Bài 3
Phương pháp Monte Carlo và sinh số ngẫu nhiên

Thống kê máy tính và ứng dụng
(Computational Statistics and Applications)

Vũ Quốc Hoàng (vqhoang@fit.hcmus.edu.vn)
FIT - HCMUS

2025

Nội dung

1. Phương pháp Monte Carlo
2. Sinh số ngẫu nhiên
3. Các bộ sinh số giả ngẫu nhiên
4. Sinh biến ngẫu nhiên rời rạc
5. Phương pháp biến đổi ngược
6. Phương pháp lấy mẫu loại bỏ
7. Các phương pháp chuyên dụng

Nội dung

1. Phương pháp Monte Carlo
2. Sinh số ngẫu nhiên
3. Các bộ sinh số giả ngẫu nhiên
4. Sinh biến ngẫu nhiên rời rạc
5. Phương pháp biến đổi ngược
6. Phương pháp lấy mẫu loại bỏ
7. Các phương pháp chuyên dụng

Phương pháp Monte Carlo

Phương pháp Monte Carlo (Monte Carlo method)

- Sinh mẫu từ mô hình thống kê bằng mô phỏng trên máy tính.
- Dùng các tính chất thống kê của mẫu để nghiên cứu mô hình.

Monte Carlo là phương pháp "giải bài toán bằng ngẫu nhiên".

Ví dụ

- Kỳ vọng của một biến ngẫu nhiên có thể được xấp xỉ bằng trung bình của một lượng lớn các mẫu của biến ngẫu nhiên.
- Xác suất của một biến cố có thể được xấp xỉ bằng tần suất biến cố xảy ra trong một lượng lớn lần lặp lại thí nghiệm.
- Chất lượng của một phương pháp phân tích có thể được đánh giá bằng cách sinh nhiều dữ liệu từ một mô hình được chọn và đối chiếu kết quả phân tích được từ phương pháp với các tính chất đã biết của mô hình.

Phương pháp Monte Carlo - Ví dụ 1

Ước lượng số π bằng phương pháp Monte Carlo (xem Wikipedia - Monte Carlo method).

Chọn điểm ngẫu nhiên (X, Y) trong hình vuông đơn vị, xác suất (X, Y) rơi vào góc tư hình tròn là

$$P(X^2 + Y^2 \leq 1) = \frac{\text{Diện tích phần tư hình tròn}}{\text{Diện tích hình vuông}} = \frac{\pi}{4}.$$

Thủ tục ước lượng

1. Sinh nhiều cặp số ngẫu nhiên $x, y \sim \mathcal{U}(0, 1)$.
2. Tính tần suất f biến cố $x^2 + y^2 \leq 1$.
3. Ước lượng $\pi \approx 4f$.

Phương pháp Monte Carlo - Ví dụ 2

Cho $X \sim \mathcal{U}(a, b)$ với hàm mật độ $\varphi(x) = \frac{1}{b-a}\mathbb{I}_{[a,b]}(x)$, ta có

$$\int_a^b f(x)dx = (b-a)\int_{-\infty}^{\infty} f(x)\varphi(x)dx = (b-a)E(f(X)).$$

Bằng cách sinh các số $x_1, ..., x_N \overset{\text{iid}}{\sim} \mathcal{U}(a, b)$, ta có ước lượng

$$\int_a^b f(x)dx \approx \frac{(b-a)}{N}\sum_{i=1}^{N} f(x_i).$$

Ví dụ. Sinh $x_1, ..., x_N \overset{\text{iid}}{\sim} \mathcal{U}(0, 2\pi)$, ta có ước lượng

$$\int_0^{2\pi} e^{\cos x}dx \approx \frac{2\pi}{N}\sum_{i=1}^{N} e^{\cos x_i}.$$

Nội dung

1. Phương pháp Monte Carlo
2. **Sinh số ngẫu nhiên**
3. Các bộ sinh số giả ngẫu nhiên
4. Sinh biến ngẫu nhiên rời rạc
5. Phương pháp biến đổi ngược
6. Phương pháp lấy mẫu loại bỏ
7. Các phương pháp chuyên dụng

Sinh số ngẫu nhiên

Việc sinh được các số ngẫu nhiên (hay các đối tượng ngẫu nhiên) theo phân phối (hay qui luật) cho trước là thủ tục cốt lõi của phương pháp Monte Carlo (hay phương pháp mô phỏng trên máy tính).

Ví dụ. Từ một đồng xu có xác suất ra ngửa là p $(0 < p < 1)$, ta không biết giá trị của p, tìm cách sinh một biến cố có xác suất là 50%.

Ứng dụng. Trọng tài tung đồng xu để chọn phần sân cho 2 đội bóng khi bắt đầu trận đấu. Làm sao đảm bảo công bằng nếu trọng tài không đáng tin (đồng xu mà trọng tài dùng có thể không cân đối)?!

Sinh số ngẫu nhiên (tt)

Giải pháp. Ta có thể sinh một biến cố có xác suất là 50% từ đồng xu bằng thủ tục sau

- Bước 1. Tung đồng xu 2 lần, gọi m_1, m_2 lần lượt là mặt ra của đồng xu ở lần 1, lần 2.
- Bước 2. Lặp lại Bước 1 cho đến khi $m_1 \neq m_2$.
- Bước 3. Trả về biến cố "m_1 là ngửa".

(Xem Notebook)

Nội dung

1. Phương pháp Monte Carlo
2. Sinh số ngẫu nhiên
3. **Các bộ sinh số giả ngẫu nhiên**
4. Sinh biến ngẫu nhiên rời rạc
5. Phương pháp biến đổi ngược
6. Phương pháp lấy mẫu loại bỏ
7. Các phương pháp chuyên dụng

Các bộ sinh số giả ngẫu nhiên

Có 2 nhóm phương pháp sinh số ngẫu nhiên khác nhau

- Các phương pháp dùng hiện tượng vật lý ngẫu nhiên để sinh các **số thật sự ngẫu nhiên** (true random number). Các phương pháp này thường chậm và tốn kém do đòi hỏi phần cứng đặc biệt.
- Các phương pháp dùng chương trình máy tính để sinh các **số giả ngẫu nhiên** (pseudo random number). Các phương pháp này thường nhanh và ít tốn kém nhưng không thể sinh số thật sự ngẫu nhiên do bản chất **tất định** (deterministic) của chúng.

Bộ sinh số giả ngẫu nhiên (Pseudo Random Number Generator - PRNG) là một thuật toán tạo ra các dãy số có thể được dùng thay cho các dãy **độc lập và cùng phân phối** (iid) các số thật sự ngẫu nhiên.

Bộ sinh đồng dư tuyến tính

Thuật toán LCG. (Linear Congruential Generator)
Input:

- $m > 1$ (modulus)
- $a \in \{1, 2, ..., m-1\}$ (multiplier)
- $c \in \{0, 1, ..., m-1\}$ (increment)
- $X_0 \in \{0, 1, ..., m-1\}$ (seed)

Output: dãy $X_1, X_2, X_3, ...$ các số giả ngẫu nhiên.

1: **for** $n = 1, 2, 3, ...$ **do**
2: $\quad X_n \leftarrow (aX_{n-1} + c) \mod m$
3: $\quad$ output X_n
4: **end for**

(Xem Notebook)

Bộ sinh đồng dư tuyến tính - Ví dụ

Với $m = 8, a = 5, c = 1$ và **mầm** (seed) $X_0 = 0$ ta có

n	$5X_{n-1} + 1$	X_n
1	1	1
2	6	6
3	31	7
4	36	4
5	21	5
6	26	2
7	11	3
8	16	0
9	1	1
10	6	6

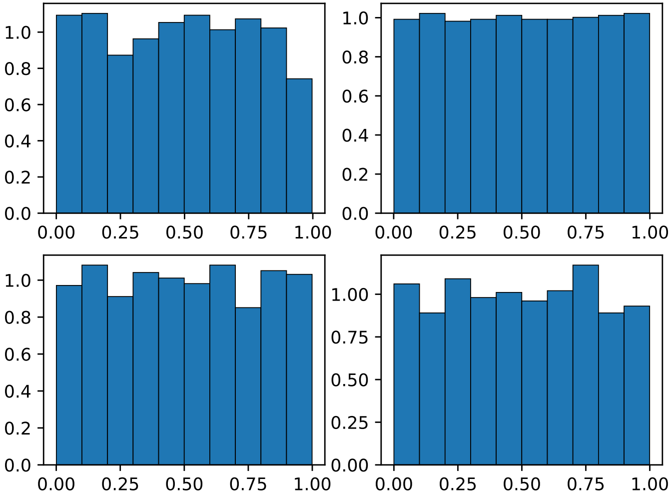
Bộ sinh đồng dư tuyến tính (tt)

- Dãy số tạo ra "trông có vẻ ngẫu nhiên" nhưng không thật sự ngẫu nhiên.
- Dãy số tạo ra **có chu kỳ** (periodic). Chẳng hạn trong Ví dụ trên, ta có $X_8 = X_0, X_9 = X_1, X_{10} = X_2, ...$
- Vì các X_n nhận giá trị trong tập $\{0, 1, ..., m-1\}$ nên **chiều dài chu kỳ** (period length) tối đa là m.
- m thường được chọn là số lớn và a, c được chọn sao cho dãy số tạo ra có chiều dài chu kỳ lớn nhất có thể.
- Một trường hợp hay dùng (ANSI C, C17, ...)

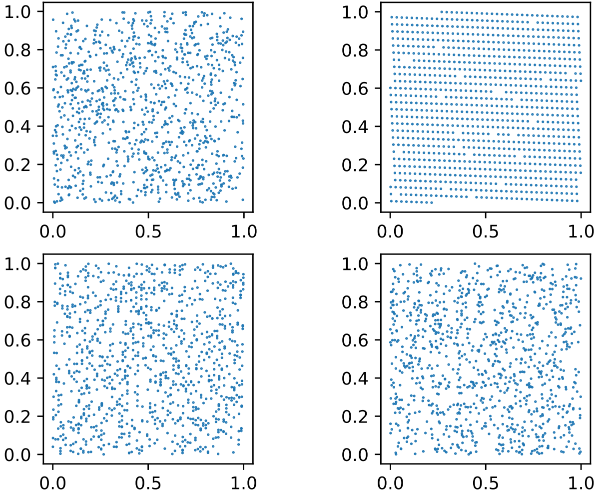
$$m = 2^{31}, a = 1103515245, c = 12345.$$

(xem Wikipedia - Linear congruential generator)

Chất lượng của các bộ sinh số giả ngẫu nhiên



Chất lượng của các bộ sinh số giả ngẫu nhiên (tt)



Các bộ sinh số giả ngẫu nhiên trong thực tế

- Nên dùng các PRNG từ các thư viện nổi tiếng hơn là tự cài đặt.
- Vì giá trị mầm (seed) quyết định dãy số được tạo ra cho nên ta nên đặt giá trị mầm
    - bằng giá trị cố định để có thể lặp lại được các kết quả có dùng dãy số được tạo ra (như trong các công bố khoa học).
    - bằng giá trị thay đổi (như thời điểm hiện tại) để có các dãy số khác nhau trong mỗi lần chạy.
- Các PRNG thường tạo ra dãy số nguyên $(X_n)_{n\in\mathbb{N}}$ phân phối đều trên tập $\{0, 1, ..., m-1\}$. Để tạo ra dãy số $(U_n)_{n\in\mathbb{N}}$ thực phân phối đều trên khoảng $(0, 1)$, tức là phân phối $\mathcal{U}(0, 1)$ hay được dùng để sinh mẫu cho các phân phối khác, ta có thể đặt

$$U_n = \frac{X_n + 1}{m + 1}.$$

Nội dung

1. Phương pháp Monte Carlo

2. Sinh số ngẫu nhiên

3. Các bộ sinh số giả ngẫu nhiên

4. **Sinh biến ngẫu nhiên rời rạc**

5. Phương pháp biến đổi ngược

6. Phương pháp lấy mẫu loại bỏ

7. Các phương pháp chuyên dụng

Sinh biến ngẫu nhiên rời rạc

Mệnh đề. Cho $p \in [0,1]$ và $U \sim \mathcal{U}(0,1)$, định nghĩa biến cố E là

$$E = \{U \leq p\}$$

thì $P(E) = p$.

Một biến ngẫu nhiên X có phân phối đều rời rạc trên tập $\{0, 1, ..., n-1\}$, kí hiệu $X \sim \mathcal{U}\{0, 1, ..., n-1\}$, nếu

$$P(X = k) = \frac{1}{n}, \forall k \in \{0, 1, ..., n-1\}.$$

Mệnh đề. Cho $U \sim \mathcal{U}(0,1)$ và $n \in \mathbb{N}$, định nghĩa biến ngẫu nhiên X là

$$X = \lfloor nU \rfloor,$$

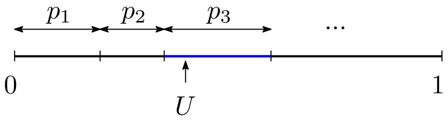
với $\lfloor . \rfloor$ kí hiệu làm tròn xuống. Khi đó, $X \sim \mathcal{U}\{0, 1, ..., n-1\}$.

Sinh biến ngẫu nhiên rời rạc (tt)

Mệnh đề. Cho $A = \{a_i : i \in I\}$ trong đó $I = \{1, 2, ..., n\}$ hoặc $I = \mathbb{N}$, và $a_i \neq a_j$ khi $i \neq j$. Cho $(p_i)_{i \in I}$ với $p_i \geq 0$ và $\sum_{i \in I} p_i = 1$. Cho $U \sim \mathcal{U}(0, 1)$, định nghĩa biến ngẫu nhiên K là

$$K = \min \left\{ k \in I : \sum_{i=1}^{k} p_i \geq U \right\}.$$

Đặt $X = a_K$ thì $P(X = a_i) = p_i, \forall i \in I$.



Sinh biến ngẫu nhiên rời rạc - Ví dụ

Biến ngẫu nhiên X có phân phối hình học với tham số p $(0 \leq p \leq 1)$ nếu X nhận các giá trị trong tập $\mathbb{N} = \{1, 2, ...\}$ với xác suất $P(X = i) = p^{i-1}(1 - p), i \in \mathbb{N}$.

Đặt $I = \mathbb{N}, a_i = i, p_i = p^{i-1}(1 - p), i \in \mathbb{N}$ ta có

$$\sum_{i=1}^{k} p_i = (1 - p) \sum_{i=1}^{k} p^{i-1} = (1 - p) \frac{1 - p^k}{1 - p} = 1 - p^k.$$

Dùng Mệnh đề trên, từ $U \sim \mathcal{U}(0, 1)$, ta có thể sinh $X = a_K = K$ với

$$K = \min \left\{ k \in I : \sum_{i=1}^{k} p_i \geq U \right\} = \min \left\{ k \in I : 1 - p^k \geq U \right\}$$
$$= \min \left\{ k \in I : k \geq \frac{\log (1 - U)}{\log p} \right\} = \left\lceil \frac{\log (1 - U)}{\log p} \right\rceil.$$

Nội dung

1. Phương pháp Monte Carlo

2. Sinh số ngẫu nhiên

3. Các bộ sinh số giả ngẫu nhiên

4. Sinh biến ngẫu nhiên rời rạc

5. Phương pháp biến đổi ngược

6. Phương pháp lấy mẫu loại bỏ

7. Các phương pháp chuyên dụng

Hàm phân phối tích lũy và hàm ngược

Hàm phân phối tích lũy (cumulative distribution function - CDF) của biến ngẫu nhiên X trên $\mathbb{R}$ được cho bởi

$$F(x) = P(X \leq x), x \in \mathbb{R}.$$

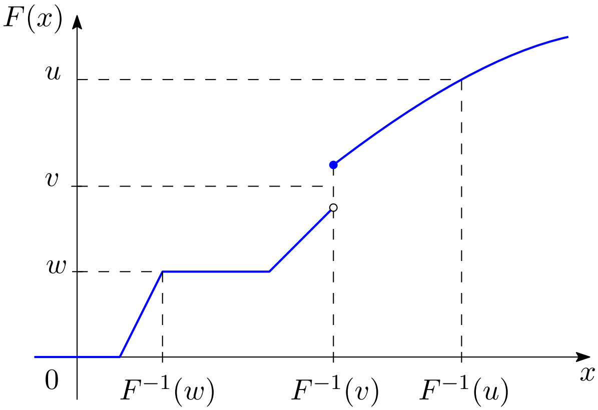
Khi đó, ta còn nói X có phân phối F, kí hiệu $X \sim F$.

Cho F là CDF của biến ngẫu nhiên X, **hàm ngược** (inverse) của F được định nghĩa là

$$F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\}, u \in (0, 1).$$

Khi F là song ánh, hàm ngược F^{-1} chính là hàm ngược của F theo nghĩa thông thường, tức là $F^{-1}(u) = x$ khi và chỉ khi $F(x) = u$.

Hàm phân phối tích lũy và hàm ngược



Phương pháp biến đổi ngược

Thuật toán ITM. (Inverse Transform Method)
Input: hàm ngược F^{-1} của CDF F
Output: $X \sim F$

1: sinh $U \sim \mathcal{U}(0,1)$
2: **return** $X = F^{-1}(U)$

Định lý. Cho $F : \mathbb{R} \to [0,1]$ là một CDF với hàm ngược F^{-1} và $U \sim \mathcal{U}(0,1)$, định nghĩa biến ngẫu nhiên

$$X = F^{-1}(U)$$

thì $X \sim F$.

Phương pháp biến đổi ngược - Ví dụ 1

Phân phối mũ Exp(λ) có hàm mật độ

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{nếu } x \geq 0 \\ 0 & \text{khác.} \end{cases}$$

Sử dụng tích phân từng phần, ta có CDF

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{0}^{x} \lambda e^{-\lambda t} dt = -e^{-\lambda t}\Big|_{t=0}^{x} = 1 - e^{-\lambda x}, \forall x \geq 0.$$

Vì hàm này tăng ngặt và liên tục nên F^{-1} là hàm ngược của F như thông thường. Ta có $1 - e^{-\lambda x} = u \Leftrightarrow -\lambda x = \log(1-u) \Leftrightarrow$ $x = -\log(1-u)/\lambda$ nên $F^{-1}(u) = -\log(1-u)/\lambda$.
Như vậy, với $U \sim \mathcal{U}(0,1)$ thì $X = -\log(1-U)/\lambda$ có phân phối Exp(λ).

Phương pháp biến đổi ngược - Ví dụ 2

Cho biến ngẫu nhiên X có hàm mật độ

$$f(x) = \begin{cases} 3x^2 & \text{nếu } x \in [0,1] \\ 0 & \text{khác.} \end{cases}$$

Ta có CDF

$$F(x) = \int_{-\infty}^{x} f(t)dt = \begin{cases} 0 & \text{nếu } x < 0 \\ x^3 & \text{nếu } 0 \leq x < 1 \\ 1 & \text{nếu } 1 \leq x. \end{cases}$$

Vì F là song ánh từ $(0,1)$ vào $(0,1)$ nên $F^{-1}(u) = u^{1/3}, \forall u \in (0,1)$.
Như vậy, với $U \sim \mathcal{U}(0,1)$ thì $X = U^{1/3}$ có phân phối như đã cho.

Phương pháp biến đổi ngược (tt)

- Phương pháp biến đổi ngược luôn có thể được dùng khi F^{-1} là dễ tính.
- Trong một số trường hợp, như với phân phối chuẩn, F^{-1} khó tính nên không thể dùng trực tiếp phương pháp này.
- Phương pháp biến đổi ngược cũng có thể dùng cho phân phối rời rạc nhưng không cần thiết.
- Hạn chế chính của phương pháp biến đổi ngược là chỉ dùng được cho phân phối 1 chiều.
- Với phân phối trên $\mathbb{R}^d$ $(d > 1)$, ta cần dùng các phương pháp lấy mẫu tinh vi hơn.

Nội dung

1. Phương pháp Monte Carlo
2. Sinh số ngẫu nhiên
3. Các bộ sinh số giả ngẫu nhiên
4. Sinh biến ngẫu nhiên rời rạc
5. Phương pháp biến đổi ngược
6. **Phương pháp lấy mẫu loại bỏ**
7. Các phương pháp chuyên dụng

Lấy mẫu loại bỏ cơ bản

Thuật toán BRS. (Basic Rejection Sampling)

Input:

- hàm mật độ xác suất g (proposal density),
- hàm p với giá trị trong khoảng $[0, 1]$ (acceptance probability).

Output: $X_{N_1}, X_{N_2}, X_{N_3}, \dots$ iid với hàm mật độ

$$f(x) = \frac{1}{Z} p(x) g(x) \text{ với } Z = \int p(x) g(x) dx.$$

for $n = 1, 2, 3, \dots$ **do**
    sinh $X_n \sim g$ #sinh đề cử (proposal)
    sinh $U_n \sim \mathcal{U}(0, 1)$
    if $U_n \leq p(X_n)$ **then**
        xuất X_n #X_n được chấp nhận (accepted) với xác suất $p(X_n)$
    end if #else: X_n bị loại bỏ (rejected)
end for

Lấy mẫu loại bỏ cơ bản (tt)

Mệnh đề. Cho $k \in \mathbb{N}$, gọi X_{N_k} là kết xuất thứ k của Thuật toán BRS, các phát biểu sau đây đúng:

1. Các phần tử của dãy $(X_{N_k})_{k\in\mathbb{N}}$ là iid với hàm mật độ

$$f(x) = \frac{1}{Z}p(x)g(x) \text{ với } Z = \int p(x)g(x)dx.$$

2. Mỗi đề cử được chấp nhận với xác suất Z và số lượng đề cử cần sinh để được mỗi X_{N_k} có phân phối hình học với kỳ vọng $1/Z$.

Lấy mẫu loại bỏ cơ bản - Ví dụ

Cho $X \sim \mathcal{U}(-1,1)$ và X được chấp nhận với xác suất

$$p(X) = \sqrt{1 - X^2}.$$

Khi đó, mẫu sinh ra từ Thuật toán BRS có mật độ

$$f(x) = \frac{1}{Z} p(x) g(x) = \frac{1}{Z} \sqrt{1 - x^2} \frac{1}{2} \mathbb{I}_{[-1,1]}(x),$$

với

$$\mathbb{I}_{[-1,1]}(x) = \begin{cases} 1 & \text{nếu } x \in [-1,1] \\ 0 & \text{khác}, \end{cases}$$

và

$$Z = \int \sqrt{1 - x^2} \frac{1}{2} \mathbb{I}_{[-1,1]}(x) dx = \frac{1}{2} \int_{-1}^{1} \sqrt{1 - x^2} dx = \frac{\pi}{4} \approx 0.7854.$$

Như vậy, phân phối của X là $f(x) = \frac{2}{\pi} \sqrt{1 - x^2} \mathbb{I}_{[-1,1]}(x)$. Phân phối này được gọi là **phân phối nửa đường tròn Wigner**.

Lấy mẫu loại bỏ theo khuôn

Thuật toán ERS. (Envelope Rejection Sampling)
Input:

- hàm f với giá trị trong khoảng $[0, \infty)$ (non-normalised target density),
- hàm mật độ xác suất g (proposal density),
- hằng số $c > 0$ sao cho $f(x) \leq cg(x), \forall x$.

Output: $X_{N_1}, X_{N_2}, X_{N_3}, \ldots$ iid với hàm mật độ

$$\tilde{f}(x) = \frac{1}{Z_f} f(x) \text{ với } Z_f = \int f(x)dx.$$

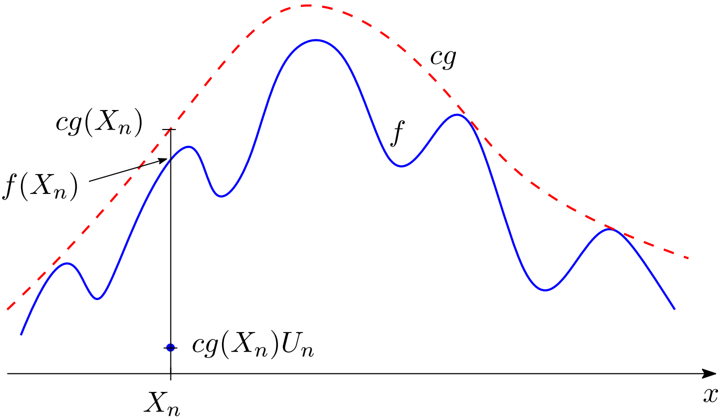
for $n = 1, 2, 3, \ldots$ **do**
    sinh $X_n \sim g$ #sinh đề cử
    sinh $U_n \sim \mathcal{U}(0, 1)$
    if $cg(X_n)U_n \leq f(X_n)$ **then**
        xuất X_n #X_n được chấp nhận với xác suất $f(X_n)/(cg(X_n))$
    end if #else: X_n bị loại bỏ
end for

Lấy mẫu loại bỏ theo khuôn (tt)

Mệnh đề. Cho $k \in \mathbb{N}$, gọi X_{N_k} là kết xuất thứ k của Thuật toán ERS. Các phát biểu sau đây đúng:

1. Các phần tử của dãy $(X_{N_k})_{k\in\mathbb{N}}$ là iid với hàm mật độ
$$\tilde{f}(x) = \frac{1}{Z_f} f(x) \text{ với } Z_f = \int f(x)dx.$$
2. Mỗi đề cử được chấp nhận với xác suất Z_f/c và số lượng đề cử cần sinh để được mỗi X_{N_k} có phân phối hình học với kỳ vọng c/Z_f.

Lấy mẫu loại bỏ theo khuôn (tt)



Lấy mẫu loại bỏ theo khuôn - Ví dụ

Dùng thuật toán lấy mẫu loại bỏ theo khuôn để sinh mẫu từ **phân phối nửa chuẩn** (half-normal distribution) có hàm mật độ

$$f(x) = \begin{cases} \frac{2}{\sqrt{2\pi}} e^{-x^2/2} & \text{nếu } x \geq 0 \\ 0 & \text{khác,} \end{cases}$$

Nếu dùng các đề cử từ phân phối mũ $\text{Exp}(\lambda)$ thì hàm mật độ đề cử là

$$g(x) = \begin{cases} \lambda e^{-\lambda x} & \text{nếu } x \geq 0 \\ 0 & \text{khác,} \end{cases}$$

Để áp dụng Thuật toán ERS, ta cần xác định hằng số $c > 0$ sao cho $f(x) \leq cg(x), \forall x \in \mathbb{R}$. Với $x < 0$ ta có $f(x) = g(x) = 0$. Với $x \geq 0$ ta có

$$\frac{f(x)}{g(x)} = \frac{2}{\sqrt{2\pi}\lambda} e^{\left(-x^2/2 + \lambda x\right)} \leq \sqrt{\frac{2}{\pi\lambda^2}} e^{\lambda^2/2} = c^*.$$

Lấy mẫu loại bỏ theo khuôn - Ví dụ (tt)

Với g và $c = c^*$ đã chọn, ta có

$$cg(x)U \le f(x) \Leftrightarrow \sqrt{\frac{2}{\pi\lambda^2}} e^{\lambda^2/2} \lambda e^{-\lambda x} U \le \frac{2}{\sqrt{2\pi}} e^{-x^2/2}$$
$$\Leftrightarrow U \le e^{-\frac{1}{2}(x-\lambda)^2}.$$

Từ đó ta có thuật toán sinh mẫu cho phân phối nửa chuẩn:

1: **for** $n = 1, 2, 3, ...$ **do**
2: sinh $X_n \sim \text{Exp}(\lambda)$
3: sinh $U_n \sim \mathcal{U}(0, 1)$
4: **if** $U_n \le e^{-\frac{1}{2}(X_n-\lambda)^2}$ **then**
5: xuất X_n
6: **end if**
7: **end for**

Lấy mẫu loại bỏ theo khuôn - Ví dụ (tt)

Vì f đã được chuẩn hóa nên $Z_f = 1$. Với $c = c^*$ và lấy $\lambda = 1$ ta có xác suất chấp nhận là

$$\frac{Z_f}{c} = \frac{1}{c^*} = \frac{1}{\sqrt{\frac{2}{\pi\lambda^2}}e^{\lambda^2/2}} = \sqrt{\frac{\pi}{2e}} \approx 76.02\%.$$

Vì phân phối chuẩn tắc $\mathcal{N}(0,1)$ đối xứng qua kỳ vọng 0 nên ta có thể lấy mẫu cho phân phối nửa chuẩn bằng cách

1: sinh $Z \sim \mathcal{N}(0,1)$
2: xuất $X = |Z|$

Phương pháp này có xác suất chấp nhận là 100%.

Ta cũng có thể lấy mẫu cho phân phối nửa chuẩn bằng cách

1: sinh $Z \sim \mathcal{N}(0,1)$
2: **if** $Z \geq 0$ **then**
3:     xuất $X = Z$

Phương pháp này có xác suất chấp nhận là 50%.

Phân phối có điều kiện

- Cho biến ngẫu nhiên X và cố định tập A với $P(X \in A) > 0$, phân phối có điều kiện $P_{X|X\in A}$ của X khi biết $X \in A$ được định nghĩa là

$$P_{X|X\in A}(B) = P(X \in B|X \in A) = \frac{P(X \in B, X \in A)}{P(X \in A)}, \forall B.$$

- Phân phối có điều kiện $P_{X|X\in A}$ mô tả phần ngẫu nhiên còn lại trong X khi đã biết $X \in A$.
- Có thể dễ dàng lấy mẫu từ các phân phối có điều kiện bằng phương pháp lấy mẫu loại bỏ.

Lấy mẫu cho phân phối có điều kiện

Thuật toán RSCD. (Rejection Sampling for Conditional Distributions)
Input: tập A với $P(X \in A) > 0$
Randomness used: $X_1, X_2, X_3, \ldots$ iid với phân phối như X (proposals)
Output: $X_{N_1}, X_{N_2}, X_{N_3}, \ldots$ iid với phân phối có điều kiện $P_{X|X \in A}$

```
for n = 1, 2, 3, ... do
    sinh X_n
    if X_n ∈ A then
        xuất X_n
    end if
end for
```

Lấy mẫu cho phân phối có điều kiện

Mệnh đề. Cho biến ngẫu nhiên X và tập A với $P(X \in A) > 0$, gọi X_{N_k} là kết xuất thứ k ($k \in \mathbb{N}$) của Thuật toán RSCD. Các phát biểu sau đây đúng:

1. Các phần tử của dãy $(X_{N_k})_{k \in \mathbb{N}}$ là iid với phân phối thỏa
$$P(X_{N_k} \in B) = P_{X|X \in A}(B), \forall B.$$
2. Mỗi đề cử được chấp nhận với xác suất $P(X \in A)$ và số lượng đề cử cần sinh để được mỗi X_{N_k} có phân phối hình học với kỳ vọng $1/P(X \in A)$.

Lấy mẫu cho phân phối có điều kiện - Ví dụ

Ta có thể dùng Thuật toán RSCD để sinh các mẫu $X \sim \mathcal{N}(0,1)$ với điều kiện $X \geq a$ bằng cách lặp lại các bước sau cho đến khi đủ số lượng mẫu:

1. sinh $X \sim \mathcal{N}(0,1)$,
2. **if** $X \geq a$ **then** xuất X.

Sự hiệu quả của phương pháp này phụ thuộc vào a. Bảng sau đây cho thấy kỳ vọng số lượng $X \sim \mathcal{N}(0,1)$ cần sinh để được mỗi $X \geq a$ (làm tròn đến số nguyên gần nhất)

a	1	2	3	4	5	6
$E(N_a)$	6	44	741	31 574	3 488 556	1 013 594 635

Lấy mẫu cho phân phối có điều kiện - Ví dụ (tt)

Ta có thể dùng Thuật toán ERS để sinh các mẫu $X \sim \mathcal{N}(0,1)$ với điều kiện $X \geq a > 0$ hiệu quả hơn Thuật toán RSCD. Hàm mật độ của phân phối có điều kiện này là

$$\tilde{f}(x) = \frac{1}{Z} e^{-x^2/2} \mathbb{I}_{[a,\infty)}(x) = \frac{1}{Z} f(x).$$

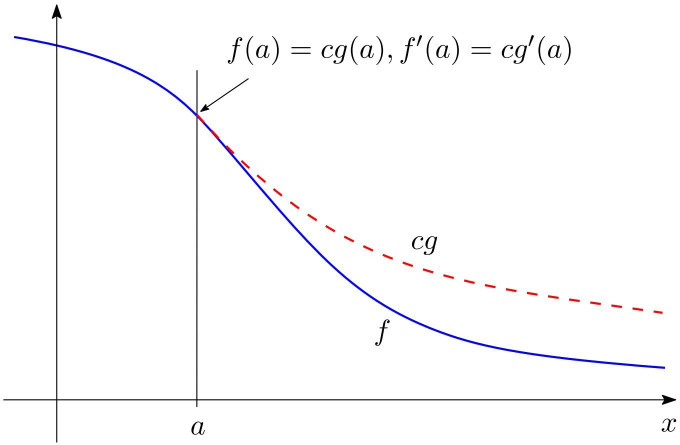
Ta có thể dùng các đề cử dạng $X = \tilde{X} + a$ với $\tilde{X} \sim \text{Exp}(\lambda)$ với hàm mật độ

$$g(x) = \lambda e^{-\lambda(x-a)} \mathbb{I}_{[a,\infty)}(x).$$

Tiếp theo ta tìm hằng số $c > 0$ sao cho $f(x) \leq cg(x), \forall x \geq a$ và chọn λ sao cho c nhỏ nhất có thể. Hình sau gợi ý cho ta lựa chọn

$$\lambda = a \ \text{ và } \ c = \frac{e^{-a^2/2}}{a}.$$

Lấy mẫu cho phân phối có điều kiện - Ví dụ (tt)



Lấy mẫu cho phân phối có điều kiện - Ví dụ (tt)

Như vậy, ta có thể dùng Thuật toán ERS để sinh các mẫu $X \sim \mathcal{N}(0,1)$ với điều kiện $X \geq a$ bằng cách lặp lại các bước sau cho đến khi đủ số lượng mẫu:

1. sinh $\tilde{X} \sim \mathrm{Exp}(a)$,
2. sinh $U \sim \mathcal{U}(0,1)$,
3. đặt $X = \tilde{X} + a$,
4. **if** $U \leq e^{-(X-a)^2/2}$ **then** xuất X.

Ta cũng có

$$E(N_a) = \frac{c}{\int f(x)dx} = \frac{\exp(-a^2/2)/a}{\int_a^\infty \exp(-x^2/2)dx} = \frac{\exp(-a^2/2)}{a\sqrt{2\pi}(1-\Phi(a))}.$$

a	1	2	3	4	5	6
$E(N_a)$	1.53	1.19	1.09	1.06	1.04	1.03

Nội dung

1. Phương pháp Monte Carlo

2. Sinh số ngẫu nhiên

3. Các bộ sinh số giả ngẫu nhiên

4. Sinh biến ngẫu nhiên rời rạc

5. Phương pháp biến đổi ngược

6. Phương pháp lấy mẫu loại bỏ

7. Các phương pháp chuyên dụng

Các phương pháp chuyên dụng

- Các phương pháp trên là các phương pháp chung, có thể dùng cho nhiều phân phối với nhiều ứng dụng khác nhau.
- Có các phương pháp chuyên dụng cho các phân phối điển hình với các ứng dụng chuyên biệt: các phương pháp này thường nhanh nhưng phức tạp.
- Tham khảo: **Chapter 4.** J. S. Dagpunar. *Simulation and Monte Carlo - With applications in finance and MCMC*. John Wiley & Sons, 2007.

Tài liệu tham khảo

Jupyter Notebook đi kèm.

Chapter 3-5. Sheldon M. Ross. *Simulation*. Elsevier, 2023.

Chapter 1. Jochen Voss. *An Introduction to Statistical Computing - A Simulation-based Approach*. John Wiley & Sons, 2014.

Chapter 2-4. J. S. Dagpunar. *Simulation and Monte Carlo - With applications in finance and MCMC*. John Wiley & Sons, 2007.