Contents

1. Basic concepts of probability

1.1. Definitions.

- A **sample space** Ω is the set of all possible outcomes.
- An **event** A is a subset of Ω.
- A **σ-algebra** $\mathcal{B}$ is a subset of the set of all subsets of Ω satisfying the following axioms
  (1) $\emptyset \in \mathcal{B}$ and $\Omega \in \mathcal{B}$;
  (2) If $B \in \mathcal{B}$ then $B^c \in \mathcal{B}$ (B^c is the complement of B in Ω, i.e., $B^c \equiv \Omega \backslash B$).

(3) If $\mathcal{A} = \{A_1, \ldots, A_n, \ldots\}$ is a finite or countable collection in $\mathcal{B}$ then

$$\bigcup_i A_i \subset \mathcal{B}.$$

Corollary: If $\mathcal{A} = \{A_1, \ldots, A_n, \ldots\}$ is a finite or countable collection in $\mathcal{B}$ then

$$\bigcap_i A_i \subset \mathcal{B}.$$

Indeed,

$$\bigcap_i A_i = \left(\bigcup_i A_i^c\right)^c.$$

Example 1 Suppose you a tossing a die. For a single throw, the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. If you are interested in particular number on the top, the natural choice of the σ-algebra is the set of all subsets of Ω. Then $|\mathcal{B}| = 2^6 = 64$. If you are interested only in where the outcome is odd or even, then a reasonable choice of σ-algebra is

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

If you are interested only whether there is an outcome or not, you can choose the coarsest σ-algebra

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

- A **probability measure** P is a function $P : \mathcal{B} \to [0, +\infty]$ such that
  (1) $P(\Omega) = 1$;
  (2) $0 \leq P(A) \leq 1$ for all $A \in \mathcal{B}$.
  (3) **Countable additivity:** If $\mathcal{A} = \{A_1, \ldots, A_n, \ldots\}$ is a finite or countable collection in $\mathcal{B}$ such that $A_i \cap A_j = \emptyset$ for any i, j, then

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i).$$

  Corollary: $P(\emptyset) = 0$. Indeed,

$$1 = P(\Omega) = P(\Omega \cup \emptyset) = P(\Omega) + P(\emptyset) = 1 + P(\emptyset).$$

  Hence, $P(\emptyset) = 0$.
- A **probability space** is the triple $(\Omega, \mathcal{B}, P)$.
- A **random variable** η is a $\mathcal{B}$-measurable function $\eta : \Omega \to \mathbb{R}$.
  A function if called $\mathcal{B}$-measurable if the preimage of any measurable subset of $\mathbb{R}$ is in $\mathcal{B}$. It is proven in analysis that it is suffices to check that

$$\{\omega \in \Omega \mid \eta(\omega) \leq x\} \in \mathcal{B} \text{ for any } x \in \mathbb{R}.$$

- A **probability distribution function** of a random variable η is defined by

$$F_\eta(x) = P(\{\omega \in \Omega \mid \eta(\omega) \leq x\}) = P(\eta \leq x).$$

Theorem 1. *If F_η is a probability distribution function then*
(1) *F is nondecreasing, i.e., $x < y \implies F(x) \le F(y)$.*
(2) *$lim_{x\to-\infty}F(x) = 0, \quad lim_{x\to\infty}F(x) = 1$.*
(3) *$F(x)$ is continuous from the right for every $x \in \mathbb{R}$, i.e.,*

$$\lim_{y\to x+0} F(y) = F(x).$$

Example 2 Suppose you a tossing a die. Consider the probability space

$$(\Omega = \{1,2,3,4,5,6\}, \mathcal{B} = 2^\Omega, P(\omega) = \tfrac{1}{6}), \tag{1}$$

where 2^Ω is the set of all subsets of Ω, and $\omega \in \Omega = \{1,2,3,4,5,6\}$. Consider the random variable $\eta(\omega) = \omega$. The probability distribution function is given by

$$F_\eta(x) = \begin{cases} 0, & x < 1, \\ j/6, & j \le x < j+1,\ j = 1,2,3,4,5 \\ 1, & x \ge 6. \end{cases}$$

- Suppose $F'_\eta(x)$ exists. Then $f_\eta(x) \equiv F'_\eta(x)$ is called the **probability density function (pdf)** of the random variable η, and

$$P(x < \eta \le x + dx) = F_\eta(x + dx) - F_\eta(x) = f_\eta(x)dx + o(dx).$$

Example 3 Gaussian density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}},$$

where m and σ are constants. m is the mean, while σ is the standard deviation.

Example 4

$$f(x) = \begin{cases} e^{-x}, & x \ge 0, \\ 0, & x < 0. \end{cases}.$$

1.2. **Expected values and moments.**

Definition 1. *Let $(\Omega, \mathcal{B}, P)$ be a probability space, and η be a random variable. Then the expected value, or mean, of the random variable η is defined as*

$$E[\eta] = \int_\Omega \eta(\omega) dP. \tag{2}$$

If Ω is a discrete set,

$$E[\eta] = \sum_i \eta(\omega_i) P(\omega_i).$$

Example 5 Suppose you a tossing a die. Consider the probability space (1) and the random variable $\eta(\omega) = \omega$, $\omega = 1, 2, 3, 4, 5, 6$. The expected value of η is

$$E[\eta] = \sum_{j=1}^{6} j \tfrac{1}{6} = 3.5$$

If a derivative of the probability distribution function F_η exists, then

$$E[\eta] = \int_{-\infty}^{\infty} x f(x) dx.$$

The integral in Eq. (2) can be rewritten using the probability distribution function $F_\eta(x)$ which we denote by $F(x)$ for brevity:

$$E[\eta] = \int_{\mathbb{R}} x P(x < \eta \leq x + dx) = \int_{-\infty}^{\infty} x dF(x).$$

If g is a continuous function defined on the range of the random variable η (on $\eta(\Omega)$), then the expected value of this function is

$$E[g(\eta)] = \int_{-\infty}^{\infty} g(x) dF(x).$$

Moments: Let us take $g(x) = x^n$.

$$E[\eta^n] = \int_{-\infty}^{\infty} x^n dF(x).$$

Central moments: Let us take $g(x) = (x - E[\eta])^n$.

$$E[(\eta - E[\eta])^n] = \int_{-\infty}^{\infty} (x - E[\eta])^n dF(x).$$

Variance = 2nd central moment:

$$\mathrm{Var}(\eta) = E[(\eta - E[\eta])^2) = \int_{-\infty}^{\infty} (x - E[\eta])^2 dF(x).$$

Example 6 Suppose you a tossing a die. Consider the probability space (1) and the random variable $\eta(\omega) = \omega$, $\omega = 1, 2, 3, 4, 5, 6$. The variance of η is

$$\mathrm{Var}(\eta) = \frac{1}{6} \sum_{j=1}^{6} (j - 3.5)^2 = \frac{35}{12} = 2.91(6).$$

The standard deviation:

$$\sigma(\eta) = \sqrt{\mathrm{Var}(\eta)}.$$

1.3. **Independence, joint distributions, covariance.**

- Two events $A, B \in \mathcal{B}$ are **independent** if

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

- Two random variables η_1 and η_2 are independent if the events

$$\{\omega \in \Omega \mid \eta_1(\omega) \le x\} \text{ and } \{\omega \in \Omega \mid \eta_2(\omega) \le y\} \tag{3}$$

are independent for all $x, y \in \mathbb{R}$.

Example 7 Suppose you are tossing a die twice. Consider the probability space

$$(\Omega = \{1,2,3,4,5,6\}^2, \mathcal{B} = 2^{\Omega^2}, P(\{\omega_1, \omega_2\}) = 1/36), \quad 1 \le \omega_1, \omega_2 \le 6. \tag{4}$$

Let η_1 and η_2 be random variables equal to the outcomes of the first and the second throws respectively. These random variables are independent. Now consider the random variables $\eta(\omega_1, \omega_2) = \omega_1$ and $\xi(\omega_1, \omega_2) = \omega_1 + \omega_2$ (see Table 1, left). We can show that η and ξ are dependent by taking e.g., $x = 1$ and $y = 2$ in Eq. (3):

$$P(\eta \le 1 \ \& \ \xi \le 2) = \tfrac{1}{36} \ne P(\eta \le 1)P(\xi \le 2) = \tfrac{1}{6} \cdot \tfrac{1}{36} = \tfrac{1}{216}.$$

Finally, we consider the random variables $\xi(\omega_1, \omega_2) = \omega_1 + \omega_2$ and $\beta(\omega_1, \omega_2) = \omega_1 - \omega_2$ (see Table 1, right). We can show that they are dependent by taking e.g., $x = 2$ and $y = -1$ in Eq. (3):

$$P(\xi \le 2 \ \& \ \beta \le -1) = 0 \ne P(\xi \le 2)P(\beta \le -1) = \tfrac{1}{36} \cdot \tfrac{15}{36} = \tfrac{5}{432}.$$

TABLE 1. Two throws of a die. Values of the random variables $\xi(\omega_1, \omega_2) = \omega_1 + \omega_2$ (left) and $\beta(\omega_1, \omega_2) = \omega_1 - \omega_2$ (right).

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

	1	2	3	4	5	6
1	0	1	2	3	4	5
2	-1	0	1	2	3	4
3	-2	-1	0	1	2	3
4	-3	-2	-1	0	1	2
5	-4	-3	-2	-1	0	1
6	-5	-4	-3	-2	-1	0

- The joint distribution function of two random variables η_1 and η_2 is given by

$$F_{\eta_1\eta_2}(x, y) = P(\{\omega \in \Omega \mid \eta_1(\omega) \le x, \ \eta_2(\omega) \le y\}) = P(\eta_1(\omega) \le x, \ \eta_2(\omega) \le y).$$

- If the second mixed derivative of $F_{\eta_1\eta_2}$ exists, it is called the **joint probability density** of η_1 and η_2 and denoted by

$$f_{\eta_1\eta_2} := \frac{\partial F_{\eta_1\eta_2}(x,y)}{\partial x \partial y}.$$

In this case,

$$F\eta_1,\eta_2(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{\eta_1\eta_2}(x,y)dxdy.$$

Exercise Show that two random variables are independent if and only if

$$F_{\eta_1\eta_2}(x,y) = F_{\eta_1}(x)F_{\eta_2}(y).$$

Furthermore, if the joint pdf $f_{\eta_1\eta_2}(x,y)$ exists, then

$$f_{\eta_1\eta_2}(x,y) = f_{\eta_1}(x)f_{\eta_2}(y).$$

- Given the joint pdf $f_{\eta_1\eta_2}$, one can obtain $f_{\eta_1}(x)$ by

$$f_{\eta_1}(x) = \int_{-\infty}^{\infty} f_{\eta_1\eta_2}(x,y)dx.$$

In this equation, f_{η_1} is called a **marginal** of $f_{\eta_1\eta_2}$, and the variable η_2 is said to have been integrated out.

- **Properties of expected value and variance** It follows from the definition, that the expected value is a linear functional:

$$E[a\eta_1 + b\eta_2] = aE[\eta_1] + bE[\eta_2]. \tag{5}$$

-

$$\mathrm{Var}(a\eta) = a^2\mathrm{Var}(\eta). \tag{6}$$

- If $\eta_1$and η_2 are independent, then

$$\mathrm{Var}(\eta_1 + \eta_2) = \mathrm{Var}(\eta_1) + \mathrm{Var}(\eta_2). \tag{7}$$

Example 8 Suppose you are tossing a die twice. Consider the probability space and random variables introduced in Example 7. Then

$$E[\xi] = E[\eta_1 + \eta_2] = E[\eta_1] + E[\eta_2] = 7.$$

$$E[\beta] = E[\eta_1 - \eta_2] = E[\eta_1] + E[-\eta_2] = 0.$$

$$Var[\xi] = Var[\eta_1 + \eta_2] = Var[\eta_1] + Var[\eta_2] = \tfrac{35}{6} = 5.8(3).$$

$$Var[\beta] = Var[\eta_1 - \eta_2] = Var[\eta_1] + Var[-\eta_2] = Var[\eta_1] + Var[\eta_2] = \tfrac{35}{6} = 5.8(3).$$

Example 9 Consider the Bernoulli random variable

$$\eta = \begin{cases} 1, & P(1) = p, \\ 0, & P(0) = 1 - p. \end{cases} \tag{8}$$

Its expected value and variance are

$$E[\eta] = 1 \cdot p + 0 \cdot (1 - p) = p,$$

$$\mathrm{Var}(\eta) = (1 - p)^2 \cdot p + (0 - p)^2 \cdot (1 - p) = p(1 - p).$$

Now consider the sum of n independent copies of η:

$$\xi := \sum_{i=1}^{n} \eta_i.$$

Using Eq. (5) we calculate $E[\xi]$:

$$E[\xi] = \sum_{\imath=1}^{n} E[\eta_i] = np.$$

Since η_i, $1 \leq i \leq n$, are independent, we can calculate $\mathrm{Var}(\xi)$ using Eq. (7):

$$\mathrm{Var}(\xi) = \sum_{i=1}^{n} \mathrm{Var}(\eta_i) = np(1 - p).$$

Finally, consider the average of n independent copies of η:

$$\zeta := \frac{1}{n} \sum_{i=1}^{n} \eta \equiv \frac{\xi}{n}.$$

Using Eqs. (5) and (6), we find

$$E[\zeta] = p,$$

$$\mathrm{Var}(\zeta) = \mathrm{Var}\left(\frac{\xi}{n}\right) = \frac{1}{n^2}\mathrm{Var}(\xi) = \frac{p(1 - p)}{n}.$$

- The **covariance** of two random variables η_1 and η_2 is defined by

$$\mathrm{Cov}(\eta_1, \eta_2) = E[(\eta_1 - E[\eta_1])(\eta_2 - E[\eta_2])].$$

Remark If η_1 and η_2 are independent, then $\mathrm{Cov}(\eta_1, \eta_2) = 0$. If $\mathrm{Cov}(\eta_1, \eta_2) = 0$ then η_1 and η_2 are uncorrelated. Note that uncorrelated random variables are not necessarily independent.

Example 10 Suppose you are tossing a die twice. Consider the probability space and random variables introduced in Example 7. As we have established in Example 7, ξ and β are dependent. However, they are uncorrelated. Indeed,

$$\mathrm{Cov}(\xi,\beta) = \sum_{1\le\omega_1\le 6,\ 1\le\omega_2\le 6} (\omega_1+\omega_2-7)(\omega_1-\omega_2)P(\{\omega_1,\omega_2\})$$

$$= \frac{1}{36}\left(\sum_{\omega_1<\omega_2}(\omega_1+\omega_2-7)(\omega_1-\omega_2) + \sum_{\omega_1>\omega_2}(\omega_1+\omega_2-7)(\omega_1-\omega_2)\right) = 0.$$

Example 11 A vector-valued random variable $\eta = [\eta_1,\dots,\eta_n]$ is jointly Gaussian if

$$P(x_1<\eta_1\le x_1+dx_1,\dots,x_n<\eta_n\le x_n+dx_n) = \frac{1}{Z}e^{-\frac{1}{2}(x-m)^TA^{-1}(x-m)}dx + o(dx),$$

where $x=[x_1,\dots,x_n]^T$, $m=[m_1,\dots,m_n]^T$, $dx = dx_1\dots dx_n$, and A is a symmetric positive definite matrix. The normalization constant Z is given by

$$Z = (2\pi)^{n/2}|A|^{1/2}, \text{ where } |A| = \det A.$$

In the case of jointly Gaussian random variables, the covariance matrix C whose entries are

$$C_{ij} = E[(\eta_i - E[\eta_i])(\eta_j - E[\eta_j])]$$

is equal to A. Two jointly Gaussian random variables are independent if and only if they are uncorrelated.

1.4. **Chebyshev inequality.**

Theorem 2. *Let η be a random variable. Suppose $g(x)$ is a nonnegative, nondecreasing function (i.e., $g(x)\ge 0$, $g(a)\le g(b)$ whenever $a<b$). Then for any $a\in\mathbb{R}$*

$$P(\eta\ge a) \le \frac{E[g(\eta)]}{g(a)}. \tag{9}$$

Proof.

$$E[g(\eta)] = \int_{-\infty}^{\infty} g(x)dF(x)$$

$$\ge \int_a^{\infty} g(x)dF(x) \ge g(a)\int_a^{\infty} dF(x) = g(a)P(\eta\ge a).$$

□

Given a random variable η we define a random variable

$$\xi := |\eta - E[\eta]|.$$

Define

$$g(x) = \begin{cases} x^2, & x \geq 0, \\ 0, & x < 0. \end{cases}.$$

Plugging this into Eq. (9) we obtain

$$P(|\eta - E[\eta]| \geq a) \leq \frac{\mathrm{Var}(\eta)}{a^2}.$$

Example 12 Suppose you are tossing a die twice. Consider the probability space and random variables introduced in Example 7. We will compare the exact probabilities with their Chebyshev estimates.

$P(|\xi - 7| \geq 1) = P(\xi \neq 7) = 1 - \frac{6}{36} = \frac{5}{6} = 0.8(3)$, $\quad \frac{\mathrm{Var}(\xi)}{1} = \frac{35}{6} = 5.8(3)$;

$P(|\xi - 7| \geq 2) = P(\xi \leq 5 \text{ or } \xi \geq 9) = \frac{20}{36} = \frac{5}{9} = 0.(5)$, $\quad \frac{\mathrm{Var}(\xi)}{4} = \frac{35}{24} = 1.458(3)$;

$P(|\xi - 7| \geq 3) = P(\xi \leq 4 \text{ or } \xi \geq 10) = \frac{12}{36} = \frac{1}{3} = 0.(3)$, $\quad \frac{\mathrm{Var}(\xi)}{9} = \frac{35}{54} = 0.6(481)$;

$P(|\xi - 7| \geq 4) = P(\xi \in \{2, 3, 11, 12\}) = \frac{6}{36} = \frac{1}{6} = 0.1(6)$, $\quad \frac{\mathrm{Var}(\xi)}{16} = \frac{35}{96} = 0.36458(3)$;

$P(|\xi - 7| \geq 5) = P(\xi \in \{2, 12\}) = \frac{2}{36} = \frac{1}{18} = 0.0(5)$, $\quad \frac{\mathrm{Var}(\xi)}{25} = \frac{35}{150} = 0.2(3)$;

Choosing $a = k\sigma$ we get

$$P(|\eta - E[\eta]| \geq k\sigma) \leq \frac{1}{k^2}.$$

This means that for *any* random variable η defined on *any* probability space we have that the probability that η deviates from its expected value by at least k standard deviations does not exceed $1/k^2$.

Note that the power of Chebyshev's inequality is its weakness at the same time. It is universal as it is valid for *any* random variable defined on *any* probability space. At the same time, due to its generality, it typically gives loose bounds. These bounds cannot be improved in principle, because they are exact for the random variable

$$\eta = \begin{cases} 1, & P = \frac{1}{2k^2}, \\ 0, & P = 1 - \frac{1}{k^2}, \\ -1, & P = \frac{1}{2k^2}. \end{cases}.$$

It is easy to check that $E[\eta] = 0$, $\mathrm{Var}(\eta) = \frac{1}{k^2}$, and

$$P(|\eta| \geq 1) = \frac{1}{k^2}.$$

Chebyshev's inequality gives the precise upper bound:

$$P(|\eta| \geq 1) \leq \frac{\mathrm{Var}(\eta)}{1^2} = \frac{1}{k^2}.$$

1.5. **Types of convergence of random variables.** Suppose we have a sequence of random variables $\{\eta_1, \eta_2, \ldots\}$. In probability theory, there exists several different notions of convergence of a sequence of random variables $\{\eta_1, \eta_2, \ldots\}$ to some limit random variable η.

- $\{\eta_1, \eta_2, \ldots\}$ **converges in distribution** or **converges weakly**, or **converges in law** to η if

$$\lim_{n\to\infty} F_n(x) = F(x) \text{ for every } x \text{ where } F(x) \text{ is continuous,} \tag{10}$$

where F_n and F are the probability distribution functions of η_n and η respectively.

Remark Convergence of pdfs $f_n(x)$ implies convergence of $F_n(x)$. The converse is not true in general. For example, consider $F_n(x) = x - \frac{1}{2\pi n}\sin(2\pi n x)$, $x \in (0,1)$. The corresponding pdf is $f_n(x) = 1 - \cos(2\pi n x)$, $x \in (0,1)$. F_ns converge to x, i.e., to the uniform distribution, while f_ns do not converge at all.

Remark In the discrete case, the convergence of probability distributions $f(k) := P(\eta = k)$ implies the convergence of the probability distribution functions.

Example 13 Consider the sum of n independent copies of the Bernoulli random variable as in Example 9:

$$\xi = \sum_{i=1}^{n} \eta_i, \text{ where } \eta_i = \begin{cases} 1, & P(1) = p, \\ 0, & P(0) = 1 - p. \end{cases} \tag{11}$$

Its probability distribution is the binomial distribution given by

$$f(k; n, p) \equiv P(\xi = k) = \begin{pmatrix} n \\ k \end{pmatrix} p^k (1-p)^{n-k}, \tag{12}$$

where $\begin{pmatrix} n \\ k \end{pmatrix}$ is the number of k-combinations of the set of n elements:

$$\begin{pmatrix} n \\ k \end{pmatrix} = \frac{n!}{k!(n-k)!}.$$

Now we let $n \to \infty$ and $p \to 0$ in such a manner that the product np (i.e., the expected value of ξ) remains constant. We introduce the parameter

$$\lambda := np.$$

Consider the sequence of random variables ξ_n where ξ_n is the sum of n independent copies of Bernoulli random variable with $p = \lambda/n$, i.e,

$$\xi_n = \sum_{\imath=1}^{n} \eta_i^{(n)}, \text{ where } \eta_i^{(n)} = \begin{cases} 1, & P(1) = \lambda/n, \\ 0, & P(0) = 1 - \lambda/n. \end{cases} \tag{13}$$

Plugging in $p = \lambda/n$ in the results of Example 9 we find the expected value and the variance:

$$E[\xi_n] = n\frac{\lambda}{n} = \lambda.$$

$$\mathrm{Var}(\xi_n) = n\frac{\lambda}{n}\left(1 - \frac{\lambda}{n}\right) = \lambda\left(1 - \frac{\lambda}{n}\right).$$

We will show that the sequence ξ_n converges to the Poisson random variable with parameter λ in distribution. Consider the limit

$$\lim_{n\to\infty} f(k; n, \lambda/n) = \lim_{n\to\infty} \frac{n(n-1)\dots(n-k+1)}{k!}\frac{\lambda^k}{n^k}\left(1 - \frac{\lambda}{n}\right)^{n-k} =$$

$$\frac{\lambda^k}{k!}\lim_{n\to\infty}\frac{n(n-1)\dots(n-k+1)}{n^k}\lim_{n\to\infty}\left(1 - \frac{\lambda}{n}\right)^{n}\lim_{n\to\infty}\left(1 - \frac{\lambda}{n}\right)^{-k}$$

The first limit in the equation above is 1 as $n(n-1)\dots(n-k+1) = n^k + O(n^{k-1})$. The second limit can be calculated using the well-known fact that

$$\lim_{n\to\infty}\left(1 + \frac{1}{n}\right)^n = e.$$

Hence

$$\lim_{n\to\infty}\left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}.$$

The third limit is 1. Therefore,

$$\lim_{n\to\infty}\frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k\left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!}e^{-\lambda},$$

which is the Poisson distribution with parameter λ.

- $\{\eta_1, \eta_2, \dots\}$ **converges in probability** to η if for any $\epsilon > 0$

$$\lim_{n\to\infty} P(|\eta_n - \eta| \geq \epsilon) = 0 \tag{14}$$

Remark Convergence in probability implies convergence in distribution.

Proof. We will prove this fact for the case of scalar random variables. We have $\lim_{n\to\infty} P(|\eta_n - \eta| \geq \epsilon) = 0$, we need to prove $\lim_{n\to\infty} P(\eta_n \leq x) = P(\eta \leq x)$ for every x where F_η is continuous. First we show an auxiliary fact that for any two random variables ξ and ζ, $x \in \mathbb{R}$ and $\epsilon > 0$

$$P(\xi \leq a) \leq P(\zeta \leq a + \epsilon) + P(|\xi - \zeta| > \epsilon). \tag{15}$$

Indeed,

$$\begin{aligned} P(\xi \le a) &= P(\xi \le x \ \&\ \zeta \le a+\epsilon) + P(\xi \le a \ \&\ \zeta > a+\epsilon) \\ &\le P(\zeta \le a+\epsilon) + P(\xi - \zeta \le a - \xi \ \&\ a - \zeta < -\epsilon) \\ &\le P(\zeta \le a+\epsilon) + P(\zeta - \xi < -\epsilon) \\ &\le P(\zeta \le a+\epsilon) + P(\zeta - \xi < -\epsilon) + P(\zeta - \xi > \epsilon) \\ &= P(\zeta \le a+\epsilon) + P(|\zeta - \xi| < \epsilon). \end{aligned}$$

Applying Eq. (15) to $\xi = \eta_n$ and $\zeta = \eta$ with $a = x$ and $a = x - \epsilon$, we get

$$\begin{aligned} P(\eta_n \le x) &\le P(\eta \le x+\epsilon) + P(|\eta_n - \eta| > \epsilon) \\ P(\eta \le x-\epsilon) &\le P(\eta_n \le x) + P(|\eta_n - \eta| > \epsilon). \end{aligned}$$

$$P(\eta \le x-\epsilon) - P(|\eta_n - \eta| > \epsilon) \le P(\eta_n \le x) \le P(\eta \le x+\epsilon) + P(|\eta_n - \eta| > \epsilon).$$

Taking the limit $n \to \infty$ and taking into account that $\lim_{i\to\infty} P(|\eta_n - \eta| \ge \epsilon) = 0$, we get

$$F_\eta(x-\epsilon) \le \lim_{n\to\infty} F_{\eta_n}(x) \le F_\eta(x+\epsilon).$$

If x is a point of continuity of F_η,

$$\lim_{\epsilon\to 0} F_\eta(x-\epsilon) = \lim_{\epsilon\to 0} F_\eta(x+\epsilon) = F_\eta(x).$$

Therefore, taking the limit $\epsilon \to 0$ we obtain the weak convergence:

$$\lim_{n\to\infty} F_{\eta_n}(x) = F_\eta(x)$$

for any x where $F_\eta(x)$ is continuous. □

Remark In the opposite direction, convergence in distribution to a *constant* random variable implies convergence in probability.

- $\{\eta_1, \eta_2, \ldots\}$ **converges almost surely** or **almost everywhere** or **with probability 1** or **strongly** to η if

$$P\left(\lim_{n\to\infty} \eta_n = \eta\right) = 1. \tag{16}$$

Remark Convergence almost surely implies convergence in probability (by Fatou's lemma) and in distribution.

- To summarize,

$$\boxed{\eta_i \to \eta \text{ almost surely}} \Rightarrow \boxed{\eta_i \to \eta \text{ in probability}} \Rightarrow \boxed{\eta_i \to \eta \text{ in distribution}} \tag{17}$$

1.6. **Laws of Large Numbers and the Central Limit Theorem.**

- Let $\{\eta_1, \eta_2, \ldots\}$ be a sequence of random variables with finite expected values $\{m_1 = E[\eta_1], m_2 = E[\eta_2], \ldots\}$. Define

$$\xi_n = \frac{1}{n}\sum_{i=1}^{n} \eta_i, \quad \bar{\xi}_n = \frac{1}{n}\sum_{i=1}^{n} m_i.$$

Definition 2. (1) *The sequence of random variables η_n satisfies the Law of Large Numbers if $\xi_n - \bar{\xi}_n$ converges to zero in probability, i.e., for any $\epsilon > 0$*

$$\lim_{n\to\infty} P(|\xi_n - \bar{\xi}_n| > \epsilon) = 0.$$

(2) *The sequence of random variables η_n satisfies the Strong Law of Large Numbers if $\xi_n - \bar{\xi}_n$ converges to zero almost surely, i.e., for almost all $\omega \in \Omega$*

$$\lim_{n\to\infty} \xi_n - \bar{\xi}_n = 0.$$

- If the random variables η_n are independent and if $\mathrm{Var}(\eta_i) \le V < \infty$, then the Law of Large Numbers holds by the Chebyshev Inequality (9):

$$\begin{aligned} P(|\xi_n - \bar{\xi}_n| > \epsilon) &= P\left(\left|\sum_{i=1}^{n} \eta_i - \sum_{i=1}^{n} m_i\right| > n\epsilon\right) \\ &\le \frac{\mathrm{Var}(\eta_1 + \ldots + \eta_n)}{\epsilon^2 n^2} \le \frac{V}{\epsilon^2 n} \to 0 \text{ as } n \to \infty. \end{aligned}$$

- **Theorem 3.** *(Khinchin) A sequence of independent identically distributed random variables $\{\eta_i\}$ with $\mathbb{E}[\eta_i] = m$ and $\mathbb{E}[|\eta_i|] < \infty$ satisfies the Law of Large Numbers.*

- **Theorem 4.** *(Kolmogorov) A sequence of independent identically distributed random variables with finite expected value and variance satisfies the Strong Law of Large Numbers.*

- **Theorem 5. (The central limit theorem)** *Let $\{\eta_1, \eta_2, \ldots\}$ be a sequence of independent identically distributed (i.i.d.) random variables with $m = E[\eta_i]$ and $0 < \sigma^2 = \mathrm{Var}(\eta_i) < \infty$, then the distributions*

$$\frac{(\sum_{i=1}^{n} \eta_i) - nm}{\sigma\sqrt{n}} \longrightarrow N(0,1) \text{ in distribution,} \tag{18}$$

i.e., converges weakly to the standard normal distribution $N(0,1)$ (i.e., the Gaussian distribution with mean 0 and variance 1) as $n \to \infty$.

A proof via Fourier transform can be found in [1]. Another proof making use of characteristic functions can be found in [2].

Remark Eq. (18) can be recasted as

$$\frac{1}{n}\sum_{i=1}^{n}\eta_i \longrightarrow N\left(m, \frac{\sigma^2}{n}\right) \text{ in distribution,} \tag{19}$$

i.e., the average of the first n i.i.d. random variables η_i converges in distribution to the Gaussian random variable with mean $m = E[\eta_i]$ and variance σ^2/n.

1.7. Conditional probability and conditional expectation.

- The conditional probability of an event B given that the event A has happened is given by

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

Note that if A and B are independent, then $P(A \cap B) = P(A)P(B)$ and hence

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

Example 14 Suppose you are tossing a die twice. Consider the probability space (4). Let A be the event that the outcome of the first throw is even, and B be the event that the sum of the outcomes is greater than 10. Then (see Table 1)

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{4/36}{1/2} = \frac{2}{9}.$$

Note that $P(B) = 1/6 < P(B|A)$. Hence the events A and B are dependent.

If the event A is fixed, then $P(B|A)$ defines a probability measure on $(\Omega, \mathcal{B})$.

- If η is a random variable on Ω, then conditional expectation of η given the event A is

$$E[\eta|A] = \int_\Omega \eta(\omega) P(d\omega|A).$$

Example 15 Suppose you are tossing a die twice. Consider the probability space (4). Let A be the event that the outcome of the first throw is even, and η be the random variable whose value is the sum of outcomes, i.e., $\eta(\{\omega_1, \omega_2\}) = \omega_1 + \omega_2$. Then

$$E[\eta|A] = \sum_{\omega_1=1}^{6}\sum_{\omega_2=1}^{6}(\omega_1+\omega_2)\frac{P(\omega_1 \ \& \ \omega_2 \ \& \ \omega_1 \in \{2,4,6\})}{1/2}$$

$$= \sum_{\omega_1 \in \{2,4,6\}}\sum_{\omega_2=1}^{6}(\omega_1+\omega_2)\frac{1/36}{1/2} = \frac{135}{18} = 7.5.$$

Note that $E[\eta] = 7 \neq E[\eta|A]$.

1.8. **Applications to statistical mechanics.** In this section, we consider some application of the concepts we have discussed to statistical mechanics.

Exercise Consider a particle in 1D in contact with a heat bath whose states follow the canonical distribution:

$$\mu(x,p) = \frac{1}{Z} e^{-\beta H(x,p)}, \quad \text{where} \quad Z = \int_{\mathbb{R}^2} e^{-\beta H(x,p)} dx dp, \tag{20}$$

where $H(x,p) = V(x) + \frac{p^2}{2}$ is its energy and $\beta = (k_B T)^{-1}$ (k_B is Boltzmann's constant). Show that the mean kinetic energy equals to $k_B T/2$, i.e., calculate the expected value of

$$E\left[\frac{p^2}{2}\right] = \frac{1}{Z} \int_{\mathbb{R}^2} \frac{p^2}{2} e^{-\beta(V(x)+p^2/2)} dx dp.$$

Use your result to show that for a system consisting of n particles with unit mass each of which is moving in 3D, the mean kinetic energy is $(3/2) n k_B T$.

2. Sampling and Monte-Carlo Integration

Reading: [1] (Chapter 3), [4] (Chapter 9). *Monte-Carlo methods* are those where one evaluates something nonrandom using pseudorandom numbers. More precisely, one evaluates a nonrandom quantity as the expected values of a random variable. On the contrary, *simulations* produce random variables with a certain distribution with the purpose of just looking at them. Typically, the error in Monte-Carlo methods decays as $n^{-1/2}$ where n is the number of samples which is worse that the error decay rate in most of deterministic methods (it is usually at least as good as n^{-1}). So, why bother? The reason is that in some important situations deterministic methods simply cannot be used due to such things as the "curse of dimensionality" or largeness of the problem. In some of these cases, Monte-Carlo methods can be efficient. For example, to find the mean magnetization in a 3D Ising model with n sites, one need to average the value of the magnetization over 2^n different spin configurations. If we are considering a 3D $10 \times 10 \times 10$ grid, then $n = 1000$, and $2^{1000} \sim 10^{301}$, a huge number, that makes the deterministic calculation infeasible. On the contrary, a Monte-Carlo calculation gives an accurate enough estimate in a reasonable time.

2.1. **Pseudorandom numbers.** Pseudorandom numbers are generated by pseudorandom number generators. A pseudorandom number generator produces a deterministic sequence of numbers starting from a seed state that can be specified by the user. Good pseudorandom number generators produce sequences that cannot be distinguished from random numbers by simple tests. In C, the operator `rand()` produces a uniformly distributed pseudorandom number in the interval [0 . . . RAND_MAX], where RAND_MAX is a constant defined in the library `stdlib.h`. It is platform-dependent. This constant might not be large enough for ambitions calculations such as sampling of random trajectories. `rand()` is claimed to be bad in [4]. The author of this notes observed the appearance of periodicity in a long but not-extremely-long sequence. Instead, one can use the C operator `random()` which is claimed to be good enough for most Monte-Carlo calculations in [4].

2.2. **Sampling random variables with given distribution.** Most programming languages have tools for generating a uniformly distributed random variable ξ on the interval $[0,1]$.

Suppose we need to sample a random variable with a pdf $f(x)$. Assume that we can integrate $f(x)$ analytically, i.e., have an analytic expression for the probability distribution function $F(x)$. We observe that

$$\int_0^\eta f(x)dx = F(\eta) = \xi. \quad \text{Hence} \quad \eta = F^{-1}(\xi),$$

where $F^{-1}(\xi)$ is the inverse function of F. It exists if $F(x)$ is strictly increasing. If ξ is uniformly distributed on $[0,1]$, i.e., its probability distribution function is

$$F_\xi(x) = P(\xi \le x) = \begin{cases} 1, & x \ge 1, \\ x, & 0 \le x < 1, \\ 0, & x < 0, \end{cases}$$

then the probability distribution of η is $F(x)$. Indeed,

$$P(\eta \le x) = P(F^{-1}(\xi) \le x) = P(\xi \le F(x)) = F(x).$$

Example 16 Suppose we need to generate an exponentially distributed random variable η with pdf $f(x) = ae^{-ax}$ where $a > 0$ is a constant. The probability distribution of ξ is given by

$$F_\eta(x) = P(\eta \le x) = \int_0^x ae^{-ay}dy = 1 - e^{-ax}.$$

Let ξ be a random variable uniformly distributed on $[0,1]$. Then η can be generated from ξ by

$$\eta = F_\eta^{-1}(\xi) = -\tfrac{1}{a}\log(1-\xi).$$

Observing that $1-\xi$ is also a random variable uniformly distributed on $[0,1]$, we can choose to generate η by

$$\eta = F_\eta^{-1}(\xi) = -\tfrac{1}{a}\log(\xi).$$

2.3. **The Box-Muller algorithm.** Suppose we need to generate a Gaussian random variable η with mean 0 and variance σ^2 while we have a built-in function for generating a random variable ξ uniformly distributed on $[0,1]$. Unfortunately, the pdf of η

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

is not analytically integrable. Therefore, we cannot use the method proposed above directly. However, we can generate pairs of independent jointly Gaussian random variables (η_1, η_2) given a pair of independent uniformly distributed on $[0,1]$ random variables (ξ_1, ξ_2). Since the pdf of jointly Gaussian independent random variables is radially symmetric, we can generate the polar radius r and polar angle θ of the pair of (η_1, η_2) and then obtain their

Cartesian coordinates from the polar ones. Since the polar angle of (η_1, η_2) is uniformly distributed on $[0, 2\pi]$, we set

$$\theta = 2\pi\xi_2. \tag{21}$$

To sample the polar radius, we calculate

$$\begin{aligned} F(a) =& P(r \le a) = \frac{1}{2\pi\sigma^2} \int_{\sqrt{x^2+y^2}\le a} e^{-\frac{x^2+y^2}{2\sigma^2}} dxdy \\ =& \frac{1}{2\pi\sigma^2} \int_0^{2\pi} d\theta \int_0^a e^{-\frac{r^2}{2\sigma^2}} r dr \\ =& \frac{1}{\sigma^2}\sigma^2 \int_0^{a^2/(2\sigma^2)} e^{-t} dt = 1 - e^{-a^2/(2\sigma^2)} = 1 - \xi_1. \end{aligned}$$

Here we used the fact that if ξ_1 is uniformly distributed on $[0, 1]$, then so is $1 - \xi_1$.

$$a = \sqrt{-2\sigma^2 \log \xi_1}. \tag{22}$$

Using Eqs. (21) and (22) we get the Box-Muller formulas for generating pairs of independent jointly Gaussian random variables with mean 0 and variance σ^2:

$$\begin{cases} \eta_1 = a\cos\theta = \sqrt{-2\sigma^2 \log \xi_1}\cos(2\pi\xi_2) \\ \eta_2 = a\sin\theta = \sqrt{-2\sigma^2 \log \xi_1}\sin(2\pi\xi_2) \end{cases} \tag{23}$$

2.4. **Monte-Carlo integration.** Suppose we need to calculate an integral of the form

$$I = \int_a^b g(x)f(x)dx, \quad \text{where}$$

$$f(x) \ge 0,\ x \in [a, b], \quad \text{and} \quad \int_a^b f(x)dx = 1.$$

Such integral can be interpreted as the expected value of the function g of the random variable η with the pdf $f(x)$, i.e.,

$$I = \int_a^b g(x)f(x)dx = \int_a^b g(x)f(x)dx = E[g(\eta)].$$

Suppose we are able to sample i. i. d. random variables η_i each of which has the pdf $f(x)$. According to the strong law of large numbers,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n g(\eta_i) = E[g(\eta)] \quad a.s.$$

The integral I is called the estimand, the random variable $g(\eta)$ is called the estimator, and the quantity

$$\frac{1}{n} \sum_{i=1}^n g(\eta_i) \tag{24}$$

is the estimate. This method of evaluating integrals is called the *Monte-Carlo integration*. According to the central limit theorem,

$$\frac{1}{n}\sum_{i=1}^{n} g(\eta_i) \;\longrightarrow\; N\left(E[g(\eta)], \frac{\mathrm{Var}(g(\eta))}{n}\right) \quad \text{in distribution.}$$

Therefore, the error of the estimate (24) is of the order of

$$err \sim \frac{\sqrt{\mathrm{Var}(g(\eta))}}{\sqrt{n}}. \tag{25}$$

Eq. (25) suggests two ways to reduce the error of the Monte-Carlo integration: (i) to increase the number of samples n, and (ii) to reduce the variance of $g(\eta)$. Note that increasing the number of samples is not very efficient approach, as the error decays as $n^{-1/2}$. A better idea is to try to reduce the variance of $g(\eta)$. One approach to the variance reduction is called the *importance sampling*.

2.5. **Importance sampling.** Suppose we need to calculate the integral

$$I = \int_a^b g(x)f(x)dx, \quad \text{where}$$

$$f(x) \geq 0,\ x \in [a,b], \quad \text{and} \quad \int_a^b f(x)dx = 1.$$

In order to reduce $\mathrm{Var}(g(\eta))$ we can try to find a function $h(x)$ with the following properties:

(1) The integral
$$I_1 = \int_a^b f(x)h(x)dx$$
is easy to evaluate;
(2) $h(x) \geq 0$;
(3) We can sample a random variable with the pdf
$$\frac{f(x)h(x)}{I_1} \quad \text{easily;}$$
(4) $g(x)/h(x)$ varies little.

Then we have

$$I = \int_a^b g(x)f(x)dx = \int_a^b \frac{g(x)}{h(x)}f(x)h(x)dx = I_1\int_a^b \frac{g(x)}{h(x)}\frac{f(x)h(x)}{I_1}dx$$
$$= I_1 E\left[\frac{g}{h}(\eta)\right] \sim \frac{I_1}{n}\sum_{i=1}^{n}\frac{g(\eta_i)}{h(\eta_i)},$$

where η has the pdf $f(x)h(x)/I_1$. See the example with

$$I = \int_0^1 \cos(x/5)e^{-5x}dx$$

in [1].

2.6. **Monte Carlo integration in higher dimensions.** Suppose we would like to evaluate the integral

$$I = \int_\Omega g(x)dx \tag{26}$$

where $\Omega \subset \mathbb{R}^n$. We proceed as we did in 1D. Let us generate a random variable η whose pdf $f_\eta(x)$ is nonzero in Ω and zero elsewhere and rewrite Eq. (26) as

$$I = \int_\Omega \frac{g(x)}{f_\eta(x)} f_\eta(x)dx \tag{27}$$

By the strong law of large numbers,

$$I = \int_\Omega \frac{g(x)}{f_\eta(x)} f_\eta(x)dx = E\left[\frac{g(x)}{f_\eta(x)}\right] \approx \frac{1}{N}\sum_{i=1}^{N} \frac{g(x_i)}{f_\eta(x_i)}, \tag{28}$$

where x_i, $1 \le i \le N$, are samples of the random variable η with pdf $f_\eta(x)$.

Suppose η is uniformly distributed in Ω. Then its pdf is given by

$$f_\eta(x) = \begin{cases} \frac{1}{|\Omega|}, & x \in \Omega \\ 0, & x \notin \Omega, \end{cases} \tag{29}$$

where $|\Omega|$ is the volume of Ω. In this case, Eq. (30) becomes:

$$I = \int_\Omega \frac{g(x)}{f_\eta(x)} f_\eta(x)dx \approx \frac{|\Omega|}{N}\sum_{i=1}^{N} g(x_i). \tag{30}$$

Similarly we proceed when we need to calculate an integral over a k-dimensional hypersurface S embedded into $\mathbb{R}^n$:

$$I = \int_S g(x)d\sigma, \tag{31}$$

where $d\sigma$ is a surface element. Let η be a random variable whose pdf is supported at the hyper surface S, i.e. $f_\eta(x) > 0$ if and only if $x \in S$. Then the integral is approximated by

$$I = \int_S g(x)d\sigma \approx \frac{1}{N}\sum_{i=1}^{N} \frac{g(x_i)}{f_\eta(x_i)}, \tag{32}$$

where x_i, $1 \le i \le N$ are samples of the random variable η. If η is uniformly distributed on the hypersurface S, then

$$I = \int_S g(x)d\sigma \approx \frac{|S|}{N}\sum_{i=1}^{N} g(x_i), \tag{33}$$

where $|S|$ is the measure (k-dimensional area) of S:

$$|S| = \int_S d\sigma. \tag{34}$$

Example 17 Consider the integral

$$I = \int_{S_{n-1}} g(x) d\sigma, \tag{35}$$

where S_{n-1} is the unit $n-1$-dimensional sphere (n-sphere) embedded into $\mathbb{R}^n$:

$$S_{n-1} = \{x = (x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_1^2 + \ldots + x_n^2 = 1\}.$$

Let us generate N samples of random variable η uniformly distributed on S_n. This can be done as follows. First we generate an array $N \times n$ of independent Gaussian random variables with mean 0 and variance 1. It is well-known that n independent Gaussian random variables with mean zero and variance 1 have the joint pdf

$$f_{\eta_1,\ldots,\eta_n}(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{x_1^2+\ldots+x_n^2}{2}} \equiv= \frac{1}{(2\pi)^{n/2}} e^{-\frac{r^2}{2}}, \tag{36}$$

where $r := \sqrt{x_1^2 + \ldots + x_n^2}$. Let us treat each row of our array as a sample of a vector random variable ξ with pdf given by Eq. (36). The distribution of ξ is spherically symmetric. Hence, we can obtain the desired random variable η uniformly distributed on the unit sphere by normalizing the radius of ξ:

$$\eta = \frac{\xi}{\sqrt{\xi_1^2 + \ldots + \xi_n^2}}. \tag{37}$$

In matlab, N samples of a random variable η uniformly distributed on the unit n-sphere can be generated by the following set of commands:

```
xi = randn(N, n);
aux = sqrt(sum(xi.^2, 2))*ones(1, n);
eta = xi./aux;
```

The surface area of the unit sphere S_{n-1} is given by

$$|S_{n-1}| = \frac{2\pi^{n/2}}{\Gamma(\frac{n}{2})}, \tag{38}$$

where

$$\Gamma(x) := \int_0^\infty t^{x-1} e^{-t} dt$$

is the Gamma-function. Thus, the integral (35) can be estimated as

```
S = 2*pi^(n/2)/gamma(n/2);
I = sum(g(eta))*S/N;
```

where n, N, and the function $y = g(x)$ must be provided. A table of exact integrals of some functions over unit hypersphere are found in [5].

- For $n = 4$ and $g(x) = x_1^2 x_2^2$, the exact integral (35) is

$$I = \int_{S_3} x_1^2 x_2^2 d\sigma = \frac{\pi^2}{12} = 0.8224670\ldots,$$

while its estimate using 10^6 samples is 0.8227420, and its error estimate is 10^{-3}.

- For $n = 10$ and $g(x) = x_1^2$, the exact integral (35) is

$$I = \int_{S_9} x_1^2 d\sigma = \frac{\pi^5}{120} = 2.550164\ldots,$$

while its estimate using 10^6 samples is 2.548990, and its error estimate is $3 \cdot 10^{-3}$.

3. Discrete time Markov chains

Think about the following problem,. Imagine a gambler who has \$1 initially. At each discrete moment of time $t = 0, 1, \ldots$, the gambler can play \$1 if he has it and win one more \$1 with probability p or lose it with probability $q = 1 - p$. If the gambler runs out of money, he is ruined and cannot play anymore. What is the probability that the gambler will be ruined?

The gambling process described in this problem exemplifies a discrete time Markov chain. In general, a discrete time Markov chain is defined as a sequence of random variables $(X_n)_{n\geq 0}$ taking a finite or countable set of values which we will denote by S and call the set of states and characterized by the Markov property: the probability distribution of X_{n+1} depends only of the probability distribution of X_n and does not depend on X_k for all $k \leq n - 1$.

Definition 3. *We say that a sequence of random variables $(X_n)_{n\geq 0}$, $X_n : \Omega \to S \subset \mathbb{Z}$, is a Markov chain with initial distribution λ and transition matrix $P = (p_{ij})_{i,j\in S}$ if*

(1) *X_0 has distribution $\lambda = \{\lambda_i \mid i \in S\}$ and*
(2) *the Markov property holds:*

$$\mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n, \ldots, X_0 = i_0) = \mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n) = p_{i_n i_{n+1}}.$$

Note that the ith row of P is the probability distribution for X_{n+1} conditioned on the fact that $X_n = i$. Therefore, all entries of the matrix P are nonnegative, and the row sums are equal to one:

$$p_{ij} \geq 0, \quad \sum_{j\in S} \mathbb{P}(X_{n+1} = j \mid X_n = i) = \sum_{j\in S} p_{ij} = 1.$$

A matrix P satisfying these conditions in called *stochastic*.

Some natural questions about a Markov chain are:

- What is the equilibrium probability distribution, i.e., the one that is preserved from step to step?
- Does the probability distribution of X_n tend to the equilibrium distribution?
- How one can find the probability to reach some particular subset of states $A \subset S$? What is the expected time to reach this subset of states?
- Suppose we have selected two disjoint subsets of states A and B. What is the probability to reach first B rather than A starting from a given state? What is the expected time to reach B starting from A?

Prior to addressing these question, we will go over some basic concepts.

3.1. **Time evolution of the probability distribution.** If the set of states S is finite, i.e., if $|S| = N$, the P^n is merely the nth power of P. If S is infinite, we define P^n by

$$(P^n)_{ij} \equiv p_{ij}^{(n)} = \sum_{k_1 \in S} \dots \sum_{k_{n-1} \in S} p_{ik_1} p_{k_1 k_2} \dots p_{k_{n-1}} p_j.$$

Theorem 6. *Let $(X_n)_{n \geq 0}$ be a Markov chain with initial distribution λ and transition matrix P. Then for all $n, m \geq 0$*

(1) $\mathbb{P}(X_n = j) = (\lambda P^n)_j$;
(2) $\mathbb{P}_i(X_n = j) = \mathbb{P}(X_{n+m} = j \mid X_m = i) = p_{ij}^{(n)}$.

Proof. (1)

$$\begin{aligned}
\mathbb{P}(X_n = j) &= \sum_{i_0 \in S} \dots \sum_{i_{n-1} \in S} \mathbb{P}(X_n = j, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) \\
&= \sum_{i_0 \in S} \dots \sum_{i_{n-1} \in S} \mathbb{P}(X_n = j \mid X_{n-1} = i_{n-1}, \dots, X_0 = i_0) \mathbb{P}(X_{n-1} = i_{n-1}, \dots, X_0 = i_0) \\
&= \sum_{i_0 \in S} \dots \sum_{i_{n-1} \in S} \mathbb{P}(X_n = j \mid X_{n-1} = i_{n-1}) \mathbb{P}(X_{n-1} = i_{n-1} \mid X_{n-2} = i_{n-1}) \dots \mathbb{P}(X_0 = i_0) \\
&= \sum_{i_0 \in S} \dots \sum_{i_{n-1} \in S} \lambda_{i_0} p_{i_0 i_1} \dots p_{i_{n-1} j} = (\lambda P^n)_j.
\end{aligned}$$

(2) The second statement is proven similarly.

□

3.2. **Communicating classes and irreducibility.** We say that state i *leads to* state j (denote it by $i \longrightarrow j$) if

$$\mathbb{P}_i(X_n = j \text{ for some } n \geq 0) > 0.$$

If i leads to j and j leads to i we say that i and j *communicate* and write $i \longleftrightarrow j$. Note that i leads to j if and only if one can find a finite sequence $k_1, \dots, k_{n-1}$ such that

$$p_{ik_1} > 0,\ p_{k_1 k_2} > 0,\ \dots,\ p_{k_{n-1} j} > 0.$$

This, in turn, is equivalent to the condition that $p_{ij}^{(n)} > 0$ for some n.

The relation $\longleftrightarrow$ is an equivalence relation as it is

(1) symmetric as if $i \longleftrightarrow j$ then $j \longleftrightarrow i$;
(2) reflective, i.e., $i \longleftrightarrow i$;
(3) transitive, as $i \longleftrightarrow j$ and $j \longleftrightarrow k$ imply $i \longleftrightarrow k$.

Therefore, the set of states is divided into equivalence classes with respect to the relation $\longleftrightarrow$ called *communicating classes.*

Definition 4. *We say that a communicating class C is closed if*

$$i \in C, i \longrightarrow j \text{ imply } j \in C.$$

Once the chain jumps into a closed class, it stays there forever.

A state i is called *absorbing* is $\{i\}$ is a closed class. In the corresponding network, the vertex i has either only incoming edges, or no incident edges at all.

Definition 5. *A Markov chain whose set of states S is a single communicating class is called irreducible.*

3.3. **Invariant distributions and measures.**

Definition 6. *A measure on a Markov chain is any vector $\lambda = \{\lambda_i \geq 0 \mid i \in S\}$. A measure is invariant (a. k. a stationary or equilibrium) if*

$$\lambda = \lambda P.$$

A measure is a distribution if, in addition, $\sum_{i\in S} \lambda_i = 1$.

Theorem 7. *Let the set of states S of a Markov chain $(X_n)_{n\geq 0}$ be finite. Suppose that for some $i \in S$*

$$\mathbb{P}_i(X_n = j) = p_{ij}^{(n)} \to \pi_j \text{ as } n \to \infty.$$

Then $\pi = \{\pi_i \mid i \in S\}$ is an invariant distribution.

Proof. Since $p_{ij}^{(n)} \geq 0$ we have $\pi_j \geq 0$. Show that $\sum_{j\in S} \pi_j = 1$. Since S is finite, we can swap the order of taking limit and summation:

$$\sum_{j\in S} \pi_j = \sum_{i\in S} \lim_{n\to\infty} p_{ij}^{(n)} = \lim_{n\to\infty} \sum_{i\in S} p_{ij}^{(n)} = 1.$$

Show that $\pi = \pi P$:

$$\pi_j = \lim_{n\to\infty} p_{ij}^{(n)} = \lim_{n\to\infty} \sum_{k\in S} p_{ik}^{(n-1)} p_{kj} = \sum_{k\in S} \lim_{n\to\infty} p_{ik}^{(n-1)} p_{kj} = \sum_{k\in S} \pi_k p_{kj}.$$

□

Remark If the set of states is not finite, then the one cannot exchange summation and taking limit. For example, $\lim_{n\to\infty} p_{ij}^{(n)} = 0$ for all i, j for a simple symmetric random walk on $\mathbb{Z}$. $\{\pi_i = 0 \mid i \in \mathbb{Z}\}$ is certainly an invariant measure, but it is not a distribution.

The existence of an invariant distribution does not guarantee convergence to it. For example, consider the two-state Markov chain with transition matrix

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

The distribution $\pi = (1/2, 1/2)$ is invariant as

$$(1/2, 1/2) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = (1/2, 1/2).$$

However, for any initial distribution $\lambda = (q, 1-q)$ where $q \in [0, 1/2) \cup (1/2, 1]$, the limit

$$\lim_{n\to\infty} P^n$$

does not exist as

$$P^{2k} = I, \quad P^{2k+1} = P.$$

In order to eliminate such cases, we introduce the concept of aperiodic states.

Definition 7. *Let us call a state i aperiodic, if $p_{ii}^{(n)} > 0$ for all sufficiently large n.*

Theorem 8. *Suppose P is irreducible and has an aperiodic state i. Then for all states j and k, $p_{jk}^{(n)} > 0$ for all sufficiently large n. In particular, all states are aperiodic.*

Proof. Since the chain is irreducible, there exist such r and s that $p_{ji}^{(r)} > 0$ and $p_{ik}^{(s)} > 0$. Then for sufficiently large n we have

$$p_{jk}^{(r+n+s)} = \sum_{i_1,\ldots,i_n \in S} p_{ji_1}^{(r)} p_{i_1 i_2} \ldots p_{i_{n-1} i_n} p_{i_n k}^{(s)} \geq p_{ji}^{(r)} p_{ii}^{(n)} p_{ik}^{(s)} > 0.$$

□

Definition 8. *We will call a Markov chain aperiodic if all its states are aperiodic.*

Theorem 9. *Suppose that $(X_n)_{n\geq 0}$ is a Markov chain with transition matrix P and initial distribution λ. Let P be irreducible and aperiodic, and suppose that P has an invariant distribution π. Then*

$$\mathbb{P}(X_n = j) \to \pi_j \text{ as } n \to \infty \text{ for all } j.$$

In particular,

$$p_{ij}^{(n)} \to \pi_j \text{ as } n \to \infty \text{ for all } i, j.$$

A proof of this theorem is found in [6]. In the case where the set of states is finite, this result can be proven by means of linear algebra. A building block of this proof is the Perron-Frobenius theorem.

Theorem 10. *Let A be an $N \times N$ matrix with nonnegative entries such that all entries of A^m are strictly positive for all $m > M$. Then*

1. *A has a positive eigenvalue $\lambda_0 > 0$ with corresponding left eigenvector x_0 where all entries are positive;*
2. *if $\lambda \neq \lambda_0$ is any other eigenvalue, then $|\lambda| < \lambda_0$.*
3. *λ_0 has geometric and algebraic multiplicity one.*

For sufficiently large n, all entries of of P^n for stochastic irreducible aperiodic matrices P become positive. The proof of this fact is similar to the one of Theorem 8. Furthermore, the largest eigenvalue of a stochastic matrix is equal to 1. Indeed, since the row sums of P are ones, $\lambda_0 = 1$ is an eigenvalue with the right eigenvector $e = [1, \ldots, 1]^T$. Show that other eigenvalues do not exceed $\lambda_0 = 1$ in absolute value. Let (λ, v) be an eigenvalue and a corresponding right eigenvector of a stochastic matrix P. We normalize v so that

$$v_i = \max_{k \in S} |v_k| = 1.$$

Since

$$\lambda v_i = \sum_{k \in S} p_{ik} v_k,$$

we have

$$|\lambda_i| = \left| \frac{1}{v_i} \sum_{k \in S} p_{ik} v_k \right| \leq \frac{1}{v_i} \sum_{k \in S} p_{ik} |v_k| \leq \sum_{k \in S} p_{ik} = 1.$$

Theorem 11. *Every irreducible aperiodic Markov chain with a finite number of states N has a unique invariant distribution π. Moreover,*

$$\lim_{n \to \infty} qP^n = \pi$$

for any initial distribution q.

Proof. The Perron-Frobenius theorem applied to a finite stochastic irreducible aperiodic matrix P implies that the largest eigenvalue of P is $\lambda_0 = 1$ and all other eigenvalues are strictly less than 1 in absolute value. The left eigenvector π, corresponding to λ_0 has positive entries and can be normalized so that they sum up to 1. Hence,

$$\pi = \pi P, \quad \sum_{i=1}^{N} \pi_i = 1.$$

To establish the convergence, we consider the eigendecomposition of P:

$$P = V \Lambda U,$$

where Λ is the matrix with ordered eigenvalues along its diagonal:

$$\Lambda = \begin{pmatrix} 1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_N \end{pmatrix}, \quad 1 > |\lambda_1| \geq \ldots \geq |\lambda_N|,$$

V is the matrix of right eigenvectors of P: $PV = V\Lambda$, such that its first column is $e = [1, \ldots, 1]^T$, and $U = V^{-1}$ is a matrix of left eigenvectors of P: $UP = \Lambda U$. The first row of U is $\pi = [\pi_1, \ldots, \pi_N]$. One can check that if $UV = I_N$, these choices of the first column of V and the first row of U are consistent. Therefore, taking into account that

$\sum_{i=1}^{N} q_i = 1$, we calculate:

$$\begin{aligned}
&\lim_{n\to\infty} qP^n \\
&= \lim_{n\to\infty} [q_1\ q_2\ \dots\ q_N] \begin{pmatrix} 1 & * & * & * \\ 1 & * & * & * \\ & \dots & & \\ 1 & * & * & * \end{pmatrix} \begin{pmatrix} 1 & & & \\ & \lambda_2^n & & \\ & & \ddots & \\ & & & \lambda_N^n \end{pmatrix} \begin{pmatrix} \pi_1 & \pi_2 & \dots & \pi_N \\ * & * & * & * \\ & \dots & & \\ * & * & * & * \end{pmatrix} \\
&= [1\ *\ \dots\ *] \begin{pmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix} \begin{pmatrix} \pi_1 & \pi_2 & \dots & \pi_N \\ * & * & * & * \\ & \dots & & \\ * & * & * & * \end{pmatrix} \\
&= [\pi_1\ \pi_2\ \dots\ \pi_N].
\end{aligned}$$

□

4. Time reversal and detailed balance

For Markov chains, the past and the future are independent given the present. This property is symmetric in time and suggests looking at Markov chains with time running backwards. On the other hand, convergence to equilibrium shows behavior that is asymmetric in time. Hence, to complete time-symmetry, we need to start with the equilibrium distribution.

For convenience, we will use the following notations: $Markov(\lambda, P)$ denotes the discrete-time Markov chain with initial distribution λ and transition matrix P.

4.1. Detailed balance.

Definition 9. *A stochastic matrix P and a measure λ are in detailed balance if*

$$\lambda_i p_{ij} = \lambda_j p_{ji}.$$

Suppose the set of states S is finite, the matrix P is irreducible, and the system is distributed according to the invariant distribution π. The condition of detailed balance means that as $n \to \infty$, one will observe equal number of transitions from i to j and from j to i for all $i, j \in S$.

The detailed balance condition gives us another way to check whether a given measure λ is invariant.

Theorem 12. *If P and λ are in detailed balance then λ is invariant for P.*

Proof.

$$(\lambda P)_i = \sum_{j\in S} \lambda_j p_{ji} = \lambda_i \sum_{j\in S} p_{ij} = \lambda_i.$$

Hence $\lambda P = \lambda$. □

Definition 10. *Let $(X_n)_{n\geq 0}$ be $Markov(\lambda, P)$ where P is irreducible. We say that $(X_n)_{n\geq 0}$ is reversible if for all $N \geq 1$, $(X_{N-n})_{0\leq n\leq N}$ is $Markov(\lambda, P)$.*

Theorem 13. *Let P be an irreducible stochastic matrix and let λ be a distribution. Suppose that $(X_n)_{n\geq 0}$ is $Markov(\lambda, P)$. Then the following are equivalent:*

(1) *$(X_n)_{n\geq 0}$ is reversible;*
(2) *P and λ are in detailed balance.*

Proof. Both (1) and (2) imply that λ is invariant for P. Then both (1) and (2) are equivalent to the statement that $\hat{P} = P$. □

5. Markov Chain Monte Carlo methods

As we have discussed, Monte Carlo methods are those where random numbers are used in order to evaluate something nonrandom. Markov Chain Monte Carlo methods (MCMC) are those where the estimation is done via constructing a Markov Chain whose invariant distribution is the desired distribution. MCMC methods are used for numerical approximation of multidimensional integrals. In particular, such integrals arise in Bayesian parameter estimation, computational physics, and computational biology. For example, consider the problem of finding the expected value of $g(\eta)$ where η is a random variable with pdf $\pi(x)$, $x \in \mathbb{R}^d$:

$$E[g(\eta)] = \int_{x\in\mathbb{R}^d} g(x)\pi(x)dx. \tag{39}$$

Or, consider the problem of finding the expected value of $g(\eta)$ in the case where Ω is a finite set, $|\Omega| = N$ where N is huge. Let $\pi(\omega)$ be the probability distribution on Ω, then

$$E[g(\eta)] = \sum_{\omega\in\Omega} g(\eta(\omega))\pi(\omega), \tag{40}$$

Note that in both of the cases, one rarely knows π per se. Instead, a measure f proportional to π is known. For example, think about the canonical pdf for n particles in 3D:

$$\mu(x,p) = \frac{1}{Z}e^{-\beta(V(x)+|p|^2/2)},\ Z = \int_{\mathbb{R}^{6n}} e^{-\beta(V(x)+|p|^2/2)}dxdp.$$

The normalization constant Z, except for some simple cases, cannot be evaluated analytically. Therefore, $\mu(x,p)$ is, strictly speaking, unknown. However, for each (x,p) one can calculate

$$f(x,p) = e^{-\beta(V(x)+|p|^2/2)}$$

that is proportional to $\mu(x,p)$

Therefore, the problem is two-fold:

- The expected value is hard-to-evaluate due to ether high dimensionality of the integral, so that numerical quadrature methods are unappreciable, or due to the huge number of summands in the sum (think about numbers like $N = 2^n$ where $n \sim 10^k$, $k = 2, 3, 4\ldots$). Moreover, π might be far from being uniform, and some

kind of importance sampling is necessary to be able to obtain a satisfactory estimate at a reasonable number of samples of η.

- The pdf or the probability distribution π is unknown. Instead, f, that is proportional to π, is given.

5.1. **Metropolis and Metropolis-Hastings algorithms.** We will explain the idea of the Metropolis algorithm on the example of the task of numerical approximation of the sum in Eq. (40) where Ω is a finite set, $|\Omega| = N$, N is huge. We wish to construct a discrete-time Markov chain $(X_n)_{n\geq 0}$, $X_n : \Omega \to \{1, \dots, N\}$, i.e., where the each random variable X_n is simply an enumeration of the set of outcomes. Therefore, we may think that the set of states S and the set of outcomes Ω are identical. In order to be able to approximate the sum in Eq. (40), we need design the transition matrix P so that the the distribution π is invariant, and for any initial distribution λ, λP^n converges to π as $n \to \infty$. Choosing P irreducible and aperiodic, we guarantee the achievement of the convergence to the unique invariant distribution. A handy way to make P to have the desired invariant measure π, it suffices to pick P to be in detailed balance with the known measure f that is proportional to π, i.e., the transition probabilities should satisfy

$$f_i p_{ij} = f_j p_{ji}.$$

Such a transition matrix is constructed in two steps. As A. Chorin puts it, first do something stupid and then improve it.

(1) Suppose $X_n = k$. Propose a move from state k according to a user supplied irreducible aperiodic transition matrix $Q = (q_{ij})_{ij\in S}$ In the original Metropolis algorithm, the matrix Q must be symmetric, i.e., $q_{ij} = q_{ji}$. Suppose the proposed move is from state k to state l.

(2) To guarantee that the condition $f_i p_{ij} = f_j p_{ji}$ holds, accept the proposed move with the probability

$$\alpha = \min\left\{\frac{f_l}{f_k}, 1\right\} \tag{41}$$

I.e., if the proposed state l is more likely than the current state k, move to the new state. Otherwise, move there with probability f_l/f_k and stay in state k with probability $1 - f_l/f_k$.

As a result, the transition probabilities p_{ij} are given by

$$p_{ij} = q_{ij} \min\left\{\frac{f_j}{f_i}, 1\right\}, \quad p_{ii} = 1 - \sum_{j\neq i} q_{ij} \min\left\{\frac{f_j}{f_i}, 1\right\}. \tag{42}$$

Let us check that P is in detailed balance with f. Assume $i \neq j$. Let $f_j/f_i \leq 1$. Then

$$f_i p_{ij} = f_i q_{ij} \frac{f_j}{f_i} = f_j q_{ij} = f_j q_{ji} = f_i p_{ji}.$$

If $f_j/f_i > 1$ then

$$f_i p_{ij} = f_i q_{ij} = f_i q_{ji} = f_i p_{ji} \frac{f_j}{f_i} = f_j p_{ij}.$$

Therefore, we have constructed a discrete-time Markov chain converging to the desired equilibrium distribution.

The Metropolis-Hastings is a generalization of the Metropolis algorithms for the case where the matrix Q is not symmetric, i.e, $q_{ij} \neq q_{ij}$ in general. It differs from the Metropolis algorithm only by the definition of the acceptance probability α: in the Metropolis-Hastings, α is given by

$$\alpha = \min\left\{ \frac{f_l}{f_k} \frac{q_{lk}}{q_{kl}}, 1 \right\} \tag{43}$$

Therefore, the transition probabilities p_{ij} are

$$p_{ij} = q_{ij} \min\left\{ \frac{f_j}{f_i} \frac{q_{ji}}{q_{ij}}, 1 \right\}, \quad p_{ii} = 1 - \sum_{j \neq i} q_{ij} \min\left\{ \frac{f_j}{f_i} \frac{q_{ji}}{q_{ij}}, 1 \right\}. \tag{44}$$

Exercise Check that $P = (p_{ij})_{i,j \in S}$ and f are in detailed balance.

References

[1] A. Chorin and O. Hald, Stochastic Tools in Mathematics and Science, 3rd edition, Springer 2013
[2] L. Koralov and Ya. Sinai, theory of probability and stochastic processes, 2nd edition, Springer, 2007
[3] C. Hartmann, J. C. Latorre, and G. Ciccotti, On two possible definitions of the free energy for collective variables, Eur. Phys. J. Special Topics 200, 73-89, 2011
[4] D. Bindel and J. Goodman, Principles of Scientific Computing, online book, 2009
[5] S. Sykora, Surface Integrals over n-Dimensional Spheres, Stan's library, DOI: 10.3247/SL1Math05.002
[6] J. R. Norris, "Markov Chains", Cambridge University Press, 1998
[7] Ralph C. Smith, "Uncertainty Quantification, Theory, Implementation, and Applications", SIAM 2014
[8] Wikipedia