Stochastic Processes and Modeling in Physics

Lecture Notes written by

S. Belan, V. Parfenyev, and M. Chertkov

20 September 2022

Abstract

The course offers as a soft and self-contained introduction to modern applied probability covering theory and application of stochastic models. Emphasis is placed on intuitive explanations of the theoretical concepts such as random walks, the law of large numbers, Markov processes, mutual information, Shannon's entropy, etc., supplemented by practical/computational implementations of basic algorithms. Most of the discussed concepts are illustrated with examples from natural sciences. To successfully master the discipline, students must have basic skills in creating programs in any programming language (Python, Julia, etc.).

Contents

Chapter 1

Random Variable. Moments. Characteristic Function

To define a random (or stochastic) variable one needs to know a *set of possible values*, which variable can take, and a *probability distribution* over this set. The set of possible values, which we denote as Ω, can be discrete, continuous or mixed. The probability to find an instance from Ω in the interval between x and $x + dx$ is $p(x)dx$, where $p(x)$ is the probability distribution density. (This is in the continuous case, in the discrete case, or in a general case, we simply call it the probability distribution.) When we want to emphasize dependence over the entire probability distribution, $p(x)$, $\forall x \in \Omega$, we denote it by X. Somehow casually, we will often say that the random variable X takes a value, x.

From the definition of $p(x)$ it is obvious that

$$p(x) \geq 0, \quad \forall x \in \Omega, \tag{1.1}$$

and normilized

$$\int_{\Omega} p(x)dx = 1. \tag{1.2}$$

Note, that in the case when Ω is mixed, the probability distribution function contains delta functions

$$p(x) = \sum_n p_n \delta(x - x_n) + \tilde{p}(x). \tag{1.3}$$

A related object of interest is the so-called cumulative distribution function, $\mathcal{P}(x)$,

which defines the total (cumulative) probability, that X has a value $\leq x$,

$$\mathcal{P}(x) = \int_{-\infty}^{x} p(x')dx'. \tag{1.4}$$

1.1 Moments

Consider a function $f(X)$ depending on a random variable X. The *average* or *expectation value* of the function $f(X)$ is

$$\mathbb{E}[f(x)] \equiv \langle f(X)\rangle = \int_{\Omega} f(x)p(x)dx. \tag{1.5}$$

In particular, the average $\mathbb{E}[X^m] \equiv \langle X^m\rangle \equiv \mu_m$ is called the m-th moment of X, and

$$\mu_1 \equiv \mathbb{E}[X] \equiv \langle X\rangle = \int_{\Omega} xp(x)dx \tag{1.6}$$

has the name *mean* or *average*. The next commonly used characteristic are called *variance*, *dispersion* or *variation*

$$\sigma^2 = \langle (X - \langle X\rangle)^2\rangle = \mu_2 - \mu_1^2, \tag{1.7}$$

which characterizes the deviation of X from its mean value $\langle X\rangle$. The quantity σ is called *standard deviation*.

1.2 Important Examples

Bernoulli Distribution is the probability distribution of a random variable which takes the value 1 (success) with probability of p and the value 0 (failure) with the remaining probability of $q = 1 - p$. The Bernoulli distribution represents (in particular) a coin toss where 1 and 0 would represent "head" and "tail" (or vice versa), respectively. The probability distribution function is

$$p(x) = p\delta(x - 1) + q\delta(x), \tag{1.8}$$

and then

$$\mu_n = \langle X^n\rangle = \int_{-\infty}^{\infty} x^n p(x)dx = p, \quad n = 1, 2, \ldots. \tag{1.9}$$

In this case the variance is $\sigma^2 = \mu_2 - \mu_1^2 = pq$.

The number k of successes in a series of N independent Bernoulli trials is described by ***Binomial Distribution*** with parameters N and p, where p is a probability of success in each trial. The probability distribution function is given by

$$B(k, N, p) = C_N^k p^k (1-p)^{N-k} = \frac{N!}{k!(N-k)!} p^k (1-p)^{N-k}, \tag{1.10}$$

and for a single trial, i.e. $N = 1$, the binomial distribution is a Bernoulli distribution (1.8).

Another important discrete distribution is the ***Poisson Distribution***. It expresses the probability of a given number of events occurring within a fixed interval of time, if these events occur with a known average rate and independently of the pre-history (the Markov independence property). The probability to observe k events within the interval is given by

$$p(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots, \quad \lambda > 0. \tag{1.11}$$

We should not forget to check that the distribution is properly normalized (1.2). The average number of events in the interval

$$\mu_1 = \sum_{k=0}^{\infty} \frac{k\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} e^{-\lambda} = \lambda \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} e^{-\lambda} = \lambda. \tag{1.12}$$

The second moment is

$$\mu_2 = \sum_{k=0}^{\infty} \frac{k^2\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} \frac{k\lambda^k}{(k-1)!} e^{-\lambda} = \lambda \sum_{n=0}^{\infty} \frac{(n+1)\lambda^n}{n!} e^{-\lambda} = \lambda(\lambda+1), \tag{1.13}$$

and then the variance is $\sigma^2 = \mu_2 - \mu_1^2 = \lambda$. Note, that the expectation value and variance of the Poisson distribution are both equal to the same value, λ.

Some examples of the Poisson distribution are: probability distribution of the number of phone calls received by a call center per hour, probability distribution of the number of meteors greater than 1 meter in diameter that strike earth in a year, probability distribution of the number of typing errors per page, and many other.

It should be noted, that the binomial distribution (1.10) converges towards the Poisson distribution (1.11) as the number of trials N goes to infinity and the probability of success p goes to zero, while the product $Np \to \lambda$ remains fixed. Indeed,

$$\begin{aligned} p(k) &= C_N^k p^k (1-p)^{N-k} = \frac{N(N-1)\ldots(N-k+1)}{k!} \left(\frac{\lambda}{N}\right)^k \left(1 - \frac{\lambda}{N}\right)^{N-k} \\ &= \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{N}\right)^N \left(1 - \frac{1}{N}\right) \ldots \left(1 - \frac{k-1}{N}\right) \left(1 - \frac{\lambda}{N}\right)^{-k} \to \frac{\lambda^k}{k!} e^{-\lambda}, \; N \to \infty. \end{aligned} \tag{1.14}$$

This statement is known as the ***law of rare events*** or ***Poisson limit theorem*** and we will discuss it in more detail in the chapter 3.

The most important continuous distribution is ***Gaussian Distribution***. The general form of its probability density function is

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{1.15}$$

The parameter μ is the mean or expectation of the distribution, while the parameter σ is its standard deviation. The variance of the distribution is σ^2. A random variable with a Gaussian distribution is said to be normally distributed.

The Gaussian distribution is often found in the world around us due to the central limit theorem, which we will talk about in the next chapter. Here we will consider a special case of it, which is known as the ***De Moivre-Laplace theorem***. It states that the normal distribution may be used as an approximation to the binomial distribution, if the probability of success in each trial $p \in (0, 1)$ and the number of trials $N \to \infty$, i.e.

$$p(k) = C_N^k p^k q^{N-k} \to \frac{1}{\sqrt{2\pi Npq}} \exp\left(-\frac{(k-Np)^2}{2Npq}\right), \quad N \to \infty. \tag{1.16}$$

The supplementary materials to the lecture illustrate how well this relation is fulfilled for different values of N.

Next, let us consider properties of another continuous distribution – the ***Lorentz or Cauchy Distribution***. The distribution plays an important role in physics, since it describes the resonance behaviour (e.g., the form of laser spectrum). The probability density function is given by expression (check that it is properly normilized)

$$p(x) = \frac{1}{\pi}\frac{\gamma}{(x-a)^2+\gamma^2}, \quad -\infty < x < +\infty. \tag{1.17}$$

The first moment is

$$\mu_1 = \frac{\gamma}{\pi} \int\limits_{-\infty}^{+\infty} \frac{x dx}{(x-a)^2+\gamma^2} = a, \tag{1.18}$$

and the second moment μ_2 is not defined (infinite). This is an example illustrating that not all probability distributions have a bounded variance. Note that strictly speaking the first moment μ_1 is also not defined, but here we can generalize the definition of moments and calculate integrals in the sense of the principal value. Sometimes this generalization is used in physics.

1.3 Probabilistic Inequalities

Intuitively one would say that it is rare for an observation to deviate greatly from the expected value. Markov's inequality and Chebyshev's inequality place this intuition on firm mathematical footings.

Markov's inequality. For a nonnegative random variable X, and for any positive real number $C > 0$:

$$P(X \geq C) \leq \frac{\mathbb{E}[X]}{C}, \tag{1.19}$$

where $P(X \geq C)$ is the probability that a random variable X has a value greater or equal to a constant C. The proof is simple and straightforward (do it as an exercise).

Chebyshev's inequality. Let X be a random variable and let $C > 0$ be any positive real number. Then:

$$P(|X - \mathbb{E}[X]| \geq C) \leq \frac{\sigma^2}{C^2}. \tag{1.20}$$

To prove it one can use the Markov's inequality for the newly introduced $Y = (X - \mathbb{E}[X])^2$.

As an example let us consider the ***Coupon Collector's Problem***. Suppose that there are n different coupons and you want to collect all of them. At every step you can get only one random coupon. What is the probability that you still do not have all coupons after t steps? The probability that we have not a particular coupon at a single step is $1 - 1/n$. The probability that a particular coupon is missing after t steps is $(1 - 1/n)^t$. Since there is n different coupons, mean/average value of coupons that we do not have after t steps is $n(1 - 1/n)^t$. Using Markov's inequality one estimates:

$$P(\text{number of coupons, still missing} \geq 1) \leq n(1 - 1/n)^t \leq ne^{-t/n}, \tag{1.21}$$

where deriving the last inequality we have used the relation $1 - x \leq e^{-x}$.

In what follows, we will also use ***Jensen's inequality.*** It states that for a convex function g and for a random variable X, we have

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]). \tag{1.22}$$

If g is concave then the reverse inequality holds. To proof, we introduce $\mu = \mathbb{E}[X]$ and let $L_\mu(x) = a + bx$ be the tangent line to the function g at $x = \mu$, i.e., $L_\mu(\mu) = g(\mu)$. By convexity we know that $g(x) \geq L_\mu(x)$ for every point x. Thus, we have that

$$\mathbb{E}[g(X)] \geq \mathbb{E}[L_\mu(x)] = a + b\mu = g(\mathbb{E}[X]). \tag{1.23}$$

To develop some intuition and better understand the formal proof, it is useful to refer to Fig. 1.1, which contains a qualitative explanation in the caption.

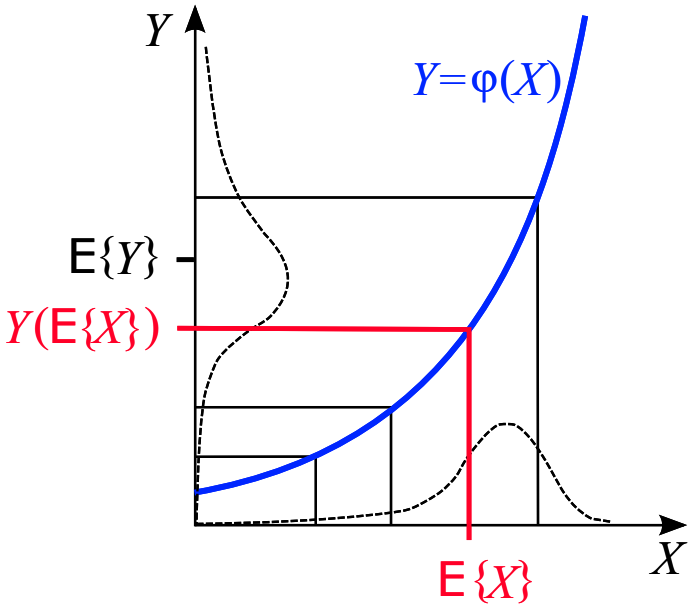


Figure 1.1: A graphical "proof" of Jensen's inequality. The dashed curve along the X axis is the hypothetical distribution of X, while the dashed curve along the Y axis is the corresponding distribution of $g(X)$ values. Note that the convex mapping $g(X)$ increasingly stretches the distribution for increasing values of X.

1.4 Characteristic Function

The characteristic function of any real-valued random variable is the Fourier-Transform of its probability distribution function,

$$G(k) = \langle e^{ikX} \rangle = \int_{-\infty}^{+\infty} e^{ikx} p(x) dx. \tag{1.24}$$

It exists for all real k and obeys relations

$$G(0) = 1, \quad |G(k)| \leq 1. \tag{1.25}$$

The characteristic function contains information about all the moments μ_m. Moreover the characteristic function allows the Taylor series representation in terms of the moments:

$$G(k) = \sum_{m=0}^{\infty} \frac{(ik)^m}{m!} \langle X^m \rangle, \tag{1.26}$$

and thus

$$\langle X^m \rangle = \frac{1}{i^m} \frac{\partial^m}{\partial k^m} G(k) \Big|_{k=0}. \tag{1.27}$$

This implies that derivatives of $G(k)$ at $k = 0$ exist up to the same m as the moments μ_m.

To illustrate the relation let us consider characteristic function of the Bernoulli distribution. Substituting Eq. (1.8) into the Eq. (1.24) one derives

$$G(k) = 1 - p + pe^{ik}, \tag{1.28}$$

and thus

$$\mu_m = \frac{\partial^m}{\partial(ik)^m}[1 - p + pe^{ik}]\Big|_{k=0} = p. \tag{1.29}$$

The result is naturally consistent with Eq. (1.9).

1.5 Cumulants

Cumulants κ_n of a probability distribution are a set of quantities that provide an alternative to the moments of the distribution. Moments determine the cumulants in the sense that any two probability distributions whose moments are identical will have identical cumulants as well, and similarly the cumulants determine the moments. In some cases theoretical treatments of problems in terms of cumulants are simpler than those using moments.

The cumulants are also defined by the characteristic function as follows

$$\ln G(k) = \sum_{m=1}^{\infty} \frac{(ik)^m}{m!}\kappa_m. \tag{1.30}$$

According to Eq. (1.25) this Taylor series start from unity. Utilizing Eqs. (1.26) and (1.30), one derives the following relations between the cumulants and the moments

$$\kappa_1 = \mu_1, \tag{1.31}$$

$$\kappa_2 = \mu_2 - \mu_1^2 = \sigma^2. \tag{1.32}$$

The procedure naturally extends to higher order moments and cumulants.

Now, consider an example of the Poisson distribution defined according to (1.11). The respective characteristic function is

$$G(p) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}e^{-\lambda}e^{ipk} = e^{-\lambda}\sum_{k=0}^{\infty} \frac{(\lambda e^{ip})^k}{k!} = exp\Big[\lambda(e^{ip} - 1)\Big], \tag{1.33}$$

and then

$$\ln G(p) = \lambda(e^{ip} - 1). \tag{1.34}$$

Next, using the definition (1.30), one finds that $\kappa_m = \lambda$, $m = 1, 2, \ldots$.

1.6 Statistical Physics

The objects like characteristic functions are very useful in the field of statistical physics. According to the *Boltzmann distribution*, the equilibrium probability $p(s)$ that a system is in a given state s

$$p(s) = \frac{1}{Z} e^{-\beta E(s)}, \quad Z = \sum_s e^{-\beta E(s)}, \tag{1.35}$$

where $\beta = 1/T$ and $E(s)$ is the energy of the state s. The normalization factor Z is called the *partition function*. In order to demonstrate utility of the partition function, let us calculate the thermodynamic value of the total energy. This is simply the expected/mean value of energy

$$\langle E \rangle = \sum_s p(s)E(s) = \frac{1}{Z} \sum_s E(s) e^{-\beta E(s)} = -\frac{1}{Z} \frac{\partial Z}{\partial \beta} = -\frac{\partial \ln Z}{\partial \beta}. \tag{1.36}$$

The variance of the energy (energy fluctuations) is

$$\Delta E^2 = \langle (E - \langle E \rangle)^2 \rangle = \frac{\partial^2 \ln Z}{\partial \beta^2}, \tag{1.37}$$

(Check it through straightforward computations.) One concludes that $\ln Z$ (compare to $\ln G$) plays an important role in statistical physics.

1.7 Problems

Problem 1. *Exponential Distribution.* The probability density function of an exponential distribution is

$$p(x) = \begin{cases} Ae^{-\lambda x}, & x \geq 0, \\ 0, & x < 0, \end{cases} \tag{1.38}$$

where the parameter $\lambda > 0$.

(1) Calculate the normalization constant A of the distribution.

(2) Calculate the *mean value* and the *variance* of the probability distribution.

The *characteristic function* of a distribution is

$$G(k) = \int_{-\infty}^{+\infty} e^{ikx} p(x) dx. \tag{1.39}$$

The characteristic function can be used to calculate high-order moments of the distribution.

(3) Calculate the characteristic function $G(k)$ of the exponential distribution.

(4) Utilizing $G(k)$, calculate the $m-$th moment of the distribution.

Problem 2. *Splitting the circle.* Randomly choose three points on a circle $x^2 + y^2 = 1$. These points form a triangle and divide the circle into three arcs.

(1) Calculate analytically the expected length of the arc containing the point $(1, 0)$.

(2) Confirm your analytical result by numerical simulations.

(3) Calculate analytically the probability that the center of the circle is contained within the triangle.

(4) Verify the answer using numerical simulations.

Problem 3. *Birthday's Problem.* What is the probability, p_m, that m people in a room all have different birthdays?

Solution: Let $(b_1, b_2, \ldots, b_m)$ forms a list of people birthdays, $b_i \in \{1, 2, \ldots, 366\}$. We slightly simplify the problem assuming that each year contains 366 days. There are 366^m different lists, and all are equiprobable. We should count the lists, which have $b_i \neq b_j,\ \forall i \neq j$. The amount of such lists is $\prod_{i=1}^{m}(366 - i + 1)$. Then, the final answer

$$p_m = \prod_{i=1}^{m}\left(1 - \frac{i-1}{366}\right). \tag{1.40}$$

The probability that at least 2 people in the room have the same birthday day is $1 - p_m$. Note that $1 - p_{23} > 0.5$ and $1 - p_{22} < 0.5$.

Problem 4. One hundred people line up to board an airplane. Each has a boarding pass with assigned seat. However, the first person to board has lost his boarding pass and takes a random seat. After that, each person takes the assigned seat if it is unoccupied, and one of unoccupied seats at random otherwise. What is the probability that the last person to board gets to sit in his assigned seat?

Problem 5. Calculate the characteristic function (1.24) of the Cauchy distribution (1.17). Show that moments do not exist.

Problem 6. Prove that $\kappa_3 = \mu_3 - 3\mu_2\mu_1 + 2\mu_1^3$.

Problem 7. A book of 500 pages contains 100 misprints. Estimate the probability that at least one page contains 5 misprints.

Chapter 2

Properties of Gaussian Distribution. Law of Large Numbers

Gaussian variables, generating function, Wick's theorem, independent random variables, characteristic function, central limit theorem.

2.1 One-Dimensional Normal Distribution

Let us consider a continuous random variable $-\infty < x < +\infty$ with Gaussian probability density function

$$p(x) = N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{2.1}$$

where μ and σ are the mean value and the variance of the distribution.

The moments $\langle x^n \rangle$ can be calculated by direct integration. Another way to find the high-order moments is via the characteristic function

$$\mathcal{G}(k) = \int e^{ikx} p(x) dx = \sum_{n=0}^{+\infty} \frac{i^n k^n}{n!} \langle x^n \rangle. \tag{2.2}$$

Then moments of x are coefficients in the Taylor series/expansion of the generating function. In the Gaussian case the characteristic function can be calculated explicitly

$$\mathcal{G}(k) = \exp\left(i\mu k - \frac{\sigma^2 k^2}{2}\right). \tag{2.3}$$

If the mean is set to zero, $\mu = 0$, one derives

$$\langle x^{2n} \rangle = \frac{(2n)!}{2^n n!} \sigma^{2n}, \quad \langle x^{2n+1} \rangle = 0. \tag{2.4}$$

Exercise 1.

Find the normalization constant A, the expected value μ and the variance σ for the following probability distribution

$$p(x) = A\exp(-x^2 + 2x). \tag{2.5}$$

Solution: Let us rewrite the distribution (2.5) as

$$p(x) = A\exp(-(x-1)^2 + 1). \tag{2.6}$$

Comparing this expression with (2.1), one derives

$$\mu = 1, \;\; \sigma = \frac{1}{\sqrt{2}}, \;\; A = \frac{\sqrt{\pi}}{e}. \tag{2.7}$$

2.2 Central limit theorem

Consider the sum

$$X_n = \frac{\sum_{i=1}^{n} x_i}{n}, \tag{2.8}$$

where the random numbers $x_1, x_2, \ldots, x_n$ are sampled i.i.d. from $p(x)$ with mean μ_x and variance σ_x both assumed finite. Statistical independence allows us to write

$$\mu_{X_n} = \mu_x, \quad \sigma^2_{X_n} = \frac{\sigma_x^2}{n}, \tag{2.9}$$

One observe that the variance (width of the probability distribution) shrinks according to $1/\sqrt{n}$ as n grows. Moreover, we observe that the shape of $P_n(X_n)$ becomes Gaussian/normal asymptotically (regardless of the shape of the original distribution):

$$P_n(X_n) \to N(\mu_x, \frac{\sigma_x^2}{n}) = \frac{\sqrt{n}}{\sqrt{2\pi}\sigma_x}\exp\left(-n\frac{(X_n-\mu_x)^2}{2\sigma_x^2}\right). \tag{2.10}$$

This statement, coined the central limit theorem, is one of the most important/fundamental results of statistics – known under the name of the ***central limit theorem***. Note, that formula (2.10) describes the behaviour of P_n only in a $|X_n - \mu_{X_n}| \lesssim \sigma_{X_n}$ vicinity of the mean, while the details of the probability distribution may be controlled by other asymptotics (of what is called the Cramer function or entropy function, see lecture notes for details).

Let us briefly sketch the proof of the theorem. It is convenient to change variables to

$$z_i = \frac{\sqrt{n}(x_i - \mu_x)}{\sigma_x}, \quad Z_n = n^{-1}\sum_{i=1}^{n} z_i = \frac{\sqrt{n}(X_n - \mu_x)}{\sigma_x}. \tag{2.11}$$

Obviously, $\mu_{Z_n} = \mu_z = 0$, $\sigma_z = \sqrt{n}$, and $\sigma_{Z_n} = 1$. The characteristic function of the probability density $P_n(Z_n)$ is defined as

$$g_n(k) = \langle e^{ikZ_n} \rangle = \int dZ_n P_n(Z_n) e^{ikZ_n}, \tag{2.12}$$

thus allowing the following representation

$$g_n(k) = \int dz_1 dz_2 \dots dz_n p(z_1) p(z_2) \dots p(z_n) e^{ik(z_1+z_2+\dots+z_n)/n} = \tag{2.13}$$

$$= \left(\int dz p(z) e^{ikz/n} \right)^n = \mathcal{G}^n(k/n). \tag{2.14}$$

where $\mathcal{G}(k)$ is the characteristic function of $p(z)$.

It follows from the definition of the characteristic function that at $k \to 0$

$$\mathcal{G}(k) = 1 - \frac{\sigma_z^2 k^2}{2} + O(k^3) = 1 - \frac{nk^2}{2} + O(k^3). \tag{2.15}$$

Therefore,

$$g_n(k) = \mathcal{G}^n(k/n) \approx \left(1 - \frac{k^2}{2n}\right)^n \approx \exp\left(-\frac{k^2}{2}\right), \tag{2.16}$$

where we have exploited the identity $\lim_{n\to\infty}(1 + x/n)^n = e^x$. One concludes that the characteristic function of $P(Z_n)$ converges to characteristic function of a normal distribution $N(0, 1)$: $P(Z_n) \to N(0, 1)$ at $n \to \infty$.

Quite often real-world quantities of interest are sums of a large number of independent random contributions. Then, CLT suggests that the resulting statistics are approximately normal. For example, repeating coin flipping many times results in a normal distribution for the total number of heads (or tails). The probability distribution of the total distance covered by a Brownian particle will also approach the normal distribution asymptotically.

Exercise 2. *Sum of uniformly distributed random variables*

Find the probability distribution $P_n(X_n)$ of the random variable $X_n = n^{-1}\sum_{i=1}^n x_i$, where $n \gg 1$ and $x_1, x_2, \dots, x_n$ are sampled i.i.d from the continuous uniform distribution

$$p(x) = \begin{cases} \dfrac{1}{b-a}, & \text{for } a \le x \le b, \\ \\ 0, & \text{for } x < a \text{ or } x > b, \end{cases} \tag{2.17}$$

Solution: First, let us calculate the mean value μ_x and variance σ_x^2 of the uniformly distributed random variable x

$$\mu_x = \int_{-\infty}^{+\infty} x p(x) dx = \frac{1}{b-a} \int_a^b x dx = \frac{a+b}{2}, \tag{2.18}$$

$$\sigma_x^2 = \int_{-\infty}^{+\infty} x^2 p(x) dx - \mu_x^2 = \frac{1}{b-a}\int_a^b x^2 dx - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}. \tag{2.19}$$

Accordingly to the central limit theorem:

$$P_n(X_n) \to \frac{2\sqrt{3n}}{\sqrt{2\pi}(b-a)} \exp\left(-6n\frac{(X_n - (a+b)/2)^2}{(b-a)^2}\right) \tag{2.20}$$

Exercise 3. *Sum of Gaussian variables*

Compute the probability distribution $P_n(X_n)$ of the random variable $X_n = n^{-1}\sum_{i=1}^n x_i$, where $x_1, x_2, \ldots, x_n$ are sampled i.i.d from the normal distribution (2.1) with $\mu = 0$.

Solution: The characteristic function of the distribution $P_n(X_n)$ is

$$g_n(k) = \mathcal{G}^n(k/n) = \exp\left(i\mu k - \frac{\sigma^2 k^2}{2n}\right), \tag{2.21}$$

Its Fourier transform is

$$P_n(X_n) = \int_{-\infty}^{+\infty} \frac{dk}{2\pi} g_n(k) e^{-ikX_n} = \int_{-\infty}^{+\infty} \frac{dk}{2\pi} \exp\left(-ik(X_n - \mu) - n\frac{\sigma^2 k^2}{2}\right) = \tag{2.22}$$

$$= \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{n(X_n - \mu)^2}{2\sigma^2}\right). \tag{2.23}$$

Exercise 4. *Violation of the central limit theorem*

Calculate the probability distribution $P_n(X_n)$ of the random variable $X_n = n^{-1}\sum_{i=1}^n x_i$, where $x_1, x_2, \ldots, x_n$ are independently chosen from a Cauchy distribution

$$p(x) = \frac{\gamma}{\pi}\frac{1}{x^2 + \gamma^2}. \tag{2.24}$$

Solution: The characteristic function of the Cauchy distribution is

$$\mathcal{G}(k) = \frac{\gamma}{\pi}\int_{-\infty}^{+\infty} \frac{dx}{x^2 + \gamma^2} e^{ikx} = e^{-\gamma|k|}. \tag{2.25}$$

The resulting characteristic functional expression is

$$g_n(k) = \mathcal{G}^n(k/n) = \mathcal{G}(k). \tag{2.26}$$

This expression shows that for any n the variable X_n is Cauchy-distributed with exactly the same width parameter as the individual samples. The CLT is "violated" because we have ignored an important requirement/condition for the CLT to hold – existence of the variance (first and second moments).

2.3 Multivariate Normal Distribution

Now let us consider M zero-mean random variables $x_1, x_2, \ldots, x_M$ sampled i.i.d. from a Gaussian distribution

$$p(x_1, \ldots, x_M) = \frac{1}{N} \exp\left(-\frac{x_i A_{ij} x_j}{2}\right), \tag{2.27}$$

where $\hat{A}$ is the symmetric positive definite matrix. If the matrix is diagonal, then one decomposes $p(x_1, \ldots, x_M)$ into a product and $x_1, x_2, \ldots, x_M$ are statistically independent.

In general, making a proper orthogonal transformation one can diagonalise $\hat{A}$, thus reducing the joint probability distribution into a product of independent Gaussians. There are some manipulations/results which are straightforward. For example one derives the normalization constant

$$N = \frac{(2\pi)^{M/2}}{\sqrt{\det A}}, \tag{2.28}$$

as well as generic expressions for the pair moments (correlation functions),

$$\mathbf{E}[x_i x_j] = A^{-1}_{ij}. \tag{2.29}$$

where $\hat{A}^{-1}$ denotes the inverse matrix.

For the high order moments the following relations are valid

$$\mathbf{E}[x_1 x_2 \ldots x_{2n}] = \sum \prod \mathbf{E}[x_i x_j], \tag{2.30}$$

$$\mathbf{E}[x_1 x_2 \ldots x_{2n+1}] = 0, \tag{2.31}$$

Notice, that in Eq. (2.31) we simply sum over all possible pairs in the set $x_1, x_2, \ldots, x_{2n}$. For example, Eq. (2.31) for the forth order moment transforms to

$$\mathbf{E}[x_i x_j x_k x_m] = \mathbf{E}[x_i x_j]\mathbf{E}[x_k x_m] + \mathbf{E}[x_i x_k]\mathbf{E}[x_j x_m] + \mathbf{E}[x_i x_m]\mathbf{E}[x_j x_k]. \tag{2.32}$$

In the probability theory, this result is known as the Isserlis' theorem, while physicists usually call it the Wick's theorem.

Exercise 5. *Joint probability distribution of the multivariate Gaussian variables*

The joint probability distribution of two random variables x_1 and x_2 is

$$p(x_1, x_2) = \frac{1}{N} \exp(-x_1^2 - x_1 x_2 - x_2^2). \tag{2.33}$$

(1) Calculate the normalization constant N.

(2) Calculate the marginal probability $p(x_1)$.

(3) Calculate the conditional probability $p(x_1|x_2)$.

(4) Calculate the statistical moments $\mathbf{E}[x_1^2 x_2^2]$, $\mathbf{E}[x_1 x_2^3]$, $\mathbf{E}[x_1^4 x_2^2]$ and $\mathbf{E}[x_1^4 x_2^4]$.

2.4 Problems

Problem 1. Assume that you play a dice game 50 times. Awards for the game are as follows

1, 3 or 5: 0$
2 or 4: 2$
6: 26$

(1) Estimate expected value of winnings

(2) Estimate standard deviation of winnings

(3) Estimate probability of winning at least 200$

(4) Estimate the probability of winning at least 50$ more than your friend who is playing the same dice game.

Problem 2. *Geometric mean.* Consider a random variable $X_n = (\prod_{i=1}^n x_i)^{1/n}$, where the non-negative random variables $x_1, x_2, \ldots, x_n$ are independently chosen from the probability distribution $p(x)$. Find the probability distribution $P(X_n)$ in the limit $n \to \infty$.

Problem 3. Consider two independent random variables $x \geq 0$ and $y \geq 0$ having the probability densities $p_1(x)$ and $p_2(y)$, respectively. Find the probability distribution $P(z)$ of the random variable $z = x + y$.

Problem 4. Propose a pair of random variables x and y such that

- x and y are linearly uncorrelated, i.e. $\langle xy \rangle - \langle x \rangle \langle y \rangle = 0$
- both have the same marginal normal distribution
- x and y are not jointly normally distributed

Chapter 3

The Bernoulli and Poisson Processes

A discrete stochastic process is simply a finite or infinite sequence of random variables. The examples include sequences of daily stock prices, scores in sport games, number of rainy days per month. If the random variables are time stamped in consecutive order, then we call it the arrival process. An arrival is broadly defined as an event that can be counted. For example, an arrival might refer to a service request, product order, device failure, arrival of e-mail message, arrival of telephone calls, etc.

3.1 Bernoulli Process

Bernoulli variable b is a random variable which has only two possible outcomes: it takes 1 (”success”) with probability p and otherwise 0 (”failure”) with probability $q = 1 - p$. The expected value of b and its variance are

$$\mathrm{E}[b] = 1 \times p + 0 \times q = p, \tag{3.1}$$

$$\mathrm{Var}[b] = (1 - p)^2 \times p + (0 - p)^2 \times q = pq. \tag{3.2}$$

Bernoulli process is a finite or infinite sequence of independent Bernoulli trials. In the case of unfair coin a trail is represented by a random variable - taking ’head’ or ’tail’ with the probability p and $1 - p$. The trials are independent because the coin does not ”remember” preceding trials.

Consider a random process consisting of N Bernoulli trials $B = \{b_1, b_2, \ldots b_N\}$. As usual, we assume that the probabilities of $b_i = 1$ and $b_i = 0$, where $1 \leq i \leq N$, are p and $q = 1 - p$, respectively. Then, the probability $B(n, N, p)$ to get exactly n successes in N

trials is given by the so-called binomial distribution

$$B(n, N, p) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}. \tag{3.3}$$

Indeed, the probability to have n successes within the sequence of N trials is $p^n q^{N-n}$. Multiplying this expression by the binomial coefficient $\binom{N}{n}$, which takes into account different ways to distribute successes, one obtains Eq. (3.3).

Next, we calculate the expected value and the variance of the random variable n

$$\begin{aligned} \mathrm{E}[n] &= \sum_{i=1}^{N} nB(n, N, p) = \sum_{i=1}^{N} \binom{N}{n} np^n q^{N-n} = p\frac{d}{dp} \sum_{i=1}^{N} \binom{N}{n} p^n q^{N-n} = \\ &= p\frac{d}{dp}(p+q)^N = Np(p+q)^N = pN, \end{aligned} \tag{3.4}$$

$$\begin{aligned} \mathrm{Var}[n] &= \sum_{i=1}^{N} n^2 B(n, N, p) - p^2 N^2 = \sum_{i=1}^{N} \binom{N}{n} n^2 p^n q^{N-n} - p^2 N^2 = \\ &= p^2 \frac{d^2}{dp^2} \sum_{i=1}^{N} \binom{N}{n} p^n q^{N-n} + p\frac{d}{dp} \sum_{i=1}^{N} \binom{N}{n} p^n q^{N-n} - p^2 N^2 = \\ &= p^2 \frac{d^2}{dp^2}(p+q)^N + p\frac{d}{dp}(p+q)^N - p^2 N^2 = pqN. \end{aligned} \tag{3.5}$$

Exercise 1.

Consider communication over a noisy channel with transmission rate of 1 symbol per second. The probability of error in a given symbol is p and the errors occurs independently for different symbols.

1) Denote as t_1 the time of the first error. Calculate the expected value of t_1.

2) Calculate the probability distribution $P(t_k)$, where t_k is the time of the kth error.

3) Calculate the probability distribution of the number of errors n in a sequence (packet) of length N.

4) Calculate the probability P that at least one symbol in in the packet of length N is an error.

Solution:

Let us introduce a Bernoulli process $b_1, b_2, \ldots$ with probability p of success in each trial. Here the success corresponds to emergence of error.

1) The probability distribution function $P(t_1)$ is given by the product of the probabilities of $t_1 - 1$ failures and one success

$$P(t_1) = p(1-p)^{t_1 - 1}. \tag{3.6}$$

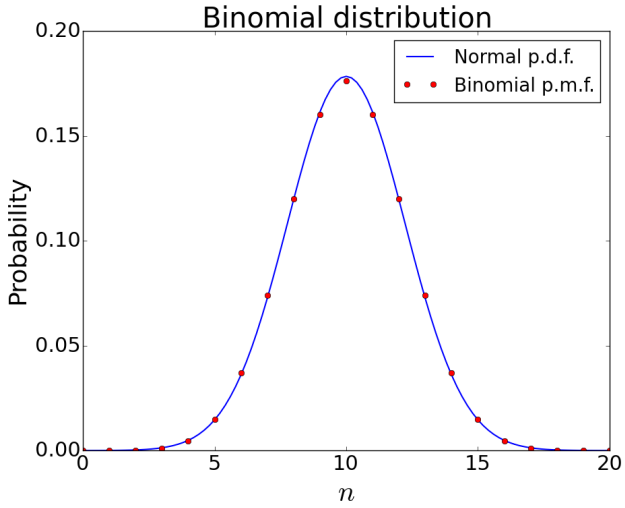


Figure 3.1: Binomial probability mass function and normal probability density function approximation for $N = 20$ and $p = 0.5$.

We obtained the so-called geometric distribution function. Then, the mean value of t_i is

$$\langle t_i \rangle = \sum_{t_1=1}^{\infty} t_i P(t_i) = \frac{1}{p}. \tag{3.7}$$

2) To estimate $P(t_k)$ one multiply the probability of observing $k-1$ errors in the packet of the first $t_k - 1$ symbols by the probability of error in t_kth symbol, i.e

$$P(t_k) = pB(k-1, t_k - 1, p) = \frac{(t_k - 1)!}{(k-1)!(t_k - k)!} p^k (1-p)^{t_k - k}. \tag{3.8}$$

This result is known as the Pascal distribution.

3) It is easy to see that the number of successes n is given by the sum of N identically distributed Bernoulli variables: $n = b_1 + b_2 + \cdots + b_N$. The central limit theorem tells us that as long as N is sufficiently large, the probability distribution of n can be approximated by a normal distribution

$$B(n, N, p) \approx N(pN, p(1-p)N) = \frac{1}{\sqrt{2\pi pqN}} \exp\left(-\frac{(n-pN)^2}{2pqN}\right).$$

The same result can be obtained directly from the binomial distribution (3.3) by exploiting the Stirling formula. Figures represents $B(n, N, p)$ in comparison with the normal approximation $N(pN, p(1-p)N)$ for $p = 0.5$ and $N = 20$.

4) $P = 1 - B(0, N, p) = 1 - (1-p)^N$.

3.2 Poisson Process

The Poisson process is used to model structureless and memoryless random arrivals in continuous time. Standard example of a Poisson process is decay of radioactive nucleus – number of decays/events/trials within a given time interval is described by the Poisson distribution.

Consider N trials which are randomly distributed within the time interval $[0, T]$. Assume that (1) each arrival is completely independent of other, and (2) the probability of arrival within an infinitesimally small time slot dt is dt/T. Let us calculate the probability $P(n, t, T)$ of n arrivals in some interval of duration $t \leq T$. The probability to observe a given arrival within this interval is t/T, while the probability that the arrival is out of this interval is $1-t/T$. Therefore, the probability that n arrivals took place is $(t/T)^n(1-t/T)^{N-n}$. Taking into account all permutations in choosing n points from N slot one derives

$$P(n, t, T) = B(n, N, t/T) = \frac{N!}{n!(N-n)!}\left(\frac{t}{T}\right)^n\left(1-\frac{t}{T}\right)^{N-n}. \tag{3.9}$$

Next let us analyze the limit $N, T \to \infty$ assuming that the average rate, i.e. frequency of arrivals, $\lambda = N/T$, is finite. One derives

$$P(n, t) = \frac{(\lambda t)^n}{n!}e^{-\lambda t}, \tag{3.10}$$

which is known as the Poisson distribution. The sequence of trials which occur randomly and independently from each other is called the Poisson point process.

Now we consider the distribution of the inter-arrival time in the previous example. Let $\{t_1, t_2, \dots\}$ be the ordered sequence of arrivals. Obviously, $t_i = T_1 + T_2 + \cdots + T_{i-1}$, where $T_i = t_{i+1} - t_i$ is the inter-arrival time. Our goal is to calculate the probability density $p(T)$ of the positive random variable T_i. One observe that the following identity holds

$$\int_T^\infty p(T')dT' = P(0, T) = e^{-\lambda T}. \tag{3.11}$$

The left hand side of this equation represents the probability that the inter-arrival time is larger than T. This probability can be also written as the probability that there are no arrivals within the interval of duration T. Therefore,

$$p(T) = \lambda e^{-\lambda T}. \tag{3.12}$$

We conclude that the Poisson process is characterized by the exponential distribution of intervals between consecutive arrivals. The parameter λ is called the rate of the process.

An important property of the Poisson process (and of the Bernoulli process) is its invariance with respect to mixing and splitting. The sum of two independent Poisson processes with rates λ_1 and λ_2 is also the Poisson process with the rate $\lambda_1 + \lambda_2$. Analogously, the Poisson process with rate λ can be split into two independent Poisson (sub)processes with rates λ_1 and $\lambda_2 = \lambda - \lambda_1$. The splitting can be performed by coin tossing: when an arrival occur we toss a coin and with probability p and $1 - p$ add the arrival to the first process or to the second process depending of the outcome. One derives, $\lambda_1 = p\lambda$ and $\lambda_2 = (1 - p)\lambda$.

Exercise 2.

Astronomers estimate that the meteors above a certain size hit the earth on average once every 1000 years, and that the number of meteor hits follows a Poisson distribution.

1) What is the probability to observe at least one large meteor next year?

2) What is the probability of observing no meteor hits within the next 1000 years?

3) Calculate the probability distribution $P(t_n)$, where the random variable t_n represents the appearance time of the nth meteor.

Solution:

The probability of observing n meteors in a time interval t is given by

$$P(n, t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, \tag{3.13}$$

where $\lambda = 0.001$ (events per year) is the average hitting rate.

1) $Pr(n > 0$ meteors next year$) = 1 - P(0, 1) = 1 - e^{-0.001} \approx 0.001$.

2) $Pr(n = 0$ meteors next 1000 years$) = P(0, 1000) = e^{-1} \approx 0.37$.

3) It is intuitively clear that

(probability that $t_n > t$) = (probability to get at least $n - 1$ arrivals in interval $[0, t]$)

Therefore

$$\int_t^\infty p(t_n) dt_n = \sum_{k=0}^{n-1} P(k, t). \tag{3.14}$$

After simple algebra we obtain

$$p(t_n) = -\frac{d}{dt} \sum_{k=0}^{n-1} P(k, t)|_{t=t_n} = \frac{\lambda^n t_n^{n-1}}{(n-1)!} e^{-\lambda t_n}. \tag{3.15}$$

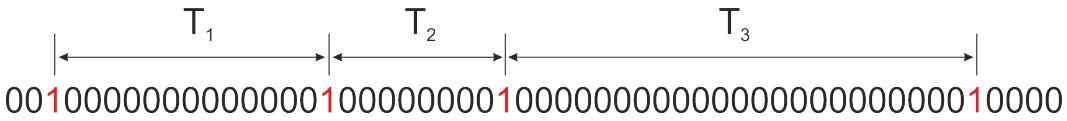


Figure 3.2: Bernoulli process with very low frequency of successes p. The distribution of the inter-arrival time t can be approximated by the Poisson distribution $p(t) = pe^{-pt}$.

3.3 Law of Rare Events

The Poisson process can be thought of as a continuous version of the Bernoulli process. Indeed, assume that the probability of success is very small, $p \ll 1$. Then the mean inter-arrival time is very large, $1/p \gg 1$. For the probability distribution of the inter-arrival time one obtains

$$P(t) = p(1-p)^{t-1} \approx pe^{-pt}. \tag{3.16}$$

Therefore, rare successes within the sequence of Bernoulli trials can be modelled as Poisson events (and vice versa).

3.4 Problems

Problem 1.

Customers arrive at a store at the Poisson rate of 10 per hour. Each is either male or female with the probability p and $1-p$, respectively.

1) Compute probability that that at least 20 customers have entered between 10 and 11 am.

2) Compute probability that exactly 10 woman entered between 10 and 11 am.

3) Compute the expected inter-arrival time of men.

4) Compute probability that there are no male customers between 2 and 4 pm.

Chapter 4

Finite Markov Chains. Efficient Mixing

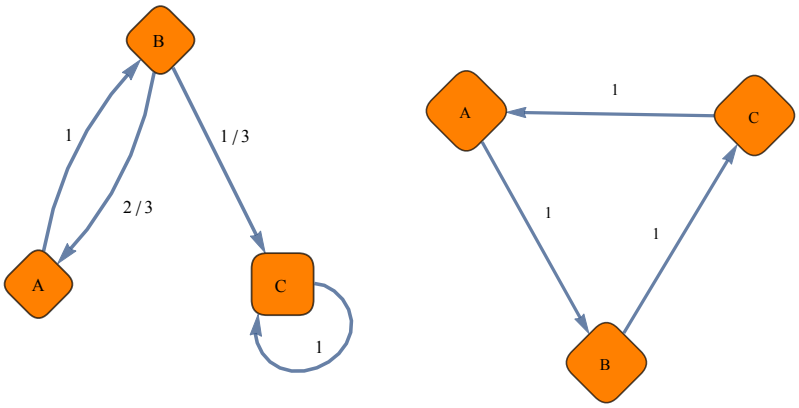
Before we give a formal definition of a Markov Chain (MC), let us watch the introduction video, which explains the origin of Markov chains and briefly describes what they are.

A Markov chain p is a stochastic process with no memory other than of its current state. We can think of a Markov chain as a random walk over a directed graph, where vertices correspond to states and edges correspond to transitions between states. Each edge $i \to j$ is associated with the probability $p(i \to j)$ of transition from the state i to the state j. A useful interactive demo can be found here.

4.1 Properties of Markov Chains

We limit our discussion to the MC with a finite number of states. Two important characteristics of a MC are ***irreducibility*** and ***aperiodicity***. A Markov chain is called irreducible, if regardless of its present state it reaches, as time progresses, any other state. We call it aperiodic if for every state i there is t such that, for all $t' \geq t$, if we start at i there is a nonzero probability of returning to i in t' steps. Aperiodicity prevents us from cycling periodically between two subsets of states and never settling down. Note that an irreducible MC with at least one self-loop is always aperiodic. Adding a self-loop is the easiest way to make an irreducible MC aperiodic.

Consider examples of MCs shown in Figure 4.1. The first example is reducible – state "C" is a trap which we reach in a finite time. In this case the stationary probability distribution corresponds to $P(C) = 1$, while the probability of finding the system in any other state is zero. Irreducibility is needed to avoid the cases with such a degenerate



(a) Reducible, Periodic (b) Irreducible, Periodic

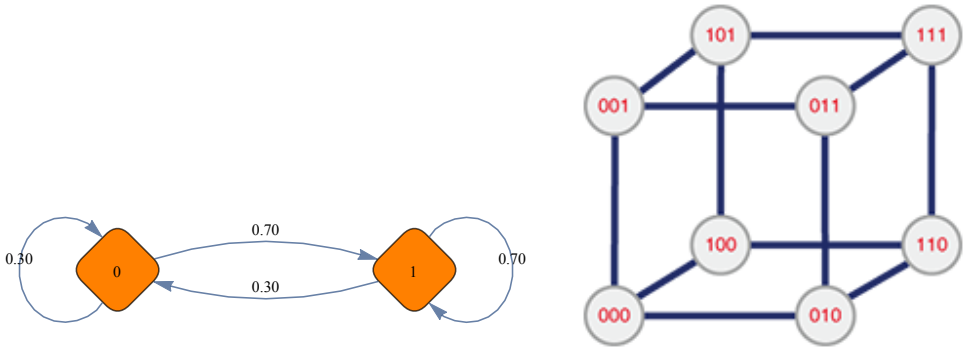
Figure 4.1: Some examples of Markov chains.

(trapped) dynamic. The second example is irreducible, but it is periodic. If we start at the state "C", we get back to the state in $3, 6, 9, \ldots$ steps. As a consequence the periodic system never forgets its initial state. One can say, that the state "C" has a period = 3. Formally, the period of state is a greatest common divisor of number of steps to return:

$$\text{period}(C) = \gcd\{n > 0 : \Pr(X_n = C | X_0 = C) > 0\}. \tag{4.1}$$

A MC is aperiodic, if and only if all its states have period = 1. The first example in Fig. (4.1) is also periodic, since the state "A" has period = 2. To make the second example aperiodic one simply needs to add a self-loop to any of the states.

Any irreducible, aperiodic Markov chain with a finite number of states will converge to a unique stationary probability distribution, no matter what initial states it starts in. This property is called ***ergodicity***, and all the Markov chains we will consider are ergodic. The opposite statement is not true, the Markov chain 4.1a is a counterexample, it has a stationary distribution and converges to it, but the MC is not irreducible and aperiodic. Note also that some MCs have stationary distributions, but they do not converge to them. The simplest example — periodic MC containing only two states. The stationary distribution is $P(A) = P(B) = 1/2$, but if you start in the state "A" you will return to it after even number of steps.



(a) Bernoulli Markov chain (b) Walking on Hypercube

Figure 4.2: Illustration of sampling idea.

4.2 Sampling

Markov chains are widely used to generate samples of some distribution. You can imagine a particle which travels on your graph according to edges' weights. After some time (for ergodic chain) the probability distribution of a particle becomes stationary (one say that the chain is mixed) and then the trajectory of the particle will represent the sample of distribution. Analyzing the trajectory you can say a lot about distribution, e.g. calculate moments and expectation values of functions.

In the Figure 4.2a you can see a Markov chain which corresponds to the Bernoulli distribution with probability of success equal to 0.7. More complicated example is shown in the Figure 4.2b. Imagine that you need to generate a random string of n bits. There is 2^n possible configurations. You can organize these configurations in a hypercube graph. The hypercube has 2^n vertices and each vertex has n neighboors, corresponding to the strings that differ from it at a single bit. Our Markov chain will walk along these edges and flip one bit at a time. The trajectory after a long time will correspond to the series of random strings. The important question is how long should we wait before our Markov chain becomes mixed (loses a memory about initial condition)? To answer this question we should look at the Markov chain from more mathematical point of view.

4.3 Stationary Distribution

The Markov process p is totally defined by a transition matrix (graph structure). Each element of this matrix corresponds to the transition probability between two states. We

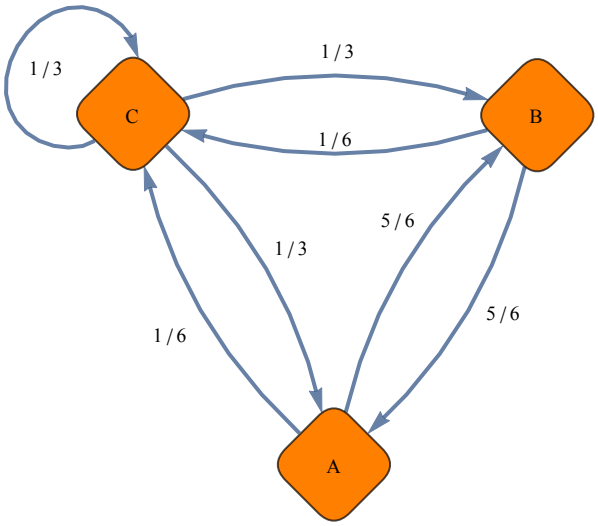


Figure 4.3: Illustration of the detailed balance.

can write the current probability distribution as a column vector π and then

$$\pi(t+1) = p\pi(t). \tag{4.2}$$

Since π is probability vector, then

$$\forall i,\ \pi_i \geq 0, \qquad \sum_i \pi_i = 1. \tag{4.3}$$

The total probability should be preserved, thus each column of p sums to 1, and all elements of p are nonnegative. Such a matrix is called ***stochastic***. Note that eigenvalues of stochastic matrix have modulus less or equal to 1. In addition, an irreducible stochastic matrix possess a simple (non-degenerate) unit eigenvalue.

Let us consider the Markov chain, which is shown in the Figure 4.3. The transition matrix is

$$p = \begin{pmatrix} 0 & 5/6 & 1/3 \\ 5/6 & 0 & 1/3 \\ 1/6 & 1/6 & 1/3 \end{pmatrix}, \tag{4.4}$$

check that the matrix is stochastic. If the initial probability distribution is $\pi(0)$, then the distribution after t steps is

$$\pi(t) = p^t\pi(0). \tag{4.5}$$

As t increases, $\pi(t)$ approaches a stationary distribution π^* (since the Markov chain is ergodic), such that

$$p\pi^* = \pi^*. \tag{4.6}$$

Thus, π^* is an eigenvector of p with eigenvalue 1, normalized according to the relation (4.3). The matrix (4.4) has three eigenvalues $\lambda_1 = 1, \lambda_2 = 1/6, \lambda_3 = -5/6$ and corresponding eigenvectors are

$$\pi^* = \left(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right)^T, \; u_2 = \left(-\frac{1}{2}, -\frac{1}{2}, 1\right)^T, \; u_3 = (-1, 1, 0)^T. \tag{4.7}$$

Suppose that we start in the state "A", i.e. $\pi(0) = (1, 0, 0)^T$. We can write the initial state as a linear combination of eigenvectors

$$\pi(0) = \pi^* - \frac{u_2}{5} - \frac{u_3}{2}, \tag{4.8}$$

and then

$$\pi(t) = p^t\pi(0) = \pi^* - \frac{\lambda_2^t}{5}u_2 - \frac{\lambda_3^t}{2}u_3. \tag{4.9}$$

Since $|\lambda_2| < 1$ and $|\lambda_3| < 1$, then in the limit $t \to \infty$ we obtain $\pi(t) = \pi^*$. The speed of convergence is defined by the eigenvalue (λ_2 or λ_3), which has the greatest absolute value.

The considered situation is typical. According to the ***Perron-Frobenius Theorem*** [5], an ergodic Markov chain with transition matrix p has a unique eigenvector π^* with eigenvalue 1, and all its other eigenvectors have eigenvalues with absolute value < 1. In general case the transition matrix p can be defective — does not have a complete basis of eigenvectors. But in this case the speed of convergence is also defined by the second largest eigenvalue [8].

4.4 Detailed Balance

We say, that a distribution π satisfies the ***detailed balance condition***, if for all pairs of states i, j

$$\pi_i p(i \to j) = \pi_j p(j \to i). \tag{4.10}$$

One can show, that if the distribution π satisfies detailed balance, then it is an p's stationary distribution, i.e. $p\pi = \pi$. Indeed, let us sum the relation (4.10) over all states i:

$$\sum_i p_{ji}\pi_i = (p\pi)_j = \sum_i p_{ij}\pi_j = \pi_j, \tag{4.11}$$

where in the last equality we have used the fact that the matrix p is stochastic. Since j is arbitrary state, we prove that $p\pi = \pi$.

If the stationary distribution π^* of a Markov chain satisfies the detailed balance, than the Markov chain is called ***reversible***. Check that the distribution π^* from our example (4.7) satisfies detailed balance. It's worth noting that the detailed balance is sufficient, but not necessary, for p to have π^* as its stationary distribution. For instance, imagine a random walk on a cycle, where we move clockwise with probability $2/3$ and counter-clockwise with probability $1/3$. This Markov chain converges to the uniform distribution, but it violates detailed balance.

A Markov chain is called ***symmetric***, if $p(i \to j) = p(j \to i)$ for all pairs of states i, j. This is a special case of detailed balance, and in the case the stationary distribution π^* is uniform.

The detailed balance is not a necessary condition for the stationary distribution. The necessary condition is a more common ***balance condition***

$$\sum_j \left(p_{ij}\pi_j - p_{ji}\pi_i\right) = 0, \tag{4.12}$$

which means that the incoming probability flux to the state i should be equal to the out-coming probability flux.

4.5 Efficient Mixing

Suppose that we want to modify a Markov chain, which is shown in the Figure 4.3. We want to obtain a faster mixing, but we need to preserve the topology of the graph and the stationary distribution. We can change the transition probabilities p_{ij}, but we cannot add a new edges to our graph. The problem is actual for some Markov Chain Monte Carlo algorithms, which we will discuss further in the course.

Here I would like to illustrate the nice idea of mixing acceleration [9]. Let me start with an analogy from the field of fluid mechanics. Consider mixing sugar in a cup of coffee, which is similar to sampling, as long as the sugar particles have to explore the entire interior of the cup (ergodicity of Markov chain). Detailed balance dynamics corresponds to the diffusion taking an enormous mixing time. Our everyday experience suggests a better solution — enhance mixing with a spoon. Spoon steering generates an out-of-equilibrium external flow which significantly accelerates mixing, while achieving

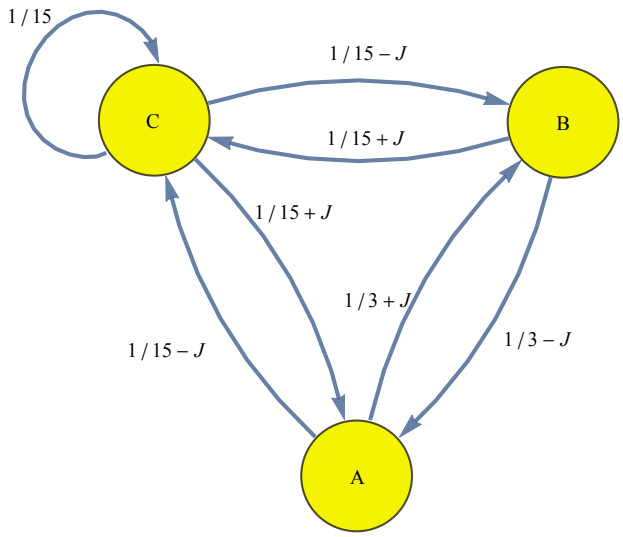


Figure 4.4: Probability fluxes for the stationary distribution π^* of the Markov chain shown in the Figure 4.3. The case $J = 0$ corresponds to the detailed balance.

the same final result — uniform distribution of sugar concentration over the cup (in our case — the same stationary distribution).

From the hydrodynamic point of view reversible Markov chains correspond to irrotational probability flows, while the violation of detailed balance relates to nonzero rotational part, e.g. correspondent to vortices contained in the flow. To understand better, look at the graph 4.4, where the edges' weights correspond to the probability fluxes $Q_{ij} = p_{ij}\pi^*_j$. We can violate the detailed balance by adding the flux J to the two cycles on our graph. We should add the flux J to both cycles, because the modified transition matrix should be stochastic. Since we know the stationary distribution π^*, we can calculate the modified transition matrix corresponding to the case of the nonzero flux J:

$$\tilde{p} = \begin{pmatrix} 0 & 5/6 - 5J/2 & 1/3 + 5J \\ 5/6 + 5J/2 & 0 & 1/3 - 5J \\ 1/6 - 5J/2 & 1/6 + 5J/2 & 1/3 \end{pmatrix}. \tag{4.13}$$

Note that all elements of the stochastic matrix $\tilde{p}$ should be nonnegative, thus we obtain the restriction on the intensity of the flux, $|J| < 1/15$.

Eigen values of the matrix (4.13) are given by expressions

$$\lambda_1 = 1, \quad \lambda_{2,3} = \frac{1}{6}\left(-2 \pm 3\sqrt{1 - 125J^2}\right). \tag{4.14}$$

The speed of mixing is defined by the value of $W = \min_J(|\lambda_2|, |\lambda_3|)$. Minimizing the quantity, we find the optimal flux $J^2_{opt} = 1/125$ and the value of $W_{opt} = 1/3$. Note that we enhance the mixing in comparison with (4.4), while the steady distribution remains unchanged.

4.6 Problems

Problem 1. *Hardy-Weinberg Law.* Consider an experiment with rabbits matting. Let us follow evolution of a particular gene that appears in two types, G or g. A rabbit has a pair of genes, either GG (dominant), Gg (hybrid — the order is irrelevant, so gG is the same as Gg) or gg (recessive). In the result of a single mating the offspring inherits a gene from each of its parents with equal probability. Thus, if a dominant parent (GG) mates with a hybrid parent (Gg), the offspring is dominant with probability $1/2$ or hybrid with probability $1/2$. Start with a rabbit of given character (GG, Gg, or gg) and assume that she mates with a hybrid. The offspring produced again mates with a hybrid, and the process is repeated for a number of generations.

1) Write down the transition matrix P of the Markov chain thus defined. Is the Markov chain irreducible and aperiodic?

2) Assume that we start with a hybrid rabbit. Let μ_n be the probability distribution of the character of the rabbit of the n-th generation. In other words, $\mu_n(GG)$, $\mu_n(Gg)$, $\mu_n(gg)$ are the probabilities that the n-th generation rabbit is GG, Gg, or gg, respectively. Compute μ_1, μ_2, μ_3. Is there a some kind of law/rule emerging?

3) Calculate P^n for general n. How does the moment, μ_n, depend on n?

4) Calculate the stationary distribution of the Markov chain. Is detailed balance hold?

Note: The first experiment of such kind was conducted in 1858 by Gregor Mendel. He started to breed garden peas in his monastery garden and analysed the offspring of these matings.

Problem 2. You want to construct a Markov chain, which mixes in the shortest time (regardless of the initial state). The state space contains N states, and desired stationary distribution is the following: the probability to be in a state i equals to p_i. What can you say about eigenvalues of the corresponding transition matrix? Construct the transition matrix explicitly.

Problem 3. Show that if M is stochastic, its eigenvalues obey $|\lambda| \leq 1$. Hint: for a vector v, let $||v||_{max}$ denote $\max_i |v_i|$, and show that $||Mv||_{max} \leq ||v||_{max}$.

Problem 4. Give an example of a Markov chain with an infinite number of states, which is irreducible and aperiodic (prove it), but which does not converge to an equilibrium probability distribution.

Chapter 5

The Ising Model and Markov Chain Monte Carlo

The Ising model was brought in as a mathematical model of ferromagnetism in statistical mechanics. In physics the traditional focus of the Ising's model analysis is on the phase transitions and, specifically, on finding and describing vicinity of the Curie point, where the system transitions from a regular/ferromagnetic behavior at low temperatures to the mixed/paramagnetic behavior at higher temperatures. However, more than 70 years after its introduction multiple applications of the Ising model in areas like neuroscience, machine learning, image analysis, economics, etc, were discovered. In this recitation we focus on some principal issues related to simulations of the Ising models.

5.1 The Ising Model

Consider a graph where a spin $s_i = \pm 1$ pointed up or down is associated with node i. We assume that energy of the spin system is a sum of local terms, measuring elongation of spins with (local) magnetic field and terms representing pair-wise interaction of spins

$$E = -\sum_{\langle ij \rangle} J_{ij} s_i s_j - \mu \sum_j s_j h_j, \tag{5.1}$$

where the first sum is over pairs of sites i, j that are graph-neighbors, J_{ij} are the interaction constants, μ is the magnetic moment, and h_j is the magnetic field acting on the spin position at the site j. Graphs common for physical applications are regular lattices. The model also has multiple application in various engineering disciplines, where the case of a regular lattice is rare.

In the following we consider square lattice with periodic boundary conditions and without external magnetic field. We will also assume in this running example the nearest neighbors have the same interaction strength $J_{ij} = 1$. Overall, the system energy is

$$E = -\sum_{\langle ij \rangle} s_i s_j. \tag{5.2}$$

If we want to minimize energy E, we can point all spins in the same direction (ferromagnetic model). But a system is not always in its lowest energy state — depending on the temperature, its energy is sometimes higher. According to the ***Boltzmann distribution***, the equilibrium probability $P_{eq}(s)$ that a system is in a given state s is

$$P_{eq}(s) = \frac{1}{Z} e^{-\beta E(s)}, \quad Z = \sum_s e^{-\beta E(s)} \tag{5.3}$$

where $\beta = 1/T$ and Z is the normalization factor called the ***partition function***. If $T \to 0$, $\beta \to \infty$, then $P_{eq}(s)$ is non-zero only at the lowest energy states. In the opposite limit of $T \to \infty$, $\beta \to 0$ all states are equally likely.

Let's lump states with the same energy together into ***macrostates***. Then the total probability of being in a macrostate with energy E is

$$\frac{W}{Z} e^{-\beta E} = \frac{1}{Z} e^{S - \beta E} = \frac{1}{Z} e^{-\beta(E - TS)}, \tag{5.4}$$

where W is the number of states in that macrostate. The quantity $S = \ln W$ is called the entropy. The likeliest macrostate minimizes the ***free energy*** $E - TS$.

5.2 Direct Sampling (by Rejection)

Now suppose that we want to generate a random state of the Ising model, according to the Boltzmann distribution (5.3). By generating a large number of such states, we can estimate some physical quantity X, e.g. an average spin $X = (1/N) \sum_i s_i$, where N is a number of spins in the system.

A naive approach is the direct ***brute-force*** sampling. We can enumerate all states, calculate its energies, the partition function, and finally calculate the equilibrium probabilities (5.3) of each state. Then we split interval $[0, 1)$ in sections, and weight sections according to the enumerated states. Finally we generate random variable ξ, uniformly distributed over the $[0, 1]$ interval, and associate each ξ with a state. The main problem here is that our algorithm is exponential in the number of spins. If our lattice contains N

spins then the number of possible states is 2^N. So, in a system sufficiently large we will not able to calculate the partition function and the set of equilibrium probabilities. The direct sampling algorithm is exponential in the number of spins with respect both memory (saving information about all the configurations) and the time (required to compute the the partition function).

Possibly a better approach is the direct sampling ***by rejection***. We can set each spin randomly with equal probability, calculate energy E of the state and then accept it as a sample with probability $p = e^{-\beta(E-E_{min})}$ (we subtract E_{min}, so $p \leq 1$). Now we do not need to calculate the partition function, but we need to know the minimal possible value of the energy E_{min}. In our simplified model it can be easily obtained theoretically (all spins have the same direction). However, for almost all states p is exponentially small, so we would have to generate an exponential number of trial states.

To construct better algorithm we should take into account the Boltzmann factor.

5.3 Metropolis-Hastings Sampling

We start from an arbitrary initial state and then perform a random walk in a state space flipping one spin at a time. Think about the algorithm as of a Markov chain defined over 2^N vertices of the hypercube. Choosing transition probabilities over the states carefully one can guarantee that the stationary state of the Markov chain reproduces the Boltzmann distribution (5.3). The resulting algorithm works as follows: at each step one, first, chooses the random site i, then compute what change ΔE in the energy would result if we flipped s_i (while other spins are kept instant), then flip the spin s_i with the following probability

$$p = \begin{cases} 1, & \text{if } \Delta E < 0 \\ e^{-\beta \Delta E}, & \text{if } \Delta E \geq 0. \end{cases} \tag{5.5}$$

This value is based on the detailed balance condition. Since our Markov chain is irreducible and aperiodic (contains self-loops), it has unique stationary distribution. And since the Boltzmann distribution (5.3) satisfies the detailed balance (check it), the end result will be convergence to the stationary distribution.

Note that if the flip is rejected one accepts the current state as a new configuration. This is the important difference with the previously discussed direct sampling by rejection. There, rejected points are discarded and have no influence on the list of samples that

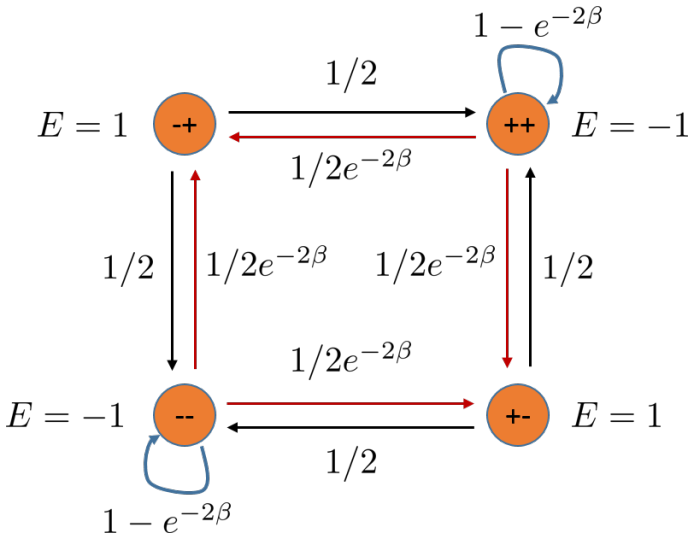


Figure 5.1: Metropolis-Hastings Markov chain example for two spins

we collect. Here the rejection into the current state being added to the list again. To understand the difference let us consider the Ising model (5.2) with only two spins. The set of possible states contain only 4 states: $--, +-, -+, ++$. The energies of states are $-1, 1, 1, -1$ correspondingly. The Markov chain is shown in the Figure 5.1. One can check that the Boltzmann distribution 5.3 satisfies the detailed balance. If the rejected configurations would be discarded the resulting distribution would be uniform.

The nontrivial question is how fast our Markov chain forgets about initial condition (how fast it mixes). A rigorous analysis of this comprehensive question is beyond the scope of this course. In practical implementations, you should continue the process till convergence, which can be verified (empirically) by checking if the expectation we compute has saturated (does not change any more). The time of convergence is (normally) polynomial in the number of spins N. If rejections do not occur often, then one can estimate mixing time following simple diffusion-in-the space state arguments. Consider two states which are farthest apart, for example all spins up vs all spins down. One can walk from one state to another in N steps — turning one spin at a time. Assuming that this walk is as convoluted as the Brownian motion one estimates that it will take N^2 steps to cover the distance N. Thus the number of steps required to generate independent samples is N^2.

Implementation of the Metropolis-Hastings algorithm as well as some additional discussion can be found in the supplemented material to this seminar (see IJulia notebook).

5.4 Gibbs Sampling

There are also other ways to enforce the detailed balance. Let us consider another example, which is called the Gibbs sampling.

Starting from a state we pick a random site i and construct two possible configurations ($s_i = 1$ and $s_i = -1$). Then we calculate the corresponding conditional (all spins except i are fixed) probabilities p_+ and p_- according to the following equations

$$p_+ + p_- = 1, \quad p_+/p_- = e^{-\beta\Delta E}, \tag{5.6}$$

where ΔE is the energy difference between the two configurations. Next, one accepts the configuration $s_i = 1$ with the probability p_+ or the configuration $s_i = -1$ with the probability p_-.

In this case our Markov chain is also defined over the hypercube. Let us check, that the calculated probabilities (5.6) and the Boltzmann distribution (5.3) satisfies detailed balance. The probability flux from the state with $s_i = 1$ to the state with $s_i = -1$ is equal to

$$Q_{-+} = \frac{1}{Z} e^{-\beta E(s_i=1)} p_-, \tag{5.7}$$

while the reversed probability flux is equal to

$$Q_{+-} = \frac{1}{Z} e^{-\beta E(s_i=-1)} p_+. \tag{5.8}$$

One finds that, indeed, the detailed balance is satisfied since $Q_{-+} = Q_{+-}$. The spirit of the Gibbs sampling is the same as in the Metropolis-Hastings Sampling. So, it is not surprising that both algorithms have comparable characteristics (e.g. mixing time).

5.5 Problems

Problem 1. Consider the Ising model (5.2) on a square lattice ($\sqrt{N} \times \sqrt{N}$) with periodic boundary conditions. Using the Gibbs sampling method, calculate the dependence of an average spin, $\langle s \rangle = (1/N) \sum_i s_i$, on the inverse temperature β and plot it. What is the critical temperature? Represent graphically the typical spin configurations below, above and near the critical temperature.

Problem 2. Consider the infinite (thermodynamic limit) two-dimensional Ising model and find the critical temperature analytically.

Problem 3. Consider the infinite (thermodynamic limit) one-dimensional Ising model and find the magnetization analytically. Is there a nontrivial critical point?

Problem 4. *Spanning Trees.* Let G be an undirected complete graph. A simple MCMC algorithm to sample uniformly from the set of spanning trees of G is as follows: Start with some spanning tree; add uniformly-at-random some edge from G (so that a cycle forms); remove uniformly-at-random some link from this cycle; repeat. Suppose now that the graph G is positively weighted, i.e., each edge e has some cost $c_e > 0$. Suggest an MCMC algorithm that samples from the set of spanning trees of G, with the probability proportional to the overall weight of the spanning for the following cases: (i) the weight of any sub-graph of G is the sum of costs of its edges; (ii) the weight of any sub-graph of G is the product of costs of its edges. In addition, (iii) estimate the average weight of a spanning tree using the algorithm of uniform sampling. Finally, (iv) implement all the algorithms on some small (but non-trivial) weighted graph of your choice. Verify that the algorithm converges to the right value.

Chapter 6

Queueing Systems

The *queueing system* model dynamical processes where individual particles/jobs advance through the system in a stochastic manner, also interacting via competition for resources (availability of servers) [6, 3]. This setting is wide spread across many disciplines, with main (traditional) applications being primarily in operation research, corresponding to industrial processes at factories, customer service at shops, cinemas, parking areas, offices and hospitals. Here we discuss basic ideas behind analysis and modelling of such queueing systems.

6.1 Parking in the Area of unlimited capacity ($M/M/\infty$ queue)

Cars enter the area of unlimited capacity with rate λ and park. One describes the state of the system at the moment of time t by the number of parked cars: $\mathcal{M} = \{0, 1, 2, \dots\}$. We will also assume that parking is modeled as another (after arrival) Poisson process. Evolution of the system state is illustrated in Fig. (6.2), and described by the following equation

$$\frac{dp_0}{dt} = -\lambda p_0 + \mu p_1, \tag{6.1}$$

$$\frac{dp_n}{dt} = \lambda p_{n-1} - (\lambda + n\mu) p_n + (n+1)\mu p_{n+1}, \quad \text{for } n \geq 1, \tag{6.2}$$

where p_n is the probability of finding n cars parked at the moment of time t.

Eq. (6.2) can be solved recursively. The steady state solution is

$$p_n = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n \exp\left(-\frac{\lambda}{\mu}\right). \tag{6.3}$$

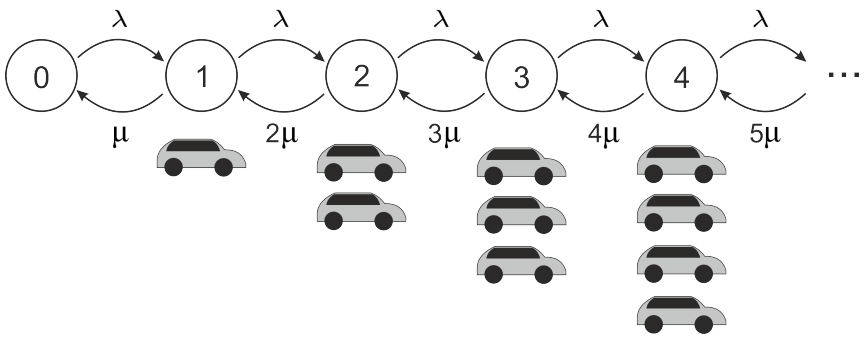


Figure 6.1: Markov Chain diagram showing evolution of the $M/M/\infty$ system state.

Then the average number of cars is

$$\langle N \rangle = \sum_{n=0}^{\infty} n p_n = \exp\left(-\frac{\lambda}{\mu}\right) \sum_{n=0}^{\infty} \frac{1}{(n-1)!} \left(\frac{\lambda}{\mu}\right)^n = \frac{\lambda}{\mu}. \tag{6.4}$$

In fact even the dynamical (transient) version of Eq. (6.3) can be integrated and analyzed. One derives

$$)p_n(t) = \frac{1}{n!} \left((1 - e^{-\mu t})\frac{\lambda}{\mu}\right)^n \exp\left(-\frac{\lambda}{\mu}(1 - e^{-\mu t})\right), \tag{6.5}$$

where one assumes that $p_n(0) = \delta_{n0}$, i.e. the queue was empty initially. Obviously, at $t \to \infty$ Eq. (6.5) converges to the steady-state distribution (6.3). The average number of cars as a function of time is given by

$$\langle N(t) \rangle = \frac{\lambda}{\mu}(1 - e^{-\mu t}). \tag{6.6}$$

6.2 Single Server Model ($M/M/1$)

Now let us shift to the queueing model describing processing of customers by a single server/teller. This queuing system is known as $M/M/1$ system — where the notation indicates that arrival and departure are both Markovian and the single server processes one particle/job at a time. We assume that the server picks next customer from the queue according to the first-come, first-served protocol. Other customers, which have already arrived the system but were not yet served, are assumed waiting in line (waiting room) of an infinite capacity (customers arriving the queue are newer rejected. Scheme of such an M/M/1 system is shown in Fig. (6.2) and the respective Markov Chain, showing transitions between states of the system, is shown in Fig. (6.3)

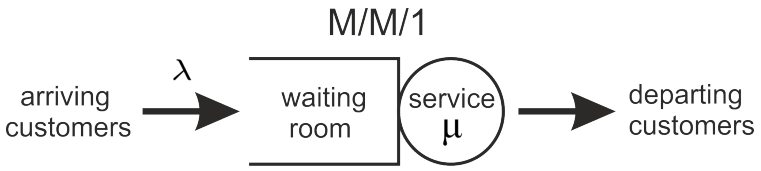


Figure 6.2: Illustration of the M/M/1 queuing system.

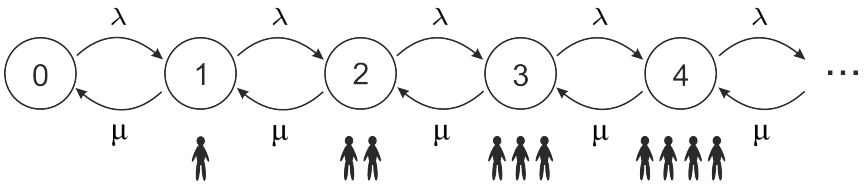


Figure 6.3: State space (Markov Chain) diagram for the $M/M/1$ queue.

State of the $M/M/1$ system evolves according to the following equation

$$\frac{dp_0}{dt} = -\lambda p_0 + \mu p_1, \tag{6.7}$$

$$\frac{dp_n}{dt} = \lambda p_{n-1} - (\lambda + \mu) p_n + \mu p_{n+1}, \quad \text{for } n \geq 1, \tag{6.8}$$

where p_n is the probability to find exactly n customers in the queue at the time t. Steady state solution of Eq. (6.8) is

$$p_n = (1 - \rho)\rho^n. \tag{6.9}$$

where $\rho = \lambda/\mu$. Obviously, the steady state exists only when $\mu > \lambda$. The expected length of the queue is

$$\langle N \rangle = (1 - \rho) \sum_{n=0}^{\infty} n\rho^n = \frac{\lambda}{\mu - \lambda}. \tag{6.10}$$

Note that the length of the queue becomes infinite at $\lambda = \mu$.

Assume that the system is initially in the state m, i.e. $p_n(0) = \delta_{nm}$ with m customers in the queue. Then probability that the system will be observed in the state n at the moment of time t is

$$p_n(t) = e^{-(\lambda+\mu)t}[\rho^{\frac{n-m}{2}} I_{n-m}(at) + \rho^{\frac{n-m-1}{2}} I_{n+m+1}(at) + \\ +(1 - \rho)\rho^n \sum_{k=n+m+2} \rho^{\frac{k}{2}} I_k(at)], \tag{6.11}$$

where $a = 2\sqrt{\mu\lambda}$.

Exercise 1

For the stationary $M/M/1$ queueing system with arrival rate λ and the service rate μ compute

(i) probability that system is empty;

(ii) average number of customers in the waiting room;

(iii) average time in the waiting room;

(iv) average time in the system;

(v) probability density function for the time a customer spends in the system.

Solution:: (i) $p_0 = 1 - \rho$.

(ii) $\langle L_q \rangle = \frac{\lambda^2}{\mu(\mu-\lambda)}$.

(iii) $\langle T_q \rangle = \frac{\lambda^2}{\mu(\mu-\lambda)}$.

(iv) $\langle T \rangle = \frac{1}{\mu-\lambda}$.

(v) $P(T) = (\mu - \lambda)e^{-(\mu-\lambda)T}$.

6.3 Tandem Queue

Now let us briefly discuss the so-called tandem system consisting of two $M/M/1$ queues operating sequentially, as shown in Fig. (6.4). Here it is assumed that a particle/job leaving the first queue immediately enters waiting room of the second queue, thus $\lambda_2 = \mu_1$. Therefore the tandem queue is characterized by three, generally distinct, parameters, λ_1, μ_1, μ_2 – the three Poisson rates.

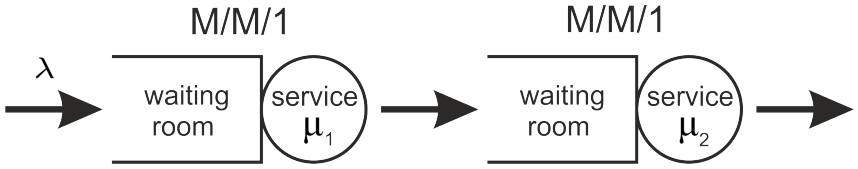


Figure 6.4: A tandem queueing system.

We have already seen on example of a single $M/M/1$ queue that if the steady state settles, i.e. if $\lambda < \mu$, departures from the single $M/M/1$ system are Poisson with rate λ. Since departures from the first queueing system turn into arrivals for the second queue, then arrivals to the second queue are also Poisson with the same rate λ. Therefore one concludes that if a steady state is established, the joint probability distribution $p(n_1, n_2)$ to find n_1 and n_2 customers in the first and second queues, respectively, is

$$p(n_1, n_2) = (1 - \rho_1)(1 - \rho_2)\rho_1^{n_1}\rho_1^{n_2}, \tag{6.12}$$

where $\rho_1 = \lambda/\mu_1$ and $\rho_2 = \lambda/\mu_2$. In other words, the tandem behaves as an independent $M/M/1$ systems. The condition of the tandem stability (existence of the steady state) is $\rho_1, \rho_2 < 1$.

Chapter 7

Brownian Motion

Brownian motion is the irregular motion of microscopic particles suspended in a fluid resulting from their collision with surrounding molecules. Einstein's paper [1] on Brownian motion along with the related works by Langevin [4] and Smoluchowski [10] have laid the foundation of the field of stochastic processes, later leading to a broad range of applications in science and engineering. Below we examine properties of the inertialess (or sometimes called overdamped), one-dimensional Brownian motion.

7.1 Langevin Equation

The inertialess, one-dimensional Brownian motion is described by the following stochastic ordinary differential equation

$$\frac{dx}{dt} = \xi(t), \tag{7.1}$$

where x is the coordinate of the particle, and $\xi(t)$ is the Gaussian white noise with zero mean and the following pair correlation function, $\langle \xi(t_1)\xi(t_2)\rangle = 2D\delta(t_1 - t_2)$. Multivariate generalization is a simple extension of the single variant process to a collection of independent processes.

Integrating Eq. (7.1) directly one derives

$$x(t) = x(0) + \int_0^t \xi(t')dt'. \tag{7.2}$$

In mathematics the random process $x(t)$ given by Eq. (7.2) is usually called the Wiener process. Let $x(0) = 0$. Since $\langle \xi(t)\rangle = 0$ by assumption, then $\langle x(t)\rangle = 0$. For the mean-

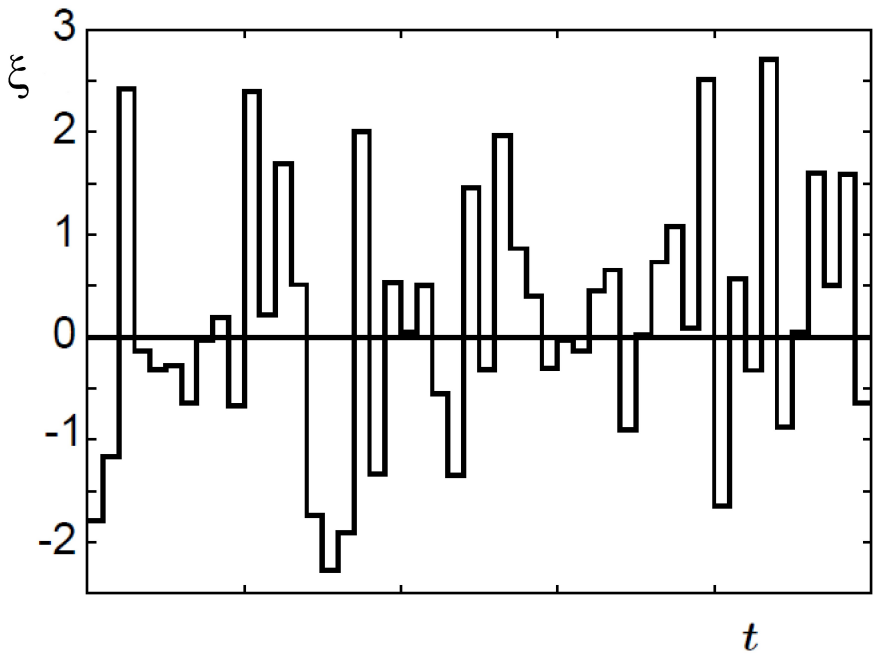


Figure 7.1: An example of the telegraph process.

square displacement one derives

$$\langle x^2(t)\rangle = \int_0^t \int_0^t \langle \xi(t')\xi(t'')\rangle dt' dt'' = 2D \int_0^t \int_0^t \delta(t_1 - t_2) dt' dt'' = 2Dt. \tag{7.3}$$

This expression describes that the displacement of a Brownian particle in time t is diffisuive, i.e. it is not proportional to the elapsed time, but rather to its square root.

7.2 Temporal Discretization

Let us consider temporal discretization of Eq. (7.1) with time step Δt so that x_i is the approximation of $x(i\Delta t)$, where $i = 0, 1, 2, ...$ Since in computer simulations one cannot realize zero correlation time, the random process ξ should be modelled in a discretized way. One option is to introduce a process whose correlation time is equal to time step duration Δt, see Fig 7.1. Specifically, the value of ξ inside of the i-th time step is assumed to be a random constant ξ_i chosen from a normal distributions $g(\xi_i) = (2\pi\sigma^2)^{-1/2} \exp(-\xi_i^2/2\sigma^2)$. Thus, the discrete-time equation of motion is as follows

$$x_{i+1} = x_i + \xi_i \Delta t. \tag{7.4}$$

Assuming $x_0 = 0$, one derives

$$\langle x_n \rangle = \Delta t \sum_{i=0}^{n-1} \langle \xi_i \rangle = 0, \tag{7.5}$$

$$\langle x_n^2 \rangle = (\Delta t)^2 \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \langle \xi_i \xi_j \rangle = (\Delta t)^2 \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \delta_{ij} \sigma^2 = n\sigma^2 (\Delta t)^2. \tag{7.6}$$

To ensure that the resulting process becomes in the continuous time limit the diffusion process with the diffusion coefficient D, one requires that, $\langle x_n^2 \rangle = 2Dn\Delta t$, and thus $\sigma^2 = 2D/\Delta t$.

Note that, thanks to the low of large numbers (central limit theory, see lecture notes and also chapter 2 of the recitation notes), one does have a flexibility in the choice of the distribution of the time-discretized ξ_i. Indeed, since the particle displacement x_n is a sum of large number identically distributed independent random variables, the central limit theorem guarantees that the resulting process, i.e. continuous time ξ, is Gaussian.

In numerical simulations it is easy to generate random variables ζ_i from the standard normal distribution ($\mu = 0$, $\sigma = 1$). In the terms of ζ_i the discrete equation of motion can be written as

$$x_{i+1} = x_i + \sqrt{2D\Delta t}\zeta_i. \tag{7.7}$$

Some simulation examples/illustrations can be found in the supplemental material to this seminar (see IJulia notebook).

7.3 Diffusion Equation

The probability density function (often we just say probability distribution, dropping the word function) for the coordinate of a Brownian particle is defined as $n(x,t) = \langle \delta(x - x(t)) \rangle$, where $x(t)$ is a particular solution of Eq. (7.1) for a given realization of the noise. The evolution of this probability density is described by the Fokker-Planck equation

$$\partial_t n = D\partial_x^2 n, \tag{7.8}$$

which is in this simple case is the diffusion equation. Obviously, $n(x,t)$ can be interpreted as the particle concentration, if one deals with a large ensemble of identical non-interacting Brownian particles. The diffusion equation (7.8) is invariant with respect to the change $x \to -x$, but it is not invariant under the time reversal transformation, $t \to -t$.

We say that the resulting process is irreversible. In the result of the temporal evolution, the probability density function broadens and become smoother, consistently with the observation that at the maximum $\partial_x^2 n < 0$ and at the minimum $\partial_x^2 n > 0$.

To solve Eq. (7.8) in an unbounded (real) domain with some initial condition $n(x, t = 0)$ one makes the Fourier transformation

$$n(x,t) = \int_{-\infty}^{+\infty} \frac{dk}{2\pi} n(k,t) e^{ikx}, \tag{7.9}$$

then to obtain $\partial_t n(k,t) = -Dk^2 n(k,t)$. Solution of the latter equation is $n(k,t) = n(k,0)e^{-Dk^2 t}$. Note that the high-order harmonics attenuate faster — consistently with the earlier remark that $n(x,t)$ gets smoother in time. Transitioning back to the Fourier to the original space (applying the inverse Fourier transform) one derives

$$\begin{aligned} n(x,t) &= \int_{-\infty}^{\infty} \frac{dk}{2\pi} \int_{-\infty}^{\infty} dx' n(x',0) e^{-ikx'} e^{-Dk^2 t} e^{ikx} \\ &= \int_{-\infty}^{+\infty} G(x-x',t) n(x',0) dx', \end{aligned} \tag{7.10}$$

where we have introduced the so-called Green function

$$G(x,t) = \frac{1}{\sqrt{4\pi D t}} \exp\left[-\frac{x^2}{4Dt}\right]. \tag{7.11}$$

which describes the probability to observe the Brownian particle at the position x in time t provided that at $t = 0$, $x = 0$.

7.4 Generating Function

In agreement with the results of the first part of this recitation, Eq. (7.11) states that the mean-square displacement of a Brownian particle grows linearly with time:

$$\langle x^2(t) \rangle = \int_{-\infty}^{+\infty} x^2 G(x,t) dx = 2Dt. \tag{7.12}$$

One may also be interested to compute high order moments of x. The simplest way to do all these computations at once (for all moments) is to analyze the generating function

$$\mathcal{Z}(\lambda, t) = \langle e^{i\lambda x(t)} \rangle = \sum_{k=0}^{+\infty} \frac{(i\lambda)^n}{n!} \langle x^k \rangle. \tag{7.13}$$

Using (7.11), one derives

$$\mathcal{Z}(\lambda, t) = \int_{-\infty}^{+\infty} e^{i\lambda x} G(x, t) dx = e^{-Dt\lambda^2}. \tag{7.14}$$

This result can also be obtained directly from diffusion equation (7.8). Indeed, it follows from (7.8) and (7.13) that the time evolution of $\mathcal{Z}(\lambda, t)$ is described by, $\partial_t \mathcal{Z} = -D\lambda^2 \mathcal{Z}$, supplemented by the initial condition, $\mathcal{Z}(\lambda, 0) = 1$ (as $x(0) = 0$). Solving this equation one arrives at the Eq. (7.14).

Now moments of x can be extracted from the Taylor expansion of (7.14) over λ. Obviously, the odd moments are all zero, $\langle x^{2k+1} \rangle = 0$, while the even moments evolve in time according to

$$\langle x^{2k} \rangle = \frac{(2k)!}{k!} (Dt)^k. \tag{7.15}$$

7.5 Wall-Bounded Brownian Motion

Next, let us consider a Brownian particle constrained not to leave the positive semi-plane, $x > 0$. Consider implementing this constraint in two different ways. One option is to introduce the so-called ***totally absorbing boundary condition*** at $x = 0$

$$n(0, t) = 0, \tag{7.16}$$

while the second option is to impose the so-called ***totally reflecting boundary condition***

$$\partial_x n(0, t) = 0 \tag{7.17}$$

Solution of the diffusion equation (7.8) with the boundary condition (7.16) and the δ-function initial condition, $n(x, 0) = \delta(x - x_0)$, becomes

$$n_a(x, t) = G(x - x_0, t) - G(x + x_0, t), \tag{7.18}$$

where G is the Green function (7.11) of the diffusion equation in free-space. The solution of the same initial value problem constrained by the boundary condition (7.17) is

$$n_r(x, t) = G(x - x_0, t) + G(x + x_0, t). \tag{7.19}$$

This method of derivation is called the image method.

7.6 Forced Brownian Motion

Let us also discuss the (overdamped) Brownian motion influenced by an external potential force, characterized by the potential $U(x)$. Generalization of the free-force overdamped Langevin equation describing this case is

$$\frac{dx}{dt} = f + \xi, \tag{7.20}$$

where $f = -dU/dx$. Then, the Fokker-Planck equation becomes

$$\partial_t n = D\partial_x^2 n - \partial_x(fn). \tag{7.21}$$

The second term on the right-hand-side of Eq. (7.21) represents the drift of the particle under action of the force f. In the case of the time-independent force the stationary solution of Eq. (7.21) becomes

$$n_0(x) \propto \exp\left(-\frac{U(x)}{D}\right). \tag{7.22}$$

This is the famous Boltzmann-Gibbs (or just Gibbs, or just Boltzmann) distribution of the equilibrium statistical physics. Here, intensity of the noise is proportional to the temperature, $D = kT$, where k is the Boltzmann constant.

As an example let us consider evolution of a Brownian particle in the parabolic potential, $U(x) = \gamma x^2/2$. In fact, this model applies broadly to the case of an overdamped dynamics in the vicinity of a minimum (then, $\gamma > 0$) or maximum (then, $\gamma < 0$) of a potential. To analyze this case one needs to solve

$$\frac{dx}{dt} + \gamma x = \xi(t), \tag{7.23}$$

which gets the following formal solution

$$x(t) = x(0)e^{-\gamma t} + \int_0^t \xi(t')e^{-\gamma(t-t')}dt'. \tag{7.24}$$

Let us assume that $x(0) = 0$. Then $\langle x(t)\rangle = 0$ and

$$\begin{aligned}\langle x^2(t)\rangle &= \int_0^t \int_0^t dt'dt''\langle\xi(t')\xi(t'')\rangle e^{-\gamma(t-t')}e^{-\gamma(t-t'')} \\ &= 2De^{-2\gamma t}\int_0^t \int_0^t dt'dt''\delta(t'-t'')e^{\gamma(t'+t'')} = \frac{D}{\gamma}(1-e^{-2\gamma t}).\end{aligned} \tag{7.25}$$

At sufficiently short times, $t \ll 1/\gamma$, the dynamics is purely diffusive, $\langle x^2(t)\rangle \simeq 2Dt$, since the particle does not feel the potential, while at the longer times, $t \gg 1/\gamma$, the dispersion (spread of the distribution) saturates, $\langle x^2(t)\rangle \simeq D/\gamma$.

In this case the Fokker-Planck equation, $\partial_t n = (\gamma\partial_x x + D\partial_x^2)n$, should be supplemented by the initial condition $n(x,0) = \delta(x)$. Then, solution (for the Green function) becomes

$$n(x,t) = \frac{1}{\sqrt{2\pi\langle x^2(t)\rangle}} \exp\left[-\frac{x^2}{2\langle x^2(t)\rangle}\right]. \tag{7.26}$$

Meaning of the latter expression is clear: the probability function $n(x,t)$ is Gaussian, but the dispersion is time-dependent. Related numerical simulations can be found in the supplemental material to this recitation (see IJulia notebook).

7.7 Problems

Problem 1. *High-order moments.* Prove that the moments, $\langle x^{2k}(t)\rangle$, for the Brownian motion in an unbounded one-dimensional space obey the following recurrent equation

$$\partial_t\langle x^{2k}\rangle = 2k(2k-1)D\langle x^{2(k-1)}\rangle. \tag{7.27}$$

Solve this equation for a particle which starts at the origin, $x = 0$, at $t = 0$.

Problem 2. *Brownian motion confined to parabolic potential.* The probability density, $n(x,t)$, of a Brownian particles confiend to potential, $U(x) = \alpha x^2/2$, is described by

$$D\partial_x^2 n + \alpha\partial_x(xn) = \partial_t n. \tag{7.28}$$

Calculate the moments $\langle x^k(t)\rangle$ under condition that $n(z,0) = \delta(x)$.

Problem 3. *Self-propelled particle.* The term "self-propelled particle" refers to an object capable of moving actively by gaining energy from the environment. Examples of such objects range from the Brownian motors and motile cells to macroscopic animals and mobile robots. The simplest two-dimensional model of a self-propelled particle is the one moving within the plane with a fixed speed v_0. Here the Cartesian components of the particle velocity v_x, v_y, stated in polar coordinates are

$$v_x = v_0\cos\varphi, \quad v_y = v_0\sin\varphi, \tag{7.29}$$

where the polar angle φ defines the direction of motion. Let us assume that φ evolves accordingly to the stochastic equation

$$\frac{d\varphi}{dt} = \xi, \tag{7.30}$$

where $\xi(t)$ stands for the Gaussian white noise with zero mean and variance, $\langle \xi(t_1)\xi(t_2)\rangle = 2D\delta(t_1 - t_2)$. The initial condition are chosen to be $\varphi(0) = 0$, $x(0) = 0$ and $y(0) = 0$.

(i) Calculate $\langle x(t)\rangle$, $\langle y(t)\rangle$.

(ii) Calculate $\langle r^2(t)\rangle = \langle x^2(t)\rangle + \langle y^2(t)\rangle$.

Problem 4. *Elementary Diffusion.*

Consider a particle jumping over nodes of the one-dimensional chain, where the states are labeled $n = 0, \pm 1, \pm 2, \ldots$. Left and right jumps are performed with the rates μ and λ respectively. Assume that at the moment of time $t = 0$ the particle was located at the node $n = 0$.

(i) Using any programming language perform and illustrate a sample of a particle path/trajectory.

(ii) Find $P(n, t)$ numerically, where $P(n, t)$ is the probability to observe a particle at the position n at the moment of time t. In order to simulate the particle motion split the time axis into discrete intervals and for any time step implement decision (on where to move next) according to the rates.

(iii) Solve (ii) analytically by solving the master equation, which is stated in continuous time as follows (this is a discrete space analog of the Fokker-Planck equation),

$$\partial_t P(n, t) = -(\lambda + \mu)P(n, t) + \mu P(n + 1, t) + \lambda P(n - 1, t). \tag{7.31}$$

(iv) Replace a discrete variable n in this equation by a continuous variable x. Under what assumptions can you do it? Solve the resulting (continuous time, continuous space) equation analytically and compare the result with the simulations performed in (ii).
Hint: if $\lambda = \mu$ then the right-hand side is just $\lambda \partial_x^2 P(x, t)$.

(v) For the original case of discrete space and setting $\lambda = 0$, solve the problem exactly. Compare the solution with (proper version of) the simulations performed in (ii).

Chapter 8

First Passage Problems

Key words: first-passage probability, first-passage time, survival probability, Kramers escape rate.

In this recitation we discuss situation when a stochastic variable (say temperature, mechanical stress, voltage etc.), descriptive of the state of the system, reaches a critical value/threshold where the system undergoes a dramatic change, call it the failure. Then, an important question is how to estimate the expected time, so-called the ***f*irst passage time**, for the system to cross the threshold first time [7]. Complementarily one is also interested to estimate probability, so called ***s*urvival probability**, that the system does not cross the threshold in time t.

8.1 First passage problem for Bernoulli processes

In fact, this subject is not new for us, as some simple examples of the first-passage problems were already discussed in the lectures and recitations (e.g., see exercises 3.1(1) and 3.2(3) of the recitation #5). Let us recall the story stated in the context of the Bernoulli random process which occurs with the probability (of success) p. Here, we define the first-passage time T as the time when a failure occurs for the first time. The probability distribution of the Bernoulli failure to occur first time at the time T is

$$P(T) = p^{T-1}(1-p). \tag{8.1}$$

Then, the mean first passage time is

$$\langle T \rangle = \sum_{T=1}^{\infty} T\, p(T) = \frac{1}{1-p}. \tag{8.2}$$

The survival probability $P(t)$, that is the probability that there have been no failure up to the time t, obeys the following equation

$$\text{(survival probability at time } t) \;=\; \text{(probability that } T > t). \tag{8.3}$$

Therefore

$$P(t) = \sum_{T_i=t+1}^{\infty} p(T_i). \tag{8.4}$$

8.2 First-passage problem for $1d$ Brownian motion

In the theory of Brownian motion an exemplary first passage problem of interest relates to computing time when a particle, initially positioned at $x = x_0$, gets to the origin, $x = 0$, the first time. The survival probability can then be stated as the probability that the particle did not visit the origin for the time t.

In numerical simulation the mean first passage time and survival probability can be both calculated by averaging over different realizations of the Brownian motion process. Consider an ensemble of $N_0 \gg 1$ non-interacting Brownian particles placed initially at x_0, i.e $x_i(0) = x_0$ where $i = 1, 2, \ldots N_0$. We should track the independent stochastic trajectories, $x_i(t)$, of each particle and measure the time T_i for a particle to reach the origin for the first time, $x(T_i = 0)$ and $x_i(t < T_i) > 0$. Then, the mean first-passage time T is just the mean value of the first passage time of a particle:

$$\langle T \rangle = \frac{1}{N} \sum_{i=1}^{N} T_i. \tag{8.5}$$

Assuming that a particle is removed from the system (dies) when it reaches the origin first time the survival probability becomes

$$P(t) = \frac{N(t)}{N_0}, \tag{8.6}$$

where $N(t)$ is the number of particles which did not get to the origin by the time t.

How to describe the first passage properties of the Brownian motion analytically? It is quite difficult from the perspective of the Langevin equation: one can write an exact expression for particle trajectory for any particular realization of noise, but it remains unclear how to extract information about the first passage. Fortunately, the problem becomes tractable in the framework of the Fokker-Planck formalism.

Indeed, the probability distribution of a particle coordinate satisfies the diffusion equation

$$\partial_t n = D\partial_x^2 n, \tag{8.7}$$

where $n(x,0) = \delta(x - x_0)$ is chosen for the initial condition. To reflect the fact that we are not interested to track the particle after it visits $x = 0$ one can simply set the zero boundary condition for the probability density at the time of the first crossing

$$n(0,t) = 0, \tag{8.8}$$

providing perfect absorbtion at the origin.

Eqs.(8.7,8.8) are solved by the image method

$$n(x,t) = \frac{1}{\sqrt{4\pi Dt}}\left[\exp\left(-\frac{(x-x_0)^2}{4Dt}\right) - \exp\left(-\frac{(x+x_0)^2}{4Dt}\right)\right]. \tag{8.9}$$

Which then allows to get the following expression for survival probability

$$P(t) = \int_0^{+\infty} n(x,t)dx = \frac{2}{\sqrt{\pi}} \int_0^{x_0/2\sqrt{Dt}} d\xi e^{-\xi^2} = \mathrm{erf}\left(\frac{x_0}{2\sqrt{Dt}}\right). \tag{8.10}$$

Exploiting the rule (8.3), one obtains

$$P(t) = \int_t^{+\infty} p(T)dT, \tag{8.11}$$

where $p(T)$ is the probability density of the first passage distribution. Then

$$p(t) = -\frac{dp(t)}{dt} = \frac{x_0}{2\sqrt{\pi D}t^{3/2}} \exp\left(-\frac{x_0^2}{4Dt}\right). \tag{8.12}$$

Since the integral $\int_0^{+\infty} tP(t)dt$ diverges, the mean first passage time is infinite. The typical (most probable) first passage time is defined as a time when $p(t)$ is at a maximum ($dp/dt = 0$), i.e. $T = x_0^2/6D$.

Exercise 2. *First passage with two thresholds*

The random process $x(t)$ is governed by the following stochastic equation

$$\frac{dx}{dt} = \xi, \tag{8.13}$$

where ξ is zero-mean Gaussian white noise with pair correlation $\langle \xi(t_1)\xi(t_2)\rangle = 2D\delta(t_1 - t_2)$. We assume that the particle is removed once it reaches $x(t) = 0$ or $x(t) = L$ starting from the initial value $x(0) = l$.

1) Calculate the survival probability $P(t)$.

2) Calculate the probability distribution $p(T)$ of the lifetime T.

3) Calculate the mean lifetime $\langle T \rangle$.

Solution:

The probability distribution $n(x,t)$ of a random variable x is recovered through solving the following initial value problem

$$\partial_t n = D\partial_x^2 n, \tag{8.14}$$

$$n(x,0) = \delta(x-l), \tag{8.15}$$

$$n(0,t) = n(L,t) = 0. \tag{8.16}$$

Then, utilizing the Fourier transform technique, one derives

$$n(x,t) = \frac{2}{L}\sum_{k=1}^{\infty} \exp\left(-\frac{\pi^2 k^2 D}{L^2}t\right)\sin\left(\frac{\pi k}{L}l\right)\sin\left(\frac{\pi k}{L}x\right). \tag{8.17}$$

1) The survival probability:

$$P(t) = \int_0^{\infty} n(x,t)dx = \frac{2}{\pi}\sum_{k=1}^{\infty}\frac{1-(-1)^k}{k}\sin\frac{\pi k l}{L}\exp\left(-\frac{\pi^2 k^2 D}{L^2}t\right). \tag{8.18}$$

2) Probability distribution of the lifetime:

$$p(T) = -\frac{dp}{dt}|_{t=T} = \frac{2\pi D}{L^2}\sum_{k=1}^{\infty}(1-(-1)^k)k\sin\frac{\pi k l}{L}\exp\left(-\frac{\pi^2 k^2 D}{L^2}T\right). \tag{8.19}$$

3) Mean lifetime:

$$\langle T \rangle = \int_0^{\infty} T p(T)dT = \frac{2}{\pi^3}\frac{L^2}{D}\sum_{k=1}^{+\infty}\frac{1-(-1)^k}{k^3}\sin\frac{\pi k l}{L} = \frac{l(L-l)}{2D}. \tag{8.20}$$

8.3 Escape rate over barrier

In the previous section we have discussed the first passage properties of the free Brownian motion. Now let us see what happens in the presence of an external potential/barrier. We are interested to understand how diffusion forces the particle to cross the barrier.

The Langevin equation for a Brownian particle in potential $U(x)$ is

$$\frac{dx}{dt} = f + \xi, \tag{8.21}$$

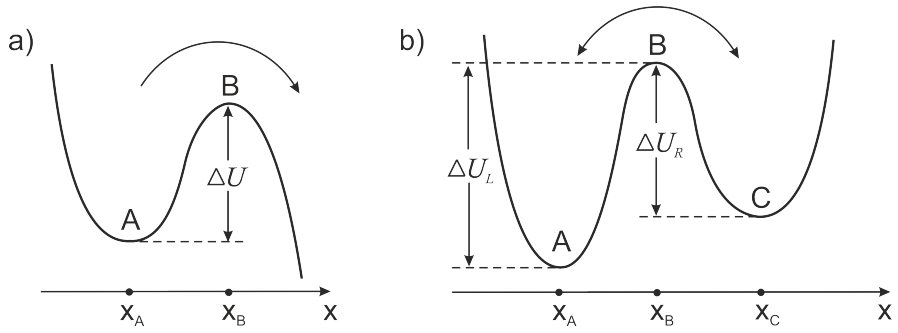


Figure 8.1: Potential energy as a function of particle coordinate in two typical cases: a) potential that allows escape to infinity; b) bistable potential.

where $f = -dU/dx$. Then corresponding Fokker-Planck equation is

$$\partial_t n = D\partial_x^2 n - \partial_x(fn). \tag{8.22}$$

The second term at the rhs represents the drift of the particle under action of the force f.

The stationary, zero flux solution of Eq. (8.22) is $\propto \exp(-U(x)/D)$. For the potential which is shown in Fig. 8.1 a, this solution is non-normalizable and, therefore, it cannot describe the probability density. The equilibrium probability distribution does not exist, since a particle placed in point A will escape from a potential well over a barrier. It is interesting to calculate the escape time, i.e. the time the Brownian particle will need to go from A to B. This first-passage problem is known as the ***Kramers problem***.

Let us consider the case of a very deep (potential) well. Namely, we assume that the mean thermal energy of the particle is much smaller than the height of the barrier, $\Delta U \gg D$. The typical noise intensity is not sufficient to drive the particle over the barrier and thus the particle may escape the well only in the result of a rare fluctuation of large amplitude. This rare event feature of the process also suggests that the time scales are well separated: the escape time is much longer than the time it takes to equilibrate around a metastable minima. However the longer one waits the more probable to observe the escape. It can be shown that the probability to escape is exponential in time

$$p(t) = \exp(-\Gamma t). \tag{8.23}$$

This means that escape is a Poisson event. The escape rate Γ is given by the Kramers formula

$$\Gamma = \frac{\omega_A \omega_B}{2\pi} \exp\left(-\frac{\Delta U}{D}\right), \tag{8.24}$$

where

$$\omega_A = \sqrt{-\frac{d^2U(x_A^2)}{dx}}, \quad \omega_B = \sqrt{-\frac{d^2U(x_B)}{dx_B^2}}. \tag{8.25}$$

Exercise 3.

Assume that a Brownian particle is placed initially at the left local minimum of the bistable potential which is shown in Fig. 8.1 b. Calculate the probabilities $p_L(t)/p_R(t)$ to find the particle at the left/right well.

Solution:

If $\Delta U_L, \Delta U_R \gg D$, then the particle spends most of the time in the vicinities of the local minimums, A and C. Jumps between two meta-stable states are Poisson events with non-equal rates

$$\Gamma_{LR} = \frac{\omega_A \omega_B}{2\pi} \exp\left(-\frac{\Delta U_L}{D}\right), \tag{8.26}$$

$$\Gamma_{RL} = \frac{\omega_C \omega_B}{2\pi} \exp\left(-\frac{\Delta U_R}{D}\right). \tag{8.27}$$

The system can be modeled with a two-state Markov chain

$$\frac{dp_L}{dt} = -\Gamma_{LR} p_L + \Gamma_{RL} p_R, \tag{8.28}$$

$$\frac{dp_R}{dt} = \Gamma_{LR} p_L - \Gamma_{RL} p_R. \tag{8.29}$$

Solution:

8.4 Problems

Problem 1. *Mortal Brownian particle*

Unstable Brownian particle moves in the interval $0 < x < L$ between two absorbing walls starting from the initial position x_0. The decay (disappearance) rate of the particle is α and the diffusion coefficient is D.

1) Find the expected lifetime of the particle analytically and by direct numerical simulations.

2) Calculate the survival probability $p(t)$ of the particle analytically.

3) What is the probability that the particle will be absorbed before it decays? Answer this question analytically or numerically.

4) Find analytically the probability that the particle will be absorbed at the left wall rather than at the right wall.

(Hint: The probability distribution of an unstable Brownian particle is described by the equation $\partial_t n = D\partial_x^2 n - \alpha n$.)

Problem 2. *Stochastic resonance*

Consider a Brownian particle moving in a periodically modulated bistable potential

$$U(x) = \frac{x^4}{4} - \frac{ax^2}{2} - \varepsilon x \cos \nu t, \tag{8.30}$$

where $a > 0$ and $\varepsilon \ll a^3$. The diffusion coefficient is small in comparison with the height of the potential barrier, $D \ll a^4$.

1) Calculate the transition rates $\Gamma_{LR}(t)$ and $\Gamma_{RL}(t)$ between two metastable states.

2) Show that at sufficiently large time the probabilities, $p_L(t)$ and $p_R(t)$, to find the particle at the left/right well is estimated as follows

$$p_L(t) = \frac{1}{2} - A\cos(\nu t + \phi_0), \quad p_R(t) = \frac{1}{2} - A\cos(\nu t + \phi_0). \tag{8.31}$$

Calculate the phase ϕ and the amplitude A in the leading order in ε.

3) Plot the amplitude A as a function of the diffusivity D for $a = 1$, $\nu = 2\pi \times 10^{-5}$. The curve has a pronounced maximum for an intermediate value $D = D_0$. Find D_0.

4) Perform numerical simulation of particle motion for $a = 1$, $\nu = 2\pi \times 10^{-5}$ and $D = D_0$. Show 10 realizations of the sequence of the first 100 transition events at the same plot.

Problem 3. *First-turn probability*

Position x of a randomly accelerated particle obeys the equation

$$\frac{d^2x}{dt^2} = \xi, \tag{8.32}$$

where $\xi(t)$ is the Gaussian white noise with zero mean and variance

$$\langle \xi(t_1)\xi(t_2)\rangle = 2D\delta(t_1 - t_2). \tag{8.33}$$

Let us assume that the particle starts its motion at $t = 0$ with initial velocity $v(0) = v_0 > 0$. Calculate the probability $P(t)$ that $v(\tau) > 0$ for any $\tau \in [0, t]$.

Chapter 9

Entropy. Mutual Information. Probabilistic Inequalities

keywords: self-information, entropy, conditional entropy, mutual information, communication channel, capacity of channel

9.1 Entropy

Let us consider a discrete random variable $x \in X$ where $X = \{x_1, \ldots, x_n\}$ and $P(x)$, as usual, is the probability mass function. The ***information content*** or ***self-information*** of an observation x_i is

$$s(x_i) = -\log_2 P(x_i). \tag{9.1}$$

We see that the smaller the probability of the outcome, the larger its self-information. Intuitively, $s(x_i)$ represents the "surprise" of seeing the outcome x_i.

The ***entropy*** of the random variable x is defined as the expected value of its self-information

$$S(X) = \mathbf{E}[s(x)] = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i). \tag{9.2}$$

The unit of entropy can be referred to as a "bit" or a "shannon".

It is straightforward to prove that

- $S(X) \geq 0$ and $S(X) = 0$ if and only if (iff) the variable X is deterministic,

i.e. a single outcome/state happens with the probability one;

- $S(X) \leq \log_2 n$ and $S(X) = \log_2 n$ iff all the outcomes are equiprobable.

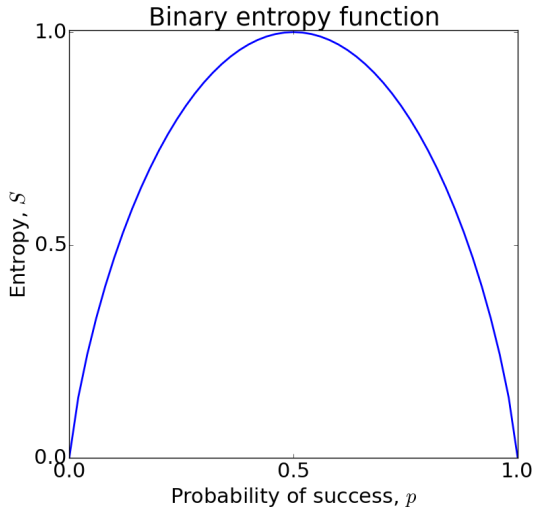


Figure 9.1: Entropy of the Bernoulli distribution as a function of the success rate, p.

This properties allow us to interpret entropy as a measure of uncertainty of the random variable x. The smaller the entropy, the larger predictability of the random process. The maximum uncertainty corresponds to the case when all outcomes have the same probability, while the minimum uncertainty occurs when the process is completely deterministic.

For the sake of illustration, let us consider the Bernoulli distribution – outcome of a potentially unfair coin tossing, where p and $q = 1 - p$ are the probabilities of observing head and tail respectively. According to the definition (9.2)

$$S_{\text{binary}}(p) = -p \log_2 p - (1 - p) \log_2 (1 - p) \tag{9.3}$$

Entropy achieves its maximum at $p = q = 1/2$ – which is the most uncertain case. The minimum uncertainty corresponds to the case $p = 1$ or $q = 1$ when the outcome of each trial is completely deterministic.

The ***joint entropy*** of a pair of discrete variables $x \in X$ and $y \in Y$ is

$$S(X, Y) = -\sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} P(x_i, y_j) \log_2 P(x_i, y_j). \tag{9.4}$$

The entropy is additive for independent random variables: $S(X, Y) = S(X) + S(Y)$ if $P(x, y) = P(x)P(y)$.

Finally, the ***conditional entropy*** is defined as

$$S(Y|X) = \sum_{i=1}^{n_X} P(x_i) S(Y|x_i) = -\sum_{i=1}^{n_X} P(x_i) \sum_{j=1}^{n_Y} P(y_j|x_i) \log_2 P(y_j|x_i) =$$

$$= -\sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} P(x_i, y_j) \log_2 P(y_j|x_i) = -\sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} P(x_i, y_j) \log_2 \frac{P(x_i, y_j)}{P(x_i)}. \tag{9.5}$$

Note, that $S(Y|X) \neq S(X|Y)$.

Exercise 1: *Properties of entropy*.

Prove that $S(X) \leq \log_2 n$, where n is the number of possible values of the random variable $x \in X$.

Solution:

The simplest proof is via Jensen's inequality. It states that if f is a convex function and u is a random variable then

$$\mathbf{E}[f(u)] \geq f[\mathbf{E}(u)]. \tag{9.6}$$

Let us define

$$f(u) = -\log_2 u, \quad u = 1/P(x). \tag{9.7}$$

Obviously, $f(u)$ is convex. Accordingly to (9.6) one obtains

$$\mathbf{E}[\log_2 P(x)] \geq -\log_2 \mathbf{E}[1/P(x)], \tag{9.8}$$

where $\mathbf{E}[\log_2 P(x)] = -S(X)$ and $\mathbf{E}[1/P(x)] = n$, so $S(X) \leq \log_2 n$.

The Jenson's inequality leads to a number of consequences for entropy, for example

$$S(X|Y) \leq S(X) \text{ with equality iff } X \text{and } Y \text{are independent}, \tag{9.9}$$

$$S(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} S(X_i) \text{ with equality iff } X_i \text{ are independent}. \tag{9.10}$$

Exercise 2: *Entropy of the English language*.

The so called Zipf's law states that the frequency of the n-th most frequent word in randomly chosen English document can be approximated by

$$p_n = \begin{cases} \frac{0.1}{n}, & \text{for } n \in 1, \ldots, 12367 \\ 0, & \text{for } n > 12367 \end{cases} \tag{9.11}$$

Under an assumption that English documents are generated by picking words at random according to Eq. (9.11) compute the entropy of the made-up English (per word).

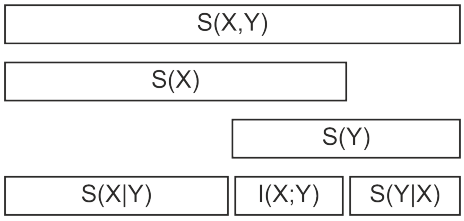


Figure 9.2: Illustration of the relations between joint entropy, marginal entropies, conditional entropies and mutual information.

Solution:

Substituting the distribution (9.11) into Eq. (9.2) one derives

$$S = -\sum_{i=1}^{12367} \frac{0.1}{n} \log_2 \frac{0.1}{n} \approx \frac{0.1}{\ln 2} \int_{10}^{123670} \frac{\ln x}{x} dx = \tag{9.12}$$

$$= \frac{1}{20 \ln 2} (\ln^2 123670 - \ln^2 10) \approx 9.9 \text{ bits}. \tag{9.13}$$

Perform the summation numerically and compare the exact result with the estimate.

Let us also calculate the entropy of English per character. The resulting entropy is fairly low $\sim$ 1 bit. Thus, the character-based entropy of a typical English text is much smaller that its entropy per word. This result is intuitively clear: after the first few letters one can often guess the rest of the word, but prediction of the next word in the sentence is a less trivial task.

9.2 Mutual Information

The ***mutual information*** of two random variables x and y, characterized by their joint distribution function, $P(x, y)$, and the marginal single-valued distribution functions, $P(x)$ and $P(y)$, is defined as follows

$$I(X;Y) = \mathbf{E}_{P(x,y)} \left[\log_2 \frac{P(x,y)}{P(x)P(y)} \right] = \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} P(x_i, y_j) \log_2 \frac{P(x_i, y_j)}{P(x_j)P(y_j)}. \tag{9.14}$$

We can also express $I(X;Y)$ in terms of respective entropies as follows

$$I(X;Y) = S(X) - S(X|Y) = S(Y) - S(Y|X) = S(X) + S(Y) - S(X,Y). \tag{9.15}$$

It is easy to see that $I(X,Y) \geq 0$, $I(X,Y) = I(Y,X)$ and $I(X,X) = S(X)$.

$P(x,y)$		X				$P(y)$
		x_1	x_2	x_3	x_4	
	y_1	1/8	1/16	1/32	1/32	1/4
Y	y_2	1/16	1/8	1/32	1/32	1/4
	y_3	1/16	1/16	1/16	1/16	1/4
	y_4	1/4	0	0	0	1/4
$P(x)$		1/2	1/4	1/8	1/8	

Table 9.1: Exemplary joint probability distribution function $P(x, y)$ and the marginal probability distributions, $P(x)$, $P(y)$, of the random variables x and y.

Mutual information is a measure of the mutual dependence between two random variables. In other words, it quantifies how much knowing one of these variables reduces uncertainty about the other. Say, if x and y are statistically independent, i.e. $P(x, y) = P(x)P(y)$, then mutual information is zero: knowing x does not give any information about y. In contrast, when y is deterministic function of x, the mutual information is maximum and equals to the entropy of x (or y), since knowing the value of x completely determines y.

Exercise 3: *Joint and Marginal entropies. Mutual information.*

The joint probability distribution $P(x, y)$ of two random variables X and Y is described in Table 9.1. Calculate the marginal probabilities $P(x)$ and $P(y)$, conditional probabilities $P(x|y)$ and $P(y|x)$, marginal entropies $S(X)$ and $S(Y)$, as well as the mutual information $I(X; Y)$.

Solution:

The probability of x_i is given by

$$P(x_i) = \sum_{j=1}^{4} P(x_i, y_j). \tag{9.16}$$

The marginal probabilities $P(x)$ and $P(y)$ are described in the Table 9.1.

Next, the single-valued marginal entropies become

$$S(X) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{8}\log_2\frac{1}{8} = \frac{7}{4}\text{ bits}, \tag{9.17}$$

$$S(Y) = -4 \times \frac{1}{4}\log_2\frac{1}{4} = 2\text{ bits}. \tag{9.18}$$

$P(x\|y)$		X			
		x_1	x_2	x_3	x_4
	y_1	1/2	1/4	1/8	1/8
Y	y_2	1/4	1/2	1/8	1/8
	y_3	1/4	1/4	1/4	1/4
	y_4	1	0	0	0

Table 9.2: Conditional probability function $P(x|y)$ for the case discussed in the exercise # 3.

The conditional probability $P(x|y)$ is

$$P(x|y) = \frac{P(x,y)}{P(y)}, \tag{9.19}$$

and the conditional entropy of x given $y = y_i$ is

$$S(X|y = y_i) = -\sum_{j=1}^{4} P(x_j|y_i) \log_2 P(x_j|y_i). \tag{9.20}$$

The results are also presented in the Table 9.2.

Now we are ready to compute the conditional entropy of X given Y:

$$S(X|Y) = \sum_{i=1}^{4} P(y_i) S(X|y = y_i) = \frac{11}{8} \text{ bits}, \tag{9.21}$$

and the mutual information

$$I(X;Y) = S(X) - S(X|Y) = \frac{7}{4} - \frac{11}{8} = \frac{3}{8} \text{ bits}. \tag{9.22}$$

9.3 Communications Over a Noise Channel

Here we consider communication over a noisy channel. A discrete memoryless channel Q is characterized by an in input alphabet $\mathcal{A}_X = \{x_1, \ldots, x_{n_X}\}$, output alphabet $\mathcal{A}_Y = \{y_1, \ldots, y_{n_Y}\}$, and a set of transition probabilities $P(y_j|x_i)$, which describes the probability to receive $y = y_j$ as an output provided that the input was $x = x_i$. We assume that the input is a random sequence of symbols $\mathcal{A}_X$ distributed according to the probability distribution function $P(x)$.

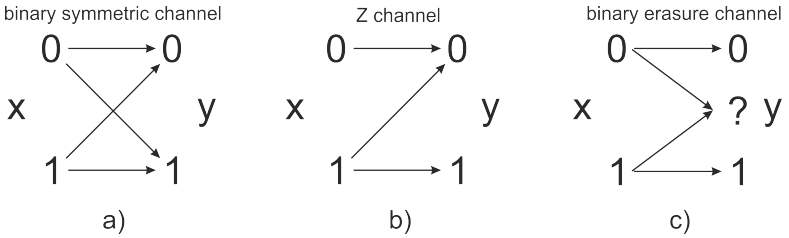


Figure 9.3: Examples of communication channels.

The capacity of a channel Q is defined as

$$C(Q) = \max_{P(X)} I(X;Y). \tag{9.23}$$

where $I(X;Y)$ is the mutual entropy of input and output.

Let us consider a couple of standard examples of noisy channels.

1. ***Binary symmetric channel***

In the case of the Binary Symmetric Channel (BSC), $\mathcal{A}_X = \mathcal{A}_Y = \{0,1\}$, i.e. both input and output alphabets are binary. When the input is 0, the output is 0 or 1 with the probabilities f and $1-f$, respectively, see Fig. (9.3) for illustration. If input is 1, the output can be 0 with the probability f or 1 with the probability $1-f$:

$$P(y=0|x=0) = 1-f, \quad P(y=0|x=1) = f, \tag{9.24}$$

$$P(y=1|x=0) = f, \quad P(y=1|x=1) = 1-f. \tag{9.25}$$

2. ***Binary erasure channel***

Alphabets: $\mathcal{A}_X = \{0,1\}$, $\mathcal{A}_Y = \{0,?,1\}$

Transition probabilities:

$$P(y=0|x=0) = 1-f, \quad P(y=0|x=1) = f, \tag{9.26}$$

$$P(y=?|x=0) = f, \quad P(y=?|x=1) = f, \tag{9.27}$$

$$P(y=1|x=0) = 0, \quad P(y=1|x=1) = 1-f. \tag{9.28}$$

3. ***Z channel***

Alphabets: $\mathcal{A}_X = \mathcal{A}_Y = \{0,1\}$

Transition probabilities:

$$P(y=0|x=0) = 1, \quad P(y=0|x=1) = f, \tag{9.29}$$

$$P(y=1|x=0) = 0, \quad P(y=1|x=1) = 1-f. \tag{9.30}$$

Exercise 4: *Binary Symmetric Channel*

Consider a BSC with the error probability, $f = 0.15$, and the following input probability distribution: $P(x = 0) = 0.9$, $P(x = 1) = 0.1$. In other words, the input signal is a Bernoulli process with $p = 0.1$.

1) Calculate the output probability distribution, $P(y)$.

2) Compute the probability $x = 1$ given $y = 1$.

3) Compute the mutual information $I(X; Y)$.

4) What is the capacity of the channel as a function of f?

Solution:

1) From the relation

$$P(y) = \sum_{j=1}^{n_X} P(y|x_j)P(x_j) \tag{9.31}$$

we derive $P(y = 1) = P(y = 1|x = 0)P(x = 0) + P(y = 1|x = 1)P(x = 1) = 0.15 \times 0.9 + 0.85 \times 0.1 = 0.22$ and $P(y = 0) = 1 - P(y = 1) = 0.78$.

2) If y is received, we do not know for sure what was an input symbol x. Can one infer the input given the output? The conditional probability $P(x|y)$ gives the posterior distribution of the input symbol x.

In accordance with the Bayes' theorem

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} = \frac{P(y|x)P(y)}{\sum_{j=1}^{n_X} P(y|x_j)P(x_j)}. \tag{9.32}$$

Then

$$P(x = 1|y = 1) = \frac{P(y = 1|x = 1)P(x = 1)}{P(y = 1|x = 0)P(x = 0) + P(y = 1|x = 1)P(x = 1)} = \\ = \frac{0.85 \times 0.1}{0.15 \times 0.9 + 0.85 \times 0.1} = 0.39. \tag{9.33}$$

We, thus, conclude that if the output was 1, then the input is also 1 with probability 0.39.

3) The mutual information $I(X; Y)$ of variables X and Y measures how much information the output conveys about the input. The larger the mutual information the more reliable the channel is. The mutual information of the channel is

$$I(X; Y) = S(Y) - S(Y|X) \tag{9.34}$$

First, the marginal entropy Y is simply $S(Y) = S_{\text{binary}}(0.22)$, where $S_{\text{binary}}(p)$ is given by 9.3. Next, the conditional entropy $S(Y|X)$ is

$$S(Y|X) = S(Y|x = 1)P(x = 1) + S(Y|x = 0)P(x = 0). \tag{9.35}$$

where

$$S(Y|x=1) = -P(y=1|x=1)\log_2 P(y=1|x=1) -$$
$$-P(y=0|x=1)\log_2 P(y=0|x=1) = -0.85\log_2 0.85 - 0.15\log_2 0.15, \tag{9.36}$$

$$S(Y|x=0) = -P(y=1|x=0)\log_2 P(y=1|x=0) -$$
$$-P(y=0|x=0)\log_2 P(y=0|x=0) = -0.15\log_2 0.15 - 0.85\log_2 0.85. \tag{9.37}$$

Therefore

$$I(X;Y) = S_{\text{binary}}(0.22) - S_{\text{binary}}(0.15) = 0.15 \text{ bits}. \tag{9.38}$$

Note, that the entropy of the input signal is $S(X) = S_{\text{binary}}(0.1) = 0.47$ bits.

4) In general

$$I(X;Y) = S_{\text{binary}}((1-f)p + (1-p)f) - S_{\text{binary}}(f). \tag{9.39}$$

Performing explicit maximization of this function over p one arrives at

$$C(Q) = \max_{P(X)} I(X;Y) = 0.39 \text{ bits}. \tag{9.40}$$

9.4 Problems

Problem 1: *Z channel*

Consider the Z channel (see Fig. 9.3c) with $f = 0.15$ and the following probability distribution of the input symbols: $P(x=0) = 0.9$, $P(x=1) = 0.1$.

(1) Compute the probability distribution of output $P(y)$.

(2) Compute the probability $x = 1$ given $y = 0$.

(3) Compute the mutual information $I(X;Y)$.

(4) What is the channel capacity?

Chapter 10

Dynamic Programming and Optimal Control Theory

Most multivariate problems are hard because of interactions between individual components — then a change in one variable affects all other variables in a global and generally unpredictable way. Anyone who has tried to pack their luggage knows what we mean. However, in some cases interactions between variables/components are factorized. The factorization may mean that solving the problem globally may actually be done in so-called greedy steps, each representing solving a much simpler sub-problem associated only with a subset of variables. Dynamic programming is computational method which allows to identify and use this factorization effectively to solve the whole problem sequentially in a greedy fashion. In this chapter we will consider some examples illustrating this approach. The discussion will follow the material of the book [5] and the article [2].

10.1 LaTeX Engine

Consider a sequence of words of varying lengths, $w_1, \ldots, w_n$, and pose the question of choosing locations for breaking the sequence at $j_1, j_2, \cdots$ into multiple lines. Once the sequence is chosen, spaces between words are stretched, so that the left and right margins are aligned. We are interested to place the line breaks in a way which would be most pleasing for the eye, which we define as associated with the least/minimal stretching.

To formalize the notion of the minimal stretching, let us introduce $c(i, j)$ denoting the cost of placing the sequence of words of lengths, $w_i, \ldots, w_j$, on a single line, and define

the total additive cost

$$c(1, j_1) + c(j_1 + 1, j_2) + \cdots + c(j_l + 1, n), \tag{10.1}$$

associated with a sequence of the line breaks. We will seek for an optimal sequence minimizing the total cost. To make description of the problem complete one needs to introduce a plausible way of "pricing" the breaks. Let us define the total length of the line as a sum of all lengths (of words) in the sequence plus the number of words in the line minus one (corresponding to the number of spaces in the line before stretching). Then, one requires the total length of the line (before stretching) to be less then the widest allowed margin, L, and otherwise define the cost to be a monotonically increasing function of the stretching factor, for example

$$c(i, j) = \begin{cases} +\infty, & L < (j - i) + \sum_{t=i}^{j} |w_t| \\ \left(\dfrac{L - (j - i) - \sum_{t=i}^{j} |w_t|}{j - i} \right)^3, & \text{otherwise} \end{cases} \tag{10.2}$$

At first glance the problem of finding the optimal sequence seems hard, that is exponential in the number of words. Indeed, formally one has to make a decision of putting if to place a break (or not) after reading each word in the sequence, thus facing the problem of choosing an optimal sequence from 2^{n-1} of possible options.

Is there a more efficient way of finding the optimal sequence? Apparently answer to this question is affirmative, and in fact, as we will see below the solution is of the dynamical programming type. The key insight is relation between optimal solution of the full problem and an optimal solution of a sub-problem consisting of an early portion of the full paragraph. One discovers that the optimal solution of the sub-problem is a sub-set of the optimal solution of the full problem. This means, in particular, that we can proceed in a greedy manner, looking for an optimal solution sequentially - solving a sequence of sub-problems, where each consecutive problem extends the preceding one incrementally.

Let $f(i)$ denote the minimum cost of formatting a sequence of words which starts from the word i and runs to the end of the paragraph. Then, the minimum cost of the entire paragraph is

$$f(1) = \min_j (c(1, j) + f(j + 1)). \tag{10.3}$$

while a partial cost satisfies the following recursive relation

$$\forall i: \quad f(i) = \min_{j: i \le j} (c(i, j) + f(j + 1)), \tag{10.4}$$

which we also supplement by the boundary condition, $f(n+1) = 0$, stating formally that no word is available for formatting when we reach the end of the paragraph. The last equation is an analog of the Bellman equation, which was discussed in the lecture. A recursive algorithm for $f(i)$ implementing Eq. (10.4) is

```
=====================================
function f(i)
begin:
    if i = n + 1 then return 0;
    f_min := +∞;
    for j = i to n do
        f_min := min(f_min, c(i, j) + f(j + 1));
    return f_min;
end
=====================================
```

Exercise: Modify this algorithm so that it returns the best location j of the next line break.

The algorithm answer the formatting question in a way smarter than naive check mentioned above. However, it is still not efficient, as it recomputes the same values of f many times, thus wasting efforts. For example, the algorithm calculates $f(4)$ whenever it calculates $f(1), f(2), f(3)$. To avoid this unnecessary, one should save the values already calculated, by placing the result just computed into the memory. Then, when we call, compute and store the functions $f(i)$ sequentially. By storing the results we win computing each $f(i)$ only once. Since we have n different values of i and the loop runs through $O(n)$ values of j, the total running time of the algorithm, relaying on the previous values stored, is $O(n^2)$.

The algorithm just discussed is of the Dynamic Programming type, where the name emphasises that you proceed sequentially/dynamically without recourse and not wasting efforts.

10.2 Shortest Path

Let us now discuss another problem. There is a number placed in each cell of a rectangular table, $N \times M$. One starts from the left-up corner and aims to reach the right-down corner. At every step one can move down or right, then "paying a price" equal to

the number written into the cell. What is the minimum amount needed to complete the task?

Solution: You can move to a particular cell (i, j) only from its left $(i-1, j)$ or up $(i, j-1)$ neighbour. Let us solve the following subproblem — find a minimal price $p[i, j]$ of moving to the (i, j) cell. The recursive formula (Bellman equation) is:

$$p[i, j] = \min(p[i-1, j], p[i, j-1]) + a[i, j], \tag{10.5}$$

where $a[i, j]$ is a table of initial numbers. The final answer is an element $p[n, m]$. Note, that you can manually add the first column and row in the table $a[i, j]$, filled with numbers deliberately larger than the content of any cell — this helps as it allows to avoid dealing with the boundary conditions. The resulting algorithm is

```
=====================================
begin:
//initial data:
    read: a[i, j];
//boundary conditions:
    for i = 2 to N do p[i, 0] := ∞;
    for i = 2 to M do p[0, i] := ∞;
    p[1, 0] = p[0, 1] = 0;
//dynamic programming:
    for i = 1 to N do
        for j = 1 to M do
            p[i, j] = min(p[i − 1, j], p[i, j − 1]) + a[i, j];
//answer:
    writeln: p[N, M];
end
=====================================
```

10.3 Markov Decision Process

Let us discuss a noisy version of the problem just discussed — suppose that at each step your greedy decision (on where to move next within the table) is probabilistic, say the suggested move is actually implemented with the probability of only 85%, while with

the probability of 15% you move to the cell on the right from the suggested move (or stay at your current cell if the move is not available).

The task here, again, is to find an optimal strategy. However, now we are discussing the optimality in average. Formally, we optimize the average over noise/uncertainty. This stochastic optimization problem, in discrete space and discrete time belongs to the class of problems called Markov Decision Processes (MDP). For additional information information on the MDP check the following on-line lecture.

10.4 Discrete Time Control

In this section we formulate and discuss the theory behind the dynamic programming and control problems. We start with the most simple control case, which is the finite horizon discrete time deterministic control problem.

The evolution of system is governed by the following equation:

$$x_{t+1} = x_t + f(t, x_t, u_t), \quad t = 0, 1, \ldots, T-1, \tag{10.6}$$

where x_t is vector describing the state of the system (the position in the shortest path problem) and u_t specifies the control or action at time t (where to move next). Note that the present equation describes a noiseless dynamics, but in principle we can add a noise term in the right-hand-side and describe MDP problems. The time T is analog of number of steps in our problem.

Next we should define the cost function:

$$C(x_0, u_{0:T-1}) = \phi(x_T) + \sum_{t=0}^{T-1} R(t, x_t, u_t). \tag{10.7}$$

Here $R(t, x, u)$ is the cost associated with taking action u at time t in the state x (corresponding number in the table in the shortest path game) and $\phi(x_T)$ is the cost (or bonus) for ending the game in the state x_T at time T (in our example we have not such a bonus because we always ends in the same position). The problem of optimal control is to find actions $u_{0:T-1}$ that minimizes $C(x_0, u_{0:T-1})$.

Let us introduce the optimal cost to go:

$$J(t, x_t) = \min_{u_{t:T-1}} \left(\phi(x_T) + \sum_{s=t}^{T-1} R(s, x_s, u_s) \right), \tag{10.8}$$

which solves the optimal control problem from an intermediate time t until the fixed end time T, starting at an arbitrary position x_t. The minimum total cost is given by $J(0, x_0)$.

One can recursively compute $J(t, x)$ from $J(t+1, x)$ for all x in the following way:

$$J(t, x_t) = \min_{u_t} \left(R(t, x_t, u_t) + J(t+1, x_t + f(t, x_t, u_t))\right) \tag{10.9}$$

and the boundary condition is $J(T, x) = \phi(x)$. Note that the minimization over the whole path $u_{0:T-1}$ has reduced to a sequence of minimizations over u_t. This simplification is due to the Markovian nature of the problem: the future depends on the past and vice versa only through the present.

The algorithm to compute the optimal control $u^*_{0:T-1}$, the optimal trajectory $x^*_{1:T}$ and the optimal cost is given by

==========

1. Initialization: $J(T, x) = \phi(x)$
2. Backwards: For $t = T-1, \ldots, 0$ and for all x compute
   $u^*_t(x) = \arg\min_u \{R(t, x, u) + J(t+1, x + f(t, x, u))\}$
   $J(t, x) = R(t, x, u^*_t) + J(t+1, x + f(t, x, u^*_t))$
3. Forwards: For $t = 0, \ldots, T-1$ compute
   $x^*_{t+1} = x^*_t + f(t, x^*_t, u^*_t(x^*_t))$

==========

The execution of the dynamic programming algorithm is linear in the horizon time T and linear in the size of the state and action spaces.

10.5 Continuous Time Control

Next let us discuss generalization of the problem to the case of continuous time. The dynamic equation becomes

$$x_{t+dt} = x_t + f(x_t, u_t, t)dt, \tag{10.10}$$

where $x_t = x(t)$. The initial state is fixed, $x(0) = x_0$, and the final state is free. The problem is to find the control signal $u(t)$, $0 < t < T$, such that the cost function

$$C(x_0, u(0 \to T)) = \phi(x_T) + \int_0^T d\tau\, R(x(\tau), u(\tau), \tau) \tag{10.11}$$

is minimal. The recursive equation (10.9) becomes

$$\begin{aligned} J(t, x) &= \min_u \left(R(t, x, u)dt + J(t + dt, x + f(t, x, u)dt)\right) \\ &\approx \min_u \left(R(t, x, u)dt + J(t, x) + \partial_t J(t, x)dt + \partial_x J(t, x) f(x, u, t)dt\right), \end{aligned} \tag{10.12}$$

where in the last line we have used the Taylor expansion of $J(t + dt, x + dx)$ to the first order in dt and dx. Finally, we obtain

$$-\partial_t J(t, x) = \min_u \left(R(t, x, u) + f(x, u, t)\partial_x J(t, x)\right), \tag{10.13}$$

which is known as Hamilton-Jacobi-Bellman equation, that describes the evolution of J as a function of x and t and must be solved with boundary condition $J(x, T) = \phi(x)$.

The optimal control at the current $x,\ t$ is given by

$$u(x, t) = \arg\min_u \left(R(t, x, u) + \partial_x J(t, x) f(x, u, t)\right). \tag{10.14}$$

Note that in order to compute the optimal control at the current state $x(0)$ at $t = 0$ one must compute $J(x, t)$ for all values of x and t.

10.6 Mass on a Spring

To illustrate introduced concepts consider a mass on a spring. The equation of motion is given by

$$\ddot{z} = -z + u, \tag{10.15}$$

where $-z$ is restoring force ($k = 1$), $\ddot{z}$ is acceleration ($m = 1$), and u is unspecified control signal with $-1 \leq u \leq 1$. We want to solve the control problem: given initial position and velocity at time $t = 0$, find the control path $u(0 \to T)$ such that $z(T)$ is maximal.

The state vector $x = (x_1, x_2)^T$ is two-dimensional, where $x_1 = z$ and $x_2 = \dot{z}$, and then

$$\dot{x} = Ax + Bu, \quad A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{10.16}$$

The problem is of the above type, with $f(x, u, t) = Ax + Bu$, $\phi(x) = C^T x$, $C^T = (-1, 0)$ and $R(x, u, t) = 0$.

The Hamilton-Jacobi-Bellman equation reads as

$$-\partial_t J = \min_u \left[(\partial_x J)^T (Ax + Bu)\right] = (\partial_x J)^T Ax - |(\partial_x J)^T B|, \tag{10.17}$$

or we can rewrite it in the following form:

$$-\partial_t J = \dot{z}\partial_z J - z\partial_{\dot{z}} J - |\partial_{\dot{z}} J|. \tag{10.18}$$

We try $J(t,x) = \psi(t)^T x + \alpha(t) = \alpha(t) + \psi_1(t)z + \psi_2(t)\dot{z}$ and then

$$\partial_t \psi_1 = \psi_2, \quad \partial_t \psi_2 = -\psi_1, \quad \partial_t \alpha = |\psi_2|. \tag{10.19}$$

These equations must be solved for all t with final boundary conditions $\psi_1(T) = -1$, $\psi_2(T) = 0$ and $\alpha(T) = 0$. Note that the optimal control (10.14) only requires $\partial_x J(x,t)$, which in this case is $\psi(t)$ and thus we do not need to solve the equation for α. The solution for ψ is

$$\psi_1(t) = -\cos(t-T), \quad \psi_2(t) = \sin(t-T), \quad 0 \le t \le T, \tag{10.20}$$

and the optimal control is

$$u(x,t) = \arg\min_u (\partial_x J(t,x) f(x,u,t)) = \arg\min_u (\psi_1 \dot{z} - \psi_2 z + \psi_2 u) = -\operatorname{sign} \psi_2(t). \tag{10.21}$$

As an example consider $x_1(0) = x_2(0) = 0$, $T = 2\pi$. Then the optimal control is

$$u = -1, \quad 0 < t < \pi, \tag{10.22}$$

$$u = 1, \quad \pi < t < 2\pi, \tag{10.23}$$

and using the equation (10.16) we can calculate the optimal trajectories:

$$x_1 = \cos t - 1, \quad x_2 = -\sin t, \qquad 0 < t < \pi, \tag{10.24}$$

$$x_1 = 3\cos t + 1, \quad x_2 = -3\sin t, \qquad \pi < t < 2\pi. \tag{10.25}$$

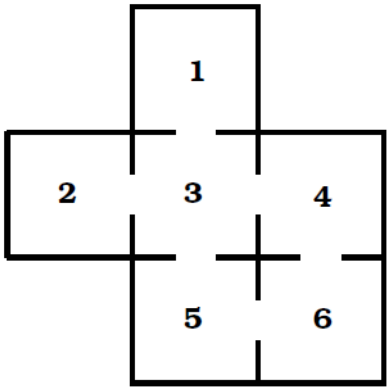
We see that in order to excite the spring to its maximal length at T, first we should push the spring down for $0 < t < \pi$ and then to push the spring up between $\pi < t < 2\pi$, taking maximally advantage of the intrinsic dynamics of the spring.

Note that since there is no cost associated with the control u and u is limited between -1 and 1, the optimal control is always either -1 or 1. This is known as bang-bang control.

10.7 Problems

Problem 1. *Labyrinth with a mousetrap.*

A mouse lives in the labyrinth shown below. At each time step the mouse chooses at random one of the doors and leaves the room trough this door. The process repeats. Formally, mouse dynamics is described fully by a Markov chain of transitions between six states.



(i) Write down the transition matrix P for this Markov chain. Is it irreducible, aperiodic, ergodic?

By definition, the stationary distribution π^* is an eigenvector of P, which corresponds to the eigenvalue $\lambda = 1$, i.e. it satisfies the equation $P\pi^* = \pi^*$.

(ii) Find the stationary distribution. Does the detailed balance hold?

Now suppose that initially at $t = 0$ the mouse was in the room 1.

(iii) What is the probability to find the mouse in the room 5 in 4 steps? In 5 steps?

(iv) Do the probabilities of finding the mouse in different rooms converge to the stationary distribution π^*?

Suppose one places a mousetrap in room 5 when the mouse is in room 1.

(v) Find the expected number of steps leading the mouse to the trap, i.e. the expected number of steps till the mouse enters the room 5 for the first time.

Hint: One (of many) ways of answering (v) is to consider the function $p(i)$ — the expected number of steps leading to the trap given that the mice is in the room i, and attempt to relate $p(i)$ with different i to each other. The resulting system of equations will be akin to the Bellman equations describing theory behind the dynamic programming.

Chapter 11

Inference and Learning over Trees

Pair-wise graphical model represents a set of random variables and their conditional dependencies via a graph: nodes correspond to variables and edges represent conditional dependencies. The random variable is conditionally independent of all other variables given its neighbors. That models are widely used in statistics and machine learning.

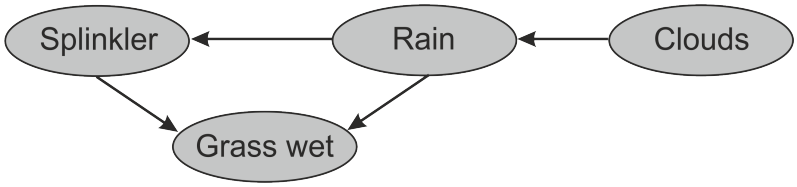


Figure 11.1: A simple example of directed graphical model. Clouds lead to rain, rain influences whether the sprinkler is activated, and both rain and the sprinkler influence whether the grass is wet.

11.1 Ising Tree Model

The main focus of this recitation is on the undirected graphical models. We assume that if two nodes are connected then they influence each other. A simple example of an undirected model is given by the Ising model representing ensemble of spins with pair-wise interaction. The probability of a given state of the system is

$$p(\sigma) = \frac{1}{Z} \exp\left(-\beta E(\sigma)\right) \tag{11.1}$$

where the energy of the state, σ, is

$$E(\sigma) = -\frac{1}{2}\sum_{(i,j)\in\mathcal{E}} \sigma_i J_{ij}\sigma_j + \sum_{i\in\mathcal{V}} h_i\sigma_i \tag{11.2}$$

and the so-called partition function (normalization factor) is

$$Z = \sum_{\sigma} \exp\left(-\beta E(\sigma)\right). \tag{11.3}$$

The undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ describes the interaction pattern of spins.

For the system of n spins, the number of possible states is 2^n, and thus computational efforts for computing the partition function, (11.3), is at least 2^n (naive enumeration). However, one hopes to improve this worst case estimate for complexity, by utilizing structure of the graph. One idea is to explore memory, hopefully avoiding many repetitive computations obviously characteristic of the naive approach.

It turns out that the problem can be solved in $O(n)$, i.e. in the number of steps scaling linearly with the number of spins, in the case when the interaction/factor graph $\mathcal{G}$ is a tree – that is a graph containing no loops. Let us illustrate the main idea of the approach on the exemplary tree graphs shown in Fig. 11.2.

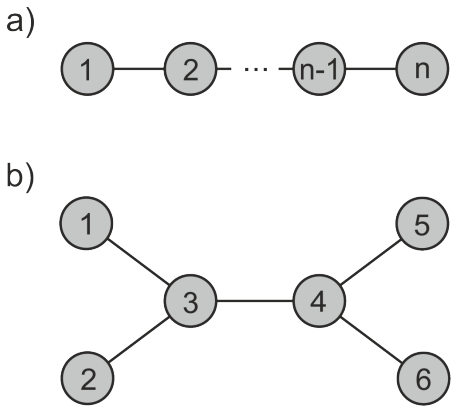


Figure 11.2: Exemplary interaction/factor graphs which are tree.

For a linear chain of n spins shown in Fig. 11.2a, the partition function is

$$Z = \sum_{\sigma_n} Z(\sigma_n), \tag{11.4}$$

where $Z(\sigma_n)$ is the newly introduced object representing sum over all but last spin in the chain, labeled by n. Z_n can be expressed as follows

$$Z(\sigma_n) = \sum_{\sigma_{n-1}} \exp(J_{n,n-1}\sigma_n\sigma_{n-1} + h_n\sigma_n) Z_{(n-1)\to(n)}(\sigma_{n-1}), \tag{11.5}$$

where $Z_{(n-1)\to(n)}(\sigma_i)$ is the partial partition function for the subtree (a shorter chain in this case) rooted at $n-1$ and built excluding the branch/link directed towards n. The newly introduced partially summed partition function contains summation over one less spins then the original chain. In fact, this partially sum object can be defined recursively

$$Z_{(i-1)\to(i)}(\sigma_{i-1}) = \sum_{\sigma_{i-2}} \exp(J_{i-1,i-2}\sigma_{i-1}\sigma_{i-2} + h_{i-1}\sigma_{i-1}) Z_{(i-2)\to(i-1)}(\sigma_{i-2}) \tag{11.6}$$

that is expressing one partially sum object via the partially sum object computed on the previous step. Advantage of this recursive approach is obvious – it allows to replace summation over the exponentially many spin configurations by summing up of only two terms at each step of the recursion. This is the essence of the method known as the Dynamic Programming. In fact we will get a special recitation (# 12) devoted primarily to this important method.

The approach just explained can be generalized from the case of the linear chain to the case of a general tree. Then, in the general case $Z(\sigma_i)$ is the partition function of the whole tree with a fixed value of the spin variable at the site/node i. Next one derives

$$Z(\sigma_i) = e^{h_i\sigma_i} \prod_{j\in\partial i} \left(\sum_{\sigma_j} e^{J_{ij}\sigma_i\sigma_j} Z_{j\to i}(\sigma_j) \right), \tag{11.7}$$

where ∂i denotes the set of neighbors of the i-th spin and

$$Z_{j\to i}(\sigma_j) = e^{h_j\sigma_j} \prod_{k\in\partial j\setminus i} \left(\sum_{\sigma_k} e^{J_{kj}\sigma_k\sigma_j} Z_{k\to j}(\sigma_j) \right) \tag{11.8}$$

is the partition function of the subtree rooted at the node j.

Let us illustrate the general scheme on example of the tree shown in Fig. 11.2b, one obtains

$$Z = \sum_{\sigma_4} Z(\sigma_4), \tag{11.9}$$

The partition function, partially summed and conditioned to the spin value at the spin, σ_4, is

$$Z(\sigma_4) = e^{h_4\sigma_4} \sum_{\sigma_5} e^{J_{45}\sigma_4\sigma_5} Z_{5\to4}(\sigma_5) \sum_{\sigma_6} e^{J_{46}\sigma_4\sigma_6} Z_{6\to4}(\sigma_6) \sum_{\sigma_3} e^{J_{34}\sigma_3\sigma_4} Z_{3\to4}(\sigma_3) \tag{11.10}$$

where

$$Z_{3\to4}(\sigma_3) = e^{h_3\sigma_3} \sum_{\sigma_1} e^{J_{13}\sigma_1\sigma_3} Z_{1\to3}(\sigma_1) \sum_{\sigma_2} e^{J_{23}\sigma_2\sigma_3} Z_{2\to3}(\sigma_2). \tag{11.11}$$

Exercise 1. Demonstrate that the ith spin is conditionally independent of all other spins given its neighbors, i.e.

$$p(\sigma_i|\sigma/\sigma_i) = p(\sigma_i|\sigma_j \sim \sigma_i), \tag{11.12}$$

where, $p(\sigma_i|\sigma/\sigma_i)$, is the probability distribution of the ith spin conditioned to the values of all other spins, and, $p(\sigma_i|\sigma_j \sim \sigma_i)$, is the probability distribution of ith spin conditioned to the spin values of its neighbors.

11.2 Properties of Undirected Tree-Structured Graphical Models

It appears that in the case of a general pair-wise graphical model over trees the joint distribution function over all variables can be expressed solely via single-node marginals and pair-vise marginals over all pairs of the graph-neighbors. To illustrate this important factorization property, let us consider examples shown in Fig. 11.3. In the case of the two-nodes example of Fig. 11.3a the statement is obvious as following directly from the Bayes formula

$$P(x_1, x_2) = P(x_1)P(x_2|x_1), \tag{11.13}$$

or, equivalently, $P(x_1, x_2) = P(x_2)P(x_1|x_2)$.

For the pair-wise graphical model shown in Fig. 11.3b one obtains

$$\begin{aligned} P(x_1, x_2, x_3) &= P(x_1, x_2)P(x_3|x_1, x_3) = P(x_1, x_2)P(x_3|x_2) = \\ &= P(x_1)P(x_2|x_1)P(x_3|x_2) = \frac{P(x_1, x_2)P(x_2, x_3)}{P(x_2)}, \end{aligned} \tag{11.14}$$

where the conditional independence of x_3 on x_1, $P(x_3|x_1, x_2) = P(x_3|x_2)$, was used.

Next, let us work it out on the example of the pair-wise graphical model shown in Fig. 11.3

$$\begin{aligned} P(x_1, x_2, x_3, x_4) &= P(x_1, x_2, x_3)P(x_4|x_1, x_3, x_4) = P(x_1, x_2, x_3)P(x_4|x_2) = \\ &= P(x_1, x_2)P(x_3|x_1, x_2)P(x_4|x_2) = P(x_1, x_2)P(x_3|x_2)P(x_4|x_2) = \\ &= P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_2) = \frac{P(x_1, x_2)P(x_2, x_3)P(x_2, x_4)}{P^2(x_2)}. \end{aligned} \tag{11.15}$$

Here one uses the following reductions, $P(x_4|x_1, x_3, x_4) = P(x_4|x_2)$ and $P(x_3|x_1, x_2) = P(x_3|x_2)$, related to respective independence properties.

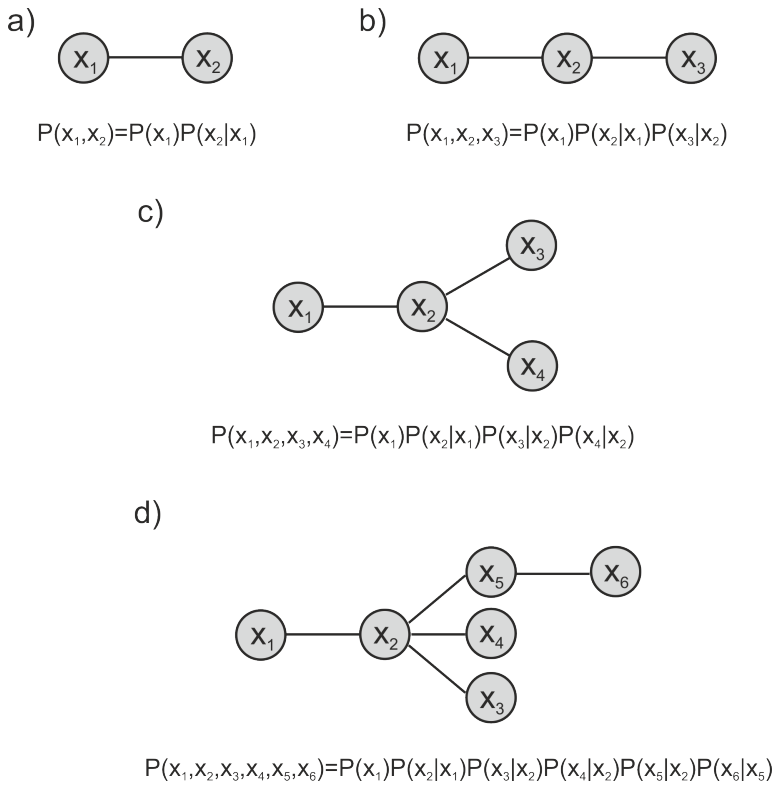


Figure 11.3: Examples of undirected tree-structured graphical models.

Finally, it is easy to verify that the joint probability distribution corresponding to the model in Fig. 11.3d is

$$P(x_1, x_2, x_3, x_4, x_5, x_6) = P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_2)P(x_5|x_2)P(x_6|x_5) = \\ = \frac{P(x_1, x_2)P(x_2, x_3)P(x_2, x_4)P(x_2, x_5)P(x_5, x_6)}{P^3(x_2)P(x_5)}. \quad (11.16)$$

In general, the joint probability distribution of a tree-like graphical model can be written as follows

$$P(x_1, x_2, \ldots, x_n) = \frac{\prod_{(i,j)\in\mathcal{E}} P(x_i, x_j)}{\prod_{i\in\mathcal{V}} P^{q_i-1}(x_i)}, \quad (11.17)$$

where q_i is the degree of the ith node. Eq. (11.17) can be proven by induction.

11.3 Learning on Tree

Eq. (11.17) suggests that knowing the structure of the tree-based graphical model allows to express the joint probability distribution in terms of the single-(node) and pairwise (edge-related) marginals. Below we will utilize this statement to solve an inverse problem.

Specifically, we attempt to reconstruct a tree representing correlations between multiple (ideally, infinitely many) snapshots of the discrete random variables $x_1, x_2, \ldots, x_n$?

A straightforward strategy to achieve this goal is as follows. First, one estimates all possible single-node and pairwise marginal probability distributions, $P(x_i)$ and $P(x_i, x_j)$, from the infinite set of the snapshots. Then, we may similarly estimate the joint distribution function and verify for a possible tree layout if the relations (11.17) hold. However, this strategy is not feasible as requiring (in the worst unlucky case) to test exponentially many, n^{n-2}, possible spanning threes. Luckily a smart and computationally efficient way of solving the problem was suggested by Chow and Liu in 1968.

Consider the probability distribution

$$P_F(x_1, x_2, \ldots, x_n) = \frac{\prod_{(i,j)\in\mathcal{E}^F} P(x_i, x_j)}{\prod_{i\in\mathcal{V}^F} P^{q_i-1}(x_i)}, \tag{11.18}$$

associated with a three-structured graph F. "Distance" between correct probability distribution P and the candidate probability distribution, P_F, can be measured in terms of the Kublack-Leibler divergence

$$D(P \parallel P_F) = -\sum_{\vec{x}} P(\vec{x}) \log_2 \frac{P(\vec{x})}{P_F(\vec{x})} \tag{11.19}$$

This measure is always positive if P and P_F are different, and is zero if these distributions are identical. Then, we are looking for a tree that minimizes the Kublack-Leibler divergence.

Substituting (11.18) into Eq. (11.19) one arrives at the following chain of explicit transformations

$$\begin{aligned}
&\sum_{\vec{x}} \mathcal{P}(\vec{x}) \left(\log \mathcal{P}(\vec{x}) - \sum_{(i,j)\in\mathcal{E}^\mathcal{F}} \log \mathcal{P}(x_i, x_j) + \sum_{i\in\mathcal{V}^\mathcal{F}} (q_i - 1) \log \mathcal{P}(x_i) \right) = \\
&= \sum_{\vec{x}} \mathcal{P}(\vec{x}) \log \mathcal{P}(\vec{x}) - \sum_{(i,j)\in\mathcal{E}^\mathcal{F}} \sum_{x_i, x_j} \mathcal{P}(x_i, x_j) \log \mathcal{P}(x_i, x_j) + \\
&+ \sum_{i\in\mathcal{V}^\mathcal{F}} (q_i - 1) \sum_{x_i} \mathcal{P}(x_i) \log \mathcal{P}(x_i) = -\sum_{(i,j)\in\mathcal{E}^\mathcal{F}} \sum_{x_i, x_j} \mathcal{P}(x_i, x_j) \log \frac{\mathcal{P}(x_i, x_j)}{\mathcal{P}(x_i)\mathcal{P}(x_j)} + \\
&+ \sum_{\vec{x}} \mathcal{P}(\vec{x}) \log \mathcal{P}(\vec{x}) - \sum_{i\in\mathcal{V}^\mathcal{F}} \sum_{x_i} \mathcal{P}(x_i) \log \mathcal{P}(x_i),
\end{aligned} \tag{11.20}$$

where the following nodal and edge marginalization relations were used, $\forall i \in \mathcal{V}^\mathcal{F}$: $\mathcal{P}(x_i) = \sum_{\vec{x}\setminus x_i} \mathcal{P}(\vec{x})$, and, $\forall (i,j) \in \mathcal{E}^\mathcal{F}$: $\mathcal{P}(x_i, x_j) = \sum_{\vec{x}\setminus x_i, x_j} \mathcal{P}(\vec{x})$, respectively. One

observes that the Kublack-Leibler divergence becomes

$$D(P \parallel P_F) = -\sum_{(i,j)\in\mathcal{E}^{\mathcal{F}}} I(x_i, x_j) + \sum_{i\in\mathcal{V}^{\mathcal{F}}} S(x_i) - S(\vec{x}), \tag{11.21}$$

where

$$I(x_i, x_j) = \sum_{x_i, x_j} P(x_i, x_j) \log_2 \frac{P(x_i, x_j)}{P(x_i)P(x_j)} \tag{11.22}$$

is the mutual information of the pair of random variables x_i and x_j.

Since the entropies $S(x_i)$ and $S(X)$ do not depend on the tree choice, minimizing the Kublack-Leibler divergence is equivalent to maximizing the following sum over branches of a tree

$$\sum_{(i,j)\in\mathcal{E}^{\mathcal{F}}} I(x_i, x_j). \tag{11.23}$$

Based on this observation, Chow and Liu have proposed a simple algorithm for tree reconstruction: at each stage of the procedure one should simply adds the maximum mutual information pair. For the set of n random variables, one ought to consider $n(n-1)/2$ possible branches. Let us index the branches according to decreasing weights $I(x_i, x_j)$ so that the weight of b_α is greater than or equal to the weight of b_β whenever $\alpha < \beta$. We then start by selecting b_1 and b_2 and add b_3 if b_3 does not form a cycle with b_1 and b_3. The process continues, where selection a branch occurs whenever the newly picked branch does not form a cycle with the set of previously selected. Otherwise, selection of the branch is rejected. This procedure produces a unique solution if the branch weights are all different (no degeneracy). Multiple solutions are possible in the degenerate case, however all the solutions show the same maximum weight.

11.4 Approximation

Eq. (11.17) is exact only in the case when it is guaranteed that the graphical model we attempt to recover forms a tree. However, the same tree ansatz can be used to recover the best tree approximation for a graphical model defined over a graph with loops. How to choose the optimal (best approximation) tree in this case? To answer this question within the aforementioned Kublack-Leibler paradigm one needs to compare the tree ansatz (11.17) and the empirical joint distribution. This reconstruction of the optimal tree is based on the Chow-Liu algorithm.

Exercise 2

Find the optimum tree approximation of a fourth-order binary distribution $P(x_1, x_2, x_3, x_4)$ listed in Table 1

Table 11.1: A binary probability distribution $P(x_1, x_2, x_3, x_4)$ in comparison with the tree approximation and the approximation based on the independence assumption.

$x_1x_2x_3x_4$	$P(x_1, x_2, x_3, x_4)$	$P(x_1)P(x_2\|x_1)P(x_3\|x_2)P(x_4\|x_1)$	$P(x_1)P(x_2)P(x_3)P(x_4)$
0000	0.100	0.130	0.046
0001	0.100	0.104	0.046
0010	0.050	0.037	0.056
0011	0.050	0.030	0.056
0100	0.000	0.015	0.056
0101	0.000	0.012	0.056
0110	0.100	0.068	0.068
0111	0.050	0.054	0.068
1000	0.050	0.053	0.056
1001	0.100	0.064	0.056
1010	0.000	0.015	0.068
1011	0.000	0.018	0.068
1100	0.050	0.033	0.068
1101	0.050	0.040	0.068
1110	0.150	0.149	0.083
1111	0.150	0.178	0.083

Solution:

To recover the best approximation tree it is sufficient to estimate from the data (thus empirical) mutual information between all pairs of random variables

$$I(x_1, x_2) = 0.079 \tag{11.24}$$

$$I(x_1, x_3) = 0.00005 \tag{11.25}$$

$$I(x_1, x_4) = 0.0051 \tag{11.26}$$

$$I(x_2, x_3) = 0.189 \tag{11.27}$$

$$I(x_2, x_4) = 0.0051 \tag{11.28}$$

$$I(x_3, x_4) = 0,0051 \tag{11.29}$$

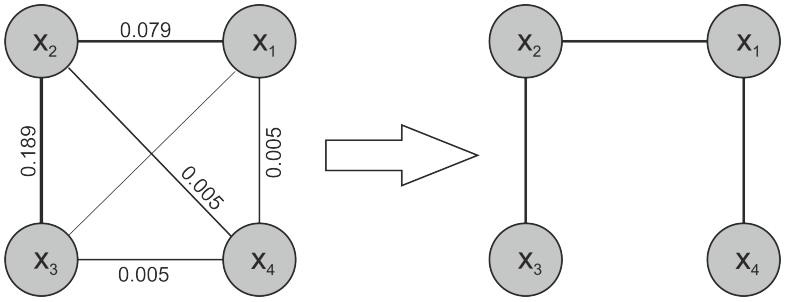


Figure 11.4: Graphical model corresponding to the fourth-order binary distribution

Since $I(x_2, x_3)$ and $I(x_1, x_2)$ are the two largest quantities, (x_2, x_3) and (x_1, x_2) constitute the first two branches of the optimum tree. To select the next branch note that $I(x_1, x_4) = I(x_2, x_4) = I(x_3, x_4)$. To brake the degeneracy one would pick arbitrarily any one of these three branches. The resulting tree provides optimal decomposition of the joint probability into product of the single-node and pair-wise conditional probabilities. For the purpose of comparison, the approximant done under assumption of statistical independence is also provided. One observes that the optimum tree approximants are closer to the true distribution. Indeed, the Kublack-Lebler measure of proximity (distance), $D(P \parallel P_{\text{approx}})$, is 0.094 for the best tree approximation in the contrast with 0.364 for the best independent-variable anzatz.

References

[1] Einstein, Albert, 1905, "On the movement of small particles suspended in a stationary liquid demanded by the molecular-kinetic theory of heart," Annalen der physik **17**, 549–560

[2] Kappen, Hilbert J, 2011, "Optimal control theory and the linear bellman equation," Inference and Learning in Dynamic Models, 363–387

[3] Kelly, Frank, and Elena Yudovina, 2014, *Stochastic network* (Cambridge University Press.)

[4] Langevin, Paul, 1908, "Sur la théorie du mouvement brownien," CR Acad. Sci. Paris **146**, 530

[5] Moore, Cristopher, and Stephan Mertens, 2011, *The nature of computation* (OUP Oxford)

[6] Nelson, Barry L, 2012, *Stochastic modeling: analysis and simulation* (Courier Corporation)

[7] Redner, Sidney, 2001, *A guide to first-passage processes* (Cambridge University Press)

[8] Schmitt, Florian, Florian Schmitt, Franz Rothlauf, and Franz Rothlauf, 2001, *On the Importance of the Second Largest Eigenvalue on the Convergence Rate of Genetic Algorithms*, Tech. Rep. (Proceedings of the 14th Symposium on Reliable Distributed Systems)

[9] Turitsyn, Konstantin S, Michael Chertkov, and Marija Vucelja, 2011, "Irreversible monte carlo algorithms for efficient sampling," Physica D: Nonlinear Phenomena **240**, 410–414

[10] Von Smoluchowski, Marian, 1906, “Zur kinetischen theorie der brownschen molekularbewegung und der suspensionen,” Annalen der physik **326**, 756–780