1. Monte Carlo Methods

Independent Sampler, Multivariate Normal Sampling, Monte Carlo

Sun Woo Lim

Mar 4, 2022

Motivation and Important Note

Motivation
That you know the CDF/PDF of a distribution does not mean you can easily sample from that distribution.
This slide deals with computational methods to generate random samples.

No such truly "random" number generator
A deterministic computer algorithm does not allow generating truly "statistically random" sample. We just generate "pseudo-random" sample: deterministic but looking statistically random.

Def)Identical in Distribution
Two random variables (generalized to random vectors easily) X and Y are identically distributed when $F_X(t) = F_Y(t), \forall t \in \mathbb{R}$

Key takeaway) Does X follow F? What distribution does X follow?

- Case1) I am who sample X knowing the sampling procedure
  $\rightarrow$ X is sample from F if CDF of X is F (this statement looks trivial but is most important!)
  $\rightarrow$ **change of variable**: powerful for sampling from exotic dist'n by transformation from known (how to sample) dist'n
  e.g) Say I can generate $U \sim U(0,1)$. Then, $V := a + (b-a)U$ is a RV from $U(a,b)$ since $V \stackrel{d}{=} W \sim U(a,b)$.
- Case2) I have the sample $X = x$ but do not know the sampling procedure
  $\rightarrow$ Use nonparametric test based on Empirical distribution (not of our interest of this slide)

Independent sample

All starts from $U(0,1)$

✓ 1. Midsquare method (von Neumann & Metropolis (1940s)

1st) Start with any 4 digit $n_0 \in \mathbb{N}$ (Usually system time: deterministic)

2nd) Take **middle** 4 digits of n_0^2 and divide by 10000.

3rd) Iterating 2nd) gives independent looking uniform (0,1) sequence.

✓ 2. Linear Congruential Method

A, C, M, X_0 be natural numbers s.t. $a, c, X_0 < M$.

1st) $X_{i+1} = (AX_i + C) mod M$ gives $X_{i+1} \in \{0, 1, ..., M-1\}$

2nd) $R_{i+1} := X_{i+1}/M$ and take $R_{i+1} \in [0, \frac{m-1}{m}]$

3rd) Iterating 1st) and 2nd) gives **looking statistically independent** and **looking** $Unif(0,1)$ sequence.

i	Z_i	U_i	$Z_i \times Z_i$
0	7182	-	51581124
1	5811	0.5811	33767721
2	7677	0.7677	58936329

(a) Mid-square method

i	X_i $X_0=1$	X_i $X_0=2$	X_i $X_0=3$	X_i $X_0=4$
0	1	2	3	4
1	13	26	39	52
2	41	18	59	36
3	21	42	63	20
4	17	34	51	4

(b) Linear Congruential Method ($A = 13, C = 0, M = 64$)

Technique 1. Inverse Transform Sampling

Fundamental Theory: Probability Integral Transform

Let $U \sim Unif(0,1)$. Let F be the CDF I want to generate sample from and assume F is differentiable. Then,
$X := F^{-1}(U) \sim F$

pf) $F_X(x) = P(X \leq x) = P(F^{-1}(U) \leq x) = P[F(F^{-1}(U)) \leq F(x)] = P(U \leq F(x)) = \int_0^{F(x)} 1dt = F(x)$.

Inverse Transformation Sampling in Continuous Case

1. Generate $U \sim Unif(0,1)$ using a method addressed in the previous slide.
2. $X := F^{-1}(U)$ is a sample from F.

✓ To generate **iid sequence** of samples from F, just generate iid sequence of $U(0,1)$ random variables.

Example: (iid sequence of) exponential distributed random variable(s)

✓ CDF of exponential distribution: $F_X(x) = 1 - exp(-\lambda x)I(x \geq 0)$.

$\rightarrow X = -\frac{1}{\lambda}log(1-U)$ is a (pseudo) random sample from exp(λ).

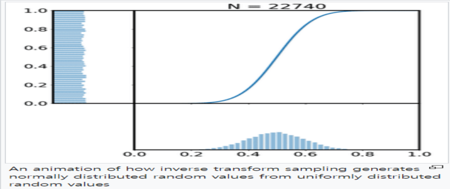


Figure: Image from https://en.wikipedia.org/wiki/Inverse_transform_sampling

Inverse Transform Sampling in Discrete Case: Use General Inverse Function

Def) General Inverse Function $F_{General}^{-1}(p) := inf\{x \in \mathbb{R} | F(x) \geq p\}$

Inverse Transformation Sampling in Discrete Case

$$X = \begin{cases} x_1 & w.p \ \mathsf{p}_1 \\ x_2 & w.p \ \mathsf{p}_2 \\ \vdots \\ x_n & w.p \ \mathsf{p}_n \end{cases}$$

1st. Generate $U \sim Unif(0,1)$ using a method addressed in the previous slide.

2nd. $X := F_{General}^{-1}(U)$ is a sample from F.

→ 2nd step meaning: Find i s.t. $\sum_{k=1}^{i-1} p_k \leq U < \sum_{k=1}^{i} p_k$ then x_i is the sampled value from F.

Examples: Discrete Uniform Distribution, Poisson Distribution, Binomial Distribution, etc

Drawbacks of Inverse Transformation Sampling

1. Hardship: many cases, hard to obtain inverse function of Continuous CDF
2. Inefficiency: discrete R.V, takes on numerous values

Technique 2. Basic Change of Variable

Fundamental Theory: Change of Variable
Let X have pdf $f_X(x)$ and $Y := g(X)$, g : monotone, invertible function. Let $f_X(x)$ be continuous on its support and $g^{-1}(y)$ has continous derivative on its support.

Then, $f_Y(y) = f_X(g^{-1}(y))|\frac{dg^{-1}(y)}{dy}|I(y \in spt(Y))$

Example)
1) Generate $U(a,b)$ from $V := a + (b-a)U, U \sim Unif(0,1)$

2) Generate exp(λ) from $X = -\frac{1}{\lambda}log(1-U) \overset{d}{=} -\frac{1}{\lambda}log(U)$ using $U \overset{d}{=} 1-U$
pf that $U \overset{d}{=} 1-U$: $Pr(U \le u) = Pr(1-u \le U) = 1-(1-u) = u$

3) Various results of addition of iid random variables: Normal, Poisson, Gamma (as sum of iid exponential), etc.

To generate standard normal random variable, use **Box-Muller method** which uses change of variable technique.

Generate iid sequence of $N(\mu,\sigma^2)$ random variables using $M \sim N(\mu,\sigma^2) \overset{d}{=} \sigma Z + \mu, Z \sim N(0,1)$

So, **identical in distribution** is all I need.

Technique 3. Factoring out the joint as marginal and conditional

$p(x_1, ..., x_n) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \dots p(x_n|x_1, \dots, x_{n-1})$.

Example) **Finite mixture of Gaussians**

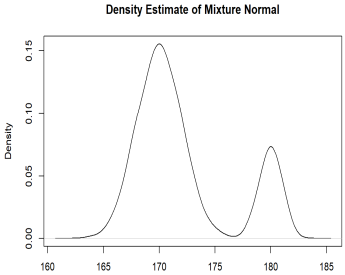
Given a finite set of pdf's $p_1(x), ..., p_H(x)$ and weights $w_1, ..., w_H$ s.t. $\sum_{i=1}^{H} w_i = 1$,
the **pdf of mixture of H distributions** is $f_X(x) = \sum_i w_i p_i(x)$. e.g) Height of total population, final scores of a typical class

Sampling from finite Gaussian mixture pdf with weights, means, variances known

$f_X(x|w_1, ..., w_H, \theta_1, ..., \theta_H, \sigma_1, ..., \sigma_H) = w_1 f(x|\theta_1, \sigma_1^2) + ... + w_H f(x|\theta_H, \sigma_H^2)$ where $f_i(x|\theta_i, \sigma_i^2) = dnorm(x, \theta_i, \sigma_i^2)$

Our example) Height of men(1) and women(2): $H = 2$, $w_1 = 0.2, w_2 = 0.8$, $\mu_1 = 180$, $\mu_2 = 170$, $\sigma_1^2 = 1, \sigma_2^2 = 4$.

```
mixture_normal_samp = function(n, group_probs, mean, sd){
  # n : number of samples
  # group_probs (= weights) w_1,..,w_H
  # mean : (theta_1,..., theta_H) : mean of each group
  # sd : (sigma_1, ..., sigma_H) : sd of each group

  H = length(group_probs) ; samples = rep(NA, n)

  group_sample = sample(1:H, prob = group_probs,
                        size = n, replace = T)

  for(i in 1:n){
    group = group_sample[i]
    samples[i] = rnorm(n = 1, mean = mean[group], sd = sd[group])
  }

  plot(density(samples),
       main = "Density Estimate of Mixture Normal")
}


mixture_normal_samp(10000, c(0.2, 0.8), c(180,170),c(1,2))
```



Technique 4. Acceptance-Rejection Sampling (Rejection Sampling)

Conditions to use A-R method

Suppose X, having pdf f be the random variable I want to generate. Density function has to be known but directly sampling from F is hard.
Need Y having pdf g that 1) I can directly generate, 2) the support of Y covers the support of X.

Acceptance Rejection Sampling

Idea) Generate $X \sim f(x)$ by accepting or rejecting sample Y from **proposal** pdf $g(y)$ where $spt(X) \subseteq spt(Y)$.

1. Get $c > 0$ s.t. $c \cdot g(x) > f(x), \forall x \in spt(Y)$. Best choice for "c" is $sup_y \frac{f(y)}{g(y)}$
2. Sample $Y \sim g(y) \perp U \sim Unif(0,1)$
3. $X = Y$ if $U \leq \frac{f(Y)}{cg(Y)}$. Else, iterate the 2nd step until the inequality is satisfied.

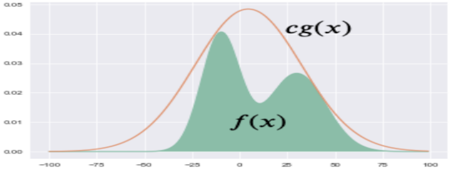


Figure: Image from https://medium.com/@msuhail153/rejection-sampling-6c4510da24f8.

Proof

- 1. Discrete case: WTS) $Pr(X = i) = Pr(Y = i|Proposal\ Accepted)$. Use def'n of conditional distribution in proof.
- 2. Continuous case: WTS: $Pr(Y \leq y|Proposal\ Accepted) = Pr(Y \leq y|U \leq \frac{f(Y)}{cg(Y)}) = F(y)$. Use Bayes thm in proof.

Key Takeaway

1. $Pr(Acceptance) = \frac{1}{c}$. $\#Proposal$ (which is a R.V!) $\sim Geom(\frac{1}{c})$ and $E(\#Proposal) = c$.
   Thus, computation is efficient if c is small as possible with the constraint: $cg(x) > f(x)$: constrained optimization!
2. High acceptance probability is always good (not always true in Metropolis Hastings Algorithm addressed later!)
   Acceptance probability is high when the proposal density $\approx$ sampling density
   ex) Sampling density: truncated standard normal distribution $f(z|z > a)$, a: large. Proposal density: standard normal
   $\rightarrow$ Accept if $Z > a$, reject if not.
   $\downarrow$ acceptance probability because the truncated normal distribution looks very different from normal distribution if $\uparrow a$.
3. Rejected proposal does not count as the next sample. cf) MH algorithm: rejected proposal counts as the next sample

Example) Sampling Standard Gaussian RV by acceptance-rejection : Note) Box-Muller is more widely used.

$p(x) = \frac{1}{\sqrt{2\pi}} exp(-x^2/2) I(x \in \mathbb{R})$

Hard to find a distribution that 1) has support of the whole real line 2) able to sample from.

Idea) Sample the absolute value of X and then sample the sign.

Let $Y := |X|$. Then, $f_Y(y) = \frac{2}{\sqrt{2\pi}} exp(-y^2/2) I(0 < y < \infty)$.

Then, take $exp(1)$ as the proposal distribution, which has support $\mathbb{R}$ and easy to sample.

$c = sup_x \frac{f(x)}{g(x)} = \frac{1}{\sqrt{2\pi}} \cdot sup_x \frac{exp(-x^2/2)}{exp(-x)} = \sqrt{\frac{2e}{\pi}}$ when $x^* = 1$.

Then, $\frac{f(x)}{cg(x)} = exp(-\frac{-x^2}{2} + x - \frac{1}{2}) = exp(-\frac{(x-1)^2}{2})$.

Steps to obtain rnorm(n, 0, 1)

1. Sample $Z \sim exp(1) = -log(U_1)$ and $U_2 \sim Unif(0, 1)$ independently.
2. If $U_2 \leq exp(-\frac{(z-1)^2}{2})$, set $Y = Z$. Else, keep repeating 1)
3. Sample $U_3 \sim Unif(0, 1)$. $X = ifelse(U_3 \leq \frac{1}{2}, Y, -Y)$
4. Repeat 1) through 3) n number of times. All you need is $3n$ number of $U(0, 1)$'s $\cdot$

Diagnostic Questions

1. How does each of U_1, U_2, U_3 work?
2. Was 'only' the A-R method used in this procedure? If not, what else is used?
3. How sample $rnorm(n, \mu, \sigma)$ for general mean and standard deviation?

Sampling Correlated Random Vector

- Until now, learned how to get iid sequence of arbitrary distribution.
- Then, how about sampling random vector (X, Y) where $X \sim N(0, 1)$, $Y \sim Gamma(2, 3)$ and $corr(X, Y) = -0.5$?
- This generally requires **copula**, which is challenging.
- Instead, sampling multivariate **normal** random vector is easy ($\&$ widely used. e.g, Normal model in Bayesian method)
  - Multivariate Normal dist'n is defined as affine transformation (linear transformation + constant) of standard normal random vector
  - Want to generate $X \sim N_p(\mu_p, \Sigma)$. Since Σ is PSD and PD (practically), $\exists$ unique M s.t. $\Sigma = MM'$.
  - Using $Y \sim N(\mu, \Sigma) \to AY + B \sim N(A\mu + B, A\Sigma A')$, generate MVN samples by
    1. Cholesky Decompose Σ as MM'
    2. Sample $Z \sim N(0_p, I_p)$
    3. return $MZ + \mu$ that follows $N(\mu, \Sigma)$
  - Note, this still belongs to iid sampling because when I sampled n MVN samples, between each sample (vector), it may be correlated but between samples, it is iid.

Monte Carlo Method for calculating summary statistics of a distribution

Monte Carlo method is a method of using computational way to generate random samples and obtain statistics : actually, whole topic in this slide!

Obtain summary statistics of a distribution using monte carlo method

1. Random sample from a distribution of interest
2. Define sample statistics (mean, quantiles, median, etc): saves effort of challenging integration

- $\#(\theta^{(s)} \leq c)/S \to \Pr(\theta \leq c|y_1, \ldots, y_n)$;
- the empirical distribution of $\{\theta^{(1)}, \ldots, \theta^{(S)}\} \to p(\theta|y_1, \ldots, y_n)$;
- the median of $\{\theta^{(1)}, \ldots, \theta^{(S)}\} \to \theta_{1/2}$;
- the α-percentile of $\{\theta^{(1)}, \ldots, \theta^{(S)}\} \to \theta_\alpha$.

Figure: Obtaining posterior summary statistics using Monte Carlo method. Image from Hoff, P. D. (2009)

Difference between (parametric) Monte Carlo method and Bootstrap

Parametric monte Carlo method requires the **form (at least, up to a normalizing constant in case of MCMC)** of F. However, bootstrap is a resampling method used when F is not known (not an interest of this slide).

1. A random sample from unknown F is given
2. Generate B samples w/ replacement **from the original random sample (treating the sample as population)**
3. Obtain arbitrary quantities from the bootstrap (re)samples.

Importance Sampling: Not a Sampling Method

Importance sampling, widely used in computational statistics, is not a method of sampling from a particular distribution.

Goal

For a random variable(vector) $X \sim f(x)$, obtain Monte Carlo integral estimate for $\theta := E_f[T(X)]$ with low variance
$T(X)$ denotes "statistics" of X.

Algorithm

1. Choose a density g that is 1) possible to get sample from and 2) $T(x) \cdot \frac{f(x)}{g(x)}$ is "similar" for all $x \in \chi$
2. From density g, sample $X_1, ..., X_n$ and return $\frac{1}{n} \sum_{i=1}^{n} T(x_i) \frac{f(x_i)}{g(x_i)}$ for large n.

Proof

$\theta := E_f[T(X)] = \int_x T(x) f(x) dx = \int T(x) \cdot \frac{f(x)}{g(x)} g(x) dx = E_g[\frac{T(x) f(x)}{g(x)}]$.

- Above only indicates that $\sum_{i=1}^{n} T(x_i) \frac{f(x_i)}{g(x_i)}$ ($\vec{x}$ is sample from g, not f!) is a consistent estimator of θ.
- $T(x) \cdot \frac{f(x)}{g(x)}$ being "similar" for all values of x is the key for **variance reduction**!

Usefulness of importance sampling

- Obtain Monte Carlo integral without sampling from f
- Smaller variance estimator: especially required in small probability(integral) estimation
  $\because$ avoid estimating θ as either 0 (most cases) or serious overestimation

References

1. https://www.mi.fu-berlin.de/inf/groups/ag-tech/teaching/2012_SS/L_19540_Modeling_and_Performance_Analysis_with_Simulation/06.pdf

2. https://en.wikipedia.org/wiki/Inverse_transform_sampling

3. https://medium.com/@msuhail153/rejection-sampling-6c4510da24f8.

4. http://www.columbia.edu/ ks20/4703-Sigman/4703-07-Notes-ARM.pdf

5. Hoff, P. D. (2009). A first course in Bayesian statistical methods (Vol. 580). New York: Springer.