Stochastic Modeling and Computations

Recitation Notes
written by

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

Skolkovo Tech
Skolkovo Institute of Science and Technology

4 April 2016

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 Σ, can be discrete, continuous or mixed. The probability to find an instance from Σ 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, $p(x)$, $\forall x \in \Sigma$, 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 of 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 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.

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 is 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

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

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. Calculate the characteristic function of the Lorentz distribution (1.13). Compute the first moment. Check that higher order moments do not exist.

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

Problem 3. *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 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. Choose, at random, three points on the circle $x^2 + y^2 = 1$. Interpret them as cuts that divide the circle into three arcs. Compute the expected length of the arc that contains the point $(1, 0)$.

Problem 6. A book of 500 pages contains 100 missprints. Estimate the probability that at least one page contains 5 missprints.

Problem 7. *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 $1 - p_{22} < 0.5$.

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 rewrite Eq. (2.8)

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

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):

$$P_n(X_n) \to N(\mu_x, \frac{\sigma^2_x}{n}) = \frac{1}{\sqrt{2\pi n}\sigma_x}\exp\left(-n\frac{(X_n-\mu_x)^2}{2\sigma^2_x}\right). \tag{2.10}$$

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 (also called the 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 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 $x_1, x_2, \dots, x_n$ 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

$$g_n(k) = \mathcal{G}^n(k/n) = \exp\left(i\mu k - \frac{\sigma^2 k^2}{2n}\right), \tag{2.17}$$

Its Fourier transform is

$$P_n(X_n) = \int_{-\infty}^{+\infty} \frac{dk}{2\pi} g_n(k) e^{-ikX_n} = \int_{-\infty}^{+\infty} \frac{dk}{2\pi} \exp\left(-ik(X_n - \mu) - n\frac{\sigma^2 k^2}{2}\right) = \tag{2.18}$$

$$= \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{n(X_n - \mu)^2}{2\sigma^2}\right). \tag{2.19}$$

Exercise 3. *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

$$P(x) = \frac{\gamma}{\pi}\frac{1}{x^2+\gamma^2}. \tag{2.20}$$

Solution: The characteristic function of the Cauchy distribution is

$$\mathcal{G}(k) = \frac{\gamma}{\pi}\int_{-\infty}^{+\infty}\frac{dx}{x^2+\gamma^2}e^{ikx} = e^{-\gamma k}. \tag{2.21}$$

The resulting characteristic functional expression is

$$g_n(k) = \mathcal{G}^n(k/n) = \mathcal{G}(k). \tag{2.22}$$

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

$$P(x_1, \ldots, x_M) = \frac{1}{N}\exp\left(-\frac{x_i A_{ij} x_j}{2}\right), \tag{2.23}$$

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

$$N = \frac{(2\pi)^{M/2}}{\sqrt{\det A}}, \tag{2.24}$$

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

$$\mathbf{E}[x_i x_j] = A^{-1}_{ij}. \tag{2.25}$$

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

$$\mathbf{E}[x_1 x_2 \dots x_{2n}] = \sum \prod \mathbf{E}[x_i x_j], \tag{2.26}$$

$$\mathbf{E}[x_1 x_2 \dots x_{2n+1}] = 0, \tag{2.27}$$

Notice, that in Eq. (2.27) we simply sum over all possible pairs in the set $x_1, x_2, \dots, x_{2n}$. For example, Eq. (2.27) for the forth order moment transforms to

$$\mathbf{E}[x_i x_j x_k x_m] = \mathbf{E}[x_i x_j]\mathbf{E}[x_k x_m] + \mathbf{E}[x_i x_k]\mathbf{E}[x_j x_m] + \mathbf{E}[x_i x_m]\mathbf{E}[x_j x_k]. \tag{2.28}$$

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

Exercise 4. *Joint probability distribution of the multivariate Gaussian variables*

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

$$P(x_1, x_2) = \frac{1}{N} \exp(-x_1^2 - x_1 x_2 - x_2^2). \tag{2.29}$$

(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 100 times. Awards for the game are as follows

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

(1) What is the expected value of your winnings?

(2) What is the standard deviation of your winnings?

(3) What is the probability you win at least 200$?

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.

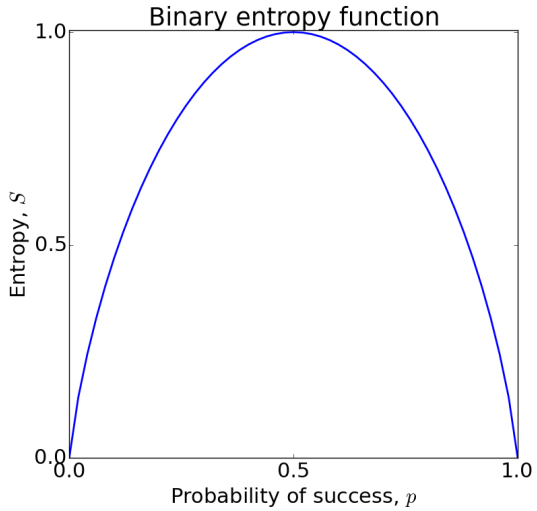

Figure 3.1: Entropy of the Bernoulli distribution as a function of the success rate, p.

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).

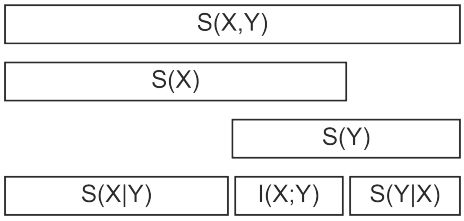

Figure 3.2: Illustration of the relations between joint entropy, marginal entropies, conditional entropies and mutual information.

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 $\sim$ 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.

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 $P(x)$.

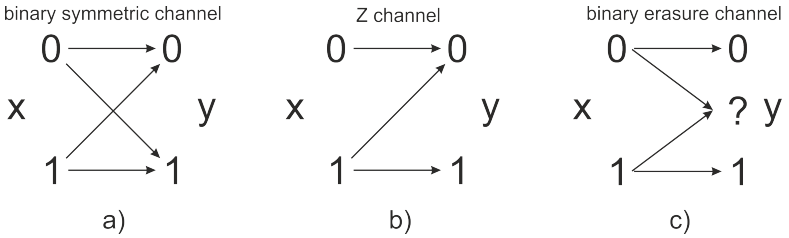

Figure 3.3: Examples of communication channels.

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

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

Then

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

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

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

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

Therefore

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

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

4) In general

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

Performing explicit maximization of this function over p one arrives at

$$C(Q) = \max_{P(X)} I(X;Y) = S_{\text{binary}}(0.5) - S_{\text{binary}}(0.5) = 0.39 \text{ bits}. \tag{3.40}$$

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?