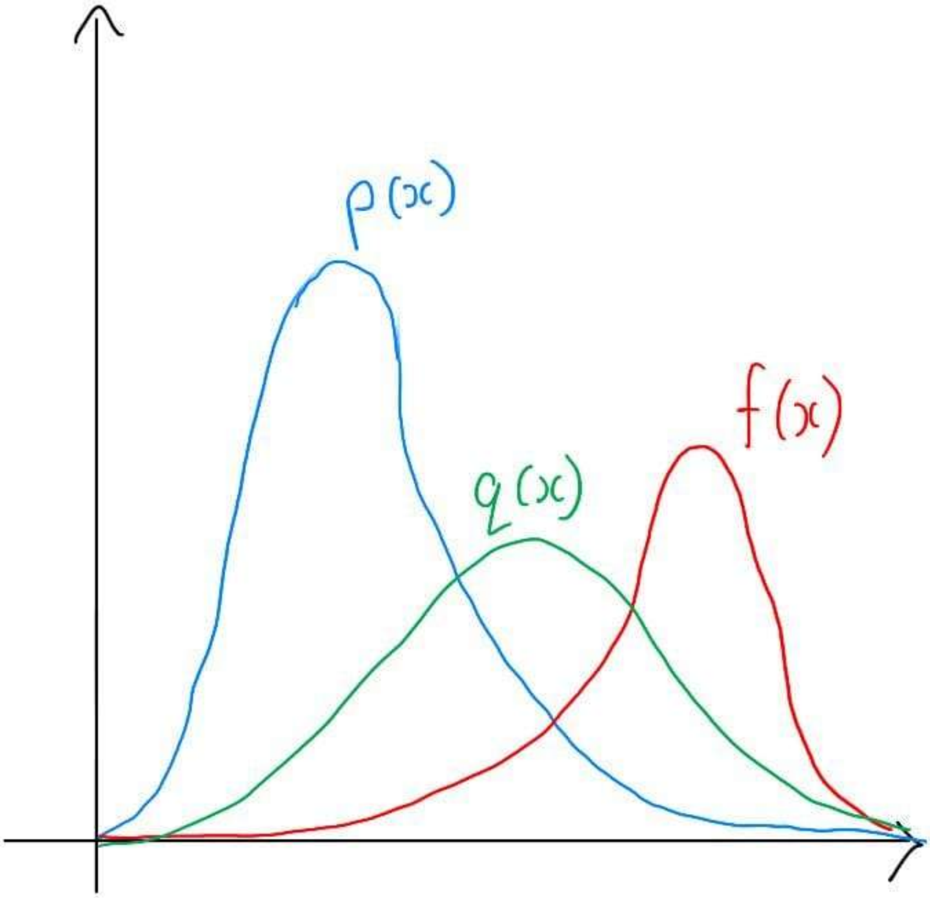
重要性采样(Importance Sampling)

目录



> 为什么要研究重要性采样(Importance Sampling)？这个问题曾经困扰了我很久。直到读了经典教科书[Robert and Casella, 1999]相关的章节, 我才真正理解了Importance Sampling。本文将用最简洁的语言阐述Importance Sampling的产生背景、意义以及若干理论结果。
>
> 本文只是科普读物，侧重于表达启发性的理解，无法代替读者去专研科学原文，特此声明。如有理解错误，请在评论区留言。

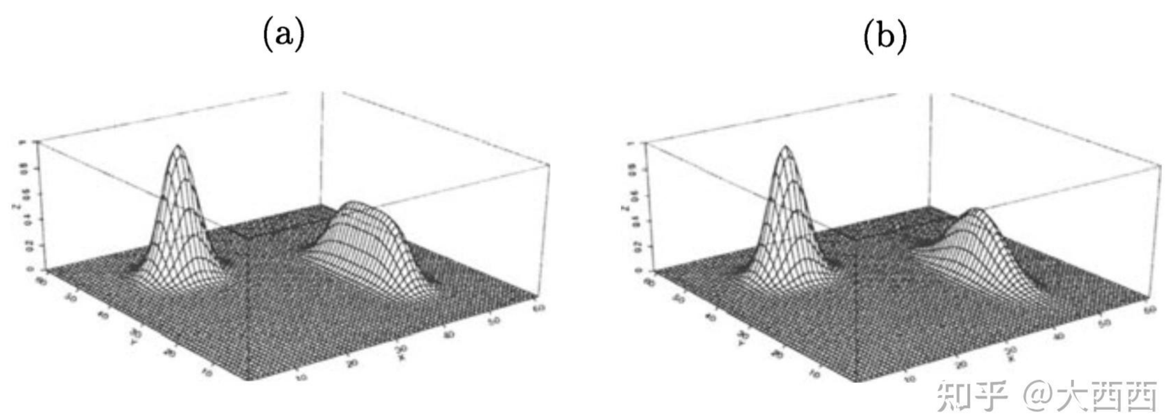
概述

> (Wikipedia)Importance sampling is a variance reduction technique that can be used in the Monte Carlo method. The idea behind importance sampling is that certain values of the input random variables in a simulation have more impact on the parameter being estimated than others. If these "important" values are emphasized by sampling more frequently, then the estimator variance can be reduced.

维基百科这三句话是教科书级别的概述。我就不逐字翻译了，因为中英文之间的差异是在太大了，逐字翻译反而不好理解。这三句从三个层面解释了Importance Sampling：(1) Importance Sampling是什么？（2）Importance Sampling的基本思想是什么？（3）Importance Sampling是如何达到其目的?

因此，可以把维基百科的这三句话提炼一下：Importance Sampling是一种减小方差的估值技术，其基本思想是挖掘出最影响估值的(重要)因素，利用好这些(重要)因素来估值就能达到减小方差的目的（从而加速推断）。

引言



蒙特卡洛积分

研究问题

很多机器学习任务需要估计下述数学期望 μ,

$$\mu = \mathbb{E}_{\mathbf{x}\sim p(\cdot)}[f(\mathbf{x})] = \int_{\mathcal{D}} f(\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x},$$ 其中 $\mathcal{D} \subset \mathbb{R}^d$， $p(\cdot)$是 $\mathcal{D}$上的概率密度函数， $f(\cdot)$是$\mathcal{D}$上的可积函数。

在随机变量清楚的情况下，简记 $\mu =: \mathbb{E}_p[f(\mathbf{x})]$。

经典蒙特卡洛估值

为了估计 μ，经典估值方法是：令 $\{\mathbf{x}_i\}_{i=1}^n \sim p(\cdot), \mathrm{i.i.d.}$，（即根据分布 p采样数据），我们可以获得如下经典蒙特卡洛估值,

$$\hat{\mu}_n =: \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i). \tag{1}$$

这个经典蒙特卡洛估值 $\hat{\mu}_n$有如下几个基本性质：

- 【无偏性】经典蒙特卡洛估值 $\hat{\mu}_n$是μ的一个无偏估计，即 $\mathbb{E}_p[\hat{\mu}_n] = \mu$。
- 【相合性】 $\forall \epsilon > 0$，$\lim_{n\to\infty} \mathbb{P}(|\hat{\mu}_n - \mu| \geq \epsilon) = 0$。这表明，只要有足够多的观察样本，经典蒙特卡洛估值 $\hat{\mu}_n$就应当逼近 μ的真值。相合性是点估计的基本要求，如果不断增加样本，都不能保证估值到任意指定精度，那么这个估值的技术就值得人们怀疑了。统计中，不满足相合性的估值方法，都不考虑。
- 【收敛速度】我们可以用两种方法来刻画收敛性。(1.)根据[中心极限定理]，可知 $\frac{\sqrt{n}(\hat{\mu}_n - \mu)}{\sigma} \xrightarrow{L} \mathcal{N}(0,1)$，这里 $\xrightarrow{L}$是指按分布收敛，这里假设 $\mathbb{V}\mathrm{ar}[f(\mathbf{x})] = \sigma^2$。因此我们得到渐进分布 $\hat{\mu}_n \dot{\sim} \mathcal{N}\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$，这表明经典蒙特卡洛估值 $\hat{\mu}_n$以 $\frac{1}{\sqrt{n}}$的速度 “接近”μ的真值。(2.)还可以根据[集中不等式]，在一些很弱的条件下，有 $\mathbb{P}\left(|\hat{\mu}_n - \mu| > \epsilon\right) \leq \exp\left\{-\frac{n\epsilon^2}{2\sigma^2}\right\}$，其中假设 $f(\cdot)$是 σ-[subgaussian]的随机函数。这是比相合性的进一步的结果，它给出了经典蒙特卡洛估值估值 $\hat{\mu}_n$的收敛速度。由于指数级的衰减，因此不需要太多的样本，经典估值 $\hat{\mu}_n$可以很快接近 μ的真值。

尽管经典估值 $\hat{\mu}_n$有三个优良的基本性质，可惜这并不是估计 μ的最优算法，这里最优性是基于方差来讲的。也即是说，经典估值 $\hat{\mu}_n$不一定是方差最小的估计。实际上后面可以证明，经典估值 $\hat{\mu}_n$通常都不是最优的，这也是要研究Importance Sampling的动机之一。这里要纠正一个说法，很多博客中认为研究Importance Sampling是由于实际问题中， f无法获得（或者 f是无法估计），这种理解是完全错误的。实际上，很多问题中 f是已知的（或者可以估计出来的），研究Importance Sampling的目的是设计更加有效的分布来采集数据，从而快速地估出 μ的真值。

事实上，不同的采样技术，会带来不同的估值速度，请看下例。

具体例子

Cauchy Tail Probability [Robert and Casella, 1999, Section 3.3]. 假设 P是柯西分布$\mathcal{C}(0,1)$在 $(2,+\infty)$上的概率，即 $P=\int_2^\infty \frac{1}{\pi(1+x^2)}\mathrm{d}x=\mathbb{E}_{x\sim\mathcal{C}(0,1)}\left[\mathbb{I}\{x>2\}\right]$。我们的目标就是把这个 $P=0.15$估计出来。

方法一：直接按照前面描述的经典估值，计算如下，

$$\hat{P}_1=\frac{1}{n}\sum_{i=1}^n \mathbb{I}\{x_i>2\},$$ 其中 $\{x_i\}_{i=1}^n\overset{\text{i.i.d}}{\sim}\mathcal{C}(0,1)$. 这个估值的方差是 $P(1-P)/n=0.127/n$。

方法二：利用柯西分布的对称性，

$$\hat{P}_2=\frac{1}{2n}\sum_{i=1}^n \mathbb{I}\{|x_i|>2\},$$ 这个估值的方差是 $P(1-2P)/2n=0.052/n$ 。

方法三： P可以改成 $P=\frac{1}{2}-\int_0^2\frac{1}{\pi(1+x^2)}\mathrm{d}x$,此时 P可以看成随机函数 $h(X)=\frac{2}{\pi(1+X^2)}$的数学期望，其中 $X\sim\mathcal{U}[0,2]$, $\mathcal{U}$是均匀分布。于是

$$\hat{P}_3=\frac{1}{2}-\frac{1}{n}\sum_{i=1}^n h(u_i),$$ 也是 P的一个无偏估计，其中 $\{u_i\}_{i=1}^n\overset{\text{i.i.d}}{\sim}\mathcal{U}[0,2]$。这个估计的方差是 $0.0285/n$。

方法四： P还可以改写成 $P=\int_0^{\frac{1}{2}}\frac{y^{-2}}{\pi(1+y^{-2})}\mathrm{d}y$,其可以看成随机函数 $\frac{1}{4}h(Y)=\frac{1}{2\pi(1+Y^2)}$的数学期望，其中 $Y\sim\mathcal{U}[0,\frac{1}{2}]$。于是

$$\hat{P}_4=\frac{1}{4n}\sum_{i=1}^n h(y_i),$$ 也是 P的一个无偏估计， $\{y_i\}_{i=1}^n\overset{\text{i.i.d}}{\sim}\mathcal{U}[0,\frac{1}{2}]$。这个估计的方差很小，只有 $0.95\times10^{-4}/n$。

经典估值$\hat{P}_1$与 $\hat{P}_4$方差之间的数量级相差是 10^3，根据集中不等式，如果经典估值$\hat{P}_1$要达到与 $\hat{P}_4$在方差上是同一个数量级，方法一就需要比方法四多采 $\sqrt{1000} \approx 33$倍数的数据样本。

这个例子给我们两点启发：

- 虽然都满足无偏性和相合性，按照方差最小原则，经典估值 $\hat{\mu}_n =: \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i), \{\mathbf{x}_i\}_{i=1}^{n} \sim p(\cdot), \mathrm{i.i.d.}$,并不是最优的估计方法。这里需要强调一下，方差越大，估值的效率往往越低。
- 方法三和方法四，都是根据新的均匀分布采样数据，然后推断柯西分布的某些性质。**这一点很关键，可升华到这个思想：运用分布 $q(\cdot)$的性质来推断与分布 $p(\cdot)$相关的性质。这是Importance Sampling思想的渊源。源于方差减小分析，最终脱离方差分析。**

大数定律

作为本节的结尾，我们列出蒙特卡洛方法的理论基石：辛钦大数定律。

> **定理 1 (辛钦大数定律)** 设 $\{\mathbf{x}_i\}_{i=1}^{n}$是独立同分布的的随机变量序列，$\mathbb{E}[\mathbf{x}_i] = \mu$，那么 $\{\mathbf{x}_i\}_{i=1}^{n}$服从大数定律，即对于任意的 $\epsilon > 0$,
>
> $$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i - \mu\right| < \epsilon\right) = 1.$$

因此，相合性和辛钦大数定律是一致的。

现在开始讨论本文的重点内容。

重要性采样

若干定义

根据前面的分析，需要引入一个新的分布 $q(\mathbb{x})$。为了保证问题是well-defined，设 $q(\mathbf{x}) \neq 0, \mathbf{x} \in \mathcal{D}$。

$$\mu=\mathbb{E}_{\mathbf{x}\sim p(\cdot)}[f(\mathbf{x})]=\int_{\mathcal{D}}f(\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x}=\int_{\mathcal{D}}q(\mathbf{x})f(\mathbf{x})\frac{p(\mathbf{x})}{q(\mathbf{x})}\mathrm{d}\mathbf{x}=\mathbb{E}_{\mathbf{x}\sim q(\cdot)}\left[f(\mathbf{x})\frac{p(\mathbf{x})}{q(\mathbf{x})}\right.$$

我们称 $\frac{p(\mathbf{x})}{q(\mathbf{x})}$为似然比(likehood ratio)， $q(\mathbb{x})$称为importance distribution (重要分布)。这个所谓的importance distribution，其实更像是一个调节权重的函数。

重要性采样估值

令 $\{\mathbf{x}_i\}_{i=1}^n\sim q(\cdot),\mathrm{i.i.d.}$，（即根据importance distribution q采样数据），根据如下数学期望，

$\mu=\mathbb{E}_{\mathbf{x}\sim q(\cdot)}\left[f(\mathbf{x})\frac{p(\mathbf{x})}{q(\mathbf{x})}\right]$，我们定义重要性采样估值 $\hat{\mu}_q$：

$$\hat{\mu}_q=:\frac{1}{n}\sum_{i=1}^{n}\frac{f(\mathbf{x}_i)p(\mathbf{x}_i)}{q(\mathbf{x}_i)}. \tag{2}$$

很显然， $\hat{\mu}_q$是 μ的一个无偏估计。

注意： 为了能计算重要性采样估值 $\hat{\mu}_q$，一个基本的前提是 $\frac{fp}{q}$是可以计算的。因此有些说法，比如" f,p是无法获得的，要选择一个简单的易获的分布 q，来估值"，这些关于importance sampling的说法都是不对的。

相合性

根据辛钦大数定律，不难验证$\hat{\mu}_q$是 μ的一个相合估计。

方差分析

本文第一重要结果：

> 定理 2 (重要性采样估值的方差) 令 $\{\mathbf{x}_i\}_{i=1}^n\sim q(\cdot),\mathrm{i.i.d.}$，重要性采样估值 $\hat{\mu}_q=:\frac{1}{n}\sum_{i=1}^{n}\frac{f(\mathbf{x}_i)p(\mathbf{x}_i)}{q(\mathbf{x}_i)}$,这里 $q(\mathbf{x})>0$。$\mu=\mathbb{E}_{\mathbf{x}\sim p(\cdot)}[f(\mathbf{x})]$。那么 $\mathbb{V}\mathrm{ar}_q[\hat{\mu}_q]=\frac{1}{n}\sigma_q^2$，其中 σ_q^2有如下两种表达：

> $$\sigma_q^2 = \mathbb{E}_{\mathbf{x}\sim q(\cdot)}\left[f^2(\mathbf{x})\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right)^2\right] - \mu^2;$$ 或者
>
> $$\sigma_q^2 = \mathbb{E}_{\mathbf{x}\sim q(\cdot)}\left[\left(\frac{f(\mathbf{x})p(\mathbf{x}) - \mu q(\mathbf{x})}{q(\mathbf{x})}\right)^2\right].$$

证明：这里罗列一些关键步骤。首先计算：

$$\mathbb{V}\mathrm{ar}[\hat{\mu}_q] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\mathrm{ar}_q\left[\frac{f(\mathbf{x}_i)p(\mathbf{x}_i)}{q(\mathbf{x}_i)}\right] = \frac{1}{n}\underbrace{\mathbb{V}\mathrm{ar}_{\mathbf{x}\sim q(\cdot)}\left[\frac{f(\mathbf{x})p(\mathbf{x})}{q(\mathbf{x})}\right];}_{=:\sigma_q^2} \quad (3$$

然后计算： σ_q^2。按照定义，可以获得定理中的两种关于 σ_q^2的等价表达（读者可以自行补充，不是很困难） 。

这个定理启发我们如何选择importance distribution q。

> **【启发1】** $\frac{p(\mathbf{x})}{q(\mathbf{x})}$不能太大。

根据第一个表达可知， σ_q^2的波动性取决于 $\mathbb{E}_{\mathbf{x}\sim q(\cdot)}\left[f^2(\mathbf{x})\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right)^2\right]$。如果 $\frac{p(\mathbf{x})}{q(\mathbf{x})}$非常大，那么Importance Sampling估值的方差就会很大，这样的 $q(\cdot)$不可取。最严重的是如果 $p(\cdot)$关于 $q(\cdot)$是重尾的(heavy tail)，那么Importance Sampling估值的方差就是无穷大： $\sigma_q^2 = +\infty$。本文这里重尾的定义是广义的： $\lim_{\|x\|\to\infty}\frac{p(\mathbf{x})}{q(\mathbf{x})} = \infty$。因此一个理想的importance distribution应该具备如下：(i)有界性 $\left|\frac{p(\mathbf{x})}{q(\mathbf{x})}\right| \le M, \mathbf{x}\in\mathcal{D}$，$f$的方差有界；或者 (ii) $p(\cdot)$是有界的， $q(\cdot)$是有下界的；这都能保证Importance Sampling估值的方差能被有效控制在一个有限的常数之内,这样的分布才有可能作为importance distribution去采样数据。

为了方便理解这个问题，我们举个例子。如果 $p(\cdot)=\mathcal{N}(0,1)$是高斯分布，一个常用做法是选取student分布作为 importance distribution： $q(\cdot)=\mathcal{T}(\nu)$。由于 $\lim_{\nu\to\infty}\mathcal{T}(\nu)=\mathcal{N}(0,1)$，见下图， $\nu=+\infty$的曲线就是标准正态分布的概率密度函数。因此 $\frac{p(\mathbf{x})}{q(\mathbf{x})}$总是有界的。这种情况也称为 p关于 q是轻尾的（light tail），看下图中尾部，这一点，也不难理解：随着随机变量的增大，$\frac{p(\mathbf{x})}{q(\mathbf{x})}\to 0$。

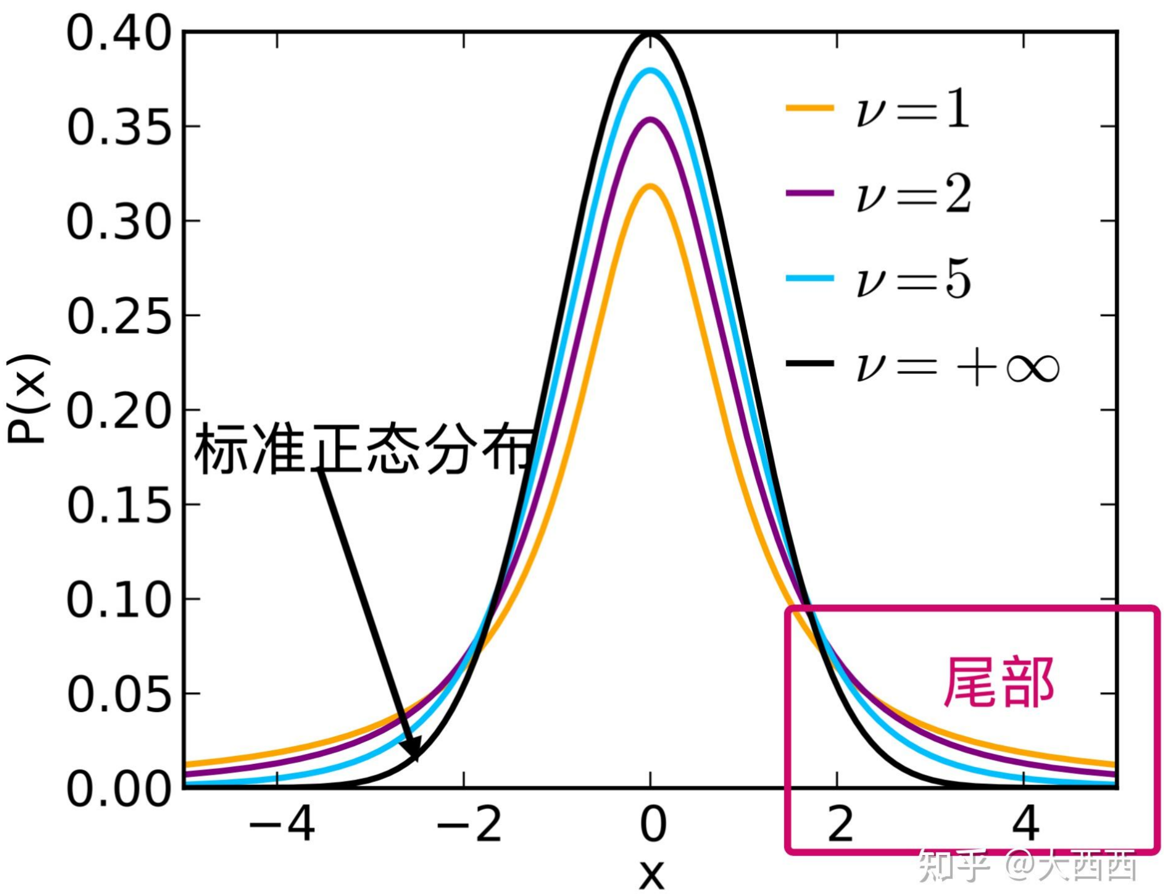


图1：Student分布的概率密度曲线。

但是，反过来，情况就发生突变了。如果 p是student分布， $q(\cdot)=\mathcal{N}(0,1)$，这就是所谓的$p$关于 q是重尾的(heavy tail)，于是 $\sigma_q^2=+\infty$（读者可以严格论证一下）。这样的 importance distribution q就很危险，要放弃。为了验证这个结论，图2展示了数值分析的结果， $\mathbb{P}[2<x<6]=\mathbb{E}_{x\sim\mathcal{C}(0,1)}[\mathbb{I}\{2<x<6\}]$，为了的到这个结果，我们用标准正态分布来采样数据。我们这里只考虑 $\nu=1$,此时的student分布就是 $\mathcal{C}(0,1)$。对于其他情况的 ν，可以做类似的讨论。

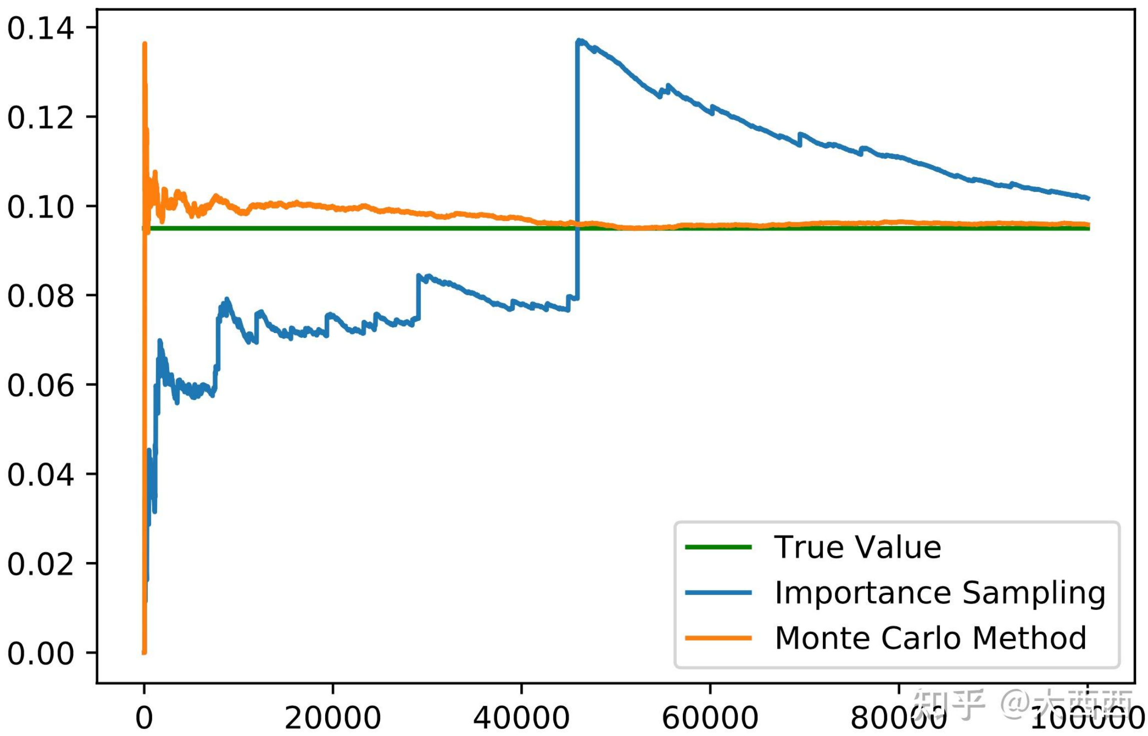


图2的结论表明：基于高斯分布采样数据来推断柯西分布的数学期望是很不靠谱的。方差很大，并且没有收敛到真值0.095。而经典的蒙特卡洛方法的收敛速度是比importance sampling方法好。图2中importance sampling方法第一次发生剧烈震荡是出现在高斯分布采样出来了一个点 $x=5.94$（采样到这个数值的概率很低很低，因为标准高斯分布有 99.73%的数值落在区间 $[-3，3]$以内），这个数值把importance radio的数值提升到了 $p/q=0.094$。回忆一下这个问题的估值公式：

$\hat{\mu}_q =: \frac{1}{n}\sum_{i=1}^{n}\frac{p(x_i)}{q(x_i)}\mathbb{I}\{2<x_i<6\}.$ 也就是说，在 数量$n=1000000$个的总样本中， $x=5.94$这个样本“发挥的作用/重要程度”是接近 $9.4\% \approx 10\%$。由此可以类推，图中每一个震荡的点都发挥了很大的作用。因此这个实验中，虽然有1000000个样本，但是实际上按照权重 p/q来看，可以认为只有少数几个样本给均值的估计贡献了力量。也就是说，大部分样本对最后的估值不起作用。比如说，采样到了真值 0.095，但是此时的权重 $p/q=0.0000001$（不一定真是数值这个，我设计得很小，有点夸张，只是为了方便读者理解），此时采样点0.095的实际贡献就不值一提。

这一段关于解释图2的震荡现象，具有普适性。事实上，当某一个样本 $\mathbf{x}_j$使得 $\dfrac{p(\mathbf{x}_j)}{q(\mathbf{x}_j)}$很大的时候，往往是由于这个样本 $\mathbf{x}_j$被采样到的概率很小很小；但是，由于 $\dfrac{p(\mathbf{x}_j)}{q(\mathbf{x}_j)}$数值比较大，这就大大增加了样本 $\mathbf{x}_j$对样本均值的“贡献”。也即是说，一个“合理”的采样 q至少能保证要在关于 p概率大的区域内尽可能多的采集样本，为什么？这是因为我们的目的是为了计算 $\int_{\mathcal{D}} f(\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x}$, 如果采样到的样本 $\mathbf{x}_j$，使得 $p(\mathbf{x}_j)$很小，那么这个样本对积分的作用就很弱，则这个样本可以视作无效的采样。再考虑到 $\dfrac{p(\mathbf{x}_j)}{q(\mathbf{x}_j)}$的数值比较大，那么这个样本 $\mathbf{x}_j$就会拉低前面有效的样本的贡献。最后，再考虑 $f(\mathbf{x}_j)\dfrac{p(\mathbf{x}_j)}{q(\mathbf{x}_j)}$，如果 $f(\mathbf{x}_j) > 0$，那么这个样本 $\mathbf{x}_j$就会剧烈地提升均值；如果 $f(\mathbf{x}_j) < 0$,那么这个样本 $\mathbf{x}_j$就会剧烈地拉低均值。这是估值不准确或者方差很大的根本原因。

但是大家也不用灰心，请看启发2。

> 【启发2】 精心的设计importance probabality q，可以提高经典蒙特卡洛积分的收敛速度。

例如，下图3是运用经典蒙特卡洛方法来计算标准正态的概率密度函数

$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} \mathrm{d}y,$ 即 $\hat{\Phi}_t = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{x_i \le t\},$ 这里 $\{x_i\}_{i=1:n} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$。

n/t	0.0	0.67	0.84	1.28	1.65	2.32	2.58	3.09	3.72
10^2	0.485	0.74	0.77	0.9	0.945	0.985	0.995	1	1
10^3	0.4925	0.7455	0.801	0.902	0.9425	0.9885	0.9955	0.9985	1
10^4	0.4962	0.7425	0.7941	0.9	0.9498	0.9896	0.995	0.999	0.9999
10^5	0.4995	0.7489	0.7993	0.9003	0.9498	0.9898	0.995	0.9989	0.9999
10^6	0.5001	0.7497	0.8	0.9002	0.9502	0.99	0.995	0.999	0.9999
10^7	0.5002	0.7499	0.8	0.9001	0.9501	0.99	0.995	0.999	0.9999
10^8	0.5	0.75	0.8	0.9	0.95	0.99	0.995	0.999	0.9999

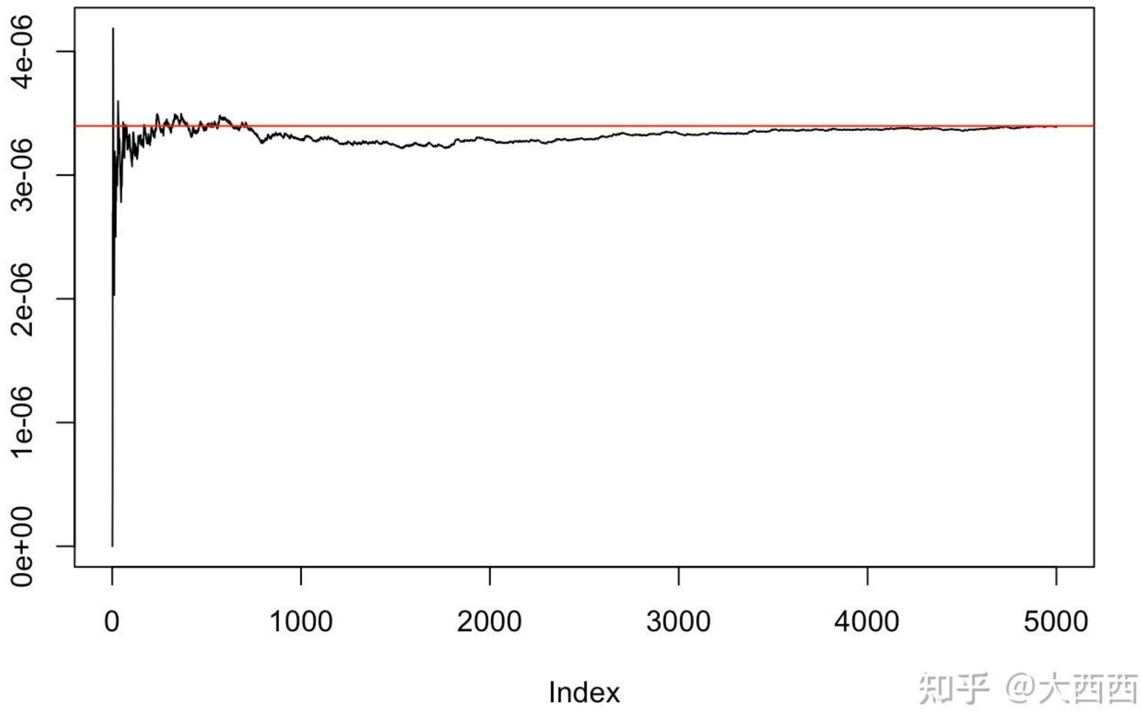
图3: 经典蒙特卡洛方法计算计算标准正态的概率密度函数数值表

因此，为了计算尾部概率 $\mathbb{P}(X>4.5), X\sim\mathcal{N}(0,1)$，根据图中的数据可以，至少需要 $10^6\sim 10^8$个样本。

如果考虑截断指数分布作为importance distribution q来采集样本，此时

$q(x)=\dfrac{e^{-y}}{\int_{4.5}^{\infty}e^{-x}\mathrm{d}x}$，那么基于上述 q的importance sampling估值是

$\hat{\mathbb{P}}(X>4.5)=\dfrac{1}{n}\sum_{i=1}^{n}\dfrac{p(x_i)}{q(x_i)}=\dfrac{1}{n}\sum_{i=1}^{n}\dfrac{\exp\{-x_i^2/2+x_i-4.5\}}{\sqrt{2\pi}}$，其中 $\{x_i\}_{i=1:n}\overset{\text{i.i.d.}}{\sim}q$。



数值结果表明，只需5000个样本，就能有效地估计出 $\mathbb{P}(X>4.5)$。

如何解释呢？这里的情况和前面正好相反。$\mathbb{P}(X>4.5), X\sim\mathcal{N}(0,1)$，这个概率的数值很低很低，几乎等于零，因此，如果要用经典的蒙特卡洛，那么必然要采很多很多数据才会出现样本 $\mathbf{x}_j>4.5$，然后估值。如果用截断指数分布作为importance distribution q来采集样本，那么出现 $\mathbf{x}_j>4.5$的样本就大大增加了。读者可以顺着这个思路，思考一下，为什么就加速了收敛？道理和【启发1】是一样的，如果能完全理解【启发1】，应该能理解这里加速的原因。

> 【启发3】 最优采样的大致模样

根据方差的第二个表达式

$$\sigma_q^2 = \mathbb{E}_{\mathbf{x}\sim q(\cdot)}\left[\left(\frac{f(\mathbf{x})p(\mathbf{x}) - \mu q(\mathbf{x})}{q(\mathbf{x})}\right)^2\right],$$ 如果 $f > 0$恒成立，那么使得上述方差等于零的最优importance distribution $q_\star(\mathbf{x}) \propto f(\mathbf{x})p(\mathbf{x})$;

如果 $f < 0$恒成立，那么使得上述方差等于零的最优importance distribution $q_\star(\mathbf{x}) \propto -f(\mathbf{x})p(\mathbf{x})$.

读者可以自行证明一下这个小结论。虽然方差等于零的估值没有意义，但是这能告诉人们，最优采样 $q_\star$大致的模样，正比于 $|f(\mathbf{x})|p(\mathbf{x})$，而不是 $\sqrt{f(\mathbf{x})p(\mathbf{x})}$ 或者 $f(\mathbf{x})p^2(\mathbf{x})$等等。下一节，给出这个结果的证明与解释。

一致最小方差估计

> 定理 （最优的重要性采样）令 $q_\star(\mathbf{x}) =: \frac{|f(\mathbf{x})|p(\mathbf{x})}{\mathbb{E}_{\mathbf{x}\sim p(\cdot)}[|f(\mathbf{x})|]}$,则 $\sigma_{q_\star}^2 \leq \sigma_q^2$对任意分布 q都成立。

证明：有两种证明方法。

第一种方法：把上述最优importance distribution $q_\star$直接带入定理2的方差中，然后运用柯西不等式，可以证明

$$\sigma_{q_\star}^2 \leq \sigma_q^2, \ \forall q.$$ 第二方法：直接计算出这个 $q_\star$:

解下述变分问题：

$$q_\star(\cdot) = \arg\min_q \left\{\int_{\mathbf{x}\in\mathcal{D}} \frac{(f(\mathbf{x})p(\mathbf{x}))^2}{q(\mathbf{x})}\mathrm{d}\mathbf{x}, \quad \text{such that} \int_{\mathbf{x}\in\mathcal{D}} q(\mathbf{x})\mathrm{d}\mathbf{x} = 1\right\}.$$

为了得到上述结果，构造Lagrange乘子：

$$\mathcal{L}(q,\lambda) =: \int_{\mathbf{x}\in\mathcal{D}} (f(\mathbf{x})p(\mathbf{x}))^2/q(\mathbf{x})\mathrm{d}\mathbf{x} + \lambda\left(\int_{\mathbf{x}\in\mathcal{D}} q(\mathbf{x})\mathrm{d}\mathbf{x} - 1\right);$$ 计算一阶稳定点 $(q_\star, \lambda_\star)$;

计算最优importance distribution

$q_\star(\mathbf{x}) = \sqrt{\frac{f^2(\mathbf{x})p^2(\mathbf{x})}{\lambda_\star}} \propto |f(\mathbf{x})|p(\mathbf{x});$最后和前面一样，然后运用柯西不等式证明 $\sigma^2_{q_\star} \leq \sigma^2_q,\ \forall q.$