1 and p_2, respectively, then μ is also absolutely continuous, with density given by

$$p(x, y) = p_1(x)p_2(y). \tag{1.33}$$

One can also understand independence from the viewpoint of expectations. Let f_1 and f_2 be two continuous functions. If X and Y are two independent random variables, then

$$\mathbb{E}f_1(X)f_2(Y) = \mathbb{E}f_1(X)\mathbb{E}f_2(Y). \tag{1.34}$$

In fact, this can also be used as the definition of independence.

This discussion can be readily generalized to multiple events and multiple random variables. A sequence of events $\{A_k\}_{k=1}^n$ are independent if for $k = 1, 2 \ldots, n$ and $1 \le i_1 < i_2 < \cdots < i_k \le n$

$$\mathbb{P}\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} \mathbb{P}(A_{i_j}).$$

Note that pairwise independence does not imply independence (see Exercise 1.9).

Definition 1.22. The random variables $X_1, \ldots, X_n$ are said to be independent if for any Borel sets $B_j \in \mathcal{R}$,

$$\mathbb{P}\left(\bigcap_{j=1}^{n} X_j^{-1}(B_j)\right) = \prod_{j=1}^{n} \mathbb{P}(X_j^{-1}(B_j)). \tag{1.35}$$

If the random variables are independent, then their joint distribution is the product measure of their distributions.

1.7. Conditional Expectation

Let X and Y be two discrete random variables, not necessarily independent, with joint distribution

$$p(i,j) = \mathbb{P}(X = i, Y = j).$$

Since $\sum_i p(i,j) = \mathbb{P}(Y = j)$, the *probability* that $X = i$ conditioned on $Y = j$ is given by

$$p(i|j) = \frac{p(i,j)}{\sum_i p(i,j)} = \frac{p(i,j)}{\mathbb{P}(Y = j)}$$

if $\sum_i p(i,j) > 0$. The convention is to set $p(i|j) = 0$ if $\sum_i p(i,j) = 0$. Now let f be a continuous function. It is natural to define the *conditional expectation* of $f(X)$ given that $Y = j$ by

$$\mathbb{E}(f(X)|Y = j) = \sum_i f(i)p(i|j). \tag{1.36}$$

A difficulty arises when one tries to generalize this to continuous random variables. Indeed, given two continuous random variables X and Y, the probability that $Y = y$ is zero for most values of y. Therefore, we need to proceed differently.

Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. Let X be a random variable such that $\mathbb{E}|X| < \infty$.

Definition 1.23 (Conditional expectation)**.** The conditional expectation Z of X given $\mathcal{G}$ is defined by the following conditions:

(i) Z is measurable with respect to $\mathcal{G}$;

(ii) for any set $A \in \mathcal{G}$,

$$\int_A Z(\omega)\mathbb{P}(d\omega) = \int_A X(\omega)\mathbb{P}(d\omega).$$

The existence of $Z = \mathbb{E}(X|\mathcal{G})$ follows from the Radon-Nikodym theorem if we consider the measure μ on $\mathcal{G}$ defined by $\mu(A) = \int_A X(\omega)\mathbb{P}(d\omega)$ (see [**Bil79**]). One can easily see that μ is absolutely continuous with respect to the measure $\mathbb{P}|_{\mathcal{G}}$, the restriction of $\mathbb{P}$ to $\mathcal{G}$. Thus Z exists and is unique up to almost sure equivalence in $\mathbb{P}|_{\mathcal{G}}$.

In addition, we have

Theorem 1.24 (Properties of conditional expectation)**.** *Assume that X, Y are random variables with $\mathbb{E}|X|, \mathbb{E}|Y| < \infty$. Let $a, b \in \mathbb{R}$. Then:*

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

(ii) $\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(X)$.

(iii) $\mathbb{E}(X|\mathcal{G}) = X$ *if* X *is* $\mathcal{G}$*-measurable.*

(iv) $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}X$ *if* X *is independent of* $\mathcal{G}$.

(v) $\mathbb{E}(XY|\mathcal{G}) = Y\mathbb{E}(X|\mathcal{G})$ *if* Y *is* $\mathcal{G}$*-measurable.*

(vi) *If* $\mathcal{H}$ *is a sub-*σ*-algebra of* $\mathcal{G}$*, then* $\mathbb{E}(X|\mathcal{H}) = \mathbb{E}(\mathbb{E}(X|\mathcal{H})|\mathcal{G})$.

Similar to Lemma 1.16 we have

Lemma 1.25 (Conditional Jensen inequality)**.** *Let* X *be a random variable such that* $\mathbb{E}|X| < \infty$ *and let* $\phi : \mathbb{R} \to \mathbb{R}$ *be a convex function such that* $\mathbb{E}|\phi(X)| < \infty$*. Then*

$$\mathbb{E}(\phi(X)|\mathcal{G}) \geq \phi(\mathbb{E}(X|\mathcal{G})). \tag{1.37}$$

These statements are straightforward. The reader may also consult [**Chu01**] for the details of the proof.

Given two random variables X and Y, the conditional expectation of X with respect to Y is defined as the conditional expectation of X with respect to the σ-algebra $\mathcal{G} = \sigma(Y)$ generated by Y

$$\mathcal{G} = \sigma(Y) := \{Y^{-1}(B), B \in \mathcal{R}\}.$$

To see that this definition reduces to the previous one for the case of discrete random variables, let $\Omega_j = \{\omega : Y(\omega) = j\}$ and

$$\Omega = \bigcup_{j=1}^{n} \Omega_j.$$

The σ-algebra $\mathcal{G}$ is simply the set of all possible unions of the Ω_j's. The measurability condition of $\mathbb{E}(X|Y)$ with respect to $\mathcal{G}$ means $\mathbb{E}(X|Y)$ is constant on each Ω_j, which is exactly $\mathbb{E}(X|Y = j)$ as we will see. By definition, we have

$$\int_{\Omega_j} \mathbb{E}(X|Y)\mathbb{P}(d\omega) = \int_{\Omega_j} X(\omega)\mathbb{P}(d\omega), \tag{1.38}$$

which implies

$$\mathbb{E}(X|Y) = \frac{1}{\mathbb{P}(\Omega_j)} \int_{\Omega_j} X(\omega)\mathbb{P}(d\omega). \tag{1.39}$$

This is exactly $\mathbb{E}(X|Y = j)$ in (1.36) when $f(X) = X$.

The conditional expectation is the optimal approximation in $L^2(\Omega)$ among all $\mathcal{G}$-measurable functions.

Proposition 1.26. *Let* g *be a measurable function. Then*

$$\mathbb{E}(X - \mathbb{E}(X|Y))^2 \leq \mathbb{E}(X - g(Y))^2. \tag{1.40}$$

Proof. We have

$$\mathbb{E}(X - g(Y))^2 = \mathbb{E}(X - E(X|Y))^2 + \mathbb{E}(E(X|Y) - g(Y))^2 + 2\mathbb{E}\Big[(X - E(X|Y))(E(X|Y) - g(Y))\Big]$$

and

$$\begin{aligned}&\mathbb{E}\Big[(X - \mathbb{E}(X|Y))(\mathbb{E}(X|Y) - g(Y))\Big]\\ &\quad = \mathbb{E}\Big[\mathbb{E}\big[(X - \mathbb{E}(X|Y))(E(X|Y) - g(Y))|Y\big]\Big]\\ &\quad = \mathbb{E}\Big[(\mathbb{E}(X|Y) - \mathbb{E}(X|Y))(E(X|Y) - g(Y))\Big] = 0,\end{aligned}$$

which implies (1.40). □

1.8. Notions of Convergence

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let μ_n be the distribution of X_n. Let X be another random variable with distribution μ. We will discuss four notions of convergence: almost sure convergence, convergence in probability, convergence in distribution, and convergence in L^p.

Definition 1.27 (Almost sure convergence)**.** X_n converges to X almost surely if

$$\mathbb{P}(\omega : X_n(\omega) \to X(\omega)) = 1. \tag{1.41}$$

We write almost sure convergence as $X_n \to X$, a.s.

Definition 1.28 (Convergence in probability)**.** X_n converges to X in probability if for any $\epsilon > 0$,

$$\mathbb{P}(\omega : |X_n(\omega) - X(\omega)| > \epsilon) \to 0, \tag{1.42}$$

as $n \to \infty$.

Definition 1.29 (Convergence in distribution)**.** X_n converges to X in distribution if for any $f \in C_b(\mathbb{R})$,

$$\mathbb{E}f(X_n) \to \mathbb{E}f(X). \tag{1.43}$$

This is denoted as $X_n \xrightarrow{d} X$ or $\mu_n \xrightarrow{d} \mu$ or $\mu_n \Rightarrow \mu$.

Definition 1.30 (Convergence in L^p)**.** X_n converges to X in L^p $(0 < p < \infty)$ if

$$\mathbb{E}|X_n - X|^p \to 0. \tag{1.44}$$

For $p = 1$, this is also referred to as convergence in mean; for $p = 2$, this is referred to as convergence in mean square.

Remark 1.31. Note that the power p does not need to be greater than 1 since it still gives a metric for the random variables. But only when $p \geq 1$ do we get a norm. The statements in this section hold when $0 < p < \infty$.

We have the following relations between different notions of convergence.

Theorem 1.32.

(i) *Almost sure convergence implies convergence in probability.*

(ii) *Convergence in probability implies almost sure convergence along a subsequence.*

(iii) *If $p < q$, then convergence in L^q implies convergence in L^p.*

(iv) *Convergence in L^p implies convergence in probability.*

(v) *Convergence in probability implies convergence in distribution.*

Proof. (i) Note that

$$\mathbb{P}(|X_n(\omega) - X(\omega)| > \epsilon) = \int_\Omega \chi_{\{|X_n - X| > \epsilon\}}(\omega)\mathbb{P}(d\omega) \to 0$$

by the almost sure convergence and dominated convergence theorems.

(ii) The proof will be deferred to Section 1.11.

(iii) This is a consequence of the Hölder inequality:

$$\mathbb{E}|X_n - X|^p \leq \left(\mathbb{E}(|X_n - X|^p)^{\frac{q}{p}}\right)^{\frac{p}{q}} = (\mathbb{E}|X_n - X|^q)^{\frac{p}{q}}, \quad p < q.$$

(iv) This is a consequence of the Chebyshev inequality:

$$\mathbb{P}(\omega : |X_n(\omega) - X(\omega)| > \epsilon) \leq \frac{1}{\epsilon^p}\mathbb{E}|X_n - X|^p$$

for any $\epsilon > 0$.

(v) Argue by contradiction. Suppose there exist a bounded continuous function $f(x)$ and a subsequence X_{n_k} such that

$$\mathbb{E}f(X_{n_k}) \not\to \mathbb{E}f(X), \quad k \to \infty. \tag{1.45}$$

From assertion (ii), there exists a further subsequence of $\{X_{n_k}\}$ (still denoted as $\{X_{n_k}\}$) such that X_{n_k} converges to X almost surely. This contradicts (1.45) by the dominated convergence theorem. □

1.9. Characteristic Function

The *characteristic function* of a random variable X is defined as

$$f(\xi) = \mathbb{E}e^{i\xi X} = \int_{\mathbb{R}} e^{i\xi x}\mu(dx). \tag{1.46}$$

Proposition 1.33. *The characteristic function has the following properties:*

(1) $\forall \xi \in \mathbb{R}$, $|f(\xi)| \leq 1$, $f(\xi) = \overline{f(-\xi)}$, $f(0) = 1$;

(2) f *is uniformly continuous on* $\mathbb{R}$.

Proof. The proof of the first statements is straightforward. For the second statement, we have

$$\begin{aligned} |f(\xi_1) - f(\xi_2)| &= |\mathbb{E}(e^{i\xi_1 X} - e^{i\xi_2 X})| = |\mathbb{E}(e^{i\xi_1 X}(1 - e^{i(\xi_2 - \xi_1)X}))| \\ &\leq \mathbb{E}|1 - e^{i(\xi_2 - \xi_1)X}|. \end{aligned}$$

Since the integrand $|1 - e^{i(\xi_2 - \xi_1)X}|$ depends only on the difference between ξ_1 and ξ_2 and since it tends to 0 almost surely as the difference goes to 0, uniform continuity follows immediately from the dominated convergence theorem. □

Example 1.34. Here are the characteristic functions of some typical distributions:

(1) Bernoulli distribution.

$$f(\xi) = q + pe^{i\xi}.$$

(2) Binomial distribution $B(n, p)$.

$$f(\xi) = (q + pe^{i\xi})^n.$$

(3) Poisson distribution $\mathcal{P}(\lambda)$.

$$f(\xi) = e^{\lambda(e^{i\xi} - 1)}.$$

(4) Exponential distribution $\mathcal{E}(\lambda)$.

$$f(\xi) = (1 - \lambda^{-1} i\xi)^{-1}.$$

(5) Normal distribution $N(\mu, \sigma^2)$.

$$f(\xi) = \exp\left(i\mu\xi - \frac{\sigma^2 \xi^2}{2}\right). \tag{1.47}$$

(6) Multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

$$f(\boldsymbol{\xi}) = \exp\left(i\boldsymbol{\mu}^T \boldsymbol{\xi} - \frac{1}{2}\boldsymbol{\xi}^T \boldsymbol{\Sigma} \boldsymbol{\xi}\right). \tag{1.48}$$

The property (1.48) is also used to define degenerate multivariate Gaussian distributions.

The following result gives an explicit characterization of the weak convergence of probability measures in terms of their characteristic functions. This is the key ingredient in the proof of the central limit theorem.

Theorem 1.35 (Lévy's continuity theorem)**.** *Let $\{\mu_n\}_{n\in\mathbb{N}}$ be a sequence of probability measures, and let $\{f_n\}_{n\in\mathbb{N}}$ be their corresponding characteristic functions. Assume that:*

(1) *f_n converges everywhere on $\mathbb{R}$ to a limiting function f.*

(2) *f is continuous at $\xi = 0$.*

Then there exists a probability distribution μ such that $\mu_n \xrightarrow{d} \mu$. Moreover f is the characteristic function of μ.

Conversely, if $\mu_n \xrightarrow{d} \mu$, where μ is some probability distribution, then f_n converges to f uniformly on every finite interval, where f is the characteristic function of μ.

For a proof, see [**Chu01**].

As for Fourier transforms, one can also define the inverse transform

$$\rho(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-i\xi x} f(\xi) d\xi.$$

An interesting question arises as to when this gives the density of a probability measure. To address this, we introduce the following notion.

Definition 1.36. A function f is called positive semidefinite if for any finite set of values $\{\xi_1, \ldots, \xi_n\}$, $n \in \mathbb{N}$, the matrix $(f(\xi_i - \xi_j))_{i,j=1}^n$ is positive semidefinite; i.e.,

$$\sum_{i,j} f(\xi_i - \xi_j) v_i \bar{v}_j \geq 0, \tag{1.49}$$

for every set of values $v_1, \ldots, v_n \in \mathbb{C}$.

Theorem 1.37 (Bochner's theorem)**.** *A function f is the characteristic function of a probability measure if and only if it is positive semidefinite and continuous at 0 with $f(0) = 1$.*

Proof. We only prove the necessity part. The other part is less trivial and readers may consult [**Chu01**]. Assume that f is a characteristic function. Then

$$\sum_{i,j=1}^n f(\xi_i - \xi_j) v_i \bar{v}_j = \int_{\mathbb{R}} \Big| \sum_{i=1}^n v_i e^{i\xi_i x} \Big|^2 dx \geq 0. \tag{1.50}$$ □

1.10. Generating Function and Cumulants

For a discrete random variable, its *generating function* is defined as

$$G(x) = \sum_{k=0}^{\infty} P(X = x_k) x^k. \tag{1.51}$$

One immediately has

$$P(X = x_k) = \frac{1}{k!} G^{(k)}(x)\Big|_{x=0}.$$

Definition 1.38. The convolution of two sequences $\{a_k\}$, $\{b_k\}$, $\{c_k\} = \{a_k\} * \{b_k\}$, is defined by

$$c_k = \sum_{j=0}^{k} a_j b_{k-j}. \tag{1.52}$$

It is easy to show that the generating functions defined by

$$A(x) = \sum_{k=0}^{\infty} a_k x^k, \quad B(x) = \sum_{k=0}^{\infty} b_k x^k, \quad C(x) = \sum_{k=0}^{\infty} c_k x^k$$

with $\{c_k\} = \{a_k\} * \{b_k\}$ satisfy $C(x) = A(x)B(x)$. The following result is more or less obvious.

Theorem 1.39. *Let X and Y be two independent random variables with probability distribution*

$$P(X = j) = a_j, \quad P(Y = k) = b_k,$$

respectively, and let A and B be the corresponding generating functions. Then the generating function of $X + Y$ is $C(x) = A(x)B(x)$.

This can be used to compute some generating functions.

(a) Bernoulli distribution: $G(x) = q + px$.

(b) Binomial distribution: $G(x) = (q + px)^n$.

(c) Poisson distribution: $G(x) = e^{-\lambda+\lambda x}$.

The moment generating function of a random variable X is defined for all values of t by

$$M(t) = \mathbb{E}e^{tX} = \begin{cases} \displaystyle\sum_x p(x)e^{tx}, & X \text{ is discrete-valued,} \\ \displaystyle\int_{\mathbb{R}} p(x)e^{tx}dx, & X \text{ is continuous,} \end{cases} \tag{1.53}$$

provided that e^{tX} is integrable. It is obvious that $M(0) = 1$.

Once $M(t)$ is defined, one can show $M(t) \in C^\infty$ in its domain and its relation to the nth moment

$$M^{(n)}(t) = \mathbb{E}(X^n e^{tX}) \quad \text{and} \quad \mu_n := \mathbb{E}X^n = M^{(n)}(0),\ n \in \mathbb{N}. \tag{1.54}$$

This gives

$$M(t) = \sum_{n=0}^{\infty} \mu_n \frac{t^n}{n!}, \tag{1.55}$$

which explains why $M(t)$ is called the moment generating function.

Theorem 1.40. *Denote $M_X(t), M_Y(t)$, and $M_{X+Y}(t)$ the moment generating functions of the random variables X, Y, and $X+Y$, respectively. If X, Y are independent, then*

$$M_{X+Y}(t) = M_X(t) M_Y(t). \tag{1.56}$$

Proof. The proof is straightforward by noticing

$$M_{X+Y}(t) = \mathbb{E}e^{t(X+Y)} = \mathbb{E}e^{tX}\mathbb{E}e^{tY} = M_X(t)M_Y(t). \qquad \square$$

The following can be obtained by direct calculation.

(a) Binomial distribution: $M(t) = (pe^t + 1 - p)^n$.

(b) Poisson distribution: $M(t) = \exp[\lambda(e^t - 1)]$.

(c) Exponential distribution: $M(t) = \lambda/(\lambda - t)$ for $t < \lambda$.

(d) Normal distribution $N(\mu, \sigma^2)$: $M(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$.

The cumulant generating function $K(t)$ is defined by

$$\Lambda(t) = \log M(t) = \log \mathbb{E}e^{tX} = \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!}. \tag{1.57}$$

With such a definition, we have $\kappa_0 = 0$ and

$$\kappa_n = \frac{d^n}{dt^n}\Lambda(0), \quad n \in \mathbb{N}. \tag{1.58}$$

It is straightforward to define moment and cumulant generating functions for random vectors:

$$M(\boldsymbol{t}) = \mathbb{E}e^{\boldsymbol{t}\cdot\boldsymbol{X}}, \ \Lambda(\boldsymbol{t}) = \log M(\boldsymbol{t}), \quad \boldsymbol{t} \in \mathbb{R}^d.$$

These notions are useful, for example, in statistical physics [**TKS95**], [**KTH95**].

1.11. The Borel-Cantelli Lemma

We now turn to a technical tool that will be useful in the next chapter. Given a sequence of events $\{A_n\}_{n=1}^{\infty}$, we are interested in the event that A_n occurs infinitely often, i.e.,

$$\{A_n \text{ i.o.}\} = \{\omega : \omega \in A_n \text{ i.o.}\} = \limsup_{k\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k.$$

The probability of such an event can be characterized by the Borel-Cantelli lemma.

Lemma 1.41.

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

(2) *If* $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ *and the* A_n*'s are mutually independent, then* $\mathbb{P}(\{A_n \text{ } i.o.\}) = 1$.

Proof. (1) We have

$$\mathbb{P}\left(\left\{\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k\right\}\right) \leq \mathbb{P}\left(\left\{\bigcup_{k=n}^{\infty} A_k\right\}\right) \leq \sum_{k=n}^{\infty} \mathbb{P}(A_k)$$

for any n. The last term goes to 0 as $n \to \infty$ since $\sum_{k=1}^{\infty} \mathbb{P}(A_k) < \infty$ by assumption.

(2) Using independence, one has

$$\mathbb{P}\left(\bigcup_{k=n}^{\infty} A_k\right) = 1 - \mathbb{P}\left(\bigcap_{k=n}^{\infty} A_k^c\right) = 1 - \prod_{k=n}^{\infty} \mathbb{P}(A_k^c) = 1 - \prod_{k=n}^{\infty}(1 - \mathbb{P}(A_k)).$$

Using $1 - x \leq e^{-x}$, this gives

$$\mathbb{P}\left(\bigcup_{k=n}^{\infty} A_k\right) \geq 1 - \prod_{k=n}^{\infty} e^{-\mathbb{P}(A_k)} = 1 - e^{-\sum_{k=n}^{\infty} \mathbb{P}(A_k)} = 1.$$

The result follows immediately. □

As an example of the application of this result, we prove

Lemma 1.42. *Let* $\{X_n\}_{n\in\mathbb{N}}$ *be a sequence of identically distributed (not necessarily independent) random variables, such that* $\mathbb{E}|X_n| < \infty$. *Then*

$$\lim_{n\to\infty} \frac{X_n}{n} = 0 \qquad a.s.$$

Proof. For any $\epsilon > 0$, define

$$A_n^{\epsilon} = \{\omega \in \Omega : |X_n(\omega)/n| > \epsilon\}.$$

Then

$$\begin{aligned}\sum_{n=1}^{\infty}\mathbb{P}(A_n^{\epsilon}) &= \sum_{n=1}^{\infty}\mathbb{P}(|X_n|>n\epsilon) = \sum_{n=1}^{\infty}\mathbb{P}(|X_1|>n\epsilon)\\ &= \sum_{n=1}^{\infty}\sum_{k=n}^{\infty}\mathbb{P}\left(k\epsilon<|X_1|\le(k+1)\epsilon\right)\\ &= \sum_{k=1}^{\infty}k\mathbb{P}\left(k\epsilon<|X_1|\le(k+1)\epsilon\right)\\ &= \sum_{k=1}^{\infty}k\int_{k\epsilon<|X_1|\le(k+1)\epsilon}\mathbb{P}(d\omega)\\ &\le \frac{1}{\epsilon}\sum_{k=1}^{\infty}\int_{k\epsilon<|X_1|\le(k+1)\epsilon}|X_1|\mathbb{P}(d\omega)\\ &= \frac{1}{\epsilon}\int_{\epsilon<|X_1|}|X_1|\mathbb{P}(d\omega)\le\frac{1}{\epsilon}\mathbb{E}|X_1|<\infty.\end{aligned}$$

Therefore if we define

$$B_{\epsilon}=\{\omega\in\Omega:\omega\in A_n^{\epsilon}\text{ i.o.}\},$$

then $\mathbb{P}(B_{\epsilon})=0$. Let $B=\bigcup_{n=1}^{\infty}B_{\frac{1}{n}}$. Then $\mathbb{P}(B)=0$, and

$$\lim_{n\to\infty}\frac{X_n(\omega)}{n}=0\quad\text{if }\omega\notin B.$$

The proof is complete. □

With the Borel-Cantelli lemma, we can give the proof of (ii) in Theorem 1.32.

Proof. Without loss of generality, we may assume that $X=0$. By the condition of convergence in probability we can take $n_k\in\mathbb{N}$ which is increasing in k such that

$$\mathbb{P}\left(|X_{n_k}|\ge\frac{1}{k}\right)\le\frac{1}{2^k}$$

for each $k\in\mathbb{N}$. For any fixed ϵ, there exists an integer k_0 such that $k_0^{-1}\le\epsilon$ and we have

$$\begin{aligned}\sum_{k=1}^{\infty}\mathbb{P}(|X_{n_k}|\ge\epsilon) &\le \sum_{k=1}^{k_0}\mathbb{P}(|X_{n_k}|\ge\epsilon)+\sum_{k=k_0+1}^{\infty}\mathbb{P}\left(|X_{n_k}|\ge\frac{1}{k}\right)\\ &\le \sum_{k=1}^{k_0}\mathbb{P}(|X_{n_k}|\ge\epsilon)+\sum_{k=k_0+1}^{\infty}\frac{1}{2^k}<\infty.\end{aligned}$$

By the Borel-Cantelli lemma we obtain

$$\mathbb{P}(|X_{n_k}| \geq \epsilon, \text{ i.o.}) = 0.$$

Thus the subsequence X_{n_k} converges to 0 almost surely by the same argument as the example above. □

Exercises

In the following, i.i.d. stands for independent, identically distributed.

1.1. Let $X = \sum_{j=1}^{n} X_j$, where the X_j are i.i.d. random variables with Bernoulli distribution with parameter p. Prove that $X \sim B(n, p)$.

1.2. Let S_n be the number of heads in the outcome of n independent tosses for a fair coin. Its distribution is given by

$$p_j = \mathbb{P}(S_n = j) = \frac{1}{2^n} \binom{n}{j}.$$

Show that the mean and the variance of this random variable are

$$\mathbb{E}S_n = \frac{n}{2}, \qquad \mathrm{Var}(S_n) = \frac{n}{4}.$$

This implies that $\mathrm{Var}(S_n)/(\mathbb{E}S_n)^2 \to 0$ as $n \to \infty$.

1.3. Prove that

$$C_n^k p^k q^{n-k} \longrightarrow \frac{\lambda^k}{k!} e^{-\lambda} \text{ for any fixed } k$$

as $n \to \infty, p \to 0$, and $np = \lambda$. This is the classical Poisson approximation to the binomial distribution.

1.4. Numerically investigate the limit processes

$$\text{Binomial } B(n,p) \longrightarrow \text{Poisson } \mathcal{P}(\lambda) \longrightarrow \text{Normal } N(\lambda, \lambda)$$

when $\lambda = np$, $n \gg 1$, and $\lambda \gg 1$, by comparing the plots for different distributions. Find the parameter regimes such that the approximation is valid.

1.5. Suppose $X \sim \mathcal{P}(\lambda)$, $Y \sim \mathcal{P}(\mu)$ are two independent Poisson random variables.

(a) Prove that $Z = X + Y \sim \mathcal{P}(\lambda + \mu)$.

(b) Prove that the conditional distribution of X (or Y), conditioning on $X + Y$ being fixed, i.e., $X + Y = N$, is a binomial distribution with parameter $n = N$ and $p = \lambda/(\lambda + \mu)$ (or $p = \mu/(\lambda + \mu)$).

1.6. Prove the following statements:
   (a) (Memoryless property of exponential distribution) Suppose $X \sim \mathcal{E}(\lambda)$. Prove that
   $$\mathbb{P}(X > s + t | X > s) = \mathbb{P}(X > t) \quad \text{for all } s, t > 0.$$
   (b) Let $X \in (0, \infty)$ be a random variable such that
   $$\mathbb{P}(X > s + t) = \mathbb{P}(X > s)\mathbb{P}(X > t) \quad \text{for all } s, t > 0.$$
   Prove that there exists $\lambda > 0$ such that $X \sim \mathcal{E}(\lambda)$.

1.7. Let $X = (X_1, \ldots, X_n)$ be an n-dimensional Gaussian random variable and let $Y = c_1 X_1 + c_2 X_2 + \cdots + c_n X_n$, where $c_1, \ldots, c_n$ are constants. Show that Y is also Gaussian.

1.8. Suppose that the multivariate Gaussian variable
$$\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{pmatrix} \right).$$
Prove that the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{Y} = \boldsymbol{y}$ is Gaussian with mean $\tilde{\boldsymbol{\mu}}$ and covariance $\tilde{\boldsymbol{\Sigma}}$:
$$\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y), \quad \tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}.$$

1.9. Pairwise independence does not imply independence. Let X, Y be two independent random variables such that
$$\mathbb{P}(X = \pm 1) = \mathbb{P}(Y = \pm 1) = \frac{1}{2},$$
and let $Z = XY$. Check that X, Y, Z are pairwise independent but not independent.

1.10. Independence does not imply conditional independence. Construct a concrete example such that
$$p(x, y) = p(x)p(y) \quad \text{for any } x, y$$
but $p(x, y|z) \neq p(x|z)p(y|z)$ for some x, y, and z, where x, y, z are the values of random variables X, Y, and Z.

1.11. Let $\mathcal{R}$ be the Borel σ-algebra on $\mathbb{R}$. For any random variable X, prove that the sets $\mathcal{B} = X^{-1}(\mathcal{R})$ form a σ-algebra.

1.12. If X_j $(j = 1, \ldots, n)$ are independent random variables and f_j $(j = 1, \ldots, n)$ are Borel measurable functions, i.e., $f_j^{-1}(B) \in \mathcal{R}$ for any $B \in \mathcal{R}$, then $f_j(X_j)$ $(j = 1, \ldots, n)$ are independent.

1.13. Prove that if the σ-algebra $\mathcal{F}$ has finite members, then there exist nonempty sets $B_1, \ldots, B_K \in \mathcal{F}$ such that $\Omega = \bigcup_{j=1}^{K} B_j$, $B_i \cap B_j = \emptyset$ for $i, j = 1, \ldots, K$, and any $B \in \mathcal{F}$ can be represented as a finite union of the sets in $\{B_j\}$.

1.14. (*Variance identity*) Suppose the random variable $\boldsymbol{X}$ has the partition $\boldsymbol{X} = (\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)})$. Prove the variance identity for any integrable function f:

$$\mathrm{Var}(f(\boldsymbol{X})) = \mathrm{Var}(\mathbb{E}(f(\boldsymbol{X})|\boldsymbol{X}^{(2)})) + \mathbb{E}(\mathrm{Var}(f(\boldsymbol{X})|\boldsymbol{X}^{(2)})).$$

1.15. Prove Theorem 1.24.

1.16. Let X be a discrete random variable and let $p_i = \mathbb{P}(X = i)$ for $i = 1, 2, \ldots, n$. Define the *Shannon entropy* of X by

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i.$$

Prove that $H(X) \geq 0$ and that the equality holds only when X is reduced to a deterministic variable; i.e., $p_i = \delta_{ik_0}$, for some fixed integer k_0 in $\{1, 2, \ldots, n\}$. Here we take the convention $0 \log 0 = 0$.

1.17. Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ be i.i.d. continuous random variables with common distribution function F and density $\rho(x) = F'(x)$. $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ are called the *order statistics* of $\boldsymbol{X}$ if $X_{(k)}$ is the kth smallest of these random variables. Prove that the density of the order statistics is given by

$$p(x_1, x_2, \ldots, x_n) = n! \prod_{k=1}^{n} \rho(x_k), \quad x_1 < x_2 < \cdots < x_n.$$

1.18. (*Stable laws*) A one-dimensional distribution $\mu(dx)$ is stable if given any two independent random variables X and Y with distribution $\mu(dx)$, there exists a and b such that

$$a(X + Y - b)$$

has distribution $\mu(dx)$. Show that $f(\xi) = e^{-|\xi|^\alpha}$ is the characteristic function of a stable distribution for $0 < \alpha \leq 2$ and that it is not a characteristic function for other values of α.

1.19. Prove that if the moment generating function $M(t)$ can be defined on an open set U, then $M \in C^\infty(U)$.

1.20. (*Wick's theorem*) For multivariate Gaussian random variables X_1, X_2, ..., X_n with mean 0, prove

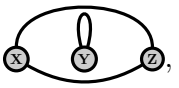
$$\mathbb{E}(X_1 X_2 \cdots X_k) = \begin{cases} \sum \prod \mathbb{E}(X_i X_j), & k \text{ is even}, \\ 0, & k \text{ is odd}, \end{cases}$$

where the notation $\sum\prod$ means summing of the products over all possible partitions of $X_1, \ldots, X_k$ into pairs; e.g., if (X, Y, Z) is jointly Gaussian, we have

$$\begin{aligned}\mathbb{E}(X^2Y^2Z^2) =&(\mathbb{E}X^2)(\mathbb{E}Y^2)(\mathbb{E}Z^2) + 2(\mathbb{E}YZ)^2\mathbb{E}X^2 + 2(\mathbb{E}XY)^2\mathbb{E}Z^2 \\ &+ 2(\mathbb{E}XZ)^2\mathbb{E}Y^2 + 8(\mathbb{E}XY)(\mathbb{E}YZ)(\mathbb{E}XZ).\end{aligned} \tag{1.59}$$

Each term in (1.59) can be schematically mapped to some graph, as shown below:

$(\mathbb{E}XY)^2\mathbb{E}Z^2 \longmapsto$  , $(\mathbb{E}YZ)^2\mathbb{E}X^2 \longmapsto$  ,

$(\mathbb{E}XZ)^2\mathbb{E}Y^2 \longmapsto$  , $(\mathbb{E}X^2)(\mathbb{E}Y^2)(\mathbb{E}Z^2) \longmapsto$  ,

$(\mathbb{E}XY)(\mathbb{E}YZ)(\mathbb{E}XZ) \longmapsto$  .

The coefficient of each term is the combinatorial number for generating the corresponding schematic combinations. This is essentially the so-called Feynman diagrams.

1.21. Suppose that the events $\{A_n\}$ are mutually independent with

$$\mathbb{P}\big(\bigcup_n A_n\big) = 1 \quad \text{and} \quad \mathbb{P}(A_n) < 1$$

for each n. Prove that $\mathbb{P}(A_n \text{ i.o.}) = 1$.

1.22. (*Girko's circular law*) If the $n \times n$ matrix $\boldsymbol{A}$ has i.i.d. entries with mean zero and variance σ^2, then the eigenvalues of $\boldsymbol{A}/\sqrt{n\sigma}$ are asymptotically uniformly distributed in the unit disk in the complex plane when $n \to \infty$. Investigate this numerically.

Notes

The axiomatic approach to probability theory starts with Kolmogorov's masterpiece [**Kol56**]. There are already lots of excellent textbooks on the introduction to elementary probability theory [**Chu01**, **Cin11**, **Dur10**, **KS07**, **Shi96**]. To know more about the measure theory approach to probability theory, please consult [**Bil79**]. More on the practical side of probability theory can be found in the classic book by Feller [**Fel68**] or the books by some applied mathematicians and scientists [**CH13**, **CT06**, **Gar09**, **VK04**].