till now - Generating $U[0,1]$

ORIE 4580/5580: Simulation Modeling and Analysis

ORIE 5581: Monte Carlo Simulation

Unit 5: Generating Non-Uniform Random Variables

Tuesday

- Inversion method

Thursday

- Gaussian rv (Box-Muller)
- Acceptance-Rejection

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 (static methods)

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

Thursday

markov-chain monte carlo (MCMC) (dynamic methods)

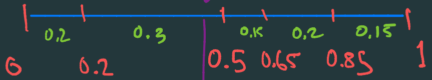
- Metropolis-Hastings
- Gibbs sampling
- advanced techniques (slice sampling, Hamiltonian MC)

inversion

warmup: simulating discrete rv

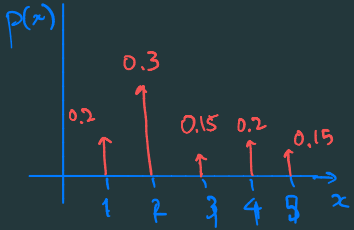
X takes values $x_1 \leq x_2 \leq \ldots \leq x_5$, $\mathbb{P}[X = x_i] = p(x_i)$

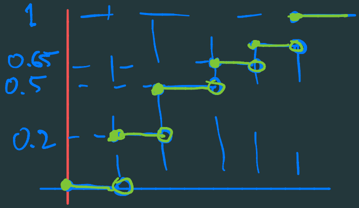
$U_1 \sim U[0,1]$



$X_1 = 2$

(ie, x_2)

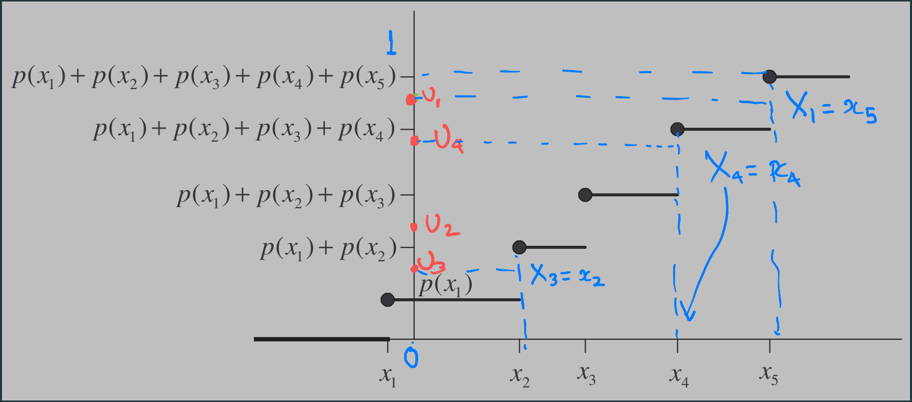




warmup: simulating discrete rv

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

$U_1, U_2, \ldots \sim U[0,1]$, iid



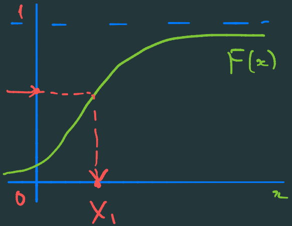
the inversion method

(or discrete or mixed)

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 $F^{-1}(\cdot)$
- $F(x) = u \iff F^{-1}(u) = x$ (always exists if F is increasing)

$U[0,1] \sim U_1$



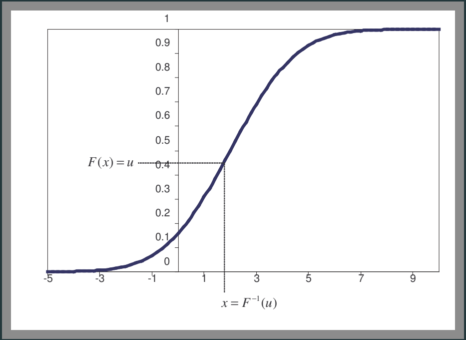
inversion method

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

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

(For F non-dec - 'generalized inverse' (see HW 3))

intuition/proof for inversion method



<u>Intuition</u> – 'limit' of staircase graph

<u>Pf</u> – Recall $X = F^{-1}(U)$, $U \sim U[0,1]$

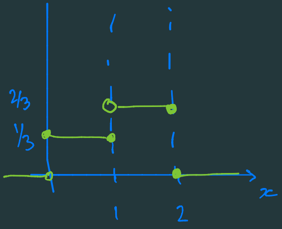
To find dist of X, <u>compute F_X</u>

$$F_X(y) = \mathbb{P}[X \le y] = \mathbb{P}[F^{-1}(U) \le y] = \mathbb{P}[U \le F(y)]$$

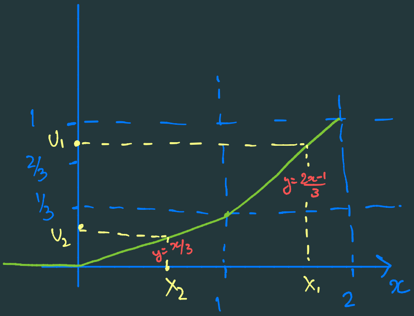
$$= F(y) \quad \left(\text{cdf of } U[0,1] \text{ is } \begin{cases} 0 ; & x \le 0 \\ x ; & x \in [0,1] \\ 1 ; & x \geq 0 \end{cases} \right)$$

example – the pdf of X is given by

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



develop an inversion method to generate samples of X.



$$F^{-1}(y) = \begin{cases} 3y & ; \; y \in [0, 1/3) \\ \dfrac{3y+1}{2} & ; \; y \in [1/3, 1] \end{cases}$$

Set $X = F^{-1}(U)$, $U \sim U[0,1]$

example (exponential rv)

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

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

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & ; x \geq 0 \\ 0 & ; x < 0 \end{cases}$$

If $y = F(x) = 1 - e^{-\lambda x}$ $(x \geq 0)$ $\Rightarrow$ $F^{-1}(y) = x = \frac{-1}{\lambda} \ln(1-y)$ $(y \in [0,1])$

$\Rightarrow$ Set $X = -\frac{1}{\lambda} \ln(1-U)$ for $U \sim U[0,1]$

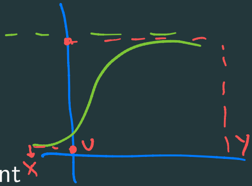
$\left(\text{or } X = \frac{-1}{\lambda} \ln(U) \right)$

clicker question: U and $1-U$

$U \sim U[0,1]$ and F is a cdf
set $X = F^{-1}(U)$, and also set $Y = F^{-1}(1-U)$; then (same U)

10 (a) X has cdf F but Y does not.

8 (b) Y has cdf F but X does not.



24 (c) both X and Y have cdf F, and X and Y are independent

60 (d) both X and Y have cdf F, and X and Y are dependent

If X big $\Leftrightarrow$ Y small (since $U \sim 0 \Leftrightarrow 1-U \sim 1$)

- What is dist of $1-U$ $\left(\mathbb{P}[1-U \le x] = \mathbb{P}[U \ge 1-x] = 1-F(1-x) = 1-(1-x) = x\right)$

$\Rightarrow 1-U \sim U[0,1]$

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

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

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

generate samples of X by using the inversion method.

If $y = 15x^4 - 24x^5 + 10x^6$, what is x?

generating normal random variables

- method 1: inversion
  - no closed form for $\phi^{-1}(x)$
  - inversion done numerically

(no closed form for $\Phi(x) = \int_{-\infty}^{x} f(x)\,dx$!)

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

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

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

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

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

4 (e) None of the above

Want to check $\mathbb{E}[X] = 0$, $Var(X) = 1$

$$\mathbb{E}\left[\frac{1}{\sqrt{12}}\underbrace{\left(\sum_{i=1}^{12}\left(U_i - \tfrac{1}{2}\right)\right)}_{0 \text{ mean}}\right] = 0$$

indep

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

$$= \frac{1}{12} \neq 1$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)$$

$\sqrt{12}$ $\sqrt{12}$

See Soln in next slide

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: $\mathbb{E}[U_1] = 0.5$, $Var(U_1) = 1/12$)

Want $X = a \sum_{i=1}^{12} U_i + b$

$\mathbb{E}[X] = a \sum_{i=1}^{12} \mathbb{E}[U_i] + b = 6a + b = 0$ (1)

$Var(X) = a^2 \sum_{i=1}^{12} Var(U_i) = a^2 = 1$ (2)

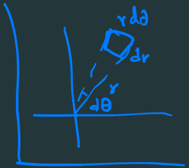
$(\because U_i$ indep$)$

$\Rightarrow \boxed{a = 1, \quad b = -6}$

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

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



Informally

$f(x)\,f(y)$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\cdot\frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, e^{-y^2/2}\, dx\, dy = 1$$

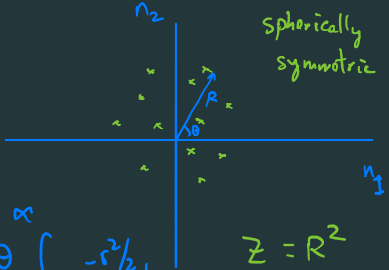
$x \quad y$

$\theta \quad r$

$$\Downarrow$$

$$\int_0^{2\pi}\int_0^{\infty} \frac{1}{2\pi}\, e^{-r^2/2}\, r\, dr\, d\theta = \int_0^{2\pi} \frac{1}{2\pi}\, d\theta \int_0^{\infty} \underbrace{r e^{-r^2/2}}_{f(r)}\, dr$$

$f(\theta) \equiv U[0, 2\pi]$



spherically symmetric

$Z = R^2$

$$\int_0^{\infty} \underbrace{\tfrac{1}{2} e^{-z/2}}_{f(z)}\, dz$$

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 $X \sim Exp(1)$ $\longleftarrow$ $\left(R^2 \sim \mathcal{E}xp(1/2)\right)$

Box-Muller Algorithm

1. generate $U_1 \sim U[0,1]$, $U_2 \sim U[0,1]$. , Set $X = \underline{-\ln(U_1)}$ $\leftarrow$ F^{-1} for $\mathcal{E}xp(1)$
2. set

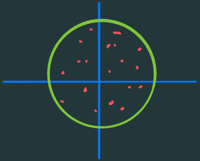
$$R = \sqrt{2X} \qquad \theta = 2\pi U_2$$

3. set

$$N_1 = R\cos\theta \qquad N_2 = R\sin\theta$$

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 $(X, Y) = (R\cos\Theta, R\sin\Theta)$

21

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

55

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

19

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

5

(e) None of the above

solution: inversion for sampling from unit disc

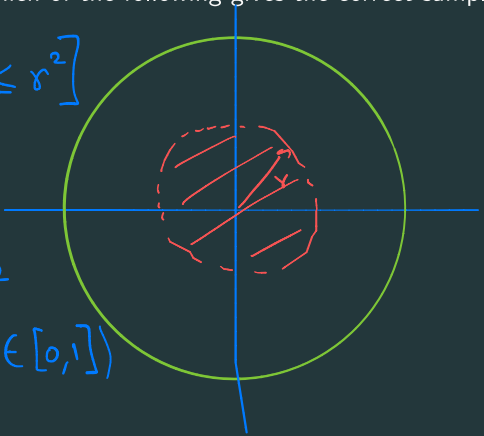
want to generate (X, Y) uniform over the unit disc, i.e., over $\{(x, y) | x^2 + y^2 \leq 1\}$
given $U, V \sim U[0, 1]$ i.i.d rvs, which of the following gives the correct sample?

$$F_R(r) = \mathbb{P}\left[X^2 + Y^2 \leq r^2\right]$$

$$= \frac{\text{Area of red } \odot}{\text{Area of green } \odot}$$

$$= \frac{\pi r^2}{\pi} = r^2 \quad (r \in [0,1])$$

$$\Rightarrow F_R^{-1}(x) = \sqrt{x}$$



acceptance-rejection

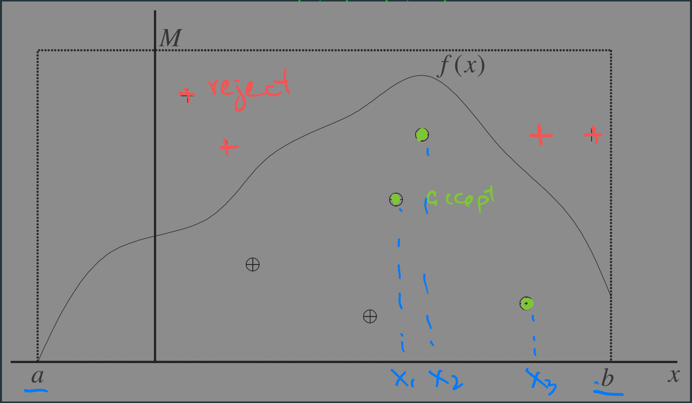
acceptance-rejection

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., $M \geq \max_{x \in [a,b]} f(x)$

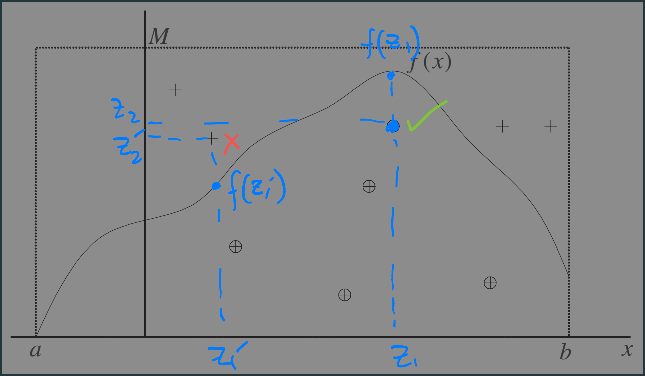
  $\Rightarrow$ can enclose pdf in the rectangle

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



acceptance-rejection

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



acceptance-rejection sampling

1. generate $U_1, U_2 \sim U[0, 1]$ want $Z_1 \sim U[a,b]$, $Z_2 \sim U[0,M]$
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 (ie, accept if below $f(z)$)

AR sampling: proof of correctness

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

- $F_{X_o}(x) = \mathbb{P}[X_o \le x] =$ $\mathbb{P}[Z_1 \le x \mid Z_2 \le f(Z_1)]$

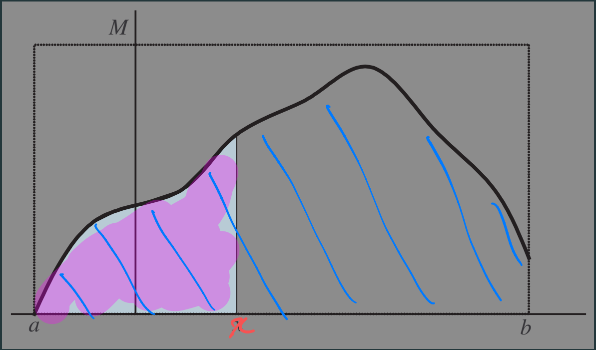
$\left(\text{ie, } \mathbb{P}[Z_1 \le x \mid (Z_1, Z_2) \text{ are accepted}]\right)$

$$= \frac{\mathbb{P}[Z_1 \le x, Z_2 \le f(Z_1)]}{\mathbb{P}[Z_2 \le f(Z_1)]}$$

$\left(\text{ie, } \mathbb{P}[(Z_1, Z_2) \text{ accepted}]\right)$

AR sampling: proof of correctness

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



$\mathbb{P}[(Z_1,Z_2) \text{ accepted}]$

$= \mathbb{P}[(Z_1,Z_2) \text{ in hatched region}]$

$= 1/c$

(where c = Area of rect $= M(b-a)$)

$$\Rightarrow F_{X_0}(x) = \frac{\mathbb{P}[(Z_1,Z_2) \text{ in shaded region}]}{\mathbb{P}[(Z_1,Z_2) \text{ in hatched region}]} = \frac{F(x)}{1} = F(x)$$

AR sampling: running time

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

- For each (z_1, z_2), need 2 $U[0,1]$ samples
- How many (z_1, z_2) do we need till one acceptance?

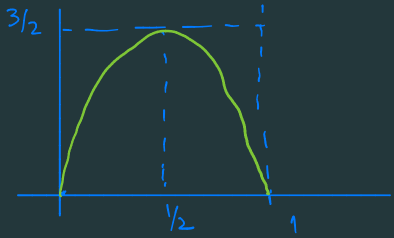
Let N = # of trials till first acceptance

$$N \sim \text{Geometric}\left(\mathbb{P}[\text{acceptance}]\right) \quad \left(\text{Geom}\left(\frac{1}{M(b-a)}\right)\right)$$

$$\Rightarrow \mathbb{E}[N] = M(b-a)$$

$\Rightarrow$ Need $2M(b-a)$ $U[0,1]$ for one sample (on average) *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

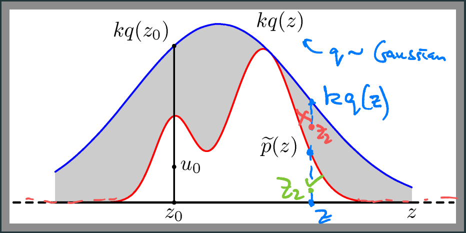
$a =$ 0 $\qquad b =$ 1 $\qquad M =$ 3/2

on average, per sample we require

$2M(b-a) = 3$ $U[0,1]$

samples on average

generalized AR sampling

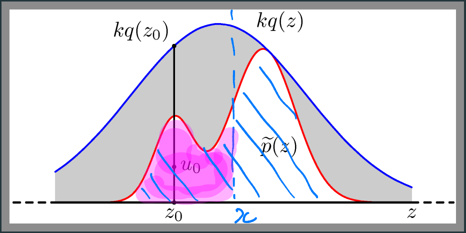


- Want $X \sim \tilde{p}(x)$ (red pdf)
- Have samples of $Z \sim q(x)$ (blue pdf)
- Suppose $\max_x \frac{\tilde{p}(x)}{q(x)} \leq k$ $(k \geq 1)$

- Sample Z_1 from pdf $q(z)$, $Z_2 \sim k\, q(Z_1)\, U[0,1]$
- Accept (Z_1, Z_2) if $Z_2 \leq \tilde{p}(Z_1)$ ← (i.e., accept with prob $\tilde{p}(Z_1)/kq(Z_1)$)
  - Output Z_1

(Claim - need $2k$ $U[0,1]$ samples on avg for 1 sample of X)

generalized AR sampling - Proof



- Let $X_0 \equiv$ output of AR

$$\Rightarrow F_{X_0}(x) = \mathbb{P}[X_0 \leq x]$$

$$= \mathbb{P}[Z_1 \leq x \mid Z_2 \leq \tilde{p}(Z_1)]$$

- $\mathbb{P}[Z_2 \leq \tilde{p}(Z_1)] = \mathbb{P}[\text{Accept}] = \frac{\text{Area in } \boxed{////}}{k}$ $=1$

k Area under blue curve

$$\mathbb{P}[Z_1 \leq x, Z_2 \leq \tilde{p}(Z_1)] = \frac{\text{Area in } \boxed{}}{k}$$

$= F(x)$

$\left(= \int_{-\infty}^{x} \tilde{p}(x)dx \right)$

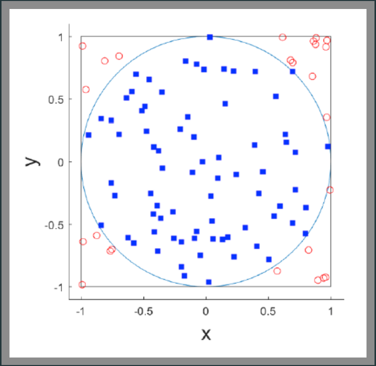
$$\Rightarrow F_{X_0}(x) = F(x)\ !$$

(even more) generalized AR sampling

For any 'event' E_1

- Suppose we accept points for which E_1 is true

Then for any other event E_2



$$\Rightarrow \quad \mathbb{P}[E_2 \text{ holds for accepted points}] = \mathbb{P}[E_2 \mid E_1]$$

clicker question: ordering conditional expectations

consider rv $X \in \mathbb{R}$ with cdf F, and any $a \in \mathbb{R}$; then (assume $\mathbb{P}[X \geq a] > 0$)

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

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

20 (c) depends on if a is positive or negative

20 (d) depends both on a and F

Thought Expt - $a = 5$, take empirical avg of $n=4$ pts

$X :=$ 1 4 6 8 3 10 5

$\mathbb{E}[X] \sim 19/4$ $\mathbb{E}[X|X>5] \sim \frac{29}{4} > \frac{19}{4}$

AR sampling: challenges in high dimensions