General and MC stuff

We consider the one-dimensional integral of an arbitrary function $g(x)$ for $x \in [0,1]$ with $0 < g(x) \leq b, \forall x$.

$$I = \int_0^1 g(x)dx.$$

To estimate the integral value, you are asked to use the following two Monte Carlo estimators and a fixed number of samples N in each estimator.

$$I_1 = \frac{1}{N}\sum_{i=1}^{N} g(x_i).$$

Here, the x_i are independent samples from a uniform distribution in $[0,1]$.

$$I_2 = \frac{1}{N}\sum_{i=1}^{N} b\, h(x_i, y_i).$$

Here, the x_i are again independent samples from a uniform distribution in $[0,1]$, whereas the y_i are independent samples from a uniform distribution in $[0,b]$. Moreover,

$$h(x,y) = \begin{cases} 1 & y < g(x), \\ 0 & \text{otherwise.} \end{cases}$$

Hint:

$$\int_0^b h(x,y)dy = \int_0^{g(x)} dy.$$

19 a) **[4 points]** Show that $\mathbb{E}[I_1] = I$

b) Show that $\mathrm{E}[I_2] = I$

c) **[8 points]** Show that $\mathrm{Var}[I_2] \geq \mathrm{Var}[I_1]$. Which Monte Carlo estimator would you choose to obtain on average the smaller error for a fixed N?

c) **[8 points]** Show that $\mathrm{Var}[I_2] \geq \mathrm{Var}[I_1]$. Which Monte Carlo estimator would you choose to obtain on average the smaller error for a fixed N?

Consider the following integral:

$$I = \int_0^1 f(x)\,dx, \quad f(x) = x.$$

a) **[4 points]** We want to solve this integral by using the following Monte Carlo estimator:

$$I_1 = \frac{1}{N}\sum_{i=1}^{N} f(x_i).$$

The x_i are the samples from a uniform distribution in $[0,1]$ and N is the total number of
20 samples used. What is the variance of the Monte Carlo estimator I_1?

b) **[8 points]** Consider the following transformation of the above integral

$$I = \frac{1}{2} \int_0^1 (f(x) + f(1-x))\, dx$$

and the corresponding estimator

$$I_2 = \frac{1}{2N} \sum_{i=1}^{N} (f(x_i) + f(1-x_i)),$$

where x_i are samples from a uniform distribution in $[0,1]$ and N is the number of samples. What is the variance of I_2?

c) **[8 points]** Explain why the variance of I_2 is lower than that of I_1.

In this exercise we fix the notation we will use during this course and refresh our memory on basic properties of random variables. Present your answers *in detail*.

a) A random variable with normal (or Gaussian) distribution $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ has probability density function (pdf) given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{1}$$

Show that the mean and the variance of X are given by $\mathbb{E}\left[X\right] = \mu$ and $\mathbb{E}\left[(X-\mu)^2\right] = \sigma^2$, respectively.

b) The probability that a random variable X with pdf f_X is less or equal than any $x \in \mathbb{R}$ is given by,

$$P(X \leq x) \;=\; F_X(x) \;=\; \int_{-\infty}^{x} f_X(z)\; \mathrm{d}z\,. \tag{2}$$

The function F_X is called the cumulative distribution function (cdf).
The Laplace distribution with parameters μ and β has pdf,

$$f(x) = \frac{1}{2\beta} \exp\left(-\frac{|x-\mu|}{\beta}\right). \tag{3}$$

i) Find the cdf of the Laplace distribution.

ii) Use the cdf to find the median of the Laplace distribution.

c) The pdf of the quotient $Q = X/Y$ of two random variables X, Y is given by,

$$f_Q(q) \;=\; \int_{-\infty}^{\infty} |x|\, f_{X,Y}(qx, x)\; \mathrm{d}x\,, \tag{4}$$

where $f_{X,Y}$ is the joint pdf of X and Y.
Assume that X and Y are independent random variables with pdfs $f_X(x) = \mathcal{N}\left(x|0,\sigma_X^2\right)$ and $f_Y(y) = \mathcal{N}\left(y|0,\sigma_Y^2\right)$.

1 i) Find the joint pdf of X and Y.

ii) Show that $Q = X/Y$ follows a Cauchy distribution with zero location parameter and scale $\gamma = \sigma_X/\sigma_Y$. The pdf of a Cauchy distribution with location parameter x_0 and scale γ is given by,

$$f(x) = \frac{1}{\pi} \frac{\gamma}{(x - x_0)^2 + \gamma^2}. \tag{5}$$

Let's say that a genetic disorder occurs in the United States population at a rate of 2% (i.e., if D is the event that an individual sampled from this population has the genetic disorder, then $P(D) = \frac{1}{50}$, $P(D^C) = 1 - \frac{1}{50} = \frac{49}{50}$).

A company that markets genetic testing services has produced a test for detecting whether an individual has this genetic disorder. According to the company, their test can accurately detect positive cases at a rate of 99% and can accurately detect negative cases at a rate of 90%:

$$P(+|D) = \frac{99}{100}, \quad P(-|D^C) = \frac{9}{10}$$

a) From the above information, let's say you take this test and receive a positive result, what is the probability that you *actually* have this disorder? In math, what is $P(D|+) =$?

b) From the company's perspective, they have no control over the rates of the disorder in the general population $(P(D), P(D^C))$. However, they do have some control over the false positive rate of the test they've produced, $P(+|D^C)$. What is the maximum false positive rate their test can produce if the company wishes for a positive test result to indicate a true positive 90% of the time (i.e., what is the maximum $P(+|D^C)$ if the company wishes for
2 $P(D|+) = \frac{9}{10}$)? What about 99% of the time?

a) Find the tangential and normal lines to the curve $y = x^2$ at the point $(1, 1)$. The *normal line* is the line that is perpendicular to the tangent line and that passes through the point of tangency.

b) Scaling t and $y(t)$ can reduce the equation

$$y_{tt} + a y_t + b y = 0$$

to the form

$$u_{\tau\tau} + \epsilon u_\tau + u = 0.$$

What is ϵ in terms of a and b (a and $b > 0$ are constants)?

c) Show that $u(t, x) = \frac{a+x}{b+t}$ is a solution of the equation $u_t + u\, u_x = 0$. What initial condition ($t = 0$) does the solution correspond to (a and $b > 0$ are constants)?

d) Consider the boundary value problem

$$-u_{xx} = 1, \quad u(0) = u(1) = 0$$

and its finite difference approximation

$$-\frac{v_{j-1} - 2v_j + v_{j+1}}{h^2} = 1, \quad j = 1, \ldots n, \quad v_0 = v_{n+1} = 0$$

where $n \geq 1$ is an integer, v_j is the finite difference approximation of $u(jh)$ and $h = 1/(n+1)$. Show that

$$v_j = (jh)(1 - jh)/2$$

is the finite difference solution. Compare it to the exact solution at the grid points $x_j = jh$.

e) The solution of the initial value problem

$$y_t = y^2, \quad y(0) = 2$$

"blows up" to infinity in a finite time. Find that time.

3

a) The concentration $c(\mathbf{x}, t)$ of a pollutant in a body of water diffuses according to

$$\frac{\partial c}{\partial t} = \nu \nabla^2 c,$$

with initial conditions $c(\mathbf{x}, 0) = c_0(\mathbf{x})$.

Show that the total concentration remains constant.

b) Show that

$$c_g(\mathbf{x}, t) = \frac{1}{\sqrt{t}} e^{-x^2/4\nu t}$$

is a solution of the diffusion equation.

4

Let $P(E)$ be the probability of an event $E \subset \Omega$ and $P(\Omega) = 1$. We denote the sample space as Ω. The three axioms of probabilities are:

- $P(E) \geq 0$ for all $E \subset \Omega$
- $P(\Omega) = 1$
- If $E_1, E_2, \cdots \subset \Omega$ and they are pairwise disjoint, i.e. $E_i \bigcap E_j = \varnothing$ when $i \neq j$, then $P(E_1 \cup E_2 \cup \ldots) = P(E_1) + P(E_2) + \cdots$

Given the three axioms, answer the following questions **[5 points each]**:

1. If $\bar{E}$ is the complement of E (i.e. $\bar{E}$ contains all the events that are not in E), show that $P(\bar{E}) = 1 - P(E)$
2. Assume, two events E_1 and E_2 are given with $E_1 \subset E_2$. Show that $P(E_2) = P(E_1) + P(E_2 \cap \bar{E}_1) \geq P(E_1)$
3. Show that $P(\cup_{i=1}^{n} E_i) \leq \sum_{i=1}^{n} P(E_i)$
4. Two event A and B are given with $0 < P(A) < 1$ and $0 < P(B) < 1$. Given A and B are disjoint, can the two events be independent ?
5. Two events A and B are given with $0 < P(A) < 1$ and $0 < P(B) < 1$. Given $A \subset B$, can the two events be independent ?

5

function:

$$f(\mathbf{x}) = 2^{-d} \sum_{i=1}^{d} x_i^2 \,,$$

where $\mathbf{x}$ is a vector $\mathbf{x} = (x_1, \ldots, x_d)$.

a) **[2 points]** Calculate the integral analytically

$$I = \int_V f(\mathbf{x}) \; \mathrm{d}\mathbf{x} \,,$$

6 where V is a d-dimensional cube with bounds $[-1, 1]^d$.

You are in charge of a factory that produces mechanical springs. A spring is produced successfully if its stiffness K is larger than 10 $[N/m]$. The stiffnesses of all produced springs are independent from each other and follow a known distribution f_K. You want to find the unknown success rate q of the production process using Monte Carlo.

a) Find a suitable Monte Carlo estimator q_N for the production sucess rate given N independent random samples from the distribution f_K.

b) Find the expected value as well as the variance of the Monte Carlo estimator q_N.

c) We define the relative error of the Monte Carlo estimator as:

$$\alpha_N = \frac{q_N - q}{q} \tag{1}$$

Find the expected value of the squared relative error. Your result should depend on both N and q.

d) Assume the unknown true success rate is $q = \frac{1}{10001}$. How many sample N are necessary

9 such that the squared relative error is smaller than 0.01 ?