Big Data Analysis
(MA60306)

Bibhas Adhikari

Spring 2022-23, IIT Kharagpur

Lecture 23
March 24, 2023

Sampling methods

Proposition The probability that $X \sim g$ is accepted is $\frac{1}{c}$, and is maximized when $c = \sup_x \frac{f(x)}{g(x)}$

Sampling methods

Proposition The probability that $X \sim g$ is accepted is $\frac{1}{c}$, and is maximized when $c = \sup_x \frac{f(x)}{g(x)}$

Proof

$$
\begin{aligned}
P\left(U \leq \frac{f(X)}{cg(X)}\right) &= \int_{-\infty}^{\infty} \int_{0}^{\frac{f(t)}{cg(t)}} g(t)du\,dt \\
&= \int_{-\infty}^{\infty} \frac{f(t)}{cg(t)} g(t)dt \\
&= \int_{-\infty}^{\infty} \frac{f(t)}{c} dt = \frac{1}{c}
\end{aligned}
$$

Sampling methods

Proposition The probability that $X \sim g$ is accepted is $\frac{1}{c}$, and is maximized when $c = \sup_x \frac{f(x)}{g(x)}$

Proof

$$
\begin{aligned}
P\left(U \leq \frac{f(X)}{cg(X)}\right) &= \int_{-\infty}^{\infty} \int_{0}^{\frac{f(t)}{cg(t)}} g(t) du\, dt \\
&= \int_{-\infty}^{\infty} \frac{f(t)}{cg(t)} g(t) dt \\
&= \int_{-\infty}^{\infty} \frac{f(t)}{c} dt = \frac{1}{c}
\end{aligned}
$$

Efficiency Since $\sup_x \frac{f(x)}{g(x)} = \sqrt{\frac{2e}{\pi}}$ is the acceptance rate, by the above result, it would be $\sqrt{\frac{\pi}{2e}} = 0.7602$. Thus if we generate 100 x-values from g, we can expect 75 of them would be retained, and the others discarded.

Sampling methods

Question What is the acceptance rate when the standard normal is sampled using Accept-Reject method through the standard Cauchy density

$$g(x) = \frac{1}{\pi(1 + x^2)}$$

Sampling methods

Question What is the acceptance rate when the standard normal is sampled using Accept-Reject method through the standard Cauchy density

$$g(x) = \frac{1}{\pi(1 + x^2)}$$

Question What is the geometry of the Accept-Reject method?

Sampling methods

Question What is the acceptance rate when the standard normal is sampled using Accept-Reject method through the standard Cauchy density

$$g(x) = \frac{1}{\pi(1 + x^2)}$$

Question What is the geometry of the Accept-Reject method?
Question How to choose $g(x)$?

Sampling methods

Question What is the acceptance rate when the standard normal is sampled using Accept-Reject method through the standard Cauchy density

$$g(x) = \frac{1}{\pi(1 + x^2)}$$

Question What is the geometry of the Accept-Reject method?

Question How to choose $g(x)$?

Some more algorithms for standard distributions

> Standard Exponential Generate $X \sim exp(1)$: generate $U(0, 1)$ and use $X = -\log U$

Sampling methods

Question What is the acceptance rate when the standard normal is sampled using Accept-Reject method through the standard Cauchy density

$$g(x) = \frac{1}{\pi(1 + x^2)}$$

Question What is the geometry of the Accept-Reject method?
Question How to choose $g(x)$?
Some more algorithms for standard distributions

Standard Exponential Generate $X \sim exp(1)$: generate $U(0, 1)$ and use $X = -\log U$

Gamma with parameters n and λ generate $X \sim G(n, \lambda)$, first generate n independent values $X_1, \ldots, X_n$ from a standard exponential, and use $X = \lambda(X_1 + \ldots + X_n)$

Sampling methods

Question What is the acceptance rate when the standard normal is sampled using Accept-Reject method through the standard Cauchy density

$$g(x) = \frac{1}{\pi(1+x^2)}$$

Question What is the geometry of the Accept-Reject method?

Question How to choose $g(x)$?

Some more algorithms for standard distributions

Standard Exponential Generate $X \sim exp(1)$: generate $U(0,1)$ and use $X = -\log U$

Gamma with parameters n and λ generate $X \sim G(n, \lambda)$, first generate n independent values $X_1, \ldots, X_n$ from a standard exponential, and use $X = \lambda(X_1 + \ldots + X_n)$

Beta with parameters m, n To generate $X \sim Be(m, n)$, generate $U \sim G(m, 1)$, $V \sim G(n, 1)$ independently, and use $X = \frac{U}{U+V}$

Sampling methods

t distribution with n degrees of freedom To generate $X \sim t(n)$, first generate $Z_1, Z_2, \ldots, Z_{n+1} \sim \mathcal{N}(0,1)$ independently, and use

$$X = \frac{Z_1}{\sqrt{\frac{Z_2^2+\ldots+Z_{n+1}^2}{n}}}$$

Sampling methods

t distribution with n degrees of freedom To generate $X \sim t(n)$, first generate $Z_1, Z_2, \ldots, Z_{n+1} \sim \mathcal{N}(0,1)$ independently, and use

$$X = \frac{Z_1}{\sqrt{\frac{Z_2^2+\ldots+Z_{n+1}^2}{n}}}$$

Wishart with general parameters To generate $S \sim W_p(n, \Sigma)$, generate $X_1, X_2, \ldots, X_n \sim \mathcal{N}_p(0, \Sigma)$ independently, and use

$$S = \sum_{i=1}^{n} X_i X_i^T$$

Uniform on the surface of the unit ball Generate $Z_1, \ldots, Z_d \sim \mathcal{N}(0,1)$ independently, and use

$$X_i = \frac{Z_i}{\sqrt{Z_1^2 + \ldots + Z_d^2}}, i = 1, \ldots, d$$

Sampling methods

Markov Chain Monte Carlo (MCMC)

→ The standard simulation techniques are usually difficult to apply, for example, when the target distribution is an unconventional one, or even worse, it is known only up to a normalizing constant such as:

$$f(x) = \frac{h(x)}{c}$$

for some explicit function h, but only c an implicit normalizing constant c because it cannot be computed exactly, or even to a high degree of accuracy

Sampling methods

→ For example, the problem of simulating from posterior densities of a parameter(s)

$$\pi(\theta | x) = \frac{f(x | \theta)\pi(\theta)}{m(x)}$$

where $f(x| \theta)$ is the likelihood function, $\pi(\theta)$ is the prior density, and $m(x)$ is the marginal density density of the observable X induced by (f, π). Thus

$$m(x) = \int_{\Theta} f(x| \theta)\pi(\theta)d\theta$$

Sampling methods

$\rightarrow$ For example, the problem of simulating from posterior densities of a parameter(s)

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)}$$

where $f(x|\theta)$ is the likelihood function, $\pi(\theta)$ is the prior density, and $m(x)$ is the marginal density density of the observable X induced by (f, π). Thus

$$m(x) = \int_{\Theta} f(x|\theta)\pi(\theta)d\theta$$

If the parameter θ is high-dimensional, and the prior density $\pi(\theta)$ is not a very conveniently chosen one, then $m(x)$ usually cannot be calculated in closed-form, or even to a high degree of numerical approximation.

All the simulation methods discussed in the previous section are useless in such a situation

Sampling methods

Markov chain Monte Carlo

Graph: A pair of sets $G = (V, E)$, $E \subseteq E \times E, V \neq \emptyset$. Two vertices $v_i, v_j \in V$ are adjacent if $(v_i, v_j) \in E$.

[1] Klafter, J. and Sokolov, I.M., 2011. First steps in random walks: from tools to applications. Oxford University Press.

Sampling methods

Markov chain Monte Carlo

Graph: A pair of sets $G = (V, E)$, $E \subseteq E \times E, V \neq \emptyset$. Two vertices $v_i, v_j \in V$ are adjacent if $(v_i, v_j) \in E$.

Walk: A sequence of vertices and edges

$$v_1, e_1, v_2, e_2, \ldots, v_k, e_k, v_{k+1}$$

such that end points of e_i are v_i and v_{i+1}, $1 \leq i \leq k$

[1] Klafter, J. and Sokolov, I.M., 2011. First steps in random walks: from tools to applications. Oxford University Press.

Sampling methods

Markov chain Monte Carlo

Graph: A pair of sets $G = (V, E)$, $E \subseteq E \times E, V \neq \emptyset$. Two vertices $v_i, v_j \in V$ are adjacent if $(v_i, v_j) \in E$.

Walk: A sequence of vertices and edges

$$v_1, e_1, v_2, e_2, \ldots, v_k, e_k, v_{k+1}$$

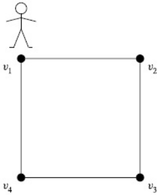
such that end points of e_i are v_i and v_{i+1}, $1 \leq i \leq k$

The problem of a random walk [1] was posed by Louis Bachelier in his thesis devoted to the theory of financial speculations in 1900. The term random walks was first introduced by Karl Pearson in 1905.

[1] Klafter, J. and Sokolov, I.M., 2011. First steps in random walks: from tools to applications. Oxford University Press.

Sampling methods

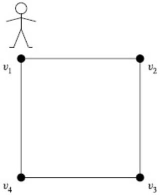
For each t, let X_t denote the index of the vertex at which the walker resides. Hence $\{X_0, X_1, \ldots\}$ is a stochastic process (Markov Chain) taking values in $\{1, 2, 3, 4\}$ stands at time t



At time 0, the random walker stands at v_1 : $p(X_0 = 1) = 1$

At time 1, flips a fair coin and moves immediately to v_2 or v_4 according to whether the coin comes up heads or tails: $p(X_1 = v_2) = \frac{1}{2} = p(X_1 = v_4)$.

Sampling methods

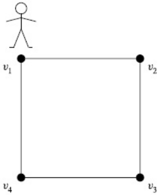


Markov property

$$
\begin{aligned}
p(X_{t+1} = v_1 | X_0 = i_0, X_1 = i_1, \ldots, X_{t-1} = i_{t-1}, X_t = v_2) &= \frac{1}{2} \\
p(X_{t+1} = v_3 | X_0 = i_0, X_1 = i_1, \ldots, X_{t-1} = i_{t-1}, X_t = v_2) &= \frac{1}{2}
\end{aligned}
$$

for any choice of $i_0, \ldots, i_{t-1}$.

Sampling methods



Markov property

$$
\begin{aligned}
p(X_{t+1} = v_1 | X_0 = i_0, X_1 = i_1, \ldots, X_{t-1} = i_{t-1}, X_t = v_2) &= \frac{1}{2} \\
p(X_{t+1} = v_3 | X_0 = i_0, X_1 = i_1, \ldots, X_{t-1} = i_{t-1}, X_t = v_2) &= \frac{1}{2}
\end{aligned}
$$

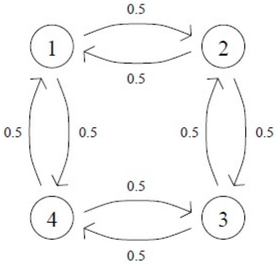
for any choice of $i_0, \ldots, i_{t-1}$.

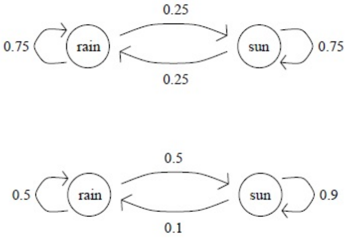
Time homogeneity

$$p(X_{t+1} | X_t = v) = c$$

for all t, for any $v \in V$

Sampling methods

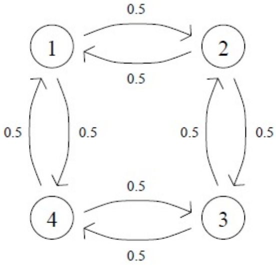


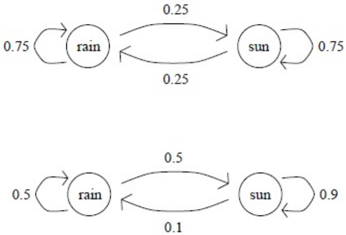


Markov chain A sequence of random variables $\{X_n\}, n \geq 0$, is said to be a Markov chain if for some countable set $S \subset \mathbb{R}$, and any $n \geq 1$, $s_{n+1}, s_n, \ldots, s_0 \in S$,

$$P(X_{n+1} = s_{n+1} | X_0 = s_0, \ldots, X_n = s_n) = P(X_{n+1} = s_{n+1} | X_n = s_n)$$

Sampling methods



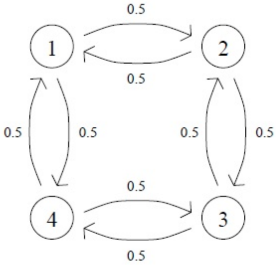


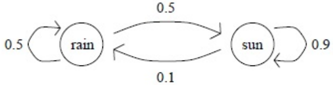
Markov chain A sequence of random variables $\{X_n\}, n \geq 0$, is said to be a Markov chain if for some countable set $S \subset \mathbb{R}$, and any $n \geq 1$, $s_{n+1}, s_n, \ldots, s_0 \in S$,

$$P(X_{n+1} = s_{n+1} | X_0 = s_0, \ldots, X_n = s_n) = P(X_{n+1} = s_{n+1} | X_n = s_n)$$

The set S is called the state space of the chain. If S is a finite set, the chain is called a finite state Markov chain. X_0 is called the initial state.

Sampling methods





Markov chain A sequence of random variables $\{X_n\}, n \geq 0$, is said to be a Markov chain if for some countable set $S \subset \mathbb{R}$, and any $n \geq 1$, $s_{n+1}, s_n, \ldots, s_0 \in S$,

$$P(X_{n+1} = s_{n+1} | X_0 = s_0, \ldots, X_n = s_n) = P(X_{n+1} = s_{n+1} | X_n = s_n)$$

The set S is called the state space of the chain. If S is a finite set, the chain is called a finite state Markov chain. X_0 is called the initial state.

Example? the indices can be treated as time and the rvs as the observation of a process: Surfing Webpages, Weather prediction

Random walks

initial distribution the distribution of the initial state X_0, which tells us how the Markov chain starts

Random walks

initial distribution the distribution of the initial state X_0, which tells us how the Markov chain starts

Distribution of the Markov Chain Let $\mu^{(0)}$ denote the initial distribution of the Markov chain, defined as

$$\mu^{(0)} = (p(X_0 = s_1), p(X_0 = s_2), \ldots, p(X_0 = s_k)),$$

with state space $S = \{s_1, \ldots, s_k\}$. Similarly,

$$\mu^{(t)} = (p(X_t = s_1), p(X_t = s_2), \ldots, p(X_t = s_k)),$$

Random walks

initial distribution the distribution of the initial state X_0, which tells us how the Markov chain starts

Distribution of the Markov Chain Let $\mu^{(0)}$ denote the initial distribution of the Markov chain, defined as

$$\mu^{(0)} = (p(X_0 = s_1), p(X_0 = s_2), \ldots, p(X_0 = s_k)),$$

with state space $S = \{s_1, \ldots, s_k\}$. Similarly,

$$\mu^{(t)} = (p(X_t = s_1), p(X_t = s_2), \ldots, p(X_t = s_k)),$$

For example, $\mu^{(0)} = (1, 0, 0, 0)$, $\mu^{(1)} = (0, \frac{1}{2}, 0, \frac{1}{2})$ in the previous example.