SI252 Reinforcement Learning: Homework #01

Due on March 9, 2025 at 11:59 a.m.(CST)

Name: **Zhou Shouchen**
Student ID: 2021533042

Problem 1

Let X, Y be jointly Gaussian random variables. Show the following equality:

$$\mathbb{E}[Y|X] = L[Y|X] = \mathbb{E}(Y) + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}(X - \mathbb{E}(X))$$

Solution

Firstly prove that $L[Y|X] = \mathbb{E}(Y) + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}(X - \mathbb{E}(X))$:

Since $L[Y|X]$ is the best linear estimator of Y given X, we can write $L[Y|X] = a + bX$. Then we define that

$$f(a,b) = \mathbb{E}[(Y - a - bX)^2] = a^2 + \mathbb{E}(Y^2) + b^2\mathbb{E}(X^2) - 2a\mathbb{E}(Y) + 2ab\mathbb{E}(X) - 2b\mathbb{E}(Y)$$

To minimizes $f(a,b)$, we need to set the first derivative of $f(a,b)$ to 0:

$$\begin{aligned}\frac{\partial f}{\partial a} &= 2a - 2\mathbb{E}(Y) + 2b\mathbb{E}(X) = 0\\ \frac{\partial f}{\partial b} &= 2b\mathbb{E}(X^2) + 2a\mathbb{E}(X) - 2\mathbb{E}[XY] = 0\end{aligned}$$

Solve the equations, we can get that

$$a = \mathbb{E}(Y) - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\mathbb{E}(X), \quad b = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$$

We also need to prove that $\nabla^2 f(a,b)$ is positive semidefinite to ensure that (a,b) is the minimum point:

$$\nabla^2 f(a,b) = \begin{bmatrix} 2 & 2\mathbb{E}(X) \\ 2\mathbb{E}(X) & 2\mathbb{E}(X^2) \end{bmatrix} \succeq 0$$

So above all

$$L[Y|X] = \mathbb{E}(Y) + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}(X - \mathbb{E}(X))$$

Then prove that $\mathbb{E}[Y|X] = L[Y|X]$:

1. $Y - L[Y|X] \perp X$, i.e. $\mathbb{E}\left[(Y - L[Y|X])\,X\right] = 0$:

$$\begin{aligned}\mathbb{E}[(Y - L[Y|X])X] &= \mathbb{E}\left[XY - X\left(\mathbb{E}(Y) + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}(X - \mathbb{E}(X))\right)\right]\\ &= \mathbb{E}(XY) - \mathbb{E}\left[X\mathbb{E}(Y)\right] - \mathbb{E}\left[\frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \cdot X^2\right] + \mathbb{E}\left[\frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \cdot X\mathbb{E}(X)\right]\\ &= [\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)] - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\left[\mathbb{E}(X^2) - \mathbb{E}^2(X)\right]\\ &= \mathrm{Cov}(X,Y) - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \cdot \mathrm{Var}(X)\\ &= 0\end{aligned}$$

2. $\mathbb{E}\left[Y - L[Y|X]\right] = 0$:

$$\begin{aligned}\mathbb{E}[Y - L[Y|X]] &= \mathbb{E}\left[Y - \mathbb{E}(Y) - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \cdot [X - \mathbb{E}(X)]\right]\\ &= \mathbb{E}(Y) - \mathbb{E}(Y) - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \cdot \mathbb{E}\left[X - \mathbb{E}(X)\right]\\ &= 0\end{aligned}$$

Problem 1 continued on next page...

3. Two Gaussian variable's Orthogonal, one of which has a 0 expectation, is equivalent to uncorrelated:
   SSince $X \perp Y$, so $\mathbb{E}(XY) = 0$, so $\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0 \Rightarrow \rho = 0$. So X, Y are uncorrelated.
   And if X, Y are uncorrelated, then $\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(XY) = 0 \Rightarrow \mathbb{E}(XY) = 0$.
   So two Gaussian variable's Orthogonal is equivalent to uncorrelated.

4. $Y - L[Y|X]$ and X are joint normal: $\forall a_1, a_2 \in \mathbb{R}$,

$$\begin{aligned} a_1(Y - L[Y|X]) + a_2 X &= a_1 Y - a_1\mathbb{E}(Y) - a_1\frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}X + a_1\frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\mathbb{E}(X) + a_2 X \\ &= \left(a_2 - a_1\frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\right)X + a_1 Y + \left[a_1\frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\mathbb{E}(X) - a_1\mathbb{E}(Y)\right] \end{aligned}$$

   Since X, Y are jointly Gaussian random variables, so $\forall a_3, a_4, a_5 \in \mathbb{R}$, $a_3 Y + a_4 X + a_5$ is also a Gaussian distribution. Here take $a_3 = \left(a_2 - a_1\frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\right), a_4 = a_1, a_5 = a_1\frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\mathbb{E}(X) - a_1\mathbb{E}(Y)$, so $Y - L[Y|X]$ and X are joint normal distribution.

5. $Y - L[Y|X]$ and X are independent: Combined with step step 1. we know that $\mathbb{E}[(Y - L[Y|X])X] = 0$. From step 2. we know that $\mathbb{E}[(Y - L[Y|X])X] = 0$, so we have $Y - L[Y|X]$ and X are uncorrelated, which is certainly independent. Thus, we can further conclude that for any function of X : $\forall \phi(X)$, $Y - L[Y|X]$ and $\phi(X)$ are independent.

6. $Y - L[Y|X]$ and $\phi(X)$ are uncorrelated: Since from step 2. $\mathbb{E}\left[Y - L[Y|X]\right] = 0$, so $\mathbb{E}\left[\left(Y - L[Y|X]\right)\phi(X)\right] = \mathbb{E}\left[Y - L[Y|X]\right]\mathbb{E}\left[\phi(X)\right] = 0 \Rightarrow \mathrm{Cov}(X,Y)$. So $Y - L[Y|X]$ and $\phi(X)$ are uncorrelated.

7. $Y - L[Y|X] \perp \phi(X)$: We have shown that $\mathbb{E}\left[\left(Y - L[Y|X]\right)\phi(X)\right] = 0, \forall \phi(X)$, so $Y - L[Y|X]$ and $\phi(X)$ are orthogonal.

8. $L[Y|X] = \mathbb{E}[Y|X]$: So we can say that $Y - L[Y|X]$ is orthogonal to any function of X, which means that $L[Y|X]$ is the projector of Y onto the space of functions of X. So $L[Y|X]$ is the best linear estimator of Y given X, which is $\mathbb{E}[Y|X]$.

   So above all, we have proved that $L[Y|X] = \mathbb{E}[Y|X]$.

Problem 2

We wish to estimate the probability of landing heads, denoted by θ, of a biased coin. We model θ as the value of a random variable Θ with a known prior PDF $f_\Theta \sim \text{Unif}(0,1)$. We consider n independent tosses and let X be the number of heads observed. Find the MMSE $\mathbb{E}[\Theta|X]$ and the LLSE $L[\Theta|X]$.

Solution

Since $\Theta \sim \text{Unif}(0,1) \sim \text{Beta}(1,1)$, accoding to the Beta-Binomial conjugacy, after n tosses with $X = x$ heads, the posterior distribution of Θ is

$$\Theta|X = x \sim \text{Beta}(x+1, n-x+1)$$

And since the expectation of $\text{Beta}(\alpha,\beta)$ is $\dfrac{\alpha}{\alpha+\beta}$, so

$$\mathbb{E}[\Theta|X=x] = \frac{x+1}{(x+1)+(n-x+1)} = \frac{x+1}{n+2} \Rightarrow \mathbb{E}_\Theta[\Theta|X] = \frac{X+1}{n+2}$$

Since $X|\Theta \sim \text{Bin}(n,\Theta)$, so $\mathbb{E}_\Theta[X|\Theta] = n\Theta$, $\text{Var}_\Theta[X|\Theta] = n\Theta(1-\Theta)$.
From Adam's Law (Law of Iterative Expectation), we can get that:

$$\mathbb{E}(X) = \mathbb{E}_\Theta\left[\mathbb{E}_X(X|\Theta)\right] = \mathbb{E}(n\Theta) = n\mathbb{E}(\Theta) = \frac{n}{2}$$

From Eve's Law (Law of Total Variance), we can get that:

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}_\Theta[\text{Var}_X(X|\Theta)] + \text{Var}_\Theta[\mathbb{E}_\Theta(X|\Theta)] \\
&= \mathbb{E}\left[n\Theta(1-\Theta)\right] + \text{Var}(n\Theta) \\
&= n\mathbb{E}(\Theta) - n\mathbb{E}(\Theta^2) + n^2\text{Var}(\Theta) \\
&= \frac{n}{2} - \frac{n}{3} + n^2\frac{1}{12} \\
&= \frac{n^2}{12} + \frac{n}{6}
\end{aligned}$$

Also, Since $\Theta \sim \text{Unif}(0,1)$ $\mathbb{E}(\Theta) = \dfrac{1}{2}$, $\text{Var}(\Theta) = \dfrac{1}{12}$, we can get that

$$\text{Var}(\Theta) = \frac{1}{12} = \mathbb{E}(\Theta^2) - \mathbb{E}(\Theta)^2 \Rightarrow \mathbb{E}(\Theta^2) = \frac{1}{3}$$

Again, use Adam's Law, we can get that:

$$\mathbb{E}(\Theta X) = \mathbb{E}_\Theta\left[\mathbb{E}_X\left(\Theta X|\Theta\right)\right] = \mathbb{E}_\Theta\left[\Theta\mathbb{E}_X(X|\Theta)\right] = \mathbb{E}\left(\Theta n\Theta\right) = n\mathbb{E}(\Theta^2) = n\text{Var}(\Theta) + n^2\mathbb{E}^2(\Theta) = \frac{n}{3}$$

Thus

$$\text{Cov}(\Theta, X) = \mathbb{E}[\Theta X] - \mathbb{E}(\Theta)\mathbb{E}[X] = \frac{n}{3} - \frac{1}{2}\cdot\frac{n}{2} = \frac{n}{12}$$

So we have:

$$L[\Theta|X] = \mathbb{E}(\Theta) + \frac{\text{Cov}(\Theta,X)}{\text{Var}(X)}(X - \mathbb{E}[X]) = \frac{1}{2} + \frac{\frac{n}{12}}{\frac{n^2}{12}+\frac{n}{6}}(X - \frac{n}{2}) = \frac{1}{2} + \frac{1}{n+2}\left(X - \frac{n}{2}\right) = \frac{X+1}{n+2}$$

We can find that this setting also satisfies $\mathbb{E}[\Theta|X] = L[\Theta|X]$.

Problem 3

Sampling from the discrete distribution.

(a) Given n positive real number $a_1, \ldots, a_n$, where $\sum_{j=1}^{n} a_j = 1$; and n i.i.d. random variables $X_1, \ldots, X_n \sim$ Gumbel$(0, 1)$. Show the following equality:

$$P\left(\log a_i + X_i = \max_{j \in \{1,\ldots,n\}} (\log a_j + X_j)\right) = a_i$$

(b) Illustrate how the above result leads to new sampling methods for the discrete distribution.

Solution

We can get the PDF and CDF of Gumbel distribution as follows:

$$F(x; \mu, \sigma) = e^{-e^{-\frac{x-\mu}{\sigma}}}, f(x; \mu, \sigma) = \frac{1}{\sigma} e^{-\frac{x-\mu}{\sigma}} e^{-e^{-\frac{x-\mu}{\sigma}}}, -\infty < x < \infty$$

Here $\mu = 0, \sigma = 1$, so

$$F(x) = e^{-e^{-x}}, f(x) = e^{-x} e^{-e^{-x}}$$

Let $Y_i = \log a_i + X_i$, then

$$\begin{aligned}
&P\left(\log a_i + X_i = \max_{j \in \{1,\ldots,n\}} (\log a_j + X_j)\right) \\
&= P(Y_i = \max_{j \in \{1,\ldots,n\}} Y_j) \\
&= \int_{-\infty}^{\infty} P(Y_i = \max_{j \in \{1,\ldots,n\}} Y_j | Y_i = x) f_{Y_i}(x) \mathrm{d}x \qquad \text{(LOTP)} \\
&= \int_{-\infty}^{\infty} P(Y_i = \max_{j \in \{1,\ldots,n\}} Y_j | Y_i = x) f_{X_i}(x - \log a_i) \mathrm{d}x \\
&= \int_{-\infty}^{\infty} P(Y_1 \le Y_1, \ldots, Y_n \le Y_i | Y_i = x + \log a_i) f(x) \mathrm{d}x \\
&= \int_{-\infty}^{+\infty} \left(\prod_{j=1, j \neq i}^{n} P(X_j \le x + \log a_i - \log a_j) \right) f(x) \mathrm{d}x \\
&= \int_{-\infty}^{+\infty} \left(\prod_{j=1, j \neq i}^{n} F(x + \log a_i - \log a_j) \right) f(x) \mathrm{d}x \\
&= \int_{-\infty}^{+\infty} \left(\prod_{j=1, j \neq i}^{n} \exp(-\exp(-x - \log a_i + \log a_j)) \right) e^{-x} e^{-e^{-x}} \mathrm{d}x \\
&= \int_{-\infty}^{+\infty} \left(\prod_{j=1, j \neq i}^{n} \exp(-e^{-x} \cdot e^{-\log a_i} \cdot e^{\log a_j}) \right) e^{-x} e^{-e^{-x}} \mathrm{d}x \\
&= \int_{-\infty}^{+\infty} \left(\prod_{j=1, j \neq i}^{n} \exp(-e^{-x} \cdot \frac{a_j}{a_i}) \right) e^{-x} e^{-e^{-x}} \mathrm{d}x \\
&= \int_{-\infty}^{+\infty} \exp\left(-e^{-x} \sum_{j=1, j \neq i}^{n} \frac{a_j}{a_i} \right) e^{-x} e^{-e^{-x}} \mathrm{d}x
\end{aligned}$$

$$
\begin{aligned}
&= \int_{-\infty}^{+\infty} \exp\left(-u \sum_{j=1, j\neq i}^{n} \frac{a_j}{a_i}\right) u e^{-u} \left(-\frac{1}{u}\right) \mathrm{d}u \qquad (u = e^{-x}) \\
&= \int_{0}^{+\infty} \exp\left(-u \sum_{j=1, j\neq i}^{n} \frac{a_j}{a_i} - u\right) \mathrm{d}u \\
&= \int_{0}^{+\infty} \exp\left[-u\left(1 + \sum_{j=1, j\neq i}^{n} \frac{a_j}{a_i}\right)\right] \mathrm{d}u
\end{aligned}
$$

Since the PDF of Expo(λ) is $\lambda e^{-\lambda x}$, here we take $\lambda = 1 + \sum_{j=1, j\neq i}^{n} \frac{a_j}{a_i}$, so

$$
\begin{aligned}
&\int_{0}^{+\infty} \exp\left[-u\left(1 + \sum_{j=1, j\neq i}^{n} \frac{a_j}{a_i}\right)\right] \mathrm{d}u \\
&= \frac{1}{\lambda} \int_{0}^{+\infty} \lambda \exp\left(-\lambda u\right) \mathrm{d}u \\
&= \frac{1}{\lambda} \\
&= \frac{1}{1 + \sum_{j=1, j\neq i}^{n} \frac{a_j}{a_i}} \\
&= \frac{a_i}{\sum_{j=1}^{n} a_j} \\
&= a_i.
\end{aligned}
$$

(b) Suppose we have a discrete random variable X whose PMF is

$$
P(X = x_i) = a_i, \quad i \in \{1, 2, \ldots, n\}
$$

We can sample X by:

1. Sample n independent values $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Gumbel}(0, 1)$
2. Let $i = \arg\max_{j \in \{1,2,\ldots,n\}} (\log a_j + X_j +)$
3. Set $X \leftarrow x_i$.

Problem 4

Adopt the Acceptance-Rejection method to estimate the value of π, then evaluate the performance of Monte Carlo algorithms with finite number of samples.

Solution

We can estimate π via Monte Carlo method. Sample $X, Y \overset{i.i.d.}{\sim} \text{Unif}(0,1)$. If (x, y) is in the unit circle in the first quadrant, then $x^2 + y^2 \leq 1$, we count the number of these points as N_{in}. Then the estimatied $\hat{\pi} = 4 \cdot \dfrac{N_{\text{in}}}{N}$, where N is the total number of sample points.

The number of sample points N are set to be $10, 10^2, 10^3, 10^4, 10^5, 10^6$ respectively. The results are shown in the following figure, the estimated $\hat{\pi}$, estimation error $|\hat{\pi} - \pi|$. We could discover that more sample points could lead to a more accurate estimation of π.

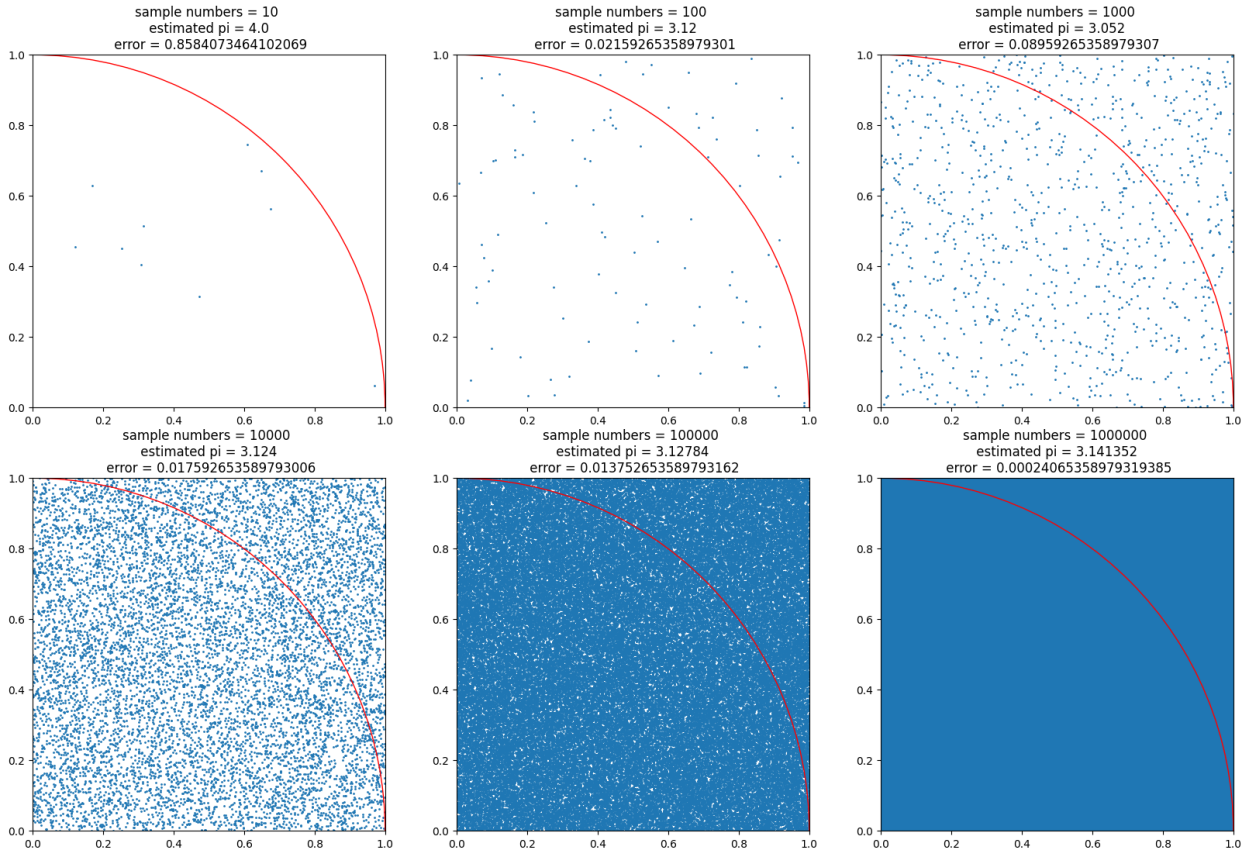

Figure 1: The estimation of π via Monte Carlo method

Let $Z_i = \mathbb{1}\{\text{The i-th sample is in the circle of the first quadrant}\}$. Then the estimated π is

$$\hat{\pi} = 4 \cdot \frac{1}{n} \sum_{i=1}^{n} Z_i$$

The error probability can be calculated as

$$P\left(|\pi - \hat{\pi}| \geq \epsilon\right) = P\left(\left|\frac{4}{n}\sum_{i=1}^{n} Z_i - \pi\right| \geq \epsilon\right) = P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \frac{\pi}{4}\right| \geq \frac{\epsilon}{4}\right)$$

According to the Hoeffding's inequality, we have

$$P\left(\left|\sum_{i=1}^{n} Z_i - \frac{n\pi}{4}\right| \geq \frac{n\epsilon}{4}\right) \leq 2\exp\left(-\frac{2n\left(\frac{\epsilon}{4}\right)^2}{(1-0)^2}\right) = 2e^{-\frac{1}{8}n\epsilon^2} = \delta$$

Where $1-\delta$ is the confidence level. So we can get that $\epsilon = \sqrt{\frac{8\ln\frac{2}{\delta}}{n}}$. Which means that

$$P\left(\pi \in \left(\hat{\pi} - \sqrt{\frac{8\ln\frac{2}{\delta}}{n}}, \hat{\pi} + \sqrt{\frac{8\ln\frac{2}{\delta}}{n}}\right)\right) \geq 1-\delta$$

Here we set $\delta = 0.05$, which means that we have a 95% confidence level that the estimated π is in the interval. The upper bound and lower bound of the interval are shown in the following figure. And we can see that the estimated $\hat{\pi}$ has mostly been in the interval.

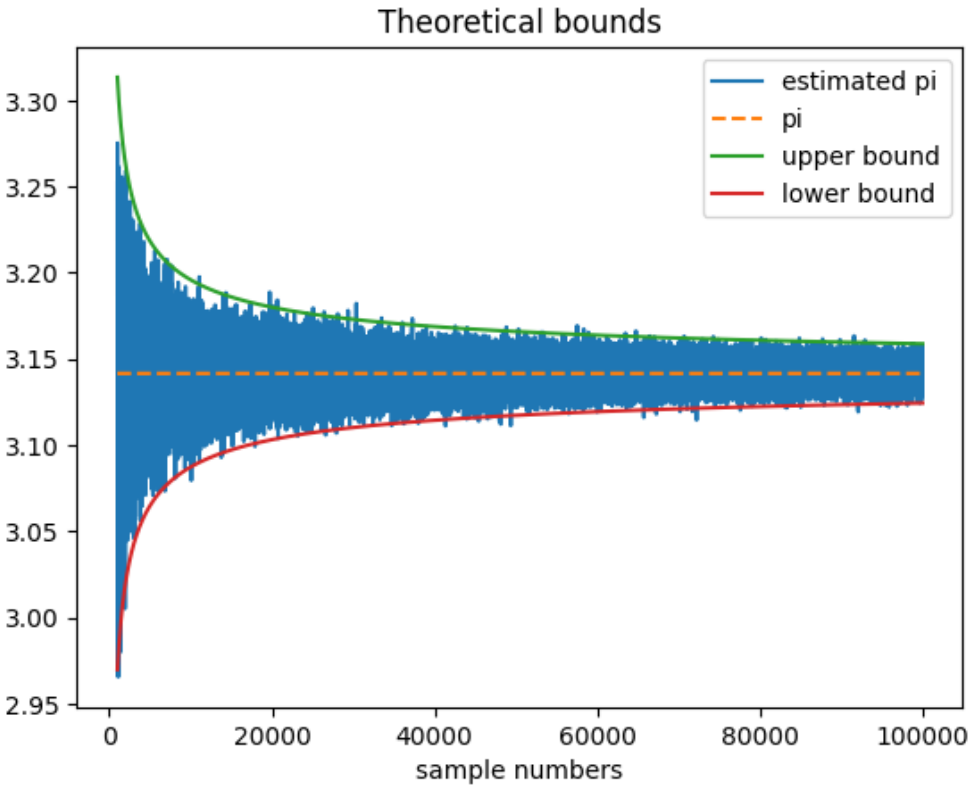

Figure 2: Analysis the performance of the estimated $\hat{\pi}$

Problem 5

Sampling from probability distributions. Show histograms and compare them to corresponding PDFs.
(a) Sampling from the standard Normal distribution with both the Box-Muller method and the Acceptance-Rejection method. Discuss the pros and cons of both methods.
(b) Sampling from the distribution with the following pdf:

$$f(x) \propto \exp\left(-\frac{1}{2}x^2\right)\left(\sin^2(6x) + 3\cos^2(x)\sin^2(4x) + 1\right)$$

Solution

(a) We could see that $X, Y \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, the estimated correlation coefficient is getting closer to, and the sampled distribution is getting closer to the theoretical distribution as the number of samples increases.

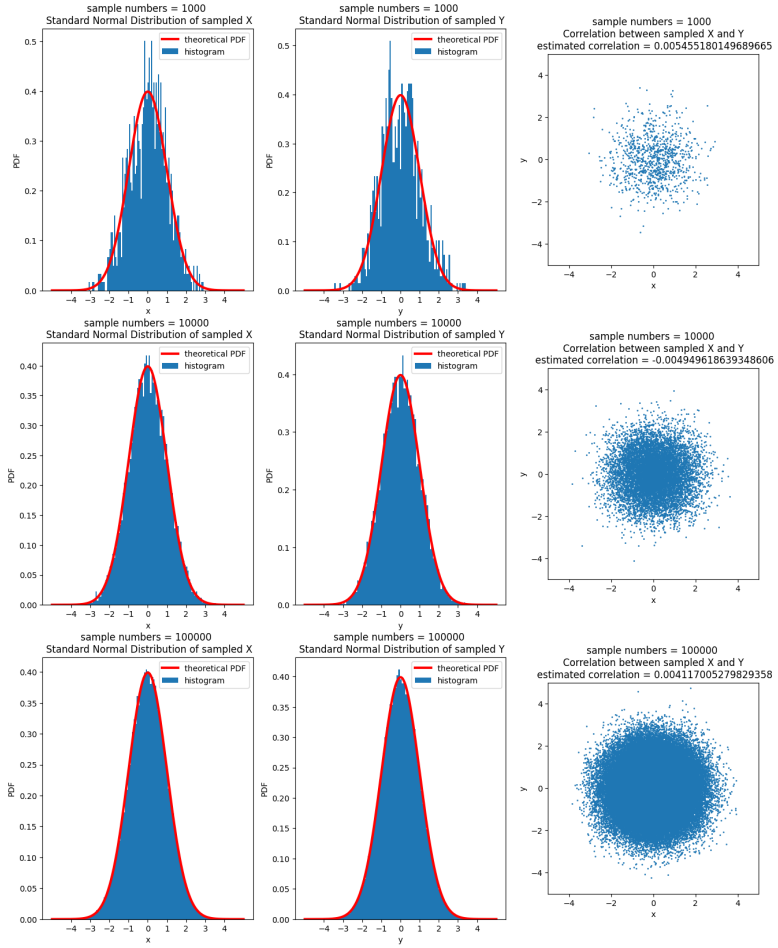

Figure 3: Sample $\mathcal{N}(0,1)$ with Box-Muller method

(b) Similarly with (a), we want to sample on $\mathcal{N}(0,1)$ with acceptance-rejection algorithm.
Let $Z \sim \mathcal{N}(0,1)$, and $X = |Z|$.
So $Z \in (-\infty, +\infty)$, $X \in (0, +\infty)$.
We can calculate the PDF of X:

$f(x) = f_X(x) = \frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, x > 0.$

And we can choose that $Y \sim Expo(1)$, and its PDF is $g(y) = f_Y(y) = e^{-y}, y > 0$.
Following the method in (a), we can get that $c \geq sup_y \frac{f(y)}{g(y)}$.

$\frac{f(y)}{g(y)} = \frac{\frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}}{e^{-y}} = \frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2+y}.$

Since $y > 0$, so we can get that when $y = 1$, $\frac{f(y)}{g(y)}$ will take the maximum value within the domain.

i.e. $sup_y \frac{f(y)}{g(y)} = \frac{f(1)}{g(1)} = \frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2}+1} = \sqrt{\frac{2e}{\pi}}.$

So we can just take $c = \sqrt{\frac{2e}{\pi}}$.

Then we can do the acceptance-rejection algorithm.
1. Generate $Y \sim Expo(1)$.
2. Generate $U \sim Unif(0, 1)$.
3. If $U \leq \frac{f(Y)}{c \cdot g(Y)} = \frac{\frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2}Y^2}}{\sqrt{\frac{2e}{\pi}} \cdot e^{-Y}} = \frac{1}{\sqrt{e}} e^{-\frac{1}{2}Y^2+Y} = e^{-\frac{1}{2}(Y-1)^2}$, then set $X = Y$
4. Else, go back to step 1.

After that, we can get the sample the distribution of X.
To sample on $Z \sim \mathcal{N}(0, 1)$, since $X = |Z|$, so we can generate $U' \sim Unif(0, 1)$,
and let $Z = \begin{cases} X, & U' \leq \frac{1}{2} \\ -X, & U' > \frac{1}{2} \end{cases}$.
After that, we can sample the distribution of $Z \sim \mathcal{N}(0, 1)$ with acceptance-rejection method.

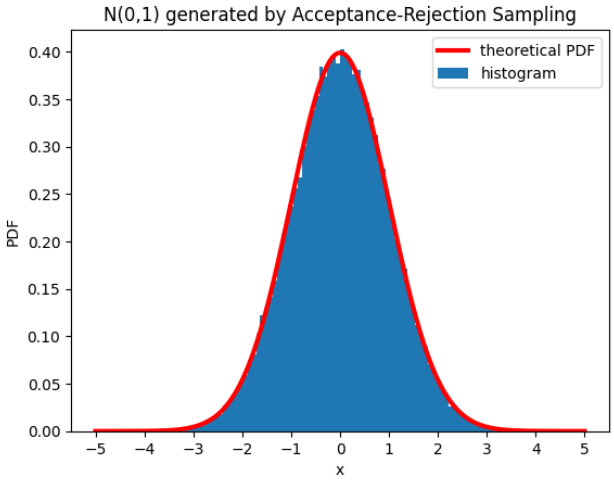

Figure 4: Sample $\mathcal{N}(0, 1)$ with Acceptance-Rejection method

Prove the correctness of the algorithm in theorem:
1. Let event $A =$ "$U \leq \frac{f(Y)}{c \cdot g(Y)}$".

From the decription of the algorithm, we know that we let $X = Y$ when A happens.

So the PDF of generated r.v.s. X is that $f_Y(y|A) = \dfrac{P(A|Y=y)}{P(A)} \cdot f_Y(y)$.

2. And $P(A|Y=y) = P(U \le \dfrac{f(Y)}{c \cdot g(Y)}|Y=y) = P(U\dfrac{f(y)}{c \cdot g(y)}|Y=y)$.

Since U and Y are independent, so $P(U \le \dfrac{f(y)}{c \cdot g(y)}|Y=y) = P(U \le \dfrac{f(y)}{c \cdot g(y)})$.

And since $U \sim Unif(0,1)$, so $P(U \le \dfrac{f(y)}{c \cdot g(y)}) = \dfrac{f(y)}{c \cdot g(y)}$.

So $P(A|Y=y) = \dfrac{f(y)}{c \cdot g(y)}$.

3. As for $P(A)$, with LOTP, we can get that

$P(A) = \int_0^{+\infty} P(A|Y=y)g(y)dy$.

From 2., we get that $P(A|Y=y) = \dfrac{f(y)}{c \cdot g(y)}$.

So $P(A) = \int_0^{+\infty} \dfrac{f(y)}{c \cdot g(y)} g(y)dy = \dfrac{1}{c}\int_0^{+\infty} f(y)dy = \dfrac{1}{c}$.

For the last step, this is because $g(y)$ is the PDF of $Expo(1)$, which support is $(0, +\infty)$, so $\int_0^{+\infty} g(y)dy = 1$.

4. Combine 2., 3. into 1., we can get that

$P(Y=y|A) = \dfrac{P(A|Y=y)}{P(A)} \cdot f_Y(y) = \dfrac{\frac{f(y)}{c \cdot g(y)}}{\frac{1}{c}} \cdot g(y) = f(y)$.

So from 1. to 4., we have prove that the PDF of generated r.v.s. X is that $f_Y(y|A) = f(y)$.

i.e. $X \sim f(y)$.

And since $X = |Z|$, so we generated a r.v.s. $U' \sim (0,1)$.

And let $Z = \begin{cases} X, & U' \le \frac{1}{2} \\ -X, & U' > \frac{1}{2} \end{cases}$.

This is because $Z \sim (0,1)$, so it is symmetric about $x = 0$.

And since $X = |Z|$, so it can be regard that X takes the value of Z with probability $\frac{1}{2}$, and takes the value of $-Z$ with probability $\frac{1}{2}$.

So we can generate X with $U' \sim (0,1)$, and let $X = Z$ when $U' \le \frac{1}{2}$, and let $X = -Z$ when $U' > \frac{1}{2}$.

So above all, we can generate $Z \sim \mathcal{N}(0,1)$ with acceptance-rejection method. The correctness have been proved.

(c) Both Acceptance-Rejection method and Box-Muller method can be used to sample on $\mathcal{N}(0,1)$. And each of them have their own pros and cons

As for the Acceptance-Rejection method:

Pros:

1. It is easy to implement.
2. It can be used to sample on any distribution, we do not have to need the exact expression of the distribution that we want to sample, we only need to know the relativeness.
3. It can sample on any distribution, not only $\mathcal{N}(0,1)$.

Cons:

1. It may be difficult to find a suitable constant c.
2. It may be slower to sample because for each time, the probability of acceptance $\sim FS(p)$, where $p = \dfrac{1}{c}$, so the expected sample time is $\dfrac{1}{p} = \dfrac{1}{\frac{1}{c}} = c$. So it may takes c times to get one sample compared with Box-Muller method.

3. The quantity of the sample points is not fixed, it depends on the constant c. If c is too big, then it may takes a lot of time to get one sample, and if it is too small, it may loss the accuracy.

As for the Box-Muller method: Pros:

1. It is easy to implement, fast, accurate
2. It is easy to generate two independent $\mathcal{N}(0,1)$ r.v.s.

Cons:

1. It can only be used to sample on $\mathcal{N}(0,1)$.
2. It requires the exact expression of the distribution that we want to sample.
3. It requires the trigonometric function and exponential functon, which is advanced, and may be difficult to implement.

So above all, Acceptance-Rejection method can be used to sample on various distributions, but it may be slow and difficult to find a suitable constant c.
Box-Muller method can only be used to sample on $\mathcal{N}(0,1)$, but it is fast and accurate.

(b) We select $Y \sim \mathcal{N}(0,1)$, which means that $g(y) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$.
Suppose that the normalization factor of $f(x)$ is Z, which means that $f(x) = Z \exp\left(-\frac{1}{2}x^2\right)\left(\sin^2(6x) + 3\cos^2(x)\sin^2(4x) + 1\right)$.
Then we can get that

$$c \geq \sup_y \frac{f(y)}{g(y)} = \frac{5\sqrt{2\pi}}{Z}$$

When $U \leq \dfrac{f(Y)}{c \cdot g(Y)} = \dfrac{\sin^2(6Y) + 3\cos^2(Y)\sin^2(4Y) + 1}{5}$, we accept $X \leftarrow Y$.
The sampled result is shown in the following figure.

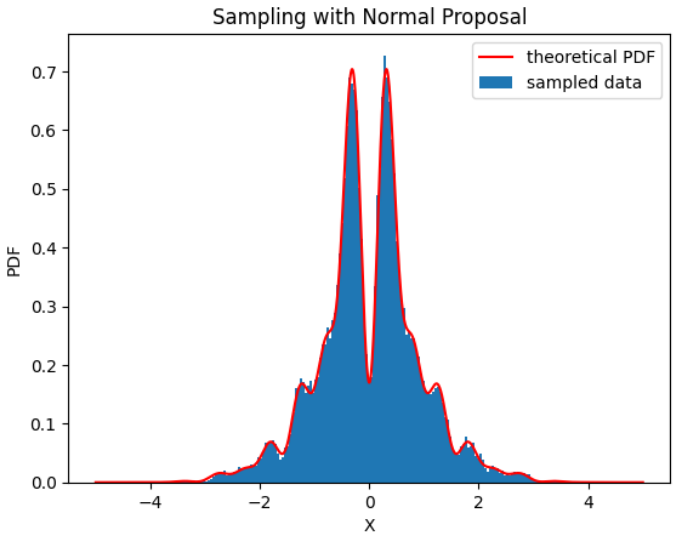

Figure 5: Sample $\mathcal{N}(0,1)$ with Acceptance-Rejection method

Problem 6

Given a random variable $X \sim \mathcal{N}(0,1)$, evaluate the tail probability $c = P(X > 8)$ by Monte Carlo methods with & without importance sampling. Discuss the pros and cons of importance sampling.

Solution

1. Without importance sampling.
The total 10^9 samples from $\mathcal{N}(0,1)$ have no single sample greater than 8.
So the tail probability estimatied is $c = P(X > 8) = 0$.

2. With importance sampling.
To calculate $c = P(Y > 8)$, where $Y \sim \mathcal{N}(0,1)$.
So the PDF of Y is that $f(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$.
It may be difficult to calculate the exact value of c using simple sampling methods (because of the 3σ's principle, the result must be very small).
So we can use the importance sampling.
Take $g \sim \mathcal{N}(8,1)$.
Let $Y_1, \cdots, Y_N \sim g$, so the PDF of g is that $g(y_j) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_j-8)^2}$.
Let $h(Y_j)$ be the indicator that whether $Y_j > 8$.

So with monty carlo method, we can get that

$$c = P(Y > 8) = \mathbb{E}[\mathbb{1}(Y > 8)] = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}(Y_j' > 8) = \frac{1}{n}\sum_{j=1}^{n} h(Y_j')$$

where $Y_j' \sim \mathcal{N}(0,1)$.

With importance sampling, since

$$I = \mathbb{E}_f[h(Y)] = \int h(y)f(y)dy = \int \frac{h(y)f(y)}{g(y)} g(y)dy = E_g\left[\frac{h(Y)f(Y)}{g(Y)}\right]$$

So $I' = \frac{1}{n}\sum_{i=1}^{n} \frac{h(Y_j)f(Y_j)}{g(Y_j)}$, where $Y_j \sim \mathcal{N}(8,1)$.
we can get that

$$\begin{aligned} c' &= \frac{1}{n}\sum_{j=1}^{n} \frac{h(Y_j)f(Y_j)}{g(Y_j)} \\ &= \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}(Y_j > 8) \cdot \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Y_j^2}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(Y_j-8)^2}} \\ &= \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}(Y_j > 8) \cdot e^{-8Y_j+32} \end{aligned}$$

So we just need to sample n times, $Y_j \sim \mathcal{N}(0,8)$, and calculate the average of $\mathbb{1}(Y_j > 8) \cdot e^{-8Y_j+32}$.

The result using important sampling method is $6.252836307280847e-16$.
Which is very close to the correct answer $6.25 * 10^{-16}$.
So we can regard that the importance sampling method is effective and provide correct answers.

Problem 6 continued on next page...

The following are the pages from jupyter notebook to show that the code can successful run out the images and calculation results we mentioned above.

3. The pros and cons of importance sampling.
Pros of Importance Sampling:

- It is efficient because it allows sampling from a distribution that is easier to sample from, which simplifies the sampling process.
- It improves accuracy for rare events, provides more accurate estimates when dealing with low-probability tail.

Cons of Importance Sampling:

- It depends on the choice of the proposal distribution. A poor choice can lead to high variance and inaccurate results.
- It requires more computations to calculate importance weights.

Problem 7

Generate uniform distributions over the following geometric objects:
(a) Ellipse $(a = 2, b = 1)$:

$$E_2(a,b) = \left\{(x,y) \in \mathbb{R}^2 : \left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 \leq 1\right\}$$

(b) Sphere $(r = 1)$:

$$S_2(r) = \left\{(x,y,z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = r^2\right\}$$

(c) Ball $(r = 1)$:

$$B_3(r) = \left\{(x,y,z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 \leq r^2\right\}$$

(d) Torus $(r_0 = 2, r = 1)$:

$$T_2\left(r_0, r\right) = \left\{(x,y,z) \in \mathbb{R}^3 : \left(r_0 - \sqrt{x^2+y^2}\right)^2 + z^2 = r^2\right\}$$

Solution

(a) 1. Acceptance-Rejection method:

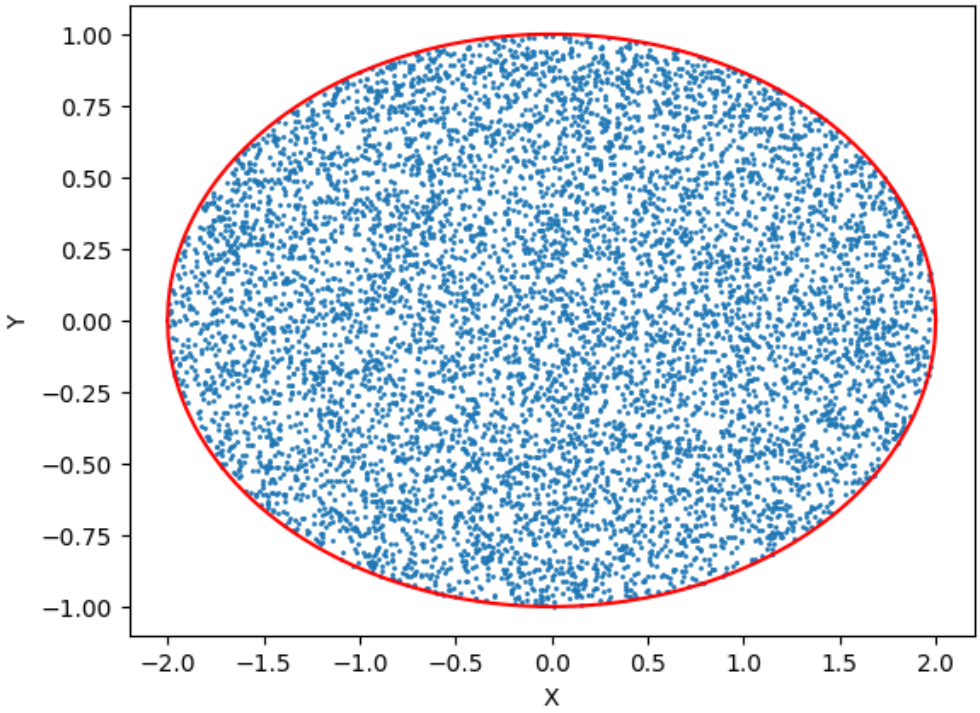

Figure 6: Ellipse sample by Acceptance-Rejection method

2. Change of variable:
Let $x = a \cdot r\cos\theta$, $y = b \cdot r\sin\theta$, where $\theta \in [0, 2\pi], r \in [0, 1]$. The Jacobian is $|J| = \left|\frac{\partial(x,y)}{\partial(r,\theta)}\right| = abr$.
Since the area of the ellipse is πab, so the PDF of the ellipse is $f_{X,Y}(x,y) = \frac{1}{\pi ab}$. Thus we have

$$f_{R,\Theta}(r,\theta) = f_{X,Y}(x,y) \cdot |J| = \frac{1}{\pi ab} \cdot abr = \frac{r}{\pi} = f_R(r) f_\Theta(\theta)$$

So R and Θ are independent variables, we can sample them respectively.

$$
\begin{aligned}
f_R(r) &= \int_0^{2\pi} f_{R,\Theta}(r,\theta)d\theta = 2r \\
F_R(r) &= \int_0^{r} f_R(r)dr = r^2 \\
F_R^{-1}(u) &= \sqrt{u} \\
f_\Theta(\theta) &= \int_0^{1} f_{R,\Theta}(r,\theta)dr = \frac{1}{2\pi} \\
F_\Theta(\theta) &= \int_0^{\theta} f_\Theta(\theta)d\theta = \frac{\theta}{2\pi} \\
F_\Theta^{-1}(u) &= 2\pi u
\end{aligned}
$$

$U_1, U_2 \overset{i.i.d.}{\sim} \text{Unif}(0,1)$, then let

$$r \leftarrow \sqrt{U_1}, \theta \leftarrow 2\pi U_2$$

Then (X, Y) can be uniformly sampled from the ellipse. $x = a \cdot r\cos\theta = 2\sqrt{U_1}\cos(2\pi U_2)$, $y = b \cdot r\sin\theta = \sqrt{U_1}\sin(2\pi U_2)$

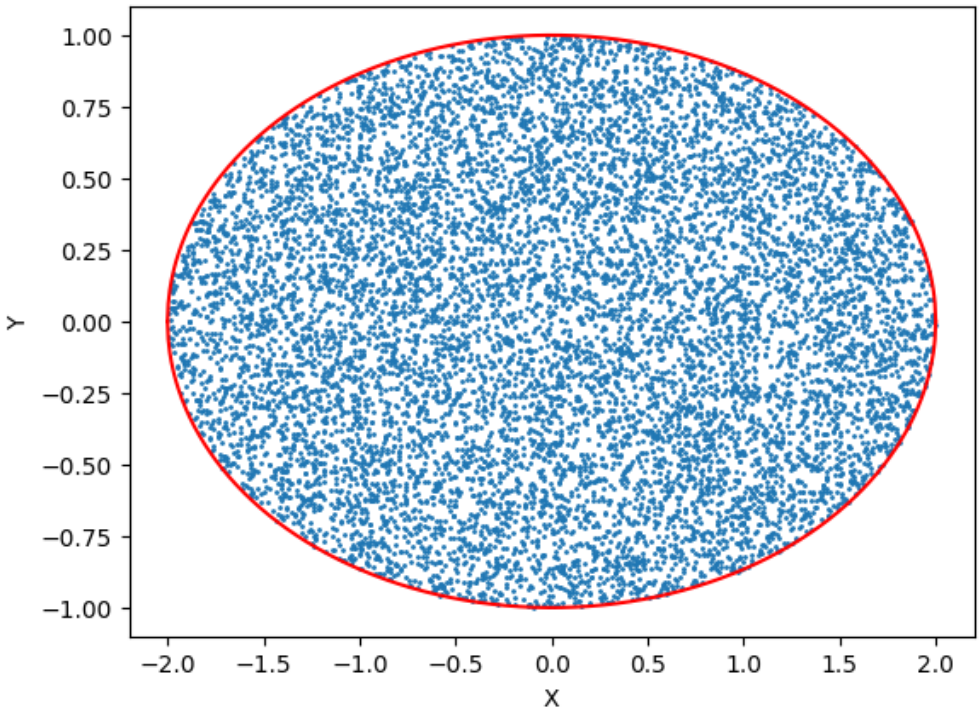

Figure 7: Ellipse sample by Change of variable

(b) Let $x = r\sin\theta\cos\phi$, $y = r\sin\theta\sin\phi$, $z = r\cos\theta$, where $\theta \in [0,\pi], \phi \in [0,2\pi]$. The Jacobian is $J = \frac{\partial(x,y,z)}{\partial(\theta,\phi)} = \begin{pmatrix} r\cos\theta\cos\phi & -r\sin\theta\sin\phi \\ r\cos\theta\sin\phi & r\sin\theta\cos\phi \\ -r\sin\theta & 0 \end{pmatrix}$ The Gram matrix $G = J^\top J = \begin{pmatrix} r^2 & 0 \\ 0 & r^2\sin^2\theta \end{pmatrix}$. The area of the sphere is $4\pi r^2$, so the PDF of the sphere is $f_{X,Y,Z}(x,y,z) = \frac{1}{4\pi r^2}$. So we have

$$f_{\Theta,\Phi}(\theta,\phi) = f_{X,Y,Z}(x,y,z) \cdot |\sqrt{\det(G)}| = \frac{1}{4\pi r^2} \cdot r^2\sin\theta = \frac{1}{4\pi}\sin\theta = f_\Theta(\theta)f_\Phi(\phi)$$

So Θ and Φ are independent variables, we can sample them respectively.

$$
\begin{aligned}
f_\Theta(\theta) &= \int_0^{2\pi} f_{\Theta,\Phi}(\theta,\phi)d\phi = \frac{1}{2}\sin\theta \\
F_\Theta(\theta) &= \int_0^{\theta} f_\Theta(\theta)d\theta = \frac{1-\cos\theta}{2} \\
F_\Theta^{-1}(u) &= \arccos(1-2u) \\
f_\Phi(\phi) &= \int_0^{\pi} f_{\Theta,\Phi}(\theta,\phi)d\theta = \frac{1}{2\pi} \\
F_\Phi(\phi) &= \int_0^{\phi} f_\Phi(\phi)d\phi = \frac{\phi}{2\pi} \\
F_\Phi^{-1}(u) &= 2\pi u
\end{aligned}
$$

Sample $U_1, U_2 \overset{i.i.d.}{\sim} \text{Unif}(0,1)$, then let

$$\theta \leftarrow \arccos(1-2U_1), \phi \leftarrow 2\pi U_2$$

Since $\cos\theta = 1-2U_1$, and $\theta \in [0,\pi]$, so $\sin\theta = 2\sqrt{U_1(1-U_1)}$. Then we set that

$x \leftarrow r\sin\theta\cos\phi = 2\sqrt{U_1(1-U_1)}\cos(2\pi U_2), y \leftarrow r\sin\theta\sin\phi = 2\sqrt{U_1(1-U_1)}\sin(2\pi U_2), z \leftarrow r\cos\theta = 1-2U_1$

Then (X,Y,Z) can be uniformly sampled from the sphere.

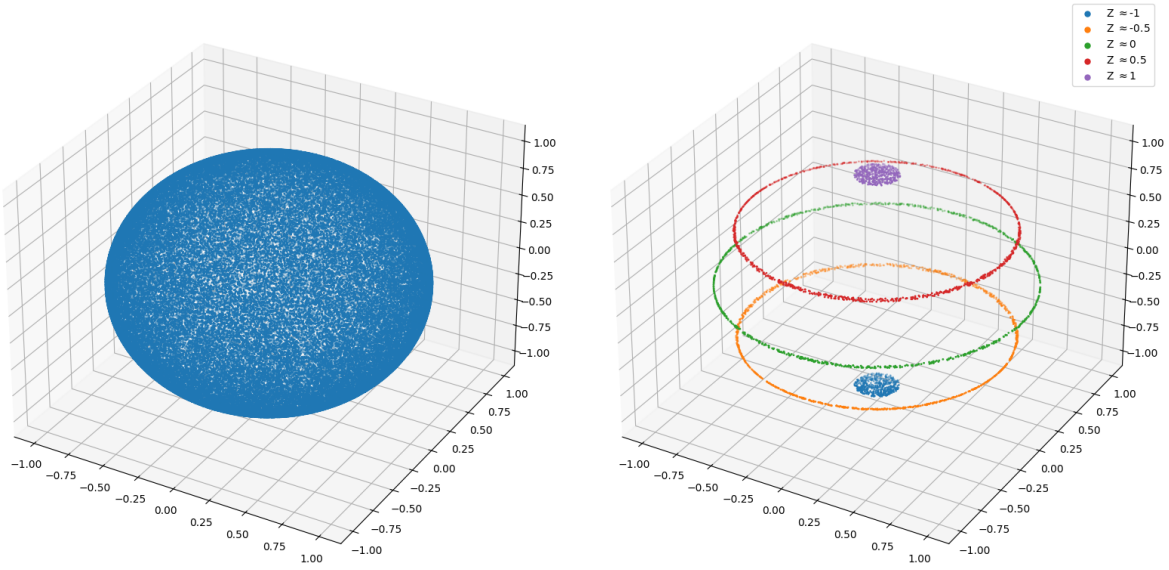

Figure 8: Left: Sphere's sample points. Right: Some contours of the Sphere to show the inner details.

(c) Let $x = r\sin\theta\cos\phi$, $y = r\sin\theta\sin\phi$, $z = r\cos\theta$, where $\theta \in [0,\pi], \phi \in [0,2\pi], r \in [0,1]$. The Jacobian is $J = \dfrac{\partial(x,y,z)}{\partial(r,\theta,\phi)} = r^2\sin\theta$.

Since the volume of the ball is $\dfrac{4}{3}\pi r^3 = \dfrac{4}{3}\pi$, so the PDF of the ball is $f_{X,Y,Z}(x,y,z) = \dfrac{3}{4\pi}$. Thus we have

$$f_{R,\Theta,\Phi}(r,\theta,\phi) = f_{X,Y,Z}(x,y,z)\cdot|J| = \frac{3}{4\pi}\cdot r^2\sin\theta = f_R(r)f_\Theta(\theta)f_\Phi(\phi)$$

So R, Θ and Φ are independent variables, we can sample them respectively.

$$
\begin{aligned}
f_R(r) &= \int_0^{2\pi}\int_0^{\pi} f_{R,\Theta,\Phi}(r,\theta,\phi)d\theta d\phi = 3r^2 \\
F_R(r) &= \int_0^{r} f_R(r)dr = r^3 \\
f_\Theta(\theta) &= \int_0^{2\pi}\int_0^{1} f_{R,\Theta,\Phi}(r,\theta,\phi)drd\phi = \frac{1}{2}\sin\theta \\
F_\Theta(\theta) &= \int_0^{\theta} f_\Theta(\theta)d\theta = \frac{1}{2}(1-\cos\theta) \\
f_\Phi(\phi) &= \int_0^{2\pi}\int_0^{1} f_{R,\Theta,\Phi}(r,\theta,\phi)drd\theta = \frac{1}{2\pi} \\
F_\Phi(\phi) &= \int_0^{\phi} f_\Phi(\phi)d\phi = \frac{\phi}{2\pi}
\end{aligned}
$$

$U_1, U_2, U_3 \overset{i.i.d.}{\sim} \mathrm{Unif}(0,1)$, then let

$$r \leftarrow \sqrt[3]{U_1}, \theta \leftarrow \arccos(1-2U_2), \phi \leftarrow 2\pi U_3$$

We can ree-write $\cos\theta = 1-2U_2$, and $\sin\theta = 2\sqrt{U_2(1-U_2)}$ to save computational complexity. Then (X,Y,Z) can be uniformly sampled from the ball.

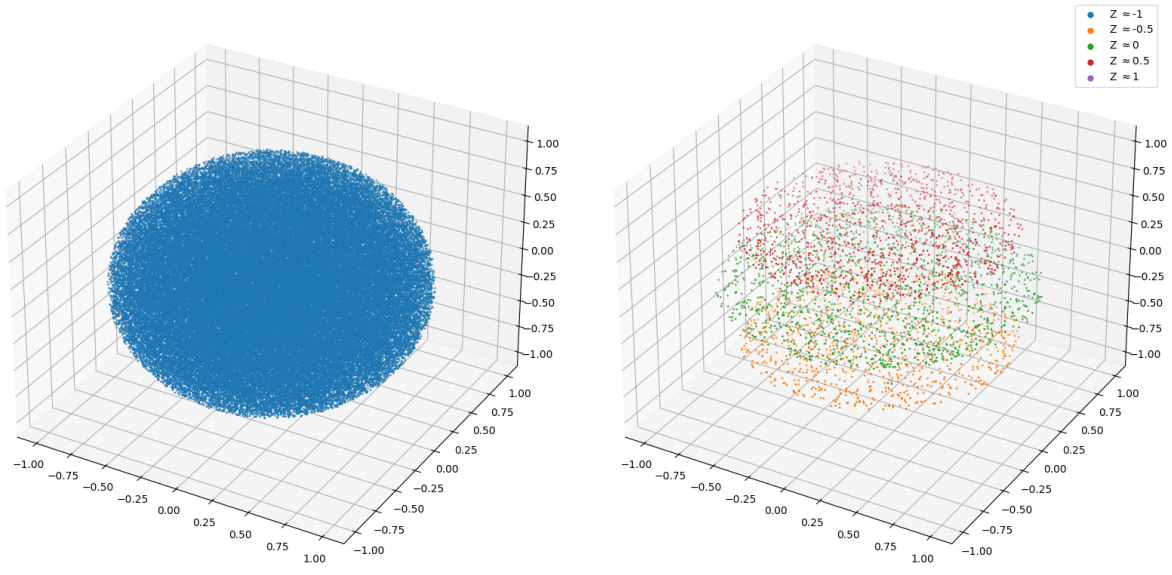

Figure 9: Left: Ball's sample points. Right: Some contours of the Ball to show the inner details.

(d) Let $x = (r_0 + r\cos\theta)\cos\phi, y = (r_0 + r\cos\theta)\sin\phi, z = r\sin\theta$, where $\theta \in [0,2\pi], \phi \in [0,2\pi], r \in [0,1]$. The Jacobian is $J = \dfrac{\partial(x,y,z)}{\partial(r,\theta,\phi)} = \begin{pmatrix} -r\sin\theta\cos\phi & -(r_0+r\cos\theta)\sin\phi \\ -r\sin\theta\sin\phi & (r_0+r\cos\theta)\cos\phi \\ r\cos\theta & 0 \end{pmatrix}$. So the Gram matrix $G = J^\top J = \begin{pmatrix} r^2 & 0 \\ 0 & (r_0+r\cos\theta)^2 \end{pmatrix}$. Since $r=1$, so $r\cos\theta > 0$

$$\mathrm{d}S = \sqrt{\det(G)}\mathrm{d}\theta\mathrm{d}\phi = r(r_0+r\cos\theta)\mathrm{d}r\mathrm{d}\theta\mathrm{d}\phi$$

$$
\begin{aligned}
S &= \oint \mathrm{d}S \\
&= \int_0^{2\pi}\int_0^{2\pi} r(r_0 + r\cos\theta)d\theta d\phi \\
&= 4\pi^2 r \cdot r_0 \\
&= 8\pi^2
\end{aligned}
$$

$f_{X,Y,Z}(x,y,z) = \dfrac{1}{4\pi^2 r \cdot r_0} = \dfrac{1}{8\pi^2}$, so

$$
f_{\Theta,\Phi}(\theta,\phi) = \frac{1}{4\pi^2 r_0}\cdot r(r_0 + r\cos\theta) = \frac{r_0 + r\cos\theta}{4\pi^2 r_0} = \frac{2+\cos\theta}{8\pi^2} = f_\Theta(\theta) f_\Phi(\phi)
$$

So Θ and Φ are independent variables, we can sample them respectively.

$$
\begin{aligned}
f_\Theta(\theta) &= \int_0^{2\pi} f_{\Theta,\Phi}(\theta,\phi)d\phi = \frac{r_0 + r\cos\theta}{2\pi r_0} = \frac{2+\cos\theta}{4\pi} \\
F_\Theta(\theta) &= \int_0^{\theta} f_\Theta(\theta)d\theta = \frac{1}{2\pi}\theta + \frac{r}{2\pi r_0}\sin\theta = \frac{1}{2\pi}\theta + \frac{1}{4\pi}\sin\theta \\
f_\Phi(\phi) &= \int_0^{2\pi} f_{\Theta,\Phi}(\theta,\phi)d\theta = \frac{1}{2\pi} \\
F_\Phi(\phi) &= \int_0^{\phi} f_\Phi(\phi)d\phi = \frac{\phi}{2\pi}
\end{aligned}
$$

$U_1, U_2 \overset{i.i.d.}{\sim} \mathrm{Unif}(0,1)$, then let θ be the solution that $\dfrac{1}{2\pi}\theta + \dfrac{1}{4\pi}\sin\theta = U_1$, $\phi \leftarrow 2\pi U_2$. θ has no closed form solution, but we can use the toolkits to solve the equation as the valid CDF must exist a solution.
Then (X,Y,Z) can be uniformly sampled from the Torus whose $r_0 = 2, r = 1$.

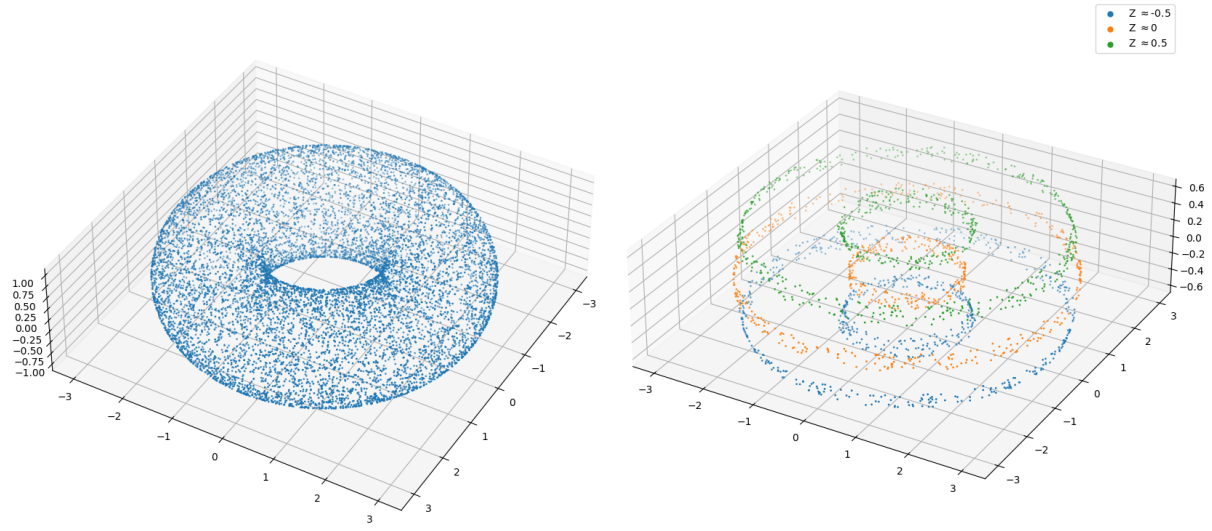

Figure 10: Left: Torus's sample points. Right: Some contours of the Torus to show the inner details.

Problem 8

The Curse and Blessing of Dimensionality. Denote $\boldsymbol{x} = (x_1, \ldots, x_d)$.
(a) The d-dimensional hyperball of radius r is denoted as

$$B_d(r) = \left\{ \boldsymbol{x} \in \mathbb{R}^d : \sum_{i=1}^{d} x_i^2 \leq r^2 \right\}$$

Find the volume of $B_d(r)$ and plot a figure to show how such volume changes with d when $r = 1$.
(b) The d-dimensional hypersphere of radius r is denoted as

$$S_{d-1}(r) = \left\{ \boldsymbol{x} \in \mathbb{R}^d : \sum_{i=1}^{d} x_i^2 = r^2 \right\}$$

When $d \gg 1$, adopt concentration inequalities to show almost all the volume of the high-dimensional hypersphere lies near its equator.
(c) The d-dimensional hypercube of radius r is denoted as

$$C_d(r)(r) = \left\{ \boldsymbol{x} \in \mathbb{R}^d : -r \leq x_i \leq r \right\}$$

When $d \gg 1$, adopt concentration inequalities to show almost all the volume of the high-dimensional cube is located in its corners.

Solution

(a) Suppose that the d-dimensional hyperball with radius r's volume is $V_d(r)$. We can get that

$$\begin{aligned} V_d(r) &= C_d(r) \cdot r^d \\ &= \int_{-r}^{r} C_{d-1}(r) \left(\sqrt{r^2 - x^2} \right)^{d-1} \mathrm{d}x \\ &= 2C_{d-1}(r) \int_{0}^{r} \left(\sqrt{r^2 - x^2} \right)^{d-1} \mathrm{d}x \\ &= 2C_{d-1}(r) \int_{0}^{\pi/2} r^d (\cos\theta)^d d\theta \end{aligned}$$

And we can get that

$$C_d(r) = 2C_{d-1}(r) \cdot \frac{(d-1)!!}{d!!} \cdot \frac{\pi}{2} \quad \text{if } d \text{ is even}$$

$$C_d(r) = 2C_{d-1}(r) \cdot \frac{(d-1)!!}{d!!} \quad \text{if } d \text{ is odd}$$

where $n!!$ is the double factorial of n, which is defined as $n!! = n \cdot (n-2) \cdot (n-4) \cdots 1$ if n is odd, and $n!! = n \cdot (n-2) \cdot (n-4) \cdots 2$ if n is even. We can get that

$$C_d(r) = \begin{cases} 2 \cdot \frac{2\pi}{3} \cdot \frac{2\pi}{5} \cdot \cdots \cdot \frac{2\pi}{d} & \text{if } d \text{ is odd} \\ \pi \cdot \frac{2\pi}{4} \cdot \frac{2\pi}{6} \cdot \cdots \cdot \frac{2\pi}{d} & \text{if } d \text{ is even} \end{cases} = \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2} + 1\right)}$$

So above all,

$$V_d(r) = \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2} + 1\right)} \cdot r^d$$

The voulme of the unit hyperball changes via dimension d is shown in the following figure, we could discover that when $d = 5$, the volume of the unit hyperball reaches its maximum value, and then it decreases as d increases.

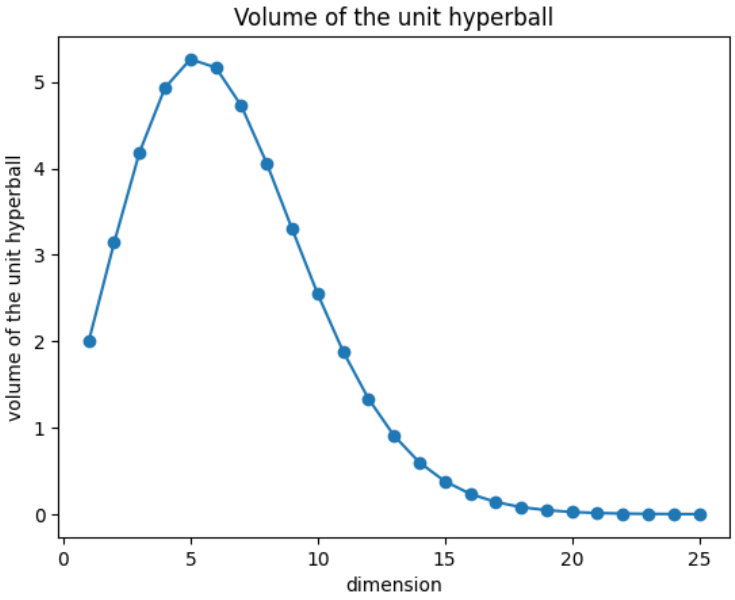

Figure 11: volume of the unit hyperball

(b) Consider uniformly sampling from the hypersphere. i.e. $\boldsymbol{x} = (x_1, \ldots, x_d)$, by symmetry, we can get that $x_1, \ldots, x_d$ has the same distribution. So $\mathbb{E}(x_1) = \ldots = \mathbb{E}(x_d) = 0$. Since $\boldsymbol{x}$ samples from hypersphere, so

$$\begin{aligned} x_1^2 + x_2^2 + \ldots + x_d^2 &= r^2 \\ \mathbb{E}\left(x_1^2 + \ldots + x_d^2\right) &= \mathbb{E}(r^2) = r^2 \\ \mathrm{Var}(x_d) = \mathbb{E}(x_d^2) &= \frac{r^2}{d} \end{aligned}$$

Let $\epsilon \ll r$, let the equator to be if $\boldsymbol{x} = \left\{(x_1, \ldots, x_d) \in S_{d-1}(r) \middle| |x_d| < \epsilon\right\}$. According to Chebyshev's Inequality:

$$\begin{aligned} P(|x_d - \mathbb{E}(x_d)| \geq \epsilon) &\leq \frac{\mathrm{Var}(x_d)}{\epsilon^2} \\ P(|x_d| \geq \epsilon) &\leq \frac{r^2}{d\epsilon^2} \end{aligned}$$

When d is very large, $\frac{r^2}{d\epsilon^2}$ becomes extremely small. Thus, it becomes highly unlikely that the sampled data will be a point where $|x_d| \geq \epsilon$. In other words, the sampled data is almost certainly located near the equator. Since the sampling is uniform over the hypersphere, we can conclude that nearly all the volume of a high-dimensional hypersphere is concentrated near its equator.

When the dimension $d \gg 1$, $\frac{r^2}{d\epsilon^2} \to 0$, which means it's very unlikely to pick a point where $|x_d| \geq \epsilon$. i.e. most points end up being near the equator. And since the points are sampled uniformly, it tells us that almost all of the hypersphere's volume in high dimensions is concentrated near its equator.

(c) The volume of the cube is $(2r)^d$. Let $x_1, x_2, \ldots, x_d \overset{i.i.d.}{\sim} \mathrm{Unif}(-r, r)$. So $\mathrm{Var}(x_i) = \frac{r^2}{3}$.
With the Chebyshev's Inequality, we can get that $\forall t > 0$:

$$P\left(\sum_{i=1}^{d} x_i^2 \leq r^2\right) \leq e^{tr^2} \left(\mathbb{E}(e^{-tx_1^2})\right)^d$$

Take $t = 1$, we have:

$$\mathbb{E}\left(e^{-x_1^2}\right) = \int_{-r}^{r} \frac{1}{2r} e^{-x_1^2} \mathrm{d}x_1 < \int_{-r}^{r} \frac{1}{2r} \mathrm{d}x_1 = 1$$

Thus, $\mathbb{E}^d\left(e^{-x_1^2}\right) \to 0$ as d increases, i.e.

$$\lim_{d\to\infty} P\left(\sum_{i=1}^{d} x_i^2 \leq r^2\right) = 0$$

It is nearly impossible to sample a point from the hyperball insiede the hypercube, nearly all of the volume of the high-dimensional cube is concentrated near its corners.

Problem 9

Uniform is the only distribution that takes the maximum entropy.

Solution

Consider the diecrete case $|\mathcal{X}| < \infty$.

1. Consider the KL-divergence $D\left(p(x)\|q(x)\right) \geq 0$.
proof:

$$\begin{aligned}
-D\left(p(x)\|q(x)\right) &= \sum_x p(x) \log \frac{q(x)}{p(x)} \\
&= \mathbb{E}_{x\sim p(x)}\left[\log \frac{q(x)}{p(x)}\right] \\
&\leq \log \mathbb{E}_{x\sim p(x)}\left[\frac{q(x)}{p(x)}\right] \qquad \text{(Jensen's Inequality)} \\
&= \log \sum_x p(x) \frac{q(x)}{p(x)} \\
&= 0
\end{aligned}$$

i.e. $D\left(p(x)\|q(x)\right) \geq 0$.
If and only if when $p(x) = q(x)$, the equality holds, this is because Jensen's Inequality holds when the function is linear.
2. Consider the entropy $H(X) = \mathbb{E}\left[-\log p(x)\right] = \sum_{x\in\mathcal{X}} p(x) \log \frac{1}{p(x)}$.
Let $u(x) = \frac{1}{|\mathcal{X}|}$, then

$$\begin{aligned}
D\left(p\|u\right) &= \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{u(x)} \\
&= \sum_{x\in\mathcal{X}} p(x) \log |\mathcal{X}| - \sum_{x\in\mathcal{X}} p(x) \log \frac{1}{p(x)} \\
&= \log |\mathcal{X}| - H(X) \\
&\geq 0
\end{aligned}$$

If and only if $p(x) = u(x)$, the equality holds, i.e. $H(X)$ takes the maximum value $\log |\mathcal{X}|$.
So above all, the uniform distribution is the only distribution that takes the maximum entropy if $|\mathcal{X}| < \infty$.