Stochastic Modeling and Computations

Recitation Notes
written by

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

Skolkovo Institute of Science and Technology

17 April 2017

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, law of large numbers, Markov processes, reversibility, sampling, etc., supplemented by practical/computational implementations of basic algorithms. In the second part of the course, the focus will shift from general concepts and algorithms per se to their applications in science and engineering with examples, aiming to illustrate the models and make the methods of solution clear, from physics, chemistry, machine learning, control and operations research.

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 $\sigma^2 = \mu_2 - \mu_1^2 = pq$.

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

$$p(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots, \quad \lambda > 0. \tag{1.10}$$

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

$$\mu_1 = \sum_{k=0}^{\infty} \frac{k\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} e^{-\lambda} = \lambda \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} e^{-\lambda} = \lambda. \tag{1.11}$$

The second moment is

$$\mu_2 = \sum_{k=0}^{\infty} \frac{k^2\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} \frac{k\lambda^k}{(k-1)!} e^{-\lambda} = \lambda \sum_{n=0}^{\infty} \frac{(n+1)\lambda^n}{n!} e^{-\lambda} = \lambda(\lambda+1), \tag{1.12}$$

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.

The most important continuous distribution is ***Gaussian Distribution***. We will discuss its properties in the chapter 2.

For now 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 line-width). The probability density function is given by expression (check that it is properly normilized)

$$p(x) = \frac{1}{\pi} \frac{\gamma}{(x-a)^2 + \gamma^2}, \quad -\infty < x < +\infty. \tag{1.13}$$

The first moment is

$$\mu_1 = \frac{\gamma}{\pi} \int_{-\infty}^{+\infty} \frac{x dx}{(x-a)^2 + \gamma^2} = a, \tag{1.14}$$

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 $C > 0$:

$$P(X \geq C) \leq \frac{\mathbb{E}[X]}{C}, \tag{1.15}$$

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:

$$P(|X - \mathbb{E}[X]| \geq C) \leq \frac{\sigma^2}{C^2}. \tag{1.16}$$

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:

$$P(\text{number of coupons, still missing} \geq 1) \leq n(1 - 1/n)^t \leq ne^{-t/n}, \tag{1.17}$$

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

1.4 Characteristic Function

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

$$G(k) = \langle e^{ikX} \rangle = \int_{-\infty}^{+\infty} e^{ikx} p(x) dx. \tag{1.18}$$

It exists for all real k and obeys relations

$$G(0) = 1, \quad |G(k)| \le 1. \tag{1.19}$$

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:

$$G(k) = \sum_{m=0}^{\infty} \frac{(ik)^m}{m!} \langle X^m \rangle, \tag{1.20}$$

and thus

$$\langle X^m \rangle = \frac{1}{i^m} \frac{\partial^m}{\partial k^m} G(k) \Big|_{k=0}. \tag{1.21}$$

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. (1.18) one derives

$$G(k) = 1 - p + pe^{ik}, \tag{1.22}$$

and thus

$$\mu_m = \frac{\partial^m}{\partial (ik)^m} [1 - p + pe^{ik}] \Big|_{k=0} = p. \tag{1.23}$$

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

$$\ln G(k) = \sum_{m=1}^{\infty} \frac{(ik)^m}{m!} \kappa_m. \tag{1.24}$$

According to Eq. (1.19) this Taylor series start from unity. Utilizing Eqs. (1.20) and (1.24), one derives the following relations between the cumulants and the moments

$$\kappa_1 = \mu_1, \tag{1.25}$$

$$\kappa_2 = \mu_2 - \mu_1^2 = \sigma^2. \tag{1.26}$$

The procedure naturally extends to higher order moments and cumulants.

Now, consider an example of the Poisson distribution defined according to (1.10). The respective characteristic function is

$$G(p) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} e^{ipk} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^{ip})^k}{k!} = exp\left[\lambda(e^{ip} - 1)\right], \tag{1.27}$$

and then

$$\ln G(p) = \lambda(e^{ip} - 1). \tag{1.28}$$

Next, using the definition (1.24), 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

$$p(s) = \frac{1}{Z} e^{-\beta E(s)}, \quad Z = \sum_s e^{-\beta E(s)}, \tag{1.29}$$

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

$$\langle E \rangle = \sum_s p(s) E(s) = \frac{1}{Z} \sum_s E(s) e^{-\beta E(s)} = -\frac{1}{Z} \frac{\partial Z}{\partial \beta} = -\frac{\partial \ln Z}{\partial \beta}. \tag{1.30}$$

The variance of the energy (energy fluctuations) is

$$\Delta E^2 = \langle (E - \langle E \rangle)^2 \rangle = \frac{\partial^2 \ln Z}{\partial \beta^2}, \tag{1.31}$$

(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

$$p(x) = \begin{cases} Ae^{-\lambda x}, & x \geq 0, \\ 0, & x < 0, \end{cases} \tag{1.32}$$

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

$$G(k) = \int_{-\infty}^{+\infty} e^{ikx} p(x) dx. \tag{1.33}$$

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 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.

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

$$p_m = \prod_{i=1}^{m} \left(1 - \frac{i-1}{366}\right). \tag{1.34}$$

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 (1.18) of the Cauchy distribution (1.13). 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

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

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

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

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

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.

In fact all expressions discussed so far (in the context of the multivariate Gaiussian distributions are Gaussian, expressions only in terms of the second moments

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

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

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.

Chapter 3

Entropy. Mutual Information. Probabilistic Inequalities

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

3.1 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

$$s(x_i) = -\log_2 P(x_i). \tag{3.1}$$

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

$$S(X) = \mathbf{E}[s(x)] = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i). \tag{3.2}$$

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.

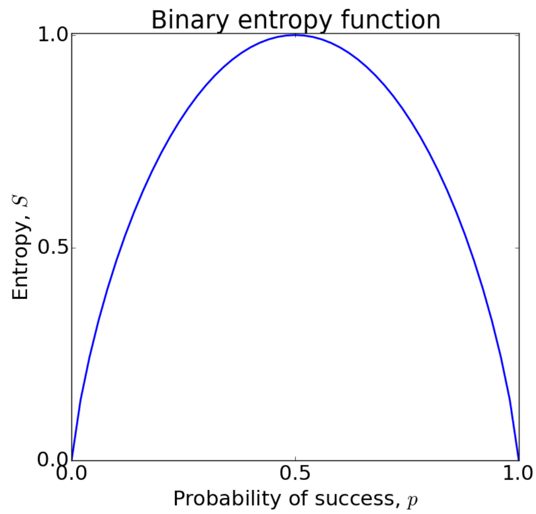


Figure 3.1: Entropy of the Bernoulli distribution as a function of the success rate, p.

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 (3.2)

$$S_{\text{binary}}(p) = -p \log_2 p - (1 - p) \log_2(1 - p) \tag{3.3}$$

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

$$S(X, Y) = -\sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} P(x_i, y_j) \log_2 P(x_i, y_j). \tag{3.4}$$

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

$$S(Y|X) = \sum_{i=1}^{n_X} P(x_i) S(Y|x_i) = -\sum_{i=1}^{n_X} P(x_i) \sum_{j=1}^{n_Y} P(y_j|x_i) \log_2 P(y_j|x_i) =$$
$$= -\sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} P(x_i, y_j) \log_2 P(y_j|x_i) = -\sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} P(x_i, y_j) \log_2 \frac{P(x_i, y_j)}{P(x_i)}. \tag{3.5}$$

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

$$\mathbf{E}[f(u)] \geq f[\mathbf{E}(u)]. \tag{3.6}$$

Let us define

$$f(u) = -\log_2 u, \quad u = 1/P(x). \tag{3.7}$$

Obviously, $f(u)$ is convex. Accordingly to (3.6) one obtains

$$\mathbf{E}[\log_2 P(x)] \geq -\log_2 \mathbf{E}[1/P(x)], \tag{3.8}$$

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

$$S(X|Y) \leq S(X) \text{ with equality iff } X \text{and } Y \text{are independent}, \tag{3.9}$$
$$S(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} S(X_i) \text{ with equality iff } X_i \text{ are independent.} \tag{3.10}$$

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

$$p_n = \begin{cases} \frac{0.1}{n}, & \text{for } n \in 1, \ldots, 12367 \\ 0, & \text{for } n > 12367 \end{cases} \tag{3.11}$$

Under an assumption that English documents are generated by picking words at random according to Eq. (3.11) compute the entropy of the made-up English (per word).

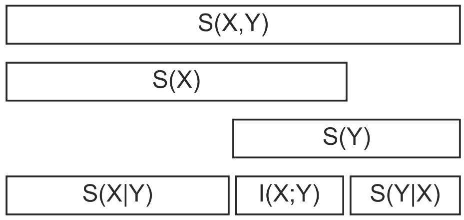


Figure 3.2: Illustration of the relations between joint entropy, marginal entropies, conditional entropies and mutual information.

Solution:

Substituting the distribution (3.11) into Eq. (3.2) one derives

$$S = -\sum_{i=1}^{12367} \frac{0.1}{n} \log_2 \frac{0.1}{n} \approx \frac{0.1}{\ln 2} \int_{10}^{123670} \frac{\ln x}{x} dx = \tag{3.12}$$

$$= \frac{1}{20 \ln 2} (\ln^2 123670 - \ln^2 10) \approx 9.9 \text{ bits.} \tag{3.13}$$

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 ~ 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.

3.2 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

$$I(X; Y) = \mathbf{E}_{P(x,y)} \left[\log_2 \frac{P(x, y)}{P(x)P(y)} \right] = \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} P(x_i, y_j) \log_2 \frac{P(x_i, y_j)}{P(x_j)P(y_j)}. \tag{3.14}$$

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

$$I(X; Y) = S(X) - S(X|Y) = S(Y) - S(Y|X) = S(X) + S(Y) - S(X, Y). \tag{3.15}$$

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 3.1: 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 3.1. 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

$$P(x_i) = \sum_{j=1}^{4} P(x_i, y_j). \tag{3.16}$$

The marginal probabilities $P(x)$ and $P(y)$ are described in the Table 3.1.

Next, the single-valued marginal entropies become

$$S(X) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{8}\log_2\frac{1}{8} = \frac{7}{4}\text{ bits}, \tag{3.17}$$

$$S(Y) = -4\times\frac{1}{4}\log_2\frac{1}{4} = 2\text{ bits}. \tag{3.18}$$

$P(x\|y)$		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 3.2: Conditional probability function $P(x|y)$ for the case discussed in the exercise # 3.

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

$$P(x|y) = \frac{P(x, y)}{P(y)}, \tag{3.19}$$

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

$$S(X|y = y_i) = -\sum_{j=1}^{4} P(x_j|y_i) \log_2 P(x_j|y_i). \tag{3.20}$$

The results are also presented in the Table 3.2.

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

$$S(X|Y) = \sum_{i=1}^{4} P(y_i) S(X|y = y_i) = \frac{11}{8} \text{ bits}, \tag{3.21}$$

and the mutual information

$$I(X; Y) = S(X) - S(X|Y) = \frac{7}{4} - \frac{11}{8} = \frac{3}{8} \text{ bits}. \tag{3.22}$$

3.3 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 $P(x)$.

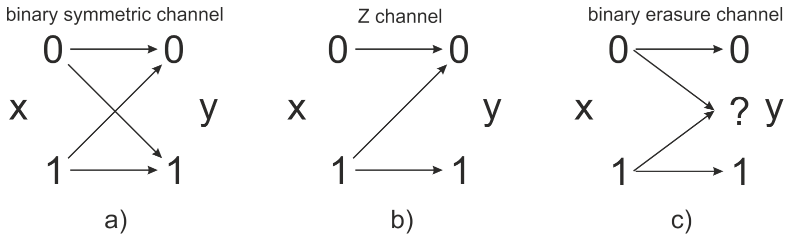


Figure 3.3: Examples of communication channels.

The capacity of a channel Q is defined as

$$C(Q) = \max_{P(X)} I(X;Y). \tag{3.23}$$

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. (3.3) for illustration. If input is 1, the output can be 0 with the probability f or 1 with the probability $1-f$:

$$P(y=0|x=0) = 1-f, \quad P(y=0|x=1) = f, \tag{3.24}$$

$$P(y=1|x=0) = f, \quad P(y=1|x=1) = 1-f. \tag{3.25}$$

2. *Binary erasure channel*

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

Transition probabilities:

$$P(y=0|x=0) = 1-f, \quad P(y=0|x=1) = f, \tag{3.26}$$

$$P(y=?|x=0) = f, \quad P(y=?|x=1) = f, \tag{3.27}$$

$$P(y=1|x=0) = 0, \quad P(y=1|x=1) = 1-f. \tag{3.28}$$

3. *Z channel*

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

Transition probabilities:

$$P(y=0|x=0) = 1, \quad P(y=0|x=1) = f, \tag{3.29}$$

$$P(y=1|x=0) = 0, \quad P(y=1|x=1) = 1-f. \tag{3.30}$$

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

$$P(y) = \sum_{j=1}^{n_X} P(y|x_j)P(x_j) \tag{3.31}$$

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

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} = \frac{P(y|x)P(y)}{\sum_{j=1}^{n_X} P(y|x_j)P(x_j)}. \tag{3.32}$$

Then

$$\begin{aligned} P(x = 1|y = 1) = \frac{P(y = 1|x = 1)P(x = 1)}{P(y = 1|x = 0)P(x = 0) + P(y = 1|x = 1)P(x = 1)} = \\ = \frac{0.85 \times 0.1}{0.15 \times 0.9 + 0.85 \times 0.1} = 0.39. \end{aligned} \tag{3.33}$$

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

$$I(X; Y) = S(Y) - S(Y|X) \tag{3.34}$$

First, the marginal entropy Y is simply $S(Y) = S_{\text{binary}}(0.22)$, where $S_{\text{binary}}(p)$ is given by 3.3. Next, the conditional entropy $S(Y|X)$ is

$$S(Y|X) = S(Y|x = 1)P(x = 1) + S(Y|x = 0)P(x = 0). \tag{3.35}$$

where

$$S(Y|x=1) = -P(y=1|x=1)\log_2 P(y=1|x=1) - \\ -P(y=0|x=1)\log_2 P(y=0|x=1) = -0.85\log_2 0.85 - 0.15\log_2 0.15, \quad (3.36)$$

$$S(Y|x=0) = -P(y=1|x=0)\log_2 P(y=1|x=0) - \\ -P(y=0|x=0)\log_2 P(y=0|x=0) = -0.15\log_2 0.15 - 0.85\log_2 0.85. \quad (3.37)$$

Therefore

$$I(X;Y) = S_{\text{binary}}(0.22) - S_{\text{binary}}(0.15) = 0.15 \text{ bits}. \quad (3.38)$$

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

4) In general

$$I(X;Y) = S_{\text{binary}}((1-f)p + (1-p)f) - S_{\text{binary}}(f). \quad (3.39)$$

Performing explicit maximization of this function over p one arrives at

$$C(Q) = \max_{P(X)} I(X;Y) = 0.39 \text{ bits}. \quad (3.40)$$

3.4 Problems

Problem 1: *Z channel*

Consider the Z channel (see Fig. 3.3c) 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 4

Finite Markov Chains. Efficient Mixing

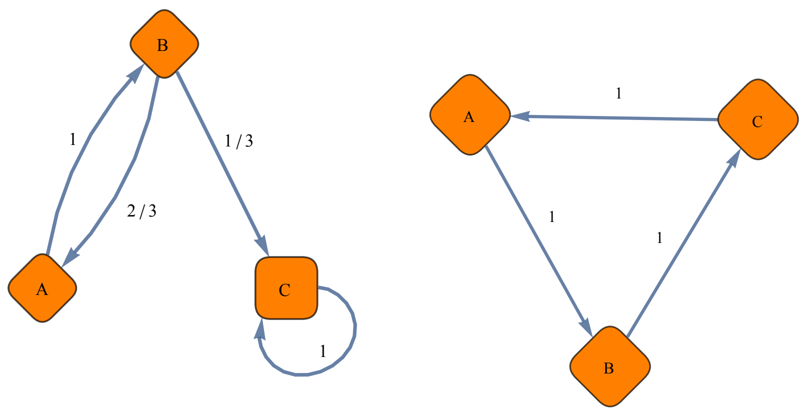
Before we give a formal definition of a Markov Chain (MC), let us watch the introduction video, which explains the origin of Markov chains and briefly describes what they are.

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

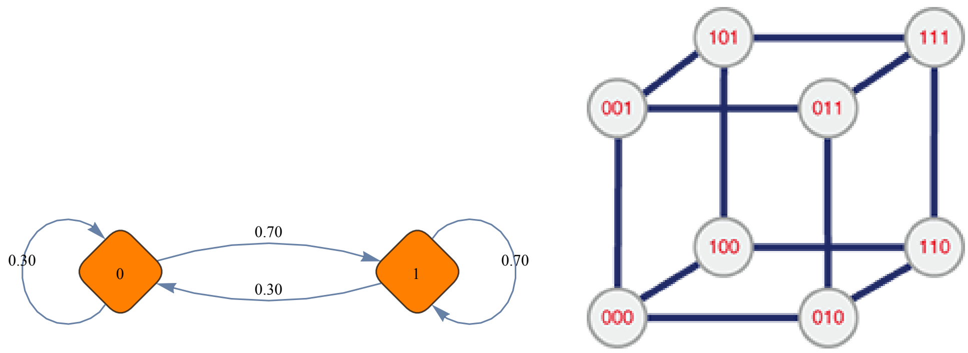
Figure 4.1: Some examples of Markov chains.

(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

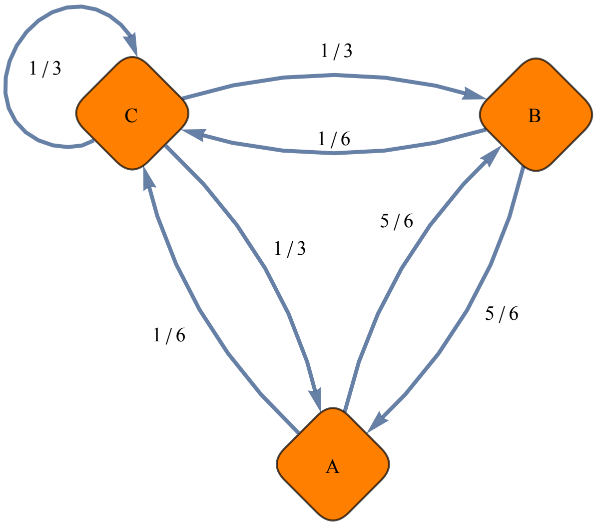


Figure 4.3: Illustration of the detailed balance.

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*** [1], 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 [2].

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 [3]. 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

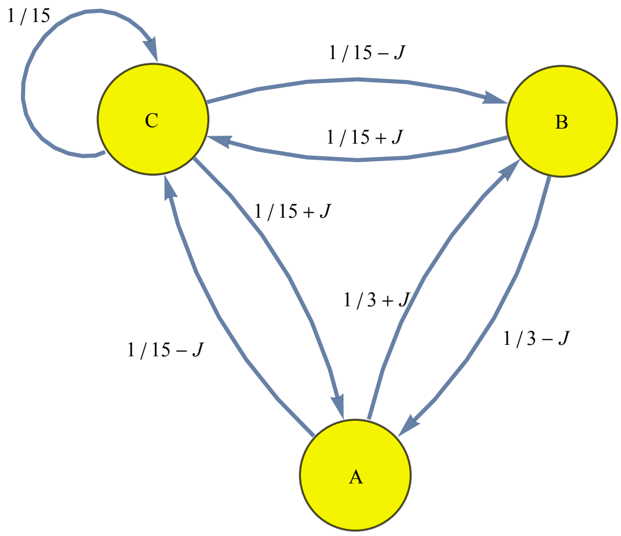


Figure 4.4: Probability fluxes for the stationary distribution π^* of the Markov chain shown in the Figure 4.3. The case $J = 0$ corresponds to the detailed balance.

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 $|\lambda| \le 1$. Hint: for a vector v, let $||v||_{max}$ denote $\max_i |v_i|$, and show that $||Mv||_{max} \le ||v||_{max}$.

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 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.

5.1 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

$$\mathrm{E}[b] = 1 \times p + 0 \times q = p, \tag{5.1}$$

$$\mathrm{Var}[b] = (1 - p)^2 \times p + (0 - p)^2 \times q = pq. \tag{5.2}$$

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

$$B(n, N, p) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}. \tag{5.3}$$

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. (5.3).

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

$$\begin{aligned} \mathrm{E}[n] &= \sum_{i=1}^{N} nB(n, N, p) = \sum_{i=1}^{N} \binom{N}{n} np^n q^{N-n} = p\frac{d}{dp}\sum_{i=1}^{N}\binom{N}{n} p^n q^{N-n} = \\ &= p\frac{d}{dp}(p+q)^N = Np(p+q)^N = pN, \end{aligned} \tag{5.4}$$

$$\begin{aligned} \mathrm{Var}[n] &= \sum_{i=1}^{N} n^2 B(n, N, p) - p^2N^2 = \sum_{i=1}^{N}\binom{N}{n} n^2 p^n q^{N-n} - p^2N^2 = \\ &= p^2\frac{d^2}{dp^2}\sum_{i=1}^{N}\binom{N}{n} p^n q^{N-n} + p\frac{d}{dp}\sum_{i=1}^{N}\binom{N}{n} p^n q^{N-n} - p^2N^2 = \\ &= p^2\frac{d^2}{dp^2}(p+q)^N + p\frac{d}{dp}(p+q)^N - p^2N^2 = pqN. \end{aligned} \tag{5.5}$$

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

$$P(t_1) = p(1-p)^{t_1-1}. \tag{5.6}$$

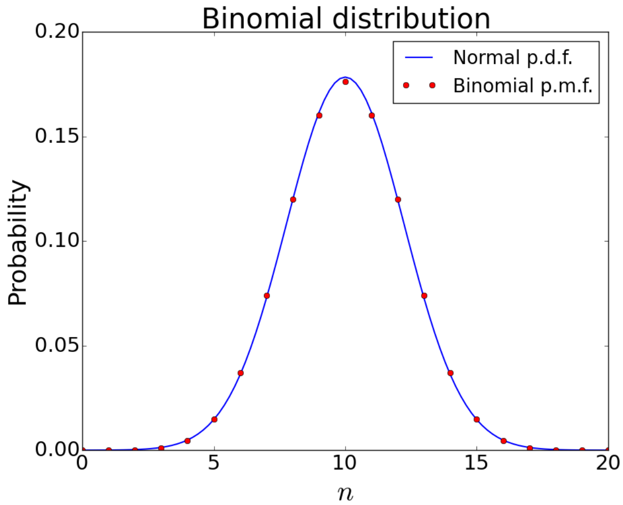


Figure 5.1: Binomial probability mass function and normal probability density function approximation for $N = 20$ and $p = 0.5$.

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

$$\langle t_i \rangle = \sum_{t_1=1}^{\infty} t_i P(t_i) = \frac{1}{p}. \tag{5.7}$$

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

$$P(t_k) = pB(k-1, t_k - 1, p) = \frac{(t_k - 1)!}{(k-1)!(t_k - k)!} p^k (1-p)^{t_k - k}. \tag{5.8}$$

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

5.2 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

$$P(n, t, T) = B(n, N, t/T) = \frac{N!}{n!(N-n)!}\left(\frac{t}{T}\right)^n\left(1-\frac{t}{T}\right)^{N-n}. \tag{5.9}$$

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

$$P(n, t) = \frac{(\lambda t)^n}{n!}e^{-\lambda t}, \tag{5.10}$$

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 $\{t_1, t_2, \ldots\}$ 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

$$\int_T^\infty p(T')dT' = P(0, T) = e^{-\lambda T}. \tag{5.11}$$

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,

$$p(T) = \lambda e^{-\lambda T}. \tag{5.12}$$

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

$$P(n,t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, \tag{5.13}$$

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

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

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

3) It is intuitively clear that

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

Therefore

$$\int_t^\infty p(t_n) dt_n = \sum_{k=0}^{n-1} P(k,t). \tag{5.14}$$

After simple algebra we obtain

$$p(t_n) = -\frac{d}{dt} \sum_{k=0}^{n-1} P(k,t)|_{t=t_n} = \frac{\lambda^n t_n^{n-1}}{(n-1)!} e^{-\lambda t_n}. \tag{5.15}$$

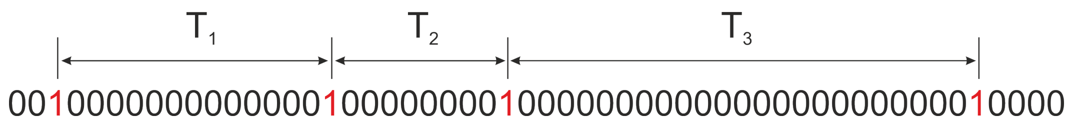


Figure 5.2: Bernoulli process with very low frequency of successes p. The distribution of the inter-arrival time t can be approximated by the Poisson distribution $p(t) = pe^{-pt}$.

5.3 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

$$P(t) = p(1-p)^{t-1} \approx pe^{-pt}. \tag{5.16}$$

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

5.4 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 6

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.

6.1 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

$$E = -\sum_{\langle ij \rangle} J_{ij} s_i s_j - \mu \sum_j s_j h_j, \tag{6.1}$$

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

$$E = -\sum_{\langle ij \rangle} s_i s_j. \tag{6.2}$$

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

$$P_{eq}(s) = \frac{1}{Z} e^{-\beta E(s)}, \quad Z = \sum_s e^{-\beta E(s)} \tag{6.3}$$

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

$$\frac{W}{Z} e^{-\beta E} = \frac{1}{Z} e^{S - \beta E} = \frac{1}{Z} e^{-\beta(E - TS)}, \tag{6.4}$$

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*** $E - TS$.

6.2 Direct Sampling (by Rejection)

Now suppose that we want to generate a random state of the Ising model, according to the Boltzmann distribution (6.3). 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 (6.3) 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.

6.3 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 (6.3). 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

$$p = \begin{cases} 1, & \text{if } \Delta E < 0 \\ e^{-\beta \Delta E}, & \text{if } \Delta E \geq 0. \end{cases} \tag{6.5}$$

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 (6.3) 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

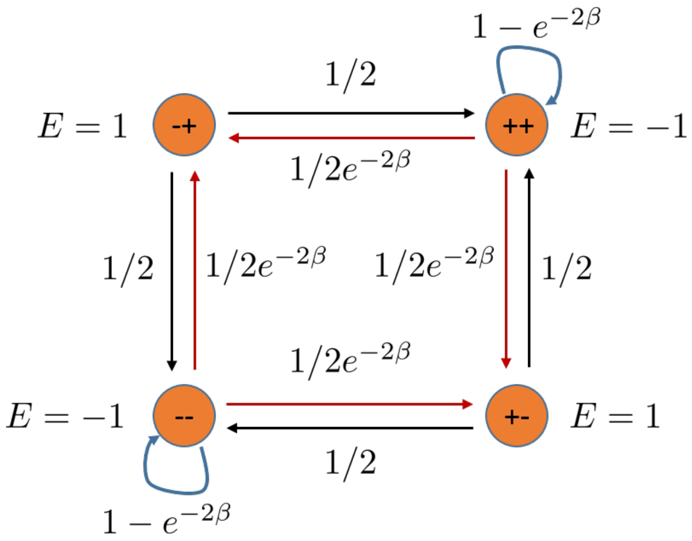


Figure 6.1: Metropolis-Hastings Markov chain example for two spins

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 (6.2) 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 6.1. One can check that the Boltzmann distribution 6.3 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).

6.4 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

$$p_+ + p_- = 1, \quad p_+/p_- = e^{-\beta\Delta E}, \tag{6.6}$$

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 (6.6) and the Boltzmann distribution (6.3) satisfies detailed balance. The probability flux from the state with $s_i = 1$ to the state with $s_i = -1$ is equal to

$$Q_{-+} = \frac{1}{Z} e^{-\beta E(s_i=1)} p_-, \tag{6.7}$$

while the reversed probability flux is equal to

$$Q_{+-} = \frac{1}{Z} e^{-\beta E(s_i=-1)} p_+. \tag{6.8}$$

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).

6.5 Problems

Problem 1. Consider the Ising model (6.2) 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.

References

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

[2] 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)

[3] Turitsyn, Konstantin S, Michael Chertkov, and Marija Vucelja, 2011, "Irreversible monte carlo algorithms for efficient sampling," Physica D: Nonlinear Phenomena **240**, 410–414