Bài 3 - Sinh số ngẫu nhiên (Random number generation)

Thống kê máy tính và ứng dụng (CLC)

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

Ngày 11 tháng 2 năm 2022

Nội dung

Nội dung

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 numbers). 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 numbers). 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

- 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 $m \approx 2^{32} \approx 4 \times 10^9$ và a, c được chọn sao cho dãy số tạo ra có chiều dài chu kỳ lớn nhất có thể.

Bộ sinh đồng dư tuyến tính

Định lý. Bộ sinh đồng dư tuyến tính có chu kỳ m khi và chỉ khi thỏa 3 điều kiện sau:

1. m và c nguyên tố cùng nhau,
2. $a - 1$ chia hết cho mọi ước nguyên tố của m,
3. nếu m là bội của 4 thì $a - 1$ cũng là bội của 4.

Hơn nữa, chiều dài chu kỳ không phụ thuộc vào mầm X_0.

Ví dụ: Với $m = 2^{32}, a = 1103515245, c = 12345$:

1. m chỉ có ước số là 2 và c lẻ nên m, c nguyên tố cùng nhau,
2. m chỉ có ước số là 2 và a lẻ nên $a - 1$ chia hết cho mọi ước nguyên tố của m,
3. m là bội của 4 và $a - 1 = 1103515244 = 4 \times 275878811$ cũng là bội của 4.

Vậy, LCG với các tham số này có chiều dài chu kỳ là 2^{32} không phụ thuộc vào mầm X_0.

Chất lượng của các bộ sinh số giả ngẫu nhiên

Các tiêu chuẩn đánh giá chất lượng của các PRNG

- Chiều dài chu kỳ
- Phân phối của mẫu
- Tính độc lập của mẫu

Phân phối của mẫu

- Các số sinh ra từ PRNG thường được dùng thay cho mẫu iid có **phân phối đều** (uniformly distributed). Vì các số sinh ra nhận giá trị trong tập hữu hạn $S = \{0, 1, ..., m-1\}$, nên trong dài hạn, với mọi tập $A \subset S$, ta cần có

$$\frac{\#\{i | 1 \leq i \leq N, X_i \in A\}}{N} \approx \frac{\#A}{\#S},$$

  với $\#A$ kí hiệu cho số lượng phần tử của tập hữu hạn A.
- Tính phân phối đều của mẫu có thể được kiểm tra qua các kiểm định thống kê như kiểm định Chi-bình phương hay kiểm định Kolmogorov–Smirnov.
- Đặc biệt, nên dùng **kiểm định 2 phía** (two-sided test) để phát hiện tính "không đồng đều" lẫn "quá đều đặn" của mẫu.

Phân phối của mẫu - Ví dụ

Kiểm tra phân phối mẫu của PRNG với $m = 1024$ dùng kiểm định Chi-bình phương với giả thuyết H_0:

$$P(X_i \in \{64j, 64j+1, 64j+2, ..., 64j+63\}) = 1/16, j = 0, 1, ..., 15.$$

Với mẫu $X_1, X_2, ..., X_N$, ta có tần số thực tế và mong đợi là:

$$O_j = \#\{i | 64j \leq X_i \leq 64(j+1)\}, E_j = N/16, j = 0, 1, ..., 15.$$

Khi cỡ mẫu N lớn, nếu H_0 đúng, giá trị thống kê kiểm định

$$Q = \sum_{j=0}^{15} \frac{(O_j - E_j)^2}{E_j}$$

có phân phối χ^2 với 15 bậc tự do.

Phân phối của mẫu - Ví dụ (tt)

Bảng sau liệt kê một vài giá trị phân vị của phân phối χ^2 với 15 bậc tự do

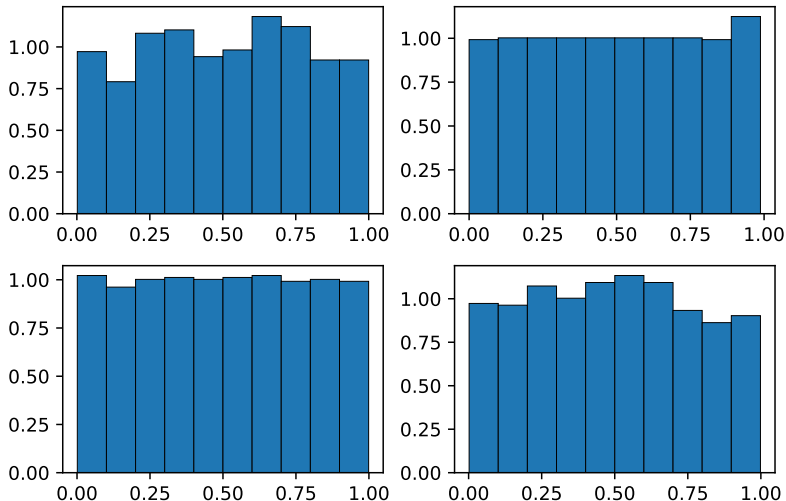
q	6.262	7.261	...	24.996	27.488
$P(Q \leq q)$	0.025	0.05	...	0.95	0.975

Xét một vài dãy số với cỡ $N = 10^6$:

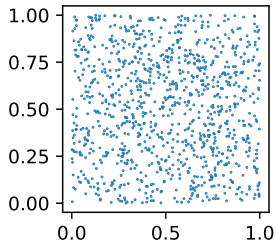
- Dãy số "quá đều đặn": $X_n = n \mod 1024$ có $Q = 0.256$
- Dãy số "không đồng đều": $X_n = n \mod 1020$ có $Q = 233.868$
- Dãy số sinh từ LCG với $m = 1024, a = 493, c = 123, X_0 = 0$ có $Q = 0.003$
- Dãy số sinh từ hàm `numpy.random.randint` của Python có $Q = 13.537$ (lưu ý, kết quả khác nhau mỗi lần chạy)

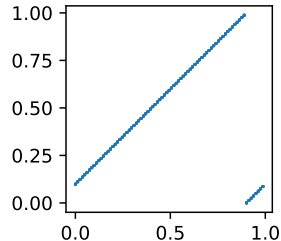
(Xem Notebook)

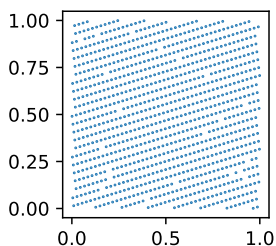
Phân phối của mẫu

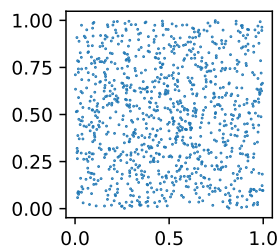


Tính độc lập của mẫu









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

Các phân phối 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

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

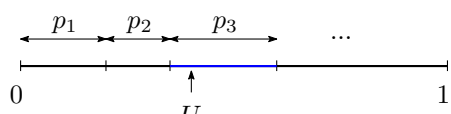
với $\lfloor . \rfloor$ kí hiệu làm tròn xuống. Khi đó, $X \sim \mathcal{U}\{0,1,...,n-1\}$.

Sinh phân phối rời rạc bất kỳ

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

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

Khi đó $X = a_K$ thỏa $P(X = a_k) = p_k, \forall k \in I$.
Tức là, X là biến ngẫu nhiên rời rạc nhận giá trị trên tập A với phân phối (p_i).



Sinh phân phối rời rạc bất kỳ - 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

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

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

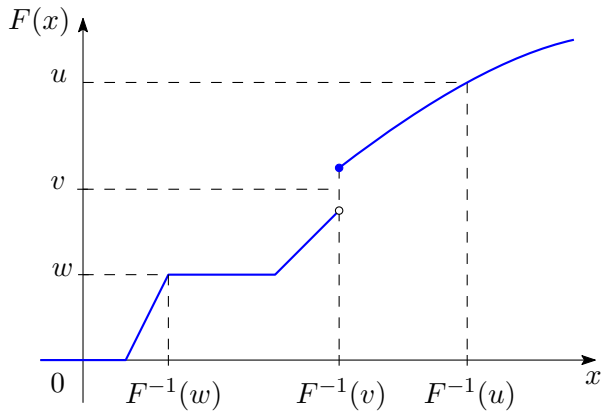
Khi đó, ta còn nói X có phân phối F, kí hiệu $X \sim F$.

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

Phân phối Rayleigh với tham số $\sigma > 0$ có hàm mật độ

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

Ta có CDF

$$F(x) = \int_0^x \frac{t}{\sigma^2} e^{-t^2/2\sigma^2} dt = -e^{-t^2/2\sigma^2}\Big|_{t=0}^{x} = 1 - e^{-x^2/2\sigma^2}, \forall x \geq 0.$$

Giải phương trình $u = F(x) = 1 - e^{-x^2/2\sigma^2}$ theo x ta có
$F^{-1}(u) = x = \sqrt{-2\sigma^2 \log(1-u)}$.
Như vậy, với $U \sim \mathcal{U}[0,1]$ thì $X = \sqrt{-2\sigma^2 \log(1-U)}$ có phân phối Rayleigh với tham số σ.

Phương pháp biến đổi ngược - Ví dụ 3

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 - Ví dụ 4

Cho biến ngẫu nhiên rời rạc X với $P(X = 0) = 0.6$ và $P(X = 1) = 0.4$. Ta có CDF

$$F(x) = \begin{cases} 0 & \text{nếu } x < 0 \\ 0.6 & \text{nếu } 0 \leq x < 1 \\ 1 & \text{nếu } 1 \leq x. \end{cases}$$

Dùng định nghĩa của hàm ngược F^{-1} ta có

$$F^{-1}(u) = \begin{cases} 0 & \text{nếu } 0 < u \leq 0.6 \\ 1 & \text{nếu } 0.6 < u < 1. \end{cases}$$

Như vậy, với $U \sim \mathcal{U}[0, 1]$ thì

$$X = \begin{cases} 0 & \text{nếu } U \leq 0.6 \\ 1 & \text{nếu } U > 0.6. \end{cases}$$

có phân phối như đã cho.

Phương pháp biến đổi ngược

- 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

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

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

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

Lấy mẫu loại bỏ cơ bản

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

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

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

Lấy mẫu loại bỏ theo khuôn

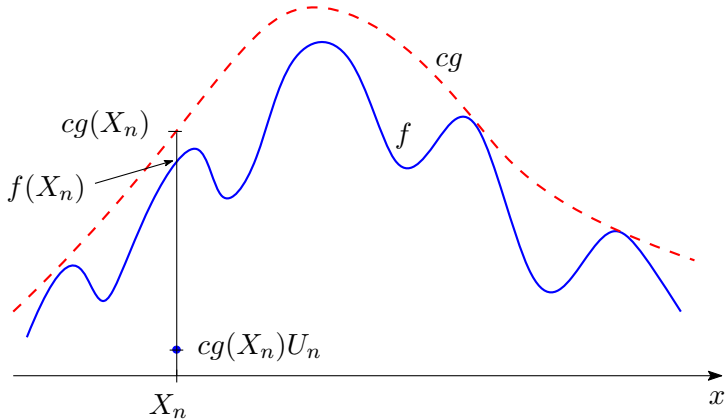
Mệnh đề. Cho $k \in \mathbb{N}$, gọi X_{N_k} là kết xuất thứ k của Thuật toán ERS. Các phát biểu sau đây đúng:

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}, $M_k = N_k - N_{k-1}$, có phân phối hình học với kỳ vọng $E(M_k) = c/Z_f$.

Lấy mẫu loại bỏ theo khuôn



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** $n = 1, 2, 3, \ldots$ **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 $P_{X|X\in A}$

1: **for** $n = 1, 2, 3, \ldots$ **do**
2: sinh X_n
3: **if** $X_n \in A$ **then**
4: xuất X_n
5: **end if**
6: **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}, $M_k = N_k - N_{k-1}$, có phân phối hình học với kỳ vọng $E(M_k) = 1/P(X \in A)$.

Lấy mẫu cho phân phối có điều kiện - Ví dụ 1

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 2

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

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

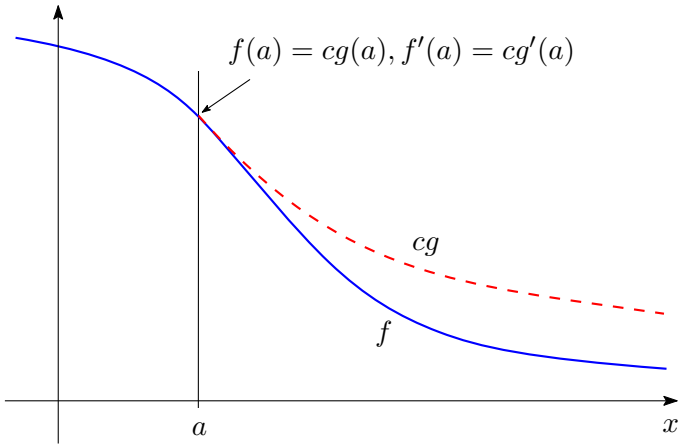
Ta có thể dùng các đề cử dạng $X = \tilde{X} + a$ với $\tilde{X} \sim \text{Exp}(\lambda)$ với hàm mật độ

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

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 2 (tt)



Lấy mẫu cho phân phối có điều kiện - Ví dụ 2 (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 $\tilde{X} \sim \text{Exp}(a)$,
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

Diễn giải hình học

Cho $A \subset \mathbb{R}^d$, kí hiệu $|A|$ chỉ "thể tích" d-chiều của A. Ví dụ:

1. Khối hộp $Q = [a, b]^3 \subset \mathbb{R}^3$ có thể tích $|Q| = (b - a)^3$,
2. Đường tròn đơn vị $C = \{(x, y) \in \mathbb{R}^2 | x^2 + y^2 \leq 1\}$ có diện tích $|C| = \pi$,
3. Đoạn thẳng $L = [a, b] \subset \mathbb{R}$ có chiều dài $|L| = (b - a)$.

Với tập $A \subset \mathbb{R}^d$, thể tích của A có thể được tính bằng tích phân

$$|A| = \int_{\mathbb{R}^d} \mathbb{I}_A(x)dx = \int ... \int \mathbb{I}_A(x_1, ..., x_d)dx_1...dx_d.$$

Một biến ngẫu nhiên có giá trị trên $\mathbb{R}^d$ được gọi là phân phối đều trên tập $A \subset \mathbb{R}^d$ với $0 < |A| < \infty$, kí hiệu $X \sim \mathcal{U}(A)$, nếu

$$P(X \in B) = \frac{|A \cap B|}{|A|}, \forall B \subset \mathbb{R}^d.$$

Diễn giải hình học

Mệnh đề G1. Cho $A \subset \mathbb{R}^d$ với $0 < |A| < \infty$, phân phối đều trên A, $\mathcal{U}(A)$, có hàm mật độ xác suất

$$f(x) = \frac{\mathbb{I}_A(x)}{|A|}, x \in \mathbb{R}^d.$$

Mệnh đề G2. Cho $X \sim \mathcal{U}(A)$ và tập B với $|A \cap B| > 0$, phân phối có điều kiện của X khi biết $X \in B$ là phân phối đều trên $A \cap B$, tức là $P_{X|X \in B} \sim \mathcal{U}(A \cap B)$.

Diễn giải hình học - Ví dụ

Lấy $(X_n), (Y_n) \sim \mathcal{U}[-1, 1]$ iid. Khi đó, các cặp (X_n, Y_n) phân phối đều trên $A = [0, 1] \times [0, 1]$. Đặt $(Z_k)_{k\in\mathbb{N}}$ là dãy con các cặp (X_{n_k}, Y_{n_k}) thỏa

$$X_n^2 + Y_n^2 \leq 1$$

thì $(Z_k)_{k\in\mathbb{N}}$ là dãy iid phân phối đều trên hình tròn đơn vị $B = \{x \in \mathbb{R}^2 : |x| \leq 1\}$. Xác suất mỗi (X_n, Y_n) được chấp nhận là

$$p = P((X_n, Y_n) \in B) = \frac{|A \cap B|}{|A|} = \frac{|B|}{|A|} = \frac{\pi}{4} \approx 78.5\%$$

và số đề cử trung bình để sinh được mỗi Z_k là $1/p \approx 1.27$.

Diễn giải hình học

Mệnh đề G3. Cho $f : \mathbb{R}^d \to [0, \infty)$ là một hàm mật độ xác suất, đặt

$$A = \{(x, y) \in \mathbb{R}^d \times [0, \infty) : 0 \leq y < f(x)\} \subset \mathbb{R}^{d+1},$$

thì $|A| = 1$ và 2 phát biểu sau tương đương:

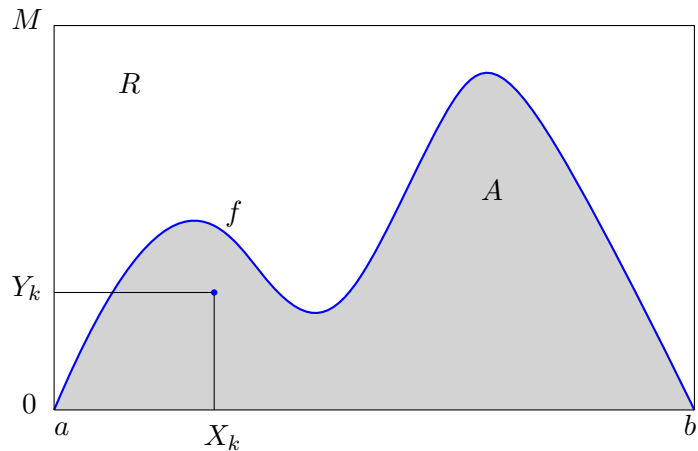
1. (X, Y) phân phối đều trên A.
2. X có phân phối với hàm mật độ f trên $\mathbb{R}^d$ và $Y = Uf(X)$ với $U \sim \mathcal{U}[0, 1]$, độc lập với X.

Diễn giải hình học

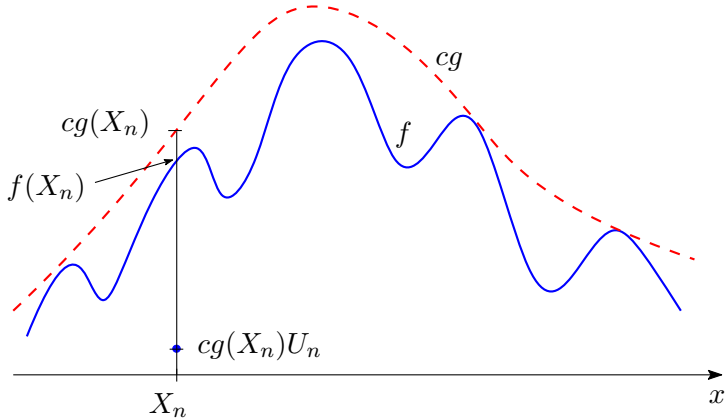
Một ứng dụng đơn giản của Mệnh đề G3 là chuyển một phân phối đều trên một tập con của $\mathbb{R}^2$ thành một phân phối trên $\mathbb{R}$ với hàm mật độ $f : [a, b] \to \mathbb{R}$ được cho. Để đơn giản, giả sử f có tập support nằm trong $[a, b]$ và $f(x) \leq M, \forall x \in [a, b]$, ta có thể sinh mẫu từ f như sau:

1. Sinh (X_k, Y_k) iid phân phối đều trên hình chữ nhật $R = [a, b] \times [0, M]$.
2. Xét tập $A = \{(x, y) \in \mathbb{R}^2 : y \leq f(x)\}$ và đặt $N = \min\{k \in \mathbb{N} : (X_k, Y_k) \in A\}$. Từ Mệnh đề G2, (X_N, Y_N) phân phối đều trên A.
3. Từ Mệnh đề G3, X_N có phân phối với hàm mật độ f.

Diễn giải hình học



Diễn giải hình học - Hiểu rõ lấy mẫu loại bỏ theo khuôn



Nội dung

Biến đổi của các biến ngẫu nhiên

Định lý TRV. (Transformation of Random Variables) Cho các tập mở $A, B \subset \mathbb{R}^d$, $\varphi : A \to B$ là một song ánh có các đạo hàm riêng liên tục, X là một biến ngẫu nhiên với giá trị trong A, $g : B \to [0, \infty)$ là một hàm mật độ xác suất, định nghĩa $f : \mathbb{R}^d \to [0, \infty)$ bởi

$$f(x) = \begin{cases} g(\varphi(x)).|\det D\varphi(x)| & \text{nếu } x \in A, \\ 0 & \text{khác.} \end{cases}$$

Ta có f là một hàm mật độ xác suất và X có hàm mật độ xác suất f khi và chỉ khi $\varphi(X)$ có hàm mật độ xác suất g.

Ở trên, $D\varphi(x)$ là **ma trận Jacobian** (Jacobian matrix) của φ tại x, là ma trận $d \times d$ với

$$D\varphi(x)_{ij} = \frac{\partial \varphi_i}{\partial x_j}(x), i, j = 1, 2, ..., d.$$

Trường hợp một chiều ($d = 1$), ta có $|\det D\varphi(x)| = |\varphi'(x)|$.

Biến đổi của các biến ngẫu nhiên - Ví dụ 1

Giả sử ta muốn sinh (X, Y) từ phân phối chuẩn tắc 2 chiều có hàm mật độ

$$g(x, y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}, (x, y) \in \mathbb{R}^2.$$

Ta có thể dùng biến đổi tọa độ cực φ cho bởi

$$\varphi(r, \theta) = (r\cos\theta, r\sin\theta), r > 0, \varphi \in (0, 2\pi).$$

φ là một song ánh từ tập mở $A = (0, \infty) \times (0, 2\pi)$ đến tập mở $B = \varphi(A) = \mathbb{R}^2 \backslash \{(x, y) : x \geq 0, y = 0\}$ trong $\mathbb{R}^2$. Lưu ý, nếu (X, Y) có phân phối chuẩn tắc 2 chiều thì

$$P\left((X, Y) \in \{(x, y) | x \geq 0, y = 0\}\right) = 0$$

nên $P\left((X, Y) \in B\right) = 1$.

Biến đổi của các biến ngẫu nhiên - Ví dụ 1 (tt)

Ma trận Jacobian của φ tại (r, θ) là

$$D\varphi(r,\theta) = \begin{pmatrix} \frac{\partial \varphi_1}{\partial r} & \frac{\partial \varphi_1}{\partial \theta} \\ \frac{\partial \varphi_2}{\partial r} & \frac{\partial \varphi_2}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix},$$

do đó $|\det D\varphi(r,\theta)| = |r\cos^2\theta + r\sin^2\theta| = r$.
Như vậy, dùng Định lý TRV ta có thể đưa việc sinh (X, Y) từ phân phối chuẩn tắc 2 chiều thành sinh (R, θ) từ phân phối có mật độ

$$f(r,\theta) = g(\varphi(r,\theta)).|\det D\varphi(r,\theta)| = \frac{1}{2\pi} e^{-r^2/2} r$$

trên $(0,\infty) \times (0, 2\pi)$.

Biến đổi của các biến ngẫu nhiên - Ví dụ 1 (tt)

Hàm mật độ $f(r, \theta)$ có thể được viết lại là

$$f(r, \theta) = \frac{1}{2\pi} e^{-r^2/2} r = \left(\frac{1}{2\pi}\right)\left(re^{-r^2/2}\right) = f_1(\theta) f_2(r).$$

Như vậy, (R, Θ) có thể được sinh từ $\Theta \sim \mathcal{U}(0, 2\pi)$ độc lập với R có hàm mật độ f_2. Kết hợp với Ví dụ trước, ta có thuật toán lấy mẫu từ phân phối chuẩn tắc 2 chiều như sau:

1. sinh $\Theta \sim \mathcal{U}(0, 2\pi)$,
2. sinh $U \sim \mathcal{U}[0, 1]$ và đặt $R = \sqrt{-2 \log U}$,
3. xuất $(X, Y) = \varphi(R, \Theta) = (R \cos \Theta, R \sin \Theta)$.

Phương pháp lấy mẫu này còn được gọi là **biến đổi Box-Muller** (Box-Muller transform).

Biến đổi của các biến ngẫu nhiên - Ví dụ 2

Giả sử ta muốn sinh Y từ phân phối có hàm mật độ

$$g(y) = \frac{3}{2}\sqrt{y}\mathbb{I}_{[0,1]}(y).$$

Ta có thể "phá căn" trong g bằng cách chọn $\varphi(x) = x^2$. Khi đó, dùng Định lý TRV với $A = B = (0,1)$ và $|\det D\varphi(x)| = |\varphi'(x)| = |2x|$, ta có

$$f(x) = g(\varphi(x)).|\det D\varphi(x)| = \frac{3}{2}x.2x = 3x^2, x \in [0,1].$$

Kết hợp với Ví dụ trước, ta có cách sinh Y như sau:

1. sinh $U \sim \mathcal{U}[0,1]$,
2. đặt $X = U^{1/3}$,
3. xuất $Y = \varphi(X) = X^2 = U^{2/3}$.

Phương pháp tỉ số đều

Định lý RUM. (Ratio-of-Uniforms Method) Cho $f : \mathbb{R}^d \to [0, \infty)$ với $Z = \int_{\mathbb{R}^d} f(x)dx < \infty$ và vector ngẫu nhiên X phân phối đều trên tập

$$A = \left\{ (x_0, x_1, ..., x_d) : x_0 > 0, \frac{x_0^{d+1}}{d+1} < f\left(\frac{x_1}{x_0}, ..., \frac{x_d}{x_0}\right) \right\} \subset [0, \infty) \times \mathbb{R}^d.$$

Ta có vector ngẫu nhiên

$$Y = \left(\frac{X_1}{X_0}, ..., \frac{X_d}{X_0}\right)$$

có hàm mật độ $\frac{1}{Z}f$ trên $\mathbb{R}^d$.

Phương pháp tỉ số đều - Ví dụ

Phân phối Cauchy có hàm mật độ

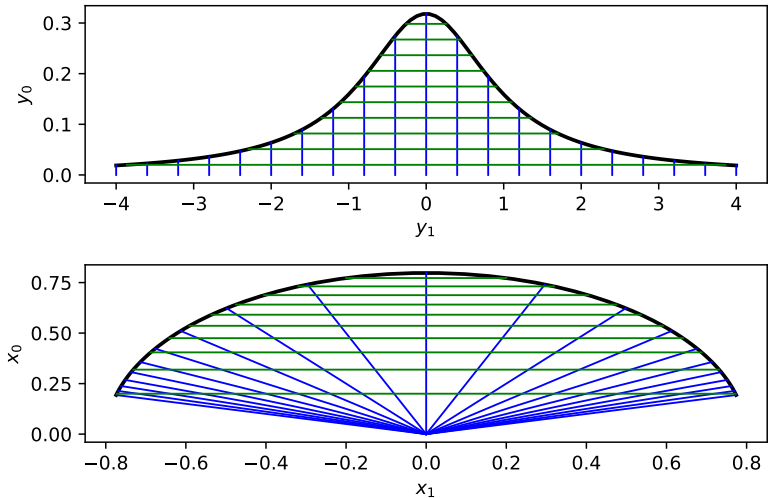
$$f(x) = \frac{1}{\pi(1+x^2)}.$$

Đặt

$$A = \left\{ (x_0, x_1) : x_0 > 0, \frac{x_0^2}{2} < f\left(\frac{x_1}{x_0}\right) = \frac{1}{\pi\left(1+\left(\frac{x_1}{x_0}\right)^2\right)} \right\}$$
$$= \left\{ (x_0, x_1) : x_0 > 0, x_0^2 + x_1^2 \leq \frac{2}{\pi} \right\}.$$

Dùng định lý RUM, nếu (X_0, X_1) phân bố đều trên A thì $Y = X_1/X_0$ có phân phối Cauchy.

Phương pháp tỉ số đều - Ví dụ (tt)



Nội dung

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

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.