Advanced Scientific Computing: Stochastic Methods for Data Analysis, Inference and Optimization

P. Koumoutsakos
323 Pierce Hall, 29 Oxford Street Cambridge,
MA 02138

Fall semester 2022

Set 1- Probability, Sampling & Monte Carlo Integration

Issued: September 14, 2022
Hand in: September 28, 2022 12:00pm

Question 1: From Ensemble Interpretation to Kolmogorov Axioms

Let $P(\omega)$ be the probability of an event $\omega \subset \Omega$ where we denote the sample space as Ω. The three axioms of probabilities are:

- $P(\omega) \geq 0$ for all $\omega \subset \Omega$
- $P(\Omega) = 1$
- If $\omega_1, \omega_2, \cdots \subset \Omega$ and they are pairwise disjoint, i.e. $\omega_i \bigcap \omega_j = \varnothing$ when $i \neq j$, then $P(\omega_1 \cup \omega_2 \cup \ldots) = P(\omega_1) + P(\omega_2) + \cdots$

The ensemble interpretation of probabilities allows for the definition of probability as a special kind of ensemble average, an ensemble simply being a finite collection of n models where each model is labelled with an index $i \in \{0, 1, ..., n-1\}$.

Recall from lecture (9/13), an indicator function

$$\chi_A(i) = \begin{cases} 1 & \text{if ensemble member labelled } i \text{ has property } A \\ 0 & \text{otherwise} \end{cases}$$

allows for the definition of the probability of A to be defined in terms of an ensemble average of $\chi_A(i)$. For example:

- Probability as the relative frequency of the members having the property A in the ensemble

$$P(A) = \sum_{i=1}^{n} \frac{\chi_A(i)}{n}$$

Starting with this ensemble interpretation, prove that the three axioms of probability must hold true.

Question 2: Mutating Genome

Consider an organism with a genome in which mutations happen as a Poisson process with rate ν. Assume the following about mutations:

- All mutations are neutral (i.e., they do not affect the rate of reproduction).
- The genome is large enough that the mutations always happen at different loci (this is known as the infinite sites model) and are irreversible.
- At $t = 0$, there are no mutations.

a) What is the probability that the genome does not obtain any new mutations within the time interval $[0, t)$?

b) What is the expected number of mutations for time T?

c) Consider a population of N individuals following the Wright-Fisher model. Each generation is formed from the previous by the following algorithm:

   - Every individual organism reproduces asexually (i.e., every organism divides forming two new organisms). The population size is now $2N$.
   - Uniformly at random sample N of these $2N$ individuals to survive. This is now the new generation.

   Find the probability $P(t)$ of two individuals having their "first" (latest chronologically) common ancestor t generations ago. Hint: Go backwards in time with discrete time steps. What is the continuum limit of this probability (i.e., the result for a large population size)?

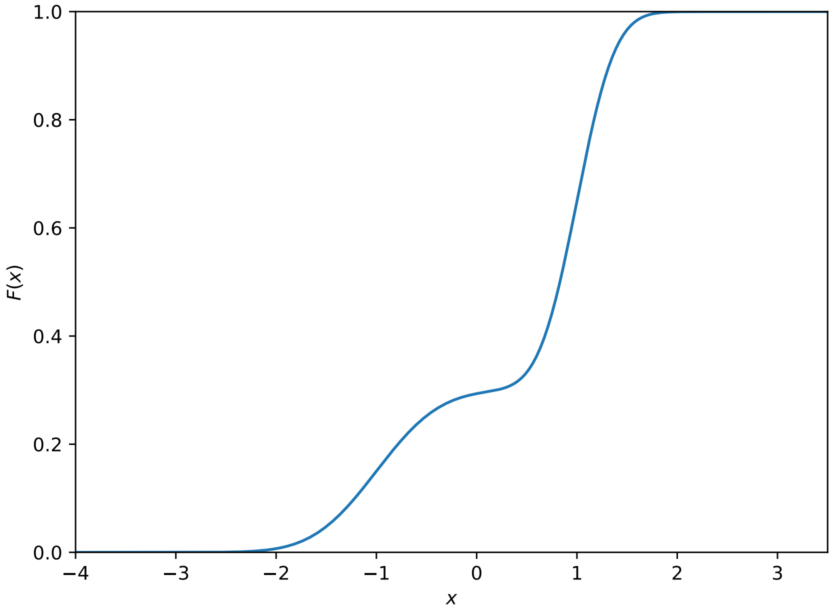
d) Now add mutations to the Wright-Fisher model. Assume we sample two individuals that followed two distinct lineages for precisely t generations (i.e., their first common ancestor occurred t generations ago). What is $P(\pi|t)$, the probability of π mutations arising during the t generations?

e) What is $P(\pi)$, the probability of two individuals being separated by π mutations after they were born from the same parent? What is the expected value of π? (You may work in the continuum limit as in (c), corresponding to a large population size $N \gg 1$).

Note that subquestions d) and e) can be solved independently of c).

Question 3: Sampling from a piecewise linear CDF

Sample a million points from the cumulative distribution function (CDF) F shown below:



The CDF is piecewise linear,

$$F(x) = \begin{cases} \frac{x-x_i}{x_{x+1}-x_i}(y_{i+1}-y_i)+y_i, & x_i \le x < x_{i+1}, \\ 0, & x \le x_1, \\ 1, & x > x_n, \end{cases}$$

where x_i and y_i, $i = 1, 2, \ldots, n$, are monotonically increasing and provided in the file `cdf.csv`. Fill-in the `TODO`s in `main.cpp` to generate the samples, compute a histogram of these samples and compute the median of the samples. Compare the empirical median (obtained from the samples) with the median of the CDF. Report the empirical median, the median of the CDF and plot the histogram. Use the provided `Makefile` to compile your code.

Question 4: Monte Carlo Integration

In this exercise we compare the convergence rate of three different integration techniques. The test function to be examined is the d-dimensional function given below:

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

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

a) Compute the analytical solution of the integral

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

where V is a d-dimensional cube with bounds $[-1, 1]^d$. Note that the solution depends on the dimension d.

b) Write a c++ program that approximates the integral using Monte-Carlo integration,

$$I_{\mathsf{MC},M} = \frac{|V|}{M} \sum_{i=1}^{M} f(x^{(i)}).$$

Here, M is the number of Monte-Carlo samples $\mathbf{x}^{(i)}$ drawn from a d-dimensional uniform distribution, and $|V|$ is the volume of the d-dimensional cube.

c) Compute the integral using importance sampling: Generate samples $\mathbf{x}^{(i)}$ normally distributed with standard deviation $\sigma = 0.5$ along each direction. The estimate of the integral is given by

$$I_{\mathsf{IS},M} = \frac{1}{M} \sum_{i=1}^{M} \frac{g(\mathbf{x}^{(i)})}{\mathcal{N}(\mathbf{x}^{(i)} | \mathbf{0}, \sigma)},$$

where

$$g(\mathbf{x}) = \begin{cases} f(\mathbf{x}), & \mathbf{x} \in V, \\ 0, & \text{otherwise.} \end{cases}$$

d) The above estimate depends highly on the choice of the distribution of the samples. Repeat the above but with samples that have a PDF

$$q(\mathbf{x}) = C \left(0.1 + g(\mathbf{x})\right) ,$$

where C is a normalization constant. Use the rejection sampling method to generate the samples $x^{(i)} \sim q(\mathbf{x})$. The estimate is then given by

$$I_{\mathsf{RS},M} = |V| \frac{\sum\limits_{i=1}^{M} w_i g(\mathbf{x}^{(i)})}{\sum\limits_{i=1}^{M} w_i},$$

where $w_i = 1/q_{\mathbf{x}^{(i)}}$.

e) For each of the three methods, compute the estimate for $M = 1, 10, 100, 1'000, 10'000, 100'000$ and $1'000'000$ samples for $d = 1, 2, 4, 8, 16$ and show that the error converges as $1/\sqrt{M}$, independently of d. Write your results to a file and plot it. Hint: Estimate the errors by computing the standard deviation over 100 runs of each estimate.

f) Write a python program for the estimate described in d). Compare the time taken by the c++ version with that taken by the python program. How much speed-up do you obtain with the c++ code? Hint: you can use the command time on UNIX environments.