Project 1: Monte Carlo methods

Due Thursday 6 October 2022

1 Warmup: life without a CLT

Consider a Pareto-distributed random variable X, with probability density

$$p_X(x) = \begin{cases} \frac{\alpha}{x^{\alpha+1}} & \text{if } x \geq 1 \\ 0 & \text{if } x < 1 \end{cases}.$$

Let $\alpha = 3/2$. (Note that the variance of X is infinite for $\alpha \leq 2$.) Write a simple Monte Carlo method to estimate the mean of X, i.e., $\mu_X \coloneqq \mathbb{E}[X]$.

Let $\bar{x}_n$ denote your n-sample Monte Carlo estimator of μ_X. The goal of this problem is to investigate, numerically, the convergence of this estimator. To do so, run many (10^3 or more) independent "replicate" sequences $(\bar{x}_n)_{n\in\mathbb{N}}$.

(a) Describe the qualitative characteristics of the each sequence $(\bar{x}_n)_{n\in\mathbb{N}}$.

(b) Examine the sampling distribution of the estimator $\bar{x}_n$ at various values of n. In particular, use quantitative diagnostics to measure how this distribution departs from normality. (For example, what are the variance, skewness, kurtosis, etc. of this distribution as a function of n, and how do they depart from those predicted by the central limit theorem? What about quantiles of the sampling distribution?) Does there seem to be an asymptotic distribution for any scaled version of this Pareto sum?[1]

[1] If you would like to understand this behavior better, after performing your numerical experiments, please see I. Zaliapin, Y. Kagan, F. Schoenberg, "Approximating the distribution of Pareto sums," *Pure and Applied Geophysics*, **162**: 1187–1228 (2005).

2 Rejection sampling versus importance sampling

Given a target density π from which one would like to simulate (or compute expectations), both *rejection sampling* and *importance sampling* rely on simulating instead from an instrumental/biasing distribution with density g. A sufficient condition for a valid g in both algorithms is the same, e.g., $\pi(x)/g(x)$ should be uniformly bounded over the support of π. Given the same g, is natural to then wonder which approach is better—i.e., which approach yields Monte Carlo estimates with smaller variance and/or mean-square error?

Here we will explore this question for one of the example problems used in lecture. Consider the following integral:

$$I = \int_{-\infty}^{\infty} x^2 \, \pi(x) \, dx$$

where $\pi(x) = \tilde{\pi}(x)/\beta$,

$$\tilde{\pi}(x) = e^{-x^2/2} \left(\sin^2(6x) + 3\cos^2(x)\sin^2(4x) + 1 \right),$$

and the normalizing constant β is chosen so that $\int_{-\infty}^{\infty} \pi(x) \, dx = 1$. In other words, $\pi(x)$ is a properly normalized density, but we presume (as is typical in many applications) that we can only evaluate the unnormalized density $\tilde{\pi}(x)$.

Choose g to be a standard normal density, i.e., $g(x) = (1/\sqrt{2\pi}) \exp(-x^2/2)$. We will use this g to construct several estimators of I and compare their performances.

(a) Using the g specified above, implement a rejection sampling method that generates n independent samples $(x_i)_{i=1}^n$from π and combines them in a Monte Carlo estimate, i.e.,

$$\hat{I}_{\mathrm{AR}}^n = \frac{1}{n} \sum_{i=1}^{n} (x_i)^2.$$

Keep track of how many samples t you must simulate from g in order to obtain n samples from π. This number t is the "stopping time" of the rejection sampler. Conversely, you can fix t and keep track of what *random* $n(t)$ you end up with. Evaluate this estimator $\hat{I}_{\mathrm{AR}}^{n(t)}$ many times for a fixed t, and then empirically estimate the variance of this estimator, $\mathbb{V}\mathrm{ar}[\hat{I}_{\mathrm{AR}}^{n(t)}]$

(b) Using the g specified above, implement a t-sample *self-normalized importance sampling* estimator of I, called $\hat{I}_{\mathrm{IS}}^t$. Again, evaluate this estimator many times for a fixed t and empirically estimate its variance. (Side question: self-normalized importance sampling also has a small and asymptotically vanishing bias, $\mathcal{O}(1/t^2)$. Is it possible to estimate this bias?)

(c) Now compare the variances of $\hat{I}_{\mathrm{AR}}^{n(t)}$ and $\hat{I}_{\mathrm{IS}}^t$ over a range of sample sizes t. Make a convergence plot. Comparing both estimators at the same t is in some sense the "fair" comparison, as both require simulating t samples from g.

(d) A third estimator of I can be produced by "recycling" the $t - n$ samples rejected by the accept/reject method. The marginal distribution of these rejected samples, up to a normalizing constant, is $Cg(x) - \tilde{\pi}(x)$, where C was the constant chosen so that $Cg(x) \geq \tilde{\pi}(x)$ for all $x \in \mathbb{R}$. Since you know this marginal density, you can construct *another* self-normalized importance sampling estimator using the rejected samples; call it $\hat{I}^{t-n}_{\text{rej}}$. This can then be combined with $\hat{I}^{n(t)}_{\text{AR}}$ to yield a third estimator of I:

$$\hat{I}^{t}_{\text{combo}} = \frac{n}{t}\hat{I}^{n(t)}_{\text{AR}} + \frac{t-n}{t}\hat{I}^{t-n}_{\text{rej}}.$$

Evaluate the performance of this estimator over a range of t values and compare it with the others.

3 Stochastic elliptic PDE

Consider a stochastic linear elliptic equation on a one-dimensional spatial domain:

$$\frac{\partial}{\partial x}\left(k(x,\omega)\frac{\partial u(x,\omega)}{\partial x}\right) = -s(x), \qquad x \in D = [0,1], \tag{1}$$

with a deterministic source term $s(x)$, a deterministic Dirichlet boundary condition at $x = 1$,

$$u(1,\omega) = u_r,$$

and a random Neumann condition at $x = 0$,

$$k(x,\omega)\frac{\partial u}{\partial x}\bigg|_{x=0} = -F(\omega).$$

The diffusivity $k(x,\omega)$ and solution $u(x,\omega)$ are stochastic processes defined on $D \times \Omega$, where $(\Omega, \mathcal{U}, \mathbb{P})$ is a probability space. $F(\omega)$ is defined on the same probability space. This stochastic elliptic equation can model a host of physical phenomena, ranging from heat conduction in a heterogeneous material (where u is proportional to temperature) to fluid flow in a porous medium (where u could denote pressure and k is proportional to permeability).

3.1 Problem parameters

Parameters and boundary conditions for equation (1) are specified as follows:

- Let $Y(x,\omega)$ be a piecewise constant random field on $D = [0,1]$, defined as:

$$Y(x,\omega) = \begin{cases} Y_1(\omega), & x \in [0, 0.25) \\ Y_2(\omega), & x \in [0.25, 0.5) \\ Y_3(\omega), & x \in [0.5, 0.75) \\ Y_4(\omega), & x \in [0.75, 1] \end{cases}$$

and let the random variables $Y_i,\ i = 1 \ldots 4$ be independent and identically distributed Gaussians with mean μ_Y and variance σ_Y^2.

The diffusivity is $k(x, \omega) = \exp(Y(x, \omega))$. In other words, the diffusivity in each segment of the domain is a log-normal random variable. Set $\mu_Y = -1.0$ and $\sigma_Y^2 = 1.0$.

- The flux F is normally distributed with mean $\mu_F = -2.0$ and variance $\sigma_F^2 = 0.5$. F and $(Y_i)_{i=1}^4$ are independent.
- The Dirichlet datum is $u_r = 1$.
- The source term is spatially uniform: $s(x) = 5$.

Before solving the rest of the problem, you will need a function that solves a deterministic version of equation (1) for any "input" realization of (F, Y_1, Y_2, Y_3, Y_4). **We have provided a matlab script, `diffusioneqn.m`, that does this.** Alternatively, feel free to implement it yourself. You can solve this elliptic equation using a finite different or finite element method. Or, since the equation is posed on 1-D domain and thus a linear ODE, you can even write the solution analytically.

3.2 Control variates for Monte Carlo variance reduction

Our first goal is to estimate the mean and the variance of the solution field u at a single point, $x = 0.6$.

(a) Use a Monte Carlo method to estimate $\mathbb{E}[u(x = 0.6, \omega)]$. Report your estimate along with its standard error (i.e., the estimated standard deviation of the estimator) and a 95% confidence interval, based a single n-sample run. Try this a few times (i.e., for independent "replicate" runs) and for different samples sizes n.

(b) Now use a Monte Carlo method to estimate $\mathbb{V}\mathrm{ar}[u(x = 0.6, \omega)]$. Report your estimate along with its standard error. Again, try this a few times, and for different samples sizes n.

(c) Now we will re-attempt the mean estimate of part (a) with a **control variate**. While many control variates could be devised for the problem, let's use a simple one: a linearization of the map $(f, y_1, y_2, y_3, y_4) \mapsto u(x = 0.6)$. (You may construct the linearization however you wish. Note that the expectation of the linear map can be computed exactly.)

Compare the variance of the Monte Carlo estimator with and without the control variate, for different numbers of samples n. What happens to the amount of variance reduction as you make μ_Y much larger or much smaller?

3.3 Importance sampling for rare events

Our second goal is to estimate the probability $p \coloneqq \mathbb{P}[u(x = 0.6, \omega) > u_0]$, with $u_0 = 40$.

(a) Use a standard Monte Carlo method to estimate p. Estimate the standard deviation $\hat{\sigma}_n$ of your Monte Carlo estimate of this probability, for different sample sizes n. As an integrated performance metric, report the *relative error per sample*, defined as $\sqrt{n}\hat{\sigma}_n/p$.

(b) Now consider using **importance sampling** to improve the efficiency of your Monte Carlo procedure. Devise a biasing distribution and use it to estimate p. Compare the relative error per sample with that of the standard Monte Carlo method in part (a).

We particularly recommend using the *cross-entropy* method (as described in lecture) to construct a good biasing distribution. For simplicity, let your biasing distribution remain Gaussian with diagonal covariance matrix. Then experiment with more complex biasing distributions (e.g., correlated Gaussian distributions, multivariate t distributions) and see if they yield any further performance gains.

4 Enrichment problem: importance sampling and large deviations

[This problem will not be graded; *it is for fun and further exploration, and designed to give you a glimpse of a topic that we won't have time to discuss in class.*]

When estimating the probability of a rare event via importance sampling, the theory of *large deviations* can provide valuable guidance for the construction of a biasing distribution. Consider a sequence of independent uniform random variables $X_i \sim U(0,1)$. We wish to estimate $\rho := \mathbb{P}\left[\sum_{i=1}^{n} X_i > nT\right]$ for $T = 0.99$ and $n = 20$. As a biasing distribution, consider the family of *exponentially shifted* probability densities:

$$q_\theta(x) = \frac{p(x)\exp(\theta x)}{M_X(\theta)},$$

where $p(x)$ is the probability density of the original random variable X, θ is a parameter to be chosen, and $M_X(\theta)$ is the *moment-generating function* of X, defined as

$$M_X(\theta) = \mathbb{E}\left[\exp(\theta X)\right].$$

Large deviations theory tells us to choose θ so that the mean of the exponentially-shifted biasing distribution is equal to T. Call this value θ_T. Simulating n independent samples from the resulting q_{θ_T} yields a biased version of $\sum_{i=1}^{n} X_i$. Call each such n-sample draw a "simulation run."

(a) Calculate θ_T. Use K simulation runs to construct an importance sampling estimator of ρ. *(Hint: you may need to employ the inverse-CDF trick to sample from q_{θ_T}.)* Run simulations and report on their behavior. Estimate the value of ρ. Roughly how many runs K do you need in order to estimate ρ with a 1% relative error?

(b) What happens to your estimator of ρ when θ departs from the optimal value θ_T? Play with this empirically and report on the behavior. You should see the impact of both underbiasing and overbiasing.

Remarks: With the right choice of θ_T, the simulation procedure described above is said to be *asymptotically efficient*, in the sense that the *rate* at which the variance of the estimator $\hat{\rho}$ goes to zero, as a function of n, is maximized. If the importance sampling procedure is not efficient in this sense, the number of simulation runs needed to achieve a given accuracy must grow exponentially with n, which is clearly undesirable. Seeking an estimator that is asymptotically efficient is often much easier than finding an estimator that is optimal at a given n; in other words, it is easier to maximize the rate at which the variance goes to zero in n than to minimize the variance of the estimator itself.

There is a rich body of work justifying the use of exponential shifting as above, quantifying variance rates, and generalizing these ideas to many other problems. For a nice introduction, see J. Bucklew, *Introduction to Rare Event Simulation*, Springer (2004).