ORIE 4580/5580: Simulation Modeling and Analysis

ORIE 5581: Monte Carlo Simulation

Unit 5: Generating Non-Uniform Random Variables

Sid Banerjee

School of ORIE, Cornell University

generating rvs with arbitrary distributions

aim: "transform" $U[0,1]$ rv to another rv with given probability distribution.

monte carlo sampling techniques

basic methods

- inversion
- acceptance-rejection
- distribution-specific techniques (Box-Muller for Gaussians)
- advanced techniques (adaptive rejection sampling, SIR)

markov-chain monte carlo (MCMC)

- Metropolis-Hastings
- Gibbs sampling
- advanced techniques (slice sampling, Hamiltonian MC)

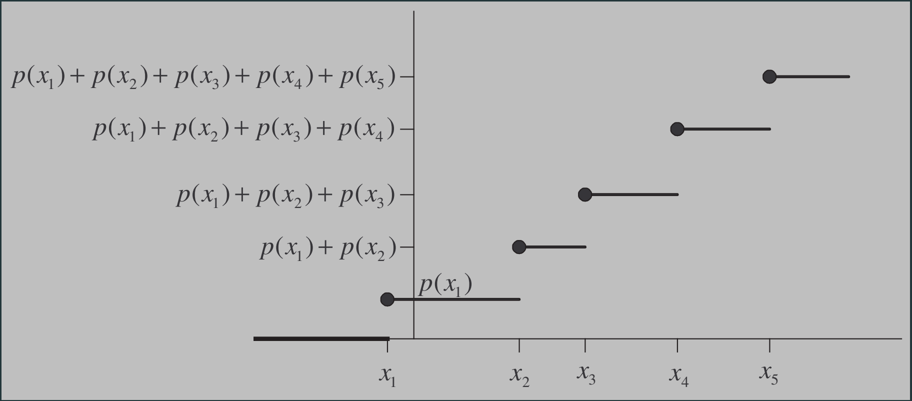
inversion

warm-up: simulating discrete rv

X takes values $x_1 \leq x_2 \leq \ldots \leq x_5$, $\mathbb{P}[X = x_i] = p(x_i)$

warmup: simulating discrete rv

X takes values $x_1 \leq x_2 \leq \ldots \leq x_5$, $\mathbb{P}[X = x_i] = p(x_i)$



the inversion method

X continuous r.v. with pdf f and c.d.f. $F(\cdot)$

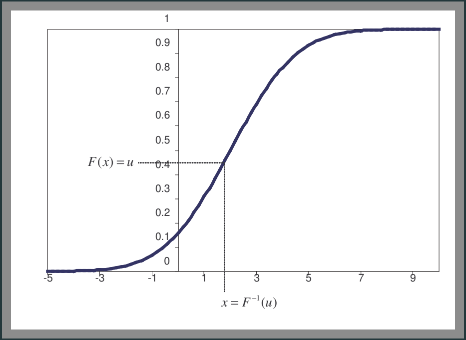
- want to generate samples of X.
- $F(\cdot)$ non-decreasing $\implies$ can define inverse $F^{-1}(\cdot)$
- $F(x) = u \iff F^{-1}(u) = x$

inversion method

given desired cdf F (continuous, increasing), generate sample $X_0 \sim F$ as:

1. generate $U \sim U[0, 1]$.
2. return $X_o = F^{-1}(U)$.

intuition/proof for inversion method



example

example – the pdf of X is given by

$$f(x) = \begin{cases} 1/3 & \text{if } 0 \leq x \leq 1 \\ 2/3 & \text{if } 1 < x \leq 2 \\ 0 & \text{otherwise.} \end{cases}$$

develop an inversion method to generate samples of X.

example (contnd)

example (exponential rv)

generate samples of an exponential r.v. with parameter λ, with cdf

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

drawback of inversion method

- inversion method may be computationally expensive.
- computing $F^{-1}(\cdot)$ may require numerical search.

example – the pdf of X is given by

$$f(x) = \begin{cases} 60x^3(1-x)^2 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

$$F(x) = 15x^4 - 24x^5 + 10x^6 \qquad \text{for } 0 \leq x \leq 1.$$

generate samples of X by using the inversion method.

generating normal random variables

- method 1: inversion
  - no closed form for $\phi^{-1}(x)$
  - inversion done numerically
- method 2: via the central limit theorem.
  - generate $U_1, U_2, \ldots$
  - scale and center appropriately
  - not exact!

clicker question: normal from Uniform

we generate 12 samples $U_1, U_2, \ldots, U_{12}$ from $U[0,1]$ distribution
to use these to generate a sample with distribution close to a $\mathcal{N}(0,1)$ rv, we should set: (note: $\mathbb{E}[U_1] = 0.5$, $Var(U_1) = 1/12$)

(a) $X = \sum_{i=1}^{12} U_i - 12$

(b) $X = \sum_{i=1}^{12} U_i - 6$

(c) $X = \frac{1}{\sqrt{12}} \left(\sum_{i=1}^{12} U_i - 6 \right)$

(d) $X = \frac{1}{12} \left(\sum_{i=1}^{12} U_i - 6 \right)$

(e) None of the above

clicker question: solution

we generate 12 samples $U_1, U_2, \ldots, U_{12}$ from $U[0,1]$ distribution
to use these to generate a sample with distribution close to a $\mathcal{N}(0,1)$ rv, we should set: (note: $\mathbb{E}[U_1] = 0.5, Var(U_1) = 1/12$)

the Box-Muller method

generates a pair of $\mathcal{N}(0,1)$ rvs

- $N_1 \sim \mathcal{N}(0,1), \quad N_2 \sim \mathcal{N}(0,1), \qquad N_1 \perp\!\!\!\perp N_2$
- the point (N_1, N_2) can be expressed in polar coordinates as

$$(N_1, N_2) = (R\cos\theta, R\sin\theta)$$

the Box-Muller method

$$(N_1, N_2) = (R\cos\theta, R\sin\theta)$$

- $\theta \sim U[0, 2\pi]$, and is independent of R.
- $R = \sqrt{N_1^2 + N_2^2} = \sqrt{2X}$, where $X \sim Exp(1)$

Box-Muller Algorithm

1. generate $U_1 \sim U[0,1]$, $U_2 \sim U[0,1]$.
2. set

   $R =$ $\theta =$

3. set

   $N_1 =$ $N_2 =$

clicker question: inversion for sampling from unit disc

want to generate (X, Y) uniform over the unit disc, i.e., over $\{(x, y)|x^2 + y^2 \leq 1\}$ given $U, V \sim U[0, 1]$ i.i.d rvs, which of the following gives the correct sample?

(a) $R = U, \Theta = 2\pi V$ and $(X, Y) = (R\cos\Theta, R\sin\Theta)$

(b) $R = \sqrt{U}, \Theta = 2\pi V$ and $(X, Y) = (R\cos\Theta, R\sin\Theta)$

(c) $R = U^2, \Theta = 2\pi V$ and $(X, Y) = (R\cos\Theta, R\sin\Theta)$

(d) $R = 2U - 1, \Theta = \pi V$ and $(X, Y) = (R\cos\Theta, R\sin\Theta)$

(e) None of the above

solution: inversion for sampling from unit disc

want to generate (X, Y) uniform over the unit disc, i.e., over $\{(x, y)|x^2 + y^2 \leq 1\}$
given $U, V \sim U[0, 1]$ i.i.d rvs, which of the following gives the correct sample?

acceptance-rejection

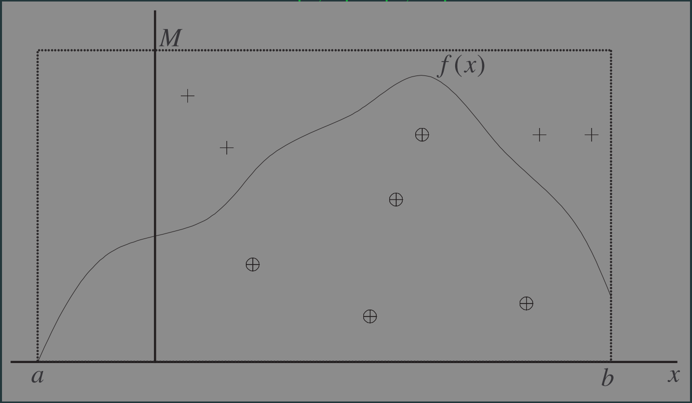
acceptance-rejection

want to generate samples of a rv X

- pdf $f(\cdot)$ of X takes positive values only over $[a, b]$
- M is an upper bound on pdf of X, i.e., $M \geq \max_{x \in [a,b]} f(x)$

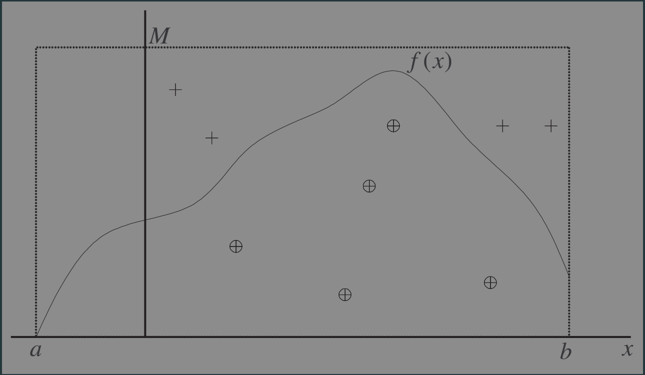
  $\Rightarrow$ can enclose pdf in the rectangle

$$[a, b] \times [0, M]$$



acceptance-rejection

want samples of a rv $X \in [a, b]$, with pdf $f(x) \leq M$



acceptance-rejection sampling

1. generate $U_1, U_2 \sim U[0, 1]$
2. set $Z_1 = a + (b - a)U_1$, $Z_2 = MU_2$
3. if $Z_2 \leq f(Z_1)$, return $X_o = Z_1$; else, reject and repeat

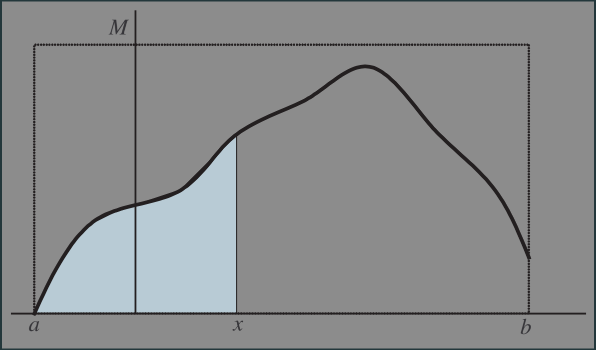
AR sampling: proof of correctness

let X_o denote the output of the AR method for cdf F

- $F_{X_o}(x) = \mathbb{P}[X_o \leq x] =$

AR sampling: proof of correctness

observe: $\mathbb{P}[Z_1 \leq x,\ Z_2 \leq f(Z_1)] = \mathbb{P}[(Z_1, Z_2) \in \text{shaded region in figure}]$



AR sampling: running time

how many $U[0, 1]$ samples do we need for one sample of X?

example: X has pdf

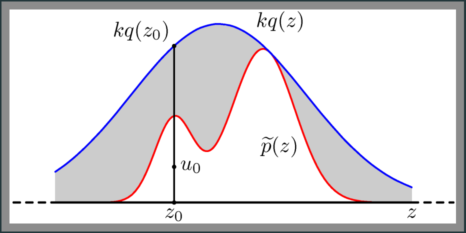
$$f(x) = \begin{cases} 6x(1-x) & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

for rejection sampling, we choose

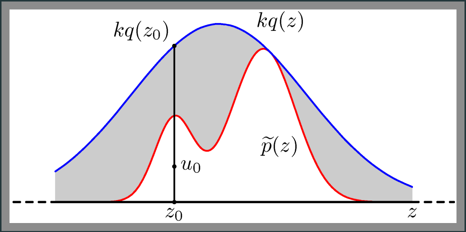
$a =$ $b =$ $M =$

on average, per sample we require

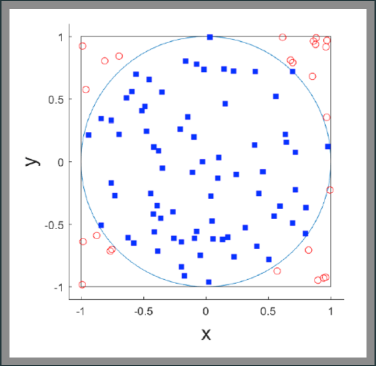
generalized AR sampling



generalized AR sampling



(even more) generalized AR sampling



clicker question: ordering conditional expectations

consider rv $X \in \mathbb{R}$ with cdf F, and any $a \in \mathbb{R}$; then

(a) $\mathbb{E}[X] \geq \mathbb{E}[X|X \geq a]$

(b) $\mathbb{E}[X] \leq \mathbb{E}[X|X \geq a]$

(c) depends on if a is positive or negative

(d) depends both on a and F

AR sampling: challenges in high dimensions