Lecture 15 Monte Carlo integration

Weinan E[1,2] and Tiejun Li[2]

[1]Department of Mathematics,
Princeton University,
weinan@princeton.edu

[2]School of Mathematical Sciences,
Peking University,
tieli@pku.edu.cn
No.1 Science Building, 1575

Outline

Monte Carlo methods: basics

Variance reduction methods

An introduction to Markov chain

Monte Carlo方法

- Monte Carlo方法近似$I(f) = \int_0^1 f(x)dx$:

$$I(f) \approx \frac{1}{N}\sum_{i=1}^{N} f(X_i) \triangleq I_N(f)$$

  这里X_i $(i = 1, 2, \ldots, N)$为独立同$[0, 1]$区间均匀分布的随机变量（以后简记为$i.i.d.$ $\mathcal{U}[0, 1]$）。

- $I_N(f)$的性质:

  显然$I_N(f)$收敛到$I(f)$，$I_N(f)$作为一个随机变量有性质

$$\begin{aligned} \mathbb{E}I_N(f) &= \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} f(X_i)\right) = \frac{1}{N}\sum_{i=1}^{N}\int_0^1 f(x)dx \\ &= I(f) \end{aligned}$$

Monte Carlo方法

- $I_N(f)$的性质:
  此时误差估计$e_N = |I_N(f) - I(f)|$仍旧是随机变量，我们估计其“均方误差”：

$$\begin{aligned}
\mathbb{E}|e_N|^2 &= \mathbb{E}(I_N(f) - I(f))^2 = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}(f(X_i) - I(f))\right)^2 \\
&= \frac{1}{N^2}\sum_{i,j=1}^{N}\mathbb{E}(f(X_i) - I(f))(f(X_j) - I(f)) \\
&= \frac{1}{N}\mathbb{E}(f(X_i) - I(f))^2 = \frac{1}{N}\mathrm{Var}(f) \qquad (1)
\end{aligned}$$

  $\mathrm{Var}(f)$为$f(X)$的方差。由Schwartz不等式:

$$\mathbb{E}|e_N| \le \sqrt{\mathbb{E}|e_N|^2} = \sqrt{\frac{\mathrm{Var}(f)}{N}} \qquad (2)$$

  如果$f(X)$有有限方差，则Monte Carlo方法在如上意义下具有半阶收敛性。

MC方法的半阶收敛性

例：考察下面的数值例子：

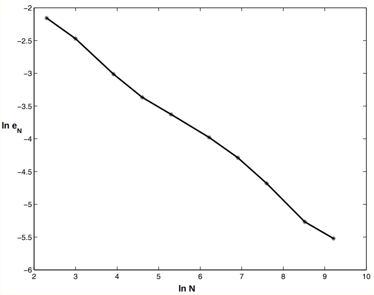
$$\int_0^{\frac{\pi}{2}} \sin x dx = 1 \tag{3}$$

由Monte Carlo方法：

$$\int_0^{\frac{\pi}{2}} \sin x dx \approx \frac{1}{N}\sum_{i=1}^{N} \frac{\pi}{2} \sin(\frac{\pi}{2} X_i) \tag{4}$$

取$X_i(i = 1, 2, \ldots, N)$为$i.i.d.\ \mathcal{U}[0, 1]$随机变量，固定$m = 100$，分别取$N = 10$, $20, 50, 100$,$200, 500$, $1000, 2000, 5000, 10000$，定义$e_N^j(j = 1, \ldots, m)$为样本量N固定的情形下m次独立计算的误差，$e_N = \frac{1}{m}\sum_{j=1}^{m} e_N^j$，以横坐标$\ln N$，纵坐标$\ln e_N$绘图如下图所示，从中不难看出半阶收敛速度。

MC方法的半阶收敛性



MC方法的半阶收敛性

MC方法的特性

- MC方法的单调性:
  MC方法的特点,随着N的增加,数值解的误差并不会单调下降!因为它的半阶收敛性是一个统计平均的结果,这一点与确定型格式是完全不同的.确定型格式往往具有单调性.
- 例子: MC方法对
$$\int_0^{\frac{\pi}{2}} \sin x dx = 1 \tag{5}$$
  的计算.

Outline

Monte Carlo methods: basics

Variance reduction methods

An introduction to Markov chain

MC方法的特性

- 我们已经了解到，Monte Carlo方法的求积误差由

$$\frac{\sigma}{\sqrt{N}}$$

来表征，这里N表示样本数量，$\sigma = (\mathrm{Var}(f))^{\frac{1}{2}}$为标准差。

- 为提高求积精度，第一种选择是将N增大，即增加样本的数量，但由于Monte Carlo方法只是半阶收敛，要得到较高精度，只有大量的增大运算量。第二种选择是通过一些理论技巧来减小方差，这正是本讲的主要内容。

重要性抽样

- 基本的Monte Carlo方法算法是产生$X_i \sim i.i.d.\ \mathcal{U}[0,1]$，于是

$$I(f) \approx I_N(f) \triangleq \frac{1}{N}\sum_{i=1}^{N} f(X_i) \tag{6}$$

- 但是我们对$I(f)$有另一种看法：

$$I(f) = \int_0^1 f(x)dx = \int_0^1 \frac{f(x)}{p(x)}p(x)dx \tag{7}$$

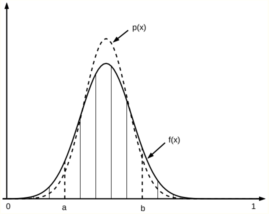
这里$p(x)$是$[0,1]$上的一个概率密度，即$\int_0^1 p(x)dx = 1, p(x) > 0$。这样我们得到了另一种形式的积分近似：

$$I(f) \approx \frac{1}{N}\sum_{i=1}^{N} \frac{f(Y_i)}{p(Y_i)} \tag{8}$$

这里$Y_i, i = 1, \dots, N$是遵循分布密度为$p(y)$的$i.i.d.$随机变量。

重要性抽样

- 以下图为例我们说明重要性抽样带来的好处:



- 设$f(x)$的值较大的部分主要集中于如图所示的$[a,b]$区域，因为$X_i \sim i.i.d.\ \mathcal{U}[0,1]$，会有相当数目的$\{X_j\}$在$[a,b]$之外，即计算过程中，有相当数目的点为计算$\int_0^a + \int_b^1 f(x)dx$去作贡献，而$\int_0^a + \int_b^1 f(x)dx$只占全部$I(f)$的很小的比例. 即我们花了相当比例的运算量去完成对最终数值结果贡献不大的部分，其效率必定低下—精度较差!
- 如果用新的方式去求$I_N(f)$，而且$p(x)$的增长、下降以及高度趋势和$f(x)$类似，可以期望，更多的点出现在$f(x)$的峰值区，从而提高计算精度，这也是“重要性”抽样这一名称的意义：$p(x)$反映了$f(x)$的“重要部分”。

重要性抽样

- 思想上，重要性抽样即是“自适应方法的随机版本”.
- 理论上，我们从方差角度分析：

$$\mathrm{Var}_X(f)=\int_0^1 (f-I(f))^2dx=\int_0^1 f^2dx-I^2(f) \tag{9}$$

$$\mathrm{Var}_Y(\frac{f}{p})=\int_0^1 (\frac{f}{p})^2pdy-I^2(f)=\int_0^1 \frac{f^2}{p}(y)dy-I^2(f) \tag{10}$$

如果适当选取$p(y)$，使得$\int_0^1 \frac{f^2}{p}(y)dy<\int_0^1 f^2dx$，则方差得到减少！特别的，如果$p(x)=\frac{f(x)}{I(f)}$，则$\mathrm{Var}_Y(\frac{f}{p})=0$！即当$p(x)$的重要性选得与$f(x)$完全一样时，方差成为0，我们得到了精确积分值！

- 但注意我们要获得$p(x)$的前提是已知精确积分值$I(f)$，这是不可能的。
- 重要性抽样对实际计算有根本思想性的指导作用，它依赖于我们对一个问题先验的了解程度并以之构造适当的抽样函数。

控制变量法

- 控制变量法的思想是利用一个性质已知的随机变量去“控制”另一个随机变量。具体而言，考察如下表达式：

$$\int_0^1 f(x)dx = \int_0^1 f(x) - g(x)dx + \int_0^1 g(x)dx \tag{11}$$

利用Monte Carlo方法：

$$I_N(f) = \frac{1}{N}\sum_{i=1}^{N}(f(X_i) - g(X_i)) + I(g) \tag{12}$$

这里$X_i \sim i.i.d.\,\mathcal{U}[0,1]$，$I(g)$是已知的。如果$\mathrm{Var}(f-g) \le \mathrm{Var}(f)$，我们得到了一个方差减小的方法。

- 显然当$f = g$时，$\mathrm{Var}(f-g) = 0$，即得到精确积分值，但这时需要事前已知$I(g)$，和重要性抽样方法一样这是不可能的.

控制变量法

- 下面给出一个具体例子说明控制变量的作用:
- **例:** 求积:

$$\begin{cases} I(f) & = & \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}(1+r)^{-1}e^{-\frac{x^2}{2}}dx \\ r & = & e^{\sigma x} \qquad (\text{常数}\sigma > 0) \end{cases} \tag{13}$$

- 注意到

$$(1+r)^{-1} \approx \begin{cases} 1 & x \leq 0 \\ 0 & x > 0 \end{cases} \triangleq h(x) \tag{14}$$

$$I(f) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} ((1+r)^{-1} - h(x))e^{-\frac{x^2}{2}}dx + \frac{1}{2} \tag{15}$$

此时$h(x)$即起到了控制变量的作用，然后可用标准正态分布作重要性抽样计算第一项的积分项.

Outline

Monte Carlo methods: basics

Variance reduction methods

An introduction to Markov chain

Random walk

- Stochastic process is a parameterized random variables $\{X_n\}_{n\in\mathbb{N}}$ (discrete time stochastic process).
- Example (1D random walk):
  Let ξ_i are $i.i.d.$ random variables such that $\xi_i = \pm 1$ with probability $\frac{1}{2}$, and let

$$X_n = \xi_1 + \xi_2 + \ldots + \xi_n$$

  $\{X_n\}$ represents a unconstrained unbiased random walk on $\mathbb{Z}$, the set of integers. Given $X_n = i$, we have

$$\begin{aligned} P\{X_{n+1} = i \pm 1 |\ X_n = i\} &= \frac{1}{2}, \\ P\{X_{n+1} = \text{anything else} |\ X_n = i\} &= 0. \end{aligned}$$

  We see that the distribution of X_{n+1} depends only on the value of X_n.

Markov chain

- The result above can be restated as the *Markov property*

$$P\{X_{n+1} = i_{n+1} | \{X_m = i_m\}_{m=1}^{n}\} = P\{X_{n+1} = i_{n+1} | X_n = i_n\},$$

and the sequence $\{X_n\}_{n=1}^{\infty}$ is called a realization of a Markov process.

- Stationary Markov chain:

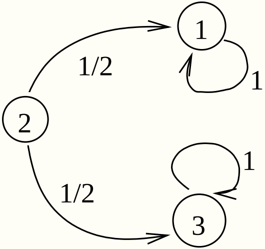
A Markov chain is called stationary if the **transition probabilities**

$$p_{jk}^{n} = P\{X_{n+1} = k | X_n = j\} = p_{jk}$$

independent of n. From now on we will discuss only stationary Markov chains and let $P = (p_{jk})_{j,k=1}^{N}$. P is called the transition probability matrix(TPM).

Graph representation

- Graph representation of Markov chains
  Any Markov chain can be sketched by their graph representation. Example 1:

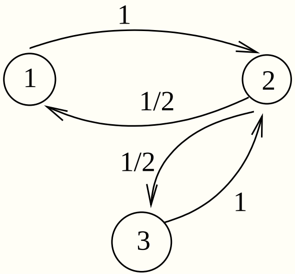


- The arrows and real numbers show the transition probability of the Markov chain. The TPM corresponds to this Figure is

$$P = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 \end{bmatrix},$$

Graph representation

- Example 2:



- The TPM corresponds to this Figure is

$$P = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \end{bmatrix}.$$

General properties

- P is also called a *stochastic matrix*, in the sense that

$$p_{ij} \geq 0, \quad \sum_{j=1}^{N} p_{ij} = 1.$$

- Given the initial distribution of the Markov chain μ_0, the distribution of X_n is then given by

$$\mu_n = \mu_0 P^n$$

- μ_n satisfies the recurrence relation $\mu_n = \mu_{n-1} P$. This equation can also be rewritten as

$$(\mu_n)_i = (\mu_{n-1})_i (1 - \sum_{j \neq i} p_{ij}) + \sum_{j \neq i} (\mu_{n-1})_j p_{ji}.$$

The interpretation is clear.

Invariant distribution

- π is called an invariant distribution if

$$\pi = \pi P$$

  This is equivalent to say that there exists a nonnegative left eigenvector of P with eigenvalue equal to 1.

- The meaning of invariant distribution is quite clear. It is a "equilibrium point" of the Markov chain.

References

- R.E. Caflish, Monte Carlo and Quasi-Monte Carlo methods, Acta Numerica, Vol. 7, 1-49, 1998.
- 汪仁官, 初等概率论, 北京大学出版社.