Probability and Stochastic Processes II
Lecture 1

Michael Evans
University of Toronto
https://utstat.utoronto.ca/mikevans/stac62/staC632024.html

2024

I. Monte Carlo

- perhaps the most useful application of probability theory

- it is a technique for approximately computing integrals (sums) that are otherwise intractable

- to begin we suppose that for any probability model $(\Omega, \mathcal{A}, P)$ there is an algorithm that can be used to generate

$$\omega_1, \omega_2, \ldots, \omega_N \overset{i.i.d.}{\sim} P$$

for N as large as is necessary

I.1 Approximate Integration

Example 1.

- suppose it is required to compute $I = \int_0^1 f(x)\,dx = \int_0^1 \frac{\cos x}{1+x^2}\,dx$

quadrature: one approach here is to use quadrature: let $x_i = i/m$ for $m \in \mathbb{N}$ and approximate I by the Riemann sum

$$I_m = \sum_{i=1}^{m} f(x_i)(x_i - x_{i-1}) = \frac{1}{m}\sum_{i=1}^{m} \frac{\cos x_i}{1+x_i^2} \to I \text{ as } m \to \infty$$

- doing this for increasing m gives the following results

m	I_m
10	0.6458649
10^2	0.679278
10^3	0.682568
10^4	0.6828965
10^5	0.6829294
10^6	0.6829327

- so it looks like with $m = 10^6$ we have 5 significant places in the answer

R code:

```
m=1000000
Im=0
f <- function(x) {
f=cos(x)/(1+x**2)
return(f)
}
# Riemann sum
for (i in 1:m){
Im=Im+f(i/m)
}
Im=Im/m
Im
```

■

Monte Carlo: alternatively we can write

$$I = E(f(\omega))$$

where $\omega \sim \text{Uniform}(0,1)$

- so generate $\omega_1, \omega_2, \ldots, \omega_N \overset{i.i.d.}{\sim} \text{Uniform}(0,1)$ and then the SLLN gives

$$I_N = \frac{1}{N}\sum_{i=1}^{N} f(\omega_i) \overset{wp1}{\rightarrow} I \text{ as } N \rightarrow \infty$$

- also we have

$$\begin{aligned} Var(f(\omega)) &= E((f(\omega) - I)^2) = E(f^2(\omega)) - I^2 \\ Var(I_N) &= Var(f(\omega)/N \end{aligned}$$

and $Var(f(\omega))$ can be estimated by (limit proved in PSPI)

$$S_N^2 = \frac{1}{N}\sum_{i=1}^{N}(f(\omega_i) - I_N)^2 = \frac{1}{N}\sum_{i=1}^{N} f^2(\omega_i) - I_N^2 \overset{wp1}{\rightarrow} Var(f(\omega))$$

- also the generalization of the CLT proved in PSPI gives

$$\frac{I_N - I}{S_N/\sqrt{N}} \xrightarrow{d} N(0,1) \text{ as } N \to \infty$$

- so for large N

$$\begin{aligned} 0.9973002 &= \Phi(3) - \Phi(-3) \approx P\left(-3 < \frac{I_N - I}{S_N/\sqrt{N}} < 3\right) \\ &= P\left(I_N - 3S_N/\sqrt{N} < I < I_N + 3S_N/\sqrt{N}\right) \end{aligned}$$

and the interval $(I_N - 3S_N/\sqrt{N}, I_N + 3S_N/\sqrt{N})$ contains the value of I with virtual certainty

- here are some results

N	I_N	$3S_N/\sqrt{N}$
10	0.6463478	0.23896600
10^2	0.6977411	0.07040631
10^3	0.6837775	0.02247795
10^4	0.6832828	0.007059335
10^5	0.6822196	0.002234954
10^6	0.683162	0.0007068541

- after $N = 10^6$ only 3 significant places, so in this case Monte Carlo is not as accurate as quadrature

R code:

```
# Monte Carlo dimension 1
N=1000000
omega=runif(N,0,1)
IN=0
IN2=0
for (i in 1:N){
fun=f(omega[i])
IN=IN+fun
IN2=IN2+fun**2
}
IN=IN/N
SN2=(IN2/N-IN**2)
error=3*sqrt(SN2/N)
IN
error
```

■

- Monte Carlo has some advantages

1. There is a natural error estimate which isn't as easy to obtain with quadrature.

2. Quadrature suffers from a dimensional effect (not as bad for MC).

Example 2.

$$I = \int_{[0,1]^{10}} \frac{\cos(x_1 x_2 \cdots x_{10})}{1 + x_1^2 + x_2^2 + \cdots + x_{10}^2}\, dx_1 \cdots dx_{10}$$

- quadrature with m subdivisions on each axis requires m^{10} function evaluations which is not feasible for even moderate m ($m = 10$ requires 10^{10} function evals) MC gives

N	I_N	$3S_N/\sqrt{N}$
10	0.1501259	0.01105813
10^2	0.2674095	0.02947328
10^3	0.2451248	0.005641905
10^4	0.2423392	0.001700063
10^5	0.2434972	0.0005625282
10^6	0.2427743	0.0001752757

R code:

```
# Monte Carlo dimension 10
N=1000000
omega=runif(N*10,0,1)
IN=0
IN2=0
for (i in 1:N){
x=1
s=1
for (j in 1:10){
x=x*omega[10*(i-1)+j]
s=s+(omega[10*(i-1)+j])**2
}
fun=cos(x)/s
IN=IN+fun
IN2=IN2+fun**2
}
IN=IN/N
SN2=(IN2/N-IN**2)
error=3*sqrt(SN2/N)
IN
error
```

■

3. Monte Carlo is flexible (but also one needs to be careful)

- suppose there is a need to approximate, for some $f : \mathbb{R}^k \to \mathbb{R}^1$, the integral

$$I = \int_{\mathbb{R}^k} f(\mathbf{x})\, d\mathbf{x} < \infty$$

- suppose g is a pdf on $\mathbb{R}^k$ such that $g(\mathbf{x}) = \mathbf{0}$ implies $f(\mathbf{x}) = 0$ and we can generate $\mathbf{x}_1, \ldots, \mathbf{x}_N \overset{i.i.d.}{\sim} g$
- then

$$\begin{aligned} E_g\left(\frac{f(\mathbf{X})}{g(\mathbf{X})}\right) &= \int_{\mathbb{R}^k} \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x})\, d\mathbf{x} = \int_{\mathbb{R}^k} f(\mathbf{x})\, d\mathbf{x} = I \\ Var_g\left(\frac{f(\mathbf{X})}{g(\mathbf{X})}\right) &= E_g\left(\left(\frac{f(\mathbf{X})}{g(\mathbf{X})} - I\right)^2\right) = \int_{\mathbb{R}^k} \frac{f^2(\mathbf{x})}{g(\mathbf{x})}\, d\mathbf{x} - I^2 \end{aligned}$$

- by the SLLN

$$
\begin{aligned}
I_N &= \frac{1}{N}\sum_{i=1}^{N}\frac{f(\mathbf{x}_i)}{g(\mathbf{x}_i)} \overset{wp1}{\to} I \\
S_N^2 &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{f(\mathbf{x}_i)}{g(\mathbf{x}_i)} - I_N\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{f(\mathbf{x}_i)}{g(\mathbf{x}_i)}\right)^2 - I_N^2 \overset{wp1}{\to} Var_g\left(\frac{f(\mathbf{X})}{g(\mathbf{X})}\right)
\end{aligned}
$$

and so again the interval $(I_N - 3S_N/\sqrt{N}, I_N + 3S_N/\sqrt{N})$ contains I with virtual certainty

- but g has to be chosen carefully: choose g such that $\int_{\mathbb{R}^k} \frac{f^2(\mathbf{x})}{g(\mathbf{x})}\, d\mathbf{x}$ is finite and as small possible

- this approach is known as *importance sampling* because you choose g so that the values $\mathbf{x}$ generated from g lie in the region where f takes its important values

Example 3.

- consider $I = \int_{-\infty}^{\infty} f(x)\, dx = \int_{-\infty}^{\infty} \frac{1}{1+x^2}\, dx$ (proportional to a Cauchy density)

- suppose we take $g(x) = \varphi(x)$ the $N(0,1)$ density

- then

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{f^2(x)}{g(x)}\, dx &= \sqrt{2\pi} \int_{-\infty}^{\infty} \frac{\exp(x^2/2)}{(1+x^2)^2}\, dx \\ &= 2\sqrt{2\pi} \int_{0}^{\infty} \frac{\exp(x^2/2)}{(1+x^2)^2}\, dx \geq \sqrt{2\pi} \int_{0}^{\infty} \frac{x^4/4}{(1+x^2)^2}\, dx = \infty\end{aligned}$$

- so $g = \varphi$ is a bad choice here ■

Theorem I.1 *(Optimal importance sampler)* For $I = \int_{\mathbb{R}^k} f(\mathbf{x})\, d\mathbf{x} < \infty$ the importance sampler that minimizes the variance is

$$g_f(\mathbf{x}) = \frac{|f(\mathbf{x})|}{\int_{\mathbb{R}^k} |f(\mathbf{x})|\, d\mathbf{x}} \text{ with variance } \left(\int_{\mathbb{R}^k} |f(x)|\, d\mathbf{x}\right)^2 - I^2.$$

Proof: Put $c = \int_{\mathbb{R}^k} |f(\mathbf{x})|\, dx$ so

$$\begin{aligned}
& Var_g\left(\frac{f(\mathbf{X})}{g(\mathbf{X})}\right) = \int_{\mathbb{R}^k} \frac{f^2(\mathbf{x})}{g(\mathbf{x})}\, d\mathbf{x} - I^2 = c^2 \int_{\mathbb{R}^k} \frac{g_f^2(\mathbf{x})}{g(\mathbf{x})}\, d\mathbf{x} - I^2 \\
= \quad & c^2 \left(\begin{array}{c} \int_{\mathbb{R}^k} \frac{g_f^2(\mathbf{x}) - 2g(\mathbf{x})g_f(\mathbf{x}) + g^2(\mathbf{x})}{g(\mathbf{x})}\, d\mathbf{x} + \\ \int_{\mathbb{R}^k} \frac{2g(\mathbf{x})g_f(\mathbf{x}) - g^2(\mathbf{x})}{g(\mathbf{x})}\, d\mathbf{x} \end{array} \right) - I^2 \\
= \quad & c^2 \left(\int_{\mathbb{R}^k} \frac{(g_f(\mathbf{x}) - g(\mathbf{x}))^2}{g(\mathbf{x})}\, d\mathbf{x} + 2\int_{\mathbb{R}^k} g_f(\mathbf{x})\, d\mathbf{x} - \int_{\mathbb{R}^k} g(\mathbf{x})\, d\mathbf{x} \right) - I^2 \\
= \quad & c^2 E_g\left(\left(\frac{g_f(\mathbf{x}) - g(\mathbf{x})}{g(\mathbf{x})} \right)^2 \right) + c^2 - I^2
\end{aligned}$$

and this is minimized as a function of g by taking $g = g_f$. ■

notes

1. When $f \geq 0$ the optimal importance sampler has variance $= 0$.

2. The expression

$$E_w\left(\left(\frac{w(\mathbf{x}) - g(\mathbf{x})}{w(\mathbf{x})}\right)^2\right)$$

is called the *chi-squared distance* between the distributions given by pdf's w and g and so we try to make this distance between g_f and g as small as we can in spite of the fact that we don't know $\int_{\mathbb{R}^k} |f(\mathbf{x})| \, d\mathbf{x}$.

3. Basically we want a g that puts the bulk of its mass in the same region where f does and the tails of g should not be shorter than the tails of f.

4. A diagnostic for the failure of a given importance sampler is given by the *coefficient of variation* (ratio of standard deviation of estimate to quantity being estimated) squared for estimating $I = \int_{\mathbb{R}^k} |f(x)|\, d\mathbf{x}$

$$
\begin{aligned}
CV_g^2(I_N) &= \frac{\frac{1}{N} Var_g\left(|f(\mathbf{X})|/g(\mathbf{X})\right)}{I^2} \approx \frac{1}{N}\frac{S_N^2}{I_N^2} = \sum_{i=1}^{N} w_i^2 - \frac{1}{N} \text{ where} \\
w_i &= \frac{|f(\mathbf{x}_i)|/g(\mathbf{x}_i)}{\sum_{j=1}^{N} |f(\mathbf{x}_j)|/g(\mathbf{x}_j)}
\end{aligned}
$$

so $0 \le w_i \le 1$ and $\sum_{i=1}^{N} w_i = 1$

- since $0 \le CV_g^2(I_N)$ we have $1/N \le \sum_{i=1}^{N} w_i^2 \le 1$ and $\sum_{i=1}^{N} w_i^2$ equals (or is close to) 1 iff $w_i = 1$ for some i (or several w_i are close to 1) as this indicates the i-th value $|f(\mathbf{x}_i)|/g(\mathbf{x}_i)$ (or just a few values) is dominating the estimate

- note - $\sum_{i=1}^{N} w_i^2 \approx 1/N$ does not mean that the importance sampling has succeeded!

I.2 Generating Random Variables

- for a given density f an efficient computer-based method is required to be able to provide a value $X \sim f$

- there are many such methods but we discuss two

1. **Inversion**

- let $F : \mathbb{R} \to [0, 1]$ given by $F(x) = P(X \leq x)$ denote the cdf of X

- the inverse cdf (quantile function) $F^{-1} : [0, 1] \to \mathbb{R}$ is given by

$$F^{-1}(u) = \inf\{x : F(x) \geq u\}$$

Theorem I.2 If $U \sim$Uniform$(0, 1)$ then $X = F^{-1}(U) \sim F$.
Proof: Note that $u \leq F(x)$ iff $F^{-1}(u) \leq x$ and so

$$P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x).$$

■

- typically we need a closed form formula for F^{-1} or F for this to be useful

Example 1. *exponential$_{rate}(\lambda)$*

- $f(x) = \lambda e^{-\lambda x}$ for $x > 0$ so $F(x) = \int_0^x \lambda e^{-\lambda z}\, dz = -e^{-\lambda z}\big|_0^x = 1 - e^{-\lambda x}$ a 1-1 increasing function on $[0, \infty)$

- so for $u \in [0, 1]$ then $u = 1 - e^{-\lambda x}$ iff $x = -\lambda^{-1}\log(1 - u) = F^{-1}(u)$ ■

Example 2. *mixtures*

- consider a weighted mixture of a $N(0, 1)$ and a Cauchy density, namely,

$$\begin{aligned} f(x) &= 0.4f_1(x) + 0.6f_2(x) = 0.4\varphi(x) + 0.6/\pi(1 + x^2) \\ F(x) &= \int_{-\infty}^{x} f(z)\, dz = 0.4F_1(x) + 0.6F_2(x) \\ &= 0.4\Phi(x) + 0.6(\arctan(x)/\pi + 0.5) \end{aligned}$$

- there isn't a closed form for Φ^{-1} but there are good computer algorithms for it and $\tan(\pi(u - 0.5))$ is the inverse cdf of the Cauchy

- to generate $X \sim F$ the following algorithm works

1. generate $U_1 \sim$Uniform$(0,1)$
2. if $U_1 \leq 0.4$ put $i = 1$ otherwise put $i = 2$
3. generate $U_2 \sim$Uniform$(0,1)$
4. return $X = F_i^{-1}(U_2)$

- then

$$\begin{aligned} P(X \leq x) &\overset{TTP}{=} P(i=1)P(X \leq x \mid i=1) + P(i=2)P(X \leq x \mid i=2) \\ &= 0.4\Phi(x) + 0.6(\arctan(x)/\pi + 0.5) = F(x) \end{aligned}$$

■

- for a multivariate distribution on $\mathbb{R}^k$ with pdf f we have

$$f(x_1, \ldots, x_k) = f_1(x_1) f_2(x_2 \mid x_1) f_3(x_3 \mid x_1, x_2) \cdots f_k(x_k \mid x_1, \ldots, x_{k-1})$$

so $\mathbf{x} \sim f$ can sometimes be accomplished by using algorithms to generate sequentially

$$\begin{aligned} x_1 &\sim f_1 \\ x_2 \mid x_1 &\sim f_2(\cdot \mid x_1) \\ &\vdots \\ x_k \mid x_1, \ldots, x_{k-1} &\sim f_k(\cdot \mid x_1, \ldots, x_{k-1}) \end{aligned}$$

2. **Rejection**

- the following algorithm to generate from (unnormalized) pdf f on $\mathbb{R}^k$ is known as *rejection*

Theorem I.3 If g is a pdf that can be generated from and c is a constant such that $f(\mathbf{x}) \leq cg(\mathbf{x})$ for every $\mathbf{x} \in \mathbb{R}^k$, then the following generates $\mathbf{X} \sim f$.

1. generate $\mathbf{Y} \sim g$ and $U \sim \text{Uniform}(0,1)$ stat. ind.,
2. if $Ucg(\mathbf{Y}) > f(\mathbf{Y})$ then go to 1, else return $\mathbf{X} = \mathbf{Y}$ and stop.

Proof: The probability of accepting at step 2 is

$$\begin{aligned} p &= P(Ucg(\mathbf{Y}) \leq f(\mathbf{Y})) \overset{TTE}{=} E_g(P(Ucg(\mathbf{Y}) \leq f(\mathbf{Y}) \,|\, \mathbf{Y})) \\ &= E_g\left(P\left(U \leq \frac{f(\mathbf{Y})}{cg(\mathbf{Y})} \,|\, \mathbf{Y}\right)\right) = E_g\left(\frac{f(\mathbf{Y})}{cg(\mathbf{Y})}\right) = \frac{\int_{\mathbb{R}^k} f(\mathbf{x})\, d\mathbf{x}}{c}. \end{aligned}$$

Since $p > 0$, the probability of stopping after finitely many steps is $\sum_{i=1}^{\infty}(1-p)^{i-1}p = p/(1-(1-p)) = 1$ and so the algorithm stops with probability 1 and returns $\mathbf{X}$. For $B \in \mathcal{B}^k$, and recall that the $(U_i, \mathbf{Y}_i)$ are *i.i.d.*,

$$P(\mathbf{X} \in B) = \sum_{i=1}^{\infty} P(\text{ algorithm stops at the } i\text{-th step and } \mathbf{Y}_i \in B)$$

$$\begin{aligned}
&= \sum_{i=1}^{\infty} P\left(U_1 > \frac{f(\mathbf{Y}_1)}{cg(\mathbf{Y}_1)}, \ldots, U_{i-1} > \frac{f(\mathbf{Y}_{i-1})}{cg(\mathbf{Y}_{i-1})}, U_i \leq \frac{f(\mathbf{Y}_i)}{cg(\mathbf{Y}_i)}, \mathbf{Y}_i \in B\right) \\
&\overset{TTP}{=} \sum_{i=1}^{\infty} P\left(U_i \leq \frac{f(\mathbf{Y}_i)}{cg(\mathbf{Y}_i)}, \mathbf{Y}_i \in B \,\middle|\, U_1 > \frac{f(\mathbf{Y}_1)}{cg(\mathbf{Y}_1)}, \ldots, U_{i-1} > \frac{f(\mathbf{Y}_{i-1})}{cg(\mathbf{Y}_{i-}\right. \\
&\qquad \times (1-p)^{i-1} \\
&= \sum_{i=1}^{\infty} P\left(U \leq \frac{f(\mathbf{Y})}{cg(\mathbf{Y})}, \mathbf{Y} \in B\right)(1-p)^{i-1} \\
&= P\left(U \leq \frac{f(\mathbf{Y})}{cg(\mathbf{Y})}, \mathbf{Y} \in B\right)\sum_{i=1}^{\infty}(1-p)^{i-1} \\
&= \frac{P\left(U \leq \frac{f(\mathbf{Y})}{cg(\mathbf{Y})}, \mathbf{Y} \in B\right)}{p}
\end{aligned}$$

and

$$
\begin{aligned}
P\left(U \leq \frac{f(\mathbf{Y})}{cg(\mathbf{Y})}, \mathbf{Y} \in B\right) &\overset{TTE}{=} E_g\left(P\left(U \leq \frac{f(\mathbf{Y})}{cg(\mathbf{Y})}, \mathbf{Y} \in B \,\middle|\, \mathbf{Y}\right)\right) \\
= \quad & E_g\left(I_B(\mathbf{Y})\frac{f(\mathbf{Y})}{cg(\mathbf{Y})}\right) = \frac{\int_B f(\mathbf{x})\,d\mathbf{x}}{c}.
\end{aligned}
$$

Therefore,

$$
P(\mathbf{X} \in B) = \frac{\int_B f(\mathbf{x})\,d\mathbf{x}}{c}\left(\frac{\int_{\mathbb{R}^k} f(\mathbf{x})\,d\mathbf{x}}{c}\right)^{-1} = \frac{\int_B f(\mathbf{x})\,d\mathbf{x}}{\int_{\mathbb{R}^k} f(\mathbf{x})\,d\mathbf{x}}
$$

as required. ■

- note the efficiency of rejection is primarily determined by

$$
p = \frac{\int_{\mathbb{R}^k} f(\mathbf{x})\,d\mathbf{x}}{c}
$$

and we want this as close to 1 as possible and expected number of iterations until acceptance is $1/p$

Example 3.

- suppose $f(x) = (x+3)^2(x+1)$ on $[0,1]$ is an unnormalized density

- $\max f$ occurs at $x = 1$ and max value is 32

- then if g is the Uniform$(0,1)$ density and $c = 32$ the conditions for rejection are satisfied and $1/p = 1.677$ (mean of a geometric(p) distribution) ■

Exercises

I.1 E&R 4.5.1

I.2 E&R 4.5.2

I.3 E&R 4.5.5

I.4 E&R 4.5.13

I.5 E&R 4.5.14

I.6 E&R 4.5.16

I.7 E&R 4.5.17

I.8 Suppose $\mathbf{X} \sim N_k(\boldsymbol{\mu}, \Sigma)$. Provide an algorithm for generating $\mathbf{X}$. (Hint: recall the relationship between such an $\mathbf{X}$ and $\mathbf{Z} \sim N_k(\mathbf{0}, \Sigma)$ and first discuss how you would generate $\mathbf{Z}$ based on generating from the $N(0,1)$ distribution.)