HW4 第四次作业解答

更新：2025 年 12 月 2 日

Exercise 1

因为依赖于 Legendre 多项式，将 $[-1,1]$ 上 $w(x)=1$ 的 Gauss 求积法则称为 *Gauss-Legendre* 求积。10 点 Gauss-Legendre 法则的节点和权重在表 1 中给出。

表 1: 范围在 $[-1,1]$ 上 10 点 Gauss-Legendre 法则的节点和权重

$\pm x_i$	A_i
0.148874338981631	0.295524224714753
0.433395394129247	0.269266719309996
0.679409568299024	0.219086362515982
0.865063366688985	0.149451394150581
0.973906528517172	0.066671344308688

1. 画出权重-节点图。
2. 求出曲线 $y=x^2$ 在 $[-1,1]$ 之间的面积（分别使用 Gauss-Legendre 法、梯形法和 Simpson 法）。将它们与实际答案比较，并评价该求积法的精确性。

解答 (1) 10 点 Gauss-Legendre 法则的节点和权重在表 1 中给出，绘制权重-节点图 1 所示。

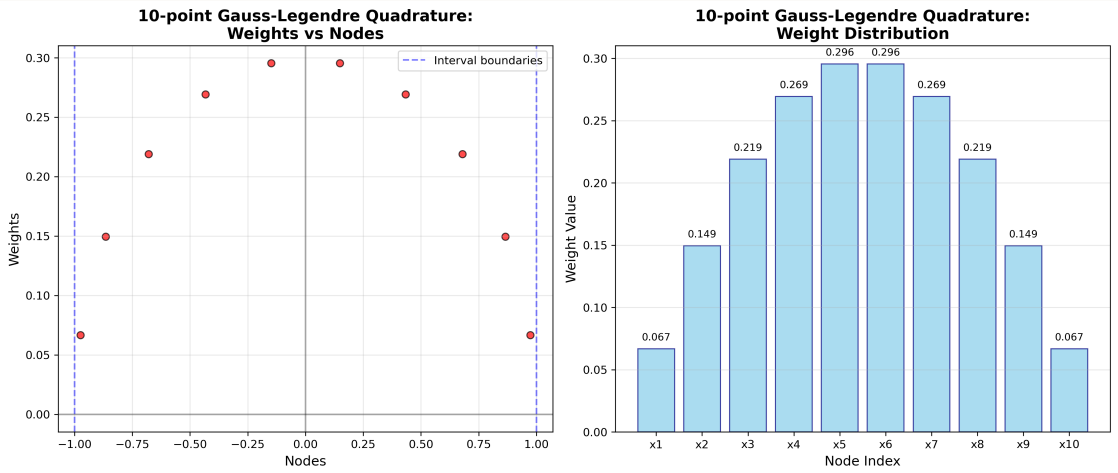

图 1: 10-point Gauss-Legendre quadrature nodes and weights

解答 (2) 实际答案为 $2/3 = 0.6666666667$，使用 Newton-Riemann, Newton-Trapezoidal, Newton-Simpson 和 Gaussian-Legendre 方法得到的结果如表 2 所示（均是只计算了 10 个函数值）。

表 2: 各类数值积分方法比较

Method	Integration	Error
Actual	0.6666666667	-
Newton-Riemann	0.6800000000	0.0133333333
Newton-Trapezoidal	0.6800000000	0.0133333333
Newton-Simpson	0.6666666667	0.0000000000
Gaussian-Legendre	0.6666666667	0.0000000000

程序计算节点-权重 x_i, A_i 使用了 `numpy.polynomial` 提供的方法，更为精确。程序运行结果见日志 `log/q1.log`。

Exercise 2

当 X 的密度函数 $f_X(x)$ 与 $q_X(x) = e^{-|x|^3/3}$ 成比例时，试估计 $\sigma^2 = \mathbb{E}(X^2)$。

1. 使用带有权重标准化的重要性采样（Importance Sampling, IS）估计 σ^2。
2. 使用拒绝抽样（Rejection Sampling, RS）估计 σ^2。

解答 (1) SIR 估计：例如选取正态分布 $N(0, b^2)$

$$g(x) = (2\pi b^2)^{-1/2} e^{-x^2/2b^2}$$

因为 e^{-x^2} 比 $e^{-|x|^3}$ 厚尾，故这样的 g 是符合要求的。步骤如下：

1. 抽取独立同分布样本 $X_1, \cdots, X_m \overset{\text{i.i.d.}}{\sim} N(0, b^2)$
2. 计算重要性
$$w^*(X_i) = \frac{f(X_i)}{g(X_i)} = \frac{C \cdot q(X_i)}{g(X_i)}$$
其中 C 为常数。
3. 标准化重要性
$$w(X_i) = \frac{w^*(X_i)}{\sum_{i=1}^n w^*(X_i)} = \frac{C \cdot q(X_i)/g(X_i)}{\sum_{i=1}^n C \cdot q(X_i)/g(X_i)} = \frac{q(X_i)/g(X_i)}{\sum_{i=1}^n q(X_i)/g(X_i)}$$
4. 从 $X_1, \cdots, X_m$ 中按照概率 $\{w(X_i)\}_{i=1}^m$ 有放回的简单随机抽样，样本容量为 $n \le m/10$，如此得到 $X_1^*, \cdots, X_n^*$。这一步不是必须的，目标若只是估计期望，则无需重抽样。
5. 计算 X_i 的二阶样本原点加权矩作为 σ^2 的估计
$$\hat{\sigma}^2 = \hat{\mathbb{E}}(h(X)) = \sum_{i=1}^n w(X_i) h(X_i) = \sum_{i=1}^n w(X_i) \cdot X_i^2$$

抽取 $N = 100$ 个随机样本，得到估计值为 $\hat{\sigma} = 0.706399$。程序结果见文件 `log/q2.log`。

解答 (2) RS 估计：注意到 $\log q(x) = -|x|^3/3$，当 $x \geq 0$ 时 $d^2 \log q/dx^2 = -2x \leq 0$；当 $x < 0$ 时，$d^2 \log q/dx^2 = 2x < 0$。所以由此可知 $q(x)$ 是 log-concave 的。

记 $r(x) = \log q(x) = -|x|^3/3$，注意到 $r(0) = 0$ 且 $r(x)$ 是偶函数。不妨用 3 条切线包络，且其中一条为 $l_2(x) = 0$，而另 2 条对称。不妨设右切点为 $(x_0, r(x_0))$，于是切线为

$$l_1(x) = r'(x_0)(x - x_0) + r(x_0) = -x_0^2 x + \frac{2x_0^3}{3}$$
$$l_2(x) = r'(-x_0)(x + x_0) + r(-x_0) = x_0^2 x + \frac{2x_0^3}{3}$$

其中 $x_0 \geq 0$。若设置上下限 $[-N, N]$ 希望曲线间面积

$$\begin{aligned} S_N(x_0) &= \int_{-N}^{0} \left(x_0^2 x + \frac{2x_0^3}{3} - r(x) \right) dx + \int_{0}^{N} \left(-x_0^2 x + \frac{2x_0^3}{3} - r(x) \right) dx \\ &= \frac{1}{6} N^4 - x_0^2 N^2 + \frac{4}{3} x_0^3 N \end{aligned}$$

尽可能小，从而抽样效率更高。解 $\partial S_N(x_0)/\partial x_0 = 0$ 得最优切点为 $x_0^* = N/2$。于是，我们得到 Envelope 函数

$$e(x) = \begin{cases} \exp\left(-x_0^2 x + \frac{2x_0^3}{3} \right), & x > 2x_0/3 \\ 1, & x \in [-2x_0/3, 2x_0/3] \\ \exp\left(x_0^2 x + \frac{2x_0^3}{3} \right), & x < -2x_0/3 \end{cases}$$

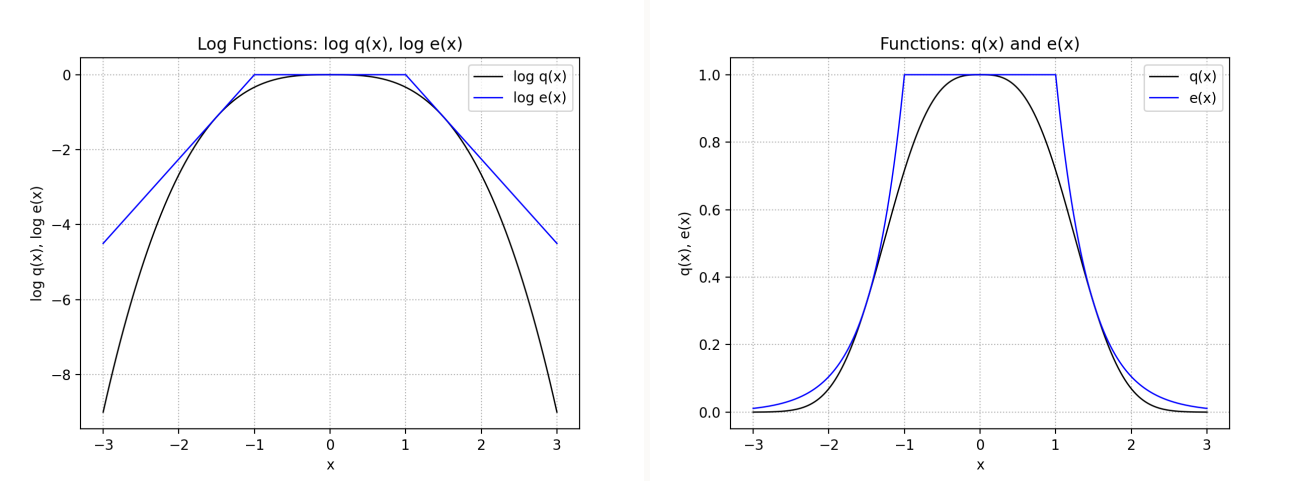

图 2: Target and Envelope Target Function, select $x_0 = 1.5$. (left: $\log q, \log e$, right: q, e)

下面构造分布 $g(x) = \alpha \cdot e(x)$ 使得 $\int g(x)\, dx = 1$，即

$$\alpha \cdot \left[\int_{-\infty}^{-2x_0/3} \exp\left(x_0^2 x + \frac{2x_0^3}{3} \right) dx + \frac{4x_0}{3} + \int_{2x_0/3}^{\infty} \exp\left(-x_0^2 x + \frac{2x_0^3}{3} \right) dx \right] = 1$$

求得 α 为

$$\alpha = \frac{3x_0^2}{4x_0^3 + 6}$$

为方便从 $g(x)$ 中抽取独立同分布样本，下面求 $g(x)$ 的概率质量函数，

$$G(x) = \begin{cases} 1 - \dfrac{\alpha}{x_0^2} \exp\left(-x_0^2 x + \dfrac{2x_0^3}{3}\right), & x > 2x_0/3 \\ \alpha x + \dfrac{1}{2}, & x \in [-2x_0/3, +2x_0/3] \\ \dfrac{\alpha}{x_0^2} \exp\left(x_0^2 x + \dfrac{2x_0^3}{3}\right), & x < -2x_0/3 \end{cases}$$

不难发现 $G^{-1}(x)$ 是容易求得的

$$G^{-1}(u) = \begin{cases} \dfrac{2x_0}{3} - \dfrac{1}{x_0^2} \log\left(\dfrac{x_0^2(1-u)}{\alpha}\right), & u \in \left[1 - \alpha/x_0^2, 1\right) \\ \dfrac{u - 1/2}{\alpha}, & u \in (\alpha/x_0^2, 1 - \alpha/x_0^2) \\ \dfrac{1}{x_0^2} \log\left(\dfrac{x_0^2 u}{\alpha}\right) - \dfrac{2x_0}{3}, & u \in (0, \alpha/x_0^2] \end{cases}$$

下面开始抽样：

1. 抽取独立同分布样本 $U_1, \cdots, U_m \overset{\text{i.i.d.}}{\sim} U(0,1)$。
2. 计算 $G^{-1}(U_1), \cdots, G^{-1}(U_m)$，于是有 $X_i = G^{-1}(U_i) \overset{\text{i.i.d.}}{\sim} g(X)$。
3. 计算 $q(X_i)$ 和 $e(X_i)$。从 $U(0,1)$ 中抽取独立同分布的 Z_i，若 $Z_i > q(X_i)/e(X_i)$ 则拒绝当前的 X_i。
4. 如此，接受的 X_i 构成 $X_1^*, \cdots, X_n^* \overset{\text{i.i.d.}}{\sim} f(X)$，进而估计 $\sigma^2 = \mathbb{E}(X^2)$

$$\hat{\sigma}^2 = \hat{\mathbb{E}}(X^2) = \frac{1}{n}\sum_{i=1}^{n} h(X_i^*) = \frac{1}{n}\sum_{i=1}^{n} [X_i^*]^2$$

使用 $q(X)$ 代替 $f(X)$ 的原因：若 $f(X) = Cq(X)$，则 $f(X)/e(X) = Cq(X)/e(X)$，可以将 $e(X)/C$ 记为新的 $e(X)$，即有等价。

抽取 $N = 100$ 个随机样本，取 $x_0 = 1.5$ 得到 $\alpha = 0.35$。最终估计值为 $\hat{\sigma} = 0.711958$。接受率为 82%，其中有 7 个样本阈值 $q(X)/e(X) \geq 1$ 占比 7%。程序结果见文件 `log/q2.log`。

Exercise 3

Let $z_0 = 0$ and $\boldsymbol{z} = (z_1, \cdots, z_n)$ be a binary sequence (0's and 1's) of Markov chain with, governed by the transition matrix

$$T = \begin{pmatrix} 0.9 & 0.1 \\ 0.05 & 0.95 \end{pmatrix}$$

where $T_{ij} = \mathbb{P}(z_{t+1} = j | z_t = i)$ is the transition probability. When $z_t = 1$, the casino uses a fair die (i.e., each side has equal probability $1/6$ to show up); whereas when $z_t = 0$, the casino uses a loaded die

that has probability $1/2$ to show "6", and probability $1/10$ to show other faces. Of course, we only got to serve the results of the die throws, denoted as the y_i's. Note that this is a simple hidden Markov model. Suppose we have observed the results $\boldsymbol{y}$ of 100 throws, as

$$
\begin{aligned}
&2,1,6,2,1,6,6,5,6,6,6,3,5,2,3,2,1,2,6,4,6,2,2,5,3,3,3,1,4,3,1,5,1,3,\\
&6,1,6,3,5,1,6,3,1,2,3,1,4,6,3,6,5,1,3,3,5,6,1,3,5,5,4,6,3,2,4,1,6,2,\\
&5,4,2,4,4,2,1,2,3,2,6,3,6,6,6,4,5,6,2,2,4,6,6,1,4,6,3,4,2,6,4,6
\end{aligned}
$$

1. Sample $10,000$ independent draws of $\boldsymbol{z}$ using the **sequential importance sampling** to estimate $\mathbb{E}(\boldsymbol{z}|\boldsymbol{y})$.
2. Based on the draws in (a), estimate $\hat{\boldsymbol{z}} = \arg\max_{\boldsymbol{z}} \mathbb{P}(\boldsymbol{z}|\boldsymbol{y})$. The true underlying hidden sequence is

$$
\begin{aligned}
\boldsymbol{z}^* = \{\ &1,1,1,1,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,\\
&1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,\\
&1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0\ \}
\end{aligned}
$$

   You can compare your estimated $\hat{\boldsymbol{z}}$ with the true one.

解答 (1, 2) 目标概率分布为 $f(\boldsymbol{z}|\boldsymbol{y}) = f(z_{1:T}|y_{1:T})$，其中 $T = n = 100$。而转移概率为 $p_z(z_t|z_{t-1}) = T_{z_{t-1}z_t}$。给定 z_t 后观测的概率为

$$
p_y(y_t|z_t) = \begin{cases} 1/6, & z_t = 1 \\ 1/2, & z_t = 0, y_t = 6 \\ 1/10, & z_t = 0, y_t \in \{1,2,3,4,5\} \end{cases}
$$

于是重要性权重为 $w_t(z_{1:t}) \propto w_{t-1}(z_{1:t-1}) \cdot p_y(y_t|z_t)$。注意到，观测值 $\boldsymbol{y}$ 已知，下面开始抽取 N 个样本：

1. 初始化。设 $z_0 = 0$，重要性权重 $w_0 = 1$。
2. 在第 t 时刻，按转移概率 $p_z(z_t|z_{t-1})$ 抽样新的分量

$$
z_t|z_{t-1} \sim \text{Bernoulli}(1, T_{z_{t-1}1})
$$

3. 此时，由观测值 y_t 计算 $p_y(y_t|z_t)$，从而更新重要性权重 $w_t(z_{1:t}) \propto w_{t-1}(z_{1:t-1}) \cdot p_y(y_t|z_t)$。
4. 回到第 2 步，直到达到 $t = T$。如此，得到单个样本序列抽样 $\boldsymbol{z} = z_{1:T}$ 和其对应的重要性权重 $w_T(z_{1:T})$。
5. 反复操作 N 次（为提高效率，也可对上述操作向量化），得到 N 个样本 $\boldsymbol{z}^1, \cdots, \boldsymbol{z}^N$ 和其权重 $w^1, \cdots, w^N$。

$\boldsymbol{z}$ 的期望可由样本加权和估计，先对权重归一化

$$
w_*^i = \frac{w^i}{\sum_{i=1}^N w^i}
$$

于是期望的估计值为

$$\hat{\mathbb{E}}(\boldsymbol{z}|\boldsymbol{y}) = \frac{1}{N}\sum_{i=1}^{N} w_*^i \cdot \boldsymbol{z}^i$$

对于 $\hat{\boldsymbol{z}} = \arg\max_{\boldsymbol{z}} \mathbb{P}(\boldsymbol{z}|\boldsymbol{y})$，即为 $\max w_*^i$ 对应的 $\boldsymbol{z}^i$。

结果见程序日志 `log/q3.log`，$\hat{\boldsymbol{z}}$ 与真实序列 $\boldsymbol{z}^*$ 相差 12 个，准确率为 88%，但有效样本量 ESS ≈ 57。若采用 Particle Filter，在设置有效样本阈值为 $N/7 \approx 1429$ 后，得到的结果：$\hat{\boldsymbol{z}}$ 与真实序列 $\boldsymbol{z}^*$ 相差 15 个，准确率为 85%，但有效样本量 ESS ≈ 2562。故采用 Particle Filter 有效样本更大，其估计 $\hat{\mathbb{E}}(\boldsymbol{z}|\boldsymbol{y})$ 的方差更小，下面给出简单地理论证明（不严格）。

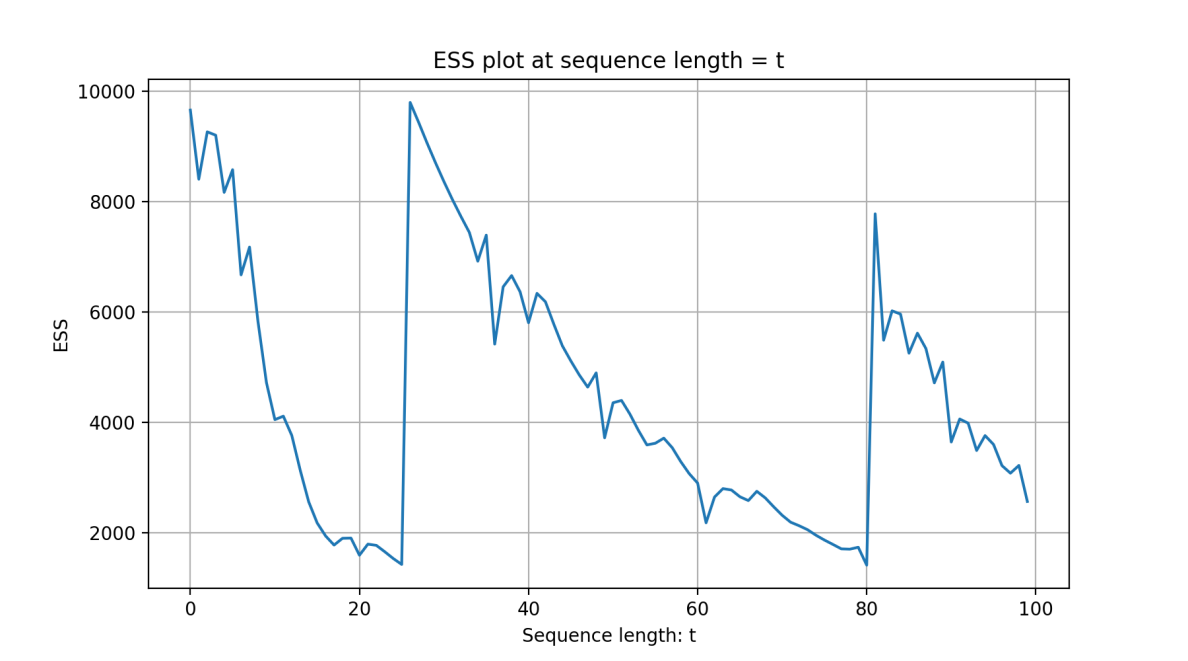

图 3: ESS Plot with Particle Filter

注意到，在这里有效样本为

$$\text{ESS} = \frac{1}{\sum_{i=1}^{N}(w_*^i)^2}$$

有效样本 ESS 越大，归一化重要性权重的平方和 $\sum_{i=1}^{N}(w_*^i)^2$ 越小。设待估计量为 $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{z}|\boldsymbol{y})$，由其估计量的形式，有估计的方差为（不妨探讨一维情形）

$$\text{Var}[\hat{\boldsymbol{\mu}}] = \frac{1}{N}\left[\text{Var}(w_*^i\boldsymbol{z}^i) + \boldsymbol{\mu}^2\text{Var}(w_*^i)\right] - 2\boldsymbol{\mu}\text{Cov}(w_*^i\boldsymbol{z}^i, w_*^i) + O(N^{-2})$$

方差估计为

$$\hat{\text{Var}}[\hat{\boldsymbol{\mu}}] = \frac{1}{N}\left[S_{w_*\boldsymbol{z}}^2 + \boldsymbol{\mu}^2 S_{w_*}^2\right] - 2\boldsymbol{\mu}S_{w_*\boldsymbol{z},w_*} + O(N^{-2})$$

注意到 $\bar{w}_* = \sum_{i=1}^{N} w_*^i/N = 1/N$，故有 w_*^i 的样本方差

$$S_{w_*}^2 \propto \sum_{i=1}^{N}(w_*^i)^2 - N(\bar{w}_*)^2 = \sum_{i=1}^{N}(w_*^i)^2 - \frac{1}{N}$$

而又 $w_*^i \geq 0, \boldsymbol{z}^i \in \{0,1\}$，于是其他样本方差和协方差

$$S_{w_*\boldsymbol{z}^i}^2 \propto \sum_{i=1}^{N}(w_*^i\boldsymbol{z}^i)^2 - N\overline{w_*\boldsymbol{z}} \leq \sum_{i=1}^{N}(w_*^i)^2$$

$$S_{w_*\boldsymbol{z},w_*} \propto= \sum_{i=1}^{N}(w_*^i\boldsymbol{z}^i)\cdot w_*^i - N\bar{w}_*\overline{w_*\boldsymbol{z}} = \sum_{i=1}^{N}(w_*^i)^2\boldsymbol{z}^i - \overline{w_*\boldsymbol{z}} \geq \sum_{i=1}^{N}(w_*^i)^2\boldsymbol{z}^i - \bar{w}_* \geq -\frac{1}{N}$$

于是，通过简略的放缩，有

$$\begin{aligned}\hat{\mathrm{Var}}[\hat{\boldsymbol{\mu}}] &\leq \frac{1}{N}\left[\sum_{i=1}^{N}(w_*^i)^2 + \boldsymbol{\mu}^2\left(\sum_{i=1}^{N}(w_*^i)^2 - \frac{1}{N}\right)\right] + \frac{2\boldsymbol{\mu}}{N} + O(N^{-2}) \\ &= \left(\frac{1}{N} + \boldsymbol{\mu}^2\right)\sum_{i=1}^{N}(w_*^i)^2 + \frac{2\boldsymbol{\mu}N - \boldsymbol{\mu}^2}{N^2} + O(N^{-2})\end{aligned}$$

即估计的方差的上界，随 $\sum_{i=1}^{N}(w_*^i)^2$ 减小而减小，而 ESS 越大，$\sum_{i=1}^{N}(w_*^i)^2$ 越小，即估计越有效。

注 此处理论证明的放缩不甚严谨，可以采用其他大样本近似的方法推导。

Exercise 4

1. 拒绝抽样（Rejection Sampling, RS）与抽样重要性重抽样（Sampling Importance Resampling, SIR）都是常用的抽样方法，但 RS 是精确方法 SIR 却是近似方法。解释 SIR 是近似方法的原因。
2. 重要性采样（Importance Sampling, IS）与抽样重要性重抽样（Sampling Importance Resampling, SIR）听起来非常相似，但它们服务于不同的目的。请对比这两种方法：
   (a). 它们各自的主要目标是什么？
   (b). 比较两种方法估计均值 μ 的方差，IS 的方差更小，从直观上解释这一现象。
   (c). 如果使用拒绝抽样（Rejection Sampling, RS）估计均值 μ，其方差与 IS 估计的方差相比如何？是否有确定的答案？
3. 随着维度增加，许多经典采样方法的效率急剧下降（维度诅咒）。例如，拒绝抽样的接受率会指数级下降。结合深度学习的发展，是否有新的思路来解决高维采样问题？

解答 (1) SIR 第一步从已知分布抽样，而非目标分布，故只能近似。而在第二步重采样，又一次引入随机性。若第一步样本量不足，从而有第二步重采样不准确。整体而言，就是近似的方法。（相比于 SIR，被 RS 接受的样本，确实是落在了目标分布中）。

解答 (2) IS 关注与对积分/期望的估计，无需真正的产生服从目标分布的样本，产生的样本是不等权的，通过加权进行后续的估计。SIR 目标是产生近似服从目标分布的样本，这些样本是等权的，可以作为目标分布的样本进行后续的操作。

SIR 在重抽样阶段又一次引入了随机性，故方差更大。

RS 在选取的已知分布对目标分布包络良好、接受率高时，产生的样本更加近似服从目标分布，估计的方差小。IS 方法估计的方差同样取决于已知分布的选取，即对权重的计算有所不同，会导致不同的方差。故没有确定的答案，需要针对特定的选取考虑

解答 (3) 例如，直接训练一个网络，使其拟合已有的样本。训练完成后，传入一个随机初始值，得到一个新的样本。

$$f(\boldsymbol{Y};\boldsymbol{\theta}_{\mathrm{NN}})\sim\boldsymbol{X}\quad\Rightarrow\quad\hat{\boldsymbol{X}}=f(\boldsymbol{Y}_0;\boldsymbol{\theta}_{\mathrm{NN}}),\quad\boldsymbol{Y},\boldsymbol{Y}_0\text{ are r.v.}$$

其中 $\boldsymbol{X}$ 是观测到的目标族（高维），$\boldsymbol{Y},\boldsymbol{Y}_0$ 作为辅助变量（低维），服从已知的一个分布。

再如，训练循环神经网络 GNN，在已有观测数据上训练。然后抽样低维 $\boldsymbol{X}_{0:b}$，对后续进行预测，得到新的高维样本

$$f_{\mathrm{GNN}}(X_0;\boldsymbol{\theta}_{\mathrm{NN}})=\boldsymbol{X}\quad\Rightarrow\quad\hat{\boldsymbol{X}}=f_{\mathrm{GNN}}(\hat{X}_0;\boldsymbol{\theta}_{\mathrm{NN}})$$

当然，也可以考虑双向 GNN。如此输入即可为中间分量 $\boldsymbol{X}_{a:b}$ 的数据。

或者，直接将高维 $\boldsymbol{X}$ 视作序列，除了上述的 GNN，还可以使用其他 Seq2Seq 模型，甚至是 Transformer 架构。产生新的样本，只需输入低维的数据，然后由网络进行预测。

当然，如上方法均未验证。考虑到网络可能存在黑盒性，且后续的估计很难写出解析式，故相关统计性质比较难评估。