APMA 1690: Computational Probability and Statistics

Milan Capoor

Fall 2023

Goal of the course: (Approximately) Compute integrals using Monte Carlo methods

Lecture 1: Sept 9

Arc of the Course

Example: Motivating the goal of the course

$$I = \int_0^1 \arccos\left(\frac{\cos(\frac{\pi x}{2})}{1 + 2\cos(\frac{\pi x}{2})}\right)\, dx = \frac{5\pi}{12}$$

This is *really* hard! But using Monte Carlo methods we can do much better!

Law of Large Numbers: Suppose $X_1, ..., X_n$ are independently and identically distributed random variables. Then, when $n \to +\infty$,

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \to \mathbb{E}[f(X_1)]$$

If we define $X_1 \sim \text{Unif}(0, 1)$, then the PDF of X_1 is 1.

Applying this to the integral above, let the integrand be denoted $f(x)$ so

$$I = \int_0^1 f(x)\, dx = \int_0^1 f(x) \cdot 1\, dx = \mathbb{E}[f(X_1)]$$

Putting all of this together,

$$I = \mathbb{E}[f(X_1)] \approx \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

which means that by doing some transformations on the integral and averaging many random outputs of the integrand, we can use the average to approximate the value of the integral with good accuracy.

In fact, with $n = 10000$, we approximate $I = 1.308827$ when in fact $I = \frac{5\pi}{12} \approx= 1.308997$ which is quite good!

A problem: Notice! This method *assumes* we are able to generate iid random variables. This introduces some questions:

1. What is "randomness"?
2. How do we generate random numbers?
3. How large is our error when using stochastic methods? How do we control this error?
4. What if the inputs are random vectors instead of random numbers? What if the problem is multi-variable?
5. How do we manage unreasonable time and memory costs?

Moving towards a solution: To address the last concern especially, we can compromise on the iid condition to generate a Markov chain where $\vec{X}_n \dot{\sim} \Pi$

Heuristic Ergodic Theorem: Suppose $\vec{X}_1, \vec{X}_2, ..., \vec{X}_n$ is a Markov chain such that $\vec{X}_n \dot{\sim} \Pi$ when n is large. Then

$$\frac{1}{n}\sum_{i=1}^{n} f(\vec{X}_i) \approx \int f(x) \cdot \Pi(x)\, dx$$

Really, this just introduces more questions:

6. What is the ergodic theorem?
7. How do we generate a Markov chain satisfying this assumption that $\vec{X}_n \dot{\sim} \Pi$?

This will lead us to two algorithms:

- Metropolis-Hastings Algorithm

- Gibbs sampling (developed by Prof Geman at Brown!)

Random Variables

Example: Coin toss

With a sample space $\Omega = \{H, T\}$, let X represent the outcome of flipping the coin once. Then really, $X : \Omega \mapsto \{0, 1\}$.

In fact, this gives us a formal definition for a *random variable*; a random variable is a function X that maps a sample space Ω to $\mathbb{R}$.

$$\text{For each fixed } \omega \in \Omega, \quad X(\omega) \in \mathbb{R}$$

Lecture 2: Sept 12

What is randomness?

Probability Theory

1. Sample Space (Ω)
2. Random Variable (X)
3. Probability ($\mathbb{P}$)

Sample space: the collection of all possible outcomes of an event *Examples:*

- For flipping a coin once, $\Omega = \{H, T\}$
- Rolling a six sided die, $\Omega = \{1, 2, 3, 4, 5, 6\}$

Remark: the essential characteristic of experiments in sample space is that the outcome in uncertain before performing the experiment.

Event: $A \subseteq \Omega$

Random Variable: a function $X : \Omega \to \mathbb{R}^d$

Deterministic/Pseudo-Random Variable: $x \in \mathbb{R}^d, \quad X(\omega) = x, \ \forall \omega \in \Omega$

Probability: a real-valued function of all subsets of omega ($\mathbb{P} : A \to \mathbb{R} \quad A \subseteq \Omega$) which satisfies the following three axioms:

1. $\forall A \subseteq \Omega, \quad \mathbb{P}(A) \geq 0$

2. $\mathbb{P}(\Omega) = 1$
3. $\forall \{A_n\}_{n=1}^{\infty} \subseteq \Omega$ such that $A_i \cap A_j = \emptyset$ $(i \neq j)$ (for any mutually exclusive infinite sequence of events),

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$

Theorem: The probability of an impossible event is 0.

Proof: Let $A_n = \emptyset \quad n = \{1, 2, ..., n\}$. Clearly, for any $i \neq j$,

$$A_i \cap A_j = \emptyset \cap \emptyset = \emptyset$$

so

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$
$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} \emptyset\right) = \mathbb{P}(\emptyset) = \sum_{n=1}^{\infty} \mathbb{P}(\emptyset)$$

Define $a := \mathbb{P}(\emptyset)$. Then,

$$a = \sum_{n=1}^{\infty} a$$

so $a = 0 \implies \mathbb{P}(\emptyset) = 0$ ■

Lecture 3: Sept 14

Indicator Functions

Definition: Let $A \subseteq \mathbb{R}^d$,

$$\mathbb{1}_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

Examples:

1. $d = 1, \quad A = (0,1) \subseteq \mathbb{R}^1$

$$\mathbb{1}_{(0,1)}(x) = \begin{cases} 0 & x \leq 0 \\ 1 & 0 < x < 1 \\ 0 & x \geq 1 \end{cases}$$

2. $d = 1, \quad A = [0, +\infty)$ (Heaviside function)

$$\mathbb{1}_{[0,\infty)} = \mathbb{1}(x \geq 0) = \begin{cases} 1 & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Interlude: Probability theory vs. Statistics

Probability seeks to establish the expected outcome of an experiment before it is performed, given $\mathbb{P}$.

Statistics seeks to infer $\mathbb{P}$ from data observed during the experiment.

Distribution

Definition: Let X be a $\mathbb{R}^d$-valued RV defined on the probability space $(\Omega, \mathbb{P})$. Then, the distribution of X according to $\mathbb{P}$ is

$$\mathbb{P}\left(\{\omega \in \Omega : X(\omega) \in A\}\right) \quad \forall A \subseteq \mathbb{R}^d$$

Remarks:

- To study the distribution of X we must examine all $A \subseteq \mathbb{R}^d$ which is very hard
- When $d = 1$ this is easier because we must only examine the interval of the form $(-\infty, t] \quad \forall t \in \mathbb{R}$

Cumulative Distribution Functions (CDFs)

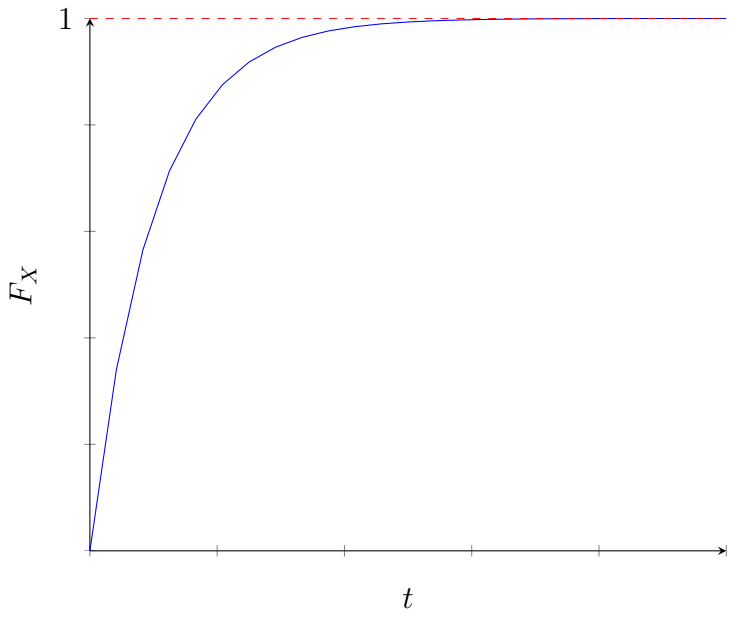
Definition: Let $X : (\Omega, \mathbb{P}) \to \mathbb{R}$. Then the CDF of X is $F_X : \mathbb{R} \to \mathbb{R}$

$$F_X(t) := \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \leq t\}\right)$$

Examples:

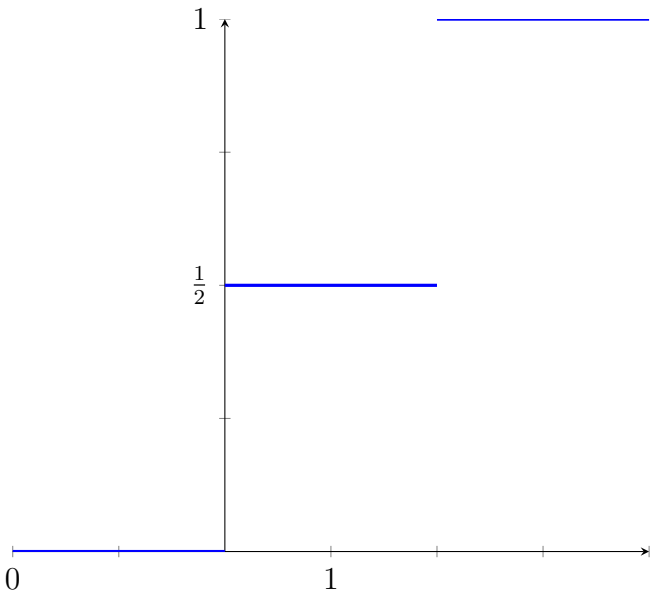
- $X \sim \text{Exp}(\lambda)$

$$F_X(t) = (1 - e^{-\lambda t}) \cdot \mathbb{1}(t \geq 0)$$



- $X \sim \text{Bernoulli}(\frac{1}{2})$

$$F_X(t) = \begin{cases} 0 & t < 0 \\ \frac{1}{2} & 0 \leq t < 1 \\ 1 & t \geq 1 \end{cases}$$



Continuous Random Variables

Definition: Let $X : (\Omega, \mathbb{P}) \to \mathbb{R}$. If its CDF is continuous and piecewise differentiable ("absolutely continuous") then X is a *continuous random variable*

Probability Density Function (PDF):

$$p_X(t) := \frac{d}{dt} F_X(t)$$

where $\frac{d}{dt}$ is the piecewise derivative.

Remarks: the CDF determines the corresponding PDF via differentiation and the PDF determines the CDF via integration

$$p_X(x) = \frac{d}{dx} F_X(x)$$
$$F_X(x) = \int_{-\infty}^{t} p_X(t)\, dt$$

Theorem: Let $X : (\Omega, \mathbb{P}) \to \mathbb{R}$.

1. $F_X(t)$ is non-decreasing: $F_X(t_1) \leq F_X(t_2)$ if $t_1 \leq t_2$
2. $\lim_{t \to -\infty} F_X(t) = 0, \quad \lim_{t \to \infty} F_X(t) = 1$

Lecture 4: Sept 19

CDFs

The CDF of $X \sim \text{Bernoulli}(\frac{1}{3})$ can be written in a number of ways.

First, by probabilities,

$$\begin{cases} \mathbb{P}(X=1) = \frac{1}{3} \\ \mathbb{P}(X=0) = \frac{2}{3} \end{cases}$$

Which yields a CDF

$$F_X(t) = \begin{cases} 0 & t < 0 \\ \frac{2}{3} & 0 \leq t < 1 \\ 1 & t \geq 1 \end{cases}$$

Which describes

$$F_X(t) = \frac{2}{3}\mathbb{1}_{(t\geq 0)}(t) + \frac{1}{3}\mathbb{1}_{(t\geq 1)}(t) = \sum_{k=0}^{1} p_k \cdot \mathbb{1}_{(t\geq x_k)}(t)$$

where $x_0 = \mathbb{P}(X=0)$ and $x_1 = \mathbb{P}(X=1)$

Discrete Random Variables

Definition: Let X be $\mathbb{R}^1$-valued RV on $(\mathbb{P}, \Omega)$. X is a discrete RV if its CDF is

$$F_X(t) = \sum_{k=0}^{K} p_k \cdot \mathbb{1}_{(t\geq x_k)}(t)$$

where $\{x_k\}_{k=0}^K$ are distinct real numbers, $\{p_k\}_{k=0}^K$ are non-negative real numbers satisfying $\sum_{k=0}^K p_k = 1$ and K is a positive integer (or $+\infty$)

Probability Mass Function (PMF): the ordered sequence $\{p_k\}_{k=0}^K$ which determines the CDF.

Theorem (Non-rigorously): For a discrete RV,

$$p_k = \mathbb{P}(X=k)$$

Mixed Random Variables (Optional)

Example: Let $Y \sim \text{Bernoulli}(\frac{1}{2})$ and $Z \sim N(0,1)$ be independent RVs on $(\Omega, \mathbb{P})$.

$$X(\omega) := Y(\omega) + (1 - Y(\omega)) \cdot Z(\omega)$$

Note that X is a RV because $X : \Omega \to \mathbb{R}$.

Expected Values

Notation: Let $g : \mathbb{R} \to \mathbb{R}$ and $X : \Omega \to \mathbb{R}$. Then $g(X(\omega)) : \Omega \to \mathbb{R}$ so we denote $g(X)$ as a random variable.

Definition: Let $g : \mathbb{R} \to \mathbb{R}$.

1. Suppose X is a continuous RV whose PDF is p_X.

   If $\int_{-\infty}^{\infty} |g(x)| \cdot p_X(x)\, dx < \infty$,

   $$\mathbb{E}[g(X)] := \int_{-\infty}^{\infty} g(x) \cdot p_X(x)\, dx$$

   otherwise, $\mathbb{E}[g(X)]$ does not exist.

2. Suppose X is a discrete RV with CDF $F_X(t) = \sum_{k=0}^{k} p_k \cdot \mathbb{1}_{(t \geq x_t)}(t)$.

   If $\sum_{k=0}^{K} |g(x_k)| \cdot p_k < \infty$, then

   $$\mathbb{E}[g(X)] := \sum_{k=0}^{K} g(x_k) \cdot p_k$$

   otherwise, $\mathbb{E}[g(X)]$ does not exist.

Lecture 5: Sept 21

Law of Large Numbers (LLN)

Theorem: Let $X_1, \ldots, X_n$ be $\mathbb{R}^1$-dimensional RVs on $(\Omega, \mathbb{P})$. If the RVs are independently and identically distributed and $\mathbb{E}[X_1]$ exists, then

$$\boxed{\mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} X_i(\omega)\right) = \mathbb{E}X_1\right\}\right) = 1}$$

Corollary: Under the same conditions,

$$\mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} X_i(\omega)\right) \neq \mathbb{E}X_1\right\}\right) = 0$$

Remarks:

$$\overline{X}_n(\omega) := \frac{1}{n}\sum_{i=1}^{n} X_i(\omega)$$

Then the following are all random variables:

- $\overline{X}_n$
- $\lim_{n\to\infty} \overline{X}_n(\omega)$
- $e_n(\omega) := |\overline{X}_n(\omega) - \mathbb{E}X_1|$

So the LLN can also be written as

$$\mathbb{P}(\{\omega \in \Omega : \lim_{n\to\infty} e_n(\omega) = 0\}) = 1$$

Law of the Iterated Logarithm

Variance:

$$\text{Var}\,(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$$

Theorem: Let $X_1, X_2, X_3, ...$ be identically and independently distributed RVs defined on $(\Omega, \mathbb{P})$ Suppose $\mathbb{E}X_1$ and $\text{Var}(X_1)$ exist. Then,

$$\mathbb{P}\left(\{\omega \in \Omega : \lim_{m\to\infty}\left[\sup_{n\geq m}\left(\frac{e_n(\omega)}{\sqrt{\text{Var}X_i\frac{2\log(\log n)}{n}}}\right)\right] = 1\}\right) = 1$$

Heuristically, when n is large,

$$\boxed{\begin{aligned}&\mathbb{P}(\{\omega \in \Omega : |e_n(\omega)| \leq \sqrt{\text{Var}X_i\frac{2\log(\log(n))}{n}}\}) \approx 1\\ &\mathbb{P}(\{\omega \in \Omega : |e_n(\omega)| > \sqrt{\text{Var}X_i\frac{2\log(\log(n))}{n}}\}) \approx 0\end{aligned}}$$

Empirical CDFs

Definition: Suppose $X_1, \ldots, X_n$ are RVs defined on $(\Omega, \mathbb{P})$.

$$F_n(\omega, t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{t \geq X_i(\omega)}(t)$$

For each fixed ω, F_n is a function of t. For each fixed t, F_n is a random variable. Therefore, F_n is a *stochastic process*

If $X_1, \ldots, X_n \overset{iid}{\sim} F$, we would like to use $F_n(\omega, t)$ to estimate F:

$$|F_n(\omega, t) - F(t)| \overset{?}{\approx} 0$$

Glivenko-Contelli Theorem (1933)

Theorem: Suppose $X_1, \ldots, X_n$ are RVs defined on $(\Omega, \mathbb{P})$. If the RVs are iid, then

$$\mathbb{P}(\{\omega \in \Omega : \lim_{n \to \infty} \max_t |F_n(\omega, t) - F(t)| = 0\}) = 1$$

In other words, "$F_n(\omega, t)$ converges to F uniformly in t with probability 1."

Lecture 6: Sept 26

True-Random Numbers

Definition: Real numbers $x_2, x_2, \ldots, x_n$ are called *random numbers* from a distribution associated with a CDF if

1. there is an underlying probability space $(\Omega, \mathbb{P})$
2. $\exists X_1, X_2, \ldots, X_n$ are iid random variables defined on $(\Omega, \mathbb{P})$ which share the CDF F
3. $$\exists \omega^* \in \Omega : \begin{cases} x_1 = X_1(\omega^*) \\ x_2 = X_2(\omega^*) \\ \quad \vdots \\ x_n = X_n(\omega^*) \end{cases}$$

Remarks:

1. Each RV X_i mainly refers to a truly random RV (X_i is not a constant function)
2. A random variable $X_i(\omega)$ is a function $\omega \to \mathbb{R}$
3. A random number $x_i = X_i(\omega^*)$ is a number in $\mathbb{R}$

Types of Random Number Generators

1. Hardware RNGs
   - *Pros:* create true random numbers
   - *Cons:* slow and expensive
2. Pseudo RNGs
   - *Pros:* Very fast
   - *Cons:* Not truly random

Pseudo-random Number Generators

Definition: Suppose F is a given CDF. Let $g : \{1, 2, 3, \ \ldots, \ \} \to \mathbb{R}$ be a function. g is called a PRNG for F if

$$\lim_{n\to\infty} \left(\sup_t \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(t \geq g(i))} - F(t) \right| \right)$$

and the outputs of g are pseudo-random numbers.

Essentially, *the numbers $g(i)$ look like true-RNs iid from F but are not.*

Glivenko-Contelli theorem vs PRNGs:

The essence of the GCT is that the empirical CDF of a sequence of iid RVs converges to the true CDF. PRNGs, meanwhile, function in the reverse direction: when the empirical CDF converges to F, the random variables look like true iid RVs.

Example: the Multiplicative Congruential Generator (MCG)

Let X be a RV defined on $(\Omega, \mathbb{P})$ and $m \in \mathbb{N}$. We say $X \sim \text{Unif}(\{1, 2, \ \ldots, \ m-1\})$ so

$$\mathbb{P}(X = i) = \frac{1}{m-1} \quad 1 \leq i \leq m-1$$

$$\implies F_X(t) = \frac{1}{m-1}\sum_{i=1}^{n} \mathbb{1}_{(t \geq i)}$$

The Algorithm:

- Input:
  1. 'properly chosen" integers m and a (in an old version of Matlab, they used $m = 2^{31} - 1$ and $a = 7^5$) where "properly chosen" involves lots of number theory
  2. a seed s
  3. a sample size n
- Output: $g(1), g(2), \ \ldots, \ g(n)$ which look like iid random numbers from Unif($\{1, 2, \ \ldots, \ m-1\}$)

Process:

```
g(1) <-- s

for i= 1, 2, ..., N
    g(i + 1) <-- ag(i) mod m
end
```

Mathematically,

$$Y(\omega) = \frac{X(\omega)}{m} \sim \text{Unif}(\{\frac{1}{m}, \frac{2}{m}, \ \ldots, \ \frac{1}{m-1}\}) \implies F_Y(t) \quad = \frac{1}{m-1}\sum_{i=1}^{m-1} \mathbb{1}_{(t \geq \frac{i}{m})}$$

$$\implies \frac{g(1)}{m}, \frac{g(2)}{m}, \ \ldots, \ \frac{g(n)}{m} \text{ all look like iid RVs from Unif}(\{\frac{1}{m}, \frac{2}{m}, \ \ldots, \ \frac{1}{m-1}\})$$

Recall that $m = 2^{31} - 1$ is a very large number. So

$$F_Y(t)\frac{1}{m-1}\sum_{i=1}^{m-1} \mathbb{1}_{(t \geq \frac{i}{m})} \approx \lim_{m \to \infty} \sum_{i=1}^{m-1} \mathbb{1}_{(t \geq \frac{i}{m})} \overset{*}{=} F_{\text{Unif(0,1)}}(t)$$

where the starred equality comes from the definition of Riemann integrals.

All together, we thus have a collection of (pseudo) random numbers that look like they were generated iid from Unif$(0, 1)$.

Lecture 7: Sept 28

A question

Let X_1 and X_2 be RVs on $(\Omega, \mathbb{P})$ with the same distribution. For any $\omega \in \Omega$, do we have $X_1(\omega) = X_2(\omega)$?

Answer: No.

Set up an experiment where we flip a coin twice. Then

$$\Omega = \{\omega = (\omega_1, \omega_2) : \omega_i \in \{H, T\}\}$$

with

$$X_i(\omega) = \begin{cases} 1 & \omega_i = H \\ 0 & \omega_i = T \end{cases} \sim \text{Bernoulli}\left(\frac{1}{2}\right)$$

Let the result of the experiment show $\omega^* = (H, T)$ Then $X_1(\omega^*) = 1$ but $X_2(\omega^*) = 0$ so

$$X_1(\omega^*) \neq X_2(\omega^*)$$

Review

Pseudo-Random Number Generator: a function $g_i : \mathbb{N} \to \mathbb{R}$ if $g(1), g(2), \ldots, g(n)$ "look like" true random numbers iid from F.

Note: "look-like" means that we fail to reject the hypothesis that the generated numbers are true RNs iid from F. (we make a Type II error)

Hypothesis testing

There are two ways to test H_0:

- Kolmogrov-Smirnov Test (1933) for continuous RVs
- χ^2-test (for discrete RVs)

Multiplicative Congruential Generator (MCG)

Using the following algorithm with s being any seed and a and m carefully chosen via number theorem, we can generate N Pseudo RNs for Unif$(0, 1)$:

```
g(1) <-- s

for i= 1, 2, ..., N
    g(i + 1) <-- ag(i) mod m
end
```

In this class, we will pretend like these are true random numbers.

Inverse CDF Method

How do we generate RNs from other distributions than Unif$(0,1)$*?*

Inverse: Let F be the CDF of the distribution of interest. Suppose F has an inverse function F^{-1}:

$$y = F(x) \iff x = F^{-1}(y)$$

Theorem: Suppose $U \sim \text{Unif}(0,1)$ is a random variable defined on $(\Omega, \mathbb{P})$. We define a random variable X by

$$X(\omega) := F^{-1}(U(\omega))$$

Then, the CDF of X is F.

Proof:

By the definitions of the CDF and X,

$$F_X(t) = \mathbb{P}(X \leq t) = F^{-1}(U(\omega))$$

Then, because the inverse of F is non-decreasing (by assumption),

$$F_X(t) = \mathbb{P}(U \leq F(t))$$

Then by the CDF of U,

$$\mathbb{P}(U \leq F(t)) = F(t)$$

so

$$F_X(t) = F(t) \quad \blacksquare$$

Remark: in the above proof we made two strong assumptions: 1) the inverse exists; 2) the inverse is non-decreasing

The first assumption is particularly bold because many CDFs do not have an inverse. The simplest example is Bernoulli($\frac{1}{2}$)

General Case Theorem:

Suppose F is any CDF. We let $U \sim \text{Unif}(0,1)$ and define new random variable,

$$G(u) := \inf \mid \{t \in \mathbb{R} : F(t) \geq u\}$$

and

$$X(\omega) := G(U(\omega))$$

Then, the CDF of X is F.

This gives us a new algorithm for generating random numbers:

1. Input a CDF F and a sample size n
2. Generate $x_1, x_2, \ldots, x_n \overset{iid}{\sim} \text{Unif}(0,1)$ via the MCG algorithm
3. Let $y_i = G(x_i) = \inf\{t \in \mathbb{R} : F(t) \geq x_i\}$ for $i \in [1, n]$

Example: We are interested in Exp(1) whose CDF is

$$F(t) = (1 - e^{-t}) \cdot \mathbb{1}_{(t>0)}$$

We can derive the formula

$$G(u) = \log\left(\frac{1}{1-u}\right)$$

Then via MCG, we generate n random numbers from the standard uniform distribution. and calculate

$$y_i = \log\left(\frac{1}{1-x_i}\right)$$

for each x_i generated from the MCG.

Thus, we have created $y_1, y_2, \ldots, y_n \overset{iid}{\sim} \text{Exp}(1)$.

Lecture 8: Oct 3

Monte Carlo Integration

If an integral of the form

$$v = \int_{-\infty}^{\infty} H(x) \cdot f(x)\, dx$$

satisfies

1. $f(x)$ is the PDF of a continuous RV
2. $\int_{-\infty}^{\infty} \left|H(x)\right| \cdot f(x)\, dx < \infty$ is finite

Remark: If $X_1, X_2, \ldots,\ X_n$ are continuous iid RVs on $(\Omega, \mathbb{P})$, and have the same PDF $f(x)$, then

$$\mathbb{E}H(X_1) = \int_{-\infty}^{\infty} H(x) \cdot f(x)\, dx$$

exists (by condition 2).

We can partition Ω such that

$$\Omega_1 = \{\omega \in \Omega : \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i(\omega)) = \mathbb{E}H(X_1)\}$$
$$\Omega_2 = \Omega_1^c$$

But by the LLN, $\mathbb{P}(\Omega_1) = 1$ and $\mathbb{P}(\Omega_0) = 0$.

Suppose we (pretend we) have generated true-RNs $x_1, x_2, \ldots, x_n$ from the given PDF $f(x)$, i.e. from CDF

$$F(t) = \int_{-\infty}^{t} f(x)\, dx$$

(such as from the inverse CDF method)

By the definition of true-RNs, $\exists \omega^* \in \Omega$ such that

$$x_1 = X_1(\omega^*), \quad x_2 = X_2(\omega^*), \quad \ldots$$

Assuming we are not extremely unlucky,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(x_i) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i(\omega^*)) = \mathbb{E}[H(X_1)] = \int_{-\infty}^{\infty} H(x) \cdot f(x)\, dx$$

All together, this allows us to define an estimator which approximates the integral v with large enough n such that

$$\boxed{\widehat{v} = \frac{1}{n} \sum_{i=1}^{n} H(x_i) \approx \int_{-\infty}^{\infty} H(x) \cdot f(x)\, dx}$$

Examples

- $H(x) = x, \quad f(x) = \text{PDF of Unif}(0,1) = \mathbb{1}(0 < x < 1)$ So

$$v = \int_{-\infty}^{\infty} H(x) f(x)\, dx = \int_0^1 H(x)\, dx = \int_0^1 x\, dx = \frac{1}{2}$$

  Generating 10000 RNs from Unif$(0,1)$, we get

$$\widehat{v} = 0.5016$$

- $H(x) = x^5, \quad f(x) = \mathbb{1}(0 < x < 1)$

$$v = \int_0^1 x^5\, dx = \frac{1}{6} \approx 0.1667$$

  A And again with $n = 10000$,

$$\widehat{v} = 0.1697$$

Estimation Error

$$\begin{aligned} e_n &= \left|\widehat{v}_n - v\right| \\ &= \left|\frac{1}{n}\sum_{i=1}^{n} H(X_i(\omega)) - \mathbb{E}[H(X_1)]\right| \\ &\overset{LIL}{\leq} \sqrt{\text{Var } H(X_1)} \cdot \sqrt{\frac{2\log(\log(n))}{n}} \end{aligned}$$

Problem:

We used the Monte Carlo method in the first place because we could not calculate $\mathbb{E}[H(X_1)]$ but

$$\text{Var } [H(X_1)] = \mathbb{E}[(H(X_1))^2] - (E[H(X_1)])^2$$

so we cannot actually calculate the error of the estimator. So, in practice, we use the *sample variance*

$$\text{Var } (H(X_1)) \approx \widehat{\sigma_n^2} = \frac{1}{n-1}\sum_{i=1}^{n} \left(H(x_i) - \widehat{v}_n\right)^2$$

Which gives us the approximation for the error

$$e_n = |\widehat{v}_n - v| \leq \sqrt{\widehat{\sigma_n^2}} \cdot \sqrt{\frac{2\log(\log n)}{n}}$$

Riemann Sum Integration

Let us focus on the specific integral

$$\int_0^1 H(x)\, dx = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(\frac{i}{n})$$

From calculus, this sum $\sum_{i=1}^{n} \frac{1}{n} H(\frac{i}{n})$ is just the sum of the area of the estimating rectangles of height $H(i/n)$ and width $1/n$.

Applying this to the same problems as the Monte Carlo integration, the Riemann Sum Estimator

$$\boxed{\widehat{R}_n = \frac{1}{n} \sum_{i=1}^{n} H(\frac{i}{n})}$$

does much better than Monte Carlo Estimator $\widehat{v}_n$:

- $v = \int_0^1 x\, dx = 0.5, \quad \widehat{v}_{10000} = 0.5016, \quad \widehat{R}_{10000} = 0.5$
- $v = \int_0^1 x^5\, dx \approx 0.1667, \quad \widehat{v}_{10000} = 0.1697, \quad \widehat{R}_{10000} = 0.1667$

So, in general, the Riemann estimator is much more accurate.

Lecture 9: Oct 5

High Dimensional Integrals

Question: If the Riemann estimator is more accurate, why would we ever use Monte Carlo integration?

Answer: High dimensional space. Consider:

$$\underbrace{\int_0^1 \cdots \int_0^1 \int_0^1}_{\text{100 integrals}} f(t_1, t_2, \ldots, t_{100})\, dt_1\, dt_2 \ldots dt_{100} \approx \underbrace{\frac{1}{n} \sum_{i=100}^{n} \ldots \frac{1}{n} \sum_{i=2}^{n} \frac{1}{n} \sum_{i=1}^{n}}_{\text{100 averages}} f(\frac{i_1}{h}, \frac{i_2}{h}, \ldots \frac{i_{100}}{h})$$

In code, this would be something like 100 nested for loops calculating n^{100} terms. In R, even 10^{12} terms is already 0.75 TB!

Interlude: High-Dimensional Probability Theory

Random Vector: $\vec{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(n)})$ is a $\mathbb{R}^d$-valued random variable if $\vec{X} : \Omega \to \mathbb{R}^d$

Note: each $X^{(i)}$ is also a random variable $X^{(i)} : \Omega \to \mathbb{R}$

CDF: the CDF of a random vector $\vec{X}$ is

$$F_{\vec{X}}(x_1, x_2, \ldots, x_d) = \mathbb{P}(\omega \in \Omega : \bigcap_{i=1}^{d} X^{(i)}(\omega) \leq x_i)$$

Continuous Random Vector: if each partial derivative $\frac{\partial}{\partial x_i} F(x_1, x_2, \ldots, x_d)$ exists piecewise, then $\vec{X}$ is a continuous random vector

Further, the *PDF* of $\vec{X}$ is

$$f(x_1, x_2, \ldots, x_d) := \frac{\partial}{\partial x_1} \frac{\partial}{\partial x_2} \ldots \frac{\partial}{\partial x_d} F(x_1, x_2, \ldots, x_d)$$

Expectation:

Let $H : \mathbb{R}^d \to \mathbb{R}$. Then $H(\vec{X}) : \Omega \to \mathbb{R}^d \to \mathbb{R}$ so it is a random variable. Thus, if $\vec{X}$ is continuous and $\int_{\mathbb{R}^d} \left|H(\vec{x})\right| \cdot f(\vec{x})\, d\vec{x} < \infty$, then

$$\mathbb{E}[H(\vec{X})] = \int_{\mathbb{R}^d} H(x_1, \ldots, x_d) \cdot f(x_1, \ldots, x_d)\, dx_1 \ldots dx_d$$

High-dimensional Monte Carlo

1. Generate n random vectors iid from the PDF f of a d-dimensional distribution
2. Then the LLN implies
$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(\vec{X}_i) = \mathbb{E}[H(\vec{X}_1)] = \int_{\mathbb{R}^d} H(\vec{x}) \cdot f(\vec{x})\, d\vec{x}$$
3. We define an estimator
$$\widehat{v}_n = \frac{1}{n} \sum_{i=1}^{n} H(\vec{X}_i)$$
4. The error of the estimator is
$$e_n = \left|\widehat{v}_n - v\right| \leq \sqrt{\text{Var } H(\vec{X}_1)} \cdot \sqrt{\frac{2\log(\log n)}{n}}$$

Generating random vectors

Note that the method described above depends on the assumption that we can generate iid random vectors from a given d-dimensional PDF in the first place. This is a very strong assumption! In general, this is not feasible.

We make the following assumptions in order to generate the vectors:

1. $f(\vec{x}) = \prod_{i=1}^{d} f_i(x_i)$ where each f_i is the PDF of a $\mathbb{R}$-valued RV.
2. If $X^{(1)} \sim f_1,\ X^{(2)} \sim f_2, \ldots$ and $X^{(1)} \ldots X^{(d)}$ are independent, then

$$\vec{X} = (X^{(1)}, X^{(2)}, \ldots,\ X^{(d)}) \sim f(x_1, x_2, \ldots,\ x_d)$$

Together, these allow us to generate the iid vectors via the following algorithm:

```
For i = 1...n
    For j = 1...d
        Generate X_i^(j) ~ f_j(x_j)
    end
end
```

This allows us to generate the vector with only $n \cdot d$ numbers instead of the n^d of Riemann integration.

Lecture 9: Oct 10

Importance Sampling

The Curse of Dimensionality: recall that

$$\left|\widehat{v}_n - v\right| \leq \sqrt{\text{Var } H(\vec{X}_1)} \cdot \sqrt{\frac{2\log(\log n)}{n}}$$

When d is large, Var $H(\vec{X}_1)$ is large so the error is large.

Importance Sampling:

$$\begin{aligned} v = \int H(\vec{x}) \cdot f(\vec{x})\, d\vec{x} &= \int \frac{H(\vec{x}) \cdot f(\vec{x})}{g(\vec{x})} \cdot g(\vec{x})\, dx \\ &= \mathbb{E}[\frac{H(\vec{x}) \cdot f(\vec{x})}{g(\vec{x})}] \\ &\approx \frac{1}{n} \sum_{i=1}^{n} \frac{H(\vec{X}_i) \cdot f(\vec{X}_i)}{g(\vec{X}_i)} \end{aligned}$$

where $\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_n \sim g(\vec{x})$, another PDF carefully chosen to minimize variance. So, defining the estimator

$$\widehat{v}_N = \frac{1}{n} \sum_{i=1}^{n} \frac{H(\vec{X}_i) \cdot f(\vec{X}_i)}{g(\vec{X}_i)} \approx v$$

We have a new, smaller variance term given a "properly -chosen" g such that

$$e_n = \left|\widehat{v}_n - v\right| \leq \sqrt{\text{Var } \frac{H(\vec{X}_1) \cdot f(\vec{X}_1)}{g(\vec{X}_1)}} \cdot \sqrt{\frac{2\log(\log n)}{n}}$$

Remark: Here, "properly chosen" means that

1. We know how to generate $X_1, X_2, \ldots, X_n \overset{iid}{\sim} g$
2. $g(\vec{x}) = 0 \implies H(\vec{x}) \cdot g(\vec{x}) = 0$
3. $$\frac{H(\vec{x}) \cdot f(\vec{x})}{g(\vec{x})} \approx \text{constant}$$

Ideally,

$$\frac{H(\vec{x}) \cdot f(\vec{x})}{g(\vec{x})} = c \implies \text{Var } \frac{H(\vec{x}) \cdot f(\vec{x})}{g(\vec{x})} = \text{Var } c = 0$$

But this is unrealistic because

$$I = \int g(\vec{x})\, dx = \frac{1}{c} \int H(\vec{x}) \cdot g(\vec{x})\, dx \implies g6*(\vec{x}) = \frac{H(\vec{x}) \cdot f(\vec{x})}{\int H(\vec{x}) \cdot f(\vec{x})\, dx}$$

where the denominator is exactly what we want to calculate in the first place; if we knew it, we would be done.

In applications, we focus on a function family $\mathfrak{F}$ of PDFs and choose the element that is "most similar" to the ideal g^*. This is chosen by minimizing the difference across the family of functions (either by the Kullback-Leibler Divergence or Wasserstein Metric)

$$\arg\min_{f\in\mathfrak{F}} D_{KL}(g^*\big|\big|f) \quad \text{or} \quad \arg\min_{f\in\mathfrak{F}} W(g^*, f)$$

Lecture 10: Oct 12

Above, we were able to generate d-dimensional random vectors for Monte Carlo integration only when we had a PDF f which could be factored into 1-dimensional PDFs

$$f(x_1, x_2, \ldots, x_n) = \prod_{j=1}^{d} f_j(x_j)$$

What if we do not have the factorization?

Example: $d = 2$

$$(X^{(1)}, X^{(2)}) = f(x_1, x_2) = \frac{f(x_1, x_2)}{\int f(x_1, x_2)\, dx_2} \cdot \int f(x_1, x_2)\, dx$$

with the PDF of $X^{(1)}$ defined as

$$f_1(x_1) := \int f(x_1, x_2)\, dx_2$$

and

$$f_{2\,|\,1}(x_2 \mid x_1) := \frac{f(x_1, x_2)}{f_1(x_1)}$$

is the conditional PDF of $X^{(2)}$ given $X^{(1)} = x_1$

Then, we can generate a 2-d random vector

$$(X^{(1)}, X^{(2)}) \sim f(x_1, x_2) = f_{2|1}(x_2 \mid x_1) \cdot f_1(x_1)$$

where $f(x_1, x_2) = f_{2|1}(x_2 \mid x_1) \cdot f_1(x_1)$ is **Baye's Law**

The Algorithm:

1. Generate $X^{(1)} \sim f_1$ (using MCG and inverse CDF)

2. Generate $X^{(2)} \sim f_{2|1}(x_2 \mid X^{(1)})$
3. Output: $\vec{X} = (X^{(1),X^{(2)}}) \sim f(x_1, x_2)$

A Problem: In step 1, we generate

$$X^{(1)} \sim f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2$$

which may be computationally expensive because we need to compute this integral for *every* x_1 (which could be infinite!)

Example: $d = 100$

Using the same process,

$$X^{(1)} \sim f_1(x_1) = \underbrace{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{\text{99 integrals}} f(x_1, x_2, \ldots, x_{100})\, dx_2\, dx_3\, \ldots\, dx_{100}$$

which again needs to be calculated for potentially infinitely many x_1!

Then in step 2, this would all need to be repeated with 98 infinite integrals!

Clearly, this is a terrible way to generate random vectors if your goal is to solve a single integral.

Conclusion: Generating a high-dimensional RV from a given high-dimensional distribution is infeasible in applications. We need a new approach.

Conditional Probability

Definition: Let $(\Omega, \mathbb{P})$ be a probability space. $A \subseteq \Omega$ and $\subseteq \Omega$ are two events.

1. If $\mathbb{P}(B) > 0$, then the conditional probability of A given B is

$$\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

2. If $\mathbb{P}(B) = 0$ then $\mathbb{P}(A \mid B)$ is not defined (in undergrad-level math)

Furthermore, we define a map

$$\tilde{\mathbb{P}} : \{A : A \subseteq \Omega\} \to \mathbb{R}$$

Claim: $\tilde{\mathbb{P}}(A) = \mathbb{P}(A \mid B)$ is a probability on Ω

Proof: HW in APMA 1655.

Law of Total Probability: Let $(\Omega, \mathbb{P})$ be a probability space with partition $\{B_1, B_2, \ldots, B_n\}$. If $\mathbb{P}(B_i) > 0$ for $i = 1, 2, \ldots, n$ then

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \mid B_i) \cdot \mathbb{P}(B_i) \quad \forall A \subseteq \Omega$$

Markov Chains

Assume the **state space** $\mathfrak{X}$ is a discrete subset of $\mathbb{R}^d$. $\{X_n\}_{n=1}^{\infty}$ is a sequence of RV $X_n : \Omega \to \mathfrak{X}$

Definition:

1. The sequence $\{X_n\}_{n=1}^{\infty}$ defined above is a Markov chain if

   $$\mathbb{P}\left(X_{n+1} = y \,\middle|\, X_n = x_1, X_{n-1} = \xi_{n-1}, \ldots, X_0 = \xi_0\right) = \mathbb{P}(X_{n+1} = y \mid X_n = x)$$

   Heuristically, the sequence is a Markov Chain if the future state depends only on the present state and not any past state for all $n = 0, 1, 2, \ldots$, all $y \in \mathfrak{X}$ and all $x_1, \xi_{n-1}, \ldots, \xi_0$ such that

   $$\mathbb{P}(X_n = x, X_{n-1} = \xi_{n-1}, \ldots, X_0 = \xi_0) > 0$$

2. Furthermore, if there exists a function $p(x, y) : \mathfrak{X}x\mathfrak{X} \to [0, 1]$ such that

   $$\mathbb{P}(X_{n+1} = y \mid X_n = x) = p(x, y)$$

   (the conditional probability does not depend on n) then the Markov Chain $\{X_n\}_{n=0}^{\infty}$ is called a **homogeneous Markov Chain**

   The function p is called the **transition probability** of the HMC. Heuristically, it is the probability of moving from x to y.

If $\{X_n\}_{n=0}^{\infty}$ is a HMC, we have a function p. Further, $X_0 : \Omega\mathfrak{X}$, whose codomain is a discrete set. We note that X_0 then has a probability mass function (PMF)

$$\mu(x) := \mathbb{P}(X_0 = x) \quad \forall x \in \mathfrak{X}$$

Together, the functions μ and p contain all of the information of the distribution of the MC:

$$\mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = \mu(x_0) \prod_{i=0}^{n-1} p(x_i, x_{i+1})$$

Example

Let $\xi_1, \xi_2, \ldots, \xi_n$ be iid $\mathbb{Z}^d$-valued RVs on $(\Omega, \mathbb{P})$, i.e. $\xi : \Omega \to \mathbb{Z}^d = \mathbb{Z} \times \cdots \times \mathbb{Z}$

We define

$$X_n(\omega) = \begin{cases} x_0 & n = 0 \\ x_0 + \sum_{i=1}^{n} \xi_i(\omega) & n \geq 1 \end{cases}$$

Then, $\{X_n\}_{n=0}^{\infty}$ is the **random walk** from x_0.

Lecture 11: Oct 17

Overview of the Monte Carlo Markov Chain

Ergodic Theorem: Suppose $\{X_n\}_{n=0}^{\infty}$ is a homogenenous Markov Chain satisfying

1. it is "recurrent"
2. it is "irreducible"
3. it is "aperiodic"
4. it has a "stationary distribution" π which is a PMF on $\mathfrak{X}$

Then we have

1. $X_n \dot{\sim} \pi$ when n is large
2. For any function f such that $E[f(X)]$ exists (with $X \sim \pi$), then

$$\mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(X_i(\omega)) = \mathbb{E}[f(X)]\right\}\right) = 1$$

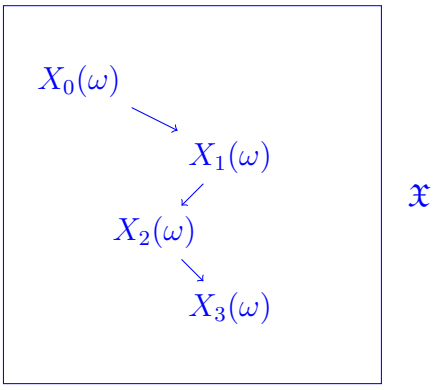
Remarks:

- The first result helps us generate RVs from π (approximately)
- The second result looks like the LLN. The difference lies in the fact that the LLN requires the random variables X_i are iid. The ergodic theorem only requires they come from the same Markov Chain. Thus, we can estimate $\mathbb{E}[f(X)]$ $(X \sim \pi)$ with the estimator

$$\widehat{v}_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

Recurrence and Transience

Suppose we have a homogenenous Markov Chain $\{X_n\}_{n=0}^{\infty}$ with $X_N : \Omega \to \mathfrak{X}$ for all n.



We call the subscript n the "time point".

With $y \in \mathfrak{X}$ fixed,

$$T_y(\omega) := \min\{n > 0 : X_n(\omega) = y\}$$

is "the time at which the Markov chain first visits y"

$$T_y(\omega) = +\infty \iff \{X_n\}_{n=0}^{\infty} \text{ will never visit } t$$

Then we define

$$\rho_{xy} = \mathbb{P}(T_y < +\infty \mid X_0 = x)$$

Definition: Let $\{X_n\}_{n=0}^{\infty}$ be a HMC with state space $\mathfrak{X}$.

1. For $y \in \mathfrak{X}$, the state y is called **recurrent** if $\rho_{yy} = 1$
2. If y is not recurrent, then it is **transient**

Interpretation: if y is recurrent and the MC starts from y, the MC will return to y with probability 1

Example: Simple Random Walks

Example 1: 1-dimensional SRW

Fix a point $x_0 \in \mathbb{Z}$.

$$\xi_1, \xi_2, \ldots, \xi_n \overset{iid}{\sim} \text{Unif}(\{-1, 1\}) \implies \mathbb{P}(\xi_i = 1) = \mathbb{P}(\xi_i = -1) = 1$$

Then

$$X_n(\omega) = \begin{cases} x_0 & n = 0 \\ x_0 + \sum_{i=1}^{n} \xi_i(\omega) & n = 1, 2, \ldots \end{cases} \implies X_{n+1} = X_n + \xi_{n+1}$$

Claim: $\{X_n\}_{n=0}^{\infty}$ is a markov chain.

Proof: a sequence of RVs is a MC if

$$\mathbb{P}(X_{n+1} = y \mid X_n = x, X_{n-1} = \xi_{n-1}, \ldots, X_0 = \xi_0) = \mathbb{P}(X_{n+1} = y \mid X_n = x)$$

First we take the LHS. By the definition of conditional probability,

$$\mathbb{P}(X_{n+1} = y \mid X_n = x, X_{n-1} = \xi_{n-1}, \ldots, X_0 = \xi_0) = \frac{\mathbb{P}(X_{n+1} = y, X_n = x, X_{n-1} = \xi_{n-1}, \ldots, X_0 = \xi_0)}{\mathbb{P}(X_n = x, X_{n-1} = \xi_{n-1}, \ldots, X_0 = \xi_0)}$$

However,

$$X_{n+1} = X_n + \xi_{n+1} \implies \xi_{n+1} = X_{n+1} - X_n = y - x$$

so

$$P(X_{n+1} = y) = \mathbb{P}(\xi_{n+1} = y - x)$$

Further, since ξ_n is independent of X_n and all X_m with $m \leq n$, we can separate out the probabilities:

$$\begin{aligned} \frac{\mathbb{P}(X_{n+1} = y, X_n = x, X_{n-1} = \xi_{n-1}, \ldots, X_0 = \xi_0)}{\mathbb{P}(X_n = x, X_{n-1} = \xi_{n-1}, \ldots, X_0 = \xi_0)} &= \frac{\mathbb{P}(\xi_{n+1} = y - x, X_n = x, X_{n-1} = \xi_{n-1}, \ldots, X_0 = \xi_0)}{\mathbb{P}(X_n = x, X_{n-1} = \xi_{n-1}, \ldots, X_0 = \xi_0)} \\ &= \frac{\mathbb{P}(\xi_{n+1} = y - x) \cdot \mathbb{P}(X_n = x, X_{n-1} = \xi_{n-1}, \ldots, X_0 =}{\mathbb{P}(X_n = x, X_{n-1} = \xi_{n-1}, \ldots, X_0 = \xi_0)} \\ &= \mathbb{P}(\xi_{n+1} = y - x) \end{aligned}$$

Now for the RHS,

$$\begin{aligned}\mathbb{P}(X_{n+1} = y \mid X_n = x) &= \frac{\mathbb{P}(X_{n+1} = y, X_n = x)}{\mathbb{P}(X_n = x)} \\ &= \frac{\mathbb{P}(\xi_{n+1} = y - x, X_n = x)}{\mathbb{P}(X_n = x)} \\ &= \frac{\mathbb{P}(\xi_{n+1} = y - x, X_n = x)}{\mathbb{P}(X_n = x)} \\ &= \mathbb{P}(\xi_{n+1} = y - x)\end{aligned}$$

Thus LHS = RHS and $\{X_n\}_{n=0}^{\infty}$ is a markov chain. ■

Further, it is recurrent.

Example 2: In a 2-dim SRW, $\rho_{\vec{0},\vec{0}} = 1$ ($\vec{0}$ is recurrent)

Example 3: In a d-dim SRW ($d \geq 3$), $\rho_{\vec{0},\vec{0}} = \mathbb{P}(T_{\vec{0}} < +\infty \mid X_0 = \vec{0}) < 1$