Efficient Computational Algorithms

Luca Pernigo

Università della Svizzera italiana[1]

Monte Carlo and MCMC
Nov 7, 2022

Outline

Table of Contents

Monte Carlo

Random numbers generator: A mechanism for producing a sequence of random variables $U_1, U_2, ...$ with the following two properties:

- each U_i follows the uniform distribution between 0 and 1
- the U_i are mutually independent

Monte Carlo

Even though these numbers may seem random, note that are just generated by deterministic algorithms.
Property two is particularly important; it implies that the value of U_i is not predictable from U_{i-1},...,U_1.

Monte Carlo

Linear congruential generators are the most widely used pseudo-random number generator.
Consider three nonnegative integers η, a, and c. An integer seed value z_0 is selected, $0 \leq z_0 < \eta$; a sequence of integers z_k and random variables u_k are obtained recursively with the following formula:

$$z_k = (a \cdot z_{k-1} + c) \mod \eta$$
$$u_k = z_k/\eta$$

Monte Carlo

Here is an example with $z_0 = 4, a = 3, c = 1, \eta = 7$

k	$az^{[k-1]} + c$	$z^{[k]}$	$u^{[k]}$
0		4	0.57142857
1	13	6	0.85714286
2	19	5	0.71428571
3	16	2	0.28571429
4	7	0	0.00000000
5	1	1	0.14285714
6	4	4	0.57142857
7	13	6	0.85714286
8	19	5	0.71428571
9	16	2	0.28571429
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Also note that at step 6 the same sequences of z_k and random numbers are generated again.

Monte Carlo

Note the periodicity of LCG; since the integers z_k are nonnegative and bounded by η, the sequence of pseudo-random number must repeat in a continual loop.

Monte Carlo

Hereafter is visualized how LCG work. This graph represent the relation between the u_i and the u_{i+1}. From the figure it is easy to see the deterministic nature of pseudo-random number generators. Note the regular pattern of how points are distributed, they all lie on parallel lines through the unit squre.

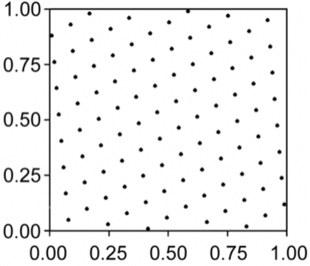


Figure: LCG with $a = 89, c = 0, \eta = 101$

Monte Carlo

Monte Carlo methods are called this way because of their use of random sampling in order to solve deterministic problems.

Monte Carlo

The idea behind Monte Carlo sampling is to use a random number generator in order to sample from the desired distribution. The Inverse Transform Method will be considered first.

Inverse Transform Method

Choosing the uniform distribution is handy because it enables us to generate random sample from any other distribution using the Inverse Transform Method.

- First simulate observations $U_i \sim U[0, 1]$
- Second calculate $Y = F^{-1}(U_i)$, where F^{-1} is the quantile function of the desired distribution

It follows that the random variable $Y_i = F^{-1}(U_i)$ is distributed according to F

Inverse Transform Method

Before illustrating the procedure, it is important to remind a useful property of the uniform distribution [0,1] in order to properly understand this procedure:

$$F(X) = \begin{cases} 0 & x < 0 \\ x & 0 < x < 1 \\ 1 & x > 1 \end{cases}$$

Hence the property $F(c) = P(X \leq c) = c$

Inverse Transform Method

To show why $Y_i \sim F$ consider:

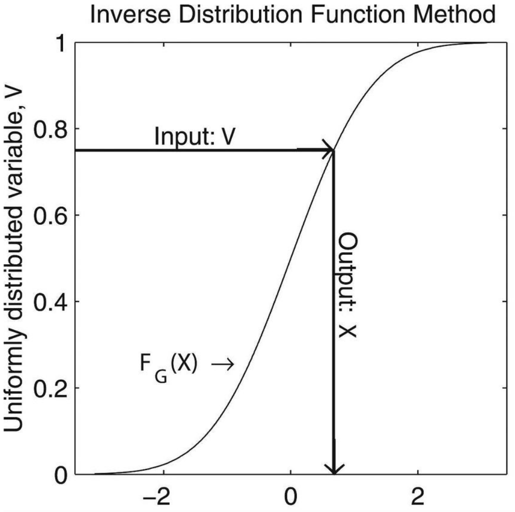
$P(Y_i \leq c) = P(F^{-1}(U_i) \leq c)$ Definition of Y

$P(F^{-1}(U_i) \leq c) = P(U_i \leq F(c))$ Apply F to both side of the disequality of the LHS

$P(U_i \leq F(c)) = F(c)$

Finally, use the fact that U_i is a random variable following the uniform distribution [0,1]

Inverse Transform Method



Random sample from an exponential distribution

Consider the exponential distribution $f(y) = \frac{e^{-y/\theta}}{\theta}$

its cumulative distribution is $F(y) = 1 - e^{-y/\theta}$

Our goal is to obtain its quantile distribution F^{-1}

$$F(y) = 1 - e^{-y/\theta} = F(F^{-1}(U)) = U$$

$$U = 1 - e^{-y/\theta}$$

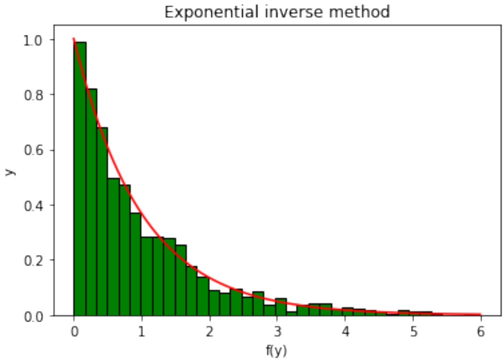
$$1 - U = e^{-y/\theta}$$

$$log(1 - U) = \frac{-y}{\theta}$$

$$y = -log(1 - U)\theta$$

Random sample from an exponential distribution

Pluggin into y the observation from a random uniform $[0,1]$ we get a random sample from an exponential with parameter θ



code1

Random sample from other distributions

Consider the following distribution $f(y)$

With cumulative distribution $F(y) = \frac{2}{\pi} arcsin(\sqrt{y})$, $\quad 0 \leq y \leq 1$

Our goal is to obtain its quantile distribution F^{-1}

$$F(y) = \frac{2}{\pi} arcsin(\sqrt{y}) = F(F^{-1}(U)) = U$$

$$U = \frac{2}{\pi} arcsin(\sqrt{y})$$

$$\frac{\pi}{2} U = arcsin(\sqrt{y})$$

$$sin(\frac{\pi}{2} U) = sin(arcsin(\sqrt{y})) = \sqrt{y}$$

$$y = sin(\frac{\pi}{2} U)^2$$

Random sample from other distributions

Pluggin into y the observation from a random uniform $[0, 1]$ we get a random sample from $f(y) = \frac{1}{\pi\sqrt{1-x}\sqrt{x}}$

Random sample from other distributions

Consider the Rayleigh distribution $f(y)$

With cumulative distribution $F(y) = 1 - e^{-x^2/2\sigma^2}$, $\qquad x \geq 0$

Our goal is to obtain its quantile distribution F^{-1}

$$F(y) = 1 - e^{-x^2/2\sigma^2} = F(F^{-1}(U)) = U$$

$$log(1 - U) = \frac{-x^2}{2\sigma^2}$$

$$x = \sqrt{-2log(1 - U)}\sigma$$

Random sample from other distributions

Pluggin into y the observation from a random uniform $[0, 1]$ we get a random sample from an Rayleigh with parameter θ

approximate Inverse Transform Method Normal

Some distributions do not even have a closed form quantile function.

approximate Inverse Transform Method Normal

Nevertheless, we can approximate it and apply the same logic to the approximation of the quantile function.

approximate Inverse Transform Method Normal

Input: u between 0 and 1
Output: x, approximation to $\Phi^{-1}(u)$.
$y \leftarrow u - 0.5$
if $|y| < 0.42$
  $r \leftarrow y * y$
  $x \leftarrow y * (((a_3 * r + a_2) * r + a_1) * r + a_0) /$
    $((((b_3 * r + b_2) * r + b_1) * r + b_0) * r + 1)$
else
  $r \leftarrow u;$
  if $(y > 0)$ $r \leftarrow 1 - u$
  $r \leftarrow \log(-\log(r))$
  $x \leftarrow c_0 + r * (c_1 + r * (c_2 + r * (c_3 + r * (c_4 +$
    $r * (c_5 + r * (c_6 + r * (c_7 + r * c_8)))))))$
  if $(y < 0)$ $x \leftarrow -x$
return x

code2

Table of Contents

MCMC

MCMC algorithms are generally used for sampling from multi-dimensional distributions, especially when the number of dimensions is high.

MCMC

Before delving into MCMC, we first have to introduce markov chain and their properties. Doing this will enable us to understand the second building block of MCMC methods.

MCMC

A stochastic process is a collection of random variables indexed by t, where t denotes time.

A Markov chain is a stochastic process; their peculiarity is their memory less property. That is the conditional distribution of X_{n+1} depends solely on X_n.

MCMC

Let P be a $k \times k$ matrix with elements $\{P_{i,j} : i, j = 1, ..., k\}$. A random process $(X_0, X_1, ...)$ with finite state spaces $S = \{s_1, ..., s_k\}$ is said to be a Markov chain with transition matrix P , if for all n, all $i, j \in \{1, ..., k\}$ and all $i_0, ..., i_{n-1} \in \{1, ..., k\}$ we have

$$\mathbf{P}(X_{n+1} = s_j | (X_0 = s_{i_0}, X_1 = s_{i_1}, ..., X_n = s_i)) = \mathbf{P}(X_{n+1} = s_j | X_n = s_i)$$

$$= \mathbf{P}_{i,j}$$

MCMC

A Markov Chain is said homogeneous if it satisfies the following property,

$$\mathbf{P}(X_{n+1}|X_n) = \mathbf{P}(X_1|X_0) \quad \forall n \geq 0$$

In simple words the probability of moving from one state to another is time invariant

MCMC

P is also called the Transition matrix. Consider for example a simple markov chain weather model. There are two kinds of weather: rain and sun, and the prediction assumes that the weather will be the same of yesterday with probability 75 %. Hence our transition matrix is

$$P = \begin{bmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{bmatrix}$$

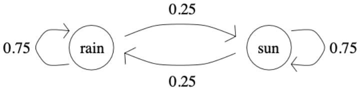


Figure: Graph of weather transition matrix

MCMC

We might be interest in obsverving the behavior of our Markov Chain as time goes on to infinity. Therefore the concept of Limit distribution of a Markov Chain is introduced:

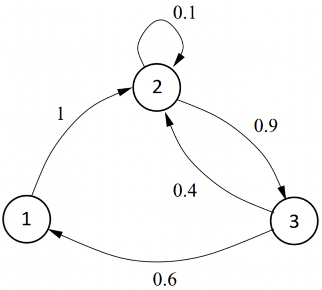
π_n is simply the probability vector for the various states at time step n

$\pi_\infty = \lim_{n\to\infty} \pi_n = \lim_{n\to\infty} \pi_0 P^n$

Note that $\pi_0 P^n$ is again just a random vector of probabilities.

MCMC

For example it can be show that this markov chain has as limiting distribution $\pi_\infty = (0.22, 0.41, 0.37)$.



To see that just multiply π_∞ by its transition matrix and notice that $\pi_\infty P = \pi_\infty$, therefore π reached its limiting distribution.
website that visualizes the concept

MCMC

Note that the limiting distribution π_∞ is unique, we reach it always independently of where we start with.
code3

MCMC

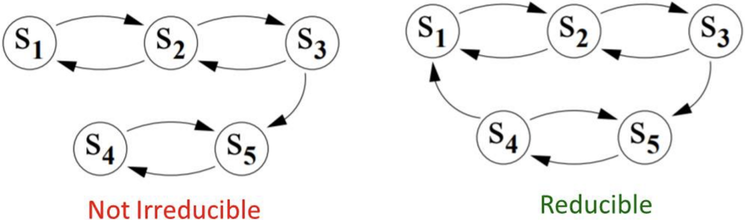
We say that s_i communicates with another state $s_j, si \to s_j$ if the chain has positive probability of ever reaching s_j starting from s_i.
If $s_i \to s_j$ and $s_j \to s_i$ then the two states intercommunicate. $s_i \leftrightarrow s_j$.

MCMC

A Markov chain $(X_0, X_1, ...)$ with state space $S = \{s_1, ..., s_k\}$ and transition matrix P is said to be irreducible if for all $s_i, s_j \in S$ we have that $s_i \leftrightarrow s_j$. Otherwise the chain is said to be reducible.

MCMC

In other words, an irreducible Markov chain is a chain where each state is reachable from any other state, in a finite number of steps.



MCMC

The period $d(s_i)$ of a state $s_i \in S$ is defined as the:

$d(s_i) = gcd\{n \geq 1 : P_{i,j}^n > 0\}$

The period s_i is the greatest common divisor of the set of times that the chain can return (has positive probability of returning) to s_i given that we start with $X_0 = s_i$

MCMC

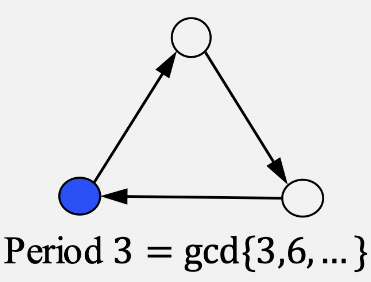


Figure: period 3

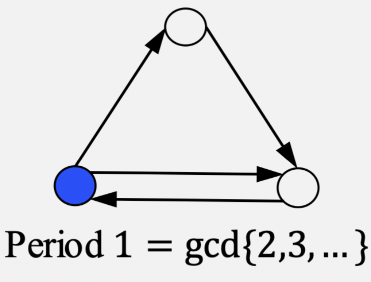


Figure: period 1

MCMC

If $d(s_i) = 1$ the state s_i is said to be aperiodic.

A Markov Chain is said to be aperiodic if all its states are aperiodic, otherwise the chain is said to be periodic.

Consider the simple weather model of before; regardless of the weather today, we have for any n a probablity greater than zero of having the same weather. Hence that Markov Chain is one example of aperiodic chain.

MCMC

A usefull property of irreducibility is that in an irreducible Markov Chains all its states have the same period.

MCMC

Let $(X_0, X_1, ...)$ be a Markov Chain with state space S and transition matrix P. A vector π is said to be a stationary distribution for the Markov Chain, it it satisfies:

(i) $\pi \geq 0$ for $i = 1, ..., k \quad \text{and} \sum_{i=1}^{k} \pi_i = 1$

(ii) $\pi P = \pi$ meaning that $\sum_{i=1}^{k} \pi_i P_{i,j} = \pi_j$ for $j = 1, ..., k$

Property (i) means that π should describe a probability distribution while property (ii) means that if the initial distribution u^0 equals π then the distribution u^1 will be equal to π and so on.

MCMC

Having laid out the foundations, we can now introduce a fundamental theorem in the theory of Markov Chains.

The Basic limit theorem
If a Markov chain is irreducible and aperiodic, then:

(i)$\exists ! \pi = (\pi_1, ..., \pi_k)$ our stationary distribution

(ii)$\lim_{n \to \infty} P(X_t = s_i) = \pi_i, \forall i$

MCMC

The Basic limit theorem implies that an irreducible and aperiodic chain will converge to its unique stationary distribution $p(x)$.
Therefore, by constructing an irreducible and aperiodic Markov Chain with a stationary distribution equal to $p(x)$, we can generate sample from our target distribution $p(x)$.

MCMC

Therefore we have to make sure that the target distribution $p(x)$ is indeed stationary with respect to our Markov Chain. To do so the concept of reversibility is introduced.

MCMC

A Markov chain is said to be reversible, if it exists a probability distribution π such that $\pi_i P_{i,j} = \pi_j P_{j,i}$ for all states i and j
similarly the probability π_i distribution is said to be reversible.

MCMC

Theorem Reversibility and Stationarity:

Let $(X_0, X_1, ...)$ be a Markov Chain. If π is a reversible distribution for the chain then it is also a stationary distribution for the chain.

MCMC

Thus, we conclude that if we can construct an irreducible, aperiodic, and reversible Markov chain with regards to a target distribution $p(x)$, then according to theorem 1 and 2 our chain will converge to $p(x)$. This is the intuition behind Markov chain Monte Carlo methods, and particularly the Metropolis-Hastings algorithm.

MCMC

We carried out our analysis in the discrete case, but these poweful results hold also in the continous case. We only have to substitute the vector of probabilites π with a density function, and substitute the transition matrix with a transition kernel K

Metropolis Hastings

Metropolis–Hastings algorithm is one of the most important Markov chain Monte Carlo algorithms. This procedure enables us to randomly sample from a certain distribution, where we know the distribution only up to a normalizing constant.

Metropolis Hastings

With normalizing constant it is meant a constant by which an everywhere non-negative function must be multiplied so that the area under its graph is 1. Hence, multiplying an everywhere non-negative function by its normalizing constant we get a probability density function or a probability mass function.

For example the normalizing constant of the standard normal distribution is $\frac{1}{\sqrt{2\pi}}$

Metropolis Hastings

Additionally note that this algorithm is one between the ten most influential algorithms of the 20^{th} century.

Metropolis Hastings

Terminology:

$p(x)$ is the target distribution, it is the distribution we want to sample from.

$q(x^*|x)$ is the proposal distribution. We choose it such that it is easy to sample from it. This distribution should share the same support as $p(x)$. Note that the at each step we sample from a proposal conditioned on the last sample x_i.

Metropolis Hastings

Initialize x_0

for i=0 to N-1

Sample u from a $U(0,1)$

Sample x^* from $q(x^*|x)$

if $u \leq \min(1, \frac{p(x^*)q(x_i|x^*)}{p(x_i)q(x^*|x_i)})$

$x_{i+1} = x^*$

else

$x_{i+1} = x_i$

Metropolis Hastings

The Metropolis–Hastings algorithm works by generating a sequence of sample values in such a way that, as more and more sample values are produced, the distribution of values approximates the desired distribution.

Metropolis Hastings

Note that the sample values are produced iteratively, doing so the distribution of the next sample depends only on the current sample value; This makes the sequence of samples a Markov chain.

Intuition:

(i) We want our chain to be irreducible; at each step we have a positive probability to reach any value in the support of the proposal distribution. From irreducibility it follows that all states have the same period.

Metropolis Hastings

(ii) We want our chain to be aperiodic. For every state x we have a positive probability of remaining at the current state and we have a positive probability of going to a new state.

Metropolis Hastings

(iii) By the basic limit theorem our irreducible and aperiodic Markov chain will converge to a unique stationary distribution. Our Markov chain is constructed such that it is reversible with respect to $p(x)$, hence $p(x)$ is the stationary distribution of our Markov chain.

Metropolis Hastings

Note why $p(x)$ and $q(x)$ must share the same support; if they do not share the same support our sample will never pick up the points of $p(x)$ that do not belong to $q(x)$.

code4

Table of Contents

Call option pricing

Monte Carlo methods are widely used in finance to simulate stocks paths.

Call option

Suppose:

S0 = 102 Stock price today

K = 100 Strike

r = 0.02 risk free rate

sigma = 0.15 volatility rate

Call option

Plug everything into our formula for the price of a stock.

$S_{t+\Delta t} = S_t e^{\mu \Delta t + \sigma \sqrt{\Delta t} \epsilon_{t+\Delta t}}$, with $\epsilon \sim \; i.i.d \; N(0,1)$

We see there is normal noise. The idea is to simulate many paths and the take the average in order to estimate the price of the stock.

code5

Table of Contents

Conclusion

Key takeaways:

(i) Monte Carlo methods; get a good approximation for the desired problem, through random sampling.

(ii) Metropolis Hastings; universal method to get a random sample from whichever distribution.

Thank you