APMA 1690: Computational Probability and Statistics

Milan Capoor

Fall 2023

Contents

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)

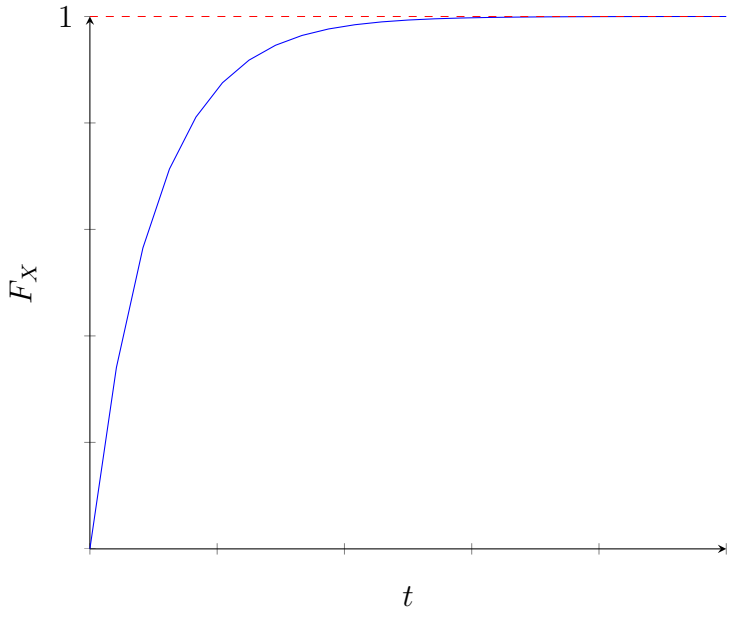
Definition: Let $X : (\Omega, \mathbb{P}) \to \mathbb{R}$. Then the CDF of X is $F_X : \mathbb{R} \to \mathbb{R}$

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

Examples:

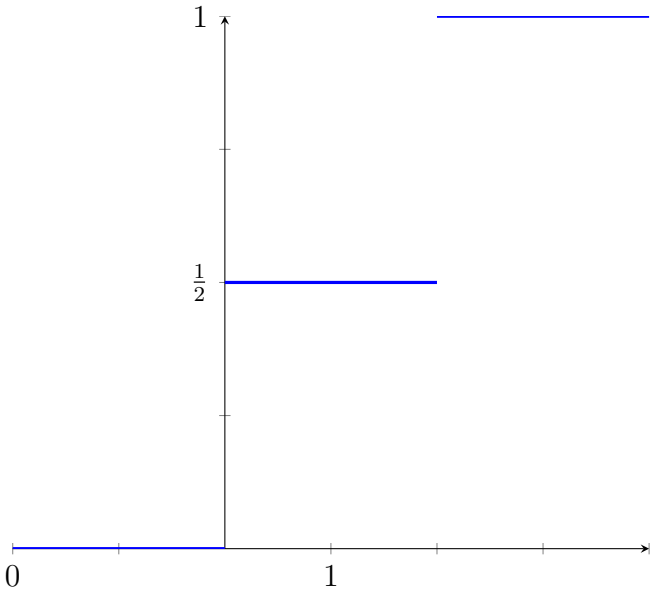
- $X \sim \text{Exp}(\lambda)$

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



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

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



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,

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

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

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

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. $\int_{-\infty}^{\infty} \big|H(x)\big| \cdot f(x)\, dx < \infty$ 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

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

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:

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

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

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

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 $p(x, y) : \mathfrak{X} \times \mathfrak{X} \to [0, 1]$ 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

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

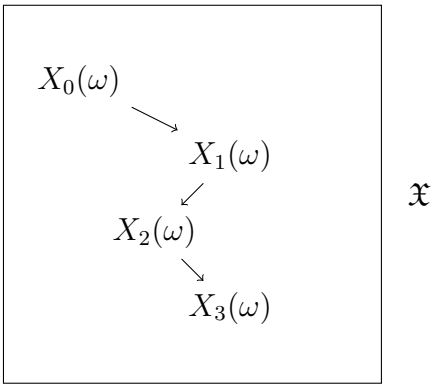
Remarks:

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

Lecture 12: Oct 19

Properties of Recurrence

Recurrence:

$$\rho_{xy} = \mathbb{P}(\text{the MC will visit y} \mid \text{it starts from x}) = \mathbb{P}(T_y < +\infty \mid X_0 = x)$$

If $\rho_{yy} = 1$ then y is recurrent.

Theorem: Let $\{X_n\}_{n=0}^{\infty}$ be a HMC with state space $\mathfrak{X}$. For any point $y \in \mathfrak{X}$,

1. y is recurrent iff
$$\sum_{n=1}^{\infty} \mathbb{P}(X_n = y \mid X_0 = y) = +\infty$$

2. If y is not recurrent (y is transient), then
$$\sum_{n=1}^{\infty} \mathbb{P}(X_n = y \mid X_0 = x) < \infty \qquad \forall x \in \mathfrak{X}$$

Theorem ("Contagion"): If x is recurrent ($\rho_{xx} = 1$) and $\rho_{xy} > 0$. Then,

1. $\rho_{yy} = 1$ (y is recurrent)
2. $\rho_{xy} = \rho_{yx} = 1$

Irreducibility

Definition: Let $\{X_n\}_{n=0}^{\infty}$ be a HMC with state space $\mathfrak{X}$. If $\rho_{xy} > 0$ for all $x, y \in \mathfrak{X}$ then the MC is *irreducible*

Total recurrence: If the MC is irreducibile and has at least one recurrent point, all states in $\mathfrak{X}$ are recurrent

Finite State Spaces

From now on, we assume all state spaces are finite.

Theorem: Let $\{X_n\}_{n=0}^{\infty}$ be an HMC with state space $\mathfrak{X}$. If $\#\mathfrak{X} < +\infty$, then $\mathfrak{X}$ has at least one recurrent state.

Proof: Suppose the theorem is false. Then,
$$\sum_{n=1}^{\infty} \mathbb{P}(X_n = y \mid X_0 = x) < \infty \quad y, x \in \mathfrak{X}$$

i.e. the sum if finite so we take the sums across the entire (finite) state space:
$$\sum_{y \in \mathfrak{X}} \sum_{n=1}^{\infty} \mathbb{P}(X_n = y \mid X_0 = x)$$

But since this is a finite sum of finite terms, the whole sum should be finite.

Exchanging the summations, we get the sum of a union of events:

$$\begin{aligned}\sum_{y\in\mathfrak{X}}\sum_{n=1}^{\infty}\mathbb{P}(X_n = y \mid X_0 = x) &= \sum_{n=1}^{\infty}\sum_{y\in\mathfrak{X}}\mathbb{P}(X_n = y \mid X_0 = x)\\ &= \sum_{n=1}^{\infty}\mathbb{P}(y \in \mathfrak{X} \mid X_0 = x)\\ &= \sum_{n=1}^{\infty} 1 = \infty\end{aligned}$$

But we assumed that $\sum_{n=1}^{\infty}\mathbb{P}(X_n = y \mid X_0 = x) < \infty$ so we have a contradiction. Thus, the theorem is true. ■

Corollary: If the state space is finite and the MC is irreducible, the contagious property tell us that all states in the state space are recurrent.

Transition matrices

Let $\{X_n\}_{n=0}^{\infty}$ be a HMC with

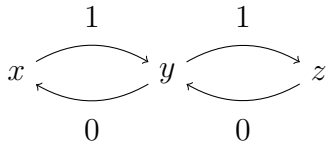
- $\mathfrak{X} = \{x_1, x_2, \ldots, x_S\}$
- transition probability: $p : \mathfrak{X} \cdot x \cdot \mathfrak{X} \to [0, 1]$

Transition Matrix:

$$P = \begin{pmatrix} p(x_1, x_1) & p(x_1, x_2) & \ldots & p(x_1, x_S) \\ p(x_2x_1) & p(x_2, x_2) & \ldots & p(x_2, x_S) \\ \vdots & \vdots & \ddots & \vdots \\ p(x_S, x_1) & p(x_S, x_2) & \ldots & p(x_S, x_S) \end{pmatrix}$$

P is a $S \times S$ matrix where $\mathbb{P}_{ij} = p(x_i, x_j) = \mathbb{P}(X_1, x_j \mid X_0 = x_i)$

Example: The chain



is represented by

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

From $\mathbb{P}(X_1 = x_j \mid X_0 = x_i) = P_{ij}$,

$$\begin{aligned} \mathbb{P}(X_2 = x_j \mid X_0 = x_i) &= \sum_{k=1}^{S} \mathbb{P}(X_2 = x_j \mid X_0 = x_i, X_1 = x_k) \cdot \mathbb{P}(X_1 = x_k \mid X_0 = x_i) \\ &= \sum_{k=1}^{S} \mathbb{P}(X_2 = x_j \mid X_1 = x_k) \cdot \mathbb{P}(X_1 = x_k \mid X_0 = x_i) \\ &= \sum_{k=1}^{S} P_{kj} \cdot P_{ik} \\ &= (P^2)_{ij} \end{aligned}$$

Then by induction,

$$\mathbb{P}(X_n = x_j \mid X_0 = x_i) = (P^n)_{ij}$$

What if we want to calculate a state without a condition?

We now define

$$\mu = \begin{pmatrix} \mathbb{P}(X_0 = x_1) \\ \mathbb{P}(X_0 = x_2) \\ \vdots \\ \mathbb{P}(X_0 = x_S) \end{pmatrix}$$

So by the law of total probability,

$$\begin{aligned} \mathbb{P}(X_n = x_j) &= \sum_{j=1}^{S} \mathbb{P}(X_n = x_j \mid X_0 = x_i) \cdot \mathbb{P}(X_0 = x_i) \\ &= \sum_{i=1}^{S} (P^n)_{ij} \cdot \mu_i \\ &= ((P^n)^T \mu)_j = (\mu^T P^n)_j \end{aligned}$$

Lecture 13: Oct 24

Review

$\{X_n\}_{n=0}^{\infty}$ is a HMC with state space $\mathfrak{X}$

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

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

A point y is *recurrent* if $\rho_{yy} = 1$

Recurrence is "contagious":

$$\begin{cases} \rho_{xx} = 1 \\ \rho_{yy} > 0 \end{cases} \implies \begin{cases} \rho_{yy} = 1 \\ \rho_{xy} = \rho_{yx} = 1 \end{cases}$$

The MC is *irreducible* if $\rho_{xy} > 0$ for all $x, y \in \mathfrak{X}$. By contagion, if there is one recurrent $y \in \mathfrak{X}$, then all $x \in \mathfrak{X}$ are recurrent.

Theorem: If $\#\mathfrak{X} < \infty$, the MC will have at least one recurrent point. Thus, all irreducible finite MCs are totally recurrent

Directed Graphs

With state space

$$\mathfrak{X} = \{x_1, \ldots, x_S\} \qquad S = \#\mathfrak{X}$$

We have a transition matrix defined by

$$P_{ij} = p(x_i, x_j)$$

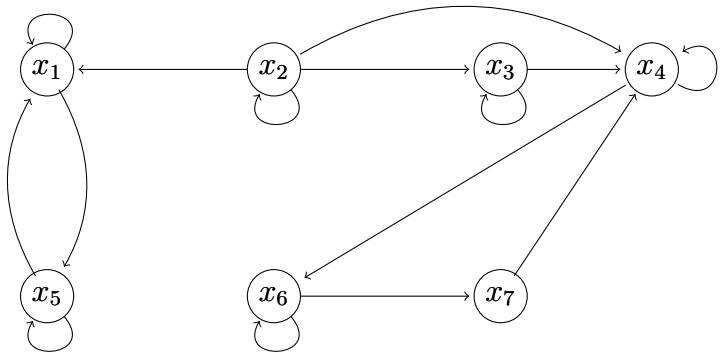
Then, the **directed graph** $G(P) = (V, E)$ where V is the set of vertices of the graph and E is the set of directed edges between vertices.

The set of vertices is precisely the state space $\mathfrak{X}$ and the set of directed edges are the connections from x_i to x_j for the points where $p(x_i, x_j) = P_{ij} > 0$

The transition matrix

$$P = \begin{pmatrix} 0.3 & 0 & 0 & 0 & 0.7 & 0 & 0 \\ 0.1 & 0.2 & 0.3 & 0.4 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0.5 & 0 \\ 0.6 & 0 & 0 & 0 & 0.4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.2 & 0.8 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

is represented by the graph



Claim: Let $\{X_n\}_{n=0}^{\infty}$ be an HMC. Its state space and transition matrix are $\mathfrak{X} = \{x_1, \ldots, x_S\}$ and P. The MC is irreducible if and only if $\forall x_i, x_j \in \mathfrak{X}$, there x a directed edge from $x_1 \to x_j$ and a directed path from $x_j \to x_i$

Stationary Distributions (SDs)

Let $X_n \dot{\sim} \pi$ for large n:

$$\pi(x) = \lim_{n \to \infty} \mathbb{P}(X_n = x)$$

By the Law of total probability,

$$\begin{aligned}\pi(x) &= \lim_{n\to\infty} \lim \mathbb{P}(X_n = x) \\ &= \lim n \to \infty \sum_{y\in\mathfrak{X}} \mathbb{P}(X_n = x \mid X_{n-1} = y) \cdot \mathbb{P}(X_{n-1} = y) \\ &= \sum_{y\in\mathfrak{X}} p(y,x) \cdot \lim_{n\to\infty} \mathbb{P}(X_{n-1} = y) \\ &= \sum_{y\in\mathfrak{X}} p(y,x) \cdot \pi(y)\end{aligned}$$

This leads us to a general formula: if $\pi(x) = \lim_{n\to\infty} \mathbb{P}(X_n = x)$ for all $x \in \mathfrak{X}$ then,

$$\pi(x) = \sum_{y\in\mathfrak{X}} \pi(y) \cdot p(y,x)$$

Definition: Let $\{X_n\}_{n=0}^{\infty}$ be a HMC whose state space and transition probability are $\mathfrak{X}$ and P. Suppose $\pi : \mathfrak{X} \to \mathbb{R}$ is a PMF. Then π is a *stationary distribution* for the MC if

$$\pi(x) = \sum_{y\in\mathfrak{X}} \pi(y) \cdot p(y,x)$$

Define $\vec{\pi} = \begin{pmatrix} \pi(x_1) \\ \vdots \\ \pi(x_S) \end{pmatrix}$. Then

$$\pi(x) = \sum_{y\in\mathfrak{X}} \pi(y) \cdot p(y,x) \iff \vec{\pi} = P^T \vec{\pi}$$

Existence of SDs

Simplex:

$$\triangle = \{(p_1, p_2, \ldots, p_S)^T : p_1 \geq 0, p_2 \geq 0 \ldots, \ p_S \geq 0 \quad \text{and} \ \sum_{k=1}^{S} p_k = 1\}$$

In 3-space, the simplex corresponds to the 2-d triangle between the vertices $(0,0,1), (0,1,0), (1,0,0)$

Theorem: If $\#\mathfrak{X} < \infty$, the HMC has at least one stationary distribution, i.e. $\exists \vec{\pi} \in \triangle : \vec{\pi} = P^T \vec{\pi}$

Proof: Brouwer fixed-point Theorem (Algebraic topology) or Perron-Frobenius theorem (Linear algebra)

Uniqueness of SDs

Theorem: Let $\{X_n\}_{n=0}^{\infty}$ be a HMC with state space $\mathfrak{X}$ and transition matrix P

1. The MC has at least one SD
2. If the MC is irreducible, the SD is unique

Lecture 14: Oct 26

Review of Irreducibility

Theorem: If $\{X_n\}_{n=0}^{\infty}$ is a HMC with finite state space $\mathfrak{X}$, then

1. The MC has at least one stationary distribution (SD) π
2. If the the MC is irreducible the SD is unique and $\pi(x) > 0, \quad \forall x \in \mathfrak{X}$

Aperiodicity

Motivation: The Ergodic theorem seeks to calculate the asymptotic distribution $\lim_{n\to\infty} \mathbb{P}(X_n = x)$

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(X_n = x) &\overset{LTP}{=} \sum_{i=1}^{S} \lim_{n\to\infty} \mathbb{P}(X_n = x_j \mid X_0 = x_i) \cdot \mathbb{P}(X_0 = x_i) \\ &= \sum_{i=1}^{S} \lim_{n\to\infty} (P^n)_{ij} \cdot \mathbb{P}(X_0 = x_i)\end{aligned}$$

So we need to caluclate $\lim_{n\to\infty} P^n$.

Example:

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \implies P^n = \begin{cases} I & \text{n even} \\ P & \text{n odd} \end{cases}$$

So there is a "periodic pattern" to powers of P and the limit does not exist.

Definition: Let P be the transition matrix of an HMC whose state space is $\mathfrak{X} = \{x_1, x_2, \ldots, x_S\}$. Suppose the MC is irreducible. Let

$$I_i := \{n \geq 1 : (P^n)_{ii} > 0\}$$

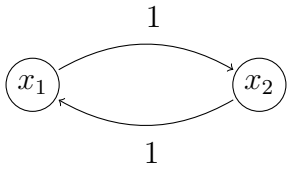
and d_i be the greatest common divisor of I_i. d_i is the *period* of x_i.

In the example above,

$$I_1 = \{n \geq 1 : (P^n)_{11} > 0\} = I_2 = \{n \geq 1 : (P^n)_{22} > 0\} = \{2, 4, 6, 8, \ldots\}$$

so $d_1 = d_2 = 2$

And in fact that example represent the simple markov chain



Theorem: $d_1 = d_2 = \cdots = d_S$

Proof: Omitted

Definition: the *period* of the irreducible MC is $d := d_1 = d_2 = \cdots = d_S$

Remark: We need the irreducibility constraint to ensure that $I_i \neq \emptyset$:irreducibility and finite state space means the MC is totally recurrent so

$$\infty = \sum_{n=1}^{\infty} \mathbb{P}(X_n = x_i \mid X_0 = x_i) = \sum_{n=1}^{\infty} (P^n)_{ii} \implies \exists i : (P^n)_{ii} > 0 \implies I_i \neq \emptyset$$

Definition: an irreducible HMC is *aperiodic* if its period is $d = 1$

1st Ergodic Theorem

1st Ergodic Theorem: Let $\{X_n\}_{n=0}^{\infty}$ be an HMC with finite state space. If the MC is irreducible and aperiodic, then

$$\lim_{n \to \infty} \mathbb{P}(X_n = x_j \mid X_0 = x_i) = \pi(x_j) \qquad \forall i, j \in \{1, 2, \ldots, S\}$$

where π is the unique SD of the MC

Proof: See lecture notes

Remark: assuming that the state space was finite let us implicitly assume that the MC is recurrent (because one point is and it is irreducible) and that there is at least one SD (which is unique by irreducibility)

Corollary: Under the conditions of the 1st Ergodic Theorem,

$$\lim_{n\to\infty} \mathbb{P}(X_n = x_j) = \lim_{n\to\infty} \mathbb{P}(X_n = x_j \mid X_0 = x_i) = \pi(x_j)$$

Proof:

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}(X_n = x_j) &= \sum_{i=1}^{n} \left[\lim_{n\to\infty} \mathbb{P}(X_n = x_j \mid X_0 = x_i)\right] \cdot \mathbb{P}(X_0 = x_i) \\
&= \sum_{i=1}^{S} \pi(x_j) \cdot \mathbb{P}(X_0 = x_i) \\
&= \pi(x_j) \sum_{i=1}^{S} \mathbb{P}(X_0 = x_i) \\
&= \pi(x_j) \cdot \mathbb{P}(x_i \in \mathfrak{X}) = \pi(x_j)
\end{aligned}$$

Application: Suppose we are interested in a distribution π. If

1. π happens to be the SD of a MC $\{X_n\}_{n=0}^{\infty}$
2. We know how to generate this MC

Then $X_n \dot{\sim} \pi$ when n is large (so we can view X_n as a RV drawn from π)

2nd Ergodic Theorem:

2nd Ergodic Theorem: Let $\{X_n\}_{n=0}^{\infty}$ be an HMC with finite state space. If the MC is irreducible, for any $f : \mathfrak{X} \to \mathbb{R}$, $\sum_{x\in\mathfrak{X}} \pi(x) \cdot \left|f(x_i)\right| < \infty$ where π is the SD of the MC (unique by Irreducibility), then

$$\mathbb{P}\left(\omega \in \Omega : \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(X_i(\omega)) = \sum_{x\in\mathfrak{X}} \pi(x) \cdot f(x)\right) = 1$$

Application: Suppose we are interested in computing $\sum_{x\in\mathfrak{X}} \pi(x)\cdot f(x)$ (perhaps in the Ising Model of statistical mechanics, $\sum_{x\in\mathfrak{X}} \pi(x)\cdot f(x) =$ "average mangetization").

We can use an estimator:

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

Markov Chain Monte Carlo

Problem: Given a distribution π, we want to derive a transition matrix P such that π is the SD of P

Note that, in general, this is a much harder problem than finding π given P!

Solutions:

- Method 1: Metropolis-Hastings Algorithm
- Method 2: Gibbs Sampling

Lecture 15: Oct 31

Review

Ergodic Theorem: Let $\{X_n\}_{n=0}^{\infty}$ be an HMC on finite state space $\mathfrak{X}$. Then, we have

1. The MC is recurrent (all $x \in \mathfrak{X}$) are recurrent
2. The MC has a unique stationary distribution $\pi : \mathfrak{X} \to \mathbb{R}$
3. $X_n \sim \pi$ at large n. i.e., $\lim_{n\to\infty} \mathbb{P}(X_n = x) = \pi(x) \quad \forall x \in \mathfrak{X}$
4. $$\mathbb{P}(\{\omega \in \Omega : \lim_{n\to\infty} \frac{1}{n} f(X_i(\omega)) = \mathbb{E}_\pi f(X)\}) = 1$$
   where $f : \mathfrak{X} \to \mathbb{R}$ and $\mathbb{E}_\pi f(X) = \sum_{x\in\mathfrak{X}} \pi(x)\cdot f(x)$

This last statement is exactly like the Law of Large Numbers except in the LLN, the RVs are required to be iid. Here, they just need to be elements of the same Markov Chain.

Markov Chain Monte Carlo

Suppose we are given a distribution $\pi : \mathfrak{X} \to \mathbb{R}$ and have derived a transition probability $p : \mathfrak{X} \times \mathfrak{X} \to \mathbb{R}$ satisfying

$$\pi(x) = \sum_{y \in \mathfrak{X}} \pi(y) \cdot p(y, x) \quad \forall x \in \mathfrak{X}$$

(i.e. it is a stationary distribution)

Then we can use the following algorithm to generate a sequence of RV $\{X_n\}_{n=0}^{\infty}$ that form a MC with transition probability p and SD π:

```
Initialize X0

for n = 1, 2, ...
```

$$X_n \sim p(X_{n-1}, \cdot)$$

```
end for
```

Remarks:

1. $p(X_{n-1}, \cdot)$ is a two-variable function with one input fixed. Thus the algorithm step corresponds to generating $X_n \sim p_{X_{n-1}}(y)$ (generating an RV from a PMF)
2. $$p(x, y) \geq 0$$ $$\mathbb{P}(X_1 = y \mid X_0 = x) \geq 0$$

Goals:

1. $X \dot{\sim} \pi$
2. Approx $\sum_{x \in \mathfrak{X}} \pi(x) \cdot f(x)$

Problem: in most applications, the dimensionality of $\pi(\vec{x})$ is very large

Summaries of Pictures

We define a *picture* as a vector $x = (x^{(1)}, x^{(2)}, \ldots, x^{(d)})$ where each $x^{(i)}$ is a *pixel*.

A *random picture* is a random vector $X = (X^{(1)}, X^{(2)}, \ldots, X^{(d)}) \sim \pi$.

In general, it is infeasible to generate a random picture from π, but we can generate an HMC $\{X_n\}_{n=0}^{\infty}$ whose SD is π so $X_n \dot{\sim} \pi$.

Then say we have $f : \mathbb{R}^d \to \mathbb{R}$ which gives a *summary of the picture*. For example,

$$f(x) = \frac{1}{d} \sum_{i=1}^{d} x^{(i)}$$

Further, the probability that a particular picture x is observed is $\pi(x) : \mathfrak{X} \to \mathbb{R}$.

So the average summary of a pictures is

$$v = \sum_{x \in \mathfrak{X}} \pi(x) \cdot f(x)$$

For a 256x256 picture, this sum would have 2^{65536} terms, so this obviously cannot be calculated theoretically.

However, using MCMC, we can create an estimator for v (assuming we have a HMC with SD π):

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

The process of creating that MC from π is the central topic of MCMC theory.

Lecture 16: Nov 2

Review of the MCMC

Our goal is to generate random numbers from a given distribution π. If we find the transition probability of an irreducible and aperiodic HMC and sample via $X_n \sim p(X_{n-1}, \cdot)$, then $\{X_n\}_{n=0}^{\infty}$ will be an HMC whose SD is π and thus the sample average will converge to the SD by the ergodic theorem. Thus, sampling from p will give us random variables from π.

Metropolis Algorithm (1953)

Assume we are given a function $\pi : \mathfrak{X} \to \mathbb{R}$, satisfying $\pi(x) > 0 \quad \forall x \in \mathfrak{X}$

We choose a $q : \mathfrak{X} \times \mathfrak{X} \to \mathbb{R}$ satisfying

1. q is the transition probability of an HMC
2. That MC is irreducible and aperiodic
3. q is symmetric: $q(x, y) = q(y, x) \quad \forall x, y \in \mathfrak{X}$
4. For each fixed x, $q(x, \cdot) = q_x(y)$ is a PMF from which we know how to generate RVs in an efficient way

Then the Metropolis Algorithm defines $p(x, y)$ by

$$p(x, y) := \begin{cases} q(x, y) \cdot \min\{1, \frac{\pi(y)}{\pi(x)}\} & x \neq y \\ 1 - \sum_{\xi \in \mathfrak{X} : \xi \neq x} p(x, \xi) & x = y \end{cases}$$

Remark: Because of the ratio $\frac{\pi(y)}{\pi(x)}$ in the Metropolis Algorithm, we do not need the exact formula of π – just the formula up to a constant which will be cancelled.

Theorem: Let q be the transition probability of an irreducible and aperiodic HMC. Suppose q is symmetric.

If $\{X_n\}_{n=0}^{\infty}$ is an HMC whose transition probability is the p of the Metropolis algorithm, then the MC is irreducible and aperiodic and its SD is π

Thus, by the Ergodic Theorem, $X_n \dot{\sim} \pi$ (for large n) so

$$\lim_{n \to \infty} \mathbb{P}(X_n = x) = \pi(x) \quad \forall x$$

Further, we can define an estimator

$$\widehat{v}_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \approx \sum_{x \in \mathfrak{X}} \pi(x) \cdot f(x) = v$$

Algorithmic Metropolis

```
Inputs: pi, q, x0
Outputs: X[0], X[1], X[2], ..., X[n]
```

```
Init X[0] <-- x0;
for n = 1, 2, 3, ...
    X' ~ q(X[n-1], ?);
    r <-- pi(X')/pi(X[n-1]);
    y ~ Bernoulli(min(1, r));
    X[n] <-- y * X' + (1 - y)*X[n-1]
end
```

Example: 2-dim Multivariate Normal Distribution

$$\begin{aligned}
\pi(x_1, x_2) &= c \cdot \exp\left(-\frac{1}{2}(x_1, x_2)\begin{pmatrix}1 & 0.8\\ 0.8 & 1\end{pmatrix}^{-1}\begin{pmatrix}x_1\\ x_2\end{pmatrix}\right)\\
&= N\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}1 & 0.8\\ 0.8 & 1\end{pmatrix}\right)\\
q(x, y) &= q((x_1, x_2), (y_1, y_2))\\
&= \left(\frac{2\pi}{25}\right)^2 \cdot \exp\left(-\frac{25}{2}\left[(x_1 - y_1)^2 + (x_2 - y_2)^2\right]\right)\\
q_{x_1}(y) &= N\begin{pmatrix}x_1\\ x_2\end{pmatrix}, \begin{pmatrix}\sigma^2 & 0\\ 0 & \sigma^2\end{pmatrix}\\
&= N\left(\begin{pmatrix}x_1\\ x_2\end{pmatrix}, \begin{pmatrix}\frac{1}{25} & 0\\ 0 & \frac{1}{25}\end{pmatrix}\right)
\end{aligned}$$

Note that this simulation is very sensitive to changes in σ. When σ is large, for example, $r = \frac{\pi(X*)}{\pi(X_{n-1})}$ tends to be small so Y tends to 0 so

$$X_n = Y \cdot X* + (1 - y)X_{n-1} = X_{n-1}$$

When σ is very small, $\text{dist}(X*, X_{n-1})$ is small so $r \to 1$, $Y \to 1$, and $X_n = X*$.

Metropolis-Hastings Algorithm (1970)

Note that sometimes, we cannot choose symmetric q so we need a different formula:

$$p(x,y) = \begin{cases} q(x,y) \cdot \min\{1, \frac{\pi(y)\cdot q(y,x)}{\pi(x)\cdot q(x,y)}\} & x \neq y,\ q(x,y) \neq 0 \\ 0 & x \neq y, q(x,y) = 0 \\ 1 - \sum_{\xi \in \mathfrak{X}: \xi \neq x} p(x,\xi) & x = y \end{cases}$$

Theorem: Let $\{X_n\}_{n=0}^{\infty}$ be an HMC whose transition prob is this p. Then, the MC is irreducible, aperiodic, and has SD π.

Lecture 17: Nov 7

Choosing q in the Metropolis Algorithm

Recall that both the Metropolis and Metropolis-Hastings algorithm require a known function $q(x,y)$ from which to sample new RVs and which is used to define p.

Now matter the choice of q, $\lim_{n\to\infty} \mathbb{P}(X_n = x) = \pi(x)$ but the rate of convergence is very strongly dependent on the choice of q.

Specifically, the convergence rates depend on the eigenvalues of

$$P = (p(x_i, x_j))_{1 \leq i,j \leq n}$$

Unfortunately, the specific relationship is outside the scope of this course.

In applications, we

1. try different q
2. for each q, generate a "long chain" $\{X_n\}_{n=0}^{N}$ for large N

If you do not want to choose q, use Gibbs Sampling instead of MCMC.

Gibbs Sampling (1984)

2-dim Gibbs Sampling ($\pi(\xi_1, \xi_2)$): Let $\pi(\xi_1, \xi_2) > 0 \quad \forall \xi_1, \xi_2$. We then define two marginal distributions

$$\pi_1(\xi_1) = \sum_{\xi_2} \pi(\xi_1, \xi_2)$$
$$\pi_2(\xi_2) = \sum_{\xi_1} \pi(\xi_1, \xi_2)$$

and condtional distributions

$$\pi_{1\,|\,2}(\xi_1 \mid \xi_2) = \frac{\pi(\xi_1, \xi_2)}{\pi_2(\xi_2)}$$
$$\pi_{2\,|\,1}(\xi_2 \mid \xi_1) = \frac{\pi(\xi_1, \xi_2)}{\pi_1(\xi_1)}$$

Further, since $\pi(x_1, \xi_2) > 0$, all four distributions will be positive.

Now let $x = (\xi_1, \xi_2)$ and $y = (\eta_1, \eta_2)$. Then

$$\begin{aligned} p(x, y) &:= \pi_{1\,|\,2}(\eta_1 \mid \xi_2) \cdot \pi_{2\,|\,1}(\eta_2 \mid \eta_1) \\ &= \frac{\pi(\eta_1, \xi_2)}{\pi_2(\xi_2)} \cdot \frac{\pi(\eta_1, \eta_2)}{\pi_1(\eta_1)} \end{aligned}$$

Explanation:

Let $\{X_n\}_{n=0}^{\infty}$ be the HMC whose transition probability is p

Claim 1: The MC is irreducible

Proof: $0 < p(x, y) = \mathbb{P}(X_1 = y \mid X_0 = x) \leq \mathbb{P}(T_y < \infty \mid X_0 = x) = \rho_{xy}$ ■

Claim 2: The MC is aperiodic.

Proof: We assume $\mathfrak{X} = \{x_1, x_2, \ldots, x_S\}$ so

$$P = (p(x_i, x_j))_{1 \leq i,j \leq S}$$

and $(P)_{ii} = (P^1)_{ii} = p(x_i, x_i) > 0$. So $1 \in I_i \implies \gcd(I_i) = 1$. ■.

Claim 3: The SD of the MC is π.

Proof: $x = (\xi_1, \xi_2)$ and $y = (\eta_1, \eta_2)$.

$$\begin{aligned}
\sum_x \pi(x) \cdot p(x,y) &= \sum_{\xi_2} \sum_{\xi_1} \pi(\xi_1, \xi_2) \cdot \frac{\pi(\eta_1, \xi_2)}{\pi_2(\xi_2)} \cdot \frac{\pi(\eta_1, \eta_2)}{\pi_1(\eta_1)} \\
&= \sum_{\xi_2} \left(\sum_{\xi_1} \pi(\xi_1, \xi_2) \right) \cdot \frac{\pi(\eta_1, \xi_2)}{\pi_2(\xi_2)} \cdot \frac{\pi(\eta_1, \eta_2)}{\pi_1(\eta_1)} \\
&= \sum_{\xi_2} \pi_2(\xi_2) \cdot \frac{\pi(\eta_1, \xi_2)}{\pi_2(\xi_2)} \cdot \frac{\pi(\eta_1, \eta_2)}{\pi_1(\eta_1)} \\
&= \sum_{\xi_2} \pi(\eta_1, \xi_2) \cdot \frac{\pi(\eta_1, \eta_2)}{\pi_1(\eta_1)} \\
&= \pi_1(\eta_1) \cdot \frac{\pi(\eta_1, \eta_2)}{\pi_1(\eta_1)} \\
&= \pi(\eta_1, \eta_2) \\
&= \pi(y)
\end{aligned}$$

Lecture 18: Nov 9

Notation

Let $\pi(\xi_1, \xi_2, \ldots, \xi_d)$ be a PMF on d-space.

We can create a marginal distribution,

$$\pi_{-i}(\xi_1, \ldots, x_{i-1}, \xi_{i+1}, \ldots, \xi_d) = \sum_{\xi_i} \pi(\xi_1, \ldots, \xi_i, \ldots, \xi_d)$$

And using the joint distribution and marginal distribution, we have a conditional distribution

$$\pi_{i|-i}(\xi_i \mid \xi_1, \ldots, \xi_{i-1}, \xi_{i+1}, \ldots, \xi_d)$$

Remark: this is a very natural extension of what we did in the 2-d Gibbs Sampling case.

2-dim Gibbs Sampler

Inputs: a given distribution $\pi(\xi_1, \xi_2)$ and an initial state $x_0 = (\xi_1, \xi_2)$

Output: a HMC $\{X_n = (\xi_1^{(n)}, \xi_2^{(n)})\}_{n=0}^{\infty}$

Algorithm:

```
for n = 1, 2, ...
```

Generate $\xi_1^{(n)} \sim \pi_{1\,|\,2}(\circ \mid \xi_2^{(n-1)})$

Generate $\xi_2^{(n)} \sim \pi_{2\,|\,1}(\circ \xi_1^{(n-1)} \mid)$

$X_n = (\xi_1^{(n)}, \xi_2^{(2)})$

```
end for
```

d-Dimensional Gibbs Sampler

Inputs: Given $\pi(\xi_1, \xi_2, \ldots, \xi_d)$, and $x_0 = (\xi_1^{(0)}, \xi_2^{(0)}, \ldots, \xi_d^{(0)})$

Outputs: a HMC $\{X_n = (\xi_1^{(n)}, \ldots, xi_d^{(n)})\}_{n=0}^{\infty}$

Algorithm:

```
X0 <-- x0
for n = 1, 2, 3, ...
    Generate ξ_1^(n) ~ π_{1|-1}(∘ | ξ_2^(n-1), ..., ξ_d^(n-1))

    for j = 2, 3, ..., d
        Generate  ξ_j^(n) ~ π_{j|-j}(∘ | xi_1^(n), ..., ξ_{j-1}^(n), ξ_{j+1}^(n-1), ..., ξ_d^(n-1))
    end for

    x_n ⟵ (ξ_1^(n), ..., ξ_d^(n))
end for
```

Problem: Generating

$$\begin{aligned}\xi_j \sim \pi_{j\,|\,-j}(\xi_j \mid \xi_1, \ldots, \xi_{j-1}, \xi_{j+1}, \ldots, \xi_d) \\ = \frac{\pi(\xi_1, \ldots, \xi_{j-1}, \xi_j, \xi_{j+1}, \ldots, \xi_d)}{\sum_{\xi_j} \pi(\xi_1, \ldots, \xi_{j-1}, \xi_j, \xi_{j+1}, \ldots, \xi_d)}\end{aligned}$$

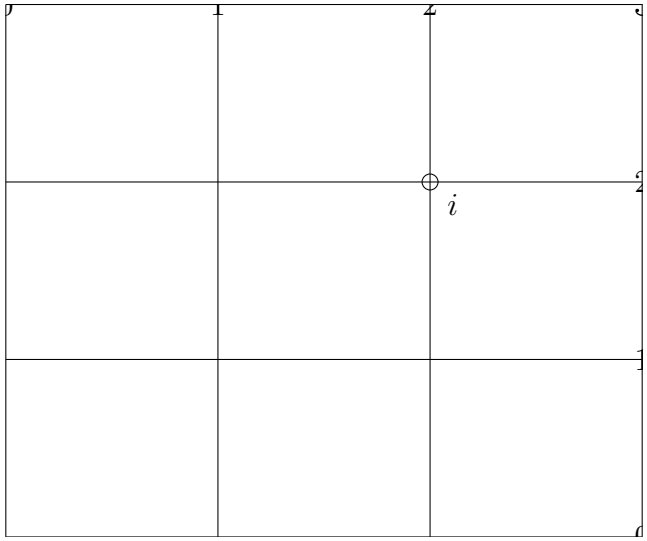
is generally infeasible.

This method only works if π is a Gibbs Random Field.

Ising Model

The most famous GRF is the Ising Model of statistical mechanics.

We construct an n-by-n lattice where each vertex has a magnetic monopole. $s_i \in \{-1, 1\}$ where -1 corresponds to a South pole and $+1$ to a North pole.



The *magnetic structure* of the lattice is

$$\vec{s} = (s_1, s_2, \ldots, s_{N^2})$$

$$f(\vec{s}) = \frac{1}{N^2} \sum_i s_i$$

- If $\left| f(\vec{s}) \right| = 0$ the lattice is not magnetic because the north and south poles cancel.
- If $\left| f(\vec{s}) \right| > 0$ the lattice is magnetic

Problem: What if $\vec{s}$ is not deterministic. i.e., $\vec{s} = \pi(s_1, \ldots, s_{N^2})$?

Each $\vec{s}$ is associated with an energy value

$$H_N(\vec{s}) = -\sum_{\langle i,j \rangle} s_i s_j$$

- If $s_i = s_j \implies s_i s_j = 1 \implies$ low energy
- If $s_i \neq s_j \implies s_i s_j = -1 \implies$ high energy

The Ising Model:

$$\pi_{N,\beta}(\vec{s}) = \frac{1}{Z(\beta)} \exp(-\beta \cdot H_N(\vec{s}))$$

where $\beta = 1/T$ (the temperature) and $Z(\beta)$ is the partition function

$$Z(\beta) = \sum_{\vec{s}} \exp\{-\beta H_N(\vec{s})\}$$

and

$$\mathbb{E}_{\pi_{N,\beta}}((f(\vec{s}))) = \sum_{\vec{s}} \pi_{N,\beta}(\vec{s}) \cdot \underbrace{\left| \frac{1}{N^2} \sum_i s_i \right|}_{m_N(\beta)}$$

In most applications, each vertex denotes an atom and $N \approx \infty$ so

$$m(\beta) = \lim_{N \to \infty} m_N(\beta)$$

- If $m(\beta) > 0$, the "infinitely large" lattice is magnetic
- If $m(\beta) = 0$, the lattice is not magnetic

Note that β is defined by a temperature. When $\beta = \beta^*$, the *Curie Temperature*, t certain materials begin to lose their permanent magnetic properties.

$$\beta^* = \frac{\log(1 + \sqrt{2})}{2}$$

Lecture 19: Nov 14

Ising Model

We have an N-by-N lattice with N^2 vertices. Each vertex i is associated with a binary $s_i \in \{-1, 1\}$. Let $\vec{s} = (s_1, s_2, \ldots, s_{N^2})$.

Then $\mathfrak{X}$ is the collection of all possible vectors $\vec{s}$:

$$\mathfrak{X} = \underbrace{\{-1,1\} \times \{-1,1\} \times \cdots \times \{-1,1\}}_{N^2 \text{ times}} \implies \#\mathfrak{X} = 2^{N^2}$$

Let $\pi : \mathfrak{X} \to \mathbb{R}$ be a PMF on $\mathfrak{X}$ expressed by

$$\pi_{N,\beta}(\vec{s}) = \frac{1}{Z(\beta)} \exp(-\beta \cdot H_N(\vec{s}))$$

where

$$H_N(\vec{s}) = -\sum_{\langle i,j \rangle} s_i s_j$$

is an energy function (the bracket notation indicates i, j are neighbours i.e., there is an edge between them), $\beta = \frac{1}{T}$ where T is the temperature, and $(\beta > 0)$

$$Z(\beta) = \sum_{\vec{s}} \exp(-\beta \cdot H_N(\vec{s}))$$

Let $f : \mathfrak{X} \to \mathbb{R}$ be a different function on the state space giving the average value of entries in a vector:

$$f(\vec{s}) = \frac{1}{N^2} \sum_{i=1}^{N^2} s_i$$

Most often in applications, we are interested in the characteristics of the full state space – not any particular vector. So we take

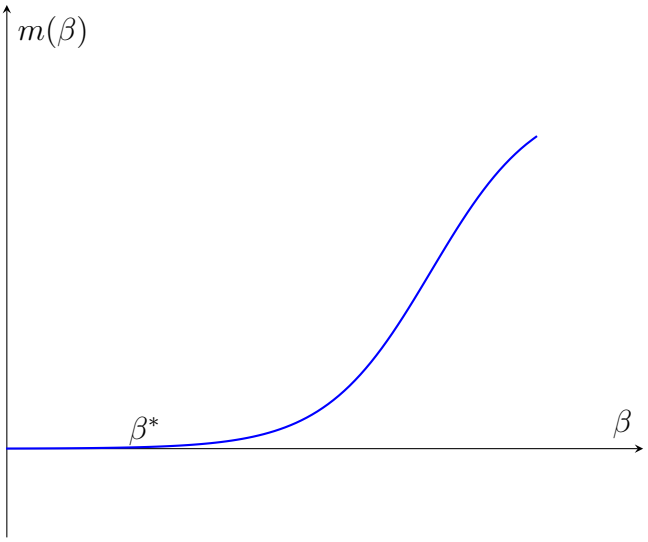
$$m_N(\beta) = \sum_{\vec{s} \in \mathfrak{X}} \pi_{N,\beta}(\vec{s}) \cdot |f(\vec{s})| = \mathbb{E}_{\pi_{N,\beta}} |f(\vec{s})|$$

Further, in physics we generally care about a huge amount of particles:

$$m(\beta) = \lim_{N \to \infty} m_N(\beta)$$

(this is the *magnetization* of the lattice)

Remark: $m(\beta)$ is a function of β and is called the *magnetization curve*. It looks like:



Theoretically,

$$\beta^* = \frac{\log(1+\sqrt{2})}{2}$$

but the calculation of this value was worthy of a Nobel Prize.

We will approximate β^*. To get β^*, we need $m(\beta)$. When N is very large,

$$m(\beta) \approx m_N(\beta) = \sum_{\vec{s}} \pi_{N,\beta} \cdot |f(\vec{s})| = \mathbb{E}_{\pi_{N,\beta}} |f(\vec{s})|$$

Using the MCMC

Suppose we have an HMC $\{X_n\}_{n=0}^{\infty}$ on the state space $\mathfrak{X}$ which is irreducible, aperiodic, and its SD is π.

By the Ergodic Theorem,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left|f(X^{(i)})\right| = \mathbb{E}_{\pi_{N,\beta}} \left|f(X^{(i)})\right| = m_N(\beta)$$

So when both N and n are large,

$$m(\beta) \approx m_N(\beta) \approx \frac{1}{n} \sum_{i=1}^{n} \left|f(X^{(i)})\right|$$

Thus, with only a small reduction in accuracy do to the approximation, we have reduced a Nobel-Prize level question into a simple high-school level average.

Problem: How do we construct an HMC $\{X_n\}_{n=0}^{\infty}$ with SD π?

Solution: Gibbs Sampling

The Conditional Distribution

The central principle of Gibbs sampling is drawing for a specific conditional distribution, $\pi_{i \mid -i}$.

Note on notation: Let $\pi = \pi_{N,\beta}$.

Then we have a PMF $\pi(s_1, \ldots, s_j, \ldots, s_{N^2})$ and we can define a conditional distribution

$$\pi_{j \mid -j}(s_j \mid s_1, \ldots, s_{j-1}, s_{j+1}, \ldots, s_{N^2}) = \frac{\pi(s_1, \ldots, s_j, \ldots, s_{N^2})}{\sum_{s_j} \pi(s_1, \ldots, s_j, \ldots, s_{N^2})}$$

Since we are applying this specifically to the Ising Model, we have an expression for π we can use to simplify the above. Focus on a single vertex j. Then from the Ising model,

$$H_N(\vec{s}) = -\sum_{\langle i,j \rangle} s_i s_j \implies H_N(\vec{s}) = -s_j s_{j,l} - s_j s_{j,r} - s_j s_{j,u} - s_j s_{j,d} - \sum_{\langle k,l \rangle : k,l \neq j} s_{kl}$$

We define $\mathcal{N}(j) = \{jl, jr, ju, jd\}$ to be the (neighborhood of j) so

$$H_N(\vec{s}) = -\sum_{j' \in \mathcal{N}(j)} s_j s_{j'} - \sum_{\langle k,l \rangle : k,l \neq j} s_k s_l$$

This gives a new expression for π:

$$\pi(\vec{s}) = \frac{1}{Z(\beta)} \cdot \exp(\beta \cdot H_N(\vec{s})) = \frac{1}{Z(\beta)} \cdot \exp\left(-\beta \sum_{j' \in \mathcal{N}(j)} s_j s_{j'}\right) \cdot \exp\left(-\beta \sum_{\langle k,l \rangle : k,l \neq j} s_k s_l\right)$$

which we can plug in to the conditional distribution:

$$\begin{aligned}
\pi_{j\mid -j}(s_j \mid s_1, \ldots,\ s_{j-1}, s_{j+1}, \ldots,\ s_{N^2}) &= \frac{\pi(s_1, \ldots,\ s_j, \ldots,\ s_{N^2})}{\sum_{s_j} \pi(s_1, \ldots,\ s_j, \ldots,\ s_{N^2})} \\
&= \frac{\frac{1}{Z(\beta)} \cdot \exp\left(-\beta \sum_{j' \in \mathcal{N}(j)} s_j s_{j'}\right) \cdot \exp\left(-\beta \sum_{\langle k,l\rangle : k,l \neq j} s_k s_l\right)}{\sum_{s_j} \frac{1}{Z(\beta)} \cdot \exp\left(-\beta \sum_{j' \in \mathcal{N}(j)} s_j s_{j'}\right) \cdot \exp\left(-\beta \sum_{\langle k,l\rangle : k,l \neq j} s_k s_l\right)} \\
&= \frac{\exp\left(-\beta \sum_{j' \in \mathcal{N}(j)} s_j s_{j'}\right)}{\sum_{s_j \in \{-1,1\}} \exp\left(-\beta \sum_{j' \in \mathcal{N}(j)} s_j s_{j'}\right)} \\
&= \frac{\exp\left(-\beta \sum_{j' \in \mathcal{N}(j)} s_j s_{j'}\right)}{\exp\left(\beta \sum_{j' \in \mathcal{N}(j)} s_{j'}\right) + \exp\left(-\beta \sum_{j' \in \mathcal{N}(j)} s_{j'}\right)}
\end{aligned}$$

For simplicity, let

$$\begin{aligned}
a &= \exp\left(\beta \sum_{j' \in \mathcal{N}(j)} s_{j'}\right) \\
b &= \exp\left(-\beta \sum_{j' \in \mathcal{N}(j)} s_{j'}\right)
\end{aligned}$$

Then we have two cases for π:

$$\begin{aligned}
\pi_{j|-j}(1 \mid \ldots) &= \frac{b}{a+b} \\
\pi_{j|-j}(-1 \mid \ldots) &= \frac{a}{a+b}
\end{aligned}$$

i.e.,

$$\begin{aligned}
\mathbb{P}(s_j = -1) &= \frac{a}{a+b} \\
\mathbb{P}(s_j = 1) &= \frac{b}{a+b}
\end{aligned}$$

Using Gibbs

Now we want to generate $X^{(n)} \sim \pi_{j|-j}(? \mid s_1, \ldots, s_{j-1}, s_{j+1}, \ldots, s_{N^2})$.

1. Compute
$$a = \exp\left(\beta \sum_{j' \in \mathcal{N}(j)} s_{j'}\right)$$
2. Compute
$$b = \exp\left(-\beta \sum_{j' \in \mathcal{N}(j)} s_{j'}\right)$$
3. Compute
$$p = \frac{b}{a+b}$$
4. Generate
$$Z \sim \text{Bernoulli}(p)$$
5. Transform $Z \in \{0,1\}$ to $s_j \in \{-1,1\}$:
$$s_j = 2Z - 1$$

Then generate $\{X^{(n)} = (X_1^{(n)}, X_2^{(n)}, \ldots, X_{N^2}^{(n)})\}_{n=1}^{\infty}$.

By the Ergodic Theorem, $X^{(n)} \dot{\sim} \pi_{N,\beta}$

Lecture 20: Nov 16

Graphs

Definition: A graph is an ordered pair $G = (V, E)$ where

1. V is the collection of vertices
2. E is the collection of (undirected) edges

Remarks:

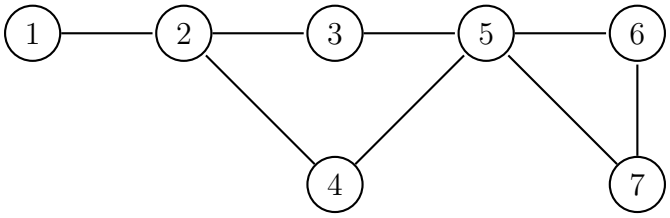
- An edge connecting vertices i and j is denoted (i, j)

- We are focused on undirected graphs so $(i, j) = (j, i)$
- For each edge $(i, j) \in E$, we require $i \neq j$

Definition: Let $G = (V, E)$ be a graph.

1. two vertices i and j are *adjacent* if $(i, j) \in E$.
2. Let $i \in V$. The *neighborhood* of i is
$$\mathcal{N}(i) = \{j \in V : (i, j) \in E\}$$

Example:



This picture shows a graph G. $\mathcal{N}(5) = \{3, 4, 6, 7\}$.

Definition: Let $G = (V, E)$ be a graph.

1. A subset $c \subseteq V$ is called a *clique* if any pair of vertices in c are adjacent.
2. Let $\mathcal{C}(G)$ be the collection of all cliques in G.

Convention: for each $i \in V$, the singleton $\{i\}$ is a clique

In the graph above, the cliques are:

- $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}$
- $\{1, 2\}, \{2, 3\}, \{3, 5\}, \{5, 6\}, \{4, 2\}, \{4, 5\}, \{7, 5\}, \{7, 6\}$
- $\{5, 6, 7\}$

Notation:

- Let $G = (V, E)$ be a graph with $V = \{1, \ldots, d\}$ and $x = (x_1, \ldots, x_d)^T$ a vector in the product space $\mathfrak{X} = \mathfrak{X}_1 \times \cdots \times \mathfrak{X}_d$
- For any subset $c = \{i_1, \ldots, i_k\} \subseteq V$ with $i_1 < i_2 < \cdots < i_k$ we denote
$$x_c := (x_{i_1}, \ldots, x_{i_k})^T$$
and $\mathfrak{X}_c := \mathfrak{X}_{i_1} \times \cdots \times \mathfrak{X}_{i_k}$

Example: Let $V = \{1, \ldots, 7\}$ and $c = \{1, 3, 5\}$. Then $x_c = (x_1, x_3, x_5)^T$ and $\mathfrak{X} = \mathfrak{X}_1 \times \mathfrak{X}_3 \times \mathfrak{X}_5$.

Gibbs Random Fields (GRFs)

Definition: LEt $\pi(x_1, \ldots, x_n)$ be a d-variable PMF and $G = (V, E)$ be a graph with vertices $V = \{1, \ldots, d\}$. If there are functions $\phi_c(x_c)$ for all $c \in \mathcal{C}(G)$ such that

$$\pi(x_1, \ldots, x_d) = \frac{1}{Z} \prod_{c \in \mathcal{C}(G)} \phi_c(x_c)$$

with

$$Z = \sum_x \prod_{c \in \mathcal{C}(G)} \phi_c(x_c)$$

then π is a GRF.

Remarks:

1. A factor $\phi_c(x_c)$ is allowed to be a constant function, $\phi_c(x_c) = 1$
2. The factorization is not unique
3. To prove that π is a GRF, you only need to find one such factorization.

Example: If $\pi(x_1, x_2) = \frac{1}{Z} x_1 x_2 \qquad x_1, x_2 \in \{1, 2\}$. Then $\mathcal{C}(G) = \{\{1\}, \{2\}, \{1, 2\}\}$. We have two approaches:

Approach 1: $\phi_{\{1\}}(x_1) = x_1, \quad \phi_{\{2\}}(x_2) = x_2, \quad \phi_{\{1,2\}}(x_1, x_2) = 1$. So

$$\pi(x_1, x_2) = \frac{1}{Z} \phi_{\{1\}}(x_1) \phi_{\{2\}}(x_2) \phi_{\{1,2\}}(x_1, x_2) = \frac{1}{Z} \prod_{c \in \mathcal{C}(G)} \phi_c(x_c)$$

Approch 2: $\phi_{\{1\}}(x_1) = 1690 x_1, \quad \phi_{\{2\}}(x_2) = \frac{1}{1690} x_2, \quad \phi_{\{1,2\}}(x_1, x_2) = 1$. So

$$\pi(x_1, x_2) = \frac{1}{Z} \prod_{c \in \mathcal{C}(G)} \phi_c(x_c)$$

Approach 3: $\phi_{\{1\}}(x_1) = \phi_{\{2\}}(x_2) = 1, \quad \phi_{\{1,2\}}(x_1, x_2) = x_1 x_2$. So

$$\pi(x_1, x_2) = \frac{1}{Z} \prod_{c \in \mathcal{C}(G)} \phi_c(x_c)$$

Example: Let G be an N-by-N lattice. Let

$$\pi(x_1, \ldots, x_{N^2}) = \frac{1}{Z(\beta)} \exp\left(\beta \sum_{(i,j)\in G} x_i x_j\right)$$

Equivalently,

$$\pi(x_1, \ldots, x_{N^2}) = \frac{1}{Z(\beta)} \prod_{\langle i,j\rangle} \exp(\beta x_i x_j)$$

For all i, j, the set $\{i, j\} \in \mathcal{C}(G)$ if i and j are neighbors. So

$$\phi_{\{i,j\}}(x_i, x_j) := \exp(\beta \cdot x_i x_j)$$

For all other cliques, we define

$$\phi_c(x_c) = 1$$

So

$$\frac{1}{Z(\beta)} \prod_{c\in\mathcal{C}(G)} \phi_c(x_c) = \frac{1}{Z(\beta)} \prod_{\langle i,j\rangle} \phi_{\{i,j\}}(x_i, x_j)$$

so the Ising Model is a GRF.

Application of Gibbs Sampling to GRFs

For ease of notation, let $x_{-j} = (x_1, \ldots,\ x_{j-1}, x_{j+1}, \ldots,\ x_j)$. So we define a conditional distribution $\pi_{j\,|\,-j}(x_j \mid x_{-j})$.

Now say

$$\pi(x_1, \ldots,\ x_d) = \frac{1}{Z}\left(\prod_{c\in\mathcal{C}(G),\, j\neq c} \phi_c(x_c)\right)\left(\prod_{c\in\mathcal{C}(G),\, j\in c} \phi_c(x_c)\right)$$

is a Gibbs Random Field from which we want to derive a conditional distribution:

$$\begin{aligned}
\pi_{j\,|\,-j}(x_j \mid x_{-j}) &= \frac{\pi(x_1, \ldots,\ x_d)}{\sum_{x_i} \pi(x_1, \ldots,\ x_d)} \\
&= \frac{\frac{1}{Z}\left(\prod_{c'\in\mathcal{C}(G),\, j\neq c'} \phi_{c'}(x_{c'})\right)\left(\prod_{c\in\mathcal{C}(G),\, j\in c} \phi_c(x_c)\right)}{\sum_{x_j} \frac{1}{Z}\left(\prod_{c'\in\mathcal{C}(G),\, j\neq c'} \phi_{c'}(x_{c'})\right)\left(\prod_{c\in\mathcal{C}(G),\, j\in c} \phi_c(x_c)\right)} \\
&= \frac{\prod_{c\in\mathcal{C}(G), j\in c} \phi_c(x_c)}{\sum_{x_j} \prod_{c\in\mathcal{C}(G), j\in c} \phi_c(x_c)}
\end{aligned}$$

So $\pi_{j\,|\,-1}(x_j \mid x_{-j})$ depends only on

$$\bigcup_{c \in \mathcal{C}(G): j \in c} c \overset{HW9}{=} \mathcal{N}(j) \cup j$$

Thus,

$$\pi_{j\,|\,-1}(x_j \mid x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d) = \pi_{j\,|\,-1}(x_j \mid x_{\mathcal{N}(j)})$$

("given the values of j's neighbors, the value of vertex j does not depend on vertices outside the neighborhood")

Lecture 21: Nov 21

Markov Random Fields (MRFs)

Definition: Let $\pi(x_1, \ldots, x_d)$ be a d-variable PMF and $G = (V, E)$ be a graph with $V = \{1, 2, \ldots, d\}$. If for every vertex $j \in V$, $\pi_{j\,|\,-1}(x_j \mid x_{-j})$ does not depend on the vertices outside $\mathcal{N}(j) \cup \{j\}$, then π is a Markov Random Field (MRF).

Hammersley-Clifford Theorem (1971): Let π be a d-variable PMF and $G = (V, E)$ be a graph with $V = \{1, 2, \ldots, d\}$. Then π is a GRF of G if and only if π is a MRF of G

Remark: this tells us that MRFs and GRFs are the same

From above, $\pi_{j\,|\,-1}(x_j \mid x_{-j})$ depends only on $\mathcal{N}(j) \cup j$. Thus, Gibbs sampling is efficient if $\max_{j \in V} \#\mathcal{N}(j)$ is small.

Example: for the Ising model, $\max_{j \in V} \#\mathcal{N}(j) = \#\mathcal{N}(j) = 4$

Proof of the Ergodic Theorem

Let $\{X_n\}_{n=0}^{\infty}$ be an HMC with state space $\mathfrak{X} = \{x_1, x_2, \ldots, x_S\}$. It has S-by-S transition matrix $P = (p(x_i, x_j))_{1 \le i,k \le S}$.

P^T has eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_S$ and eigenvectors $v_1, \ldots, v_S$.

Suppose the MC is irreducible and aperiodic. Further, it satisfies (the nonessential but convenient conditions):

- The eigenvalues are real

- $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_S$
- $\lambda_1 = 1$
- $\lambda_S > -1$

Now let

$$\mu = \begin{pmatrix} \mathbb{P}(X_0 = x_1) \\ \mathbb{P}(X_0 = x_2) \\ \vdots \\ \mathbb{P}(X_0 = x_S) \end{pmatrix} \in \mathbb{R}^S \implies \mu = \sum_{i=1}^{S} \alpha_i v_i$$

(Recall that $((P^T)^n \mu)_i = \mathbb{P}(X_n = i)$) so

$$P^T \mu = \sum_{i=1}^{S} \alpha_i P^T v_i = \sum_{i=1}^{S} \alpha_i \lambda_i^n v_i \implies \lim_{n\to\infty} (P^T)^n \mu = \sum_{i=1}^{S} \alpha \left(\lim_{n\to\infty} \lambda_i^n \right) v_i$$

But because $|\lambda_i| < 1$,

$$\lim_{n\to\infty} \lambda_i^n = \begin{cases} 1 & i = 1 \\ 0 & i \geq 2 \end{cases}$$

So

$$\lim_{n\to\infty} (P^T)^n \mu = \alpha_1 v_1 = \pi$$

where π is the solution to $\pi = P^T \pi$.

Convergence Rates of the MCMC

Let $x, y \in \mathbb{R}^S$. We define

$$||x - y|| := \sqrt{\sum_{i=1}^{S} (x_i - y_i)^2}$$

Then using the Triangle Inequality,

$$\left|\left|(P^T)^n\mu - \pi\right|\right| = \left|\left|\alpha_1 v_1 + \sum_{i=2}^{S} \alpha_i \lambda_i^n v_i - \alpha_1 v_1\right|\right| \tag{1}$$

$$= \left|\left|\sum_{i=2}^{S} \alpha_i \lambda_i^n v_i\right|\right| \tag{2}$$

$$\leq \sum_{i=2}^{S} |\alpha_i| \, |\lambda_i^n| \, ||v_i|| \tag{3}$$

$$\leq \sum_{i=2}^{S} |\alpha_i| \, ||v_i|| \cdot \max_i\{|\lambda_i|^n\} \tag{4}$$

$$\tag{5}$$

But since $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_S$,

$$\sum_{i=2}^{S} |\alpha_i| \, ||v_i|| \cdot \max_i\{|\lambda_i|^n\} = \sum_{i=2}^{S} |\alpha_i| \, ||v_i|| \cdot \max_i\{|\lambda_2|^n, |\lambda_S|^n\}$$

Since the left factor does not depend on n, we can call it a constant C. Thus,

$$\left|\left|(P^T)^n\mu - \pi\right|\right| \leq C \cdot \max_i\{|\lambda_2|^n, |\lambda_S|^n\}$$

However, in applications, the size of P^T can be very large. So we want to find a way to bound the convergence rate of the MCMC without having to compute the eigenvalues of P^T. For example, in the 100-by-100 Ising model, $S = 2^{10000}$, so P is a 2^{10000}-by-2^{10000} matrix.

Lecture 22: Nov 28

Dimension Reduction

A toy example:

$$Z^{(i)} = \begin{pmatrix} Z_1^{(i)} \\ Z_2^{(i)} \end{pmatrix}, \quad i = 1, 2, \ldots, 10000$$

with

$$Z_1^{(1)}, Z_1^(2), \ldots, Z_1^{(n)}, Z_2^{(1)}, Z_2^{(2)}, \ldots \mathbb{Z}_2^{(n)} \overset{iid}{\sim} N(0, \frac{1}{4})$$

so $Z^{(1)}, \ldots, Z^{(n)}$ are 2-dim data points.

Let $g : \mathbb{R}^2 \to \mathbb{R}^3$ defined by $g(\xi_z, \xi_2) = (\xi_1, \xi_2, \xi_1^2 + \xi_2^2)$ such that the image of g is a paraboloid.

Now we define a sequence $X^{(1)}, X^{(2)}, \ldots, X^{(10000)} \in \mathbb{R}^3$ in the image of g by

$$X^{(i)} = g(Z^{(i)}) \in \mathbb{R}^3$$

Are they 2-dim or 3-dim?

- If we want them to be 3-d, this makes sense because they are in $\mathbb{R}^3$ but they live only on the 2d surface
- If we want them to be 2-d, this also makes sense because they are on a 2-d surface but they live in $\mathbb{R}^3$

More generally, if we have a random vector $Z^{(i)} \in \mathbb{R}^d$, a function $g : \mathbb{R}^d \to \mathbb{R}^D$ and a transformed random vector $X^{(i)} = g(Z^{(i)}) \in \mathbb{R}^D$ (with $d < D$) we call d the *intrinsic dimension* and D the *extrinsic dimension*.

In most applications, only $X^{(i)}$ is observed and $Z^{(i)}$ is hidden. Worse, there is usually a high-dimensional noise term:

$$\underbrace{X^{(i)}}_{\text{D-dim (observed)}} = \underbrace{g(Z^{(i)})}_{\text{underlying structure}} + \underbrace{\varepsilon^i}_{\text{D-dim noise}}$$

We are interested in the image of g,

$$M = \{g(z) : z \in \mathbb{R}^d\} \subset \mathbb{R}^D$$

(M is a manifold)

Our ultimate goal is to learn M.

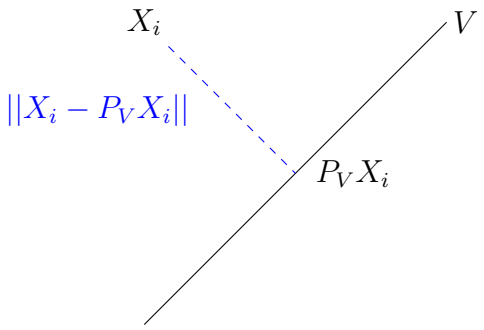
Two Branches of Dimension Reduction

1. Linear DR (*Assumption: M* has zero-curvature)
    - e.g. Principal Component Analysis (Pearson, 1901)

2. Nonlinear DR/Manifold Learning (*Assumption: M may have curvature*)
    - e.g. Principal Curves (Hastie, 1984), Isomap (2000), Local-Linear Embedding (2000), Principal Manifolds (2001, 2021)

Principal Component Analysis (PCA)

We ket V be a hyperplane and $X_i \in \mathbb{R}^d$. We consider the projection of the X_i onto V:



We want to fit/represent X_i by $P_v X_i$ because$D$-dim space is complex but a d-dim hyperplane is simpler.

Fitting error: $||X_i - P_V X_i||^2$ where

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} p \qquad \text{(D-dim distribution)}$$

Average fitting error:

$$\frac{1}{n}\sum_{i=1}^{n} ||X_i - P_V X_i||^2 \approx \mathbb{E}(||X - P_V X||^2)$$

Thus, the *optimal hyuperplane* V^* is the hyperplace that minimizes the average fitting error.

$$V^* := \arg\min_V \{\mathbb{E}(||X_i - P_V X_i||^2)\}$$

The essence of PCA is to make the assumption that $M = V^*$

Question: How do we compute V^*?

Some Preparation: Let X be an n-dim RV.

Covariance: $\text{Cov}(X_i, X_j) = \mathbb{E}[X_i X_j] - (\mathbb{E}X_i)(\mathbb{E}X_j)$

We can create a matrix

$$\mathbb{V}(X) = \begin{pmatrix} \text{Cov}(X_1, X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Cov}(X_2, X_2) & \dots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \dots & \text{Cov}(X_n, X_n) \end{pmatrix}$$

which has some very nice properties:

- $\mathbb{V}(X)$ is symmetric ($\mathbb{V}(X) = (\mathbb{V}(X))^T$)
- $\mathbb{V}(X)$ is positive semi-definite (*Proof:* use the fact that for an $m \times n$ matrix A, $\mathbb{V}(AX) = A \cdot \mathbb{V}(X) \cdot A^T$)
- Since $\mathbb{V}(X)$ is positive semi-definite it has real eigenvectors $v_1, \dots, v_d$ and eigenvalues which we can arrange in decreasing order ($\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_D \geq 0$)

Answer:

$$V^* = \arg\min_V \left\{ \mathbb{E}(||X_i - P_V X_i||^2) \right\} = \text{Span}\{v_1, v_2, \dots, v_d\} = \left\{ \sum_{i=1}^{d} \alpha_i v_i \middle| \alpha_i, \alpha_2, \dots, \alpha_d \in \mathbb{R} \right\}$$

We call v_i the i-th *principal component*

Remark: part of this requires guessing/suspecting a value of d. We can apply a condition, such as if

$$\frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{D} \lambda_i} > 95\%$$

then we believe the intrinsic dimension is d.

Lecture 23: Nov 30

Linear Manifold Learning (Principal Component Analysis)

Let d be the intrinsic dimension and D be the extrinsic dimension with $d < D$. D is always known. In this class, we assume d is known too.

Let $X_1, X_2, \dots, X_n \overset{iid}{\sim}$ a D-dim distribution so each of the random vectors is in $\mathbb{R}^D$.

For simplicity, we assume that the random vectors have been *centralized* so

$$\mathbb{E}X_i = (0, 0, \ldots, 0) = \vec{0} \qquad \forall i = 1, 2, \ldots, \ n$$

(This is reasonable because if they are not centralized, we may just update each coordinate by $X_i \to X_i - \frac{1}{n}\sum_{k=1}^{n} X_k$ since they are iid.)

We want to project X_i into a hyperplane $V \in \mathbb{R}^d$ and we want to find the optimal hyperplane V^* that minimizes the average fitting error:

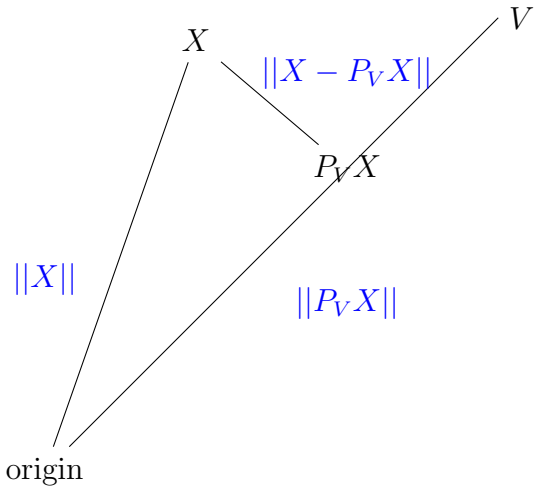
$$\frac{1}{n}\sum_{i=1}^{n} ||X_i - P_V X_i||^2 \approx \mathbb{E}[||X - P_V X_i||^2]$$

where X is a D-dim RV which shares the same distribution as the X_i.

i.e.,

$$V^* = \arg\min_V \mathbb{E}[||X - P_V X_i||^2]$$

PCA by Variance Maximization



By the Pythagorean Theorem,

$$||X||^2 = ||P_V X||^2 + ||X - P_V X||^2$$

and

$$\min_V \mathbb{E}[||X - P_V X||^2] = \mathbb{E}(||X||^2) - \max_V \underbrace{\mathbb{E}(||P_V X||^2)}_{\text{variance}}$$

so

$$V^* = \arg\min_V \mathbb{E}(||X - P_V X||^2) = \arg\max_V \mathbb{E}(||P_V X||^2)$$

Now let $\mathbb{V}(X)$ be the D-by-D (symmetric and semidefinite) covariance matrix of X with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq \cdots \geq \lambda_D \geq 0$ and eigenvectors $v_1, v_2, \ldots, v_D$.

This means that we can write

$$V^* = \text{Span}\{v_1, v_2, \ldots, v_d\} = \left\{ \sum_{k=1}^{d} \alpha_k v_k \middle| \alpha_1, \ldots, \alpha_d \in \mathbb{R} \right\}$$

Choosing d

From HW 10,

$$\mathbb{E}(||P_V^* X||^2) = \sum_{k=1}^{d} \lambda_k$$

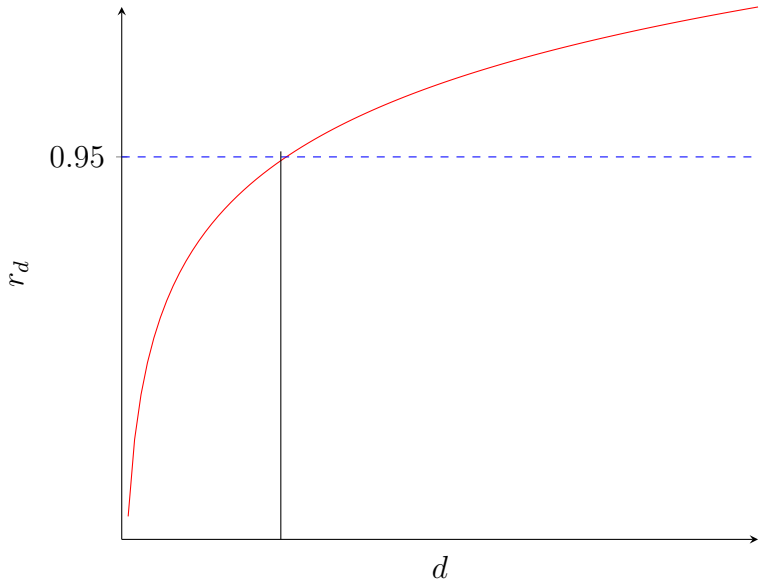
$$\mathbb{E}(\left|\left|X^2\right|\right|) = \sum_{k=1}^{D} \lambda_k$$

$$r_d = \frac{\mathbb{E}(||P_V^* X||^2)}{\mathbb{E}(||X^2||)} = \frac{\sum_{k=1}^{d} \lambda_k}{\sum_{k=1}^{D} \lambda_k}$$

where r_d is the proportion of variance preserved by V^*. This allows us to set some criterion. Usually, we want

1. $r_d \geq 95\%$
2. $r_{d-1} < 95\%$

So



Further Interpretation of r_d

In the generic formulation of manifold learning, we have $g : \mathbb{R}^d \to \mathbb{R}^D$ and

$$X = g(Z) + \varepsilon$$

i.e., there is an independent D-dim noise term.

In PCA, it is practically the same except for the fact that g is a linear function (a matrix):

$$X = AZ + \varepsilon$$

where $A \in \mathbb{R}^{D \times d}$, $X \in \mathbb{R}^D$, $Z \in \mathbb{R}^d$, and $\varepsilon \in \mathbb{R}^D$.

So

$$\begin{aligned}\mathbb{V}(X) &= \mathbb{V}(LZ) + \mathbb{V}(\varepsilon) \\ &= L\mathbb{V}(Z)L^T + \mathbb{V}(\varepsilon) \\ \text{rank}(L\mathbb{V}(Z)L^T) &\leq \text{rank}\mathbb{V}(Z) \leq d \\ \implies L\mathbb{V}(Z)L^T &= U \begin{pmatrix} \tilde{\lambda}_1 & & & & & \\ & \ddots & & & & \\ & & \tilde{\lambda}_d & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix} U^T\end{aligned}$$

where $UU^T = I$

Further, we assume

$$\mathbb{V}(\varepsilon) = \begin{pmatrix} \sigma^2 & & & \\ & \sigma^2 & & \\ & & \ddots & \\ & & & \sigma^2 \end{pmatrix} = \sigma^2 I = U(\sigma^2 I)U^T$$

$$\implies \mathbb{V}(X) = U \begin{pmatrix} \tilde{\lambda}_1 + \sigma^2 & & & & & \\ & \ddots & & & & \\ & & \tilde{\lambda}_d + \sigma^2 & & & \\ & & & \sigma^2 & & \\ & & & & \ddots & \\ & & & & & \sigma^2 \end{pmatrix}$$

Since the eigenvalue of $\mathbb{V}(X)$ are $\lambda_1 \geq \cdots \geq \lambda_D$,

$$\begin{aligned}\lambda_1 &= \tilde{\lambda}_1 + \sigma^2 \\ &\vdots \\ \lambda_d &= \tilde{\lambda}_d + \sigma^2 \\ \lambda_{d+1} &= \sigma^2 \\ &\vdots \\ \lambda_D &= \sigma^2\end{aligned}$$

So the first d eigenvalues are determined by Z and ε while the rest are determined only by the noise.

In general, if the noise is small, $\lambda_d >> \lambda_{d+1}$ so it suffices to look only at the first d eigenvalues.

This recontextualizes r_d:

$$r_d = \frac{\sum_{k=1}^{d} \lambda_k}{\sum_{k=1}^{D} \lambda_k} = \frac{\sum_{k=1}^{d} \tilde{\lambda}_k + d\sigma^2}{\sum_{k=1}^{d} \tilde{\lambda}_k + D\sigma^2}$$

if $\varepsilon = 0$ (i.e $\sigma^2 = 0$), $r_d = 1$ so r_d is a metric of the *contamination* from noise.

A Small Example

Let $D = 3$ and $d = 1$ with

- $Z \sim N(0, 1)$
- $L = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$
- $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$ with $\varepsilon_i \sim N(0, 0.04)$

So $X = LZ + \varepsilon$ and

$$\begin{aligned} \mathbb{V}(X) &= L\mathbb{V}(Z)L^T + \mathbb{V}(\varepsilon) \\ &= \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} + \begin{pmatrix} 0.04 & 0 & 0 \\ 0 & 0.04 & 0 \\ 0 & 0 & 0.04 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} + 0.04 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{aligned}$$

Numerical decomposition gives,

$$\begin{cases} \tilde{\lambda}_1 = 3 \\ \tilde{\lambda}_2 = 0 \\ \tilde{\lambda}_3 = 0 \end{cases} \implies \begin{cases} \lambda_1 = 3.04 \\ \lambda_2 = 0.04 \\ \lambda_3 = 0.04 \end{cases}$$

Lecture 23: Dec 05

Review of Probability Theory

The three building blocks of probability thoery:

1. *Sample space* Ω*:* the collection of all possible outcomes
2. *Random Variables:* $X : \Omega \to \mathbb{R}^d$
3. *Probability:* $\mathbb{P} : \{A \subset \Omega\} \to [0, 1]$ satisfying 3-axioms

Law of Large Numbers: Let $X_1, X_2, \ldots, X_n$ be iid scalar-valued RVs. If $\mathbb{E}(|X_i|) < \infty$,

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

so we have an estimator

$$\widehat{v}_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \approx v = \mathbb{E}f(X) = \begin{cases} \int f(x) \cdot p(x)\, dx & \text{if continuous} \\ \sum_x f(x) \cdot p(x) & \text{if discrete} \end{cases}$$

Law of the Iterated Logarithm: If $\text{Var } X_i < \infty$, then

$$|\widehat{v}_n - v| = \left| \frac{1}{n} \sum_{i=1}^{n} X_i(\omega) - \mathbb{E}X_1 \right| \leq \sqrt{\text{Var } (X_1) \cdot \frac{2 \log(\log n)}{n}}$$

Generating (Scalar-valued) Random Numbers from PRNGs

Multiplicative Congruential Generator: generates $u_1, u_2, \ldots, u_n$ that look like they are iid from Unif$(0, 1)$

Inverse CDF: If $F(t)$ is a CDF, we define $G(u) = \inf\{t \in \mathbb{R} : F(t) \geq u\}$ where $u \sim \text{Unif}(0, 1)$. Then $G(u) \overset{iid}{\sim} F$.

Thus,

$$\text{MCG} \longrightarrow \text{PRNs from Unif}(0, 1) \xrightarrow{\text{Inverse CDF}} \text{PRNs from } F$$

Monte Carlo Integration

Using the LLN, the LIL, and the PRNG, we can approximate the integral of a function $f : \mathbb{R}^d \to \mathbb{R}$:

1. The PRNG gives us $X_1, X_2, \ldots, X_n \overset{iid}{\sim} p$
2. Using the LLN, we calculate $\widehat{v}_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i) \approx \mathbb{E}f(X_1) = v$ (where v can be a very complicated integral)
3. The LIL quantifies this error

Importance Sampling

Goal: compute

$$v = \underbrace{\int \cdots \int}_{dintegrals} G(x_1, \ldots, x_d)\, dx_1, \ldots dx_d = \int_{\mathbb{R}^d} H(\vec{x})\, dx$$

We choose a d-dim PDF $f(\vec{x})$ such that $f(\vec{x})$ is "similar to" $\frac{H(\vec{x})}{\int H(\vec{x})\, d\vec{x}}$. This allows us to write

$$v = \int H(\vec{x})\, d\vec{x} = \int \frac{H(\vec{x})}{f(\vec{x})} \cdot f(\vec{x})\, d\vec{x} = \mathbb{E}[\frac{H(\vec{X})}{f(\vec{X})}]$$

(where the final $\vec{X} \sim f$)

When

$$f(\vec{x}) = \prod_{i=1}^{d} x_i$$

we can generate random numbers such that ... so

$$\widehat{v}_n = \frac{1}{n}\sum_{i=1}^{n} \frac{H(\vec{X}_i)}{f(\vec{X}_i)} \approx \int H(\vec{x})\, d\vec{x}$$

This method is preferred when d is large.

Markov Chain Monte Carlo

Motivation: In general, it is infeasible to generate random vectors from a high dimensional distribution

Ergodic Theorem: Let $\{X_n\}_{n=0}^{\infty}$ be a homogeneous Markov Chain with state space $\mathfrak{X} = \{x_1, x_2, \ldots, x_S\}$. If the MC is

1. irreducible
2. aperiodic
3. has a unique stationary distribution $\pi : \mathfrak{X} \to [0, 1]$

then

1. $$\lim_{n\to\infty} \mathbb{P}(X_n = y \mid X_0 = x) = \pi(y) quad x, y \in \mathfrak{X}$$
2. $$\lim_{n\to\infty} \mathbb{P}(X_n = y) = \pi(y)$$
3. $$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(X_i) = \sum_{x\in\mathfrak{X}} \pi(x) \cdot f(x) = \mathbb{E}[f(\xi)]$$

   $(\xi \sim \pi)$

Conclusion: If $\{X_n\}_{n=0}^{\infty}$ is such a MC,

- When n is large, $X_n \dot{\sim} \pi$
- $$\widehat{v}_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \approx \sum_{x\in\mathfrak{X}} \pi(x) \cdot f(x)$$

That is, given an HMC $\{X_n\}_{n=0}^{\infty}$, we can calculate $p(x, y) = \mathbb{P}(X_{n+1} = y \mid X_n = x)$ to find *the* solution to

$$\pi(x) = \sum_{y\in\mathfrak{X}} \pi(y) \cdot p(y, x) \quad \forall x \in \mathfrak{X}$$

To go the other direction (to find an HMC given π) is harder but can be accomplished with

- The Metropolis Hastings Algorithm
- Gibbs Sampling

Gibbs Sampling

Gibbs sampling works only if the model in question is a Gibbs Random Field.

Gibbs Random Field: Let $\mathfrak{X} = \{x_1, x_2, \ldots, \ x_S\}$ be a finite set of states. A probability distribution π on $\mathfrak{X}$ is a *Gibbs Random Field* on a graph $G = (V, E)$ with $V = 1, 2, \ldots, \ d$ if

$$\pi(x_1, \ldots, \ x_d) = \frac{1}{Z}\pi_{c \in C(G)}\phi_c(x_c)$$

If

$$\pi_{i \mid -i}(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, \ x_d)$$

depends only on x_j with $j \in N(i) \cup \{i\}$, then π is called a Markov Random Field (MRF) WRT G.

Theorem (Hammersley-Clifford): Iff π is a GRF with respect to a graph, it is a MRF