Assignment 0: O Brother, How Far Art Thou?

Computational Statistics
Instructor: Luiz Max de Carvalho
Student: Henrique Ennes

September 29, 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 $s/3$.

2. Show that Problem A is equivalent to computing

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

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

   Hint: Draw a picture.

3. Compute I in closed-form.

   Hint: Look up *Crofton's mean value theorem* or *Crofton's formula*.

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

Solution (1)**.** Suppose x_1, x_2 are independent random samples from the uniform distribution $U[0, s]$. Therefore, the average distance between the random variables X_1 and X_2 thus sampled is

$$\mathbb{E}|X_1 - X_2|.$$

However, by the Law of Total Probability

$$\begin{aligned}
\mathbb{E}|X_1 - X_2| &= \mathbb{E}[\mathbb{E}(|X_1 - x_2| \mid X_1 = x_1)] \\
&= \int_0^s \frac{1}{s} \left(\int_0^s \frac{|x_1 - x_2|}{s} dx_1 \right) dx_2 \\
&= \int_0^s \frac{1}{s} \left(\int_0^{x_2} \frac{x_2 - x_1}{s} dx_1 + \int_{x_2}^s \frac{x_1 - x_2}{s} dx_1 \right) dx_2 \\
&= \int_0^s \frac{1}{s} \left(\frac{s}{2} - x_2 + \frac{x_2^2}{s} \right) dx_2 \\
&= \frac{s}{3}.
\end{aligned}$$

Solution (2)**.** Now, let $\mathbf{p}, \mathbf{q}$ be independent random samples from a uniform distribution on the disk C_R, $U(C_R)$, implying both having density $\frac{1}{\pi R^2}$. Again, by the same reasoning of last problem, the average distance between the random variables $\mathbf{P}$ and $\mathbf{Q}$ is

$$\mathbb{E}\|\mathbf{P} - \mathbf{Q}\|.$$

We therefore, again by the Law of Total Probability,m, have

$$\begin{aligned}\mathbb{E}\|\mathbf{P}-\mathbf{Q}\| &= \mathbb{E}[\mathbb{E}(\|\mathbf{P}-\mathbf{q}| \mid \mathbf{Q}=\mathbf{q})] \\ &= \int_{C_R} \frac{1}{\pi R^2}\left(\int_{C_R} \frac{1}{\pi R^2}\|\mathbf{p}-\mathbf{q}\| d\mathbf{p}\right) d\mathbf{q} \\ &= \frac{1}{\pi^2 R^4}\int_{C_R}\left(\int_{C_R}\sqrt{(p_x-q_x)^2+(p_y-q_y)^2} dp_x dp_y\right) dq_x dq_y.\end{aligned}$$

This integral is, however, better expressed in polar coordinates. For that, we transform

$$\begin{aligned} p_x &= r_1\cos\theta_1 \quad p_y = r_1\sin\theta_1 \\ q_x &= r_2\cos\theta_2 \quad p_y = r_2\sin\theta_2,\end{aligned}$$

and the integration limits become

$$\begin{aligned} 0 \le r_1 \le R \quad 0 \le \theta_1 \le 2\pi \\ 0 \le r_2 \le R \quad 0 \le \theta_2 \le 2\pi,\end{aligned}$$

and the measure of each integral becomes $r_1 dr_1 d\theta_1$ and $r_2 dr_2 \theta_2$, given the Jacobian of the transformation from Cartesian to polar coordinates as $|J| = |\cos\theta(-r\cos\theta) - (-r\sin\theta)\sin\theta| = r$.

Therefore, by Fubini's theorem, as $\|\mathbf{P}-\mathbf{Q}\|$ is clearly integrable, given it is bounded by $2R$, and making use of the trigonometry identities given through

$$\begin{aligned}\sin^2\phi + \cos^2\phi &= 1 \\ \cos(\phi_1-\phi_2) &= \cos\phi_1\cos\phi_2 + \sin\phi_1\sin\phi_2,\end{aligned}$$

for any $\phi, \phi_1, \phi_2 \in \mathbb{R}$, we have that

$$\begin{aligned}\mathbb{E}\|\mathbf{P}-\mathbf{Q}\| &= \frac{1}{\pi^2 R^4}\int_{C_R}\left(\int_{C_R}\sqrt{(p_x-q_x)^2+(p_y-q_y)^2} dp_x dp_y\right) dq_x dq_y \\ &= \frac{1}{\pi^2 R^4}\int_0^R\int_0^{2\pi}\int_0^R\int_0^{2\pi}\sqrt{(r_1\cos\theta_1 - r_2\cos\theta_2)^2 + (r_1\sin\theta_1 - r_2\sin\theta_2)^2} r_1 r_2 d\theta_1 dr_1 d\theta_2 dr_2 \\ &= \frac{1}{\pi^2 R^4}\int_0^R\int_0^{2\pi}\int_0^R\int_0^{2\pi}(r_1^2\cos^2\theta_1 + r_2^2\cos^2\theta_2 - 2r_1r_2\cos\theta_1\cos\theta_2 \\ &\quad + r_1^2\sin^2\theta_1 + r_2^2\sin^2\theta_2 - 2r_1r_2\sin\theta_1\sin\theta_2)^{1/2} r_1 r_2 d\theta_1\theta_2 dr_1 dr_2 \\ &= \frac{1}{\pi^2 R^4}\int_0^R\int_0^{2\pi}\int_0^R\int_0^{2\pi}[r_1^2(\sin^2\theta_1+\cos^2\theta_1) + r_2^2(\sin^2\theta_1+\cos^2\theta_1) \\ &\quad - 2r_1r_2(\cos\theta_1\cos\theta_2 + \sin\theta_1\sin\theta_2)]^{1/2} r_1 r_2 d\theta_1 d\theta_1 dr_1 dr_2 \\ &= \frac{1}{\pi^2 R^4}\int_0^R\int_0^{2\pi}\int_0^R\int_0^{2\pi}\sqrt{r^1 + r^2 - 2r_1r_2\cos(\theta_1-\theta_2)} r_1 r_2 d\theta_1 d\theta_2 dr_1 dr_2 \\ &= \frac{1}{\pi^2 R^4}\int_0^R\int_0^{2\pi}\int_0^R\int_0^{2\pi}\sqrt{r^1 + r^2 - 2r_1r_2\cos\phi(\theta_1,\theta_2)} r_1 r_2 d\theta_1 d\theta_2 dr_1 dr_2,\end{aligned}$$

as $\phi(\theta_1,\theta_2) = \theta_1 - \theta_2$, that is, is the angle between r_1 and r_2.

Solution (3)**.** Instead of evaluating the integral in equation (1) directly we can, instead, use Crofton's Theorem on Mean Values [Mat99], stated as following:

Theorem (Crofton's Theorem on Mean Values)**.** *Let D be a domain in $\mathbb{R}^k$ of volume V. If μ is an invariant expected value of a function of $x_1, ..., x_n$, which are random independent uniform samples in D and μ_1, the same μ when exactly one point lies in the boundary of D, D_1, and $n-1$ are internal in D, then*

$$\frac{d\mu}{dV} = \frac{n}{V}(\mu_1 - \mu). \tag{2}$$

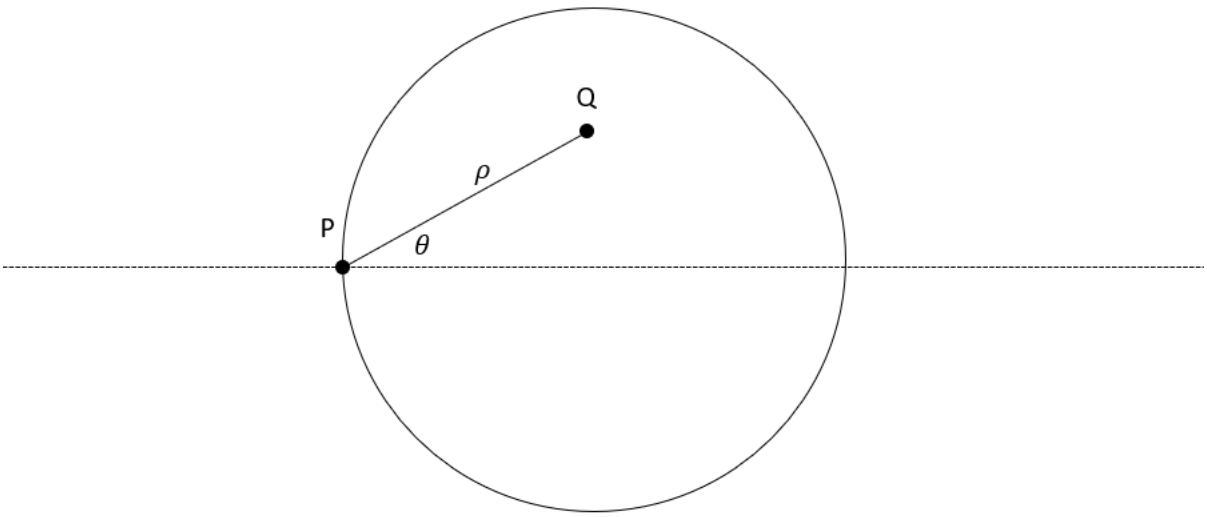

Figure 1: Depiction of the geometric construction when P lies in the boundary of the disk C_R. Notice that the axis falls within the line joining P and the disk's center. Additionally, we define ρ as the distance between P and the random point Q and θ as the angle between Q and the axis (i.e. between the disk center, P and Q).

Let us calculate μ_1, that is, the expected value of the distance between P and Q, given one of them (without loss of generality, say P) in the boundary of the disk $C_R = D$. Let us align the axis so that both P and the disk's center lie on it. Also, assign ρ as the distance between P and the random point Q and θ as the angle between Q and the axis, as shown in Figure 1.

Notice that $0 \leq \rho \leq 2R$ and $\frac{-\pi}{2} \leq \theta \leq \frac{\pi}{2}$, we get that, in polar coordinates, now using the measure $\rho d\rho d\theta$ and density $\frac{1}{\pi R^2}$

$$\begin{aligned}
\mu_1 &= \int_{-\pi/2}^{\pi/2} \int_0^{2R\cos\theta} \left(\frac{1}{\pi R^2}\,\rho\right) \rho\, d\rho\, d\theta \\
&= \frac{8R}{3\pi} \int_{-\pi/2}^{\pi/2} \cos^3\theta d\theta \\
&= \frac{32R}{9\pi}.
\end{aligned}$$

Applying this into equation (2), with $n = 2$, gives

$$d\mu = \frac{2}{\pi R^2}\left(\mu - \frac{32R}{9\pi}\right) dV,$$

as $dV = 2\pi R dR$. Therefore,

$$d\mu = \frac{4}{R}\left(\mu - \frac{32R}{9\pi}\right) dR.$$

Noticing, however, that $d(\mu R^4) = R^4 d\mu + 3R^3 \mu dR$, we have, multiplying both sides of the above by R^4

$$R^4 d\mu = 4R^3 \mu dR - 4R^3 \frac{32R}{9\pi} dR,$$

or, rearranging terms

$$d(\mu R^4) = R^4 d\mu + 4R^3 \mu dR = \frac{128R^4}{9\pi} dR.$$

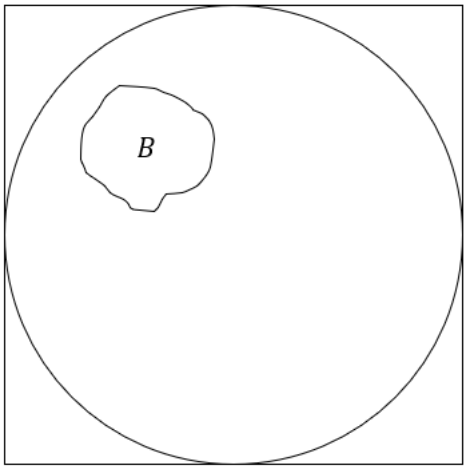

Figure 2: Depiction of the Borel set B contained in the circle C_R, itself contained in a squared of side R.

Integrating both sides in dR from 0 to R gives

$$R^4\mu = \frac{128R^5}{45\pi} + c.$$

However, as clearly $\mu \to 0$ when $R \to 0$ (the circle becomes a point as $R \to 0$, so both P and Q must coincide), $c = 0$. Therefore

$$\mu = \frac{128R}{45\pi} \tag{3}$$

solves equation (1).

Solution (4). For the computational part [1] of the problem, we shall first, for any fixed radius R, sample uniformly on the circle of disk C_R. We will do so by using the following very naive algorithm for sampling n points That this algorithm indeed produces a uniform distribution on

Algorithm 1 Uniform sample on the disk C_R

```
while length(points) ≤ n do
    x ← uniform random(−R, R)
    y ← uniform random(−R, R)
    if x² + y² ≤ R² then
        points ← points.append((x, y))
    end if
end while
```

the circle is easy to be verified[2]. For such, suppose B is a Borel set of measure A fully contained in C_R and P is a point sampled from Algorithm 1 (Figure 2). Therefore, the probability of P being in B is given by

$$\mathbb{P}(P \in B) = \mathbb{P}(P \in B | P \in C_R) = \frac{\mathbb{P}(P \in B \cap P \in C_R)}{\mathbb{P}(P \in C_R)} = \frac{\mathbb{P}(P \in B)}{\mathbb{P}(P \in C_R)} = \frac{\frac{A}{R^2}}{\frac{\pi R^2}{R^2}} = \frac{A}{\pi R^2},$$

which is exactly the probability of $P \in B$ in the distribution U_{C_R}.

[1] The Python code used for defining the functions necessary for simulation is available at https://github.com/HLovisiEnnes/Practice_problems/blob/main/Assignment_0 — Computational%20Statistics/Assignment_0.py. The diagrams and tables presented below were done using simple packages, as *matplotlib* and *pandas*.

[2] In fact, this turns out to be a specific case of rejection sampling, but as we have not seen that method before this assignment presented, I devised a proof on my own.

R	I	Estimator of I	Beginning of Confidence Interval	End of Confidence Interval
0.1	0.090541	0.090072	0.088221	0.091924
0.5	0.452707	0.451224	0.441951	0.460497
1.0	0.905415	0.895565	0.877122	0.914008
2.0	1.810830	1.811611	1.774385	1.848837
10.0	9.054148	8.963193	8.778571	9.147814
20.0	18.108296	18.170211	17.797306	18.543116
30.0	27.162444	27.166269	26.607613	27.724925

Figure 3: Table depicting the simulated value of I, $\hat{I}_n$, for distinct R values and $n = 2000$, together with 5% confidence intervals calculated through equation(5).

Finally, to calculate the integral in equation (1), we shall use the estimator

$$\hat{I}_{n,m} = \frac{1}{n}\frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m}\|P_i - Q_j\|, \tag{4}$$

where P_i and Q_j are distinct samples drawn through the Algorithm 1 above. Through the usual convergence theorems (Strong Law of Large Numbers, Central Limit Theorem, etc.), we know $\hat{I}_n$ to be a consistent efficient estimator of $\mathbb{E}\|P - Q\|$, which, as we have seen in part (2), is given by the I defined through equation (1); that is, we have $\hat{I}_{n,m}$ the Monte Carlo estimator for I. For practical purposes, we choose $m = n$ and denote the estimator $\hat{I}_{n,n}$ by $\hat{I}_n$.

With that, we show in Figure 3, for different R values, the theoretically expected value of I through equation (3), together with the simulated value calculated through equation (4) with $n = m = 2000$, where the points P and Q are sampled according to Algorithm 1, together with a 95% confidence interval. The confidence interval is calculated through the Central Limit Theorem applied to our estimator, where we know, at the large n limit

$$\frac{\sqrt{n}}{\sigma}(\hat{I}_n - I) \approx N(0,1),$$

where $\sigma^2 = \frac{1}{R^2\pi^2}\int_{C_R}\int_{C_R}(\|P - Q\| - I)^2 d\mathbf{p}\, d\mathbf{q}$. Also recalling that

$$\hat{S}_n^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\|P_i - Q_j\| - \hat{I}_n)^2$$

is a maximum likelihood estimator for σ^2 and, consequently, is efficient, at the large n limit [Kee10], $S_n^2 \approx \sigma^2$. Therefore, given c the $\frac{5\%}{2} = 2.5\%$ quantile of the standard normal $N(0,1)$, we have

$$P\left(|\hat{I}_n - I| > c\frac{\sigma}{\sqrt{n}}\right) \approx 0.95,$$

giving the confidence interval approximately as

$$\left(\hat{I}_n - c\frac{\hat{S}_n}{\sqrt{n}}, \hat{I}_n + c\frac{\hat{S}_n}{\sqrt{n}}\right), \tag{5}$$

where we will be taking $c = 1.96$ and $n = 2000$.

Figure 4 demonstrates, for $R = 1$, the dependency of the estimated value on n. Notice the expected decrease of the error $|\hat{I}_n - I|$ as $\frac{1}{\sqrt{n}}$ is visible.

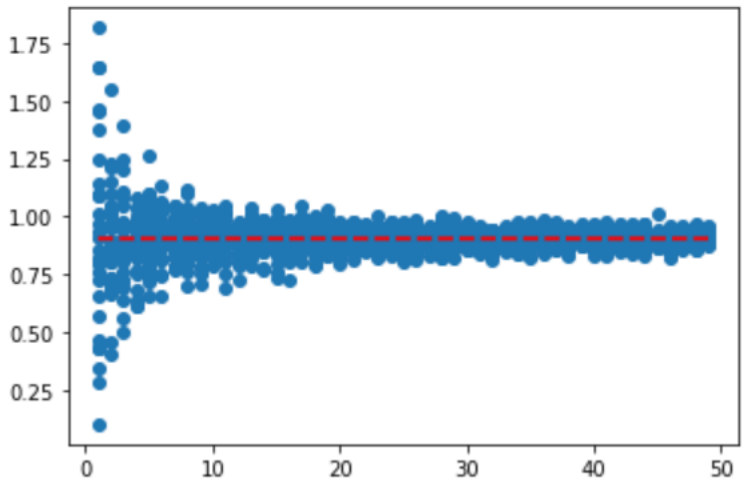

Figure 4: Plot of the simulated $\hat{I}_n$ as a function of n (blue points) for $R = 1$, together with the theoretically predicted value of $I = \frac{128}{45\pi}$ (red dashed line). Notice the convergence of predictions as $\frac{1}{\sqrt{n}}$ to I.

Bibliography

[Kee10] Robert W. Keener. *Theoretical Statistics Topics for a Core Course; Chapter 9*. Springer Science+Business Media, LLC, 2010.

[Mat99] A. M. Mathai. *An introduction to geometrical probability: Distributional aspects with applications; Section 2.2.5*. Gordon and Breach, Science Pub., 1999.