Assignment 0: O Brother, How Far Art Thou?

Computational Statistics
Instructor: Luiz Max de Carvalho

September 25, 2021

Hand-in date: 06/10/2020.

General guidance

- State and prove all non-trivial mathematical results necessary to substantiate your arguments;
- Do not forget to add appropriate scholarly references *at the end* of the document;
- Mathematical expressions also receive punctuation;
- Please hand in a single PDF file as your final main document.

  Code appendices are welcome, *in addition* to the main PDF document.

Background

A large portion of the content of this course is concerned with computing high-dimensional integrals *via* simulation. Today you will be introduced to a simple-looking problem with a complicated closed-form solution and one we can approach using simulation.

Suppose you have a disc C_R of radius R. Take $p = (p_x, p_y)$ and $q = (q_x, q_y) \in C_R$ two points in the disc. Consider the Euclidean distance between p and q, $||p - q|| = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2} = |p - q|$.

Problem A: What is the *average* distance between pairs of points in C_R if they are picked uniformly at random?

Questions

1. To start building intuition, let's solve a related but much simpler problem. Consider an interval $[0, s]$, with $s > 0$ and take $x_1, x_2 \in [0, s]$ *uniformly at random*. Show that the average distance between x_1 and x_2 is $s/3$.

   Solution. We want to calculate $\mathbb{E}[|x_1 - x_2|]$ such that $x_1, x_2 \overset{iid}{\sim} \text{Unif}(0, s)$. A long and default way to solve this problem is to introduce the random variable $Y = x_1 - x_2$, derive its density, and its absolute mean from the density. Let
   $$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(x_1 \leq y + x_2).$$
   Suppose first that $y \geq 0$. Given x_2,
   $$\mathbb{P}(x_1 \leq y + x_2 | x_2) = \begin{cases} \dfrac{y + x_2}{s} & 0 \leq x_2 \leq s - y \\ 1 & s - y < x_2 \leq s. \end{cases}$$
   Therefore,
   $$\begin{aligned} F_Y(y) &= \int_0^{s-y} \frac{y + x_2}{s^2}\, dx_2 + \int_{s-y}^{s} \frac{1}{s}\, dx_2 \\ &= \frac{y(s-y) + 0.5(s-y)^2}{s^2} + \frac{y}{s} \\ &= \frac{(s-y)(y+s)}{2s^2} + \frac{y}{s} = -\frac{y^2}{2s^2} + \frac{y}{s} + \frac{1}{2}. \end{aligned}$$
   Now suppose $y < 0$. Then
   $$\mathbb{P}(x_1 \leq y + x_2 | x_2) = \begin{cases} \dfrac{y + x_2}{s} & -y \leq x_2 \leq s \\ 0 & 0 \leq x_2 < -y. \end{cases}$$
   Therefore,
   $$\begin{aligned} F_Y(y) &= \int_{-y}^{s} \frac{y + x_2}{s^2}\, dx_2 \\ &= \frac{ys + 0.5s^2}{s^2} - \frac{-y^2 + 0.5y^2}{s^2} \\ &= \frac{y^2}{2s^2} + \frac{y}{s} + \frac{1}{2}. \end{aligned}$$

Deriving these expressions we get the density with respect to the Lebesgue measure,

$$f_Y(y) = \begin{cases} \frac{1}{s} - \frac{y}{s^2}, & \text{if } 0 \le y \le s \\ \frac{1}{s} + \frac{y}{s^2}, & \text{if } -s \le y < 0. \end{cases}$$

Finally,

$$\begin{aligned} \mathbb{E}[|x_1 - x_2|] = \mathbb{E}[|Y|] &= \int_0^s \frac{y}{s} - \frac{y^2}{s^2}\, dy - \int_{-s}^0 \frac{y}{s} + \frac{y^2}{s^2}\, dy \\ &= \frac{s^2}{2s} - \frac{s^3}{3s^2} + \frac{s^2}{2s} - \frac{s^3}{3s^2} \\ &= \frac{s}{2} - \frac{s}{3} + \frac{s}{2} - \frac{s}{3} = \frac{s}{3}, \end{aligned}$$

as we wished to prove. Now I give a more geometric approach. Consider (x_1, x_2) uniformly distributed over $[0,1]^2$. Then, let $t \in [0, s]$,

$$\mathbb{P}(|x_1 - x_2| > t) = \mathbb{P}(x_1 > t + x_2) + \mathbb{P}(x_2 > t + x_1),$$

since they are disjoint events for $t \ge 0$. Note that the first region is delimited by the straight lines $x_1 = t + x_2$, $x_1 = s$, and $x_2 = 0$, what gives a triangle with points $(t, 0), (s, 0)$, and $(s, s - t)$ with area $(s-t)^2/2$. The second region is delimited by the straight lines $x_2 = t + x_1$, $x_2 = s$, and $x_1 = 0$, what gives the triangle $(0, t), (0, s)$ and $(s - t, s)$, with area $(s-t)^2/2$. We conclude that

$$\mathbb{P}(|x_1 - x_2| > t) = \frac{(s-t)^2}{s^2} \text{ if } t \in [0, s],$$

using the fact that the density of (x_1, x_2) is s^{-2}. This implies that

$$\mathbb{E}[|x_1 - x_2|] = \int_0^{+\infty} \frac{(s-t)^2}{s^2} 1_{\{t \le s\}}\, dt = -\frac{(s-t)^3}{3s^2}\Big|_0^s = \frac{s}{3},$$

a simpler prove. □

2. Show that Problem A is equivalent to computing

$$\frac{1}{\pi^2 R^4} \int_0^R \int_0^R \int_0^{2\pi} \int_0^{2\pi} \sqrt{r_1^2 + r_2^2 - 2r_1 r_2 \cos\phi(\theta_1, \theta_2)} r_1 r_2\, d\theta_1\, d\theta_2\, dr_1\, dr_2,$$

where $\phi(\theta_1, \theta_2)$ is the central angle between r_1 and r_2.

Solution. Let $p = (p_x, p_y)$ e $q = (q_x, q_y)$ points in the disc of radius R. We are interested in the quantity

$$\mathbb{E}\left[\sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}\right].$$

Transforming to polar coordinates,

$$p_x = r_1 \cos(\theta_1), p_y = r_1 \sin(\theta_1), q_x = r_2 \cos(\theta_2), q_y = r_2 \sin(\theta_2).$$

And we have the following relation.

$$\begin{aligned}(p_x - q_x)^2 + (p_y - q_y)^2 &= r_1^2\cos(\theta_1)^2 + r_2^2\cos(\theta_1) - 2r_1r_2\cos(\theta_1)\cos(\theta_2) \\ &+ r_1^2\sin(\theta_1)^2 + r_2^2\sin(\theta_1) - 2r_1r_2\sin(\theta_1)\cos(\theta_2) \\ &= r_1^2 + r_2^2 - 2r_1r_2\cos(\theta_1 - \theta_2),\end{aligned}$$

using the appropriate trigonometric relations. Let $p = (p_x, p_y)$ e $q = (q_x, q_y)$ pontos aleatórios com distribuição uniforme no disco de raio R. Consider a system of coordinates in which the origin is the center of the disc. We know that the density of the distribution of p with respect to the Lebesgue measure is

$$f(p) = \frac{1}{\pi R^2} 1_{\{||p|| \leq R\}}.$$

The same for q. Now we want to transform this distribution to polar coordinates, in which $\theta_1 \in [0, 2\pi]$ and $r_1 \in [0, R]$. The Jacobian of the inverse transformation is given by

$$\begin{bmatrix}\cos(\theta_1) & -r_1\sin(\theta_1) \\ \sin(\theta_1) & r_1\cos(\theta_1),\end{bmatrix}$$

in which the absolute value of the determinant is r_1. By the Change of Variables in probability density function,

$$g_1(r_1, \cos(\theta_1)) = \frac{1}{\pi R^2} r_1.$$

Notice that the same happens with $(r_2, \cos(\theta_2))$ with

$$g_2(r_2, \cos(\theta_2)) = \frac{1}{\pi R^2} r_2.$$

By the Law of the unconscious statistician and the transformations made above, $\mathbb{E}\left[\sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}\right] =$

$$\int_{A_1}\int_{A_2} \sqrt{r_1^2 + r_2^2 - 2r_1r_2\cos(\theta_1 - \theta_2)}\, g_1(r_1, \theta_1) g_2(r_2, \theta_2)\, dA_2\, dA_1,$$

such that $A_i = [0, R] \times [0, 2\pi]$, for $i = 1, 2$. Given that the integrand is non negative, by Tonelli's Theorem (Rosenthal, 2006), the integral is

$$\frac{1}{\pi^2 R^4}\int_0^R\int_0^R\int_0^{2\pi}\int_0^{2\pi} \sqrt{r_1^2 + r_2^2 - 2r_1r_2\cos(\theta_1 - \theta_2)}\, r_1 r_2\, d\theta_1\, d\theta_2\, dr_1\, dr_2.$$

□

3. Compute I in closed-form.

Solution. Following the hint, we shall use the Crofton's mean value theorem. However, instead of only using it, we will derive it from the beginning for this particular case to gain intuition. Let $I(R)$ be the integral above

as a function of R. Consider the mapping $x \mapsto Rx$ from the unit circle to the R-circle. It is a linear function, so it is bijective. Besides that,

$$||Rx - Ry|| = R||x - y||,$$

for all x, y in the unit circle. This implies that

$$I(R) = RI(1) \text{ and } I'(R) = I(1).$$

Let $\epsilon > 0$ and define A° the region delimited by the circle of radius $R - \epsilon$, A de region delimited by the circle of radius R, and $\partial A = A/A^\circ$. By the Law of Total Expectation,

$$\begin{aligned} I(R) = I(R-\epsilon)\mathbb{P}(p, q \in A^\circ) + 2J(A^\circ, \partial A)\mathbb{P}(p \in A^\circ, q \in \partial A) \\ + K(\partial A)\mathbb{P}(p, q \in \partial A), \end{aligned} \tag{1}$$

such that $J(A^\circ, \partial A)$ is the expected distance between a point in the inner circle and the other in the outer circle, and $K(\partial A)$ the expected distance between points in the outer circle. We will take ϵ infinitely small afterwards. For that, note the following relations,

$$\begin{aligned} \mathbb{P}(p, q \in A^\circ) &= \left(\frac{\pi(R-\epsilon)^2}{\pi R^2}\right)^2 \\ &= \frac{(R-\epsilon)^4}{R^4} = \sum_{k=0}^{4} \binom{4}{k} (-1)^k R^{-k} \epsilon^k \\ &= 1 - \frac{4\epsilon}{R} + o(\epsilon), \end{aligned}$$

in which $o(\cdot)$ is the little o-notation,

$$\begin{aligned} \mathbb{P}(p \in A^\circ, q \in \partial A) &= \frac{(R-\epsilon)^2}{R^2}\left(1 - \frac{(R-\epsilon)^2}{R^2}\right) \\ &= \frac{(R^2 - 2R\epsilon + \epsilon^2)(2R\epsilon - \epsilon^2)}{R^4} \\ &= \frac{2\epsilon}{R} + o(\epsilon), \end{aligned}$$

and

$$\begin{aligned} \mathbb{P}(p, q \in \partial A) &= \left(1 - \frac{(R-\epsilon)^2}{R^2}\right)^2 \\ &= \frac{(2R\epsilon - \epsilon^2)^2}{R^4} \\ &= o(\epsilon). \end{aligned}$$

The formula given by (1) can be rewritten as

$$I(R) = I(R-\epsilon)\left(1 - \frac{4\epsilon}{R} + o(\epsilon)\right) + J(A^\circ, \partial A)\left(\frac{4\epsilon}{R} + o(\epsilon)\right) + K(\partial A)o(\epsilon),$$

then

$$I(1) = I'(R) = \lim_{\epsilon \to 0} \frac{I(R) - I(R-\epsilon)}{\epsilon} = -\frac{4}{R} I(R) + \frac{4}{R} J(A, A),$$

using the fact that I and J are continuos function (they are integrals of continuos functions). The above formula is what is known for Crofton's formula (Solomon, 1978, p. 100).

It remais to calculate the expected distance between a point in the boundary of the circle and a point in the interior, $J(A, A)$. Without loss of generality, suppose that p is in the border and q in the interior of the circle. By its symmetry, and using the fact that distances are invariant to translations, we place p at the origin and the center of the circle the x-axis. That said, the distance between the points is only r_2. Then,

$$J(A, A) = \frac{1}{\pi R^2} \int_A r_2 dA.$$

In polar coordinates, q can be defined by (r_2, θ_2). Notice that

$$-\pi/2 \leq \theta_2 \leq \pi/2.$$

Besides that, let $c = \sup_{q \in \text{circle}} r_2$, then c is the distance between p and another point in the boundary of the circle. This distance can be calculated using the Law of Cosines for the triangle delimited by these points and the center. Then, $c = \sqrt{R^2 + R^2 - 2R^2 \cos(\pi - 2\theta)} = R\sqrt{2(1 + \cos(2\theta))}$. Since $\cos(2x) = \cos(x)^2 - \sin(x)^2 = 2\cos(x)^2 - 1$, we have that $c = R\sqrt{4\cos(\theta)^2} = 2R\cos(\theta)$, that is,

$$0 \leq r_2 \leq 2R\cos(\theta),$$

and (applying the Jacobian again),

$$\begin{aligned}
J(A, A) &= \frac{1}{\pi R^2} \int_{-\pi/2}^{\pi/2} \int_0^{2R\cos(\theta)} r_2^2 \, dr_2 \, d\theta_2 \\
&= \frac{1}{\pi R^2} \int_{-\pi/2}^{\pi/2} \frac{8R^3}{3} \cos(\theta_2)^3 \, d\theta_2 \\
&= \frac{1}{\pi R^2} \int_{-\pi/2}^{\pi/2} \frac{8R^3}{3} (1 - \sin(\theta_2)^2) \cos(\theta_2) \, d\theta_2 \\
&= \frac{8R}{3\pi} \int_{-1}^{1} (1 - u^2) \, du \\
&= \frac{32R}{9\pi}.
\end{aligned}$$

Finally,

$$I(1) = -\frac{4}{R} I(R) + \frac{4}{R} \frac{32R}{9\pi} = -4I(1) + \frac{128}{9\pi} \implies I(1) = \frac{128}{45\pi},$$

and

$$I(R) = \frac{128}{45\pi} R. \tag{2}$$

We conclude that I has a closed form given by (2). □

4. Propose a simulation algorithm to approximate I. Provide point and interval estimates and give theoretical guarantees about them (consistency, coverage, etc).

   Solution. Consider the following transformation,

   $$p_x = \sqrt{r_1}\cos(\theta_1), p_y = \sqrt{r_1}\sin(\theta_1), q_x = \sqrt{r_2}\cos(\theta_2), q_y = \sqrt{r_2}\sin(\theta_2).$$

   The Jacobian of this transformation is

   $$\begin{bmatrix} \frac{1}{2\sqrt{r_1}}\cos(\theta_1) & -\sqrt{r_1}\sin(\theta_1) \\ \frac{1}{2\sqrt{r_1}}\sin(\theta_1) & \sqrt{r_1}\cos(\theta_1), \end{bmatrix}$$

   whose determinant is $1/2$, then the distribution is

   $$g(r_1, \theta_1) = \frac{1}{2\pi R^2} 1_{[0,R^2]\times[0,2\pi]},$$

   what is the uniform distribution in $[0, R^2] \times [0, 2\pi]$. Then the algorithm can be as follows

Algorithm 1 Simulation for problem A

Require: $R > 0$, n_samples positive integer
  $r_1 \leftarrow$ Uniform(lower = 0, upper = 1, length = n_samples)
  $r_2 \leftarrow$ Uniform(lower = 0, upper = 1, length = n_samples)
  $\theta_1 \leftarrow$ Uniform(lower = 0, upper = 1, length = n_samples)
  $\theta_2 \leftarrow$ Uniform(lower = 0, upper = 1, length = n_samples)
  dist $\leftarrow \sqrt{r_1^2 + r_2^2 - 2r_1 r_2 \cos(2\pi(\theta_1 - \theta_2))}$
  mean_dist $\leftarrow R \cdot$ mean(dist)

   We want to evaluate the following integral, which is equivalent to I, as already mentioned previously

   $$\frac{1}{4\pi^2 R^4}\int_0^{R^2}\int_0^{R^2}\int_0^{2\pi}\int_0^{2\pi}\sqrt{r_1 + r_2 - 2\sqrt{r_1 r_2}\cos(\theta_1 - \theta_2)}\, d\theta_1\, d\theta_2\, dr_1\, dr_2.$$

   We use the following Monte Carlo simulation, let

   $$\{r_1^{(n)}\}_{1\le n\le N}, \{r_2^{(n)}\}_{1\le n\le N}, \{\theta_1^{(n)}\}_{1\le n\le N}, \{\theta_2^{(n)}\}_{1\le n\le N} \overset{iid}{\sim} \text{Unif}(0,1)$$

   and

   $$I_N^{MC} = \frac{R}{N}\sum_{k=1}^{N}\sqrt{r_1^{(k)} + r_2^{(k)} - 2\sqrt{r_1^{(k)} r_2^{(k)}}\cos(2\pi(\theta_1^{(k)} - \theta_2^{(k)}))}.$$

   By the Strong Law of Large Numbers, I_N^{MC} converges almost surely to I. Now we want to calculate the variance of the desired quantity. The second moment is given by

$$
\begin{aligned}
&\frac{1}{4\pi^2 R^4}\int_0^{R^2}\int_0^{R^2}\int_0^{2\pi}\int_0^{2\pi} r_1 + r_2 - 2\sqrt{r_1 r_2}\cos(\theta_1-\theta_2)\,d\theta_1\,d\theta_2\,dr_1\,dr_2.\\
&= \frac{1}{4\pi^2 R^4}\int_0^{R^2}\int_0^{R^2}\int_0^{2\pi} 2\pi(r_1+r_2)\,d\theta_2\,dr_1\,dr_2\\
&= \frac{1}{4\pi^2 R^4}\int_0^{R^2}\int_0^{R^2} 4\pi^2(r_1+r_2)\,dr_1\,dr_2\\
&= \frac{1}{R^4}\int_0^{R^2} \frac{1}{2}R^4 + R^2 r_2\,dr_2\\
&= \frac{1}{R^4}(\frac{1}{2}R^6 + \frac{1}{2}R^6) = R^2,
\end{aligned}
$$

therefore

$$
V = \mathrm{Var}\left[\sqrt{(p_x-q_x)^2+(p_y-q_y)^2}\right] = \left(1-\left(\frac{128}{45\pi}\right)^2\right)R^2,
$$

which can also be estimated by Monte Carlo Simulation. This expression tell us that the variance increases quadratically with R. The rate of convergence of this estimator is $O(n^{-1/2})$, but it also depends on R, as already mentioned. In order to provide interval estimates, by the Central Limit Theorem,

$$
\sqrt{N}(I_N^{MC} - I) \xrightarrow{d} \mathrm{Normal}(0, V).
$$

Supposing the standard deviation known, the confidence interval for I is

$$
I_N^{MC} \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\frac{\sqrt{V}}{\sqrt{N}},
$$

such that α is the level (commonly $\alpha = 0.05$) and Φ is the normal cumulative distribution. When it is unknown,

$$
I_N^{MC} \pm T_{N-1}^{-1}\left(1-\frac{\alpha}{2}\right)\frac{\sigma'}{\sqrt{N}},
$$

with T_{n-1} being the cumulative distribution of t-Student with $n-1$ degrees of freedom and σ'^2 which is the unbiased Monte Carlo estimator for V.

In Table 1, we provide some information about the simulations with $R = 5$.The confidence intervals converge to the same value with $N = 100$, which shows the Monte Carlo convergence for the variance. The point estimates also approximate to the tru value 4.527. Figure 1 presents these results in the general aspect.

□

N	**Time(s)**	**Point estimate**	**95 CI (known V)**	**95 CI**
10^1	0.0001161	4.9346	(3.62, 6.25)	(3.59, 6.28)
10^2	0.0001843	4.6448	(4.23, 5.06)	(4.24, 5.04)
10^3	0.0002418	4.5683	(4.36, 4.62)	(4.36, 4.62)
10^4	0.001179	4.5466	(4.51, 4.59)	(4.51, 4.59)
10^5	0.01124	4.5247	(4.51, 4.54)	(4.51, 4.54)
10^6	0.1318	4.5264	(4.52, 4.53)	(4.52, 4.53)

Table 1: Simulations results for some values of N.

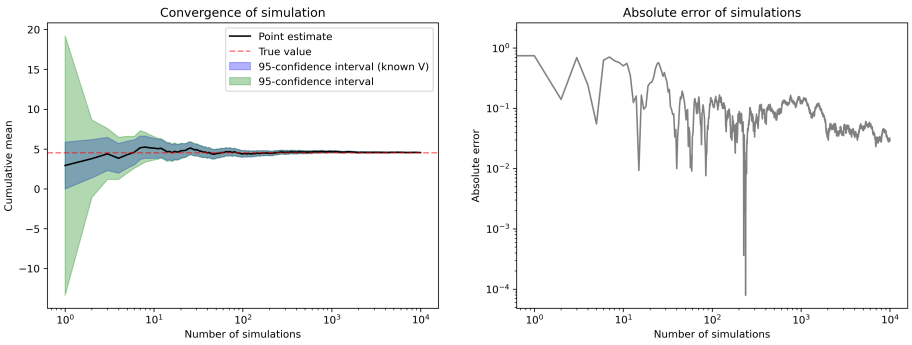

Figure 1: Convergence figure and absolute error for different values of N.

Bibliography

Rosenthal, J. S. (2006). *First Look At Rigorous Probability Theory, A*. World Scientific Publishing Company.

Solomon, H. (1978). *Geometric probability*. SIAM.