第三周：重要性采样

重要性采样可以在教材的 2.4 节找到参考内容。

问题引入：为什么需要重要性采样？

实际工程中往往需要计算积分 / 期望 …… 等等，这些计算通常可以抽象为计算如下的表达：

$$I = \int_a^b f(x)p(x)dx$$

上述积分中，如果将 $p(x)$ 视为概率，那么积分值就可以理解为：x 为概率密度函数为 $p(x)$ 的随机变量，f 是和 x 相关的函数（有连续性等要求），整个 $f(x)$ 为一个新的随机变量，计算其的期望。

由于无法显式的算出上述式子的值，必须要使用蒙特卡洛算法来进行近似计算。那么如何来近似计算呢？需要离散！

$$I = E_{x\sim p}[f(x)] = \frac{1}{N}\sum_{x_i\sim p, 1\le i\le N, a<x_i<b} f(x_i) = \frac{1}{N}\sum_{x_i\sim p, 1\le i\le N} \mathbf{I}_{a<x_i<b}\cdot f(x_i)$$

注意到，微分算符 dx 离散为了 $1/N$，由于和 i 无关因此提到前面也可以。

公式中的 x_i 应该服从概率密度为 $p(x)$ 的原始分布。示例代码：

计算积分：

$$I = \int_4^\infty \frac{e^{-x^2/2}}{\sqrt{2\pi}}dx = \frac{1}{2}erfc(2\sqrt{2}) \approx 3.16712418e-5$$

上面的积分即为 $f = \mathbf{I}_{x>4}$，$p(x)$ 为标准正态密度。

int_eg.m

```
%% 重要性采样计算尾概率 P(X>4), X~N(0,1)
```

```
clc, clear
% 真实概率密度函数（标准正态）
p = @(x) exp(-x.^2/2) / sqrt(2*pi);

true_p = erfc(2*sqrt(2)) / 2;

% 方法1：简单蒙特卡洛（效率极低）
fprintf('=== 方法比较：计算 P(X>4) ===\n\n');

N = 100000;
x_mc = randn(N, 1);  % 从标准正态采样
success_mc = sum(x_mc > 4);
p_mc = success_mc / N;
fprintf('简单蒙特卡洛:\n');
fprintf('  估计值: %.6e\n', p_mc);
fprintf('  有效样本数: %d/%d (效率: %.4f%%)\n', success_mc, N,
100*success_mc/N);
fprintf('  相对误差: %.2f%%\n\n', 100*abs(p_mc - true_p)/true_p);
```

改变 N，可以发现精度很低，提高精度所需要的N过多。

为什么结果不好？

因为积分值中占主导地位的是 $x = 4$ 附近的点，而标准正态分布采样到 4 附近的概率不是很大。即便是较大的 N 也很难采样到对积分值能有明显贡献的样本。

Example: Company Salary Estimation

Consider a company with $n = 10{,}000$ employees and we want to estimate the average salary. However the salaries are very imbalanced; the CEO makes way more than the typical employee does. Say we know the CEO makes at most 2 million a year, but the other 9,999 employees make at most 50 thousand a year.

Using just uniform sampling of $k = 100$ employees, we can apply a Chernoff-Hoeffding bound to estimate the average salary $\hat{w}$ from the true average salary $\bar{w}$ with error more than \$8,000 with probability

$$\mathbf{Pr}[|\hat{w} - \bar{w}| \geq 8{,}000] \leq 2\exp\left(\frac{-2(8{,}000)^2 \cdot 100}{(2\text{ million})^2}\right) = 2\exp\left(\frac{-2}{625}\right) \approx 1.99$$

This is a useless bound, since the probability is greater than 1. If we increase the error tolerance to half a million, we still only get a good estimate with probability 0.42. The problem hinges on if we sample the CEO, and our estimate is too high; if we do not, then the estimate is too low.

上面两个例子表明了直接蒙特卡洛模拟事实上无法胜任计算这样的积分，如果需要预定精度，将会付出无法承受的计算代价。此时有必要引入重要性采样。

重要性采样

重要性采样的思想非常简单，既然采样到 4 附近的点概率不高，那么换一个分布来采样 x 不就可以采样到 4 附近了。

什么分布生成的样本有较大概率在 4 附近？

$$N(4.5, 1)$$

上面的分布即为建议分布 $q(x)$.

$$q(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-4.5)^2}{2}\right)$$

原积分可以表示为：

$$E_p[f(x)] = \int_4^\infty f(x)p(x)dx = \int_4^\infty \frac{p(x)}{q(x)}q(x)dx = E_q[\frac{p(x)}{q(x)}f(x)]$$

其中 $p(x)/q(x)$ 即为 x 的重要性权重。

```
% 方法2：重要性采样（使用平移的正态分布作为建议分布）
% 选择建议分布：N(4.5, 1)，集中在重要区域
mu_q = 4.5;
sigma_q = 1;
q = @(x) exp(-(x-mu_q).^2/(2*sigma_q^2)) / (sigma_q*sqrt(2*pi));

% 从建议分布采样
x_is = mu_q + sigma_q * randn(N, 1);

% 计算重要性权重
weights = p(x_is) ./ q(x_is);

% 计算指示函数和加权平均
indicator = (x_is > 4);
p_is = mean(indicator .* weights);

fprintf('重要性采样:\n');
fprintf('  估计值: %.6e\n', p_is);
fprintf('  有效样本数: %d/%d (效率: %.2f%%)\n', sum(x_is>4), N,
100*sum(x_is>4)/N);
fprintf('  相对误差: %.2f%%\n\n', 100*abs(p_is - true_p)/true_p);
```

可以发现精度提高很多。

重要性采样的意义：

- 有效降低了离散积分值的方差——参考教材 p.37，结合 **Chernoff – Hoeffding** 不等式
- 可以使用一个已知分布 $q(x)$ 产生的样本点来推断未知分布 $p(x)$ 的信息

重要性采样的缺点：

- 需要先验信息来选择建议分布，很多时候我们对 p 所知甚少

很多时候，我们会直接使用均匀分布作为建议分布，它取到4附近的概率也不低，而且可以在没有任何先验知识的情况下来近似积分。不过这个例子不适用，因为被积区域总长度为∞。

例如：

计算：

$$I = \int_4^8 \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx$$

使用均匀分布可以离散为：

$$\frac{8-4}{N} \sum_{1 \le i \le N, x_i \sim U[4,8]} \frac{e^{-x_i^2/2}}{\sqrt{2\pi}}$$

使用 linspace 可以做到完美的采样。

```
p = @(x) exp(-x.^2/2) / sqrt(2*pi);
true_p = 1/2 * (erf(4 * sqrt(2)) - erf(2 * sqrt(2)));

a = 4;
b = 8;

N = 100000;
%% 使用均匀分布作为建议分布，计算积分
x_mc = rand(N, 1) * (b - a) + a;  % 从均匀分布中采样
success_mc = sum((x_mc >= a) .* (x_mc <= b));  % 所有点都在被积区域内
p_mc = mean(p(x_mc)) * (b - a);

fprintf('重要性采样（均匀分布）:\n');
fprintf('  估计值: %.6e\n', p_mc);
fprintf('  有效样本数: %d/%d (效率: %.4f%%)\n', success_mc, N,
100*success_mc/N);
fprintf('  相对误差: %.2f%%\n\n', 100*abs(p_mc - true_p)/true_p);


%% 使用linspace代替均匀分布采样
x_mc = linspace(a, b, N);
```

```
success_mc = sum((x_mc >= a) .* (x_mc <= b));
p_mc = mean(p(x_mc)) * (b - a);

fprintf('重要性采样（linspace）:\n');
fprintf('  估计值: %.6e\n', p_mc);
fprintf('  有效样本数: %d/%d (效率: %.4f%%)\n', success_mc, N,
100*success_mc/N);
fprintf('  相对误差: %.2f%%\n\n', 100*abs(p_mc - true_p)/true_p);
```

可以发现，均匀分布的结果也不错，使用linspace则更好。

思考：

- 均匀分布作为建议分布时适用面非常广，可以有效的采样到整个区域，而且可以使用 linspace 达到完美的采样效果，不会漏掉任何对积分值有重要贡献的点。什么情况下好用，什么情况下不好用？

低维情况下好用，高维下不好用。
高维 $x \in \mathbb{R}^D, D \gg 1$ 时，均匀分布的采样效率很低。假设每个维度长度差不多，均采1000个点，那么使用 linspace 将需要生成 1000^D 个样本点，这是无法计算的。直接使用均匀分布也面临同样的问题，高维空间过大，均匀分布的采样概率不集中，无法保证能采样到对积分有贡献的点附近。

拉丁超立方采样策略可以解决这个问题——Latin hypercube sampling。

一般情况下还是推荐换用其他的采样分布。

建议分布的选择方法（了解）

$$E_p[f(x)] = \int_a^b f(x)p(x)dx = \int_a^b \frac{p(x)}{q(x)}q(x)dx = E_q[\frac{p(x)}{q(x)}f(x)]$$

设想如果 q 的选择非常完美，恰好使得 $q = Cp(x)f(x)$. 那么右端的期望内部，将直接变为一个常数，仅需要一个样本，就可以完美估算这个积分结果！

例子：

计算如下积分：

$$I = \int_R \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

上式是标准正态的密度函数的实数域积分，精确值等于1。如果使用蒙特卡洛模拟计算这个积分，可以使用标准正态分布作为采样分布，计算如下的期望：

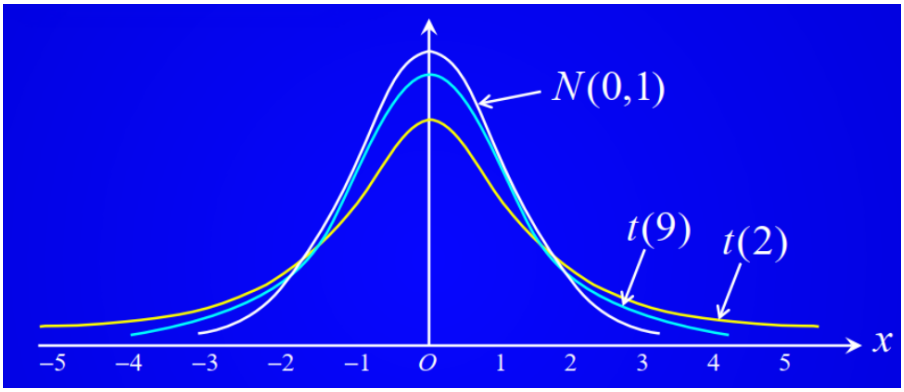
$$I = E_{x \sim N(0,1)}[1]$$

随便采样什么结果都是完全准确的，方差为0.

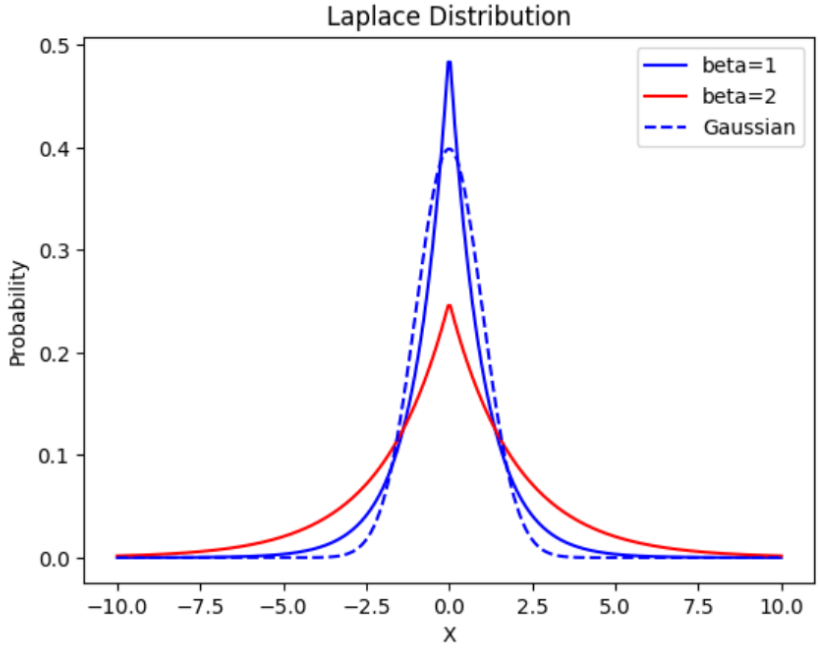
不过有这么好的建议分布（采样分布）往往是可遇而不可求的，因此我们的目的就在于：选择一个形状上近似于 $f(x) \mid p(x)$ 的、且易于采样的分布$q(x)$ 。

根据不同的场景，可以采用以下策略：

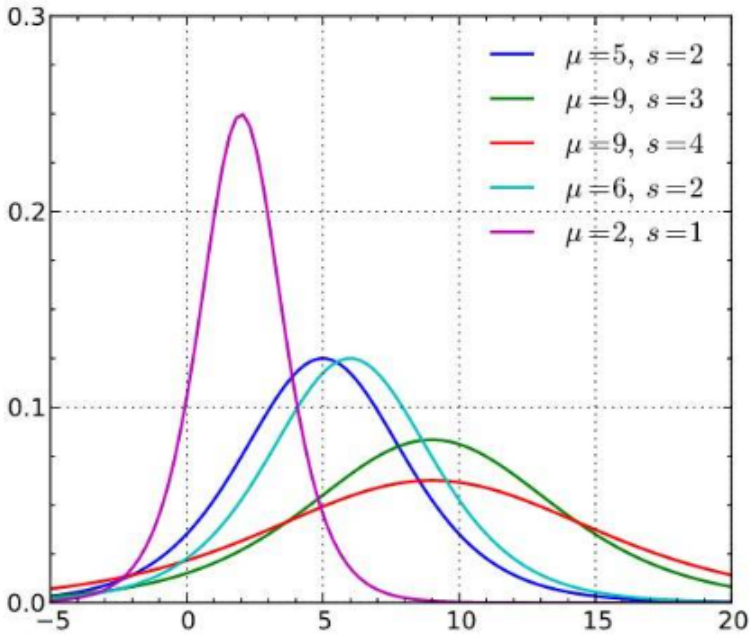
1. 基于领域知识进行启发式选择
    - 方法：分析目标分布 p(x) 和被积函数 f(x) 的性质。如果知道 p(x) 是单峰的，可以用一个更“分散”的高斯分布或t分布来覆盖它。如果知道 f(x) 在某些区域值很大，可以尝试让 q(x) 在这些区域有更高的概率质量。
    - 优点：简单直接。
    - 缺点：依赖于经验和运气，不一定是最优的。
2. 使用重尾分布
    - 方法：当对 p(x) 了解不多，或者 p(x) 本身是重尾分布时，一个安全的选择是使用尾部更厚的分布作为建议分布。常用的重尾分布包括：
        - **t**-分布：自由度越低，尾部越厚。当自由度=1时，就是柯西分布，尾部非常厚。



        - 拉普拉斯分布



- 对数逻辑分布等



- 优点：能有效避免因尾部问题导致的方差爆炸，非常稳健。
- 缺点：可能会过于保守，因为其形状可能与 ｜f(x)｜p(x)｜ 相差较远，导致效率不是最高。

3. 参数化建议分布与自适应重要性采样

- 方法：这是更高级和自动化的一种方法。我们选择一个参数化的分布族作为 q(x;θ)（例如高斯混合模型），然后通过迭代采样来优化参数 θ，使其不断逼近最优的建议分布。
- 常见算法：
    - 自适应重要性采样：用前一轮采样的加权结果来更新 q(x;θ) 的参数（如均值和方差）。

- 交叉熵方法：通过最小化 q(x;θ) 与最优建议分布 q(x) 之间的KL散度，来迭代更新参数 θ。
- 优点：可以自动找到接近最优的建议分布，效率高。
- 缺点：实现复杂，计算成本较高。

4. 混合分布/防御性重要性采样

- 方法：为了确保建议分布的尾部足够厚，可以将其与一个重尾分布（如一个宽的高斯分布或均匀分布）进行混合。

$$q(x) = \gamma q_{main}(x) + (1-\gamma) q_{heavy-tailed}(x)$$

其中 γ∈(0,1) 是一个混合系数。

- 优点：结合了主分布 qmain(x) 的高效率和重尾分布 $q_{heavy-tailed}(x)$ 的稳健性。即使 q_{main} 选择不当，重尾的“防御”成分也能保证采样不会完全失败。
- 缺点：需要选择混合系数和第二个分布。

作业

使用蒙特卡洛计算如下积分：

$$I = \int_2^{\infty} e^{2x} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \frac{e^2}{2}$$

1. 分析分布的形状特征，为重要性采样方法选择合适的建议分布并说明理由，尽量可视化【例图1~3】
2. 给出原始蒙特卡洛、重要性采样方法的估计值方差比较。图6.
3. 给出相对误差结果、效率。

optional: 分析收敛速度、分析建议分布的最优参数。

提交：

- 【必须】word / pdf 文档：包含图片和分析
- 【必须】matlab 代码文件
- 9.29日