APMA1690: Homework # 4 (Due by 11pm Oct 19)

1 Review

Please read the review section before delving into the problem set.

1.1 Random Variables

Let $\boldsymbol{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(d)})$ be a $\mathbb{R}^d$-valued random variable defined on the probability space $(\Omega, \mathbb{P})$, i.e.,

$$\begin{aligned} \boldsymbol{X}: \ & \Omega \to \mathbb{R}^d, \\ & \omega \mapsto \boldsymbol{X}(\omega) = \Big(X^{(1)}(\omega), X^{(2)}(\omega), \ldots, X^{(d)}(\omega)\Big), \end{aligned}$$

where each $X^{(i)}$ is a $\mathbb{R}^1$-valued random variable. When $d > 1$, $\boldsymbol{X}$ is also referred to as a "random vector."

Suppose H is a d-variable function which takes values in $\mathbb{R}$, i.e.,

$$\begin{aligned} H: \ & \mathbb{R}^d \to \mathbb{R}, \\ & \boldsymbol{x} = (x_1, x_2, \ldots, x_d) \mapsto H(\boldsymbol{x}) = H(x_1, x_2, \ldots, x_d). \end{aligned}$$

Then, we have the $\mathbb{R}^1$-valued random variable $H(\boldsymbol{X})$ defined as follows

$$\begin{aligned} H(\boldsymbol{X}): \ & \Omega \to \mathbb{R}^1, \\ & \omega \mapsto H\left(\boldsymbol{X}(\omega)\right) = H\Big(X^{(1)}(\omega), X^{(2)}(\omega), \ldots, X^{(d)}(\omega)\Big). \end{aligned}$$

1.2 Setup

Suppose our goal is to compute the following multiple integral[1]

$$v = \int H(\boldsymbol{x})\, d\boldsymbol{x} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} H(x_1, x_2, \ldots, x_d)\, dx_1 dx_2 \cdots dx_d,$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$ and $d\boldsymbol{x} = dx_1 dx_2 \cdots dx_d$. This integral can be represented as follows

$$\boxed{v = \int H(\boldsymbol{x})\, d\boldsymbol{x} = \int \frac{H(\boldsymbol{x})}{f(\boldsymbol{x})} \cdot f(\boldsymbol{x})\, d\boldsymbol{x} = \mathbb{E}\left[\frac{H(\boldsymbol{X})}{f(\boldsymbol{X})}\right],}$$

where $f(\boldsymbol{x})$ is a d-dimensional PDF, and the random vector $\boldsymbol{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(d)}) \sim f(\boldsymbol{x})$. Additionally, the PDF $f(\boldsymbol{x})$ satisfies the following conditions

[1] We assume that all means and variances utilized herein do exist.

1. $\{\boldsymbol{x} \in \mathbb{R}^d \,|\, H(\boldsymbol{x}) \neq 0\} \subset \{\boldsymbol{x} \in \mathbb{R}^d \,|\, f(\boldsymbol{x}) \neq 0\}$;

2. We know how to generate random vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \overset{iid}{\sim} f(\boldsymbol{x})$;

3. $f(\boldsymbol{x})$ is similar to the "optimal" PDF $\frac{1}{\int H(\boldsymbol{x}')d\boldsymbol{x}'} \cdot H(\boldsymbol{x})$. This similarity makes $\operatorname{Var}\left(\frac{H(\boldsymbol{X}_1)}{f(\boldsymbol{X}_1)}\right)$ small. (Since the integral $v = \int H(\boldsymbol{x}')d\boldsymbol{x}'$ is unavailable at this point, the optimal PDF is not achievable.)

1.3 Importance Sampling

We generate random vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \overset{iid}{\sim} f(\boldsymbol{x})$ and compute the following estimator of v

$$\boxed{\widehat{v}_n = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{H(\boldsymbol{X}_i)}{f(\boldsymbol{X}_i)}\right].} \tag{1.1}$$

Then, we have

1. Law of large numbers $\implies \widehat{v}_n \approx v$ when the sample size n is sufficiently large;

2. Law of the iterated logarithm $\implies |\widehat{v}_n - v| \leq \sqrt{\operatorname{Var}\left(\frac{H(\boldsymbol{X}_1)}{f(\boldsymbol{X}_1)}\right)} \cdot \sqrt{\frac{2\log(\log n)}{n}}$, where "$\leq$" holds in an approximate way.

A good reference for importance sampling is Chapter 7 of Wang (2012).

1.4 Markov Chains

Roughly speaking, a Markov chain is a sequence of random variables $\{X_n\}_{n=0}^{\infty}$ satisfying the Markov property.[2] In APMA 1690, we assume that all random variables take values in a generic countable set $\mathcal{X}$, i.e., $\mathcal{X}$ can be expressed as $\mathcal{X} = \{\xi_0, \xi_1, \ldots, \xi_n, \ldots\}$. The countability assumption of $\mathcal{X}$ heavily simplifies the theory of Markov chains. The following is the definition[3] of Markov chains.

Definition 1.1. • *A sequence of random variables $\{X_n\}_{n=0}^{\infty}$ taking values in $\mathcal{X}$ is called a* ***Markov chain*** *if this sequence satisfies*

$$\boxed{\mathbb{P}(X_{n+1} = y \,|\, X_n = x, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = \mathbb{P}(X_{n+1} = y \,|\, X_n = x)} \tag{1.2}$$

for all $n = 0, 1, \ldots,$ all $y \in \mathcal{X}$, and all the $x_0, x_1, \ldots, x_{n-1}, x \in \mathcal{X}$ such that $\mathbb{P}(X_n = x, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) > 0$. The property in Eq. (1.2) is called the ***Markov property****.*

[2] For Markov chains in this course, we always let the index n go from 0 instead of 1. It is just a convention.

[3] The definition herein works only for the scenario where a Markov chain takes values in a countable space $\mathcal{X} = \{\xi_0, \xi_1, \ldots, \xi_n, \ldots\}$. It is one of the reasons that we assume $\mathcal{X}$ is countable. For the general definition of Markov chains and the relevant details, see Definition 17.1 and Remark 17.2 of Klenke (2020). Since the materials of general Markov chains involve too much real analysis knowledge (see APMA 2110), we skip the general Markov chains in this course.

- *Furthermore, if there exists a bivariate function* $p : \mathcal{X} \times \mathcal{X} \to [0, 1]$ *such that*

$$\boxed{p(x, y) = \mathbb{P}(X_{n+1} = y \,|\, X_n = x) \qquad \textit{(this function does not depend on } n\textit{)},}$$

  for all $n = 0, 1, \ldots$, *then* $\{X_n\}_{n=0}^{\infty}$ *is called a* ***homogeneous*** *Markov chain, and the function* $p(\cdot, \cdot)$ *is called the* ***transition probability*** *of this Markov chain.*

- *In literature,* $\mathcal{X}$ *is usually referred to as the* ***state space*** *of the Markov chain* $\{X_n\}_{n=0}^{\infty}$; *each element in* $\mathcal{X}$ *is referred to as a* ***state****.*

Throughout this course, all the Markov chains will be homogeneous. Hence, **we will omit the word "homogeneous" hereafter**. In Eq. (1.2), if we call X_n as "the present," X_{n+1} as "the future," and $X_{n-1}, \ldots, X_0$ as "the past," the Markov property means, "given the present, the future does not depend on the past" — the condition $X_{n-1} = x_{n-1}, \ldots, X_0 = x_0$ in Eq. (1.2) does not play any role!

The value of $p(x, y)$ is the probability of "transiting from x to y," so it is called a transition probability. For each fixed $x \in \mathcal{X}$, the univariate function $p(x, \cdot) : y \mapsto p(x, y)$ is a PMF.

2 Problem Set

1. The unit ball in d-dimensional space is defined as follows

$$\mathbf{B}^d = \left\{ (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d \,\middle|\, x_1^2 + x_2^2 + \cdots + x_d^2 < 1 \right\}.$$

The boundary of this unit ball is the following $(d-1)$-dimensional sphere

$$\mathbb{S}^{d-1} = \left\{ (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d \,\middle|\, x_1^2 + x_2^2 + \cdots + x_d^2 = 1 \right\}.$$

(1 point) Please explain (not rigorously prove) the claim, "***when the dimension*** d ***is large, most of the volume of the unit ball*** $\mathbf{B}^d$ ***is concentrated near the sphere*** $\mathbb{S}^{d-1}$***.***"

As d gets large, each individual x_i^2 must get smaller for their sum to be less than 1. This means, however, that for large d, the volume added by each successive x_i gets smaller and smaller ($x_{10000} \approx x_{10001}$) as the sum tends to 1.

Heuristically, one can imagine the 3-sphere which has volume $\frac{4}{3}\pi r^3$. Small increases in radius lead to very large increases in volume because of the cubic factor; elements towards the edge of the boundary contribute more to the total volume.

2. Suppose we are interested in the following multiple integral

$$v_d = \int H(\boldsymbol{x})d\,\boldsymbol{x}, \quad \text{where } H(\boldsymbol{x}) = \mathbb{1}_{\{x_1^2+x_2^2+\cdots+x_d^2<1\}},$$

and we want to estimate the v_d using the importance sampling approach. To do so, we need to choose a d-variable PDF which is simiar to $\frac{1}{v_d} \cdot \mathbb{1}_{\{x_1^2+x_2^2+\cdots+x_d^2<1\}}$.

We restrict our attention to the following collection of multivariable normal PDFs indexed by $\sigma > 0$

$$f_\sigma(\boldsymbol{x}) = (2\pi\sigma^2)^{-\frac{d}{2}} \cdot \exp\left\{-\frac{x_1^2+\cdots+x_d^2}{2\sigma^2}\right\},$$

and we need to choose a proper parameter σ^* such that $f_{\sigma^*}(\boldsymbol{x})$ is similar to $\frac{1}{v_d}\cdot\mathbb{1}_{\{x_1^2+x_2^2+\cdots+x_d^2<1\}}$ in some way.

Because of the claim, "*when the dimension d is large, most of the volume of the unit ball $\mathbf{B}^d$ is concentrated near the sphere $\mathbb{S}^{d-1}$*," we choose a parameter σ^* such that

$$\begin{aligned} &\mathbb{E}\left[\left(X^{(1)}\right)^2 + \left(X^{(2)}\right)^2 + \cdots + \left(X^{(d)}\right)^2\right] = 1, \\ &\text{where } \boldsymbol{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(d)}) \sim f_{\sigma^*}(\boldsymbol{x}). \end{aligned} \tag{2.1}$$

That is, σ^* ensures that the average mass of the PDF $f_{\sigma^*}(\boldsymbol{x})$ lies in the region near the sphere $\mathbb{S}^{d-1}$. In this manner, sample points from $f_{\sigma^*}(\boldsymbol{x})$ concentrate in the region of importance, i.e., the vicinity of the unit sphere $\mathbb{S}^{d-1}$.

Questions:

(a) (3 points) Show that the "good choice" σ^* satisfying Equation (2.1) is

$$\sigma^* = \frac{1}{\sqrt{d}}.$$

(Hint: You may consider using the Chi-squared distribution.)

We seek to find a $\sigma^* > 0$ for which random variables $X_1, \ldots, X_n$ sampled from

$$f_{\sigma^*}(x) = (2\pi(\sigma^*)^2)^{-\frac{d}{2}} \cdot \exp\left(-\frac{1}{2(\sigma^*)^2}\sum_{i=1}^d x_i^2\right)$$

satisfy

$$\mathbb{E}[\sum_{i=1}^d X_i^2] = 1$$

We immediately notice that the RV $\vec{X} = (X_1, X_2, \ldots, X_n)$ is drawn from a normal distribution and we can define a new random variable by

$$Q = \sum_{i=1}^d X_i^d$$

If $Q \sim \chi^2$, then $\mathbb{E}Q = 1$ which is precisely the condition we would like f_{σ^*} to have.
When does a sum of squares follow a chi-squared distribution? Precisely when the random variables $X_1, X_2, \ldots, X_n$ are independent standard normal random variables. For the multivariate normal distribution, being "standard" means that each random variable has zero mean and unit variance.
That is, with

$$f(\vec{x}) = \frac{1}{\sqrt{(2\pi)^k \det \Sigma}} \cdot \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1}(\vec{x}-\vec{\mu})\right)$$

we have $\vec{\mu} = \vec{0}$ and $\Sigma_{ij} = c$, i.e.,

$$f(\vec{x}) = \frac{1}{\sqrt{(2\pi)^k}} \cdot \exp\left(-\frac{1}{2}x^T cIx\right) = \frac{1}{\sqrt{(2\pi)^k}} \cdot \exp\left(-\frac{c}{2}\sum_{i=1}^{k} x_i^2\right)$$

Now looking back at $f_{\sigma^*}(x)$, we have

$$f_{\sigma^*}(x) = \frac{1}{\sqrt{(2\pi(\sigma^*)^2)^d}} \cdot \exp\left(-\frac{1}{2(\sigma^*)^2}\sum_{i=1}^{d} x_i^2\right)$$

Setting the two equal, we can solve for σ^*:

$$\frac{c}{\sqrt{(2\pi)^k}} \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{k} x_i^2\right) = \frac{1}{\sqrt{(2\pi(\sigma^*)^2)^d}} \cdot \exp\left(-\frac{1}{2(\sigma^*)^2}\sum_{i=1}^{d} x_i^2\right)$$

So we let $c = d$ and $\sigma^* = \frac{1}{\sqrt{d}}$. ■

(b) (2 points) Compute v_{100} (i.e., the volume of the unit ball in 100-dimensional space) using the importance sampling method (see Equation (1.1)) and the multivariable normal PDF $f_{\sigma^*}(\boldsymbol{x})$ satisfying Equation (2.1). Provide your estimated value and the code for generating the value. You may use the sample size $n = 100000$. (Please feel free to use your preferred programming languages. You can also use any built-in functions of these programming languages that generate random numbers/vectors.)

Unfortunately, Python (even with extra libraries) is not very good at ultra-high precision floating point math. My code generated a volume estimate of

$$2.3911226283950689723312848 44e-40$$

which is quite far from the true value of

$$1.0329397732669577e-64$$

It does, however, correctly approximate the volume of a 2-ball and 3-ball, leading me to conclude that the problem indeed lies with Python and not my own implementation.

Homework > HW 4 > HW4.py > ...

```python
import numpy as np
import scipy
import math
from decimal import Decimal

def H(vec_x):
    sum_squares = sum(list(map(lambda x: x**2, vec_x)))
    if sum_squares < 1:
        return 1
    else:
        return 0

def f(vec_x):
    d = len(vec_x)
    sum_squares = sum(list(map(lambda x: x**2, vec_x)))

    return Decimal((2*math.pi/d)**(-d/2))*Decimal(-sum_squares/(2/d)).exp()

def volume(dim, n):
    summation = 0
    for i in range(1, n + 1):
        vec_x = scipy.stats.multivariate_normal.rvs(mean=None, cov=(1/dim), size=dim)
        summation += Decimal(H(vec_x))/Decimal(f(vec_x))

        if i % 1000 == 0:
            print(f"On iteration {i}/{n}\r", end="")

    return Decimal(summation)/Decimal(n)


print(f"Estimated volume: {volume(100, 100000)}")
```

3. (4 points) Let $\{X_n\}_{n=0}^{\infty}$ be a homogeneous Markov chain with a discrete state space $\mathcal{X}$. Let μ denote the PMF of X_0, i.e., $\mu(x) = \mathbb{P}(X_0 = x)$, for all $x \in \mathcal{X}$; furthermore, $p(x, y)$ denotes the transition probability of the Markov chain, i.e.,

$$p(x, y) = \mathbb{P}(X_{n+1} = y \,|\, X_n = x) = \mathbb{P}(X_1 = y \,|\, X_0 = x), \quad \text{for all } x, y \in \mathcal{X}.$$

Prove the following identity

$$\mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = \mu(x_0) \cdot p(x_0, x_1) \cdot p(x_1, x_2) \ldots p(x_{n-1}, x_n),$$

for all $n = 1, 2, \ldots$ and $x_0, x_1, \ldots, x_n \in \mathcal{X}$. (Hint: use the law of total probability/definition of conditional probability, and the Markov property in Eq. (1.2).)

By the law of total probability and the Markov property,

$$\begin{aligned}
\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) &= \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}, \ldots,\ X_0 = x) \cdot \mathbb{P}(X_{n-1} = x_{n-1}, \ldots,\ X_0 = x) \\
&= \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}) \cdot \mathbb{P}(X_{n-1} = x_{n-1}, \ldots,\ X_0 = x) \\
&= p(x_{n-1}, x_n) \cdot \mathbb{P}(X_{n-1} = x_{n-1}, \ldots,\ X_0 = x) \\
&= p(x_{n-1}, x_n) \cdot p(x_{n-1}, x_{n-2}) \ldots \mathbb{P}(X_1 = x_1, X_0 = x_0) \\
&= p(x_{n-1}, x_n) \cdot p(x_{n-1}, x_{n-2}) \ldots \mathbb{P}(X_1 = x_1 \mid X_0 = x_0) \cdot \mathbb{P}(X_0 = x_0) \\
&= p(x_{n-1}, x_n) \cdot p(x_{n-2}, x_{n-1}) \ldots p(x_0, x_1) \cdot \mu(x_0) \\
&= \mu(x_0) \cdot p(x_0, x_1) \cdot p(x_1, x_2) \ldots p(x_{n-1}, x_n) \quad \blacksquare
\end{aligned}$$

References

A. Klenke. *Probability theory: a comprehensive course, 3rd Edition*. Springer Science & Business Media, 2020.

H. Wang. *Monte Carlo simulation with applications to finance*. CRC Press, 2012.