Stochastic Processes
and
Monte-Carlo methods

University of Massachusetts: Fall 2007

Luc Rey-Bellet

Contents

Chapter 1

Random Variables

1.1 Review of probability

In this section we briefly review the basic terminology of probability and statistics, see any elementary probability book for reference.

Any real-valued random variable X is described by its *cumulative distribution function* (abbreviated c.d.f) of X, i.e., the function $F_X : \mathbf{R} \to [0,1]$ defined by

$$F_X(x) \;=\; P(X \leq x)\,.$$

If there exists a continuous function $f : \mathbf{R} \to [0,\infty)$ such that $F_X(x) = \int_{-\infty}^{x} f_X(y)\,dy$ then X is said to be *continuous* with *probability density function* (abbreviated p.d.f) f_X. By the fundamental theorem of calculus the p.d.f of X is obtained from the c.d.f of X by differentiating, i.e.,

$$f_X(x) = F_X'(x)\,.$$

On the other hand if X takes values in the set of integers, or more generally in some countable or finite subset $S$of the real numbers, then the random variable X and its c.d.f. are completely determined by its *probability distribution function* (p.d.f), i.e. by $p : S \to [0,1]$ where

$$p(i) \;=\; P(X = i)\,, \quad i \in S\,.$$

In this case X is called a *discrete* random variable.

The p.d.f. f of a continuous random variable satisfies $\int_{-\infty}^{\infty} f(x)\,dx = 1$ and the p.d.f of a discrete random variable satisfies $\sum_{i\in S} p_i = 1$. Either the c.d.f or p.d.f describes the *distribution* of X and we compute the probability of any *event* $A \subset \mathbf{R}$ by

$$P(X \in A) \;=\; \begin{cases} \int A f(x)\,dx & \text{if } X \text{ is continuous}\,, \\ \sum_{i\in A} p(i)\,dx & \text{if } X \text{ is discrete}\,. \end{cases}$$

Let $\mathbf{X} = (X_1, \cdots, X_d)$ be a *random vector*, i.e., $X_1, \cdots X_d$ are real-valued random variables with some joint distribution. Often the joint distribution can be described

by the multiparameter analogue of the p.d.f. For example if there is a function $f_{\mathbf{X}} : \mathbf{R}^d \to [0, \infty)$ such that

$$P(\mathbf{X} \in A) = \int \cdots \int_A f_{\mathbf{X}}(x_1, \cdots, x_d) dx_1 \cdots dx_d$$

then $\mathbf{X}$ is a continuous random vector with p.d.f $f_{\mathbf{X}}$. Similarly a discrete random vector $\mathbf{X}$ taking values $\mathbf{i} = (i_1, \cdots, i_d)$ is described by

$$p(i_1, \cdots, i_d) = P(X_1 = i_1, \cdots X_d = i_d)\,.$$

A collection of random variables $X_1, \cdots, X_d$ are *independent* if

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= f_{X_1}(x_1) \cdots f_{X_d}(x_d)\,, \quad \text{continuous case} \\ p_{\mathbf{X}}(\mathbf{i}) &= p_{X_1}(i_1) \cdots p_{X_d}(i_d)\,, \quad \text{discrete case} \end{aligned} \tag{1.1}$$

If $\mathbf{X}$ is a random vector and $g : \mathbf{R}^d \to \mathbf{R}$ is a function then $Y = g(\mathbf{X})$ is a real random variable. The *mean* or *expectation* of a real random variable X is defined by

$$E[X] = \begin{cases} \int_{-\infty}^{\infty} x f_X(x)\, dx & \text{if } X \text{ is continuous} \\ \sum_{i \in S} i\, p_X(i) & \text{if } X \text{ is discrete} \end{cases}$$

More generally if $Y = g(\mathbf{X})$ then

$$E[Y] = E[g(\mathbf{X})] = \begin{cases} \int_{\mathbf{R}^d} g(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})\, dx & \text{if } X \text{ is continuous} \\ \sum_{\mathbf{i}} g(\mathbf{i})\, p_{\mathbf{x}}(\mathbf{i}) & \text{if } X \text{ is discrete} \end{cases}$$

The *variance* of a random variable X, denoted by $\mathrm{var}(X)$, is given by

$$\mathrm{var}(X) = E\left[(X - E[X])^2\right] = E[X^2] - E[X]^2\,.$$

The mean of a random variable X measures the average value of X while its variance is a measure of the spread of the distribution of X. Also commonly used is the *standard deviation* $\sqrt{\mathrm{var}(X)}$.

Let X and Y be two random variables then we have

$$E[X + Y] = E[X] + E[Y]\,.$$

For the variance a simple computation shows that

$$\mathrm{var}(X + Y) = \mathrm{var}(X) + 2\mathrm{cov}(X, Y) + \mathrm{var}(Y)$$

where $\mathrm{cov}(X, Y)$ is the *covariance* of X and Y and is defined by

$$\mathrm{cov}(X, Y) = E\left[(X - E[X])(Y - E[Y])\right]\,.$$

In particular if X and Y are independent then $E[XY] = E[X]E[Y]$ and so $\text{cov}(X, Y) = 0$ and thus $\text{var}(X_1 + X_2) = \text{var}(X_1) + \text{var}(X_2)$.

Another important and useful object is the *moment generating function* (m.g.f.) of a random variable X and is given by

$$M_X(t) \;=\; E\left[e^{tX}\right] .$$

Whenever we use a m.g.f we will always assume that $M_X(t)$ is finite at least in an interval around 0. Note that this is not always the case.

If the moment generating function of X is known then one compute all *moments* of X, i.e. $E[X^n]$ by repeated differentiation of the function $M_X(t)$ with respect to t. The n^{th} derivative of $M_x(t)$ is given by

$$M_x^{(n)}(t) \;=\; \left[X^n e^{tX}\right]$$

and therefore

$$E[X^n] \;=\; M^{(n)}(0)\,.$$

In particular $E[X] = M_X'(0)$ and $\text{var}(X) = M_X''(0) - (M_X'(0))^2$. It is often very convenient to compute the mean and variance of X using these formulas (see the examples below).

An important fact is the following (its proof is not that easy!)

Theorem 1.1.1 *Let X and Y be two random variables and suppose that $M_X(t) = M_Y(t)$ for all $t \in (-\delta, \delta)$ then X and Y have the same distribution.*

Another important property of the m.g.f is

Proposition 1.1.2 *If X and Y are independent random variable then the m.g.f of $X + Y$ satisfies*

$$M_{X+Y}(t) = M_X(t)M_Y(t)\,,$$

i.e., the m.g.f of a sum of independent random variable is the product of the m.g.f.

Proof: We have

$$E\left[e^{t(X+Y)}\right] \;=\; E\left[e^{tX}e^{tY)}\right] \;=\; E\left[e^{tX}\right]E\left[e^{tY)}\right] ,$$

since e^{tX} and e^{tY} are independent. ∎

1.2 Some Common Distributions

We recall some important distributions together with their basic properties. The following fact are useful to remember.

Proposition 1.2.1 *We have*

1. *Suppose X is a continuous random variable with p.d.f $f(x)$. For any real number a the p.d.f of $X+a$ is $f(x-a)$.*
2. *Suppose X is a continuous random variable with p.d.f $f(x)$. For any non zero real number b the p.d.f of bX is $\frac{1}{b}f\left(\frac{x}{b}\right)$.*
3. *Suppose X is a random variable. For any real number a and b we have $M_{bX+a}(t) = e^{at}M_X(bt)$*

Proof: The c.d.f of $X+a$ is

$$F_{X+a}(x) \;=\; P(X+a \le x) \;=\; P(X \le x-a) \;=\; F_X(x-a)\,.$$

Differentiating with respect to x gives

$$f_{X+a}(x) = F'_{X+a}(x) = f_X(x-a)\,.$$

This shows (i).

To prove (ii) one proceeds similarly. For $b>0$

$$F_{bX}(x) \;=\; P(bX \le x) \;=\; P(X \le x/b) \;=\; F_X(x/b)\,.$$

Differentiating gives $f_{bX}(x) = \frac{1}{b}f\left(\frac{x}{b}\right)$. The case $b<0$ is left to the reader.

To prove (iii) note that

$$M_{bX+a}(t) \;=\; E\left[e^{t(bX+a)}\right] \;=\; e^{ta}E\left[e^{tbX}\right] \;=\; e^{ta}M_X(bt)\,.$$

■

1) Uniform Random Variable

Consider real numbers $a<b$. The *uniform random variable on* $[a,b]$ is the continuous random variable with p.d.f

$$f(x) \;=\; \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

The moment generating function is

$$E\left[e^{tX}\right] = \int_a^b e^{tx}\,dx \;=\; \frac{e^{tb}-e^{ta}}{t(b-a)}\,.$$

and the mean and variance are

$$E[X] = \frac{b-a}{2}\,, \quad \mathrm{var}(X) \;=\; \frac{(b-a)^2}{12}\,.$$

We write $X = U[a,b]$ to denote this random variable.

2) Normal Random Variable

Let μ be a real number and σ be a positive number. The *normal random variable with mean μ and variance σ^2* is the continuous random variable with p.d.f

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

The moment generating function is (see below for a proof)

$$E\left[e^{tX}\right] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\;dx \;=\; e^{\mu t + \frac{\sigma^2 t^2}{2}}\,. \tag{1.2}$$

and the mean and variance are

$$E[X] = \mu\,, \quad \mathrm{var}(X) \;=\; \sigma^2\,.$$

We write $X = N(\mu,\sigma^2)$ to denote this random variable. The *standard normal random variable* is the normal random variable with $\mu = 0$ and $\sigma = 1$, i.e., $N(0,1)$

The normal random variable has the following property

$$X = N(0,1) \quad \text{ if and only if } \quad \sigma X + \mu = N(\mu,\sigma^2)$$

To see this one applies Proposition 1.2.1 (i) and (ii) and this tells us that the density of $\sigma X + \mu$ is $\frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$.

To show that the formula for the moment generating function we consider first $X = N(0,1)$. Then by completing the square we have

$$\begin{aligned} M_X(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-\frac{x^2}{2}}\;dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{t^2}{2}} e^{-\frac{(x-t)^2}{2}}\;dx \\ &= e^{\frac{t^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-t)^2}{2}}\;dx \\ &= e^{\frac{t^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}\;dy \\ &= e^{\frac{t^2}{2}} \end{aligned} \tag{1.3}$$

This proves the formula for $N(0,1)$. Since $N(\mu,\sigma^2) = \sigma N(0,1) + \mu$, by Proposition 1.2.1, (iii) the moment generating function of $N(\mu,\sigma^2)$ is $e^{t\mu}e^{\frac{\sigma^2 t^2}{2}}$ as claimed.

3) Exponential Random Variable

Let λ be a positive number. The *exponential random variable with parameter* λ is the continuous random variable with p.d.f

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} .$$

The moment generating function is

$$E\left[e^{tX}\right] = \lambda \int_0^\infty e^{tx} e^{-\lambda x} = \begin{cases} \frac{\lambda}{\lambda - t} & \text{if } \lambda < t \\ +\infty & \text{otherwise} \end{cases}$$

and the mean and variance are

$$E[X] = \frac{1}{\lambda}, \quad \text{var}(X) = \frac{1}{\lambda^2}.$$

We write $X = Exp(\lambda)$ to denote this random variable. This random variable will play an important role in the construction of continuous-time Markov chains. It often has the interpretation of a waiting time until the occurrence of an event.

4) Gamma Random Variable

Let n and λ be positive numbers. The *gamma random variable with parameters* n *and* λ is the continuous random variable with p.d.f

$$f(x) = \begin{cases} \lambda e^{-\lambda x} \frac{(\lambda x)^{n-1}}{(n-1)!} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} .$$

The moment generating function is

$$E\left[e^{tX}\right] = \lambda \int_0^\infty e^{tx} \lambda e^{-\lambda x} \frac{(\lambda x)^{n-1}}{(n-1)!} = \begin{cases} \left(\frac{\lambda}{\lambda - t}\right)^n & \text{if } t < \lambda \\ +\infty & \text{otherwise} \end{cases} .$$

and the mean and variance are

$$E[X] = \frac{n}{\lambda}, \quad \text{var}(X) = \frac{n}{\lambda^2}.$$

We write $X = Gamma(n,\lambda)$ to denote this random variable.

To compute the m.g.f note that for any $\alpha > 0$

$$\int_0^\infty e^{-\alpha x}\, dx = \frac{1}{\alpha}.$$

and differentiating repeatedly w.r.t. α gives the formula

$$\int_0^\infty e^{-\alpha x} x^{n-1}\, dx \;=\; \frac{(n-1)!}{\alpha^n}\,.$$

Note that $Gamma(1,\lambda) = Exp(\lambda)$. Also the m.g.f of $Gamma(n,\lambda)$ is the m.g.f of $Exp(\lambda)$ to the n^{th} power. Using Theorem 1.1.1 and Proposition 1.1.2 we conclude that if $X_1, \cdots, X_n$ are exponential random variables with parameters λ then $X_1 + \cdots + X_n = Gamma(n,\lambda)$.

5) Bernoulli Random Variable A Bernoulli random variable models the toss a (possibly unfair coin), or more generally any random experiment with exactly two outcomes. Let p be a number with $0 \le p \le 1$. The *Bernoulli random variable with parameter* p is the discrete random variable taking value in $\{0,1\}$ with

$$p(0) = 1-p\,, \quad p(1) = p$$

The moment generating function is

$$E\left[e^{tX}\right] \;=\; 1 - p + pe^t\,,$$

and the mean and the variance are

$$E[X] = p\,, \quad \mathrm{var}(X) \;=\; p(1-p)\,.$$

A typical example where Bernoulli random variable occur is the following. Let Y be any random variable, let A be any event, the indicator random variable $\mathbf{1}_A$ is defined by

$$\mathbf{1}_A \;=\; \begin{cases} 1 & \text{if } Y \in A \\ 0 & \text{if } Y \notin A \end{cases}$$

Then $\mathbf{1}_A$ is a Bernoulli random variable with $p = P\{Y \in A\}$.

6) Binomial Random Variable Consider an experiment which has exactly two outcomes 0 or 1 and is repeated n times, each time independently of each other (i.e., n *independent trials*). The binomial random variable is the random variable which counts the number of 1 obtained in the n trials. Let p be a number with $0 \le p \le 1$ and let n be a positive integer. The *Bernoulli random variable with parameters* n *and* p is the random variable which counts the number of 1 occurring in the n outcomes. The p.d.f is

$$p(i) = \binom{n}{i} p^i (1-p)^{n-i}\,, \quad i = 0, 1, \cdots, n\,.$$

The moment generating function is

$$E\left[e^{tX}\right] \;=\; ((1-p) + pe^t)^n\,,$$

and the mean and the variance are

$$E[X] = np\,, \quad \text{var}(X) \;=\; np(1-p)\,.$$

We write $X = B(n,p)$ to denote this random variable.

The formula for the m.g.f can be obtained directly using the binomial theorem, or simply by noting that by construction $B(n,p)$ is a sum of n independent Bernoulli random variables.

7) Geometric Random Variable Consider an experiment which has exactly two outcomes 0 or 1 and is repeated as many times as needed until a 1 occurs. The geometric random describes the probability that the first 1 occurs at exactly the n^th trial. Let p be a number with $0 \leq p \leq 1$ and let n be a positive integer. The *Geometric random variable with parameter* p is the random variable with p.d.f

$$p(n) = (1-p)^{n-1}p\,, \quad n = 1, 2, 3, \cdots$$

The moment generating function is

$$E\left[e^{tX}\right] \;=\; \sum_{n=1}^{\infty} e^{tn}(1-p)^{n-1}p \;=\; \begin{cases} \frac{pe^t}{1-e^t(p-1)} & \text{if } e^t(p-1) < 1 \\ 0 & \text{otherwise} \end{cases}\,,$$

The mean and the variance are

$$E[X] = \frac{1}{p}\,, \quad \text{var}(X) \;=\; \frac{1-p}{p^2}\,.$$

We write $X = Geometric(p)$ to denote this random variable.

8) Poisson Random Variable Let λ be a positive number. The *Poisson random variable with parameter* λ is the discrete random variable which takes values in $\{0, 1, 2, \cdots\}$ and with p.d.f

$$p(n) \;=\; e^{-\lambda}\frac{\lambda^n}{n!} \quad n = 0, 1, 2, \cdots\,.$$

The moment generating function is

$$E\left[e^{tX}\right] \;=\; \sum_{n=0}^{\infty} e^{tn}\frac{\lambda^n}{n!}e^{-\lambda} \;=\; e^{\lambda(e^t-1)}\,.$$

The mean and the variance are

$$E[X] = \lambda\,, \quad \text{var}(X) \;=\; \lambda\,.$$

We write $X = Poisson(\lambda)$ to denote this random variable.

1.3 Simulating Random Variables

In this section we discuss a few techniques to simulate a given random variable on a computer. The first step which is built-in in any computer is the simulation of a *random number*, i.e., the simulation of a uniform random variable $U([0,1])$, rounded off to the nearest $\frac{1}{10^n}$.

In principle this is not difficult: take ten slips of paper numbered $0, 1, \cdots, 9$, place them in a hat and select successively n slips, with replacement, from the hat. The sequence of digits obtained (with a decimal point in front) is the value of a uniform random variable rounded off to the nearest $\frac{1}{10^n}$. In pre-computer times, tables of random numbers were produced in that way and still can be found. This is of course not the way a actual computer generates a random number. A computer will usually generates a random number by using a deterministic algorithm which produce a pseudo random number which "looks like" a random number For example choose positive integers a, c and m and set

$$X_{n+1} = (aX_n + c) \bmod(m)\,.$$

The number X_n is either $0, 1, \cdots, m-1$ and the quantity X_n/m is taken to be an approximation of a uniform random variable. One can show that for suitable a, C and m this is a good approximation. This algorithm is just one of many possibles and used in practice. The issue of actually generating a good random number is a nice, interesting, and classical problem in computer sciences. For our purpose we will simply content ourselves with assuming that there is a "black box" in your computer which generates $U([0,1])$ in a satisfying manner.

We start with a very easy example, namely simulating a discrete random variable X.

Algorithm 1.3.1 (Discrete random variable) *Let X be a discrete random variable taking the values $x_1, x_2, \cdots$ with p.d.f. $p(j) = P\{X = x_j\}$. To simulate X,*

- *Generate a random number $U = U([0,1])$.*
- *Set*

$$X = \begin{cases} x_1 & \text{if } U < p(1) \\ x_2 & \text{if } p(1) < U < p(1) + p(2) \\ \vdots & \vdots \\ x_n & \text{if } p(1) + \cdots + p(n-1) < U < p(1) + \cdots p(n) \\ \vdots & \vdots \end{cases}$$

Then X has the desired distribution.

We discuss next two general methods simulating continuous random variable. The first is called the inverse transformation method and is based on the following

Proposition 1.3.2 *Let $U = U([0,1])$ and let $F = F_X$ be the c.d.f of the continuous random variable X. Then*

$$X = F^{-1}(U)\,,$$

and also

$$X = F^{-1}(1-U)\,.$$

Proof: By definition the c.d.f of the random variable X is a continuous increasing function of F, therefore the inverse function F^{-1} is well-defined and we have

$$P\{F^{-1}(U) \leq a\} \;=\; P\{U \leq F(a)\} = F(a)\,.$$

and this shows that the c.d.f of $F^{-1}(U)$ is F and thus $X = F^{-1}(U)$. To prove the second formula simply note that U and $1-U$ have the same distribution. ∎

So we obtain

Algorithm 1.3.3 (Inversion method for continuous random variable) *Let X be a random variable with c.d.f $F = F_X$. To simulate X*

- **Step 1** *Generate a random number $U = U([0,1])$.*
- **Step 2** *Set $X = F^{-1}(U)$.*

Example 1.3.4 (Simulating an exponential random variable) If $X = Exp(\lambda)$ then its c.d.f if

$$F(x) = 1 - e^{-\lambda x}\,.$$

The inverse function F^{-1} is given by

$$1 - e^{-\lambda x} = u \quad \text{iff} u = -\frac{1}{\lambda}\log(1-u)\,.$$

Therefore we have $F^{-1}(u) = -\frac{1}{\lambda}\log(1-u)$. So if $U = U([0,1])$ then

$$Exp(\lambda) = -\frac{1}{\lambda}\log(1-U) = -\frac{1}{\lambda}\log(U)\,.$$

The inversion method is most straightforward when there is an explicit formula for the inverse function F^{-1}. In many examples however a such a nice formula is not available. Possible remedies to that situation is to solve $F(X) = U$ numerically for example by Newton method.

Another method for simulating a continuous random variable is the *rejection method*. Suppose we have a method to simulate a random variable with p.d.f $g(x)$ and that we want to simulate the random variable with p.d.f $f(x)$. The following algorithm is due to Von Neumann.

Algorithm 1.3.5 (Rejection method for continuous random variable). *Let X be a random variable with p.d.f $f(x)$ and let Y be a random variable with p.d.f $g(x)$. Furthermore assume that there exists a constant C such that*

$$\frac{f(y)}{g(y)} \leq C\,, \quad \text{for all } y\,.$$

To simulate X

- **Step 1** *Simulate Y with density g.*
- **Step 2** *Simulate a random number U.*
- **Step 3** *If*

$$U \leq \frac{f(Y)}{g(Y)C}$$

set $X = Y$. Otherwise return to Step 1.

That the algorithm does the job is the object of the following proposition.

Proposition 1.3.6 *The random variable X generated by the rejection method has p.d.f $f(x)$.*

Proof: To obtain a value of X we will need in general to iterate the algorithm a random number of times We generate random variables $Y_1, \cdots, Y_N$ until Y_N is accepted and then set $X = Y_N$. We need to verify that the p.d.f of X is actually $f(x)$.

Then we have

$$\begin{aligned}
P(X \leq x) &= P(Y_N \leq x) \\
&= P\left(Y \leq x \,|\, U \leq \frac{f(Y)}{Cg(Y)}\right) \\
&= \frac{P\left(Y \leq x\,, U \leq \frac{f(Y)}{Cg(Y)}\right)}{P\left(U \leq \frac{f(Y)}{Cg(Y)}\right)} \\
&= \frac{\int_{-\infty}^{\infty} P\left(Y \leq x\,, U \leq \frac{f(Y)}{Cg(Y)} \,|\, Y = y\right) g(y)\, dy}{P\left(U \leq \frac{f(Y)}{Cg(Y)}\right)} \\
&= \frac{\int_{-\infty}^{x} P\left(U \leq \frac{f(y)}{Cg(y)}\right) g(y)\, dy}{P\left(U \leq \frac{f(Y)}{Cg(Y)}\right)} \\
&= \frac{\int_{-\infty}^{x} \frac{f(y)}{Cg(y)} g(y)\, dy}{P\left(U \leq \frac{f(Y)}{Cg(Y)}\right)} \\
&= \frac{\int_{-\infty}^{x} f(y)\, dy}{CP\left(U \leq \frac{f(Y)}{Cg(Y)}\right)}\,.
\end{aligned}$$

If we let $x \to \infty$ we obtain that $CP\left(U \leq \frac{f(Y)}{Cg(Y)}\right) = 1$ and thus

$$P(X \leq x) \;=\; \int_{-\infty}^{x} f(x)\, dx\,.$$

and this shows that X has p.d.f $f(x)$. ∎

In order to decide whether this method is efficient of not, we need to ensure that rejections occur with small probability. The above proof shows that at each iteration the probability that the results is accepted is

$$P\left(U \leq \frac{f(Y)}{Cg(Y)}\right) = \frac{1}{C}$$

independently of the other iterations. Therefore the number of iterations needed is $Geom(\frac{1}{C})$ with mean C. Therefore the ability to choose a reasonably small C will ensure that the method is efficient.

Example 1.3.7 Let X be the random variable with p.d.f

$$f(x) \;=\; 20(1-x)^3\,, \quad 0 < x < 1\,.$$

Since the p.d.f. is concentrated on $[0,1]$ let us take

$$g(x) = 1 \quad 0 < x < 1\,.$$

To determine C such that $f(x)/g(x) \leq C$ we need to maximize the function $h(x) \equiv f(x)/g(x) = 20x(1-x)^3$. Differentiating gives $h'(x) = 20\left((1-x)^3 - 3x(1-x)^2\right)$ and thus the maximum is attained at $x = 1/4$. Thus

$$\frac{f(x)}{g(x)} \leq 20\frac{1}{4}\left(\frac{3}{4}\right)^3 = \frac{135}{64} \equiv C\,.$$

We obtain

$$\frac{f(x)}{Cg(x)} \;=\; \frac{256}{27}x(1-x)^3$$

and the rejection method is

- **Step 1** Generate random numbers U_1 and U_2.
- **Step 2** If $U_2 \leq \frac{256}{27}U_1(1-U_1)^3$, stop and set $X = U_1$. Otherwise return to step 1.

The average number of accepted iterations is 135/64.

Example 1.3.8 (Simulating a normal random variable) Note first that to simulate a normal random variable $X = N(\mu, \sigma^2)$ it is enough to simulate $N(0,1)$ and then set $X = \sigma N(0,1) + \mu$.

Let us first consider the random variable Z whose density is

$$f(x) \;=\; \frac{2}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,, \quad 0 \le x \le \infty\,.$$

One can think of Z as the absolute value of $N(0,1)$.

We simulate Z by using the rejection method with

$$g(x) = e^{-x} \quad 0, x < \infty\,,$$

i.e., $Y = Exp(1)$. To find C we note that

$$\frac{f(x)}{g(x)} \;=\; \sqrt{\frac{2e}{\pi}} e^{-\frac{(x-1)^2}{2}} \;\le\; \frac{2e}{\pi} \equiv C\,.$$

One generates Z using the rejection method. To generate $X = N(0,1)$ from Z one generate a discrete random variable S with takes value $+1$ and -1 with probability $\frac{1}{2}$ and then set $X = SZ$. The random variable S is $S = 2B(1, \frac{1}{2}) - 1$.

- **Step 1** Generate a random numbers U, an exponential random variable Y and a Bernoulli random variable B.
- **Step 2** If $U \le \exp -\frac{(Y-1)^2}{2}$ set $Z = Y$ and $X = (2B-1)Z$

For particular random variables many special techniques have been devised. We give here some examples.

Example 1.3.9 (Simulating a geometric random variable) The c.d.f of the geometric random variable $X = Geom(p)$ is given by

$$F(n) \;=\; P(X \le n) \;=\; 1 - P(X > n) \;=\; 1 - \sum_{k=n+1}^{\infty} (1-p)^{n-1} p \;=\; 1 - (1-p)^n$$

The exponential random variable $Y = Exp(\lambda)$ has c.d.f $1 - e^{-\lambda x}$.

For any positive real number let $\lceil x \rceil$ denote the smallest integer greater than or equal to x, e.g. $\lceil 3.72 \rceil = 4$. Then we claim that if $Y = Exp(\lambda)$ then

$$\lceil Y \rceil = Geom(p) \qquad \text{with } p = 1 - e^{-\lambda}\,.$$

Indeed we have

$$P(\lceil Y \rceil \le n) = P(Y \le n) \;=\; 1 - e^{-\lambda n}\,.$$

Thus we obtain

Algorithm 1.3.10 (Geometric random variable)

- **Step 1** *Generate a random number* U.
- **Step 2** *Set* $X = \lceil \frac{\log(U)}{\log(1-p)} \rceil$

Then $X = Geom(p)$.

Example 1.3.11 (Simulating the Gamma random variable) Using the fact that $Gamma(n, \lambda)$ is a sum of n independent $Exp(\lambda)$ one immediately obtain

Algorithm 1.3.12 (Gamma random variable)

- **Step 1** *Generate* n *random number* $U_1, \cdots, U_n$.
- **Step 2** *Set* $X_i = -\frac{1}{\lambda}\log(U_i)$
- **Step 3** *Set* $X = X_1 + \cdots + X_n$.

Then $X = Gamma(n, p)$.

Finally we give an elegant algorithm which generates 2 independent normal random variables.

Example 1.3.13 (Simulating a normal random variable: Box-Müller) We show a simple way to generate 2 independent standard normal random variables X and Y. The joint p.d.f. of X and Y is given by

$$f(x, y) \;=\; \frac{1}{2\pi} e^{-\frac{(x^2+y^2)}{2}} \,.$$

Let us change into polar coordinates (r, θ) with $r^2 = x^2 + y^2$ and $\tan(\theta) = y/x$. The change of variables formula gives

$$f(x, y)\, dxdy \;=\; re^{-\frac{r^2}{2}} dr \frac{1}{2\theta} d\theta \,.$$

Consider further the change of variables set $s = r^2$ so that

$$f(x, y)\, dxdy \;=\; \frac{1}{2} e^{-\frac{s}{2}}\, ds \frac{1}{2\theta} d\theta \,.$$

The right-hand side is iasily seen to be the joint p.d.f of the two independent random variables $S = Exp(1/2)$ and $\Theta = U([0, 2\pi])$.

Therefore we obtain

Algorithm 1.3.14 (Standard normal random variable)

- **Step 1** *Generate two random number U_1 and U_2*
- **Step 2** *Set*

$$X = \sqrt{-2\log(U_1)}\cos(2\pi U_2) \tag{1.4}$$
$$Y = \sqrt{-2\log(U_1)}\sin(2\pi U_2) \tag{1.5}$$

Then X and Y are 2 independent $N(0,1)$.

1.4 Markov, Chebyshev, and Chernov

We start by deriving simple techniques for bounding the *tail distribution* of a random variable, i.e., bounding the probability that the random variable takes value far from the its mean.

Our first inequality, called *Markov's inequality* simply assumes that we know the mean of X.

Proposition 1.4.1 (Markov's Inequality) *Let X be a random variable which assumes only nonnegative values, i.e. $P(X \geq 0) = 1$. Then for any $a > 0$ we have*

$$P(X \geq a) \leq \frac{E[X]}{a}.$$

Proof: For $a > 0$ let us define the random variable

$$I_a = \begin{cases} 1 & \text{if } X \geq a \\ 0 & \text{otherwise} \end{cases}.$$

Note that, since $X \geq 0$ we have

$$I_a \leq \frac{X}{a} \tag{1.6}$$

and that since I_a is a binomial random variable

$$E[I_a] = P(X \geq a).$$

Taking expectations in the inequality (1.6) gives

$$P(X \geq a) = E[I_a] \leq E\left[\frac{X}{a}\right] = \frac{E[X]}{a}. \quad \blacksquare$$

Example 1.4.2 (Flipping coins) Let us flip a fair coin n times and let us define the random variables X_i, $i = 1, 2, \cdots, n$ by

$$X_i = \begin{cases} 1 & \text{if the } i^{th} \text{ coin flip is head} \\ 0 & \text{otherwise} \end{cases} .$$

Then each X_i is a Bernoulli random variable and $S_n = X_1 + \cdots X_n = B(n, \frac{1}{2})$ is a binomial random variable.

Let us the Markov inequality to estimate the probability that at least 75% of the n coin flips are head. Since $E[S_n] = \frac{n}{2}$ the markov's inequality tells us that

$$P(S_n \geq \frac{3n}{4}) \leq \frac{E[S_n]}{3n/4} = \frac{n/2}{3n/4} = \frac{2}{3} .$$

As we will see later this is a terrible, really terrible, bound but note that we obtained it using only the value of the mean and nothing else.

Our next inequality, which we can derive from Markov's inequality, involves now the variance of X. This is called *Chebyshev's inequality.*

Proposition 1.4.3 (Chebyshev's Inequality) *Let X be a random variable with $E[X] = \mu$ and $Var(X) = \sigma^2$. Then for any $a > 0$ we have*

$$P(|X - \mu| \geq a) \leq \frac{\text{var}(X)}{a^2} .$$

Proof: Observe first that

$$P(|X - \mu| \geq a) = P((X - \mu)^2 \geq a^2) .$$

Since $(X - \mu)^2$ is a nonnegative random variable we can apply Markov's inequality and obtain

$$P(|X - \mu| \geq a) \leq \frac{E[(X - \mu)^2]}{a^2} = \frac{\text{var}(X)}{a^2} . \quad \blacksquare$$

Let us apply this result to our coin flipping example

Example 1.4.4 (Flipping coins, cont'd) Since S_n has mean $n/2$ and variance $n/4$ Chebyshev's inequality tells us that

$$\begin{aligned} P\left(S_n \geq \frac{3n}{4}\right) &= P\left(S_n - \frac{n}{2} \geq \frac{n}{4}\right) \\ &\leq P\left(\left|S_n - \frac{n}{2}\right| \geq \frac{n}{4}\right) \\ &\leq \frac{n/4}{(n/4)^2} = \frac{4}{n} . \end{aligned} \tag{1.7}$$

This is significantly better that the bound provided by Markov's inequality! Note also that we can do a bit better by noting that the distribution of S_n is symmetric around its mean and thus we can replace $4/n$ by $2/n$.

We can better if we know all moments of the random variable X, for example if we know the moment generating function $M_X(t)$ of the random variable X. We have

Proposition 1.4.5 (Chernov's bounds) *Let X be a random variable with moment generating function $M_X(t) = E[e^{tX}]$.*

- *For any a and any $t > 0$ we have*

$$P(X \geq a) \leq \min_{t \geq 0} \frac{E[e^{tX}]}{e^{ta}}.$$

- *For any a and any $t < 0$ we have*

$$P(X \leq a) \leq \min_{t < 0} \frac{E[e^{tX}]}{e^{ta}}.$$

Proof: This follows from Markov inequality. For $t > 0$ we have

$$P(X \geq a) = P(e^{tX} > e^{ta}) \leq \frac{E[e^{tX}]}{e^{ta}}.$$

Since $t > 0$ is arbitrary we obtain

$$P(X \geq a) \leq \min_{t \geq 0} \frac{E[e^{tX}]}{e^{ta}}.$$

Similarly for $t < 0$ we have

$$P(X \leq a) = P(e^{tX} > e^{ta}) \leq \frac{E[e^{tX}]}{e^{ta}},$$

and thus

$$P(X \geq a) \leq \min_{t \leq 0} \frac{E[e^{tX}]}{e^{ta}}. \blacksquare$$

Let us consider again our flipping coin examples

Example 1.4.6 (Flipping coins, cont'd) Since S_n is a binomial $B(n, \frac{1}{2})$ random variable its moment generating function is given by $M_{S_n}(t) = (\frac{1}{2} + \frac{1}{2}e^t)^n$. To estimate $P(S_n \geq 3n/4)$ we apply Chernov bound with $t > 0$ and obtain

$$P\left(S_n \geq \frac{3n}{4}\right) \leq \frac{(\frac{1}{2} + \frac{1}{2}e^t)^n}{e^{\frac{3nt}{4}}} = \left(\frac{1}{2}e^{-\frac{3t}{4}} + \frac{1}{2}e^{\frac{t}{4}}\right)^n.$$

To find the optimal bound we minimize the function $f(t) = \frac{1}{2}e^{-\frac{3t}{4}} + \frac{1}{2}e^{\frac{t}{4}}$. The mimimum is at $t = \log 3$ and

$$f(\log(3)) \;=\; \frac{1}{2}(e^{-\frac{3}{4}\log(3)} + e^{\frac{1}{4}\log(3)}) \;=\; \frac{1}{2}e^{\frac{1}{4}\log(3)}(e^{-\log 3} + 1) \;=\; \frac{2}{3}3^{\frac{1}{4}} \;\simeq\; 0.877$$

and thus we obtain

$$P\left(S_n \geq \frac{3n}{4}\right) \;\leq\; 0.877^n\,.$$

This is course much better than $2/n$. For $n = 100$ Chebyshev inequality tells us that the probability to obtain 75 heads is not bigger than 0.02 while the Chernov bounds tells us that it is actually not greater than 2.09×10^{-6}.

1.5 Limit Theorems

In this section we study the behavior, for large n of a sum of independent identically distributed variables. That is let $X_1, X_2, \cdots$ be a sequence of independent random variables where all X_i's have the same distribution. Then we denote by S_n the sum

$$S_n = X_1 + \cdots + X_n\,.$$

Under suitable conditions S_n will exhibit a universal behavior which does not depend on all the details of the distribution of the X_i's but only on a few of its charcteristics, like the mean or the variance.

The first result is the weak law of large numbers. It tells us that if we perform a large number of independent trials the average value of our trials is close to the mean with probability close to 1. The proof is not very difficult, but it is a very important result!

Theorem 1.5.1 (The weak Law of Large Numbers) *Let $X_1, X_2, \cdots$ be a sequence of independent identically distributed random variables with mean μ and variance σ^2. Let*

$$S_n = X_1 + \cdots + X_n$$

Then for any $\epsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) \;=\; 0\,.$$

Proof: : By the linearity of expectation we have

$$E\left[\frac{S_n}{n}\right] \;=\; \frac{1}{n}E[X_1 + \cdots + X_n] \;=\; \frac{n\mu}{n} \;=\; \mu\,.$$

i.e. the mean of S_n/n is μ. Furthermore by the independence of $X_1, \cdots, X_n$ we have

$$\mathrm{var}\left(\frac{S_n}{n}\right) \;=\; \frac{1}{n^2}\mathrm{var}(S_n)\frac{1}{n^2}\mathrm{var}(X_1+\cdots+X_n) \;=\; \frac{n\sigma^2}{n^2} \;=\; \frac{\sigma^2}{n}\,.$$

Applying Chebyshev's inequality we obtain

$$P\left(\left|\frac{S_n}{n}-\mu\right|\geq\epsilon\right) \;\leq\; \frac{\mathrm{var}\left(\frac{S_n}{n}\right)}{\epsilon^2} \;=\; \frac{\sigma^2}{\epsilon^2}\frac{1}{n}\,,$$

and for any $\epsilon > 0$ the right hand sides goes to 0 as n goes to ∞. ∎

Usually we refer to the quantity $\frac{S_n}{n}$ as the empirical average. The weak law of large numbers tells us that if we perform a large number of independent trials then the average value of our trials is close to the mean with probability close to 1. The proof is not very difficult, but it is a very important result!

We discuss now several refinements of the Law of Large numbers. A look at the proof shows that the probability to observe a deviation from the mean behaves like $1/n$, and that we have used only the fact that the variance is finite. One would expect that we know that higher moments $E[X^n]$ are finite one should obtain better estimates.

For this we need some preparation. Let X be a random variable with m.g.f $\phi(t) = E[e^{tX}]$. It will be useful to consider the logarithm of ϕ which we denote by u

$$u(t) = \log\phi(t) = \log E[e^{tX}]$$

and referred to as the *logarithmic moment generating function*.

Recall that a function $f(t)$ is called *convex* if for any $0 \leq \alpha \leq t$ we have

$$f(\alpha t_1 + (1-\alpha)t_2) \;\leq\; \alpha f(t_1) + (1-\alpha)f(t_2)\,.$$

Graphically it means that the graph of $f(t)$ for $t_1 \leq t \leq t_2$ lies below the line passing through the $(t_1, f(t_1)$ and $(t_1, f(t_2)$. From calculus f is convex iff $f'(t)$ is increasing iff $f''(t)$ is nonnegative (provided the derivatives exist).

Lemma 1.5.2 *The logarithmic moment generating function $u(t) = \log\phi(t)$ is a convex function which satisfies*

$$u(0) = 0\,, \quad u'(0) = \mu\,, \quad u''(0) = \sigma^2$$

Proof: We will prove the convexity using Hölder inequality which states that if $1/p + 1/q = 1$ then $E[XY] \leq E[X^p]^{1/p}E[Y^q]^{1/q}$. We choose $p = \frac{1}{\alpha}$ and $q = \frac{1}{1-\alpha}$ and obtain

$$E\left[e^{(\alpha t_1+(1-\alpha)t_2)X}\right] \;=\; E\left[\left(e^{t_1X}\right)^{\alpha}\right]E\left[\left(e^{t_2X}\right)^{(1-\alpha)}\right] \;\leq\; E\left[\left(e^{t_1X}\right)\right]^{\alpha}E\left[\left(e^{t_2X}\right)\right]^{(1-\alpha)}\,.$$

Taking logarithms proves the convexity.

Note further that

$$\begin{aligned} u'(t) &= \frac{\phi'(t)}{\phi(t)} \\ u''(t) &= \frac{\phi''(t)\phi(t) - \phi'(t)^2}{\phi(t)^2}\,. \end{aligned}$$

If $t = 0$ we find that

$$\begin{aligned} u'(0) &= \frac{\phi'(0)}{\phi(0)} = \mu \\ u''(0) &= \frac{\phi''(0)\phi(0) - \phi'(0)^2}{\phi(0)^2} = \sigma^2\,. \end{aligned}$$

■

Definition 1.5.3 *The legendre transform of a function $f(t)$ is the function defined by*

$$f^*(z) = \sup_t (zt - f(t)) \tag{1.8}$$

Note that the supremum in Eq. 1.8 can be equalt to $+\infty$. If the the supremum is finite we can comute f^* can using calculus if f is differentiable. The supremum is attained at the point t^* such that the derivative of $zt - f(t)$ vanishes, i.e.,

$$z = f'(t^*)\,.$$

Then solving for t^* and inserting in the l.h.s. of (1.8) gives $f^*(t)$.

For future use let us compute the Legendre transform of some logarithmic moment generating functions.

Example 1.5.4 Let $\phi(t) = e^{\mu t + \sigma^2 t^2/2}$ be the m.g.f of $N(0,1)$ and let $u(t) = \log \phi(t) = \mu t + \sigma^2 t^2/2$. Given z the maximum of $zt - \mu t - \sigma^2 t^2/2$ is attained if t^* satisfies

$$z - \mu - \sigma^2 t^* = 0\,, \quad t^* = \frac{z-\mu}{\sigma^2}$$

and thus

$$u^*(z) = zt^* - \mu t^* - \sigma^2 (t^*)^2/2 = \frac{(z-\mu)^2}{2\sigma^2}\,.$$

We see that $u^*(z)$ is a parabola centered around μ.

Example 1.5.5 Let $\phi(t) = (1-p) + pe^t$ be the m.g.f of $B(1,p)$ and let $u(t) = \log \phi(t) = \log((1-p) + pe^t)$. We distinguish three cases

- If $z > 1$ then the function $zt - \log((1-p) + pe^t)$ is increasing since its derivative is $z - \frac{pe^t}{(1-p)+pe^t} > 0$ for all t. The maximum is attained as $t \to \infty$ and is equal to $+\infty$ since

$$\lim_{t\to\infty} zt - log((1-p) + pe^t) \;=\; \lim_{t\to\infty} z(t-1) - log((1-p)e^{-t} + p) \;=\; +\infty\,.$$

- If $z < 0$ then the function $zt - \log((1-p) + pe^t)$ is decreasing for all t and thus the supremum is attained as $t \to -\infty$. The supremum is $+\infty$.

- For $0 \le z \le 1$ the the maximum is attained if t^* satisfies

$$z = \frac{pe^t}{(1-p) + pe^t}\,, \quad t^* = \log\left(\frac{z}{1-z}\frac{1-p}{p}\right)\,,$$

and we obtain

$$u^*(z) \;=\; z\log\left(\frac{z}{p}\right) + (1-z)\log\left(\frac{1-z}{1-p}\right)\,.$$

A simple computation shows that $u^*(z)$ is strictly convex and that $u^*(z)$ has its minimum at $z = p$.

Lemma 1.5.6 *Let $u(t)$ be the logarithmic moment generating function of the random variable X. Then the Legendre transform $u^*(z)$ of $u(t)$ is a convex function which satisfies $u(z) \ge 0$. If $\sigma^2 > 0$ then $u(z) = 0$ if $z = \mu$, i.e. $u^*(z)$ is nonnegative and takes its unique minimum (which is equal to 0) at the mean μ of X.*

Moreover if $z > \mu$ then

$$u^*(z) = \sup_{t\ge 0}(tz - u(t))\,.$$

and if $z < \mu$ then

$$u^*(z) = \sup_{t\le 0}(tz - u(t))\,.$$

Proof:
1) The convexity of $u^*(z)$ follows from

$$\begin{aligned}\alpha u^*(z_1) + (1-\alpha)u^*(z_2) &= \sup_t(\alpha z_1 t - \alpha u(t)) + \sup_t((1-\alpha)z_2 t - (1-\alpha)u(t)) \\ &\ge \sup_t((\alpha z_1 + (1-\alpha)z_2))t - u(t)) \\ &= u^*(\alpha z_1 + (1-\alpha)z_2)\,. \end{aligned} \tag{1.9}$$

2) Next note that $u^*(z) \ge 0z - u(0) = 0$ and thus $u^*(z)$ is nonnegative.
3) Suppose that $u^*(z_0) = 0$ for some z_0. Then $\sup_t(tz - u(t)) = 0$. The supremum is attained at t^* which satisfies the equation $z_0 = u'(t^*)$ and thus we must have

$$f(t^*) = u(t^*) - t^*u'(t^*) \;=\; 0 \tag{1.10}$$

This equation has one solution, namely for $t^* = 0$ since $u(0) = 0$. In that case $z_0 = u'(0) = \mu$. Let us show that this is the only solution. The function $f(t) = u(t) - tu'(t)$ is 0 at $t = 0$ and its derivative is $f'(t) = -tu''(t)$. If $u''(0) = \sigma^2 > 0$ by continuity $u''(t) > 0$ for $t \in (-\delta, \delta)$. Thus $f'(t) > 0$ for $t \in (0, \delta)$ and $f'(t) < 0$ for $t \in (-\delta, 0)$. Therefore 0 is the only solution of $f(0) = 0$.
4) If $z > \mu$ then for $t < 0$ we have

$$zt - u(t) \;\leq\; \mu t - u(t) \;\leq\; \sup_t(\mu t - u(t)) \;=\; u^*(\mu) = 0\,.$$

Since $u^*(z) > 0$ we conclude that the supremum is attained for some $t \geq 0$. One argues similarly for $z < \mu$. ∎

Theorem 1.5.7 (One-half of Cramer's theorem) *Let $X_1, X_2, \cdots$ be a sequence of independent and identically distributed random variables. Assume that the moment generating function $\phi(t)$ of X_i exists and is finite in a neighborhood of 0. Let $u(t) = \log \phi(t)$ and $u^*(z) = \sup_z(zt - u(t))$. Then for any $a > \mu$ we have*

$$P\left(\frac{S_n}{n} > a\right) \;\leq\; e^{-nu^*(a)}$$

and for any $a < \mu$ we have

$$P\left(\frac{S_n}{n} < a\right) \;\leq\; e^{-nu^*(a)}\,.$$

Proof: We use Chernov inequality. Let $a > \mu$, for $t > 0$ we have

$$\begin{aligned}
P\left(\frac{S_n}{n} \geq a\right) &= P\left(\S_n \geq an\right) \\
&= \min_{t\geq 0} e^{-ant} E\left[e^{tS_n}\right] \\
&= \min_{t\geq 0} e^{-ant} (\phi(t))^n \\
&= \min_{t\geq 0} e^{-n(at-u(t)} \\
&= e^{-n\sup_{t\geq 0}(at-u(t)} \\
&= e^{-n\sup_t(at-u(t)} \;=\; e^{-nu^*(a)}\,.
\end{aligned}$$

One proceeds similarly for $a < 0$.

There is a strengthening of the weak law of large numbers called the strong law of large numbers

Theorem 1.5.8 (Strong Law of Large Numbers) *Let $X_1, X_2, \cdots$ be a sequence of independent identically distributed random variables with mean μ. Then S_n/n converges to μ with probability 1, i.e.,*

$$P\left(\lim_{n\to\infty} \frac{S_n}{n} = \mu\right) = 1\,.$$

The strong law of large numbers is useful in many respects. Imagine for example that you are simulating a sequence of i.i.d random variables and that you are trying to determine the mean μ. The strong law of large numbers tells you that it is enough to do ONE simulation for a sufficiently long time it will always reproduce the mean. The weak law of large numbers tells you something a little weaker: with very large probability you will obtain the mean. Based on the weak law of large numbers only you might want to repeat your experiment a number of times to make sure you were not unlucky and hit an event of small probability. The strong law of large numbers tells you not to worry.

Proof: The proof of the strong law of large numbers use more advanced tools that we are willing to use here.

Finally we discuss the Central Limit Theorem. The Law of large number and Cramer's theorem deals with large fluctuations for S_n/n, that is with the probability that S_n/n is at a distance away from the mean which is of order 1. In particular these fluctuations vanish when $n \to \infty$. For example we can ask if there are non trivial fluctuations of order $\frac{1}{n^\alpha}$ for some $\alpha > 0$. One can easily figure out which power α has to be chosen. Since $E[S_n] = n\mu$ $\mathrm{var}(S_n) = n\sigma^2$ we see that the ratio

$$\frac{S_n - n\mu}{\sqrt{n}\sigma}$$

has mean 0 and variance 1 for all n. This means that fluctuation of order $1/\sqrt{n}$ may be non trivial. The Central limit theorem shows not the fluctuation of order $1/\sqrt{n}$ of S_n are in fact universal: for large n they behave like a normal random variable, that is

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \sim N(0,1)\,,$$

or

$$\frac{S_n}{n} \sim \mu + \frac{1}{\sqrt{n}} N(0,\sigma^2)\,.$$

What we exactly mean by $\sim$ is given in

Theorem 1.5.9 (Central Limit Theorem) *Let $X_1, X_2, \cdots$ be a sequence of independent identically distributed random variables with mean μ and variance $\sigma^2 > 0$. Then for any $-\infty \le a \le b\infty$ we have*

$$\lim_{n\to\infty} P\left(a \le \frac{S_n - n\mu}{\sqrt{n}\sigma} \le b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2} 2\, dx\,.$$

Proof: We will not give the complete proof here but we will prove that the moment generating function of $\frac{S_n - n\mu}{\sqrt{n}\sigma}$ converges to the moment generating of $N(0,1)$ as $n \to \infty$.

Let by $X_i^* = \frac{X_i - \mu}{\sigma}$ then $E[X_i^*] = 0$ and $\text{var}(X_i^*) = 1$. If $S_n^* = S_1^* + \cdots X_n^*$ then

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \;=\; \frac{S_n^*}{\sqrt{n}}\,.$$

Therefore without loss of generality we can assume that $\mu = 0$ and $\sigma = 1$.

If ϕ be the moment generating function of X_i then $\phi'(0) = 0$ and $\phi''(0) = 1$. We denote the moment generating function of $S_n/\sqrt{n}$ by $\phi_n(t)$. Using independence we have

$$\phi_n(t) \;=\; E\left[e^{t\frac{S_n}{\sqrt{n}}}\right] \;=\; E\left[e^{\frac{t}{\sqrt{n}}(X_1+\cdots X_n)}\right] \;=\; \left(\phi\left(\frac{t}{\sqrt{n}}\right)\right)^n\,.$$

Recall that the m.g.f. of $N(0,1)$ is given by $e^{t^2/2}$, so we need to show that $\phi_n(t) \to e^{t^2/2}$ as $n \to \infty$. To show this we set $u(t) = \log\phi(t)$ and $u_n(t) = \log\phi_n(t)$ and show that $u_n(t) \to t^2/2$ as $n \to \infty$. We have

$$u_n(t) \;=\; \log\phi_n(t) \;=\; n\log\phi\left(\frac{t}{\sqrt{n}}\right)\,.$$

Note that

$$\begin{aligned}
u(0) &= \log\phi(0) \;=\; 0\\
u'(0) &= \frac{\phi'(0)}{\phi(0)} \;=\; \mu \;=\; 0\\
u''(0) &= \frac{\phi''(0)\phi(0) - (\phi'(0))^2}{(\phi(0))^2} \;=\; \sigma^2 = 1\,.
\end{aligned}$$

By using L'Hospital rule twice we obtain

$$\begin{aligned}
\lim_{n\to\infty} u_n(t) &= \lim_{s\to\infty} \frac{\phi(t/\sqrt{s})}{s^{-1}}\\
&= \lim_{s\to\infty} \frac{\phi'(t/\sqrt{s})t}{2s^{-1/2}}\\
&= \lim_{s\to\infty} \phi''(t/\sqrt{s})\frac{t^2}{2} \;=\; \frac{t^2}{2}\,.
\end{aligned}$$

Therefore $\lim_{n\to\infty}\phi_n(t) = e^{t^2}2$. From this it seems at least plausible that the c.d.f of $S_n/\sqrt{n}$ converges to the c.d.f of $N(0,1)$ but proving this requires some serious work and some Fourier analysis. ∎

1.6 Monte-Carlo methods

The basic Monte-Carlo method uses sum of independent random variables and the law of large numbers to estimate a deterministic quantity. In order to illustrate the method let us start by an example.

Example 1.6.1 (Estimating the number π.) We construct a random algorithm to generate the number π. Consider a circle of radius 1 that lies inside a 2×2 square. The square has area 4 and the circle has area π. Suppose we pick a point at random within the circle and define

$$X = \begin{cases} 1 & \text{if the point is inside the circle} \\ 0 & \text{otherwise} \end{cases}$$

and $P(X = 1) = \pi/4$. We repeat this experiment n times. That is we we select n points inside the circle independently. The number of points S_n within the circle can be written as $S_n = X_1 + \cdots + X_n$ where the X_i's are independent copies of X. So $S_n = B(n, \frac{\pi}{4})$ and $E[S_n] = n\pi/4$.

Suppose now that perform $n = 10'000$ trials and observe $S_n = 7932$, then our estimator for π is $4\frac{7932}{10000} = 3.1728$. We try to determine how accurate our estimator is.

By the Central Limit Theorem, $\frac{S_n - n\mu}{\sqrt{n}\sigma}$ has for sufficiently large n a normal distribution $N(0, 1)$. Therefore

$$P\left(\frac{S_n}{n} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \frac{S_n}{n} + 1.96\frac{\sigma}{\sqrt{n}}\right) \cong 0.95\,,$$

for sufficiently large n. The value $x = 1.96$ is such that $P(|N(0, 1)| \leq x) = 0.95$. For this reason we call the interval $\left[\frac{S_n}{n} - 1.96\frac{\sigma}{\sqrt{n}}\,,\, \frac{S_n}{n} + 1.96\frac{\sigma}{\sqrt{n}}\right]$ a 95% confidence interval.

In our case a 95% confidence interval for $\pi/4$ is

$$\left[\frac{S_n}{n} - 1.96\frac{\sigma}{\sqrt{n}}\,,\, \frac{S_n}{n} + 1.96\frac{\sigma}{\sqrt{n}}\right]\,.$$

where $\sigma = \sqrt{\frac{\pi}{4}(1 - \frac{\pi}{4})}$ which we can't really evaluate since we do not know π. There are several ways to proceed

1. Use the simple bound $x(1 - x) \leq \frac{1}{4}$ so $\sigma \leq \frac{1}{2}$ and thus

$$1.96\frac{\sigma}{\sqrt{n}} \leq 1.96\frac{1}{2\sqrt{n}} = 0.0098\,.$$

This gives the interval $[3.1336, 3.2120]$ for a conservative 95% confidence interval.

2. We can simply use our estimate for π into the formula $\sigma = \sqrt{\frac{\pi}{4}(1-\frac{\pi}{4})} \cong 0.405$. This gives a confidence interval of $[3.1410, 3.2046]$.

3. Another way to estimate the variance when σ^2 is unknown is to use the *sample variance* given by

$$V_n^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \frac{S_n}{n})^2 \,.$$

The sample is an unbiased estimator since we have $E[V_n^2] = \sigma^2$. To see this note we can assume that $\mu = 0$ and then we have

$$E[V_n^2] \;=\; \frac{1}{n-1}\sum_{i=1}^{n} E\left[X_i^2 - 2X_i\frac{S_n}{n} + \left(\frac{S_n}{n}\right)^2\right] \;=\; \frac{n}{n-1}\sigma^2(1-\frac{2}{n}+\frac{1}{n^2}n) \;=\; \sigma^2 \,.$$

This example is a particular case of the *hit-or-miss method*. Suppose you want to estimate the volume of the set B in $\mathbf{R}^d$ and that you know the volume of a set A which contains B. The hit-or-miss method consists in choosing n points in A uniformly at random and use the fraction of the points that land in B as an estimate for the volume of B.

Another class of examples where Monte-Carlo methods can be applied is the computation of integrals. Suppose you want to compute the integral

$$I_1 \;=\; \int_0^1 \frac{e^{\sqrt{x}} - e^{cos(x^3)}}{3+\cos(x)}\,dx \,.$$

or more generally

$$I_2 \;=\; \int_S h(\mathbf{x})d\mathbf{x}$$

where S is a subset of $\mathbf{R}^d$ and h is a given real-valued function on S. A special example is the function $h = 1$ on S in which case you are simply trying to compute the volume of S. Another example is

$$I_3 \;=\; \int_{\mathbf{R}^d} h(\mathbf{x})f(\mathbf{x})\,d\mathbf{x} \,.$$

where h is a given real-valued function and f is a p.d.f of some random vector on $\mathbf{R}^d$. All these examples can be written as expectations of a suitable random variable. Indeed we have

$$I_3 \;=\; E[h(\mathbf{X})] \quad \text{where } \mathbf{X} \text{ has p.d.f } f(\mathbf{x}) \,.$$

We have also

$$I_1 \;=\; E[h(U)] \quad \text{where } U = U([0,1]) \,.$$

To write I_2 has an expectation choose a random vector such that its p.d.f f satisfies $f(\mathbf{x}) > 0$ for every $\mathbf{x} \in S$. Extend k to $\mathbf{R}^d$ by setting $k = 0$ if $x \notin S$. Then

$$I_2 \;=\; \int_{\mathbf{R}^d} k(\mathbf{x})\,dx \;=\; \int_{\mathbf{R}^d} \frac{k(\mathbf{x})}{f(\mathbf{x})} f(\mathbf{x})\,d\mathbf{x} \;=\; E\left[\frac{k(\mathbf{X})}{f(\mathbf{X})}\right] \,.$$

Note that you have a considerable freedom in choosing f and this is what lies behind the idea of *importance sampling* (see Example 1.6.3 below).

Many many other problems can be put in the form (maybe after some considerable work)

$$I = E\left[h(X)\right] ,$$

and the random variable X could also be a discrete random variable.

Algorithm 1.6.2 (Simple sampling) *In order to estimate $I = E\left[h(X)\right]$ the simple samling consists in generating i.i.d. random variables $X_1, X_2, \cdots, X_n$ and set*

$$I_n \equiv \frac{1}{n}\sum_{i=1}^{n} h(X_i) .$$

The quantity I_n gives an unbiased estimate of I (i.e. $E\left[I_n\right] = I$). By the strong law of large numbers I_n converges to I with probability 1 as $n \to \infty$.

The variance of the simple sampling estimate is

$$\mathrm{var}(I_n) \;=\; \frac{\mathrm{var}(h(X))}{n} .$$

Note that the variance $\mathrm{var}(I_n)$ can be used to determine the accuracy of our estimate, for example by determining a 95% confidence interval as in Example 1.6.1. If we denote $\sigma^2 = \mathrm{var}(h(X))$ then the half length of 95% confidence interval is given by $1.96\sigma/\sqrt{n}$. If we wish our confidence to half length ϵ we need to choose n such that

$$n \geq \frac{\epsilon^2}{(1.96)^2\sigma^2} .$$

So, as a rule the accuracy of the method is of order $\sqrt{n}$.

We consider next another example which illustrate one technique through which one can reduce the variance considerably (*variance reduction*). The technique we will use goes under the name of *importance sampling*. Suppose we want to compute $E[h(X)]$. We can use simple sampling by simulating i.i.d random variables with p.d.f. $f(x)$. Instead of using X we can choose another random variable Y with p.d.f $g(x)$ and write

$$E[h(x)] \;=\; \int h(x)f(x)dx \;=\; \int \frac{h(x)f(x)}{g(x)}g(x)\,dx \;=\; E\left[\frac{h(Y)f(Y)}{g(Y)}\right] .$$

We then simulate i.i.d rndom variables Y_i with p.d.f g and this gives a new estimator

$$J_n \;=\; \frac{1}{n}\sum_{j=1}^{n} \frac{h(Y_j)f(Y_j)}{g(Y_j)} ,$$

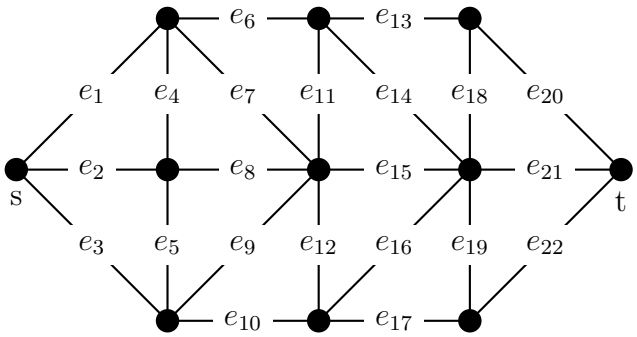

Figure 1.1: A graph with 22 edges

The variance is given by

$$\begin{aligned}\mathrm{var}(J_n) &= \frac{1}{n}\mathrm{var}\left(\frac{h(Y)f(Y)}{g(Y)}\right)\\ &= \frac{1}{n}\left(\int\left(\frac{h(x)f(x)}{g(x)}\right)^2 g(x)\,dx - \left(\int\frac{h(x)f(x)}{g(x)}h(x)f(x)\right)^2\right)\end{aligned}$$

The idea of importance sampling is to choose Y such that

$$\mathrm{var}\left(\frac{h(Y)f(Y)}{g(Y)}\right) < \mathrm{var}(h(X))\,,$$

and thus to improve the efficiency of our method.

There are many other methods to reduce the variance and some are touched upon in the exercises.

Example 1.6.3 (Network reliability) Let us consider an application of simple sampling to nework reliability. Consider a connected graph as in Figure 1.1. Each edge as a probability q of failing and all edges are independent. Think of q has a very small number, to fix the idea let $q = 10^{-2}$. Fix two vertices s and t and we want to compute the disconnection probability

$$p_D \equiv P\,(s \text{ is not connected to } t \text{ by working edges})$$

This can be computed by hand for very small graphs but even for the graph shown in Figure 1.1 this is hardly doable. Our graph here has 22 edges and let $\mathcal{E} = \{e_1 \cdots e_{22}\}$ denote the set of all edges. Let X denote the set of edges that fail, so X is a random subset of $\mathcal{E}$. So for every $B \subset \mathcal{E}$ we have

$$P(X = B) = q^{|B|}(1-q)^{|\mathcal{E}|-|B|}$$

where $|A|$ denotes the cardinality of A. If we denote by $\mathcal{S}$ the set of all subsets of $\mathcal{E}$ then X is a random variable which takes value in $\mathcal{S}$.

Let us define the function $k : \mathcal{S} \to \mathbf{R}$ by

$$k(B) = \begin{cases} 1 & \text{if } s \text{ is not connected to } t \text{ when the edges of } B \text{ fail} \\ 0 & \text{if } s \text{ is connected to } t \text{ when the edges of } B \text{ fail} \end{cases}$$

Then we have

$$p_D = \sum_{B\,;\,k(B)=1} P(X = B) = \sum_B k(B)P(X = B) = E[k(X)]\,.$$

The simple sampling estimator for p_D is

$$\frac{1}{n}\sum_{i=1}^{n} k(X_i)$$

where $X_1, \cdots X_n$ are i.i.d copies of X. Each X_i can be generated by tossing an unfair coin 22 times. Then our estimator is simply the fraction of those simulated networks that fail to connect s and t.

In order to get an idea of the number involved let us give a rough estimate of p_D. It is easy to see that at least 3 nodes must fail for s not to be connected to t. So we have

$$p_D \leq P(|X| \geq 3) = 1 - \sum_{j=0}^{2} \binom{22}{j} q^j (1-q)^{22-j} \cong 0.00136\,,$$

since $|X| = B(22, q)$.

On the other hand we can get a lower bound for p_D by noting that

$$p_D \geq P(e_1, e_2, e_3 \text{ fail}) = q^3 = 10^{-6}\,.$$

Therefore p_D is between 10^{-2} and 10^{-6} which is very small. We will thus need very tight confidence intervals. To compute $\text{var}(I_n)$ note that $k(X)$ is a Bernoulli random variable with parameter p_D. Hence

$$\text{var}(I_n) = \frac{1}{n} p_D(1 - P_D) \cong p_D\,,$$

since p_D is small. To get a meaningful confidence interval we need its half length $2\sqrt{p_D/n}$ to be at the very least less than $p_D/2$. This implies however that we must choose $n > 16/p_D$, and thus we need millions of iterations for a network which is not particularly big.

Let us use importance sampling here. Note that $E[k(x)]$ is very small which means that typical X have $k(X) = 0$. The basic idea is to choose the sampling variable in

such a way that we sample more often the X for which $k(X) = 1$ (i.e., large in our case).

A natural try is take the random variable Y to have a distribution $\phi(D) = P(y = B) = \theta^B(1-\theta)^{22-|B|}$ with a well chosen θ. Since $k(Y) = 0$ whenever $|Y| > 3$ we can for example choose θ such that $E[|Y|] = 3$. Since $|Y| = B(22, \theta)$ this gives $E[|Y|] = 22\theta$ and thus $\theta = 3/22$.

The estimator is now

$$J_n = \frac{1}{n}\sum_{i=1}^{n} \frac{k(Y_i)p(Y_i)}{\phi(Y_i)}$$

where Y_j are i.i.d with distribution $\phi(Y)$. Let us compute the variance of J_n. We have

$$\begin{aligned} \mathrm{var}(J_n) &= \frac{1}{n}\left(\sum_B \frac{k(B)^2 p(B)^2}{\phi(B)^2}\phi(B) - p_D^2\right) \\ &= \frac{1}{n}\left(\sum_{B\,:\,k(B)=1} \frac{p(B)}{\phi(B)} p(B) - p_D^2\right). \end{aligned} \tag{1.11}$$

Note that

$$\frac{p(B)}{\phi(B)} = \frac{q^B(1-q)^{22-|B|}}{\theta^B(1-\theta)^{22-|B|}} = \left(\frac{1-q}{1-\theta}\right)^{22}\left(\frac{q(1-\theta)}{\theta(1-q)}\right)^{|B|} = 20.2 \times (0.064)^{|B|}.$$

In Eq. (1.11) all terms with $k(B) = 1$ have $|B| \geq 3$. For those B we have

$$\frac{p(B)}{\phi(B)} \leq 20.2 \times (0.064)^3 \leq 0.0053$$

So we get

$$\mathrm{var}(J_n) \leq \frac{1}{n}\sum_{B\,:\,k(B)=1} 0.0053\; p(B) = \frac{0.0053\, p_D}{n}$$

This means that we have reduced the variance by a factor approximately of 200. So for the same n the confidence interval is going to be about $\sqrt{200} \cong 14$ times smaller. Alternatively a given confidence interval for I_n can be obtained for $J_{n/200}$. This is good.

1.7 Problems

Exercise 1

1. For positive numbers a and b, the $Pareto(a, b)$ distribution has p.d.f $f(x) = ab^a x^{-a-1}$ for $x \geq b$ and $f(x) = 0$ for $x < b$. Apply the inversion method to generate $Pareto(a, b)$.

2. The standardized logistic distribution has the p.d.f $f(x) = \frac{e^{-x}}{(1+e^{-x})^2}$. Use the inversion method to generate a random variable having this distribution.

Exercise 2 Consider the technique of generating a $Gamma(n, \lambda)$ random variable by using the rejection method with $g(x)$ being the p.d.f of an exponential with parameter λ/n.

1. Show that the average number of iterations of the algorithm is $n^n e^{1-n}/(n-1)!$.
2. Use Stirling formula to show that for large n the answer in 1. is approximately $e\sqrt{(n-1)/2\pi}$.
3. Show that the rejection method is equivalent to the following
   - **Step 1:** Generate Y_1 and Y_2 independent exponentials with parameters 1.
   - **Step 2:** If $Y_1 < (n-1)[Y_2 - \log(Y_2) - 1]$ return to step 1.
   - **Step 3:** Set $X = nY_2/\lambda$.

Exercise 3 (Generating a uniform distribution on the permutations) In this problem we will use the following notation. If x is positive real number we denote by $[x]$ the integer part of x, i.e. $[x]$ is the greatest integer less than or equal x. For example $[2.37] = 2$.

Consider the following algorithm to generate a random permutation of the elements $1, 2, 3, \cdots, n$. We will denote by S(i) the element in position i. For example for the permutation $(2, 4, 3, 1, 5)$ of 5 elements we have $S(1) = 2$, $S(2) = 4$, and so on.

1. Set $k = 1$
2. Set $S(1) = 1$
3. If $k = n$ stop. Otherwise let $k = k + 1$.
4. Generate a random number U, and let

$$S(k) = S([kU] + 1)\,,$$

$$S([kU] + 1) = k\,.$$

   Go to step 3.

Explain, in words, what the algorithm is doing. Show that at iteration k, – i.e. when the value of $S(k)$ is initially set– $S(1), S(2), \cdots, S(k)$ is a random permutation of $1, 2, \cdots, k$.
Hint: Relate the probability P_k obtained at iteration k with the probability P_{k-1} obtained at iteration $k-1$.

Exercise 4 Compute the Legendre transform u^* of the logarithmic m.g.f u for the random variables $Poisson(\lambda)$ and $Exp(\lambda)$. Discuss in details where u^* is finite or not.

Exercise 5 We have seen in class that if X_i are independent $B(1,p)$ Bernoulli random variable and $S_n = X_1 + \cdots + X_n$ then for $a > p$

$$P\left(\frac{S_n}{n} \geq a\right) \leq e^{-nu_p^*(a)}$$

where

$$u_p^*(z) = \begin{cases} z\log\left(\frac{z}{p}\right) + (1-z)\log\left(\frac{1-z}{1-p}\right) & \text{if } 0 \leq \text{z} \leq 1 \\ +\infty & \text{otherwise} \end{cases} .$$

and a similar bound for $a < p$. In order to make that bound more easy to use in practice

1. Show that for $a > 0$ and $0 < p < 1$ we have

$$u_p^*(z) - 2(z-p)^2 \geq 0\,.$$

Hint: Differentiate twice.

2. Show that for any $\epsilon > 0$

$$P\left(\left|\frac{S_n}{n} - p\right| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2}\,.$$

Exercise 6 On a friday night you enter a fast food restaurant which promises that every customer is served within a minute. Unfortunately there are 30 customers in line and you an appointment will force you to leave in 40 minutes. Being a probabilist you assume that the waiting time of each customer is exponential is mean 1. Estimate the probability that you will miss your appointment if you wait in line until you are served using (a) Chebyshev inequality, (b) The central limit theorem, (c) Cramer's theorem.

Exercise 7 Consider the problem of estimating π that we have considered in class. Estimate the number of trials you should perform to ensure that with probability $1-\delta$ your result is at distance no more than ϵ from the true value of π. Do this in two ways: (1) Use the central limit theorem, (b) Use the estimates of Exercise 5 and Cramer's theorem.

Exercise 8 (Hit-or-missmethod)

1. Suppose that you wish to estimate the volume of a set B contained in the Euclidean space $\mathbf{R}^k$. You know that B is a subset of A and you know the volume of A. The "hit-or-miss" method consists in choosing n independent points uniformly at random in A and use the fraction of points which lands in B to get an estimate of the volume of B. (We used this method to compute the number π in class.) Write down the estimate I_n obtained with this method and compute $\text{var}(I_n)$.

2. Suppose now that D is a subset of A and that we know the volume of D and the volume of $D \cap B$. You decide to estimate the volume of B by choosing n points at random from $A \setminus D$ and counting how many land in B. What is the corresponding estimator I'_n of the volume of B for this second method? Show that this second method is better than the first one in the sense that $\mathrm{var}(I'_n) \leq \mathrm{var}(I_n)$.

3. How would use this method concretely to the estimation of the number π? Compute the corresponding variances.

Exercise 9

Suppose f is a function on the interval $[0,1]$ with $0 < f(x) < 1$. Here are two ways to estimate $I = \int_0^1 f(x)dx$.

1. Use the "hit-or-miss" from the previous problem with $A = [0,1] \times [0,1]$ and $B = \{(x,y) : 0 \leq x \leq 1\,,\, 0 \leq y \leq f(x)\}\,.$

2. Let $U_1, U_2, \cdots$ be i.i.d. uniform random variables on $[0,1]$ and use the estimator

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} f(U_i)\,.$$

   Show that that $\hat{I}_n$ has smaller variance than the estimator of (a).

Exercise 10 (Antithetic variables) In this problem we describe an example of a method to reduce the variance of the simple sampling method.

1. Suppose that k and h are both nondecreasing (or both nonincreasing) functions then show that

$$\mathrm{cov}(k(X), h(X)) \geq 0\,.$$

   Hint: Let Y be a random variable which is independent of X and has the same distribution as X. Then by our assumption on h, k we have $(k(X)-k(Y))(h(X)-h(Y)) \geq 0$. Take then expectations.

2. Consider the integral $I = \int_0^1 k(x)dx$ and assume that k is nondecreasing (or nonincreasing). The simple sampling estimator is

$$I_n = \frac{1}{n}\sum_{i=1}^{n} k(U_i)\,.$$

   where U_i are independent $U([0,1])$ random variables. Consider now the alternative estimator: for n even set

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n/2} k(U_i) + k(1-U_i)\,.$$

where U_i are independent $U([0,1])$ random variables. Show that I_n is an estimator for I and that $\text{var}(\hat{I_n}) \leq \text{var}(I_n)$.
Hint: Use part 1. to show $\frac{1}{2}\text{var}(k(U_1) + k(1-U_1)) \leq \text{var}(k(U_1))$.

3. Let $k(x) = 4\sqrt{1-x^2}$. Then $I = \pi$. Compute $\text{var}(\hat{I_n})$ and $\text{var}(I_n)$.