Big Data Analysis
(MA60306)

Bibhas Adhikari

Spring 2022-23, IIT Kharagpur

Lecture 22
March 23, 2023

Sampling methods

Sampling from standard normal Suppose $T : \mathbb{R}^n \to \mathbb{R}^n$ is a differentiable transformation. Then the Jacobian of the transformation T, denoted by J_T is defined as the determinant of the matrix whose ij-th entry is $\frac{\partial T_i}{\partial x_j}(x)$

Sampling methods

Sampling from standard normal Suppose $T : \mathbb{R}^n \to \mathbb{R}^n$ is a differentiable transformation. Then the Jacobian of the transformation T, denoted by J_T is defined as the determinant of the matrix whose ij-th entry is $\frac{\partial T_i}{\partial x_j}(x)$

If $y = Tx$ has the inverse $T^{-1}(y)$ then we can prove that

$$J_{T^{-1}}(y) = \frac{1}{J_T(T^{-1}y)}$$

i.e. the matrix of partial derivatives with entries $\frac{\partial T_i}{\partial x_j}(x)$ is invertible, and its inverse is the matrix of partial derivatives whose entries are $\frac{\partial T_i^{-1}}{\partial x_j}(y)$

Sampling methods

Sampling from standard normal Suppose $T : \mathbb{R}^n \to \mathbb{R}^n$ is a differentiable transformation. Then the Jacobian of the transformation T, denoted by J_T is defined as the determinant of the matrix whose ij-th entry is $\frac{\partial T_i}{\partial x_j}(x)$

If $y = Tx$ has the inverse $T^{-1}(y)$ then we can prove that

$$J_{T^{-1}}(y) = \frac{1}{J_T(T^{-1}y)}$$

i.e. the matrix of partial derivatives with entries $\frac{\partial T_i}{\partial x_j}(x)$ is invertible, and its inverse is the matrix of partial derivatives whose entries are $\frac{\partial T_i^{-1}}{\partial x_j}(y)$

Change of variables for integration Let $y = T(x)$, $x \in S \subseteq \mathbb{R}^n$ be a differentiable one-to-one transformation from S to $T(S)$. Then

$$\int_S f(x)dx = \int_{T^{-1}(S)} f(T^{-1}(y))J_{T^{-1}}(y)dy$$

Sampling methods

Let X and Y be independent standard normal variables and consider the vector $(X, Y) \in \mathbb{R}^2$.

Sampling methods

Let X and Y be independent standard normal variables and consider the vector $(X, Y) \in \mathbb{R}^2$. Define the transformation

$$x = \sqrt{d}\cos\theta, y = \sqrt{d}\sin\theta, d > 0$$

Now since X, Y are independent, then the joint distribution of (X, Y) is

$$f(x, y) = \frac{1}{2\pi}e^{-(x^2+y^2)/2}$$

Sampling methods

Let X and Y be independent standard normal variables and consider the vector $(X, Y) \in \mathbb{R}^2$. Define the transformation

$$x = \sqrt{d}\cos\theta, y = \sqrt{d}\sin\theta, d > 0$$

Now since X, Y are independent, then the joint distribution of (X, Y) is

$$f(x, y) = \frac{1}{2\pi}e^{-(x^2+y^2)/2}$$

Since $d = x^2 + y^2$, and $J_T(x, y) = 1/2$, then from the change-of-variables formula that D and Θ have the joint distribution

$$f(d, \theta) = \frac{1}{2\pi} \cdot \frac{1}{2}e^{-d/2}$$

Sampling methods

This implies that D has the exponential distribution with $\lambda = 1/2$, and Θ has the uniform distribution over $[0, 2\pi]$.

Sampling methods

This implies that D has the exponential distribution with $\lambda = 1/2$, and Θ has the uniform distribution over $[0, 2\pi]$.

When sampling from these two distributions, the standard normal Y can be samples from the equation

$$Y = \sqrt{d} \sin \theta$$

Sampling methods

This implies that D has the exponential distribution with $\lambda = 1/2$, and Θ has the uniform distribution over $[0, 2\pi]$.

When sampling from these two distributions, the standard normal Y can be samples from the equation

$$Y = \sqrt{d} \sin \theta$$

Thus Y can be sampled from

$$Y = \sqrt{-2 \ln U} \sin 2\pi V$$

where U, V are uniform distributions over $(0, 1)$

Sampling methods

This implies that D has the exponential distribution with $\lambda = 1/2$, and Θ has the uniform distribution over $[0, 2\pi]$.

When sampling from these two distributions, the standard normal Y can be samples from the equation

$$Y = \sqrt{d} \sin \theta$$

Thus Y can be sampled from

$$Y = \sqrt{-2 \ln U} \sin 2\pi V$$

where U, V are uniform distributions over $(0, 1)$
Homework Is the proof correct? Justify.

Sampling methods

Sampling from multivariate normal Recall that if $X \sim \mathcal{N}_d(\mu, \Sigma)$ then

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right)$$

Sampling methods

Sampling from multivariate normal Recall that if $X \sim \mathcal{N}_d(\mu, \Sigma)$ then

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right)$$

Then a direct way of generating random vectors from the multivariate normal distribution is to generate a d-vector iid standard normal $z = (z_1, z_2, \ldots, z_d)$ and then form the vector

$$x = R^T z + \mu$$

where R is a $d \times d$ matrix such that $R^T R = \Sigma$ (Choleskey factor of Σ). Then x has $\mathcal{N}_d(\mu, \Sigma)$ distribution

Sampling methods

Sampling from multivariate normal Recall that if $X \sim \mathcal{N}_d(\mu, \Sigma)$ then

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right)$$

Then a direct way of generating random vectors from the multivariate normal distribution is to generate a d-vector iid standard normal $z = (z_1, z_2, \ldots, z_d)$ and then form the vector

$$x = R^T z + \mu$$

where R is a $d \times d$ matrix such that $R^T R = \Sigma$ (Choleskey factor of Σ). Then x has $\mathcal{N}_d(\mu, \Sigma)$ distribution

Another way Generate x_1 from $\mathcal{N}(0, \sigma_{11})$, generate x_2 conditionally on x_1, generate x_3 conditionally on x_1 and x_2, and so on

Sampling methods

Accept-Reject method This method is useful when it is difficult to directly simulate from a given target density $f(x)$ on the real line, however there exists another density $g(x)$ which is easier to simulate from such that $\frac{f(x)}{g(x)}$ is uniformly bounded

Sampling methods

Accept-Reject method This method is useful when it is difficult to directly simulate from a given target density $f(x)$ on the real line, however there exists another density $g(x)$ which is easier to simulate from such that $\frac{f(x)}{g(x)}$ is uniformly bounded

Then we can simulate X from g, and retain it or discard it according to some specific rule. Because an X value is either retained or discarded, depending on whether it passes the admission rule, the method is called the accept–reject method. The density g is called the envelope density.

Sampling methods

Accept-Reject method This method is useful when it is difficult to directly simulate from a given target density $f(x)$ on the real line, however there exists another density $g(x)$ which is easier to simulate from such that $\frac{f(x)}{g(x)}$ is uniformly bounded

Then we can simulate X from g, and retain it or discard it according to some specific rule. Because an X value is either retained or discarded, depending on whether it passes the admission rule, the method is called the accept–reject method. The density g is called the envelope density.
Method:

1. Find a density function g and find a constant c such that $\frac{f(x)}{g(x)} \leq c$ for all x

Sampling methods

Accept-Reject method This method is useful when it is difficult to directly simulate from a given target density $f(x)$ on the real line, however there exists another density $g(x)$ which is easier to simulate from such that $\frac{f(x)}{g(x)}$ is uniformly bounded

Then we can simulate X from g, and retain it or discard it according to some specific rule. Because an X value is either retained or discarded, depending on whether it passes the admission rule, the method is called the accept–reject method. The density g is called the envelope density.
Method:

1. Find a density function g and find a constant c such that $\frac{f(x)}{g(x)} \leq c$ for all x
2. Generate $X \sim g$

Sampling methods

Accept-Reject method This method is useful when it is difficult to directly simulate from a given target density $f(x)$ on the real line, however there exists another density $g(x)$ which is easier to simulate from such that $\frac{f(x)}{g(x)}$ is uniformly bounded

Then we can simulate X from g, and retain it or discard it according to some specific rule. Because an X value is either retained or discarded, depending on whether it passes the admission rule, the method is called the accept–reject method. The density g is called the envelope density.
Method:

1. Find a density function g and find a constant c such that $\frac{f(x)}{g(x)} \leq c$ for all x
2. Generate $X \sim g$
3. Generate $U(0,1)$, independent of X

Sampling methods

Accept-Reject method This method is useful when it is difficult to directly simulate from a given target density $f(x)$ on the real line, however there exists another density $g(x)$ which is easier to simulate from such that $\frac{f(x)}{g(x)}$ is uniformly bounded

Then we can simulate X from g, and retain it or discard it according to some specific rule. Because an X value is either retained or discarded, depending on whether it passes the admission rule, the method is called the accept–reject method. The density g is called the envelope density.

Method:

1. Find a density function g and find a constant c such that $\frac{f(x)}{g(x)} \leq c$ for all x
2. Generate $X \sim g$
3. Generate $U(0,1)$, independent of X
4. Retain this generated X value if $U \leq \frac{f(x)}{cg(x)}$

Sampling methods

Accept-Reject method This method is useful when it is difficult to directly simulate from a given target density $f(x)$ on the real line, however there exists another density $g(x)$ which is easier to simulate from such that $\frac{f(x)}{g(x)}$ is uniformly bounded

Then we can simulate X from g, and retain it or discard it according to some specific rule. Because an X value is either retained or discarded, depending on whether it passes the admission rule, the method is called the accept–reject method. The density g is called the envelope density.

Method:

1. Find a density function g and find a constant c such that $\frac{f(x)}{g(x)} \leq c$ for all x
2. Generate $X \sim g$
3. Generate $U(0,1)$, independent of X
4. Retain this generated X value if $U \leq \frac{f(x)}{cg(x)}$
5. Repeat the steps until the required number of n values of X has been obtained

Sampling methods

Theorem Let $X \sim g$, and U, independent of X, uniformly distributed over $[0, 1]$. Then the conditional density of X given that $U \leq \frac{f(x)}{cg(x)}$ is f

Sampling methods

Theorem Let $X \sim g$, and U, independent of X, uniformly distributed over [0, 1]. Then the conditional density of X given that $U \leq \frac{f(x)}{cg(x)}$ is f

Proof denote the CDF of f by F. Then

$$P\left(X \leq x | U \leq \frac{f(x)}{cg(x)}\right) = \frac{P\left(X \leq x, U \leq \frac{f(x)}{cg(x)}\right)}{P\left(U \leq \frac{f(x)}{cg(x)}\right)}$$

Sampling methods

Theorem Let $X \sim g$, and U, independent of X, uniformly distributed over [0, 1]. Then the conditional density of X given that $U \leq \frac{f(x)}{cg(x)}$ is f

Proof denote the CDF of f by F. Then

$$P\left(X \leq x | U \leq \frac{f(x)}{cg(x)}\right) = \frac{P\left(X \leq x, U \leq \frac{f(x)}{cg(x)}\right)}{P\left(U \leq \frac{f(x)}{cg(x)}\right)}$$

$$= \frac{\int_{-\infty}^{x} \int_{0}^{\frac{f(t)}{cg(t)}} g(t) du\, dt}{\int_{-\infty}^{\infty} \int_{0}^{\frac{f(t)}{cg(t)}} g(t) du\, dt}$$

Sampling methods

Theorem Let $X \sim g$, and U, independent of X, uniformly distributed over $[0, 1]$. Then the conditional density of X given that $U \leq \frac{f(x)}{cg(x)}$ is f

Proof denote the CDF of f by F. Then

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

Sampling methods

Theorem Let $X \sim g$, and U, independent of X, uniformly distributed over $[0, 1]$. Then the conditional density of X given that $U \leq \frac{f(x)}{cg(x)}$ is f

Proof denote the CDF of f by F. Then

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

Sampling methods

Sampling normal distribution via Accept-reject Suppose we want to generate $X \sim \mathcal{N}(0,1)$ and we denote the density as f. We need to find an envelope density g such that $\frac{f(x)}{g(x)}$ is uniformly bounded, and it should be easy to sample from g

Sampling methods

Sampling normal distribution via Accept-reject Suppose we want to generate $X \sim \mathcal{N}(0,1)$ and we denote the density as f. We need to find an envelope density g such that $\frac{f(x)}{g(x)}$ is uniformly bounded, and it should be easy to sample from g
Consider $g(x) = \frac{1}{2}e^{-|x|}$. Then

$$\frac{f(x)}{g(x)} = \frac{\frac{1}{\sqrt{2\pi}}e^{-x^2/2}}{\frac{1}{2}e^{-|x|}} = \sqrt{\frac{2}{\pi}}e^{|x|-x^2/2} \leq \sqrt{\frac{2e}{\pi}}$$

for all real x

Sampling methods

Sampling normal distribution via Accept-reject Suppose we want to generate $X \sim \mathcal{N}(0,1)$ and we denote the density as f. We need to find an envelope density g such that $\frac{f(x)}{g(x)}$ is uniformly bounded, and it should be easy to sample from g
Consider $g(x) = \frac{1}{2}e^{-|x|}$. Then

$$\frac{f(x)}{g(x)} = \frac{\frac{1}{\sqrt{2\pi}}e^{-x^2/2}}{\frac{1}{2}e^{-|x|}} = \sqrt{\frac{2}{\pi}}e^{|x|-x^2/2} \leq \sqrt{\frac{2e}{\pi}}$$

for all real x
Set $c = \sqrt{\frac{2e}{\pi}}$ in the accept-reject scheme and, g has the standard double exponential density. The scheme works out to the following: generate U, and a double exponential value of X, and retain X if

$$U \leq e^{|X|-X^2/2-\frac{1}{2}}$$

Sampling methods

Generation of double exponential value by several ways:

(a) Generate a standard exponential value Y and assign it a random sign ($+$ or $-$ with an equal probability).

Sampling methods

Generation of double exponential value by several ways:

(a) Generate a standard exponential value Y and assign it a random sign ($+$ or $-$ with an equal probability).

(b) Generate two independent standard exponential values Y_1, Y_2 and set $X = Y_1 - Y_2$.

Sampling methods

Generation of double exponential value by several ways:

(a) Generate a standard exponential value Y and assign it a random sign ($+$ or $-$ with an equal probability).

(b) Generate two independent standard exponential values Y_1, Y_2 and set $X = Y_1 - Y_2$.

Question Can we understand the Accept-Reject method through a graph?

Sampling methods

Generation of double exponential value by several ways:

(a) Generate a standard exponential value Y and assign it a random sign ($+$ or $-$ with an equal probability).

(b) Generate two independent standard exponential values Y_1, Y_2 and set $X = Y_1 - Y_2$.

Question Can we understand the Accept-Reject method through a graph?

For example plot the graph $u = e^{|x|-x^2/2-\frac{1}{2}}$, and the generated X value is retained if and only if the pair (X, U) is below the graph of the function. Then we can see that one of the two generated values will be accepted, and the other rejected

Sampling methods

Generation of double exponential value by several ways:

(a) Generate a standard exponential value Y and assign it a random sign ($+$ or $-$ with an equal probability).

(b) Generate two independent standard exponential values Y_1, Y_2 and set $X = Y_1 - Y_2$.

Question Can we understand the Accept-Reject method through a graph?

For example plot the graph $u = e^{|x|-x^2/2-\frac{1}{2}}$, and the generated X value is retained if and only if the pair (X, U) is below the graph of the function. Then we can see that one of the two generated values will be accepted, and the other rejected

Question What is the acceptance percentage of the Accept-Reject scheme?