

# Monte Carlo Methods - Introduction

Prof. Dr. Narayan Prasad Adhikari  
Central Department of Physics  
Tribhuvan University Kirtipur, Kathmandu, Nepal

January 28, 2024



## ■ Warm up!!!

- Important Discoveries
- What is it?
- Why do we need it in Data Science?

## ■ Warm up - what is really it?!!!

You may face multidimensional integration. It comes in Mathematics as well as in Physics, Statistics ....

$$\int \int \dots \int f(x_1) f(x_2) \dots f(x_n) dx_1 dx_2 \dots dx_n \quad (1)$$

In above equation n may be huge ... 100,  $10^4$ , or  $10^6$ ,....?

How to handle such a complex problem? Just think about it....

Does our ways of doing till now work?

# ■ Warm up - what is really it?!!!

Activities Google Chrome •  
"Monte Carlo" - Google Search Estimation of Absorbed Dose Why the Monte Carlo method... •  
scholar.google.com/scholar?hl=en&as\_sdt=0%2C5&q="Monte+Carlo"&btnG=

Gmail YouTube Maps

≡ Google Scholar "Monte Carlo" 

Articles About 4,740,000 results (0.06 sec)

Any time Since 2022 Since 2021 Since 2018 Custom range...

The monte carlo method N Metropolis, S Ulam - Journal of the American statistical association, 1949 - Taylor & Francis  
We shall present here the motivation and a general description of a method dealing with a class of problems in mathematical physics. The method is, essentially, a statistical approach ...  
☆ Save 99 Cite Cited by 8220 Related articles All 16 versions [PDF] hedibert.org

Sort by relevance Sort by date

Monte Carlo theory and practice F James - Reports on progress in Physics, 1980 - iopscience.iop.org  
... in the Monte Carlo approach. The aim of this review is, first, to lay a theoretical basis for both the 'traditional' Monte Carlo and quasi-Monte Carlo ... of Monte Carlo, quasi-Monte Carlo and ...  
☆ Save 99 Cite Cited by 723 Related articles All 15 versions [PDF] iop.org

include patents  include citations

Create alert

[book] Monte carlo methods J Hammersley - 2013 - books.google.com  
This monograph surveys the present state of Monte Carlo methods. we have dallied with certain topics that have interested us Although personally, we hope that our coverage of the ...  
☆ Save 99 Cite Cited by 6688 Related articles All 6 versions

[PDF] Introduction to monte carlo simulation RL Harrison - AIP conference proceedings, 2010 - aip.scitation.org  
... This paper reviews the history and principles of Monte Carlo simulation, emphasizing techniques commonly used in the simulation of medical imaging. ... This paper gives an ...  
☆ Save 99 Cite Cited by 286 Related articles All 8 versions [PDF] scitation.org

Related searches

"monte carlo" simulation "monte carlo" dose  
"monte carlo" calculations kinetic "monte carlo"  
markov chain "monte carlo" sequential "monte carlo"

# ■Warm up - Health

[View PDF](#)[Download Full Issue](#)

Physica A: Statistical Mechanics and its  
Applications

Volume 574, 15 July 2021, 126014



## A random walk Monte Carlo simulation study of COVID-19-like infection spread

[...]

S. Triambak <sup>a</sup> D.P. Mahapatra <sup>b</sup>

Show more

+ Add to Mendeley Share Cite

---

<https://doi.org/10.1016/j.physa.2021.126014>

Get rights and content

### Abstract

Recent analysis of early COVID-19 data from China showed that the number of confirmed cases followed a subexponential power-law

# ■ Warm up - Health

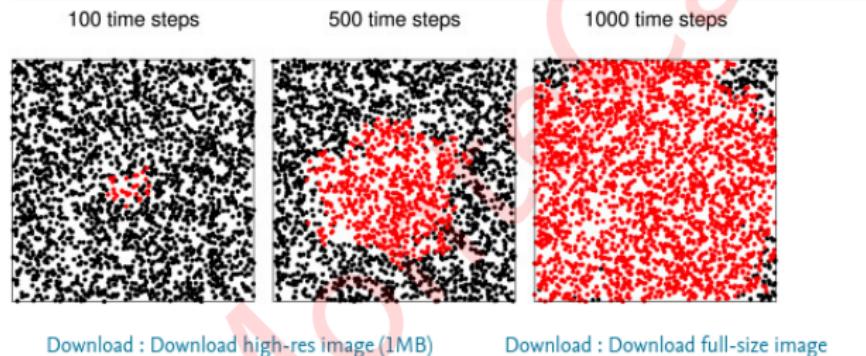


Fig. 1. An example of proximity-based infection spread obtained using the random walk Monte Carlo simulations described in this work. Each of the panels shown above has a population of 2.5k over a unit area. The average distance  $\langle r \rangle$  between any two points is = 0.02 units. In this case every point (walker) takes randomly directed steps of length  $l = 0.25 \langle r \rangle$ . Further details are described in the text below.

# ■ Warm up - Health

Estimation of Absorbed Dose Distribution in Different Organs during the CT Scan: Monte Carlo Study

1 / 3 | - 100% + | Open Access

Austin Journal of Radiology

Austin Publishing Group

Special Article - Therapeutic Radiology

## Estimation of Absorbed Dose Distribution in Different Organs during the CT Scan: Monte Carlo Study

Umit Kara<sup>a\*</sup> and Hacerin Ozan Tekin<sup>b</sup>

<sup>a</sup>Medical Imaging, Bilecik Sultani Devlet University, Turkey  
<sup>b</sup>Department of Radiotherapy, Usak University, Turkey

\*Corresponding author: Umit Kara, Suleyman Demirel University, Vocational School of Health Services, Medical Imaging, 32100, Isparta, Turkey

Received: March 08, 2017; Accepted: March 30, 2017;  
Published: April 04, 2017

This work aimed to validate the accuracy of a Monte Carlo source model of the GE LightSpeed CT scanner using organ doses measured in specific Human adult phantoms. The x-ray dose from the GE LightSpeed multi-slice CT scanner was simulated using the Monte Carlo method. The resulting source model was able to perform various real simulated scan model helical Computed Tomography (CT) scans of varying scan parameters such as kVp, mAs, pitch, and slice thickness. The results were compared with the corresponding real Computed Tomography (CT) protocols to patients in public hospitals in Turkey. We used Monte Carlo simulation methods with real height and weight of patients and Computed Tomography (CT) scanners real parameters. Absorbed organ doses were calculated by using the Monte Carlo simulation. The data showed that mAs value is significantly important for obtaining the risk of cancer from dose rates. Additionally, the dose received by each organ has been calculated and the results showed that Monte Carlo is a strong and effective tool in radiological dosimetry.

**Keywords:** CT scan, Absorbed dose, Monte-carlo simulation

### Introduction

Diagnostic radiology is a significant tool for clinical diagnosis and includes general X-ray, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), ultrasound, and nuclear medicine. Diagnostic radiology is part of medicine that uses medical imaging facility and technology to diagnose disease and helps define the structures inside a human body. Diagnostic radiology uses the imaging technologies of X-ray radiography, computed tomography, magnetic resonance imaging, ultrasound, nuclear medicine, and Computed Tomography of CT scans are abdominal, breast and vascular system. CT scanning is the one of most used units in diagnostic radiology and these units consist the use of X-rays and computer processing to generate tomographic images of the body. The use of CT in radiology and medical imaging has been growing in the world and especially in the last decade due to its high resolution picture of the entire body. CT scanners, low dose, and accurate CT scan uses X-ray equipment and computers to produce medical images that often can make more detailed image than conventional radiography. CT scans use ionizing radiation as used in other X-rays units. In clinical, the benefits of a correct diagnosis preponderance most accurate, reliable, and versatile is in accomplishing this task [1-4].

In the Monte Carlo method, the patient and CT scanner are simulated using a computational anatomical model of the patient and an X-ray source model representing the scanner's beam output. However, to ensure the accuracy of these calculations, these CT source models must be benchmarked and validated against actual experimental measurements made on the scanners they simulate. In the past, most phantom studies have been conducted using the AAPM Task Group 210 (TG210) phantoms, but in recent years, anthropomorphic phantoms have been increasingly utilized [6,10]. Mathematical phantoms or MIRD phantoms have been produced [11] which were the first models of human phantom to be widely used in dosimetry studies involving X-ray exposure by the Monte Carlo method.

However, the organs in these phantoms are described by mathematical equations with limited representation of the actual structure of a human body and its chemical and physical characteristics. These phantoms have been developed for tomography images provide a geometry which adequately represents a patient including internal organs, displacements, and deformations. These phantoms are recommended for dosimetric studies with the Monte

whyMCM\_fin.pdf

1949Metrop...pdf

7

## ■ Warm up - Health

- Case of Siamese Twins
- Pharmacy
- MC dose calculation in the radiotherapy treatment

# ■ Warm up - Artificial intelligence

Outline  
Highlights  
Abstract  
Keywords  
1. Introduction  
2. Adapting MCTS  
3. Examples of usage  
4. Quality of decisions  
5. Conclusion & future work  
[Author statement](#)



Computers in Industry  
Volume 128, June 2021, 103433



## Monte Carlo Tree Search for online decision making in smart industrial production

Richard Senington , Bernard Schmidt , Anna Syberfeldt

[Show more](#)



[Download Full Issue](#)

[References](#)

Under a Creative Commons license

Open access

Artic

Monte Carlo methods are also pervasive in artificial intelligence and machine learning. Many important technologies used to accomplish machine learning goals are based on drawing samples from some probability distribution and using these samples to form a Monte Carlo estimate of some desired quantity.

# ■ Warm up - Risk Analysis

[View PDF](#)[Access through your institution](#)[Purchase PDF](#)

Reliability Engineering &amp; System Safety

Volume 199, July 2020, 106792



Recom

Optimal

Reliability

[Purch](#)

An integ

Reliability

[Purch](#)

Bayesian

Reliability

[Purch](#)

e work

g interest

y materials

Risk analysis of an underground gas storage facility using a physics-based system performance model and Monte Carlo simulation

Zaki Syed , Yuri Lawryshyn

[Show more](#) [+ Add to Mendeley](#) [Get rights and content](#)<https://doi.org/10.1016/j.ress.2020.106792>

Article

Citation:

Citation I

Capture

Readers:



## Highlights

- A risk analysis model of underground gas storage facility operational reliability.
- Physics-based performance model combined with Monte Carlo simulation of disruptions.

# ■ Warm up - Sensitivity

[View PDF](#) | [Access through your institution](#) | [Purchase PDF](#) | [Search](#)

---

View  
Issues  
About  
All articles (6)

 European Journal of Operational Research  
Volume 298, Issue 1, 1 April 2022, Pages 229-242 

---

Stochastics and Statistics

## Sensitivity estimation of conditional value at risk using randomized quasi-Monte Carlo

Zhijian He  

Show more 

+ Add to Mendeley  Share  Cite 

---

<https://doi.org/10.1016/j.ejor.2021.11.013> [Get rights and content](#)

---

### Abstract

Conditional value at risk (CVaR) is a popular measure for quantifying portfolio risk. Sensitivity analysis of CVaR is common in risk management and gradient-based optimization algorithms. In this paper, we study the infinitesimal perturbation analysis estimator for CVaR sensitivity using randomized quasi-Monte Carlo (RQMC) simulation.

# ■ Warm up - powerplant

 View PDF     Access through your institution    Purchase PDF

 ELSEVIER

ISA Transactions  
Volume 100, May 2020, Pages 171-184 

R  
A  
N  
T  
F  
N  
T  
A  
C  
T  
A  
C  
T  
A  
C  
T  
C  
C  
C  
R  
C  
R

Research article

## Nonlinear robust fault diagnosis of power plant gas turbine using Monte Carlo-based adaptive threshold approach

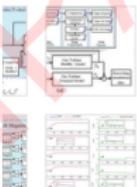
Saeed Amirkhani <sup>a,b</sup>, Ali Chaibakhsh <sup>a,b,✉</sup>, Ali Ghaffari <sup>c</sup>

Show more ▾

+ Add to Mendeley  Share  Cite

<https://doi.org/10.1016/j.isatra.2019.11.035> [Get rights and content](#)

---



### Highlights

- The adaptive threshold approach is used for the gas turbine fault diagnosis.
- Adaptive threshold bounds are determined based on Monte Carlo simulations.
- The robustness of fault detection is analysed through

12

# ■ Warm up - Finance

[View PDF](#) [Download Full Issue](#)

 Procedia Computer Science  
Volume 1, Issue 1, May 2010, Pages 2381-2389 

International Conference on Computational Science, ICCS 2010  
**Complex systems in finance: Monte Carlo evaluation of first passage time density functions**

O. Tsviliuk<sup>a</sup>, D. Zhang<sup>b</sup>, R. Melnik<sup>a</sup>

Show more ▾

+ Add to Mendeley  Cite 

<https://doi.org/10.1016/j.procs.2010.04.268> [Get rights and content](#)  Open access

Under a Creative Commons license

Abstract

Many examples of complex systems are provided by applications in finance and economics areas. Some of intrinsic features of such systems lie with the fact that their parts are interacting in a non-trivial dynamic

Part of ICCS 2010

Other:

- Evaluating May 2010 
- The latest May 2010 
- Genetic I May 2010 

View more

Recom

Article

Citation:

Monte Carlo analysis is useful in risk analysis because many investment and business decisions are made on the basis of one outcome. In other words, many analysts derive one possible scenario and then compare that outcome to the various impediments to that outcome to decide whether to proceed.

## ■ Warm up - Physics

- diffusion Limited Aggregation

## ■ Warm up - what is really it?!!!

- Monte Carlo methods, or Monte Carlo experiments, are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results. The underlying concept is to use randomness to solve problems that might be deterministic in principle.
- In a Monte Carlo simulation (method) we attempt to follow the 'time dependence' of a model for which change, or growth, does not proceed in some rigorously predefined fashion (e.g. according to Newton's equation of motion) but rather in a stochastic manner which depends on a sequence of random numbers which is generated during the simulation

## ■ Warm up - what is really it?!!!

- Monte Carlo method is used to estimate the possible outcomes of an uncertain event. The Monte Carlo Method was invented by John von Neumann and Stanislaw Ulam during World War II to improve decision making under uncertain conditions. It was named after a well-known casino town, called "Monte Carlo" in Monaco, since the element of chance is core to the modeling approach, similar to a game of roulette.
- Monte Carlo Simulations have assessed the impact of risk in many real-life scenarios, such as in artificial intelligence, stock prices, sales forecasting, project management, and pricing.

## ■ Warm up - What is it!!!!

- When Ulam was in hospital he was playing cards (just for time passing) and got the idea of random sampling. He then applied this idea along with Neuman to solve the problem of "neutron diffusion" in the Manhattan project
- Monte Carlo algorithms are simple, flexible, and scalable. When applied to physical systems, Monte Carlo techniques can reduce complex models to a set of basic events and interactions, opening the possibility to encode model behavior through a set of rules which can be efficiently implemented on a computer.
- However he published paper about MC simulation only in 1949 from LANL (Los Alamos National Lab)

## ■ Warm up - What is it!!!!

- Take a random sample of given population. Calculate many different outcomes and their probabilities of occurrence
- This outcome represents the desired results.

## ■ Warm up - What is it!!!!

- Consider a simple example of rolling dice. Assume that you want to determine the probability of rolling a seven using two dice with values one through six. There are 36 possible combinations for the two dice, six of which will total seven, as shown in the following image.

	Column-1	Column-2	Column-3	Column-4	Column-5	Column-6
Row-1						
Row-2						
Row-3						
Row-4						
Row-5						
Row-6						

## ■ Warm up - What is it!!!!

- This means that mathematical probability of rolling a seven is six in 36, or 16.67 percent.
- But is the mathematical probability the same as the actual probability? Or are there other factors that might affect the mathematical probability, such as the design of the dice themselves, the surface on which they are thrown, and the technique that is used to roll them?
- To determine the actual probability of rolling a seven, you might physically roll the dice 100 times and record the outcome each time. Assume that you did this and rolled a seven 17 out of 100 times, or 17 percent of the time. Although this result would represent an actual, physical result, it would still represent an approximate result. If you continued to roll the dice again and again, the result would become less and less approximate.

## ■ Warm up - What is it!!!!

- A Monte Carlo simulation is the mathematical representation of this process. It allows you to simulate the act of physically rolling the dice and lets you specify how many times to roll them. Each roll of the dice represents a single iteration in the overall simulation; as you increase the number of iterations, the simulation results become more and more accurate. For each iteration, variable inputs are generated at random to simulate conditions such as dice design, rolling surface, and throwing technique. The results of the simulation would provide a statistical representation of the physical experiment described above.

## ■ Warm up - What is it!!!!

- Bayesian Statistics is fundamentally all about modifying conditional probabilities – it uses prior distributions for unknown quantities which it then updates to posterior distributions using the laws of probability. In fact Bayesian statistics is all about probability calculations!
- MC method is some how based on the Bayesian Statistics
- Bayesian Statistics is a theory in the field of statistics based on the Bayesian interpretation of probability where probability expresses a degree of belief in an event

# Random numbers - Introduction

Prof. Dr. Narayan Prasad Adhikari  
Central Department of Physics  
Tribhuvan University Kirtipur, Kathmandu, Nepal

January 30, 2024



## ■ Warm up!!!

- What are random numbers?
- Can we really get random numbers?
- Pseudorandom numbers
- Generating random numbers in your own laptop

## ■What are random numbers?

- **What is a random number?** As the term suggests, a random number is a number chosen by chance i.e., randomly, from a set of numbers.
- Random numbers play vital role in Monte Carlo Methods (simulation).
- The earliest methods for generating random numbers, such as dice, coin flipping and roulette wheels, are still used today, mainly in games and gambling as they tend to be too slow for most applications in statistics and cryptography.

## ■Who generated first random numbers?

- John von Neuman gave idea to generate random numbers in 1946
- His idea was to start with an initial random seed value, square it, and slice out the middle digits. If you repeatedly square the result and slice out the middle digits, you'll have a sequence of numbers that exhibit the statistical properties of randomness.
- An example: Consider any large numbers say - 2934; square is:8608356; you pick 083 as a random number; square of 83 is 6889; next random number became 88 ...

## ■ Main properties of random numbers

- Good random number generator should be
  - (a) random
  - (b) reproducible
  - (c) portable
  - (d) efficient

## ■ Random numbers (RN)- Background

- MC methods are heavily dependent on the fast, efficient production of streams of random numbers.
- Physical processes such as white noise - a random signal having equal intensity at different frequencies, generation from electrical circuits are too slow
- If you are interested in MC you must be able to generate your random numbers (that too in sequence)
- Since such sequences are actually deterministic, the random number sequences we produce in our laptop/computers are only "pseudo-random"
- It is important for you to understand the limitations of pseudo random number generators (PRNG)

## ■ Random numbers (RN)- Background

- In our context - "random numbers" (RN) means "pseudo-random numbers (PRN)"
- These deterministic features of PRN are not always negative.
- For example - for testing a program it is often useful to compare the results with a previous run made using exactly same random numbers.

## ■Monte Carlo (MC) methods

- MC simulations are subject to both statistical and systematic errors from multiple sources.
- If your RN are of poor quality it leads to systematic errors
- In fact the testing as well as the generation of random numbers remain important problems that have not been fully solved yet.  
So its for you ....
- As mentioned above RN sequences which are needed in MC should be uniform, uncorrelated, and of extremely long period i.e. do not repeat over quite long intervals.
- Also if you use parallel computing (of course you must to handle large data), you must insure all the random numbers sequences generated are distinct and uncorrelated

## ■Generation of PRNs- Congruential method

- Most popular method - multiplicative OR congruential method
- **Main idea:**A fixed number  $c$  is chosen along with a given seed and subsequent numbers are generated by simple multiplication

$$X_n = (c \times X_{n-1} + a_0) MOD N_{max}$$

where  $X_n$  is an integer between 1 and  $N_{max}$ .

## ■Generation of PRNs- Congruential method

- Experience has shown that a good congruential generator is the 32-bit linear congruential (CONG) algorithm:

$$X_n = (16807 \times X_{n-1}) MOD (2^{31} - 1)$$

- Some people call the number "16807" a **A Miracle Number**
- Even though CONG showed some drawbacks it is still popular being simplest way to generate random numbers

## ■Generation of PRNs- Congruential method: algorithm

You need to produce random numbers from seed using above formula  
Use following algorithm:

1. Start
2. For loop (I mean to produce many random numbers) set count 0
3. Define seed ( A large number)
4. start loop (while or any other you like)
5. Calculate  $\text{ran} = 16807 * \text{seed}$
6. Set seed equal to ran for the next iteration of the loop
7. Print random number you generated
8. Increase count by 1
9. End program

## ■Generation of PRNs- Congruential method: algorithm

The code in Python looks like:

```
count=0  
seed=1982537  
while (count <100):  
    ran=(16807*seed)%(2**31-1)  
    seed =ran  
    print (ran)  
    count=count+1
```

## ■Generation of PRNs- Congruential method: algorithm

For following codes each time you must write an algorithm.  
You can change the first one to add another one

CWI: Write an algorithm to open a file and write above random numbers in that file.

CWII: Now write two random numbers in the file at a time (say ran1 and ran2)

CWIII: Now convert ran1 and ran2 to lie in the range of (0,1).

CWIV: Plot ran2 vs ran1.

CWV: Check the distribution of random numbers.

## ■Generation of PRNs-Python random()

Now write a code (python) using following algorithm:

1. Start
2. import random
3. open file
4. start a loop as before
5. get random numbers from python's intrinsic function random()
6. Write them in a file (generate two columns)
7. End loop
8. Close file
9. End program

## ■Generation of PRNs-Python random()

The code looks like:

```
import random
f = open("rand.dat", "w")
count = 0
while (count < 100):
    print (random.random(),random.random())
    f.write("{} {} \n".format(random.random(), random.random()))
    count = count + 1
f.close()
```

## ■Generation of PRNs - Python random()

HWI: Now you compare the distribution of random numbers you generated using congruent method and built in function of python. Compare them and discuss.

I will evaluate this HW for your grading

## ■Generation of PRNs -Other algorithms

- HW: Now you try to understand at least one more PRNGs algorithm
- Can you convert uniform distribution to gaussian distribution?
- For this: pick any two random numbers  $x_1$  and  $x_2$  from uniform distribution

$$y_1 = (-2 \ln(x_1))^{1/2} \cos(2\pi x_2) \quad (1)$$

$$y_2 = (-2 \ln(x_1))^{1/2} \sin(2\pi x_2) \quad (2)$$

HW: Given a sequence of uniformly distributed random numbers  $y_i$  show how sequences  $x_i$  distributed according to  $x^2$  would be produced.

## ■ HW: Properties of selected RNGs

- The underlying PDF for the generation of random numbers is the uniform distribution, meaning that the probability for finding a number  $x$  in the interval  $[0, 1]$  is  $p(x) = 1$ .
- A random number generator should produce numbers which uniformly distributed in this interval.
- Just think about different ways to check this distribution
- One way: by plotting as before.
- Another way: You just find number of random numbers between  $0.0 - 0.1$ ,  $0.1-0.2$ , ...,  $0.9-1.0$  for say large numbers of random numbers say 100000. You develop a code to read random numbers generated and find RNs between those limits.

## ■ HW: Properties of selected RNGs

- Two additional measures are the s.d.  $\sigma$  and the mean  $\mu = \langle x \rangle$ .
- For the uniform distribution with  $N$  points we have that the average  $\langle x^k \rangle$ . is

$$\langle x^k \rangle = \frac{1}{N} \sum_{i=1}^N x_i^k p(x_i) \quad (3)$$

and taking the limit  $N \rightarrow \infty$  we have

$$\langle x^k \rangle = \int_0^1 dx p(x) x^k = \int_0^1 dx x^k = \frac{1}{k+1} \quad (4)$$

as  $p(x) = 1$ .

$$\therefore \mu = \langle x \rangle = \frac{1}{2} \quad (5)$$

## ■ HW: Properties of selected RNGs

- Similarly standard deviation is

$$\sigma = \sqrt{\langle x^2 \rangle - \mu^2} = \frac{1}{\sqrt{12}} = 0.2886 \quad (6)$$

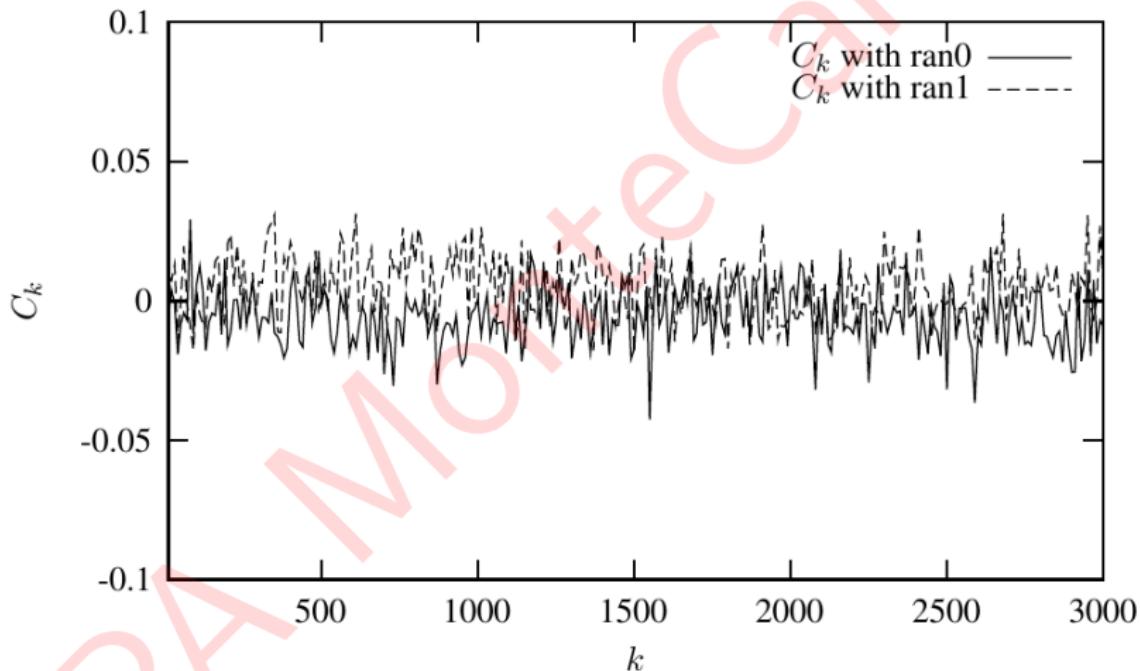
HW: Now you write a code to check your random numbers's distributions. Also evaluate mean and variance of them.

- Auto-Correlation function:** Since our random numbers, which are typically generated via a linear congruential algorithm, are never fully independent, we can then define an important test which measures the degree of correlation, namely the so-called auto-correlation function  $C_k$

$$C_k = \frac{\langle x_{i+k}x_i \rangle - \langle x_i \rangle^2}{\langle x_i^2 \rangle - \langle x_i \rangle^2} \quad (7)$$

with  $C_0 = 1$ . The non-vanishing of  $C_k$  for  $k \neq 0$  means that the random numbers are not independent. The independence of the random numbers is crucial in the evaluation of other expectation values.

## ■ HW: Properties of selected RNGs



HW: Now you calculate auto correlation functions for three different RNs and plot them. Can you explain the fluctuations as shown in above figure.

## ■ HW: Properties of selected RNGs

- The expectation values which enter the definition of  $C_k$  are given by

$$\langle x_{i+k} x_i \rangle = \frac{1}{N-k} \sum_{i=1}^{N-k} x_i x_{i+k} \quad (8)$$

# Monte Carlo - Basics

Prof. Dr. Narayan Prasad Adhikari  
Central Department of Physics  
Tribhuvan University Kirtipur, Kathmandu, Nepal

March 15, 2024



# ■Markov Chain and Master Equation

- Random variables
- Concept of errors
- Estimation of errors
- Markov Chain
- Master Equation

# ■ Random variables

- Consider an elementary event with a countable set of random outcomes,  $A_1, A_2, \dots, A_k$  (e.g. you can consider a rolling dice OR a set of "Khodkhode". )
- You are data scientist so you need to consider this event occurring repeatedly say  $N$  times such that  $N \ggg 1$  and we count how often the outcome  $A_k$  is observed ( $N_k$ ).
- The probabilities  $p_k$  for outcome  $A_k$  is

$$p_k = \lim_{N \rightarrow \infty} \left( \frac{N_k}{N} \right) \quad (1)$$

with  $\sum_k p_k = 1$ .

Obviously  $0 \leq p_k \leq 1$

You are familiar with conditional probability  $P(j/i)$ , average of any outcomes of such random events  $x_i$ , its variances and so on.

## ■ Statistical errors

- Suppose the quantity  $A$  is distributed according to a Gaussian with mean value  $\langle A \rangle$  and width  $\sigma$ . We consider  $n$  statistically independent observations  $\{A_i\}$  of this quantity  $A$ .
- An unbiased estimator of the mean  $\langle A \rangle$  of this distribution is

$$\bar{A} = \frac{1}{n} \sum_{i=1}^n A_i \quad (2)$$

and the standard error of this estimate is

$$\text{error} = \frac{\sigma}{\sqrt{n}} \quad (3)$$

# ■ Statistical errors

- The variance is obtained from mean square deviation

$$\delta \bar{A}^2 = \frac{1}{n} \sum_{i=1}^n (\delta A_i)^2 = \bar{A}^2 - (\bar{A})^2 \quad (4)$$

The expectation value of this quantity is easily related to  $\sigma^2 = \langle A^2 \rangle - \langle A \rangle^2$  as

$$\langle \delta \bar{A}^2 \rangle = \sigma^2 (1 - 1/n) \quad (5)$$

$$\therefore \text{error} = \sqrt{\frac{\delta \bar{A}^2}{(n-1)}} = \sqrt{\frac{\sum_{i=1}^n (\delta A_i)^2}{(n(n-1))}} \quad (6)$$

## ■ Ingredients of MC

- As mentioned before there are at least four ingredients which are crucial in order to understand the basic MC strategy.
  - (i) Random variables
  - (ii) Probability distribution functions (PDF),
  - (iii) Moments of a PDF
  - (iv) and pertinent variance  $\sigma$

## ■Random Variables

- Let us first demistify the somewhat obscure concept of a random variable. The example we choose is the classic one, the tossing of two dice, its outcome and the corresponding probability. In principle, we could imagine being able to determine exactly the motion of the two dice, and with given initial conditions determine the outcome of the tossing. Alas, we are not capable of pursuing this ideal scheme. However, it does not mean that we do not have a certain knowledge of the outcome. This partial knowledge is given by the probability of obtaining a certain number when tossing the dice. To be more precise, the tossing of the dice yields the following possible values

$$[ 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12.] \quad (7)$$

## ■ Random Variables

- These values are called the domain. To this domain we have the corresponding probabilities

$$[1/36, 2/36, 3/36, 4/36, 5/36, 6/36, 5/36, 4/36, 3/36, 2/36, 1/36] \quad (8)$$

- These values are called the domain. To this domain we have the corresponding probabilities
- The numbers in the domain are the outcomes of the physical process tossing the dice. We cannot tell beforehand whether the outcome is 3 or 5 or any other number in this domain. This defines the randomness of the outcome, or unexpectedness or any other synonymous word which encompasses the uncertainty of the final outcome.

## ■Random Variables

- The only thing we can tell beforehand is that say the outcome 2 has a certain probability. If our favorite hobby is to spend an hour every evening throwing dice and registering the sequence of outcomes, we will note that the numbers in the above domain

$$[ 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12.] \quad (9)$$

appear in a random order.

after (say) 11 throws the results may look like

$$[ 10, 8, 6, 3, 6, 9, 11, 8, 12, 4, 5] \quad (10)$$

- Eleven new attempts may results in a totally different sequence of numbers and so forth. Repeating this exercise the next evening, will most likely never give you the same sequences. Thus, we say that the outcome of this hobby of ours is truely random.

## ■ Random Variables

- *Random variables are hence characterized by a domain which contains all possible values that the random value may take. This domain has a corresponding PDF.*

## ■MC Illustration - Integration

- Consider an integration

$$I = \int_0^1 f(x)dx \simeq \sum_{i=1}^N w_i f(x_i) \quad (11)$$

where  $w_i$  are the weights determined by specific integration methods like Trapezoid, Simpson etc. In the crudest approach here in MC integration we set up  $w_i = 1$  then above eq becomes

$$I = \int_0^1 f(x)dx \simeq \frac{1}{N} \sum_{i=1}^N f(x_i) \quad (12)$$

Now introduce the concept of the average of the function  $f$  for a given PDF  $p(x)$  as

$$\langle f \rangle = \frac{1}{N} \sum_{i=1}^N f(x_i)p(x_i) \quad (13)$$

## ■MC Illustration - Integration

Now identify  $p(x_i) = 1$  with the uniform distribution when  $x \in [0, 1)$  and zero for all other values of  $x$ . Then

$$I = \int_0^1 f(x)dx \simeq \frac{1}{N} \sum_{i=1}^N f(x_i) \simeq \langle f \rangle \quad (14)$$

Similarly the variance (which is also important in MC methods) is

$$\sigma_f^2 = \frac{1}{N} \sum_{i=1}^N (f(x_i) - \langle f \rangle)^2 p(x_i) \quad (15)$$

After inserting value of  $p(x_i)$  we get

$$\sigma_f^2 = \frac{1}{N} \sum_{i=1}^N f^2(x_i) - (\langle f \rangle)^2 \quad (16)$$

## ■MC Illustration - Integration:Algorithm

- Choose the number of Monte Carlo samples N.
- Perform a loop over N and for each step generate a random number  $x_i$  in the interval [0, 1] through a call to a random number generator. Translate the random numbers to other required interval if it needs.
- Use this number to evaluate  $f(x_i)$ .
- Evaluate the contributions to the mean value and the standard deviation for each loop.
- After N samples calculate the final mean value and the standard deviation.

## ■MC Illustration - Integration

- As an example evaluate following by MC method:

$$I = \int_0^1 \exp(x) dx \quad (17)$$

and

$$I = \int_1^3 \exp(x) dx \quad (18)$$

Also Compare the final results with the correct and hence estimate the errors.

## ■MC Illustration - Integration

```
import random
import numpy as np
import math
a = 0.
b = 1.
integral = 0.0
i=0
while i<1000:
    x=random.random()
    integral += math.exp(x)
    i=i+1
ans=integral*(b-a)/float(N)
print ("The value calculated by monte carlo integration is {"
        }.format(ans))
```

**HW:** You also estimate it following above algorithm. Further estimate integral for different set of random numbers and hence estimate  $\sigma_N$ .

## ■MC Illustration - Estimate $\pi$

1. Initialize circle points, square points and interval to 0.
2. Generate random point x.
3. Generate random point y.
4. Calculate  $d = x*x + y*y$ .
5. If  $d \leq 1$ , increment circle points.
6. Increment square points.
7. Increment interval.
8. If increment < NOOFITERATIONS, repeat from 2.
9. Calculate  $\pi = 4 * (\text{circle points} / \text{square points})$ .
10. Terminate.

Also follow the discussion in my lecture

## ■ HW: Estimate $\pi$

1. Estimate value of  $\pi$  also from

$$\pi = \int_0^1 4 \frac{dx}{1+x^2} \quad (19)$$

2. What we did in previous slide was to estimate value of  $\pi$  from area of a circle in 2 dimensions. Can you think of similar methods for higher dimensions? You may need volume of a hypersphere of radius  $R$  in  $n$  dimensions:

$$V_n(R) = \frac{\pi^{n/2}}{\left(\frac{n}{2}\right)!} R^n \quad (20)$$

what you did in 2D is just a special case of above equation in  $n=2$ .

## ■Concept of Importance Sampling

- Till now we discussed about 'simple sampling' of MC
- In principle, MC integrations and other simulations can be performed using the simple sampling techniques we discussed till now. Unfortunately most of the samples produced in this fashion will contribute relatively little to the equilibrium (time independent) averages and more sophisticated methods are required if we are to obtain results of sufficient accuracy to be useful.
- One of such a methods is "Importance Sampling". For this we need to discuss change of variables.

# ■Concept of Importance Sampling

- With improvements we think of a smaller variance and the need for fewer Monte Carlo samples, although each new Monte Carlo sample will most likely be more times consuming than corresponding ones of the brute force method (Simple sampling). For this we consider two topics.
- The first topic deals with change of variables, and is linked to the cumulative function  $P(x)$  of a PDF  $p(x)$ . Obviously, not all integration limits go from  $x = 0$  to  $x = 1$ , rather, in DATA Science we are often confronted with integration domains like  $x \in [0, \infty]$  or  $x \in [-\infty, \infty]$  etc. Since all random number generators give numbers in the interval  $x \in [0, 1]$ , we need a mapping from this integration interval to the explicit one under consideration.

## ■Concept of Importance Sampling

- The next topic deals with the shape of the integrand itself. Let us for the sake of simplicity just assume that the integration domain is again from  $x = 0$  to  $x = 1$ . If the function to be integrated  $f(x)$  has sharp peaks and is zero or small for many values of  $x \in [0, 1]$ , most samples of  $f(x)$  give contributions to the integral  $I$  which are negligible. As a consequence we need many  $N$  samples to have a sufficient accuracy in the region where  $f(x)$  is peaked. What do we do then? We try to find a new PDF  $p(x)$  chosen so as to match  $f(x)$  in order to render the integrand smooth. The new PDF  $p(x)$  has in turn an  $x$  domain which most likely has to be mapped from the domain of the uniform distribution.

## ■ Importance Sampling -Change of variables

- Consider uniform distribution

$$\begin{aligned} p(x)dx &= dx \text{ (for } 0 \leq x \leq 1) \\ &= 0 \text{ else} \end{aligned} \tag{21}$$

with  $p(x) = 1$  and satisfying

$$\int_{-\infty}^{\infty} p(x)dx = 1 \tag{22}$$

All random number generators provided in the program library generate numbers in this domain. When we attempt a transformation to a new variable  $x \rightarrow y$  we have to conserve the probability

$$p(y)dy = p(x)dx \tag{23}$$

which for the uniform distribution implies

$$p(y)dy = dx \tag{24}$$

## ■ Importance Sampling -Change of variables

Let us assume that  $p(y)$  is a PDF different from the uniform PDF  $p(x) = 1$  with  $x \in [0, 1]$ . If we integrate the last expression we arrive at

$$x(y) = \int_0^y p(y') dy' \quad (25)$$

which is nothing but the cumulative distribution of  $p(y)$ , i.e.

$$x(y) = P(y) = \int_0^y p(y') dy' \quad (26)$$

This is an important result which has consequences for eventual improvements over the brute force Monte Carlo.

## ■ Change of variables- an example

Suppose we have the general uniform distribution

$$p(y)dy = \begin{cases} \frac{dy}{b-a} & (\text{for } a \leq y \leq b) \\ 0 & \text{else} \end{cases} \quad (27)$$

If we wish to relate this distribution to the one in the interval  $x \in [0, 1]$  we have

$$p(y)dy = \frac{dy}{b-a} (\text{for } a \leq y \leq b) = dx \quad (28)$$

and integrating we obtain the cumulative function

$$x(y) = \int_a^y \frac{dy'}{b-a} \quad (29)$$

yielding

$$y = a + (b - a)x \quad (30)$$

## ■ Importance Sampling

- With the aid of the above variable transformations we address now one of the most widely used approaches to Monte Carlo integration, namely importance sampling. It will be helpful to sample a function which has peak as we need to consider many more sampling points near the peak.
- Let us assume that  $p(y)$  is a PDF whose behavior resembles that of a function  $F$  defined in a certain interval  $[a, b]$ . The normalization condition is

$$\int_a^b p(y)dy = 1 \quad (31)$$

We can rewrite our integral as

## ■ Importance Sampling

$$I = \int_a^b F(y)dy = \int_a^b p(y) \frac{F(y)}{p(y)} dy \quad (32)$$

Since random numbers are generated for the uniform distribution  $p(x)$  with  $x \in [0, 1]$ , we need to perform a change of variables  $x \rightarrow y$  through

$$x(y) = \int_a^y p(y')dy' \quad (33)$$

where we used

$$p(x)dx = dx = p(y)dy \quad (34)$$

If we can invert  $x(y)$ , we find  $y(x)$  as well. With this change of variables we can express the integral of Eq. 32 as

$$I = \int_a^b p(y) \frac{F(y)}{p(y)} dy = \int_a^b \frac{F(y(x))}{p(y(x))} dx \quad (35)$$

## ■ Importance Sampling

meaning that a Monte Carlo evalutaion of the above integral gives

$$\int_a^b \frac{F(y(x))}{p(y(x))} dx = \sum_{i=1}^N \frac{F(y(x_i))}{p(y(x_i))} \quad (36)$$

The advantage of such a change of variables in case  $p(y)$  follows closely  $F$  is that the integrand becomes smooth and we can sample over relevant values for the integrand. It is however not trivial to find such a function  $p$ .

The conditions on  $p$  which allow us to perform these transformations are

1.  $p$  is normalizable and positive definite,
2. it is analytically integrable and
3. the integral is invertible, allowing us thereby to express a new variable in terms of the old one.

## ■ Important Note

- The average is over  $y(x)$  distribution.

Therefore above equation 35 can be rewritten as

$$I = \int_a^b p(y) \frac{F(y)}{p(y)} dy = \int_a^b p(y) \left\{ \frac{F(y)}{p(y)} \right\} dy = E_{p(y)} \left\{ \frac{F(y)}{p(y)} \right\} \quad (37)$$

is actually expectation value of

$$\left\{ \frac{F(y)}{p(y)} \right\}$$

with distribution  $p(y)$ .

**Please note that the average is over the distribution  $p(y)$  not over  $p(x)$ .**

Therefore in the importance sampling integration you first find the  $p(y)$  corresponding to  $p(x) \in [0, 1]$ . Then find average 36 with distribution  $p(y)$ . See Example below.

## ■ Importance Sampling - Examples

(1) Consider the integral

$$I = \int_0^1 \exp(-x^2) dx \quad (38)$$

Evaluate  $I$  using (i) brute force (simple sampling) MC with  $p(x) = 1$  and (ii) importance sampling with  $p(x) = a \exp(-x)$ .

**Important Note:** You first write Algorithm in each case then write code in python language following the Hints

- (a) Obtain average of  $\exp(-x^2)$  for  $x \in [0, 1]$
- (b) Find its variance too.

These are results of Simple sampling.

for importance sampling:

- (c) Find  $p(y)$  corresponding to  $p(x) = a \exp(-x)$  from equation 33 where  $a$  is normalization constant.
- (d) Then find the expectation value of  $\left[ \frac{\exp(-x^2)}{a \exp(-x)} \right]$  with distribution  $p(y)$ .
- (e) Find variance also.
- (f) Compare both errors or variances and comments on your results.

## ■ Monte Carlo - More on above examples

- Solution of above example (Importance sampling):

$$I = \int_0^1 \exp(-x^2) dx \quad (39)$$

with chosen pdf (probability distribution function)  $p(x) = a \exp(-x)$  such that  $x \in (0, 1)$  and

$$\int_0^1 p(x) dx = \int_0^1 a \exp(-x) dx = 1.$$

resulting  $a = \frac{e}{e-1}$ .

$$\Rightarrow p(x) = \frac{\exp(-x)}{1 - \frac{1}{e}} \quad (40)$$

## ■ Monte Carlo - More on above examples

Also check whether  $p(x)$  fulfills the criteria for pdf. for this we find  $\frac{F(0)}{p(0)}$  and  $\frac{F(1)}{p(1)}$ . They have to be equal.  
Now

$$\frac{F(0)}{p(0)} = \frac{e}{e-1} = \frac{F(1)}{p(1)} \quad (41)$$

Since our pdf fulfills the criteria lets find  $y(x)$ . For this we perform then the change of variables (via the Cumulative distribution function)

$$y(x) = \int_0^x p(x') dx' = \int_0^x dx' \exp(-x') \times \frac{e}{e-1} \quad (42)$$

## ■ Monte Carlo - More on above examples

$$\Rightarrow y(x) = \left( \frac{e}{e-1} \right) \left( 1 - e^{-x} \right) \quad (43)$$

after solving for  $x$  we get;

$$\Rightarrow x = -\ln \left( 1 - y(1 - e^{-1}) \right) \quad (44)$$

which gives  $y = 0$  for  $x = 0$  and  $y = 1$  for  $x = 1$  as required for the property of pdf.

Now we need to find expectation value of

$$\left[ \frac{\exp(-x^2)}{\exp(-x)} \right]$$

with distribution  $y(x)$ . That is we need to evaluate

$$\int_0^1 \exp(-x^2) dx = \left\langle \frac{\exp(-y^2(x))}{\exp(-y(x))} \right\rangle$$

## ■ Monte Carlo - More on above examples

Algorithm:

1. Start
2. import required libraries like random, numpy ..
3. Define n, functions, initialize summ etc
4. start loop over n
5. generate random numbers  $x \in (0, 1)$
6. define function  $y(x)$  using above formula from x as  
$$y(x) = \left(\frac{e}{e-1}\right)(1 - e^{-x})$$
7. Get sum of the function  $\frac{\exp(-y^2(x))}{\exp(-y(x))}$
8. close the loop
9. Find the integration value i.e.  $\left\langle \frac{\exp(-y^2(x))}{\exp(-y(x))} \right\rangle$
10. You also find variance and hence the error.

The python code is in next page.

Please note that the code contains the simple sampling also. Compare both results.

# ■ Monte Carlo - More on above examples

```
In [3]: import numpy as np
import random
from scipy.stats import norm
#define the number of MC steps
n=10000

# Standard (simple sampling Monte Carlo
sum=0.0
i=0
summ=0.
while i<n:
    x = random.random()
    g = np.exp(-x**2)
    sum+=g
    summ+=g*g
    i=i+1
MC=sum/n
std_MC = summ/n-MC**2
print('Standard Monte-Carlo estimate of given function: ' + str(MC))
print('Standard deviation of simple sampling Monte Carlo: ' + str(std_MC))
print(' ')
#Importance sampling
i=0
sum=0.0
summ=0.0
while i<n:
    x = random.random()
    y=(np.exp(1.)/(np.exp(1.0)+1.)*(1-np.exp(-x)))
    f = np.exp(-y**2)/np.exp(-y)
    sum+=f
    summ+=f*f
    i=i+1
meanf=sum/n
MCI=meanf*(1.0-np.exp(-1.0))
std_MCI=summ/n-meanf**2
print('Importance Sampling Monte-Carlo estimate of given function: ' + str(MCI))
print('Standard deviation of Impostance sampling Monte Carlo: ' + str(std_MCI))
print(' ')
```

Standard Monte-Carlo estimate of given function: 0.7469875583049324  
Standard deviation of simple sampling Monte Carlo: 0.04042059388000929

Importance Sampling Monte-Carlo estimate of given function: 0.7442631266953907  
Standard deviation of Impostance sampling Monte Carlo: 0.00770205070157007

## ■ Importance Sampling - Examples

Now compare the variances OR errors due to simple sampling and importance sampling both.

(2) Consider the integral

$$I = \int_0^{\pi} \frac{1}{x^2 + \cos^2 x} dx \quad (45)$$

Evaluate  $I$  using importance sampling with  $p(x) = a \exp(-x)$  where  $a$  is a constant.

**Important Note:** You first write Algorithm then write code in python language following the Algorithm. Can you find the value of  $a$  which minimizes the variance.

## ■ Importance Sampling - Examples

Evaluate

$$\int_0^{10} \exp(-2|x-5|)dx \quad (46)$$

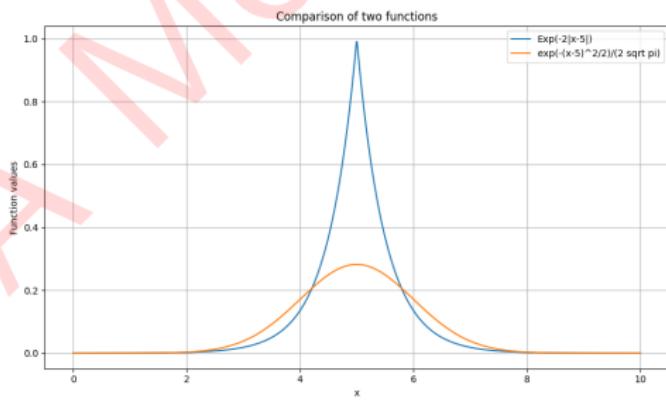
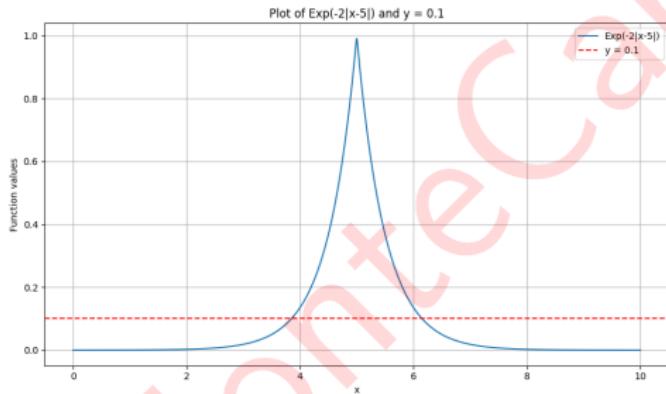
The true value of this integral is about 1. Simple way is Simple Sampling and integrating it.

However the error decreases if we take

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp(-(x-5)^2/2) \quad (47)$$

You plot the integrand and the density function  $p(x)$  defined above in the same plot. Explain why error decreases by choosing  $p(x)$ .

# ■ Importance Sampling - Examples



# ■ Monte Carlo Integration - Homework

Consider

$$I = \int_0^1 dx_1 \int_0^1 dx_2 \dots \int_0^1 dx_n g(x_1, x_2, \dots, x_n) \quad (48)$$

with  $x_i$  defined in the interval  $[a_i, b_i]$  we would typically need a transformation of variables of the form

$$x_i = a_i + (b_i - a_i) * t_i$$

if we were to use the uniform distribution on the interval  $[0, 1]$ .

As an example, evaluate

$$I = \int_{-5}^5 d\mathbf{x} d\mathbf{y} g(\mathbf{x}, \mathbf{y}) \quad (49)$$

with

$$g(\mathbf{x}, \mathbf{y}) = \exp(-\mathbf{x}^2 - \mathbf{y}^2 - (\mathbf{x} - \mathbf{y})^2/2)$$

Again you write Algorithm and code.

## ■ Monte Carlo Acceptance-Rejection method

It is simple and appealing method after von Neumann. Assume that we are looking at an interval  $x \in [a, b]$ , this being the domain of the PDF  $p(x)$ . Suppose also that the largest value our distribution function takes in this interval is  $M$ , that is

$$p(x) \leq M \quad x \in [a, b] \quad (50)$$

Then we generate a random number  $x$  from the uniform distribution for  $x \in [a, b]$  and a corresponding number  $s$  for the uniform distribution between  $[0, M]$ . If

$$p(x) \geq s \quad (51)$$

we accept the new value of  $x$ , else we generate again two new random numbers  $x$  and  $s$  and perform the test in the latter equation again.

## ■ Acceptance-Rejection method: an example

- Actually Acceptance-Rejection sampling is the conceptually simplest way to generate samples of some arbitrary probability function without having to do any transformations.
- No integration, no trickery, you simply trade computational efficiency away to keep everything as simple as possible.
- Consider an example:

$$f(x) = 1.2 - x^4$$

You want to sample points in the given function for  $x \in (0, 1)$ . Well, if you integrate  $f(x)$  between 0 and 1 you get 1.

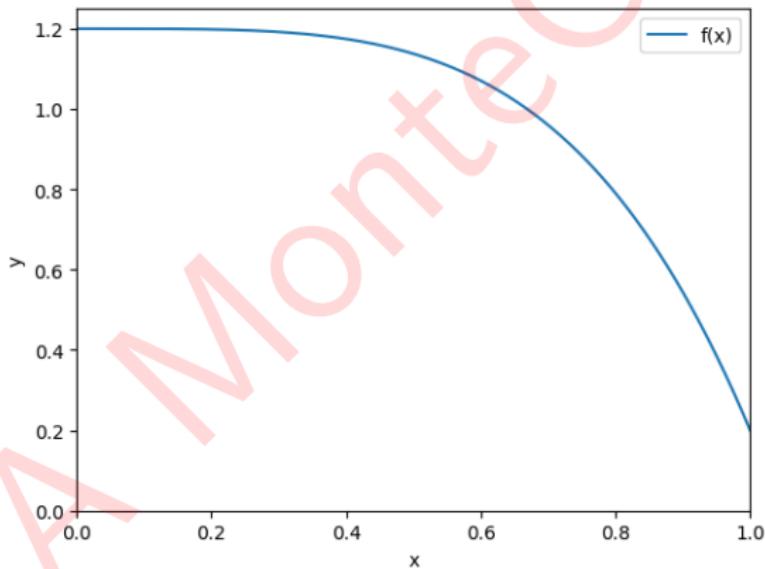
## ■ Acceptance-Rejection method: an example

Let's first plot the function just to see how does it look like

```
import numpy as np
import matplotlib.pyplot as plt
def f(x):
    return 1.2 - x**4
xs = np.linspace(0, 1, 1000)
ys = f(xs)

plt.plot(xs, ys, label="f(x)") plt.xlim(0, 1), plt.ylim(0, 1.25),
plt.xlabel("x"), plt.ylabel("y"), plt.legend();
```

## ■ Acceptance-Rejection method: an example



## ■ Acceptance-Rejection method: an example

- **Algorithm**
- Pick two random numbers. One for  $x$  (between 0 and 1), one for  $y$  (between 0 and 1.2).
- If the  $y$  value we randomly picked is less than  $f(x)$ , keep it, otherwise go back to above step

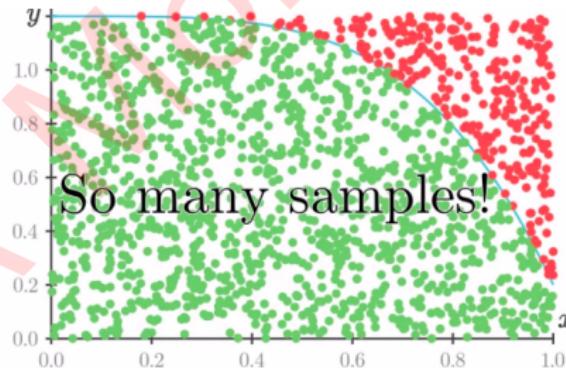


Figure: Green points accepted and red one rejected

## ■ Acceptance-Rejection method: an example

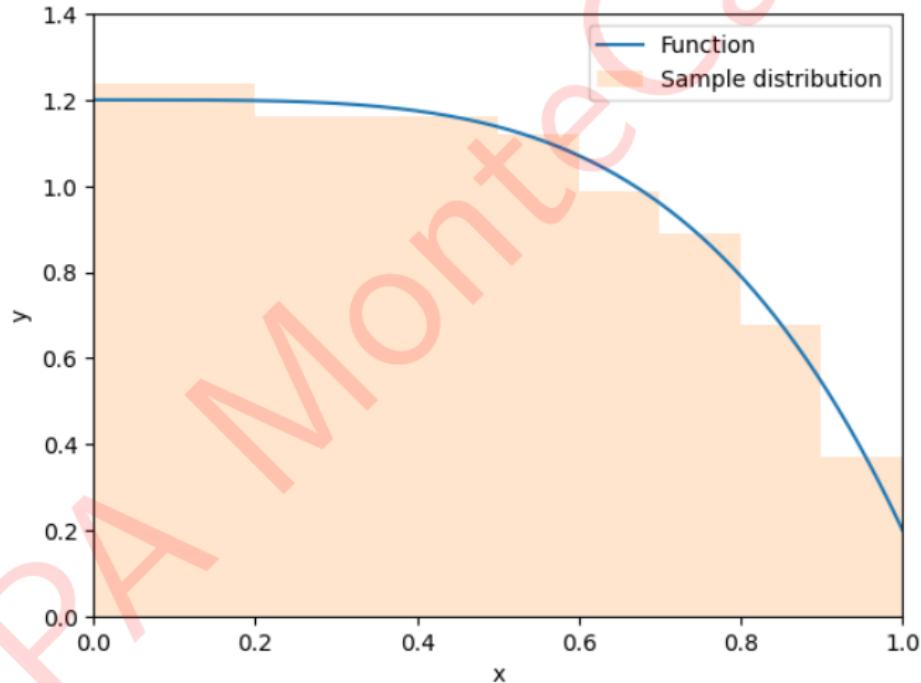
- So you can see that the reason this is so straightforward is that we get samples according to the function by simply throwing away the right number of samples when the function has a smaller value.
- In our function, this means if we get a small  $x$  value, we'd normally keep the sample (and indeed the distribution is pretty flat for  $x < 0.5$ ), but for values close to  $x = 1$ , we'd throw them out most of the time

## ■ Acceptance-Rejection method: an example

```
def sample(function, xmin=0, xmax=1, ymax=1.2):
    while True:
        x = np.random.uniform(low=xmin, high=xmax)
        y = np.random.uniform(low=0, high=ymax)
        if y < function(x):
            return x

samps = [sample(f) for i in range(10000)]
plt.plot(xs, ys, label="f(x)")
plt.hist(samps, density=True, alpha=0.2, label="Sample
distribution")
plt.xlim(0, 1), plt.ylim(0, 1.4), plt.xlabel("x"), plt.ylabel("y"),
plt.legend();
```

## ■ Acceptance-Rejection method: an example



# Markov Chain

Prof. Dr. Narayan Prasad Adhikari  
Central Department of Physics  
Tribhuvan University Kirtipur, Kathmandu, Nepal

February 22, 2025



# ■Contents

- Random variables
- Definition and Transition Probabilities
- Decomposition of the State Space
- Stationary distributions
- Limiting Theorems
- Reversible chains
- Continuous State Spaces

# ■ Random variables

- Consider an elementary event with a countable set of random outcomes,  $x_1, x_2, \dots, x_k$  (e.g. you can consider a rolling dice OR a set of "Khodkhode", for example you can consider the outcome of two dices after rolling )
- You are data scientist so you need to consider this event occurring repeatedly say  $N$  times such that  $N \ggg 1$  and we count how often the outcome  $x_k$  is observed ( $N_k$ ).
- The probabilities  $p_k$  for outcome  $x_k$  is

$$p_k = \lim_{N \rightarrow \infty} \left( \frac{N_k}{N} \right) \quad (1)$$

with  $\sum_k p_k = 1$ .

Obviously  $0 \leq p_k \leq 1$

You are familiar with conditional probability  $P(j/i)$ , average of any outcomes of such random events  $x_i$ , its variances and so on.

## ■ Notation

This lecture contains Markov Chain Stochastic Simulation so its about probabilities and statistics (I presume that you are familiar with probability density function).

- Distributions are identified with their density or probability functions.
- Variables are generally treated as if they are continuous.
- Posterior densities are denoted by  $\pi$  and their approximations by  $q$ .
- $x, y, \dots$  denote the observed quantities whereas unobserved quantities or parameters are denoted by Greek letters  $\theta, \phi, \dots$

## ■ Notation

- No distinctions between a random variable and its observed value
- Scalars and vectors both are denoted by small letters whereas capital letters represent matrices.
- The transpose of  $x$  is denoted by  $x'$ . Dimension of matrices is denoted by  $d$ .
- The component of  $A$  is  $\bar{A}$ .
- The probability of an event  $A$  is denoted by  $Pr(A)$ .
- Expectation and variance of a quantity  $x$  are  $E(x)$  and  $Var(x)$ .
- The covariance and correlation between random quantities  $x$  and  $y$  are  $Cov(x, y)$  and  $Cor(x, y)$
- The number of elements of a set  $A$  is denoted by  $\#A$
- . denotes approximations and they are with appropriate symbols

## ■Markov Chains - Introduction



Figure: Andrei Andreivich Markov - Russian Mathematician (1856-1922)

- Markov dependence is a concept attributed to the Russian mathematician Andrei Andreivich Markov that at the start of the 20th century investigated the alternance of vowels and consonants in the poem *Onegin* by Poeshkin
- He developed a probabilistic model where successive results depended on all their predecessors only through the immediate predecessor. The model allowed him to obtain good estimates of the relative frequency of vowels in the poem.

# ■Markov Chains - Introduction

- A Markov chain is a special type of stochastic process (random and usually dependent on time), which deals with characterization of sequences of random variables. Special interest is paid to the dynamic and the limiting behaviors of the sequence. A stochastic process can be defined as a collection of random quantities  $\{\theta^{(t)} : t \in T\}$  for some set  $T$ .
- The set  $\{\theta^{(t)} : t \in T\}$  for some set  $T$ , is said to be a stochastic process with state space  $S$  and index (or parameter) set  $T$ .  $T$  is taken to be countable, defining a discrete time stochastic process , i.e.  $T \in N$ , with  $N$  the set of natural numbers. State Space is the set of all possible and known states of a system. In state-space, each unique point represents a state of the system. For example, Take a pendulum moving in to and fro motion. Then its state is represented by its angle and angular velocity. Similarly consider rolling of two dices. Then the state space gives any of the outcomes  $\{2, 3, \dots, 12\}$  (see state space of

## ■ Random variables & PDF

- The state space will be a subset of  $R^d$  representing support of a parameter vector.
- Find a formula for the probability distribution of the total number of heads obtained in four tosses of a balanced coin.
- The sample space, probabilities and the value of the random variable ( $X$  where  $X$  is the number of heads obtained in four tosses) are given in table.

From the table we can determine the probabilities as

$$P(X = 0) = \frac{1}{16}, P(X = 1) = \frac{4}{16}, P(X = 2) = \frac{6}{16}, P(X = 3) = \frac{4}{16}, P(X = 4) = \frac{1}{16} \quad (2)$$

Notice that the denominators of the five fractions are the same and the numerators of the five fractions are 1, 4, 6, 4, 1. The numbers in the numerators is a set of binomial coefficients.

## ■ Random variables & PDF

$$\begin{aligned}\frac{1}{16} &= {}^4C_0 \frac{1}{16}, \quad \frac{4}{16} = {}^4C_1 \frac{1}{16} \\ \frac{6}{16} &= {}^4C_2 \frac{1}{16}, \quad \frac{4}{16} = {}^4C_3 \frac{1}{16}, \quad \frac{1}{16} = {}^4C_4 \frac{1}{16}\end{aligned}\tag{3}$$

We can then write the probability mass function as

# ■Random variables & PDF

TABLE 1. Probability of a Function of the Number of Heads from Tossing a Coin Four Times.

Element of sample space	Probability	Value of random variable X (x)
HHHH	1/16	4
HHHT	1/16	3
HHTH	1/16	3
HTHH	1/16	3
THHH	1/16	3
HHTT	1/16	2
HTHT	1/16	2
HTTH	1/16	2
THHT	1/16	2
THTH	1/16	2
TTHH	1/16	2
HTTT	1/16	1
THTT	1/16	1
TTHT	1/16	1
TTTH	1/16	1
TTTT	1/16	0

## ■Definition and Transition Probabilities

- A Markov chain is a stochastic process where given the present state, past and future states are independent.
- This property can be formally stated through

$$Pr(\theta^{(n+1)} \in A | \theta^{(n)} = x, \theta^{(n-1)} \in A_{n-1}, \dots, \theta^{(0)} \in A_0) \quad (4)$$

$$= Pr(\theta^{(n+1)} \in A | \theta^{(n)} = x) \quad (5)$$

for all sets  $A_0, A_1, \dots, A_{n-1}, A \in S$ . All these  $\theta$ 's may be random numbers from set  $\{2, 3, \dots, 12\}$  in case of rolling of two dices.

## ■ Definition and Transition Probabilities

- The state space in this case is

$$S = \left\{ (1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), \right. \\ \left. (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), \right. \\ \left. (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), \right. \\ \left. (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), \right. \\ \left. (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), \right. \\ \left. (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6) \right\}$$

## ■Definition and Transition Probabilities

Markovian property equation 5 can be rewritten as

$$E \left[ f(\theta^{(n)} | \theta^{(m)}, \theta^{(m-1)}, \dots, \theta^{(0)}) \right] = E \left[ f(\theta^{(n)} | \theta^{(m)}) \right] \quad (6)$$

for all bounded functions  $f$  and  $n > m \geq 0$ .

## ■ Definition and Transition Probabilities

Equivalently,

$$Pr(\theta^{(n+1)} = y | \theta^{(n)} = x, \theta^{(n-1)} = x_{n-1}, \dots, \theta^{(0)} = x_0) \quad (7)$$

$$= Pr(\theta^{(n+1)} = y | \theta^{(n)} = x) \quad (8)$$

for all  $x_0, x_1, \dots, x_{n-1}, x, y \in S$ . This form is obviously appropriate only for discrete state spaces.



- If a sequence of numbers follows the above graphical model, it is a Markov Chain.
  - That is,  $p(x_5 | x_4, x_3, x_2, x_1) = p(x_5 | x_4)$ .
  - So the probability of a certain state being reached, depends only on the previous state of the chain.

## ■Definition and Transition Probabilities

In general, the probabilities in 5 depend on  $x, A$  and  $n$ . When they do not depend on  $n$ , the chain is said to be homogeneous. In this case, a transition function or kernel  $P(x, A)$  can be defined as:

1. for all  $x \in S$ ,  $P(x, .)$  is a probability distribution over  $S$ ;
2. for all  $A \in S$ , the function  $x \mapsto P(x, A)$  can be evaluated.

It is also useful when dealing with discrete state space to identify  $P(x, \{y\}) = P(x, y)$ . This function is called a transition probability and satisfies:

- $P(x, y) \geq 0, \forall x, y \in S$ ;
- $\sum_{y \in S} P(x, y) = 1, \forall x \in S$ ; as any probability distribution  $P(x, .)$  should

## ■ Example 1: Random Walk

Consider a particle moving independently left and right on the line with successive displacements from its current position governed by a probability function  $f$  over the integers and  $\theta^{(n)}$  representing its position at instant  $n$ ,  $n \in N$ . Initially,  $\theta^{(0)}$  is distributed according to some distribution  $\pi^0$ . The positions can be related as

$\theta^{(n)} = \theta^{(n-1)} + w_n = w_1 + w_2 + \dots + w_n$  where the  $w_i$  are independent random variables with probability function  $f$ . So,  $\{\theta^{(n)} : n \in N\}$  is a Markov chain in  $Z$ .

## ■ Example 1: Random Walk

The position of the chain at instant  $t = n$  is described probabilistically by the distribution of  $w_1 + w_2 + \dots + w_n$ . If  $f(1) = p$ ,  $f(-1) = q$  and  $f(0) = r$  with  $p + q + r = 1$  then the transition probabilities are given by

$$P(x, y) = \begin{cases} p, & \text{if } y = x + 1 \\ q, & \text{if } y = x - 1 \\ r, & \text{if } y = x \\ 0, & \text{if } y \neq x, x - 1, x + 1 \end{cases} \quad (9)$$

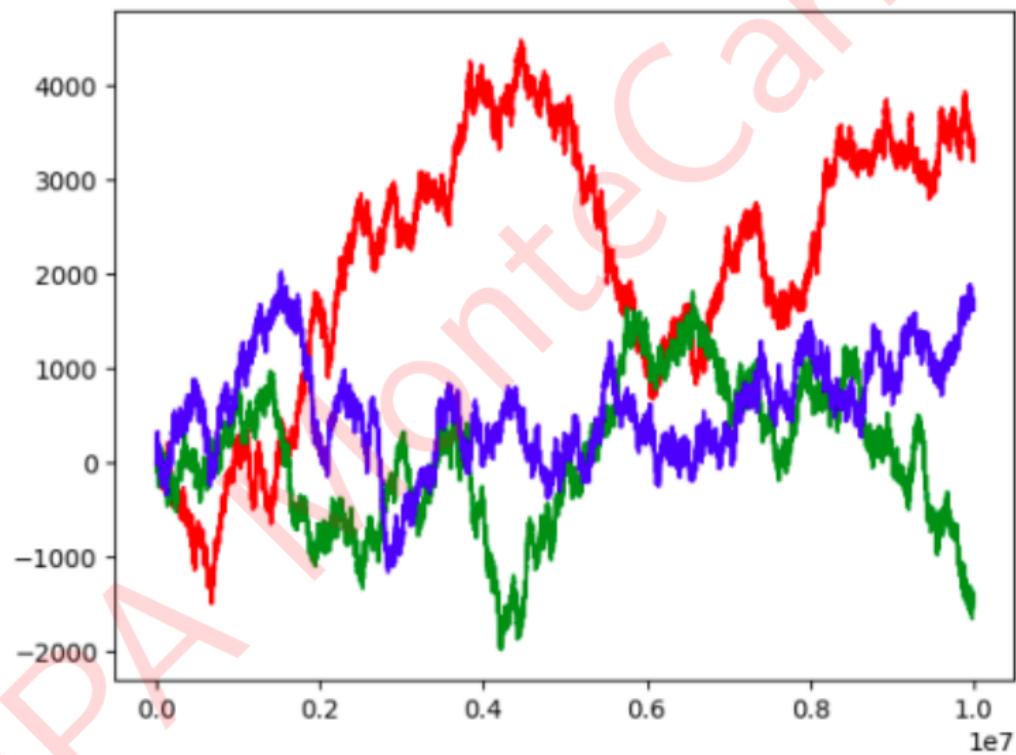
# ■ Example 1: Random Walk

- Consider a random walk - forward direction +1, backward direction -1 and remaining in the same position as 0 as described in previous page.
- We can use python code to generate a random walk. After n steps the displacement will be zero but root mean squared deviation (RMSD) will not be. After a large numbers of such walks the average displacement will be RMSD.

```
: import numpy as np
import matplotlib.pyplot as plt
import random
def rwID(n):
    x, y = 0, 0
    # Generate the time points [1, 2, 3, ..., n]
    timepoints = np.arange(n + 1)
    positions = [y]
    for i in range(1, n + 1):
        # Randomly select either UP or DOWN
        step = random.random()

        # Move the object up or down
        if step <= 0.5:
            y += 1
        elif step > 0.5:
            y -= 1
        # Keep track of the positions
        positions.append(y)
    return timepoints, positions
rw1 = rwID(10000000)
rw2 = rwID(10000000)
rw3 = rwID(10000000)
plt.plot(rw1[0], rw1[1], 'r-', label="rw1")
plt.plot(rw2[0], rw2[1], 'g-', label="rw2")
plt.plot(rw3[0], rw3[1], 'b-', label="rw3")
plt.show()
```

## ■ Example 1: Random Walk



## ■Definition and Transition Probabilities

**CW/HW:** You understand the meaning of  $P(x, y)$ . Also as an example discuss Random Walk problems in 1 and 2 dimensions. Students can write a python code for random walk in 1 & 2 dimensions.

## ■Markov chain Ehrenfest model - book

Consider a total of  $r$  balls distributed in two urns with  $x$  balls in the first urn and  $r - x$  in the second urn. Take one of the  $r$  balls at random and put it in the other urn. Repeat the random selection process independently and indefinitely. This procedure was used by Ehrenfest to model the exchange of molecules between two containers. If  $X^{(n)}$  represents the number of balls in the first urn after  $n$  exchanges then  $\{X^{(n)} : n \in N\}$  is a Markov chain with state space  $S = \{0, 1, 2, \dots, r\}$  and transition probabilities

$$P(x, y) = \begin{cases} x/d, & \text{if } y = x + 1 \\ 1 - x/d, & \text{if } y = x - 1 \\ 0, & \text{if } |y - x| \neq 1 \end{cases} \quad (10)$$

## ■ Simulation of Ehrenfest model (Explanation)

Gas molecules move about randomly in a box which is divided into two halves symmetrically by a partition. A hole is made in the partition. Suppose there are  $N$  molecules in the box. Think of the partitions as two urns containing balls labeled 1 through  $N$ . Molecular motion can be modeled by choosing a number between 1 and  $N$  at random and moving the corresponding ball from the urn it is presently in to the other. This is a historically important physical model introduced by Ehrenfest in the early days of statistical mechanics to study thermodynamic equilibrium.

The set of states of the Markov chain is  $S = \{0, 1, 2, \dots, N\}$  representing the number of molecules in one partition of the box.

## ■ Simulation of Ehrenfest model (HW)

Do a computer simulation of the Markov chain for  $N = 100$ . Start from state 0 (one of the partitions is empty) and follow the chain up to 1000 steps. Draw a graph of the number of molecules in the initially empty partition as a function of the number of steps. On the basis of your simulation, would you expect to observe during the course of the simulation a return to state 0?

It takes  $2^{200}$  steps to make a box empty or to observe state  $S = 0$ . If a molecule remains in a box for say 1/1000 seconds then it takes around  $10^{20}$  years!!! WOW?

## ■ Transition probabilities

In the case of discrete state spaces  $S = x_1, x_2, \dots$ , a transition matrix  $P$  with  $(i, j)$ th element given by  $P(x_i, x_j)$  can be defined. If  $S$  is finite with  $r$  elements, the transition matrix  $P$  is given by

$$P = \begin{pmatrix} P(x_1, x_1) & \dots & P(x_1, x_r) \\ \dots & \dots & \dots \\ \dots & \dots & \dots \\ P(x_r, x_1) & \dots & P(x_r, x_r) \end{pmatrix} \quad (11)$$

- Transition matrices have all lines summing to one. Such matrices are called stochastic and have a few interesting properties.
- For instance, at least one eigenvalue of a stochastic matrix equals one and the product of stochastic matrices always produces a stochastic matrix. Of course, countable state spaces will lead to an infinite number of eigenvalues.

## ■ Transition probabilities

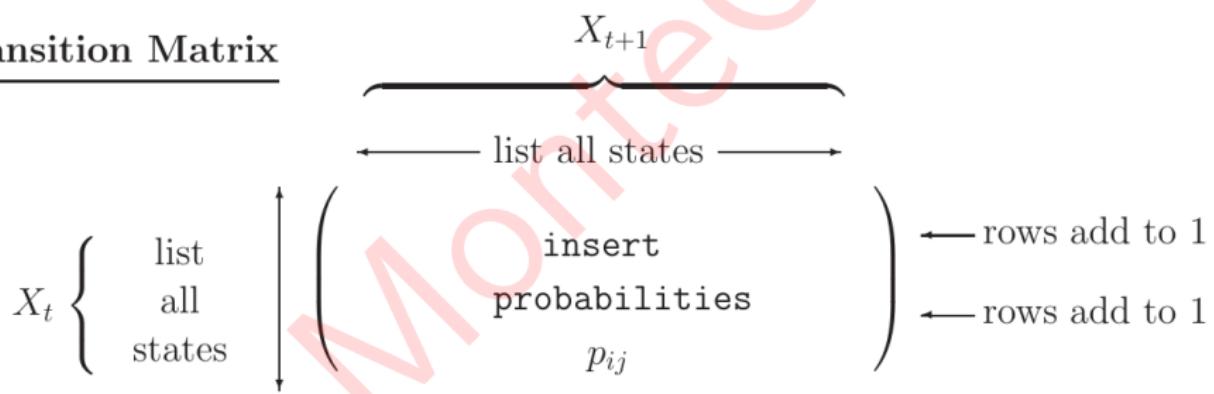
In the transition matrix P:

- the ROWS represent NOW, or FROM ( $X_t$ );
- the COLUMNS represent NEXT, or TO ( $X_{t+1}$ );
- entry  $(i, j)$  is the CONDITIONAL probability that NEXT =  $j$ , given that

NOW =  $i$ : the probability of going FROM state  $i$  TO state  $j$ .

# ■ Transition probabilities

Transition Matrix



# ■ Transition probabilities

Notes:

1. The transition matrix  $P$  must list all possible states in the state space  $S$ .
2.  $P$  is a square matrix ( $N \times N$ ), because  $X_{t+1}$  and  $X_t$  both take values in the same state space  $S$  (of size  $N$ ).
3. The rows of  $P$  should each sum to 1:

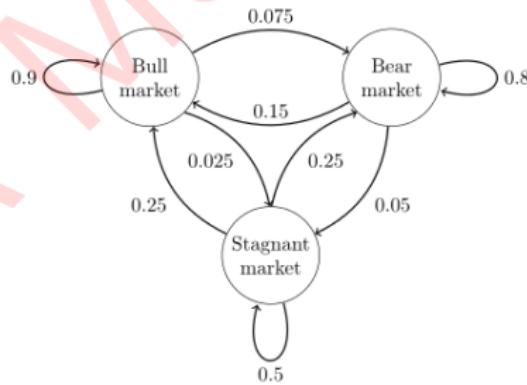
$$\sum_{j=1}^N p_{ij} = \sum_{j=1}^N \mathbb{P}(X_{t+1} = j \mid X_t = i) = \sum_{j=1}^N \mathbb{P}_{\{X_t = i\}}(X_{t+1} = j) = 1.$$

This simply states that  $X_{t+1}$  *must* take one of the listed values.

4. The columns of  $P$  do not in general sum to 1.

## ■ Transition probabilities

A state diagram for a simple example is shown in the figure using a directed graph to picture the state transitions. The states represent whether a hypothetical stock market is exhibiting a bull market, bear market, or stagnant market trend during a given week. According to the figure, a bull week is followed by another bull week 90% of the time, a bear week 7.5% of the time, and a stagnant week the other 2.5% of the time. Labeling the state space  $\{1 = \text{bull}, 2 = \text{bear}, 3 = \text{stagnant}\}$  the transition matrix for this example is



## ■ Transition probabilities

$$P = \begin{pmatrix} & Bull & Bear & stagnate \\ Bull & 0.9 & 0.075 & 0.025 \\ Bear & 0.15 & 0.8 & 0.05 \\ Stagnant & 0.25 & 0.25 & 0.5 \end{pmatrix} \quad (12)$$

The distribution over states can be written as a stochastic row vector  $x$  with the relation  $x^{(n+1)} = x^{(n)}P$ . So if at time  $n$  the system is in state  $x^{(n)}$ , then three time periods later, at time  $n + 3$  the distribution is

$$x^{n+3} = x^{n+2}P = x^{n+1}P^2 = x^n P^3 \quad (13)$$

If at time  $n$  the system is in state 2 (bear), then at time  $n + 3$  the distribution is

## ■ Transition probabilities

- Transition probabilities from state  $x$  to state  $y$  over  $m$  steps, denoted by  $P^m(x, y)$ , is given by the probability of a chain moving from state  $x$  to state  $y$  in exactly  $m$  steps. It can be obtained for  $m \geq 2$  as

$$\begin{aligned} P^m(x, y) &= Pr\left(\theta^{(m)} = y | \theta^{(0)} = x\right) \\ &= \sum_{x_1} \dots \sum_{x_{m-1}} Pr\left(\theta^{(m)} = y, \theta^{(m-1)} = x_{m-1}, \dots, \theta^{(1)} = x_1 | \theta^{(0)} = x\right) \\ &= \sum_{x_1} \dots \sum_{x_{m-1}} Pr\left(\theta^{(m)} = y, \theta^{(m-1)} = x_{m-1}\right) \dots Pr\left(\theta^{(1)} = x_1 | \theta^{(0)} = x\right) \\ &= \sum_{x_1} \dots \sum_{x_{m-1}} P(x, x_1)P(x_1, x_2)\dots P(x_{m-1}, y) \end{aligned} \tag{14}$$

where the second equality is due to the Markovian property of the process.

## ■ Transition probabilities

- The last equality means that the matrix containing elements  $P_m(x, y)$  is also a stochastic matrix and is given by  $P^m$  obtained by the matrix product of the transition matrix  $P$   $m$  times. Also, for completeness,  $PI(x, y) = P(x, y)$  and  $P^0(x, y) = I(x = y)$ .

The above derivation can be used to establish that

$$\begin{aligned} & P^{n+m}(x, y) \\ &= \sum_z Pr(\theta^{(n+m)} = y | \theta^{(n)} = z, \theta^{(0)} = x) Pr(\theta^{(n)} = z, \theta^{(0)} = x) \\ &= \sum_z P^n(x, z) P^m(z, y) \end{aligned} \tag{15}$$

## ■ Transition probabilities

Equations 15 are usually called Chapman-Kolmogorov equations. All summations are with respect to the elements of the state space  $S$  and results are valid for any stage of the chain due to the assumed homogeneity. Higher transition matrices can be formed with these higher transition probabilities and it can be shown that they satisfy the relation  $P^{n+m} = P^n P^m$  and, in particular,  $P^{n+1} = P^n P$ .

## ■ Transition probabilities

The marginal distribution (marginal distribution describes the probability distribution of a single variable from a set of related variables, ignoring the other variables) of the  $n$ th stage can be defined by the row vector  $\pi^{(n)}$  with components  $\pi^{(n)}(x_i)$ , for all  $x_i \in S$ . For finite state spaces, this is a  $r$ -dimensional vector

$$\pi^{(n)} = (\pi^{(n)}(x_1), \dots, \pi^{(n)}(x_r)) \quad (16)$$

When  $n = 0$ , this is the initial distribution of the chain. Then

$$\begin{aligned} \pi^{(n)}(y) &= Pr(\theta^{(n)} = y) \\ &= \sum_{x \in S} Pr(\theta^{(n)} = y | \theta^{(0)} = x) Pr(\theta^{(0)} = x) \\ &= \sum_{x \in S} P^n(x, y) \pi^{(0)}(x) \end{aligned} \quad (17)$$

## ■ Transition probabilities

The above equation can be written in matrix notation as

$\pi^{(n)} = \pi^{(0)} P^n$ . Also, since the same is valid for  $n - 1$ ,

$$\pi^{(n)} = \pi^{(0)} P^{n-1} P = \pi^{(n-1)P}.$$

The probability of any event  $A \in S$  for a Markov chain starting at  $x$  is denoted by  $Pr_x(A)$ . The hitting time of  $A$  is defined as

$$T_A = \min\{n \geq 1 : \theta^{(n)} \in A\} \text{ if } \theta^{(n)} \in A \text{ for some } n > 0.$$

Otherwise,  $T_A = \infty$ . If  $A = \{a\}$ , the notation  $T_{\{a\}} = T_a$  is used.

# What is a Transition Probability Matrix?

- Describes the probabilities of transitioning from one state to another.
- Square matrix with non-negative entries.
- Each row sums to 1.

# Example: Weather Forecasting

- States:
  - Sunny (S)
  - Rainy (R)
- Transition probabilities:
  - If today is sunny:
    - 70% chance of sunny tomorrow
    - 30% chance of rainy tomorrow
  - If today is rainy:
    - 40% chance of sunny tomorrow
    - 60% chance of rainy tomorrow

# Transition Probability Matrix

The transition probability matrix is:

$$P = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$$

Where:

- $P_{11} = 0.7$ : Sunny  $\rightarrow$  Sunny
- $P_{12} = 0.3$ : Sunny  $\rightarrow$  Rainy
- $P_{21} = 0.4$ : Rainy  $\rightarrow$  Sunny
- $P_{22} = 0.6$ : Rainy  $\rightarrow$  Rainy

## Probability After Two Days

If today is sunny, the probability distribution after two days is calculated as:

$$\text{Initial State} = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

$$\text{After 2 days} = \begin{bmatrix} 1 & 0 \end{bmatrix} P^2$$

Where  $P^2$  is the matrix product of  $P$  with itself.

# Applications

- Weather prediction
- Customer behavior analysis
- Stock market analysis
- Biological systems (e.g., gene transitions)

# Conclusion

- Transition probability matrices model dynamic systems.
- Useful for predicting state transitions over time.
- Widely used in various fields including science and economics.

# ■ Decomposition of the State Space

A few quantities of interest are important in the classification of states of a Markov chain with state space  $S$  and transition matrix  $P$ :

- (i) The probability of the chain starting from state  $x$  hitting state  $y$  at any posterior step is  $\rho_{xy} = Pr_x(T_y < \infty)$  where  $T_y$  is the hitting time to  $y$  state;
- (ii) The number of visits of a chain to a state  $y$  is

$$N(y) = \#\{n > 0 : \theta^{(n)} = y\} = \sum_{n=1}^{\infty} I(\theta^{(n)} = y)$$

where  $I$  is indicator function

$$I(x \in A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

**HW:** Show that  $E(T_y | \theta^0 = x) = \sum_{n=0}^{\infty} Pr_x(T_y > n)$  and  $E(N(y) | \theta^0 = x) = \sum_{n=0}^{\infty} P^n(x, y)$ .

## ■ Decomposition of the State Space

A state  $y \in S$  is said to be recurrent if the Markov chain, starting in  $y$ , returns to  $y$  with probability 1 ( $\rho_{yy} = 1$ ) and is said to be transient if it has positive probability of not returning to  $y$  ( $\rho_{yy} < 1$ ). An absorbing state  $y \in S$  is recurrent because

$$Pr_y(T_y = 1) = Pr_y(\theta^{(1)} = y) = P(y, y) = 1$$

and therefore ( $\rho_{yy} = 1$ ).

If a Markov chain starts at a recurrent state  $y$ , the hitting (or return, in this case) time of  $y$ ,  $T_y$ , is a finite random quantity whose mean  $\mu_y$  can be evaluated. If this mean is finite, the state  $y$  is said to be positive recurrent and otherwise the state is said to be null recurrent. Positive recurrence is a very important property for establishing limiting results (Next lecture).

## ■ Decomposition of the State Space: Recurrent and Transient state

*Recurrent state:* A state in a Markov chain where, if you start there, you are guaranteed to return to it at least once in the future (probability of returning = 1).

*Transient state:* A state in a Markov chain where, if you start there, there's a chance you might never return (probability of returning < 1).

Analyzing these states helps understand the long-term behavior of Markov chains.

## ■ Decomposition of the State Space

An important result describing analytically the difference between a recurrent and a transient state is that

- if  $y \in S$  is a transient state then, for all  $x \in S$ ,

$$Pr_x(N(y) < \infty) = 1 \text{ and } E[N(y)|\theta^{(0)} = x] = \frac{\rho_{xy}}{1 - \rho_{yy}} < \infty$$

- if  $y \in S$  is a recurrent state then, for all  $x \in S$ ,

$$Pr_x(N(y) = \infty) = 1 \text{ and } E[N(y)|\theta^{(0)} = y] = \infty$$

So, recurrent states are infinitely often (i.o.) visited with probability one.

The expected number of visits is finite if the state is transient.

## ■ Decomposition of the State Space

It is interesting to investigate possible decompositions of  $S$  in subsets of recurrent and transient states. From this decomposition, probabilities of the chain hitting a given set of states can be evaluated. For states  $x$  and  $y$  in  $S$ ,  $x \neq y$ ,  $x$  is said to hit  $y$ , denoted  $x \rightarrow y$ , if  $\rho_{xy} > 0$ . A set  $C \subseteq S$  is said to be closed if  $\rho_{xy} = 0$  for  $x \in C$  and  $y \notin C$ .

In obvious nomenclature, it is said to be irreducible if  $x \rightarrow y$  for every pair  $x, y \in C$ . A chain is said to be irreducible if  $S$  is irreducible. An irreducible Markov chain is one where all states communicate with each other. This means that, starting from any state in the chain, it's possible to eventually reach any other state with a non-zero probability of transition, no matter how many steps it may take.

## ■ Decomposition of the State Space

It is not difficult to show that the condition  $\rho_{xy} > 0$  is equivalent to  $P^n(x, y) > 0$  for some  $n > 0$ . This can be used to show that if  $x \in S$  is recurrent and  $x \rightarrow y$  then  $y$  is also recurrent. In this case,  $y \rightarrow x$  and one can write  $x \leftrightarrow y$  when  $x \rightarrow y$  and  $y \rightarrow x$ . In other words, recurrence defines an equivalence class with respect to the  $\leftrightarrow$  operation. Also,  $\rho_{xy} = \rho_{yx} = 1$ . In fact, a stronger result is valid: null recurrence and positive recurrence also define equivalence classes. If  $C \subseteq S$  is a closed, finite, irreducible set of states then all states of  $C$  are recurrent.

# Transition Matrix

Consider the transition matrix:

$$P = \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0.2 & 0.3 & 0.5 \\ 0 & 0 & 1 \end{bmatrix}$$

We analyze the recurrence and transience of each state.

# Transition Matrix

Consider the transition matrix:

$$P = \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0.2 & 0.3 & 0.5 \\ 0 & 0 & 1 \end{bmatrix}$$

## P Squared ( $P * P$ )

$$P^2 = P \cdot P = \begin{bmatrix} 0.35 & 0.4 & 0.25 \\ 0.16 & 0.25 & 0.59 \\ 0 & 0 & 1 \end{bmatrix}$$

This matrix represents the probabilities of transitioning between states in *two steps*.

## P Cubed ( $P * P * P$ )

$$P^3 = P \cdot P \cdot P = \begin{bmatrix} 0.255 & 0.305 & 0.44 \\ 0.132 & 0.197 & 0.671 \\ 0 & 0 & 1 \end{bmatrix}$$

This matrix represents the probabilities of transitioning between states in *three* steps.

# State 1 Analysis

Transition probabilities for State 1:

$$P_{11}^1 = 0.5, \quad P_{11}^2 = 0.35, \quad P_{11}^3 = 0.255, \dots$$

General pattern:

$$P_{11}^n = 0.5 \times (0.7)^{n-1}$$

Summing the series:

$$\sum_{n=1}^{\infty} P_{11}^n = \frac{0.5}{1 - 0.7} = \frac{5}{3} \approx 1.67$$

Since the sum is finite, **State 1 is transient.**

## State 2 Analysis

Transition probabilities for State 2:

$$P_{22}^1 = 0.3, \quad P_{22}^2 = 0.25, \quad P_{22}^3 = 0.197, \dots$$

General pattern:

$$P_{22}^n = 0.3 \times (0.833)^{n-1}$$

Summing the series:

$$\sum_{n=1}^{\infty} P_{22}^n = \frac{0.3}{1 - 0.833} = \frac{0.3}{0.167} \approx 1.8$$

Since the sum is finite, **State 2 is transient.**

# State 3 Analysis

Transition probabilities for State 3:

$$P_{33}^1 = 1, \quad P_{33}^2 = 1, \quad P_{33}^3 = 1, \dots$$

General pattern:

$$P_{33}^n = 1 \text{ for all } n$$

Summing the series:

$$\sum_{n=1}^{\infty} P_{33}^n = \sum_{n=1}^{\infty} 1 = \infty$$

Since the sum is infinite, **State 3 is recurrent.**

# Summary of Results

- **State 1:** Sum =  $\frac{5}{3}$  (Finite)  $\Rightarrow$  Transient
- **State 2:** Sum  $\approx 1.8$  (Finite)  $\Rightarrow$  Transient
- **State 3:** Sum =  $\infty$  (Infinite)  $\Rightarrow$  Recurrent

# Conclusion

**State 1 and State 2 are Transient. State 3 is Recurrent.** This is concluded from the expected number of visits being finite or infinite.

# ■ Examples of Recurrent and Transient states

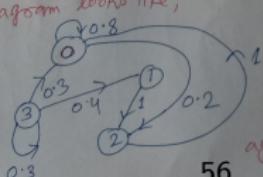
Example of Transient and Recurrent states

$$P_{xx} \equiv P_{x|x} = \sum_{n=1}^{\infty} P_{x|x}^{(n)} = \text{Prob. of recurrence to } x.$$

Then a state  $x$  is recurrent if  $P_{xx} = 1$  and a state  $x$  is transient if  $P_{xx} < 1$ . As discussed before, if  $x \leftrightarrow y$  and  $x$  is recurrent  $\Rightarrow y$  recurrent.

Similarly, if  $x \leftrightarrow y$  and  $x$  is transient  $\Rightarrow y$  transient.

Consider four states (say) Bhatbhatri, Bigmant, Salewars and Danaz represented by states 0, 1, 2 and 3. The transition diagram looks like;



This means a person if shopping in Bhatbhatri, the prob that he goes to bhatbhatri again is 80%.

# ■ Examples of Recurrent and Transient states

(S) He goes to Salway  $\otimes$  (probability) is 20%, and 80 cm.

$P = \begin{pmatrix} \text{BBM} & \text{Bigant} & \text{Salway} & \text{Dmz} \\ \text{BBM} & 0.8 & 0 & 0.2 \\ \text{Bigant} & 0 & 0 & 1 \\ \text{Salway} & 1 & 0 & 0 \\ \text{Dmz} & 0.3 & 0.5 & 0.3 \end{pmatrix}$

To check whether Bhatbhateni is recurrent state.

$$\text{For this find prob. } P_{xx} = P_{xx} = \sum_{n=1}^{\infty} P_{xx}^{(n)}$$

$$\therefore \text{here } P_{x0}^{(1)} = 0.8$$

$$P_{x0}^{(2)} = 0.2 \times 1 = 0.2$$

$$P_{x0}^{(3)} = 0$$

$$\therefore P_{x0} = P_0 = P_{x0}^{(1)} + P_{x0}^{(2)} + P_{x0}^{(3)}$$

$$= 1.0 \quad \therefore \text{Bhatbhateni}$$

is recurrent State.

Consider Bigant:  $P_{x1}^{(1)} = 0$

$$P_{x1}^{(2)} = 0$$

$$P_{x1}^{(3)} = 0$$

$$\sum_{n=1}^{\infty} P_{x1}^{(n)} = 0 \Rightarrow P_{x1} = 0 \leftarrow 1 \text{ means}\right.$$

Transient State.

# ■ Decomposition of the State Space

## - Example 4.6 (Textbook)

AS,  $0 \leftrightarrow 2$  and  $0$  is recurrent.  
 $2$  i.e. sideways is also recurrent.

Data:  $P_{03}^{(1)} = 0.3$ ;  $P_{33}^{(2)} = 0 = P_{33}^{(3)}$

$\therefore P_{03} = 0.3 + 0 = 0.3 < 1$   
 $\Rightarrow 3$  also recurrent state.

NOW our aim is to decompose the state space into Recurrent and Transient States.

Consider example from Text book (4.6)

Consider  $S = \{0, 1, 2, \dots, T\}$ ,  
and  $P = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 \\ 1 & 1/2 & 1/4 & 1/4 \\ 2 & 0 & 1/3 & 2/3 \end{pmatrix}$ . Now represent the transition matrix by transition state diagram.

NOW represent the pair  $(x, y)$  by  $+ \text{ or } -$  such that  
+ means  $x \rightarrow y$   
- means  $x \not\rightarrow y$ .

# ■ Decomposition of the State Space

## - Example 4.6 (Textbook)

Then we get;

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{pmatrix} \Rightarrow \begin{pmatrix} + & + & + \\ + & + & + \\ + & + & + \end{pmatrix}$$

This means all the states  $S = \{0, 1, 2\}$  are recurrent. Hence  $S_R = S_R$ .

Now you prove that

$$(1, 0, 0) P^2 = ((1, 0, 0) P) P > 0.$$

Actually  $0 \rightarrow 2$  after operating

$(1, 0, 0)$  by  $P^2$  and seeing that the prob. to state 2 is non zero.

$$\overbrace{(1 \ 0 \ 0) \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{pmatrix}}^{\text{represents state } 0} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

$$\text{Again, } \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{pmatrix} = \frac{1}{4} + \frac{1}{4} \quad \frac{1}{4} + \frac{1}{8} \quad \frac{1}{8}$$

$$= \left( \frac{1}{2}, \frac{5}{8}, \frac{1}{8} \right) \text{ means } 0 \rightarrow 2.$$

## ■ Example 4.6(b)

(b)

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1/4 & 1/2 & 1/4 & 0 & 0 & 0 \\ 0 & 1/5 & 2/5 & 1/5 & 0 & 1/5 \\ 0 & 0 & 0 & 1/6 & 1/3 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1/4 & 0 & 3/4 \end{pmatrix} \quad (18)$$

CW: Draw the transition probability diagram as in previous case. Then write above matrix in terms of + and - as before.

$$P = \begin{pmatrix} + & - & - & - & - & - \\ + & + & + & + & + & + \\ + & + & + & + & + & + \\ - & - & - & + & + & + \\ - & - & - & + & + & + \\ - & - & - & + & + & + \end{pmatrix} \quad (19)$$

## ■ Example 4.6(b)

From above figures we get  $S_R = \{0\} \cup \{3, 4, 5\}$  and  $S_T = \{1, 2\}$ .

# ■ Decomposition of the State Space

## -Example

**Birth and Death Processes** Consider a Markov chain that from the state  $x$  can only move in the next step to one of the neighboring states  $x - 1$ , representing a death,  $x$  or  $x + 1$ , representing a birth. The transition probabilities are given by

$$P(x, y) = \begin{cases} p_x, & \text{if } y = x + 1 \\ q_x, & \text{if } y = x - 1 \\ r_x, & \text{if } y = x \\ 0, & \text{if } |y - x| > 1 \end{cases}$$

where  $p_x$ ,  $q_x$ , and  $r_x$  are non-negative with  $p_x + q_x + r_x = 1$  and  $q_0 = 0$ . Note also that Ehrenfest model is special case of birth and death processes. Irreducible chains are obtained when  $p_x > 0$  for  $x \geq 0$  and  $p_x > 0$  for  $x > 0$ .

## ■ Decomposition of the State Space -Example

It is possible to determine if a state  $y$  is recurrent or transient even for an infinite state space by studying the convergence of the series  $\sum_{y=0}^{\infty} \gamma_y$  where

$$\gamma_y = \begin{cases} 1 & \text{if } y = 0 \\ \frac{q_1 \dots q_y}{p_1 \dots p_y} & \text{if } y > 0 \end{cases}$$

If the sum diverges, the chain is recurrent. Otherwise, the chain is transient.

If  $S$  is finite and 0 is an absorbing state, the absorption probability is

$$\rho\{0\}(x) = \rho_{x0} = \frac{\sum_{y=x}^{d-1} \gamma_y}{\sum_{y=0}^{d-1} \gamma_y}, \quad x = 1, \dots, d-1. \quad (20)$$

This example just discussed is optional.

## ■Stationary (Equilibrium) Distributions

A fundamental problem for Markov chains in the context of simulation is the study of the asymptotic behavior of the chain as the number of steps or iterations  $n \rightarrow \infty$ . A key concept is that of a stationary distribution  $\pi$ . A distribution  $\pi$  is said to be a stationary distribution of a chain with transition probabilities  $P(x, y)$  if

$$\sum_{x \in S} \pi(x)P(x, y) = \pi(y), \quad \forall y \in S \quad (21)$$

Equation 21 can be written in matrix form as

$$\pi P = \pi \quad (22)$$

The reason of the name is clear from the above equation. If the marginal distribution at any given step  $n$  is  $\pi$  then the distribution at the next step is  $\pi P = \pi$ .

## ■Stationary (Equilibrium) Distributions

Once the chain reaches a stage where  $\pi$  is the distribution of the chain, the chain retains this distribution for all subsequent stages. This distribution is also known as the invariant or equilibrium distribution for similar interpretations.

One can show that if the stationary distribution  $\pi$  exists and  $\lim_{n \rightarrow \infty} P^n(x, y) = \pi(y)$  then, independently of the initial distribution of the chain,  $\pi^{(n)}$  will approach  $\pi$  as  $n \rightarrow \infty$ . In this sense, the distribution is also referred to as the limiting distribution.

## ■Equilibrium Distributions- Example

Consider  $\{\theta^{(n)} : n \geq 0\}$ , a Markovian chain in  $S = \{0, 1\}$  with initial distribution  $\pi^{(0)}$  given by  $\pi^{(0)} = (\pi^{(0)}(0), \pi^{(0)}(1))$  and transition matrix  $P$  given by

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} \quad (23)$$

The stationary distribution  $\pi$  is the solution of the system  $\pi P = \pi$  that gives the equations

$$\pi(0)P(0, y) + \pi(1)P(1, y) = \pi(y), \quad y = 0, 1 \quad (24)$$

The solution is  $\pi = (q, p)/(p + q)$ , a distribution that can be shown to be invariant for the stages of the chain.

**CW/HW: Prove above solution. Also write a python code to get numerical values of  $\pi$  for given  $q=0.5$ ,  $p=0.5$ . Also try different values of  $p$  and  $q$ . Do you still get stationary values of  $\pi$  for  $(p + q) = 2$ ?**

## ■Equilibrium Distributions

```
In [1]: import numpy as np

def get_stationary(n):
    row = n
    pi = np.full((1, row), 1 / row)
    P = np.array([[1/4, 1/2, 1/4],
                  [1/3, 0, 2/3],
                  [1/2, 0, 1/2]])
    while True:
        new_pi = np.dot(pi, P)
        if np.allclose(pi, new_pi):
            return pi
        break
    pi = new_pi
print(get_stationary(3))
```

```
[[0.37500019 0.18750166 0.43749814]]
```

# ■Equilibrium Distributions

Let us consider a stochastic process at discrete times labeled consecutively  $t_1, t_2, t_3, \dots$ , for a system with a finite set of possible states  $S_1, S_2, S_3, \dots$ , and we denote by  $X_t$  the state the system is in at time  $t$ . We consider the conditional probability that  $X_{t_n} = S_{i_n}$ ,

$$P(X_{t_n} = S_{i_n} | X_{t_{n-1}} = S_{i_{n-1}}, X_{t_{n-2}} = S_{i_{n-2}}, \dots, X_{t_1} = S_{i_1}) \quad (25)$$

given that at the preceding time the system state  $X_{t_{n-1}}$  was in state  $S_{i_{n-1}}$ , etc. Since this process is Markov process the conditional probability is in fact independent of all states but the immediate predecessor i.e.  $P = P(X_{t_n} = S_{i_n} | X_{t_{n-1}} = S_{i_{n-1}})$ .

Above conditional probability can be interpreted as the transition probability to move from state  $i$  to state  $j$ ,

$$P_{ij} = P(S_i \rightarrow S_j) = P(X_{t_n} = S_j | X_{t_{n-1}} = S_i) \quad (26)$$

## ■Equilibrium Distributions and Master Equation

We further require that  $P_{ij} \geq 0$ ,  $\sum_j P_{ij} = 1$  as usual for the transition probabilities. We may then construct the total probability  $P(X_{t_n} = S_j)$  that at time  $t_n$  the system is in state  $S_j$  as

$$\begin{aligned} P(X_{t_n} = S_j) &= P(X_{t_n} = S_j | X_{t_{n-1}} = S_i)P(X_{t_{n-1}} = S_i) \\ &= P_{ij}P(X_{t_{n-1}} = S_i) \end{aligned} \tag{27}$$

The master equation considers the change of this probability with time  $t$  (treating time as continuous rather than discrete variable and writing then  $P(X_{t_n} = S_j) = P(S_j, t)$ )

## ■Equilibrium Distributions and Master Equation

$$\frac{dP(S_j, t)}{dt} = - \sum_i P_{ij} P(S_j, t) + \sum_i P_{ij} P(S_i, t) \quad (28)$$

This equation also represents continuity of probability as it says total probability  $\sum_j P(S_j, t) = 1$  at all times. All probability of a state  $i$  that is 'lost' by transition to state  $j$  is gained in the probability of that state, and vice versa.

Basic property of Markov process: knowledge of state at time  $t$  completely determines the future time evolution. The main significance of equation 28 is that the importance sampling Monte Carlo process can be interpreted as a Markov process, with a particular choice of transition probabilities.

## ■Equilibrium Distributions and Master Equation

Hence for equilibrium probability  $P_{eq}$  we get

$$P_{ji}P_{eq}(S_j) = P_{ij}P_{eq}(S_i) \quad (29)$$

called *detailed balance equation*. This means the master equation yields

$$\frac{dP_{eq}(S_j, t)}{dt} \equiv 0 \quad (30)$$

This equation ensures that gain and loss terms in equation 28 cancel exactly.

## ■Equilibrium Distributions: Example - Gibbs Sampler

This example provides a very simple special case of the Gibbs sampler. The complete form of the Gibbs sampler will be our next topics, here let us consider a simple and special case of Gibbs Sampler. In this special case, the state space is  $S = \{0, 1\}$  and define a probability distribution  $\pi$  over  $S$  as

		$\theta_2$
$\theta_1$	0	1
0	$\pi_{00}$	$\pi_{01}$
1	$\pi_{10}$	$\pi_{11}$

The probability vector  $\pi$  contains the above probabilities in any fixed order, say  $(\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$ .

The chain now consists of a bidimensional vector  $\theta^{(n)} = (\theta_1^{(n)}, \theta_2^{(n)})$ . Although this introduces some novelties in the presentation they can

## ■Equilibrium Distributions: Example - Gibbs Sampler

be removed by considering a scalar chain  $\psi^{(n)}$  that assumes values that are in correspondence with the  $\theta^{(n)}$  chain, e.g.

$\psi^{(n)} = 10 \theta_1^{(n)} + \theta_2^{(n)}$ . This is always possible for discrete state spaces. Therefore we do not make any distinction between scalar and vector chains.

Consider the following transition probabilities:

- For the first component  $\theta_1$ , the transition probabilities are given by the conditional distribution  $\pi_1$  of  $\theta_1 | \theta_2 = j$ ,

$$\pi_1(0|j) = \frac{\pi_{0j}}{\pi_{+j}} \text{ and } \pi_1(1|j) = \frac{\pi_{1j}}{\pi_{+j}}$$

where  $\pi_{+j} = \pi_{0j} + \pi_{1j}, j = 0, 1$ .

- For the second component  $\theta_2$ , the transition probabilities are given by the conditional distribution

## ■Equilibrium Distributions: Example - Gibbs Sampler

$\pi_2$  of  $\theta_2 | \theta_1 = i$ ,

$$\pi_2(0|i) = \frac{\pi_{i0}}{\pi_{i+}} \text{ and } \pi_2(1|i) = \frac{\pi_{i1}}{\pi_{i+}}$$

where  $\pi_{i+} = \pi_{i0} + \pi_{i1}, i = 0, 1$ .

The overall transition probability of the chain is

$$\begin{aligned} P((i,j), (k,l)) &= Pr(\theta^{(n)} = (k,l) | \theta^{(n-1)} = (i,j)) \\ &= Pr(\theta_2^{(n)} = l | \theta_1^{(n)} = i) Pr(\theta_1^{(n)} = k | \theta_1^{(n)} = j) \\ &= \frac{\pi_{kl}}{\pi_{k+}} \frac{\pi_{kj}}{\pi_{+j}} \end{aligned} \tag{31}$$

for  $(i,j), (k,l) \in S$ . Thus a  $4 \times 4$  matrix  $P$  can be formed.

## ■Equilibrium Distributions

The existence and uniqueness of stationary distributions can be studied through weaker results. Let  $N_n(y)$  be the number of visits to state  $y$  in  $n$  steps and define  $G_n(x, y) = E_x[N_n(y)]$ , the average number of visits of the chain to state  $y$  and  $m_y = E_y(T_y)$ , the average return time to state  $y$ . Then,  $G_n(x, y) = \sum_{k=1}^n P^k(x, y)$  and  $\lim_{n \rightarrow \infty} G_n(x, y)/n$  provides the limiting occupation of state  $y$  in a chain observed for an infinitely long number of steps.

## ■ Limiting Theorems

- There are situations where stationary distributions are available but limiting distributions are not (See above examples).
- In order to establish limiting results, one characterization of states still absent and that must be introduced. This is the notion of periodicity.
- The period of a state  $x$ , denoted by  $d_x$  is the largest common divisor of the set

$$\{n \geq 1 : P^n(x, x) > 0\}$$

It is obvious that  $P(x, x) > 0$  implies that  $d_x = 1$  and that if  $x \leftrightarrow y$  then  $d_x = d_y$ . Therefore, the states of an irreducible chain have the same period.

## ■ Limiting Theorems

- **Aperiodic state:** A state  $x$  is aperiodic if  $d_x = 1$
- **Ergodic state:** An aperiodic and positive recurrent state is said to be ergodic state.
- A chain is periodic with period  $d$  if all its states are periodic with period  $d > 1$  and aperiodic if all its states are aperiodic. Finally, a chain is ergodic if all its states are ergodic.
- In an ergodic scenario, the average outcome of the group is the same as the average outcome of the individual over time. An example of an ergodic systems would be the outcomes of a coin toss (heads/tails). If 100 people flip a coin once or 1 person flips a coin 100 times, you get the same outcome.

## ■ Limiting Theorems

- Once ergodicity of the chain is established, important limiting theorems can be stated. The first and most important one is the ergodic theorem. The ergodic average of a real-valued function  $t(\theta)$  is the average

$$\bar{t}_n = \left( \frac{1}{n} \right) \sum_{i=1}^n t(\theta^{(i)}).$$

- If the chain is ergodic and  $E_\pi[t(\theta)] < \infty$  for the unique limiting distribution  $\pi$  then

$$\bar{t}_n \rightarrow (\text{a.s.}) E_\pi[t(\theta)] \text{ as } n \rightarrow \infty \quad (32)$$

where *a.s.* refers to almost sure.

## ■ Limiting Theorems

- This result is a Markov chain equivalent of the law of large numbers. It states that averages of chain values also provide strongly consistent estimates of parameters of the limiting distribution  $\pi$  despite their dependence.
- If  $t(\theta) = I(\theta = x)$  then the ergodic averages are simply counting the relative frequency of values of  $x_s$  in realizations of the chain. By the ergodic theorem, this relative frequency converges almost surely to  $\pi(x) = \frac{1}{m_x}$ , the average frequency of visits to state  $x$ .

**HW/CW: You collect sales of shoes from a store say Bhatbhateni for three months. It may increase but if you analyze difference it remains almost stationary.**

## ■ Reversible Chains

- Let  $(\theta^{(n)})_n \geq 0$  be an homogeneous Markov chain with transition probabilities  $P(x, y)$  and stationary distribution  $\pi$ . Assume that one wishes to study the sequence of states  $\theta^{(n)}, \theta^{(n-1)}, \dots$  in reverse order. One can show that this sequence satisfies

$$Pr(\theta^{(n)} = y | \theta^{(n+1)} = x, \theta^{(n+2)} = x_2, \dots) = Pr(\theta^{(n)} = y | \theta^{(n+1)} = x)$$

which defines a Markov chain. The Transition probabilities are

$$\begin{aligned} P_n^*(x, y) &= Pr(\theta^{(n)} = y | \theta^{(n+1)} = x) \\ &= \frac{Pr(\theta^{(n+1)} = x | \theta^{(n)} = y) Pr(\theta^n = y)}{Pr(\theta^{(n+1)} = x)} \\ &= \frac{\pi^{(n)}(y) P(y, x)}{\pi^{(n+1)}(x)} \end{aligned} \tag{33}$$

and in general the chain is not homogeneous.

## ■Reversible Chains

- If  $n \rightarrow \infty$  or alternatively,  $\pi^{(0)} = \pi$ , then  $P_n^*(x, y) = P * (x, y) = \pi(y)P(y, x)/\pi(x)$  and the chain becomes homogeneous. If  $P * (x, y) = P(x, y)$  for all  $x$  and  $y \in S$ , the time reversed Markov chain has the same transition probabilities as the original Markov chain. Markov chains with such a property are said to be reversible and the reversibility condition is usually written as

$$\pi(x)P(x, y) = \pi(y)P(y, x) \quad \text{for all } x, y \in S \quad (34)$$

## ■ Reversible Chains

- It can be interpreted as saying that the rate at which the system moves from  $x$  to  $y$  when in equilibrium,  $\pi(x)P(x, y)$ , is the same as the rate at which it moves from  $y$  to  $x$ ,  $\pi(y)P(y, x)$ . For that reason, equation 34 is sometimes referred to as the detailed balance equation; balance because it equates the rates of moves through states and detailed because it does it for every possible pair of states.

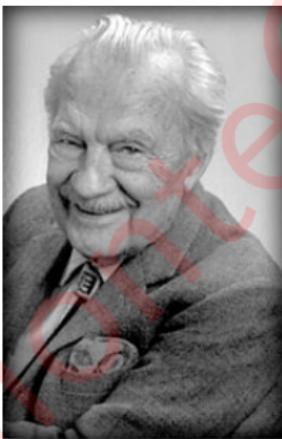
**HW/CW: You can prove that the irreducible birth and death chains are reversible.**

**Note that the detailed balance is required for getting and maintaining the stationary (equilibrium) Distribution.**

## ■ Reversible Chains

- Reversible chains are useful because if there is a distribution  $\pi$  satisfying equation 34 for an irreducible chain, then the chain is positive recurrent, reversible with stationary distribution  $\pi$ . This is easily obtained by summing over  $y$  both sides of 34 to give 21. Construction of Markov chains with a given stationary distribution  $\pi$  reduces to finding transition probabilities  $P(x, y)$  satisfying 34.

## ■Reversible Chains- Example : Metropolis Algorithm



**Figure:** Metropolis was a Greek-American physicist. In Los Alamos he led the group in the Theoretical Division that designed and built the MANIAC I computer in 1952 that was modeled on the IAS machine, and the MANIAC II in 1957. At Los Alamos in the late 1940s and early 1950s a group of researchers led by Metropolis, including John von Neumann and Stanislaw Ulam, developed the Monte Carlo method.

## ■Reversible Chains- Example : Metropolis Algorithm

- Consider a given distribution  $p_x$ ,  $x \in S$  with  $\sum_x p_x = 1$  where the state space  $S$  can be a subset of the line or even a  $d$ -dimensional subset of  $R^d$ . The problem posed and solved by Metropolis and coworkers in 1953 was how to construct a Markov chain with stationary distribution  $\pi$  such that  $\pi(x) = p_x$ ,  $x \in S$ .

Let  $Q$  be any irreducible transition matrix on  $S$  satisfying the symmetry condition  $Q(x, y) = Q(y, x)$ , for  $x, y \in S$ . Define a

Markov chain  $(\theta^{(n)})_n \geq 0$  as having transition from  $x$  to  $y$  proposed according to the probabilities  $Q(x, y)$ . This proposed value for  $(\theta^{(n+1)})$  is accepted with probability  $\min\{1, \frac{p_y}{p_x}\}$  and rejected otherwise, leaving the chain in state  $x$ .

## ■ Reversible Chains- Example : Metropolis Algorithm

- The transition probabilities  $P(x, y)$  of the above chain  $(\theta^{(n)})_n \geq 0$  are

$$\begin{aligned} P(x, y) &= Pr(\theta^{(n+1)} = y, TA | \theta^{(n)} = x) \\ &= Pr(\theta^{(n+1)} = y, |\theta^{(n)} = x) Pr(TA) \\ &= Q(x, y) \min\left\{1, \frac{p_y}{p_x}\right\} \end{aligned} \tag{35}$$

for  $y \neq x$  and TA denotes the event [transition accepted.]

## ■ Reversible Chains- Example : Metropolis Algorithm

- If  $y = x$ , then

$$\begin{aligned} P(x, x) &= Pr(\theta^{(n+1)} = x, TA | \theta^{(n)} = x) + Pr(\theta^{(n+1)} \neq x, \bar{T}A | \theta^{(n)} = x) \\ &= Pr(\theta^{(n+1)} = x, |\theta^{(n)} = x) Pr(TA) + \sum_{y \neq x} Pr(\theta^{(n+1)} = y, \bar{T}A | \theta^{(n)} = x) \\ &= Q(x, x) + \sum_{y \neq x} Q(x, y)[1 - \min\{1, p_y/p_x\}] \end{aligned} \tag{36}$$

In the above case  $y = x$ , we need to consider state  $y = x$  with  $\text{prob}(TA)$  as well as  $y \neq x$  with prob  $1 - \text{prob}(TA) \equiv \text{prob}(\bar{T}A)$ .

## ■Reversible Chains- Example : Metropolis Algorithm

The first step to obtaining the stationary distribution of this chain is to prove that the probabilities  $p_x$  satisfy the reversibility condition. For  $x = y$ , equation

$$\pi(x)P(x, y) = \pi(y)P(y, x) \quad (37)$$

that is

$$p_x P(x, y) = p_y P(y, x) \quad (38)$$

for all  $x, y \in S$  is trivially satisfied.

## ■Reversible Chains- Example : Metropolis Algorithm

For  $y \neq x$ , there will be two cases. (i)  $p_y > p_x$  and case (ii)  $p_x > p_y$ .  
Case (i)  $p_y > p_x$ :

$$p_x P(x, y) = p_x Q(x, y) = Q(y, x) \min\{1, p_x/p_y\} p_y = p_y P(y, x) \quad (39)$$

In writing above equation, we used  $p_x = \min(1, \frac{p_x}{p_y}) * p_x$  as we are considering  $p_y > p_x$  from  $\min(1, \frac{p_x}{p_y})$  one gets  $\frac{p_x}{p_y}$  and hence  $p_x = \min(1, \frac{p_x}{p_y}) * p_x$ .

Case (ii) can be followed analogously. Therefore the chain is reversible and the probabilities  $p_x, x \in S$  provide the stationary distribution of the chain. If  $Q$  is aperiodic, so will be  $P$  and the stationary distribution is also the limiting distribution.

## ■ Example : Metropolis Algorithm

One can use Metropolis algorithm to generate "target distribution" say gaussian distribution from "initial (random) distribution".

The algorithm is simple.

- **Initialization (Input):**
- *randomsamples* : A list containing initial random values.
- *mu* : The mean of the target Gaussian distribution.
- *sigma* : The standard deviation of the target Gaussian distribution.
- *numsteps* : The number of iterations to run the Metropolis algorithm.
- **Output:** states: A list containing the final samples that approximate a Gaussian distribution.

## ■ Example : Metropolis Algorithm

- Create a copy of the *randomsamples* list to avoid modifying the original data (store it in *states*).
- Iterate for *numsteps* : - Randomly select an index (*index*) from the *states* list.
- - Propose a new state by adding a random Gaussian noise (*deltax*) to the current state at the selected index (*states[index]*).
- - Calculate the acceptance probability based on the target Gaussian distribution:
- - Calculate the probability of the new state (*targetprobnew*) using its distance from the target mean (*mu*) and standard deviation (*sigma*).

## ■ Example : Metropolis Algorithm

- - Calculate the probability of the old state (*targetprobold*) using the same approach. - Divide *targetprobnew* by *targetprobold* to get the acceptance probability.
- - Generate a random number between 0 and 1.
- - If the random number is less than the acceptance probability, accept the proposed state and update *states[index]* with the new value.
- Return the final list *states* containing the samples that resemble a Gaussian distribution.
- Plot initial and final distributions

Discuss the results from code (20240315MetropolisExample.ipynb) - initial distribution "uniform" and final distribution "gaussian". Also check supplementary lectures (20240315Metropolis.tex)

## ■ Continuous state spaces

In continuous state spaces, sequences of random quantities that form a Markov chain in  $\mathbb{R}$  but still retain a discrete parameter space  $T$ . There are a few changes required with respect to the discrete case but the main results of the previous sections are still valid. In particular, convergence to the limiting distribution, the ergodic theorem and the central limit theorem need basically technical changes in the conditions of the chain to hold.

## ■ Continuous state spaces- transition kernels

Markov chains can be defined analogously as discrete case. If the conditional probabilities do not depend on the step  $n$ , the chain is homogeneous.

Then the transition kernel  $P(x, A)$  is again used to define the chain. The analogy with the discrete case breaks when trying to consider  $P(x, \{y\})$ , which is always null in the continuous case and not useful in this context. Therefore, transition matrices cannot be constructed and transition kernels must be used instead.

## ■ Continuous state spaces- transition kernels

However, given that  $P(x, \cdot)$  defines a probability distribution, the notation  $P(x, y)$  can be used as

$$P(x, y) = \Pr(\theta^{(n+1)} \leq y | \theta^n = x) = \Pr(\theta^{(1)} \leq y | \theta^0 = x) \quad (40)$$

for  $x, y \in S$  when  $P$  is absolutely continuous with respect to  $y$ . Also associated with this conditional distribution, one can obtain the conditional density

$$p(x, y) = \frac{\partial P(x, y)}{\partial y} \quad (41)$$

for  $x, y \in S$ . This density can be used to define the transition kernel of the chain instead of  $P(x, A)$ . The state space  $S$  does not need to be the entire line. It can be any interval or collection of intervals for results below to hold.

## ■ Continuous state spaces- transition kernels

The conditional transition probability over m steps is given by

$$P^m(x, y) = \Pr(\theta^{(m+n)} \leq y | \theta^{(n)} = x) \text{ for } x, y \in S \quad (42)$$

and the transition kernel over to steps is given by

$$P^m(x, y) = \frac{\partial P^m(x, y)}{\partial y} \text{ for } x, y \in S \quad (43)$$

## ■ Continuous state spaces- Stationary distribution

The stationary or invariant distribution  $p(x, y)$  must satisfy

$$\pi(y) = \int_{-\infty}^{\infty} \pi(x)p(x, y)dx \quad (44)$$

which is continuous version of stationary distribution.

# Gibbs Sampling

**Prof. Dr. Narayan Prasad Adhikari**  
Central Department of Physics  
Tribhuvan University Kirtipur, Kathmandu, Nepal

March 18, 2025



# ■ Assigned Problems

- Introduction
- Definition and Properties
- Implementation and optimization
- Forming the Sample
- Scanning strategies
- Using the sample
- Reparametrization
- Convergence diagnostics
- Applications

# ■Introduction

- Gibbs sampling was originated in the context of image processing. In this context, the posterior of interest for sampling is a Gibbs distribution. Borrowing concepts from Mechanical Statistics, the density of the Gibbs distribution can be written as

$$f(x_1, x_2, \dots, x_d) \propto \exp\left[\frac{-E(x_1, x_2, \dots, x_d)}{kT}\right] \quad (1)$$

where  $k$  is a positive constant and  $T$  is absolute temperature.  $E$  is the energy of the system, a positive function, and  $x_i$  is the characteristic of interest for the  $i$ th component of the system,  $i = 1, \dots, d$ . In Mechanical Statistics,  $x_i$  is the position or perhaps the velocity and position of the  $i$ th particle and in image processing it is (an indicator of) the colour of the  $i$ th pixel of an image.

## ■Introduction

- The energy function  $E$  is commonly given by a sum of potential functions  $V$ . These sums operate over collections of subgroups of components over which each potential function is evaluated.
- The subgroups generally obey some neighboring relationship in their definition. This leads to a probability specification based on local properties, useful for modelling spatial interaction between components. The main drawback is the difficulty in the determination of the global properties, such as the normalizing constant.
- Gibbs sampling scheme could in fact be used for a host of other posterior distributions.

# ■ In Physics

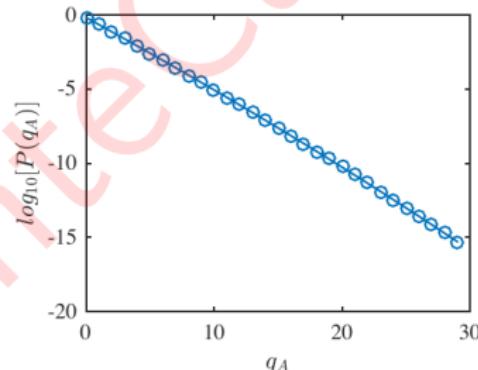
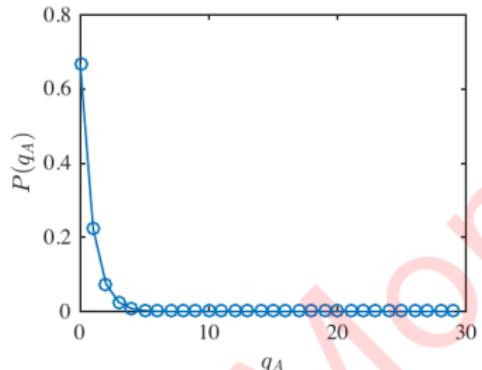


Figure: The Gibbs distribution gives how the energy is distributed in the system.

## ■Definition and Properties

- The Gibbs sampler is a very useful tool for simulations of Markov processes for which the transition matrix cannot be formulated explicitly
- Gibbs sampling is a Markov Chain Monte Carlo (MCMC) scheme where the transition kernel is formed by the full conditional distributions. Let us assume as before that the distribution of interest is  $\pi(\theta)$  where  $\theta = (\theta_1, \dots, \theta_d)'$ . Each one of the components  $\theta_i$  can be a scalar, a vector or a matrix. However in our case we consider them as scalar.
- Consider also that the full conditional distributions  $\pi_i(\theta_i) = \pi(\theta_i | \theta_{-i})$ ,  $i = 1, 2, \dots, d$  are available.
- This means that they are completely known and can be sampled from.

## ■Definition and Properties

- The problem to be solved is to draw from  $\pi$  when direct generation schemes are costly, complicated or simply unavailable but when generations from the  $\pi_i$ , are possible.
- Gibbs sampling provides an alternative generation scheme based on successive generations from the full conditional distributions.
- **To carry on the Gibbs sampling we use following algorithm (See next page)**

## ■Definition and Properties -Algorithm

- Initialize the iteration counter of the chain  $j = 1$  and set initial values

$$\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_d^{(0)})'$$

- Obtain a new value  $\theta^{(j)} = (\theta_1^{(j)}, \theta_2^{(j)}, \dots, \theta_d^{(j)})'$  from  $\theta^{(j-1)}$  through successive generations of values

$$\theta_1^{(j)} = \pi(\theta_1 | \theta_2^{(j-1)}, \theta_3^{(j-1)}, \dots, \theta_d^{(j-1)})'$$

$$\theta_2^{(j)} = \pi(\theta_2 | \theta_1^{(j)}, \theta_3^{(j-1)}, \dots, \theta_d^{(j-1)})'$$

.....

.....

$$\theta_d^{(j)} = \pi(\theta_d | \theta_1^{(j)}, \theta_2^{(j)}, \dots, \theta_{d-1}^{(j)})'$$

- Change counter  $j$  to  $j + 1$  and return to above step until convergence is reached.

## ■Definition and Properties

- When convergence is reached, the resulting value  $\theta(j)$  is a draw from  $\pi$ . As the number of iterations increases, the chain approaches its equilibrium condition. Convergence is then assumed to hold approximately.
- The obvious form to obtain a sample of size  $n$  from  $\pi$  is to replicate  $n$  chains until convergence.
- Alternatively, after convergence all draws from a chain come from the stationary distribution. Therefore  $n$  successive values from this chain after the burn-in period will also provide a sample from  $\pi$ .

## ■Algorithm - two variables case

- Data distribution:
- Assume that  $y|\theta \sim N(\theta, \Sigma)$  is a bivariate normal distribution with unknown mean  $\theta = (\theta_1, \theta_2)$  and known covariance matrix (in covariance matrix variances are given by diagonal elements whereas covariance is given by non-diagonal elements)

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

- Prior distribution: The prior for  $\theta$  is an improper uniform over the real line i.e.  $p(\theta_1, \theta_2) \propto 1$ .
- Posterior distribution: Assuming we observe a single observation  $y = (y_1, y_2)$ ,

$$\theta|y \sim N(y, \Sigma)$$

## ■Algorithm - two variables case

- Full conditional distribution:

$$\theta_1 | \theta_2, y \sim N(y_1 + \rho(\theta_2 - y_2), 1 - \rho^2)$$

$$\theta_2 | \theta_1, y \sim N(y_2 + \rho(\theta_1 - y_1), 1 - \rho^2)$$

Sample from the posterior distribution using a Gibbs sample assuming  $y = (0, 0)$  and  $\rho = 0.8$

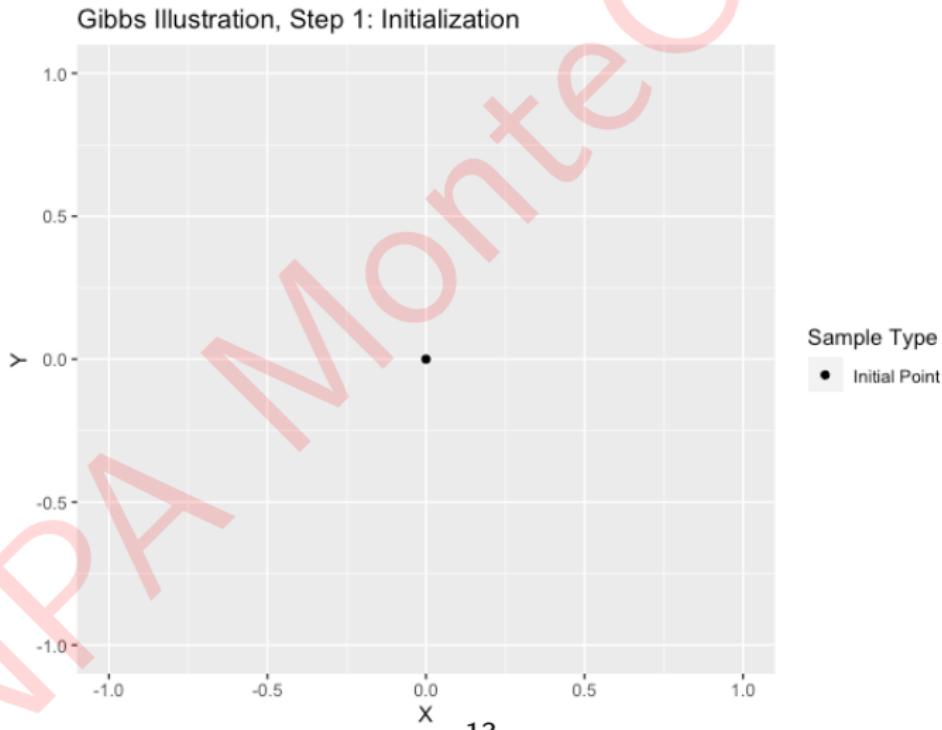
- **Algorithm and code**

## ■Algorithm - two variables case

- The gist of the Gibbs sampler is simple: sample from known conditional distributions, and use that resulting value to sample the next random variable from the following conditional probability distribution, ad infinitum.
- **Algorithm:**
  - 1. Initialize  $(x_0, y_0)$  and set time  $n$  or  $t = 0$
  - 2. Draw  $x_t$  from conditional distribution  
 $X_t | (Y_{t-1} = y_{t-1}) \sim N(\rho y_{t-1}, 1 - \rho^2)$
  - 3. Draw  $y_t$  from conditional distribution  
 $y_t | (x_t = x_t) \sim N(\rho x_t, 1 - \rho^2)$
  - 4. Increase  $t=t+1$
  - 5. Return to step 2

## ■Algorithm - two variables case

- Lets consider the case for  $\rho = 0.9$  Step 1: Initialize  $x_0 = 0.$ ;  $y_0 = 0.$  also set the iteration counter  $n$  OR  $(t)$  to 0.

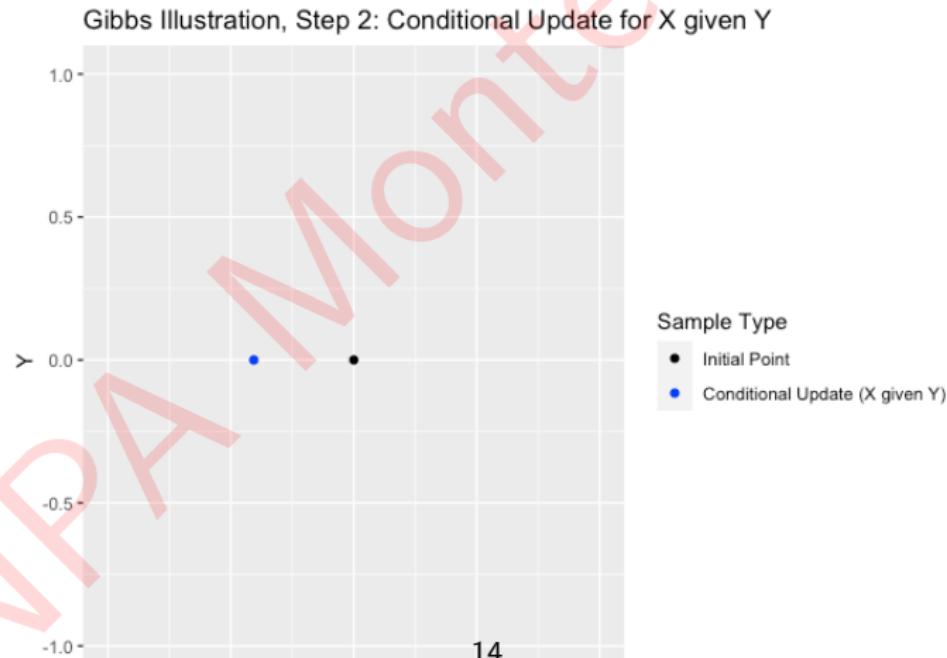


## ■Algorithm - two variables case

- Step 2: Conditional update of X given Y

$$X_1 | (Y_0 = 0) \sim N(0 \times \rho, 1 - \rho^2)$$

In one of the case it was -0.4 as shown below.



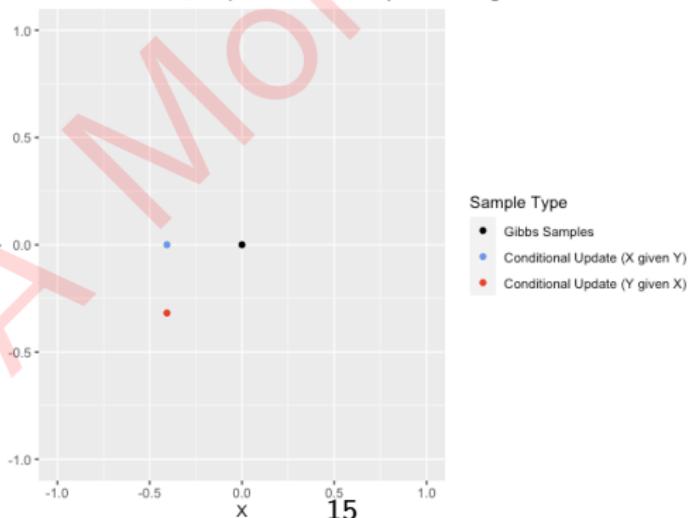
## ■Algorithm - two variables case

- Step 3: Conditional update of Y given X

$$Y_1 | (X_1 = -0.4) \sim N(-0.4 \times \rho, 1 - \rho^2)$$

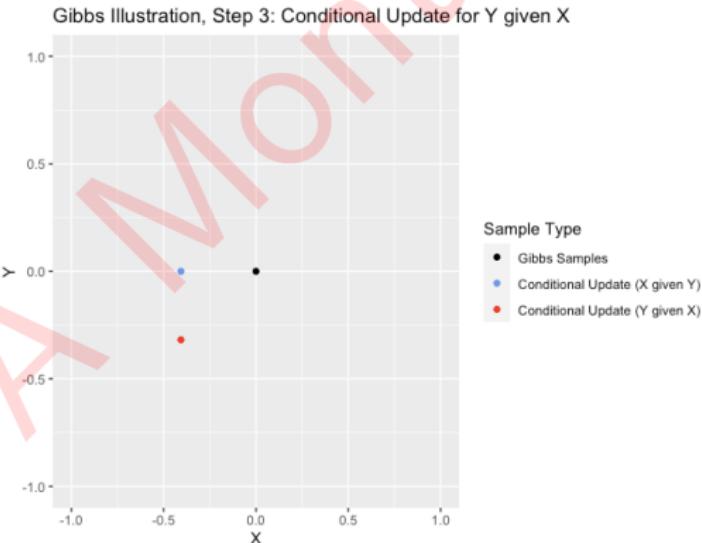
In one of the case it was -0.32 as shown below. This time, the X coordinate of our new point is the same as the X coordinate of the point from step 2.

Gibbs Illustration, Step 3: Conditional Update for Y given X



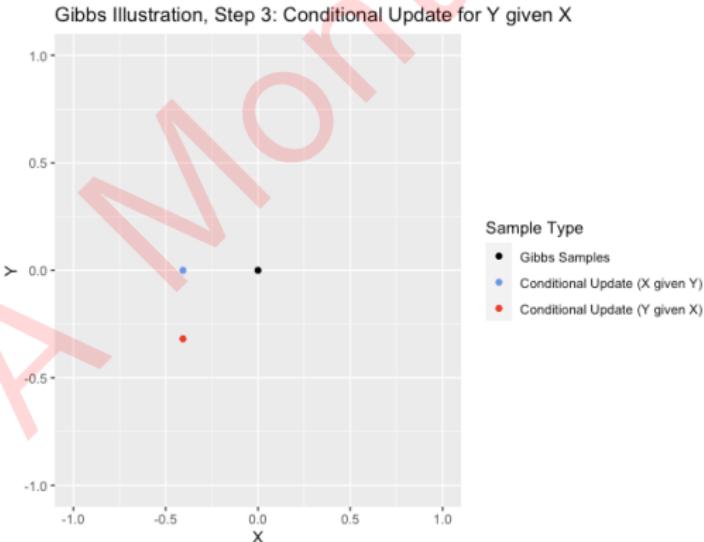
## ■Algorithm - two variables case

- Now we consider more steps repeating above processes. In this way one can sample the conditional probability of X given Y and vice versa. For multivariate systems its not easy to handle analytically. Even in this bivariate case for many steps its not possible to solve this problem analytically.



## ■ Python code - two variables case

- Now we consider more steps repeating above processes. In this way one can sample the conditional probability of X given Y and vice versa. For multivariate systems its not easy to handle analytically. Even in this bivariate case for many steps its not possible to solve this problem analytically.



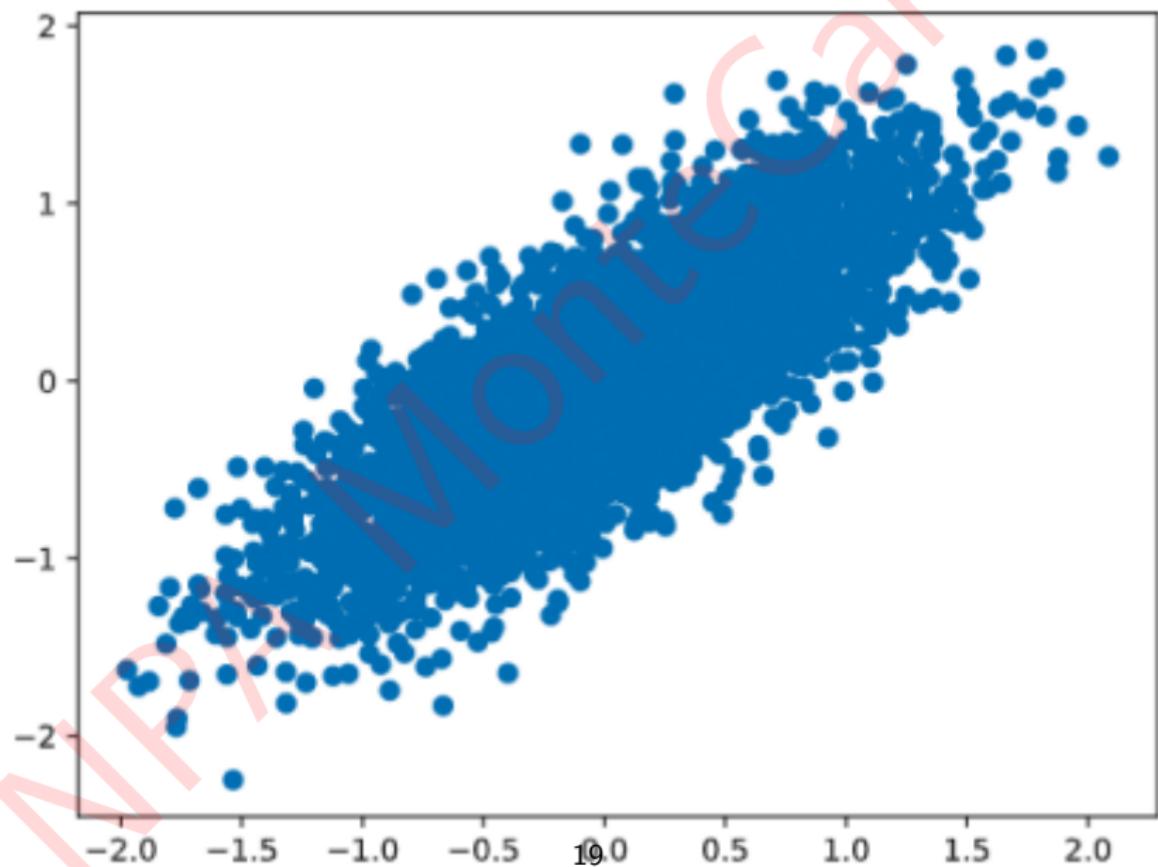
# ■ Python code - two variables case

```
In [11]: # import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
#matplotlib inline
#config InlineBackend.figure_format = 'svg'
np.random.seed(42)
mus = np.asarray([0, 0])
sigmas = np.asarray([[1, .8], [.8, 1]])
def gibbs_sampler(mus, sigmas, n_iter=5000):
    samples = []
    y = mus[1]
    for _ in range(n_iter):
        x = p_x_given_y(y, mus, sigmas)
        y = p_y_given_x(x, mus, sigmas)
        samples.append([x, y])
    return samples
def p_x_given_y(y, mus, sigmas):
    mu = mus[0] + sigmas[1, 0] / sigmas[0, 0] * (y - mus[1])
    sigma = sigmas[0, 0] - sigmas[1, 0] / sigmas[1, 1] * sigmas[1, 0]
    return np.random.normal(mu, sigma)

def p_y_given_x(x, mus, sigmas):
    mu = mus[1] + sigmas[0, 1] / sigmas[1, 1] * (x - mus[0])
    sigma = sigmas[1, 1] - sigmas[0, 1] / sigmas[0, 0] * sigmas[0, 1]
    return np.random.normal(mu, sigma)

samples = gibbs_sampler(mus, sigmas); samples[:5000]
burn = 100
x, y = zip(*samples[burn:])
plt.plot(x, y, "o")
plt.show()
```

## ■ Python code - two variables case



## ■Classwork/Homework

- Solve the problem of bivariate normal distribution with  $x_0 = y_0 = 0$  and  $\rho = 0.8$
- Solve the problem of bivariate normal distribution with  $x_0 = 0, y_0 = 0.5$  and  $\rho = 0.8$
- Solve the problem of bivariate normal distribution with  $x_0 = 0.5, y_0 = 0$  and  $\rho = 0.8$
- Solve the problem of bivariate normal distribution with  $x_0 = 0.5, y_0 = 0.25$  and  $\rho = 0.8$
- Solve the problem of bivariate normal distribution with  $x_0 = 0.25, y_0 = 0.5$  and  $\rho = 0.8$
- Now repeat above problems with  $\rho = 0.0$
- Again repeat above problems with  $\rho = 1.0$

## ■Classwork/Homework

- You can study the distribution of  $x$  and  $y$  separately by plotting the histograms
- It must be Gaussian distribution.
- Understand the role of  $\rho$  and initial values.
- What is the role of mean values?
- How do you obtain mean values of  $x$  and  $y$ ?

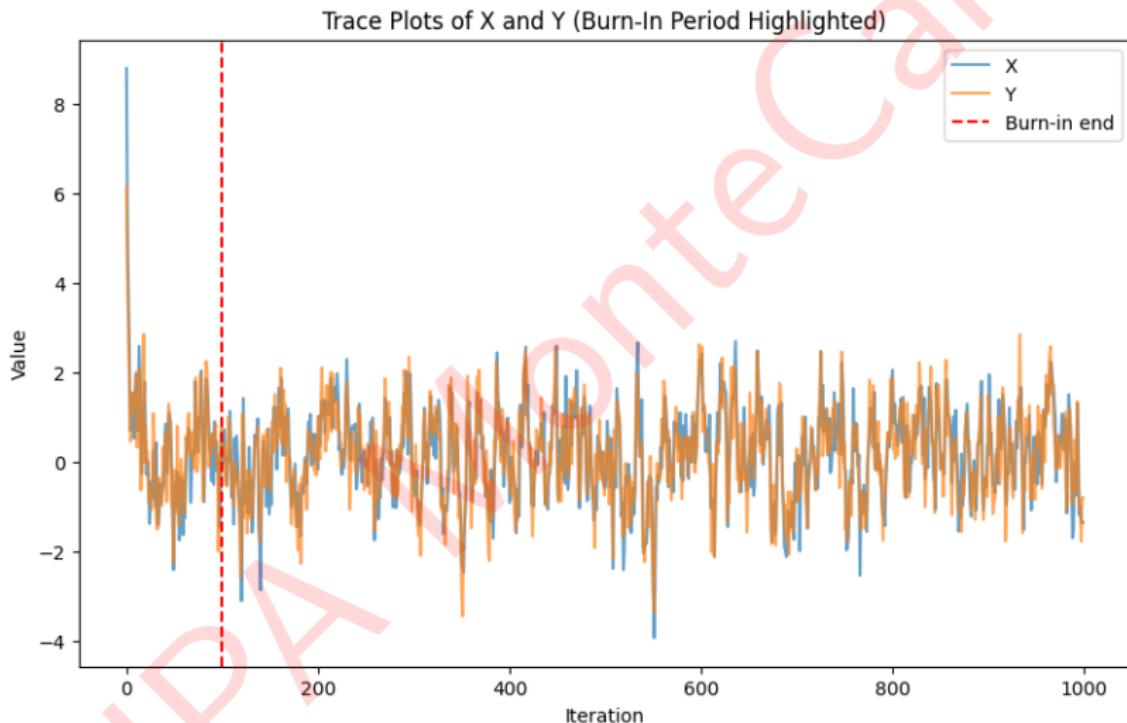
## ■Classwork/Homework

- Now generate one chain for long time and remove the burn-in period (the time required to reach the chain in its equilibrium distribution). Then study the distributions of  $x$  and  $y$ . Did you get the same results as before?
- Now instead of one long chain, you generate many chains and remove the burn-in periods (the time required to reach the chain in its equilibrium distribution) of each chain. Then study the distributions of  $x$  and  $y$ . Did you get the same results as before?
- Which way looks better?

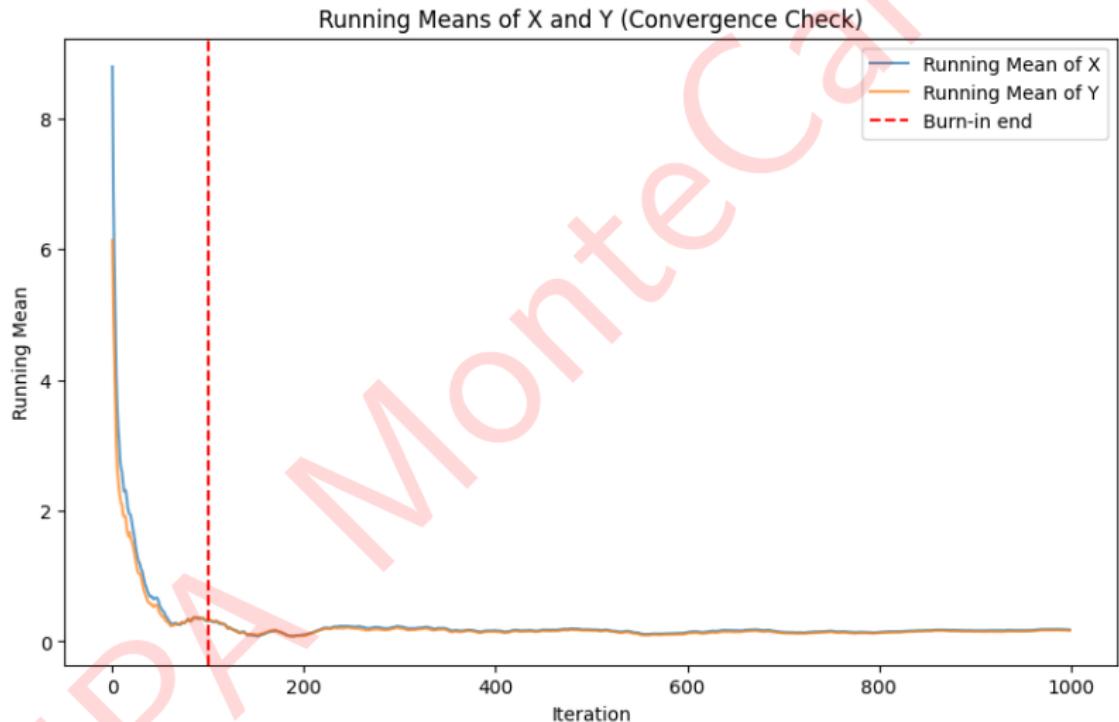
## ■Classwork/Homework: Burn-in Period

- Take starting point  $x_0 = y_0 = 0.0$  and carry on Gibbs sampling as above.
- Plot trace values i.e. x versus number of iterations and y versus number of iterations
- Plot running average versus time for time = (say)100.
- The results are shown in figures.
- Now explain the significance of burn-in period.

# Classwork/Homework: Burn-in Period



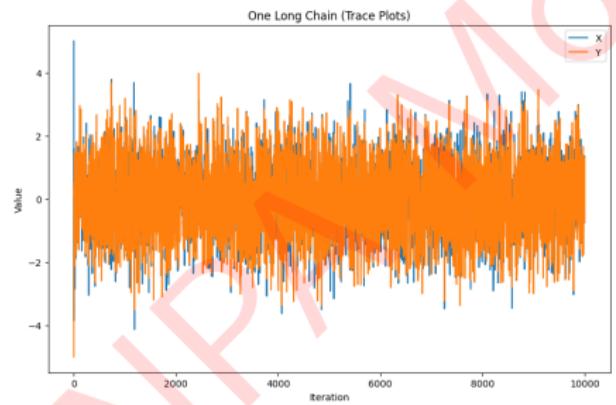
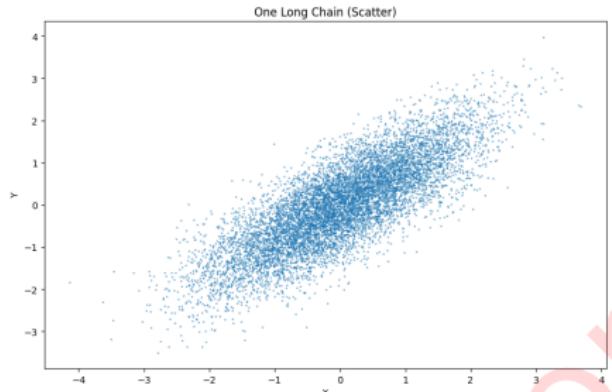
# Classwork/Homework: Burn-in Period



## ■Classwork/Homework: Burn-in Period

- Now you form a long chain of samples say 5000. Analyze it and get mean & variances of all x & y.
- You form 5 different chains of lengths say 1000 each. Analyze results - mean & variances.
- Explain/workout the conclusions
- Which way do you prefer now on? Explore your idea

# Classwork/Homework: Forming Samples



## ■Classwork/Homework: Forming Samples

One Long Chain:

Mean X: 0.0116, Mean Y: 0.0026

Variance X: 1.0604, Variance Y: 1.0587

Multiple Short Chains:

Mean X: 0.0068, Mean Y: 0.0097

Variance X: 1.0083, Variance Y: 1.0052

## ■ Homework to be submitted within a week

- Study the Gibbs sampler for the case of bivariate normal distribution. Analyze the convergence and forming sample and prepare a report.

## ■ Implementation and Optimization

Despite the theoretical results ensuring the convergence of the Gibbs sampler, its practical implementation may be complicated by the potential complexity of the models considered. Convergence of the sampler becomes difficult to characterize. Given that it is a numeric and iterative method, practical strategies to improve the efficiency of the method may have a considerable impact on its computational cost. Efficiency broadly consists of reducing the number of burn-in iterations and the amount of arithmetic operations required at each iteration. The techniques presented are related to the basic MCMC methods.

## ■Implementation and Optimization- Forming the sample

There are two forms to obtain a sample of size  $n$  from the posterior distribution  $\pi$ .

(i) The obvious one is to process  $n$  chains in parallel until convergence, say after  $m$  iterations, and take as sample elements the  $m$ th chain value from each of the  $n$  chains. The generation procedure will then require  $mn$  generations from the chain. If chains are initialized independently, the sample consists of independent values from  $\pi$ . Independence is easier to establish if the initial values are all different and preferably with larger dispersion than in the posterior.

## ■Implementation and Optimization- Forming the sample

(ii) Another form is to consider a single chain and explore ergodic results. After convergence, all chain values have marginal distribution given by the equilibrium distribution  $\pi$ . So, a sample of size  $n$  may be formed by  $n$  successive values from this chain. This generation will require  $m + n$  generations from the chain. This is substantially less than independent sampling. The difficulty here is that the sample elements are no longer independent due to chain dependence. Ergodic theorems ensure that inference based on this sample is still valid. From a practical point of view, there may be problems if the chain autocorrelation is too high and the sample is not large enough to acknowledge it. In these cases, chains may take too long to adequately cover the entire parameter space appropriately. As a result, some relevant regions may be underrepresented in the sample.

## ■Implementation and Optimization- Forming the sample

An alternative approach accomodating independence is to take for the sample chain values at every  $k$ th iteration after the burn-in period. Markovian processes only have first order dependence. As the lag between iterations increases, chain values become less and less correlated and are virtually independent for a large enough value of the lag  $k$ . A sample of size  $n$  with quasi-independent elements thus requires  $m + kn$  generations from the chain. The value of  $k$  is typically smaller than  $m$  and again an improvement over independent sampling is obtained. There is no gain in efficiency. This procedure is advantageous if computer storage of values is limited.

## ■Implementation and Optimization- Forming the sample

There is no general agreement on the subject although it is generally agreed that running  $n$  parallel chains in practice is computationally inefficient and unnecessary. The main debate is whether a few parallel chains are needed. If the convergence properties of the chain are well understood then clearly a single chain suffices. As these characteristics are hard to obtain, prudence suggests that a few pilot parallel chains should be run. If they quickly settle around common values then a single chain can be safely used to extract a large sample for inference. Otherwise, there may be minor characteristics of the posterior distribution such as secondary modes far from the mode that require very large samples to be noticed.

## ■Implementation and Optimization- Scanning strategies

The Gibbs sampler described above involved a complete scan over the components. All iterations consisted of visits to update the components in the same deterministic order, typically  $1 \rightarrow 2 \rightarrow \dots \rightarrow d$ . There are many other possible scanning or updating strategies for visiting the components of  $\theta$ . There are a few schemes for the purpose.

- (i) Geman and Geman proved convergence to the joint distribution in a discrete setting for all visiting schemes that guarantee that all components are visited infinitely often when the chain is run indefinitely. The reversible Gibbs sampler where at each iteration each component is visited in a fixed order and then visited again in reversed order satisfies this property.

## ■Implementation and Optimization- Scanning strategies

- (ii) Another scheme where an i.o. schedule is guaranteed draws a number  $i$  from  $\{1, 2, \dots, d\}$  with fixed positive probabilities at each iteration and only updates the  $\theta_i$  at that iteration. To make it more comparable with the deterministic scan, an iteration of these random scans can be defined by a collection of  $d$  such updates.
- (iii) In another scheme we consider a **random permutation scan** where at each iteration a permutation of  $\{1, 2, \dots, d\}$  is selected and components are visited in that order.

## ■ Implementation and Optimization- Scanning strategies

Assume now that  $\pi$  is a multivariate normal distribution with precision matrix  $\Phi = (\phi_{ij})$ . For this setting, convergence for the deterministic scan is faster than for the random scan if  $\Phi$  is tridiagonal ( $\pi(\theta_i|\theta_{-i}) = \pi(\theta_i|\theta_{i-1}, \theta_{i+1})$ , for all i) OR if  $\Phi$  has nonnegative partial correlations ( $\phi_{ij} \leq 0$ ). This result is particularly important because both dynamic and hierarchical models lead to tridiagonal matrices if variances are known. Their results also indicate that more precise distributions lead to faster convergence both for the deterministic and random scans.

## ■Implementation and Optimization- Using the sample

Whatever the scheme chosen for forming the sample, after it is used a sample of vectors  $\theta_1, \dots, \theta_n$  generated from the posterior distribution  $\pi$  is available. Assume also the more general case where these are successive values from a single Markov chain. A sample from the  $i$ th component of  $\theta$  is given by  $\theta_{1i}, \dots, \theta_{ni}$ . Marginal point or interval summaries of any real function  $\psi = t(\theta)$  are estimated by their corresponding estimators based on the sample. This is always a consistent estimator by the **ergodic theorem**.

## ■Implementation and Optimization- Using the sample

The posterior mean of  $\psi$  is estimated by  $\hat{E}(\psi) = \hat{\psi} = (1/n) \sum_{j=1}^n \psi_j$  where  $\psi_j = t(\theta_j)$ ,  $j = 1, \dots, n$ . The posterior variance of  $\psi$  is similarly estimated by noting that

$$\sigma_\psi^2 = \text{Var}(\psi) = E(\psi^2) - [E(\psi)]^2.$$

The expectation value is obtained by  $\bar{t} \rightarrow E_\pi[t(\theta)]$  as  $n \rightarrow \infty$ .  
Similarly the  $\sigma_\psi^2$  is estimated by  $\hat{\sigma}_\psi^2$  where

$$\hat{\sigma}_\psi^2 = \hat{E}(\psi^2) - [\hat{E}(\psi)]^2 = \frac{1}{n} \sum_{j=1}^n (\psi_j \hat{\psi})^2 \quad (2)$$

the sample variance.

## ■Implementation and Optimization- Using the sample

The marginal densities  $\pi(\theta_i)$  can be estimated by (a smoothed version of) the histogram of sampled values of  $\theta_i$ . Better estimators can be obtained by using conditional distributions. Recalling that

$$\pi(\theta_i) = \int \pi(\theta_i | \theta_{-i}) \pi(\theta_{-i}) d\theta_{-i}$$

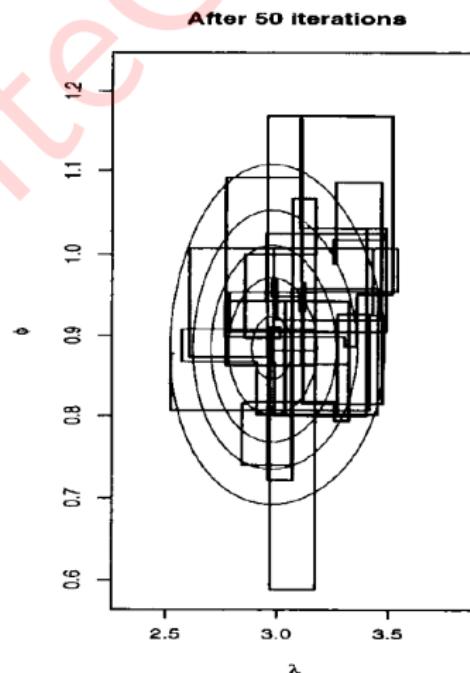
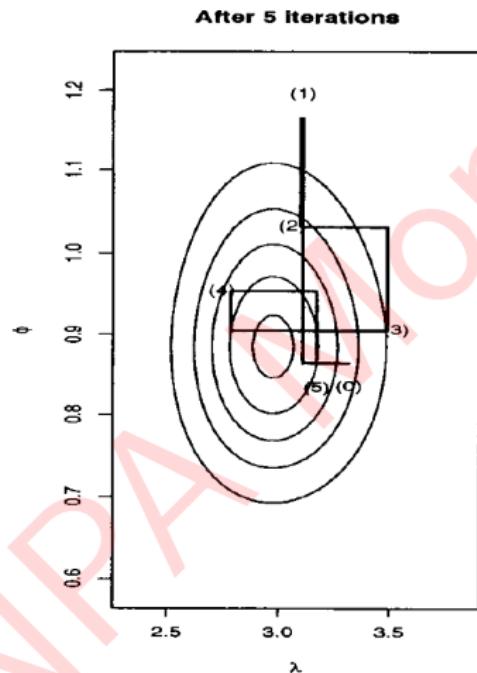
a Monte Carlo estimator is given by

$$\hat{\pi}(\theta_i) = \frac{1}{n} \sum_{j=1}^n \pi(\theta_i | \theta_{j,-i}) \quad (3)$$

where  $\theta_{j,-i}$ ,  $j = 1, \dots, n$  are samples from marginal  $\pi(-\theta_i)$ .

# ■ Implementation and Optimization-Reparametrization

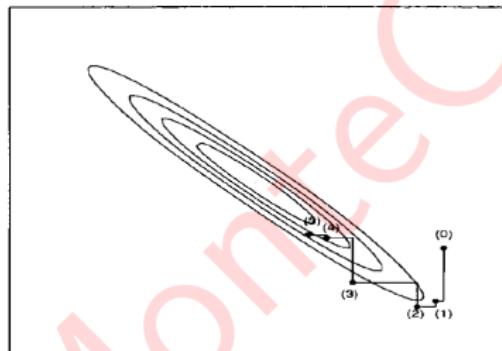
Consider the figure shown below.



## ■ Implementation and Optimization- Reparametrization

- An iteration is formed by moves along the coordinate axes of the components of  $\theta$ .
- If there is weak dependence between the components, the moves will be ample. Often, the posterior structure leads to high correlation between some of the components of  $\theta$
- Figure 2 illustrates this point for a bidimensional parameter. The contours of the posterior show strong dependence between the components of  $\theta$  and chain moves, governed by the conditional densities, will be small. The chain will take many iterations to adequately cover the parametric space and as a result convergence is slow. In this case, the Gibbs sampler will be inefficient. Examples can be constructed in larger dimension models where convergence can be slowed to any arbitrary amount of iterations

## ■ Implementation and Optimization- Reparametrization



**Figure:** Contour lines of a bivariate posterior density with components highly correlated. A possible chain trajectory is also depicted to illustrate slow convergence, with iterations in parentheses. The contours are from a bivariate normal distribution with marginal distributions  $\theta_1 \sim N(2, 1)$  and  $\theta_2 \sim N(3, 1)$  and correlation -0.97. The trajectory is obtained by sampling from the full conditional distributions  $\theta_2|\theta_1$  and  $\theta_1|\theta_2$

## ■ Reparametrization- An Example

- We want to estimate the joint distribution of two variables, X and Y, where:
- $X \sim \text{Uniform}(-2, 2)$  (uniform distribution between -2 and 2)
- $Y = X^2$  ( $Y$  is simply the square of  $X$ )
- We face following Problem:
- While sampling X from a uniform distribution is straightforward, directly sampling Y from its conditional distribution (given X) is difficult. The typical Gibbs sampling approach would involve:

## ■ Reparametrization- An Example

- Sample X from its conditional distribution (which is easy in this case).
- Sample Y given the sampled X (which requires evaluating a complex function - squaring X).
- However, the issue here is that for most values of X, the corresponding Y value will be very close to 0 (since  $X^2$  is small for small X). This creates a "bottleneck" effect around  $Y = 0$ , making the chain mix slowly between high and low Y values.

## ■ Reparametrization- Algorithm of example

- One of the ways to solve above problem is Reparametrization
- To improve convergence, we can reparametrize Y and X both.
- Define the number of samples (*numsamples*).
- **Initialize arrays**
- Initialize arrays to store samples for both cases: with and without reparametrization
- $xsamplesnoreparam, ysamplesnoreparam$ : Arrays to store samples without reparametrization.
- $xsamplesreparam, ysamplesreparam$ : **Arrays to store samples with reparametrization.**

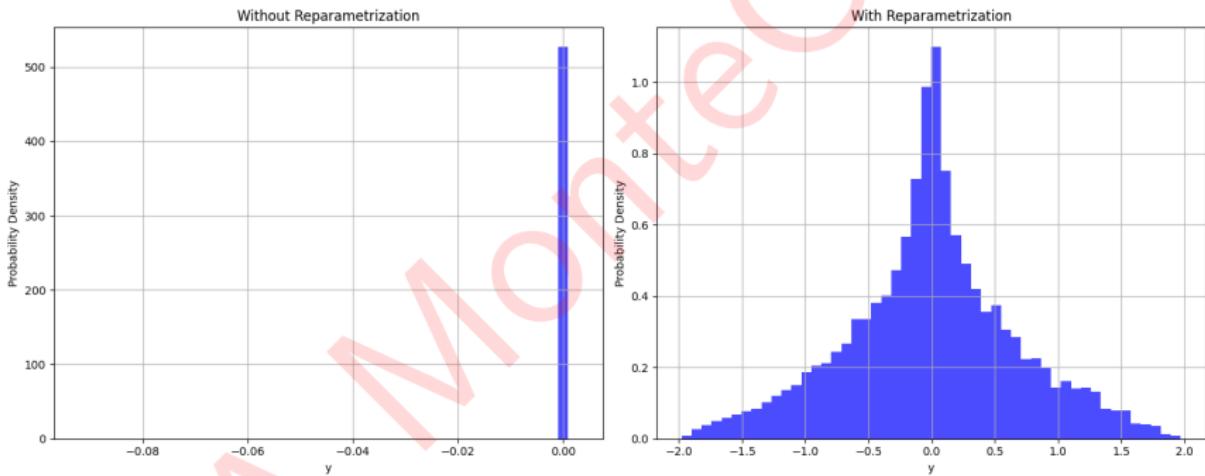
## ■ Reparametrization- Algorithm of example

- Without Reparametrization:
- Initialize starting values for  $x$  and  $y$  using uniform distribution within the specified range.
- Implement Gibbs sampling iterations:
  - i. Sample  $x$  from its conditional distribution given  $y$ .
  - ii. Sample  $y$  from its conditional distribution given  $x$ .
  - iii. Store the samples in the respective arrays.

## ■ Reparametrization- Algorithm of example

- With Reparametrization:
- Initialize starting values for  $u$  and  $v$  using uniform distribution within the range  $[0, 1]$ .
- Reparametrize  $u$  and  $v$  to the range of  $x$  and  $y$ .
- $x$  from  $(4*u-2)$  &  $y$  from  $(2*v-1)*x$
- Implement Gibbs sampling iterations:
  - i. Sample  $u$  from its conditional distribution given  $v$ .
  - ii. Sample  $v$  from its conditional distribution given  $u$ .
  - iii. Reparametrize  $u$  and  $v$  to the range of  $x$  and  $y$ .
  - iv. Store the samples in the respective arrays.
- plot all the data obtained.

## ■ Reparametrization- Algorithm of example



## ■ Implementation and Optimization- Reparametrization

A simple and sometimes effective way to reduce convergence time is to use **reparametrizations**. Adequate transformations in the parameter space may produce situations of near independence that are ideal for fast convergence of the chain. Unfortunately, there are no rules to determine suitable transformations but frequently linear transformations that produce a diagonal variance matrix provide good results.

## ■ Reparametrization - An example

### Random Effects (Hierarchical) model

consider the simple random effects model:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

with  $i=1,\dots,I$ ; and  $j=1,\dots,J$  where  $\alpha_i \sim N(0, \sigma_\alpha^2)$  and  $\epsilon_{ij} \sim N(0, \sigma_y^2)$ .

For a flat prior on  $\mu$ , the Gibbs sampler implemented for the  $(\mu, \alpha_1, \dots, \alpha_I)$  parametrization exhibits high correlations and consequent slow convergence if  $\sigma_y^2/(IJ\sigma_\alpha^2)$  is large.

On the other hand, if the model is rewritten as the hierarchy

$$Y_{ij} \sim N(\eta_i, \sigma_y^2), \eta_i \sim N(\mu, \sigma_\alpha^2)$$

, the correlations between the  $\eta_i$ 's and between  $\mu$  and the  $\eta_i$ 's are lower so converges faster than before.

## ■ Sampling from the full conditional distributions

In some cases, the form of the full conditional distribution is not recognizable which prevents sampling via the conventional algorithms.

Ritter and Tanner developed yet another sampling scheme from difficult full conditionals. Their approach is similar to adaptive rejection by being based on the evaluation of the full conditional at a few selected points. For that reason, they called it the griddy Gibbs sampler. Let  $\pi_i(\theta_{ij})$ , be a difficult full conditional distribution. Then, sampling from  $\pi_i$  can be approximately performed as follows:

## ■ Sampling from the full conditional distributions

1. Take a grid of points  $\theta_{i1}, \dots, \theta_{im}$ , evaluate  $\pi_i(\theta_{ij})$ ,  $j=1,\dots,m$  and normalize them to obtain weights  $w_1, \dots, w_m$ .
2. Use the weights  $w_1, \dots, w_m$  to construct a simple approximation to the distribution function of  $\pi_i$ .
3. Draw a value from  $\pi_i$  by the probability integral transform method.

## ■ Convergence Diagnostics

A value from the distribution of interest  $\pi$  is only obtained when the number of iterations of the chain approaches infinity. In practice this is not attainable and a value obtained at a sufficiently large iteration is taken instead of being drawn from  $\pi$ . The difficulty is the determination of how large this iteration should be. There is no simple answer to this question and most efforts have been directed at studying as close as possible the convergence characteristics of the chain.

## ■ Convergence Diagnostics

- There are two main ways to approach the study of convergence.
- The first one is more theoretical and tries to measure distances and establish bounds on distribution functions generated from a chain. In particular, one can study the total variation distance between the distribution of the chain at iteration  $j$  and the limiting distribution  $\pi$ . Special aspects derived from the probabilistic structure of the chain can also be studied.

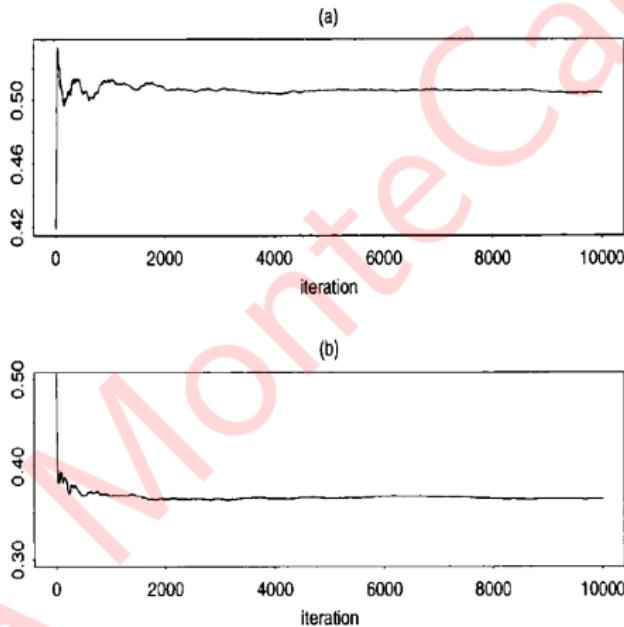
## ■ Convergence Diagnostics

- Another approach is statistical perspective i.e., by analyzing the properties of the observed output from the chain. This is an empirical as opposed to a theoretical treatment of the problem and is obviously more practical. The difficulty with this approach is that it can never guarantee convergence because it is only based on observations from the chain
- The first one is more difficult than the second one.

## ■ Convergence Diagnostics

- One can observe the trajectory of a chain exhibiting the same qualitative behavior through iterations after a transient initial period is an indication of convergence. Similarly, the trajectory of the ergodic averages can be evaluated and plotted. An asymptotic behavior over many successive iterations indicates convergence.
- Figure next page shows the ergodic averages of variance components in a model data.

## ■ Convergence Diagnostics



**Figure:** Ergodic averages of two parameters with number of iterations of the chain. The parameters are standard deviations of random effects at: (a) individual; (b) unit level in a longitudinal study of epilepsy treatment

## ■ Convergence Diagnostics

These, graphical, techniques must be used with caution and should always be accompanied by some theoretical reasoning. Graphical techniques may be deceptive indicating constancy that may not be so evident under a different scale. More importantly, there are many chains that exhibit every indication of convergence without actually achieving it. They are called metastable chains and are the subject of much research in probability theory.

## ■ Convergence Diagnostics- prescription

- Raftery and Lewis proposed a method to establish the length of a chain required for a M CM C run. More specifically, the methodology suggests values of  $m$ , the number of burn-in iterations,  $k$ , the number of iterations to be skipped between stored chain values and  $n$ , the size of the sample values that must be stored to achieve a given Monte Carlo precision of estimates.

## ■ Convergence Diagnostics- prescription

- The setting for these choices is the estimation of  $u$ , the  $q$  quantile (percentile) of a given function  $\psi = t(\theta)$ , i.e.  $q = \Pr_{\pi}(\psi \leq u)$ . The method requires that the Monte Carlo estimate  $\hat{q}$  satisfies  $\Pr(|\hat{q} - q| \leq r) = s$ . A common choice is the tail probability with  $q = 0.025$  in which case  $u$  is the lower limit of the equal tail 95% posterior credibility interval for  $\psi$ . One may require that the value of this probability be estimated in a MCMC run with error smaller than  $r = 0.01$  with confidence  $s = 0.99$ . So, 95% posterior intervals would be given by intervals with posterior probabilities between 93% and 97% with 99% confidence. This confidence level is due to the estimation of  $q$  by MCMC and should not be confused with posterior uncertainty about  $\psi$ , governed by  $\pi$ .

## ■ Convergence Diagnostics- Formal methods

- The methods presented here diagnose convergence based on exploration of the statistical properties of the observed chain. The methods here attempt to decide whether convergence can be safely assumed to hold rather than prescribing the run length to achieve convergence. There have been many methods presented in the literature. We consider only *Time series analysis*.

## ■ Convergence Diagnostics- Formal methods

- Consider a real function  $\psi = t(\theta)$  and its trajectory  $\psi^{(1)}, \psi^{(2)}, \dots$  obtained from  $\psi^{(j)} = t(\theta^{(j)})$ ,  $j = 1, 2, \dots$ . This trajectory defines a time series and ergodic averages of this series can be evaluated.
- Geweke suggested the use of tests on ergodic averages to verify convergence of the chain based on the series  $\psi^{(j)}$ .

## ■ Convergence Diagnostics- Formal methods

- Assume observation of the chain for to  $m + n$  iterations and form averages

$$\bar{\psi}_b = \frac{1}{n_b} \sum_{j=m+1}^{m+n_b} \psi^{(j)} \quad (4)$$

and

$$\bar{\psi}_a = \frac{1}{n_a} \sum_{j=m+n-n_a+1}^{m+n} \psi^{(j)} \quad (5)$$

where  $n_b + n_a < n$ .

- If  $m$  is the length of the burn-in period, then  $\psi_a$  and  $\psi_b$  are the ergodic averages at the end and beginning of the convergence period and should behave similarly. As  $n$  gets large and the ratios  $n_a/n$  and  $n_b/n$  remain fixed then

## ■ Convergence Diagnostics- Formal methods

$$z_G = \frac{\psi_a - \psi_b}{\sqrt{\hat{Var}(\psi_a) + \hat{Var}(\psi_b)}} \rightarrow N(0, 1) \quad (6)$$

So, the standardized difference  $z_G$  between the ergodic averages at the beginning and at the end of the convergence period should not be large if convergence has been achieved. Large differences indicate lack of convergence but small differences do not imply convergence. Geweke suggested the use of values  $n_b = 0.1n$  and  $n_a = 0.5n$  and used spectral density estimators for the variances. This is a univariate technique.

## ■ Applications of Gibbs Sampling: Hierarchical Model

- Hierarchical models have the following structure - first we specify that the data come from a distribution with parameters  $\theta$

$$X \sim f(X|\theta)$$

and that the parameters themselves come from another distribution with hyperparameters  $\lambda$

$$\theta \sim g(\theta|\lambda)$$

and finally that  $\lambda$  comes from a prior distribution

$$\lambda \sim h(\lambda)$$

- More levels of hierarchy are possible - i.e you can specify hyper-hyperparameters for the distribution of  $\lambda$  and so on.

## ■ Applications of Gibbs Sampling: Hierarchical Model

- The essential idea of the hierarchical model is because the  $\theta$ s are not independent but rather are drawn from a common distribution with parameter  $\lambda$ , we can share information across the  $\theta$ s by also estimating  $\lambda$  at the same time.
- As an example, suppose we have data about the proportion of heads after some number of tosses from several coins, and we want to estimate the bias of each coin. We also know that the coins come from the same mint and so might share some common manufacturing defect.

## ■ Applications of Gibbs Sampling: Hierarchical Model

- There are two extreme approaches - we could estimate the bias of each coin from its coin toss data independently of all the others, or we could pool the results together and estimate the same bias for all coins. Hierarchical models provide a compromise where we shrink individual estimates towards a common estimate.
- Note that because of the conditionally independent structure of hierarchical models, Gibbs sampling is often a natural choice for the MCMC sampling strategy.

## ■ Applications of Gibbs Sampling: Hierarchical Model

- Suppose we have data of the number of failures ( $y_i$ ) for each of 10 pumps in a nuclear plant. We also have the times ( $i$ ) at which each pump was observed. We want to model the number of failures with a Poisson likelihood, where the expected number of failure  $\lambda_i$  differs for each pump. Since the time which we observed each pump is different, we need to scale each  $\lambda_i$  by its observed time  $t_i$
- We now specify the hierarchical model - note change of notation from the overview above - that  $\theta$  is  $\lambda$  (parameter) and  $\lambda$  is  $\beta$  (hyperparameter) simply because  $\lambda$  is traditional for the Poisson distribution parameter.

## ■ Applications of Gibbs Sampling: Hierarchical Model

The likelihood  $f$  is

$$\prod_{i=1}^{10} \text{Poisson}(\lambda_i t_i) \quad (7)$$

we let the prior  $g$  for  $\lambda$  be

$$\lambda \sim \text{Gamma}(\alpha, \beta) \quad (8)$$

with  $\alpha = 1$  and let  $\beta$  to be a random variable to be estimated from the data

$$\beta \sim \text{Gamma}(\gamma, \delta) \quad (9)$$

with  $\gamma = 0.01$  and  $\delta = 1$ .

There are 11 unknown parameters (10  $\lambda$ s and  $\beta$ ) in this hierarchical model.

## ■ Applications of Gibbs Sampling: Hierarchical Model

The posterior is

$$p(\lambda, \beta | y, t) = \prod_{i=1}^{10} \text{Poisson}(\lambda_i t_i) \times \text{Gamma}(\alpha, \beta) \times \text{Gamma}(\gamma, \delta) \quad (10)$$

with the conditional distributions needed for Gibbs sampling given by

$$p(\lambda_i | \lambda_{-i}, \beta, y, t) = \text{Gamma}(y_i + \alpha, t_i + \beta) \quad (11)$$

and

$$p(\beta | \lambda, y, t) = \text{Gamma}(10\alpha + \gamma, \delta + \sum_{i=1}^{10} \lambda_i) \quad (12)$$

**HW: Write Algorithm and hence code to solve above problem. You can take data as follows:**

# Markov Chain Monte Carlo: Metropolis-Hastings Algorithm

Prof. Dr. Narayan Prasad Adhikari

Central Department of Physics

Tribhuvan University Kirtipur, Kathmandu, Nepal

April 6, 2025



# ■Introduction

- The original paper by Metropolis et al. (1953) deals with the calculation of properties of chemical substances and was published in the Journal of Chemical Physics. Nevertheless, it later proved itself to have a great impact in Statistics and Simulation.

 Publishing The Journal of Chemical Physics

HOME BROWSE INFO FOR AUTHORS COLLECTIONS ACCEPTED MANUSCRIPTS

SIGN UP FOR ALERTS

Home > The Journal of Chemical Physics > Volume 21, Issue 6 > 10.1063/1.1699114  
No Access • Submitted: 06 March 1953 • Published Online: 23 December 2004  
◀ PREV ▶ NEXT ▷

## Equation of State Calculations by Fast Computing Machines

J. Chem. Phys. 21, 1087 (1953); <https://doi.org/10.1063/1.1699114>

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, and Augusta H. Teller  
• Los Alamos Scientific Laboratory, Los Alamos, New Mexico  
Edward Teller  
more...

View Contributors

PDF ABSTRACT CITED BY TOOLS SHARE METRICS

**TOPICS**

- Monte Carlo methods
- Statistical mechanics
- Equations of state

### ABSTRACT

A general method, suitable for fast computing machines, for investigating such properties as equations of state for substances consisting of interacting individual molecules is described. The method consists of a modified Monte Carlo integration over configuration space. Results for the two-dimensional rigid-sphere system have been obtained on the Los Alamos MANIAC and are presented here. These results are compared to the free volume equation of state and to a four-term virial coefficient expansion.

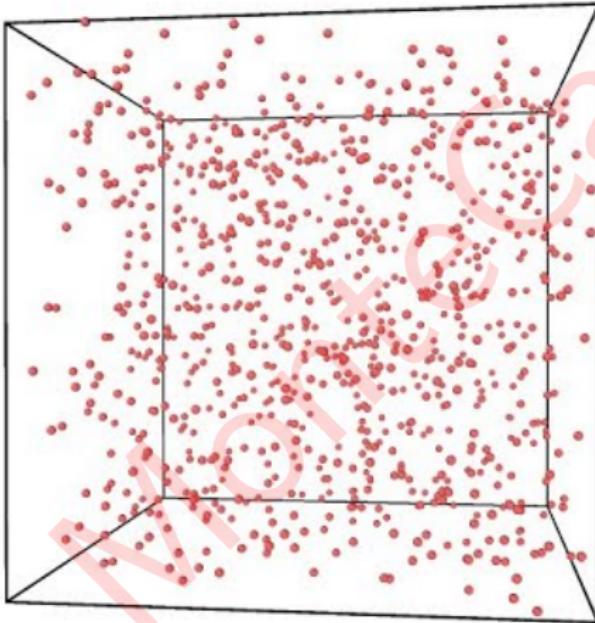


- Consider a substance with  $d$  molecules positioned at  $\theta = (\theta_1, \dots, \theta_d)'$ . In this case, the component  $\theta_i$  is formed by the bidimensional vector of positions in the plane of the  $i$ th molecule. From Statistical Mechanics, the density of these positions is given by Equation

$$f(x_1, \dots, x_d) \propto \exp(-E(x_1, \dots, x_d)/kT) \quad (1)$$

vectors where a potential  $V$  between molecules can be defined. The potential energy of the substance is then given by  $E(\theta) = \sum_{i,j} V(\theta_i, \theta_j)/2$ .

- The calculation of the equilibrium value of any chemical property is given by the expected value of this property with respect to the distribution of the vector of positions. Direct calculation of the expectation is not feasible for  $d$  large and is replaced by a Monte Carlo estimate.



- Metropolis et al. (1953) suggested a method to deal with the difficult problem of sampling from this density.

# ■ Metropolis Algorithm

- ① Start with any initial configuration  $\theta^{(0)} = (\theta_1^{(0)}, \dots, \theta_d^{(0)})'$  and set the iteration counter  $j = 1$ .
- ② Move the particles from previous positions  $\theta^{(j-1)} = (\theta_1^{(j-1)}, \dots, \theta_d^{(j-1)})'$  according to a uniform distribution centered at these positions in order to obtain new positions  $\phi = (\phi_1, \dots, \phi_d)'$ .
- ③ Calculate the change  $\Delta E$  in the potential energy caused by the move. The move in step 2 is accepted with probability  $\min \{1, \exp(-\Delta E)/kT\}$ , with If the move is accepted,  $\theta^{(j)} = \phi$  Otherwise,  $\theta^{(j)} = \theta^{(j-1)}$
- ④ Change the counter from  $j$  to  $j + 1$  and return to step 2 until convergence is reached.

After convergence, the vector of positions generated by the method has density according to probability density given by equation 1.

# ■ Metropolis Algorithm

- It is evident that the above method defines a Markov chain as the transitions depend only on the positions at the previous stage. However, it is not obvious that the method converges to an equilibrium distribution and that this distribution is given by equation 1.

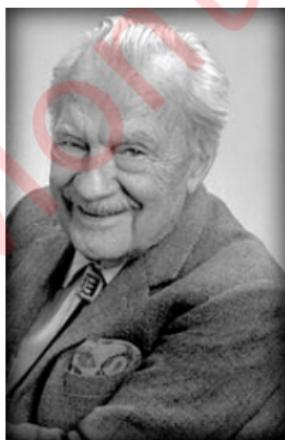


Figure: Nicholas Metropolis who invented Metropolis Algorithm

# ■ Metropolis Algorithm



Figure: Wilfred Keith Hastings (July 21, 1930 – May 13, 2016[1]) was a statistician. He was noted for his contribution to the Metropolis–Hastings algorithm, the most commonly used Markov chain Monte Carlo method (MCMC)

# ■ Metropolis Algorithm

- Metropolis and coworkers present a heuristic proof (non-formal a simple argument) of this result. The same proof is valid for the case where the moves to  $\phi$  are made according to any symmetric distribution centered at previous positions. This defines a transition kernel  $q$  that depends on  $(\theta, \phi)$  through  $(|\theta - \phi|)$ .
- Hastings in 1970 referred to the above algorithm in this extended form as the Metropolis method.

# ■Hastings Algorithm

*Biometrika* (1970), 57, 1, p. 97  
Printed in Great Britain

97

## Monte Carlo sampling methods using Markov chains and their applications

BY W. K. HASTINGS

*University of Toronto*

### SUMMARY

A generalization of the sampling method introduced by Metropolis *et al.* (1953) is presented along with an exposition of the relevant theory, techniques of application and methods and difficulties of assessing the error in Monte Carlo estimates. Examples of the methods, including the generation of random orthogonal matrices and potential applications of the methods to numerical problems arising in statistics, are discussed.

### 1. INTRODUCTION

For numerical problems in a large number of dimensions, Monte Carlo methods are often

# ■ Metropolis Algorithm

- Note that the above algorithm includes an additional step that was not present in the chains previously presented. The transition mechanism now depends on a proposed transition  $q$  and a subsequent step of evaluation of this proposal. Also note that the proposed positions are completely unrelated from the equilibrium distribution but this is represented in the overall transition through the acceptance probability because

$$\frac{\pi(\phi)}{\pi(\theta^{(j-1)})} = \frac{\exp(-E(\phi)/kT)}{\exp(-E(\theta^{(j-1)})/kT)} = \exp(-\Delta E/kT) \quad (2)$$

# ■ Metropolis Algorithm

- Another important point is that the resulting chain may remain in a low energy (or equivalently, high density) position for many iterations. In this case, it is likely that the proposal will lead to very high energy (very low density) points and  $\Delta E \ggg 0$  forcing an acceptance probability very close to 0. Computationally, this is not desirable and transition kernels must be carefully chosen to avoid such low acceptance rates.

## ■Definition and Properties

- Consider a distribution  $\pi$  from which a sample must be drawn via Markov chains. Again, it is worth stressing that this task will only make sense if the non-iterative generation of  $\pi$  is very complicated or expensive. In this case, a transition kernel  $p(\theta, \phi)$  must be constructed in a way such that  $\pi$  is the equilibrium distribution of the chain. A simple way to do this is to consider reversible chains where the kernel  $p$  satisfies

$$\pi(\theta)p(\theta, \phi) = \pi(\phi)p(\phi, \theta) \quad \forall(\theta, \phi) \quad (3)$$

As previously seen, this is the reversibility condition of the chain. It is also called *detailed balance equation*.

- Even though this is not a necessary condition for convergence, it is a sufficient condition in order that  $\pi$  be the equilibrium distribution of the chain.

## ■Definition and Properties

- The kernel  $p(\theta, \phi)$  consists of 2 elements: (i) an arbitrary transition kernel  $q(\theta, \phi)$  and a probability  $\alpha(\theta, \phi)$  such that

$$p(\theta, \phi) = q(\theta, \phi)\alpha(\theta, \phi), \text{ if } \theta \neq \phi \quad (4)$$

So, the transition kernel defines a density  $p(\theta, \cdot)$  for every possible value of the parameter different from  $\theta$ . Consequently, there is a positive probability left for the chain to remain at  $\theta$  given by

$$p(\theta, \theta) = 1 - \int q(\theta, \phi)\alpha(\theta, \phi)d\phi \quad (5)$$

These two forms can be grouped in the general expression

$$p(\theta, A) = \int_A q(\theta, \phi)\alpha(\theta, \phi)d\phi + I(\theta \in A) \left[ 1 - \int q(\theta, \phi)\alpha(\theta, \phi)d\phi \right] \quad (6)$$

for any subset A of the parameter space.

## ■Definition and Properties

- So, the transition kernel defines a mixed distribution for the new state  $\phi$  of the chain. For  $\theta \neq \phi$ , this distribution has a density and for  $\phi = \theta$ , this distribution has a probability atom.
- Hastings proposed to define the acceptance probability in such a way that when combined with the arbitrary transition kernel, it defines a reversible chain. The expression most commonly cited for the acceptance probability is

$$\alpha(\theta, \phi) = \min \left\{ 1, \frac{\pi(\phi)q(\phi, \theta)}{\pi(\theta)q(\theta, \phi)} \right\} \quad (7)$$

Algorithms based on chains with transition kernel given by equation 6 and acceptance probability equation 7 will be referred to as Metropolis-Hastings algorithms. Hastings referred to the ratio appearing in eq. 7 as the test ratio.

## ■Justification/Proof of Metropolis-Hastings Algorithm

- The purpose of the Metropolis–Hastings algorithm is to generate a collection of states according to a desired distribution  $p(\theta)$ . To accomplish this, the algorithm uses a Markov process, which asymptotically reaches a unique stationary distribution  $\pi(\theta)$  such that  $\pi(\theta) = p(\theta)$
- A Markov process is uniquely defined by its transition probabilities  $p(\phi|\theta)$ , the probability of transitioning from any given state  $\theta$  to any other given state  $\phi$ . It has a unique stationary distribution  $\pi(\theta)$  when the following two conditions are met.

# ■ Justification/Proof of Metropolis-Hastings Algorithm

- ① *Existence of stationary distribution:* there must exist a stationary distribution  $\pi(\theta)$ . A sufficient but not necessary condition is detailed balance, which requires that each transition  $\theta \rightarrow \phi$  is reversible: for every pair of states  $\theta, \phi$ , the probability of being in state  $\theta$  and transitioning to state  $\phi$  must be equal to the probability of being in state  $\phi$  and transitioning to state  $\theta$ ,

$$\pi(\theta)p(\phi|\theta) = \pi(\phi)p(\theta|\phi)$$

- ② *Uniqueness of stationary distribution:* the stationary distribution  $\pi(\theta)$  must be unique. This is guaranteed by ergodicity of the Markov process, which requires that every state must (1) be aperiodic—the system does not return to the same state at fixed intervals; and (2) be positive recurrent—the expected number of steps for returning to the same state is finite.

## ■Justification/Proof of Metropolis-Hastings Algorithm

The Metropolis–Hastings algorithm involves designing a Markov process (by constructing transition probabilities) that fulfills the two above conditions, such that its stationary distribution  $\pi(\theta)$  is chosen to be  $p(\theta)$ . The derivation of the algorithm starts with the condition of detailed balance:

$$p(\theta)p(\phi|\theta) = p(\phi)p(\theta|\phi)$$

which is written as

$$\frac{p(\phi|\theta)}{p(\theta|\phi)} = \frac{p(\phi)}{p(\theta)}. \quad (8)$$

## ■Justification/Proof of Metropolis-Hastings Algorithm

The approach is to separate the transition in two sub-steps; the proposal and the acceptance-rejection. The proposal distribution  $q(\phi | \theta)$  is the conditional probability of proposing a state  $\phi$  given  $\theta$  , and the acceptance distribution  $\alpha(\phi, \theta)$  is the probability to accept the proposed state  $\phi$ . The transition probability can be written as the product of them:

$$p(\phi|\theta) = q(\phi | \theta)\alpha(\phi, \theta)$$

Inserting this relation in the previous equation, we get

$$\frac{\alpha(\phi, \theta)}{\alpha(\theta, \phi)} = \frac{p(\phi)q(\theta | \phi)}{p(\theta)q(\phi | \theta)} \quad (9)$$

## ■Justification/Proof of Metropolis-Hastings Algorithm

The next step in the derivation is to choose an acceptance ratio that fulfills the condition above. One common choice is the Metropolis choice:

$$\alpha(\phi, \theta) = \min \left\{ 1, \frac{p(\phi)q(\theta | \phi)}{p(\theta)q(\phi | \theta)} \right\} \quad (10)$$

For this Metropolis acceptance ratio  $\alpha$ , either  $\alpha(\phi, \theta) = 1$  or  $\alpha(\theta, \phi) = 1$  and, either way, the condition is satisfied.

## ■ Metropolis-Hastings Algorithm

In practical terms, simulation of a draw from  $n$  using the Markov chain defined by the transition given by equation 6 can be set up as follows:

- ① Initialize the iteration counter  $j = 1$  and set an arbitrary initial value  $\theta^{(0)}$ .
- ② Move the chain to a new value  $\phi$  generated from the density  $q(\theta^{(j-1)}, \cdot)$
- ③ Evaluate the acceptance probability of the move  $\alpha(\theta^{(j-1)}, \phi)$  given by equation 7. If the move is accepted, set  $\theta^{(j)} = \phi$ . If it is not accepted,  $\theta^{(j)} = \theta^{(j-1)}$  and the chain does not move.
- ④ Change the counter from  $j$  to  $j + 1$  and return to step 2 until convergence is reached.

## ■ Metropolis-Hastings Algorithm

Step 3 is performed after the generation of an independent uniform quantity  $u$  - a random number (in practice). If  $u < \alpha$ , the move is accepted and if  $u > \alpha$  the move is not allowed. The transition kernel  $q$  defines only a possible move that can be confirmed according to the value of  $\alpha$ . For that reason,  $q$  is generally referred to as the proposal kernel or proposal (conditional) density when looked upon as a (conditional) density  $q(\theta, \cdot)$ . Other terms sometimes used are probing kernel or density.

In any of the forms of the Metropolis algorithm,  $q$  defines a symmetric transition around the previous positions of the molecules. Therefore,

$$q(\theta, \phi) = q(\phi, \theta)$$

## ■ Metropolis-Hastings Algorithm

for every  $(\theta, \phi)$  and the acceptance probability becomes

$$\alpha(\theta, \phi) = \min \left\{ 1, \frac{\pi(\phi)}{\pi(\theta)} \right\} \quad (11)$$

depending only on a simplified test ratio  $\frac{\pi(\phi)}{\pi(\theta)}$ , the ratio of the posterior density values at the proposed and previous positions of the chain.

Note also that the chain may remain in the same state for many iterations. A useful monitoring device of the method is given by the average percentage of iterations for which moves are accepted. Hastings suggests that this acceptance rate should always be computed in practical applications.

## ■ Metropolis-Hastings Algorithm

It is also crucial that the proposal kernels are easy to draw from as the method replaces the difficult generation of  $\pi$  by many generations proposed from  $q$ . Another less obvious but equally important requirement to be met by  $q$  is the correct tuning of the moves it proposes to ensure that moves covering the parameter space can be made and accepted in real computing time.

The test ratio can be rewritten as

$$\frac{\pi(\phi)/q(\theta, \phi)}{\pi(\theta)/q(\phi, \theta)} \quad (12)$$

Acceptance of proposed values is based on the ratio of target and proposed density. So, there is a connection here with the resampling schemes. In resampling schemes, the proposal density  $q$  was to be chosen as similar as possible to  $n$  to increase acceptance rates but the methods were not iterative. Also, for the rejection method, the rejection probability depended only on the numerator in 12.

## ■ Metropolis-Hastings Algorithm

The target distribution  $\pi$  enters the algorithm through the test ratio  $\pi(\phi)/\pi(\theta)$  in the form of the ratio as in the resampling methods. So again, the complete knowledge of  $\pi$  is not required. In particular, proportionality constants are not needed. When  $\pi$  is a posterior density, even though its functional form is always known, the value of the proportionality constant is rarely known. So, the algorithm is particularly useful for applications to Bayesian inference.

Many of the comments made about Gibbs sampling in the previous chapter are also valid for the Metropolis-Hastings algorithm. So, the discussion about single long against multiple chains is just as relevant here.

## ■ Metropolis-Hastings Algorithm

Formal and informal convergence techniques described in **Gibbs Sampling** can all be used here. The exception is made up of those based on complete knowledge of conditional densities. Typically, but not necessarily, Metropolis-Hastings algorithms are used when these are not completely known and hence difficult to sample from. When the complete conditional densities are known, Gibbs sampling is generally used.

**The algorithm has the interesting feature of not needing the value of the normalization constant in the probability density function. This is the most interesting feature of Metropolis-Hastings Algorithm.**

## ■ Metropolis-Hastings Algorithm: Example

We define three functions: (1) *Normal*, which evaluates the probability density of any observation given the parameters mu and sigma. (2) *Random\_coin*. And (3) *Gaussian\_mcmc*, which samples executes the algorithm as described.

We're not calling any Gaussian or normal function from numpy, scipy, etc. In the third function, we initialize a current sample as an instance of the uniform distribution (where the lower and upper boundaries are  $+/- 5$  standard deviations from the mean.) Likewise, movement is defined in the same way. Lastly, we move (or stay) based on the random event's observed value in relation to acceptance, which is the probability density comparison discussed at length elsewhere.

# ■ Metropolis-Hastings Algorithm: Code

```
: import numpy as np
import random
import matplotlib.pyplot as plt
def normal(x,mu,sigma):
    numerator = np.exp((- (x-mu)**2)/(2*sigma**2))
    denominator = sigma * np.sqrt(2*np.pi)
    return numerator/denominator

def random_coin(p):
    unif = random.uniform(0,1)
    if unif>=p:
        return False
    else:
        return True

def gaussian_mcmc(hops,mu,sigma):
    states = []
    burn_in = int(hops*0.2)
    current = random.uniform(-5*sigma+mu,5*sigma+mu)
    for i in range(hops):
        states.append(current)
        movement = random.uniform(-5*sigma+mu,5*sigma+mu)

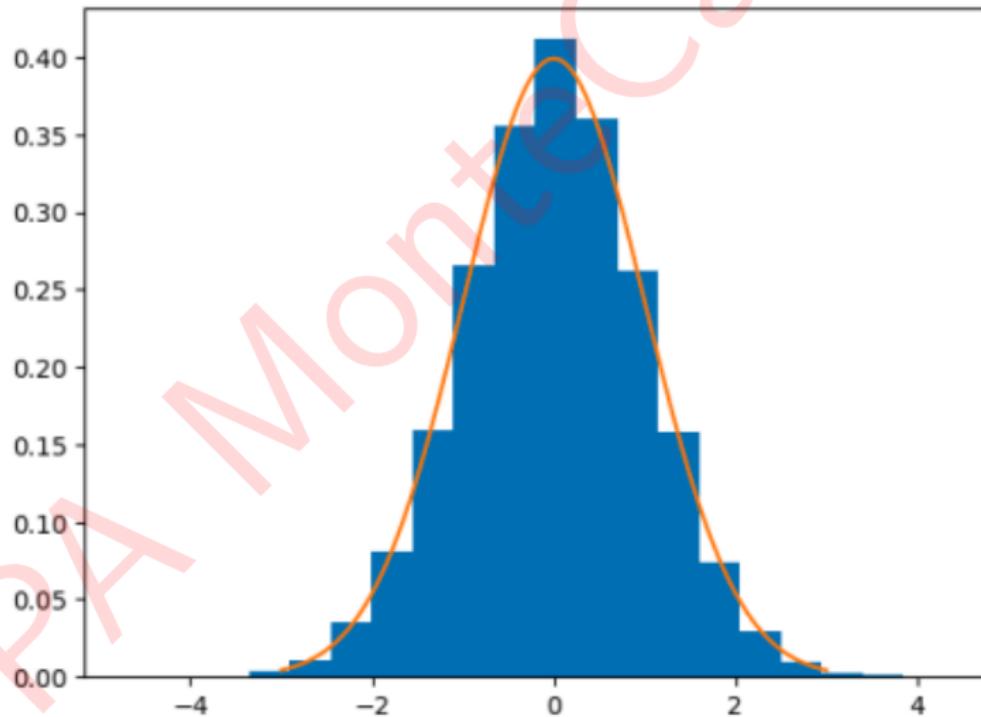
        curr_prob = normal(x=current,mu=mu,sigma=sigma)
        move_prob = normal(x=movement,mu=mu,sigma=sigma)

        acceptance = min(move_prob/curr_prob,1)
        if random_coin(acceptance):
            current = movement
    return states[burn_in:]

lines = np.linspace(-3,3,1000)
normal_curve = [normal(l,mu=0,sigma=1) for l in lines]
dist = gaussian_mcmc(100_000,mu=0,sigma=1)
plt.hist(dist,density=1,bins=20)
plt.plot(lines,normal_curve)
```

## ■ Metropolis-Hastings Algorithm: Code

Out[8]: [`<matplotlib.lines.Line2D at 0x7f975a1e1af0>`]



# Markov Chain Monte Carlo: Approaches for Statistical Inference

**Prof. Dr. Narayan Prasad Adhikari**

Central Department of Physics

Tribhuvan University Kirtipur, Kathmandu, Nepal

This lecture note is based on Text Book

January 3, 2023



# ■Outline

- Introduction
- Motivating vignettes
- Defining the approaches
- Bayes vs frequentist approach
- Some basic Bayesian models

## ■ Thomas Bayes



**Figure:** Thomas Bayes (1701 – 7 April 1761) was an English statistician, philosopher and Presbyterian minister who is known for formulating a specific case of the theorem that bears his name: Bayes' theorem

# ■ Bayes Theorem

## LIKELIHOOD

The probability of "B" being True, given "A" is True

## PRIOR

The probability "A" being True. This is the knowledge.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

## POSTERIOR

The probability of "A" being True, given "B" is True

## MARGINALIZATION

The probability "B" being True.

# ■Introduction

- The practicing statistician faces a variety of challenges: designing complex studies, summarizing complex data sets, fitting probability models, drawing conclusions about the present, and making predictions for the future. Statistical studies play an important role in scientific discovery, in policy formulation, and in business decisions. Applications of statistics are ubiquitous, and include clinical decision making, conducting an environmental risk assessment, setting insurance rates, deciding whether (and how) to market a new product, and allocating federal funds. Currently, most statistical analyses are performed with the help of commercial software packages, most of which use methods based on a classical, or frequentist, statistical philosophy.

# ■Introduction

- The Bayesian approach to statistical design and analysis is emerging as an increasingly effective and practical alternative to the frequentist one. Indeed, due to computing advances that enable relevant Bayesian designs and analyses, the philosophical battles between frequentists and Bayesians that were once common at professional statistical meetings are being replaced by a single, more eclectic approach.

# ■Motivating vignettes

## Personal Probability

- Suppose you have submitted your first manuscript to a journal and have assessed the chances of its being accepted for publication. This assessment uses information on the journal's acceptance rate for manuscripts like yours (let's say around 30%), and your evaluation of the manuscript's quality.
- Subsequently, you are informed that the manuscript has been accepted (congratulations!). What is your updated assessment of the probability that your next submission (on a similar topic) will be accepted?

## ■ Motivating vignettes

- The direct estimate is of course 100% (thus far, you have had one success in one attempt), but this estimate seems naive given what we know about the journal's overall acceptance rate (our external, or prior, information in this setting). You might thus pick a number smaller than 100%; if so, you are behaving as a Bayesian would because you are adjusting the (unbiased, but weak) direct estimate in the light of your prior information. This ability to formally incorporate prior information into an analysis is a hallmark of Bayesian methods, and one that frees the analyst from ad hoc adjustments of results that "don't look right."

## Motivating vignettes

### Missing Data

- Consider Table (below), reporting an array of stable event prevalence or incidence estimates scaled per 10,000 population, with one value (indicated by “★”) missing at random. We may think of them as geographically aligned disease prevalences, or perhaps as death rates cross-tabulated by clinic and age group.

79	87	83	80	78
90	89	92	99	95
96	100	★	110	115
101	109	105	108	112
96	104	92	101	96

## ■ Motivating vignettes

- With no direct information for  $\star$ , what would you use for an estimate? Does 200 seem reasonable? Probably not, since the unknown rate is surrounded by estimates near 100. To produce an estimate for the missing cell you might fit an additive model (rows and columns) and then use the model to impute a value for  $\star$ , or merely average the values in surrounding cells. These are two examples of borrowing information. Whatever your approach, some number around 100 seems reasonable.
- Now assume that we obtain data for the  $\star$  cell and the estimate is, in fact, 200, based on 2 events in a population of 100 ( $200 = 10000 \times 2/100$ ). Would you now estimate  $\star$  by 200 (a very unstable estimate based on very little information), when with no information a moment ago you used 100?

## ■ Motivating vignettes

- While 200 is a perfectly valid estimate (though its uncertainty should be reported), some sort of weighted average of this direct estimate (200) and the indirect estimate you used when there was no direct information (100) seems intuitively more appealing. The Bayesian formalism allows just this sort of natural compromise estimate to emerge.
- Finally, repeat this mental exercise assuming that the direct estimate is still 200 per 10,000, but now based on 20 events in a population of 1000, and then on 2000 events in a population of 100,000. What estimate would you use in each case? Bayes and empirical Bayes methods structure this type of statistical decision problem, automatically giving increasing weight to the direct estimate as it becomes more reliable.

## Motivating vignettes

**Bioassay: measurement of the concentration or potency of a substance by its effect on living cells or tissues.**

- Consider a carcinogen bioassay where you are comparing a control group (C) and an exposed group (E) with 50 rodents in each (see Table below). In the control group, 0 tumors are found; in the exposed group, there are 3, producing a non-significant, one-sided Fisher exact test p-value (p-values give the probability that the null hypothesis is true) of approximately 0.125. However, your colleague, who is a veterinary pathologist, states, "I don't know about statistical significance, but three tumors in 50 rodents is certainly biologically significant!"

	C	E	Total
Tumor	0	3	3
No Tumor	50	47	97
	120	50	100

## Motivating vignettes

- This belief may be based on information from other experiments in the same lab in the previous year in which the tumor has never shown up in control rodents. For example, if there were 400 historical controls in addition to the 50 concurrent controls, none with a tumor, the one-sided p-value becomes 0.001 (see Table below). Statistical and biological significance are now compatible. In general, it can be inappropriate simply to pool historical and concurrent information. However, Bayes and empirical Bayes methods may be used to structure a valid synthesis

	C	E	Total
Tumor	0	3	3
No Tumor	450	47	497
	450	50	500

# Motivating vignettes

## Attenuation Adjustment

- In a standard errors-in-variables simple linear regression model, the least squares estimate of the regression slope ( $\beta$ ) is biased toward 0, an example of attenuation (the reduction of the force, effect, or value of something). More formally, suppose the true regression is  $Y = \beta x + \epsilon$ ,  $\epsilon \sim N(0, \sigma_\epsilon^2)$ , but  $Y$  is regressed not on  $x$  but on  $X \equiv x + \delta$ , where  $\delta \sim N(0, \sigma_\delta^2)$ . Then the least squares estimate  $\hat{\beta}$  has expectation  $E[\hat{\beta}] \simeq \rho\beta$ , with

$$\rho = \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_\delta^2}$$

If  $\rho$  is known or well-estimated, one can correct for attenuation and produce an unbiased estimate by using  $\hat{\beta}/\rho$  to estimate  $\beta$ .

## Motivating vignettes

- Though unbiasedness is an attractive property, especially when the standard error associated with the estimate is small, in general it is less important than having the estimate “close” to the true value. The expected squared deviation between the true value and the estimate (mean squared error, or MSE) provides an effective measure of proximity. Fortunately for our intuition, MSE can be written as the sum of an estimator’s sampling variance and its squared bias,

$$MSE = \text{variance} + (\text{bias})^2$$

- The unbiased estimate sets the second term to 0, but it can have a very large MSE relative to other estimators; in this case, because dividing  $\hat{\beta}$  by  $\rho$  inflates the variance as the price of eliminating bias. Bayesian estimators typically strike an effective tradeoff between variance and bias.

## ■ Defining Approaches

- Three principal approaches to inference guide modern data analysis: frequentist, Bayesian, and likelihood.
- The **frequentist** evaluates procedures based on imagining repeated sampling from a particular model (the likelihood), which defines the probability distribution of the observed data conditional on unknown parameters. Properties of the procedure are evaluated in this repeated sampling framework for fixed values of unknown parameters; good procedures perform well over a broad range of parameter values.

## ■Defining Approaches

- The **Bayesian** requires a sampling model and, in addition, a prior distribution on all unknown quantities in the model (parameters and missing data). The prior and likelihood are used to compute the conditional distribution of the unknowns given the observed data (the posterior distribution), from which all statistical inferences arise. Allowing the observed data to play some role in determining the prior distribution produces the empirical Bayes (EB) approach. The Bayesian evaluates procedures over repeated sampling of unknowns from the posterior distribution for a given data set. The empirical Bayesian may also evaluate procedures under repeated sampling of both the data and the unknowns from their joint distribution.

## ■Defining Approaches

- The likelihoodist (or Fisherian) develops a sampling model but not a prior, as does the frequentist. However, inferences are restricted to procedures that use the data only as reported by the likelihood, as a Bayesian would. Procedure evaluations can be from a frequentist, Bayesian, or EB point of view.

## ■The Bayes-frequentist controversy

- While probability has been the subject of study for hundreds of years (most notably by mathematicians retained by rich noblemen to advise them on how to maximize their winnings in games of chance), statistics is a relatively young field. Linear regression first appeared in the work of Francis Galton in the late 1800s, with Karl Pearson adding correlation and goodness-of-fit measures around the turn of the last century. The field did not really blossom until the 1920s and 1930s, when R.A. Fisher developed the notion of likelihood for general estimation, and Jerzy Neyman and Egon Pearson developed the basis for classical hypothesis testing. A flurry of research activity was energized by the World War II, which generated a wide variety of difficult applied problems and the first substantive government funding for their solution in the United States and Great Britain.

## ■The Bayes-frequentist controversy

- By contrast, Bayesian methods are much older, dating to the original 1763 paper by the Rev. Thomas Bayes, a minister and amateur mathematician. The area generated some interest by Laplace, Gauss, and others in the 19th century, but the Bayesian approach was ignored (or actively opposed) by the statisticians of the early 20th century. Fortunately, during this period several prominent non-statisticians, most notably Harold Jeffreys (a physicist) and Arthur Bowley (an econometrician), continued to lobby on behalf of Bayesian ideas (which they referred to as “inverse probability”). Then, beginning around 1950, statisticians such as L.J. Savage, Bruno de Finetti, Dennis Lindley, and many others began advocating Bayesian methods as remedies for certain deficiencies in the classical approach. The following example discusses the case of interval estimation.

## ■The Bayes-frequentist controversy

**Example 1.1** Suppose  $X_i \stackrel{iid}{\sim} N(\theta, \sigma^2)$ ,  $i = 1, \dots, n$ , where  $N$  denotes the normal (Gaussian) distribution and *iid* stands for “independent and identically distributed.” We desire a 95% interval estimate for the population mean  $\theta$ . Provided  $n$  is sufficiently large (say, bigger than 30), a classical approach would use the confidence interval

$$\delta(\mathbf{x}) = \bar{x} \pm 1.96s/\sqrt{n},$$

## ■The Bayes-frequentist controversy

where  $\mathbf{x} = (x_1, \dots, x_n)$ ,  $\bar{x}$  is the sample mean, and  $s$  is the sample standard deviation. This interval has the property that, on average over repeated applications,  $\delta(\mathbf{x})$  will fail to capture the true mean  $\theta$  only 5% of the time. An alternative interpretation is that, *before* any data are collected, the probability that the interval contains the true value is 0.95. This property is attractive in the sense that it holds for *all* true values of  $\theta$  and  $\sigma^2$ .

## ■The Bayes-frequentist controversy

On the other hand, its use in any single data-analytic setting is somewhat difficult to explain and understand. After collecting the data and computing  $\delta(\mathbf{x})$ , the interval either contains the true  $\theta$  or it does not; its coverage probability is not 0.95, but either 0 or 1. After observing  $\mathbf{x}$ , a statement like, “the true  $\theta$  has a 95% chance of falling in  $\delta(\mathbf{x})$ ,” is not valid, though most people (including most statisticians irrespective of their philosophical approach) interpret a confidence interval in this way. Thus, for the frequentist, “95%” is not a conditional coverage probability, but rather a tag associated with the interval to indicate either how it is likely to perform before we evaluate it, or how it would perform over the long haul. A 99% frequentist interval would be wider, a 90% interval narrower, but, conditional on  $\mathbf{x}$ , all would have coverage probability 0 or 1. ■

## ■The Bayes-frequentist controversy

By contrast, Bayesian confidence intervals (known as “credible sets,” and discussed further in Subsection 2.3.2) are free of this awkward frequentist interpretation. For example, conditional on the observed data, the probability is 0.95 that  $\theta$  is in the 95% credible interval. Of course, this natural interpretation comes at the price of needing to specify a (possibly quite vague) prior distribution for  $\theta$ .

# ■The Bayes-frequentist controversy

**Example 1.2** Consider the following simple experiment, originally suggested by Lindley and Phillips (1976), and reprinted many times. Suppose in 12 independent tosses of a coin, I observe 9 heads and 3 tails. I wish to test the null hypothesis  $H_0 : \theta = 1/2$  versus the alternative hypothesis  $H_a : \theta > 1/2$ , where  $\theta$  is the true probability of heads. Given only this much information, two choices for the sampling distribution emerge:

1. *Binomial*: The number  $n = 12$  tosses was fixed beforehand, and the random quantity  $X$  was the number of heads observed in the  $n$  tosses. Then  $X \sim \text{Bin}(12, \theta)$ , and the likelihood function is given by

$$L_1(\theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} = \binom{12}{9} \theta^9 (1 - \theta)^3. \quad (1.2)$$

2. *Negative binomial*: Data collection involved flipping the coin until the third tail appeared. Here, the random quantity  $X$  is the number of heads required to complete the experiment, so that  $X \sim \text{NegBin}(r = 3, \theta)$ ,

## ■The Bayes-frequentist controversy

with likelihood function given by

$$L_2(\theta) = \binom{r+x-1}{x} \theta^x (1-\theta)^r = \binom{11}{9} \theta^9 (1-\theta)^3. \quad (1.3)$$

Under either of these two alternatives, we can compute the  $p$ -value corresponding to the rejection region, “Reject  $H_0$  if  $X \geq c$ .” Doing so using the binomial likelihood (1.2), we obtain

$$\alpha_1 = P_{\theta=\frac{1}{2}}(X \geq 9) = \sum_{j=9}^{12} \binom{12}{j} \theta^j (1-\theta)^{12-j} = .075,$$

while for the negative binomial likelihood (1.3),

$$\alpha_2 = P_{\theta=\frac{1}{2}}(X \geq 9) = \sum_{j=9}^{\infty} \binom{2+j}{j} \theta^j (1-\theta)^3 = .0325.$$

## ■The Bayes-frequentist controversy

Thus, using the “usual” Type I error level  $\alpha = .05$ , we see that the two model assumptions lead to two different decisions: we would reject  $H_0$  if  $X$  were assumed negative binomial, but not if it were assumed binomial. But there is no information given in the problem setting to help us make this determination, so it is not clear which analysis the frequentist should regard as “correct.” In any case, assuming we trust the statistical model, it does not seem reasonable that how the experiment was monitored should have any bearing on our decision; surely only its *results* are relevant! Indeed, the likelihood functions tell a consistent story, since (1.2) and (1.3) differ only by a multiplicative constant that does not depend on  $\theta$ . ■

## ■ The Bayes-frequentist controversy

- A Bayesian explanation of what went wrong in the previous example would be that the Neyman-Pearson approach allows unobserved outcomes to affect the rejection decision. That is, the probability of  $X$  values “more extreme” than  $\theta$  (the value actually observed) was used as evidence against  $H_0$  in each case, even though these values did not occur. More formally, this is a violation of a statistical axiom known as the *Likelihood Principle*
- **The Likelihood Principle** states that once the data value  $x$  has been observed, the likelihood function  $L(\theta|x)$  contains all relevant experimental information delivered by  $x$  about the unknown parameter  $\theta$ .

## ■ The Bayes-frequentist controversy

- In the previous example,  $L_1$  and  $L_2$  are proportional to each other as functions of  $\theta$ , hence are equivalent in terms of experimental information (recall that multiplying a likelihood function by an arbitrary function  $h(x)$  does not change the MLE ( $\hat{\theta}$ )). Yet in the Neyman-Pearson formulation (power of hypothesis), these equivalent likelihoods lead to two different inferences regarding  $\theta$ . Put another way, frequentist test results actually depend not only on what  $x$  was observed, but on how the experiment was stopped.

## ■Advantages of Bayesian Approach

- ① Bayesian methods provide the user with the ability to formally incorporate prior information.
- ② Inferences are conditional on the actual data.
- ③ The reason for stopping the experimentation does not affect Bayesian inference
- ④ Bayesian answers are more easily interpretable by nonspecialists (a concern in Example 1.1)
- ⑤ All Bayesian analyses follow directly from the posterior; no separate theories of estimation, testing, multiple comparisons, etc. are needed.
- ⑥ Any question can be directly answered through Bayesian analysis.
- ⑦ Bayes procedure possesses numerous optimality properties.

## ■ Some basic Bayesian models

NPA MonteCarlo

# Markov Chain Monte Carlo: The Bayes Approaches

**Prof. Dr. Narayan Prasad Adhikari**

Central Department of Physics

Tribhuvan University Kirtipur, Kathmandu, Nepal

This lecture note is based on Text Book

January 11, 2023



# ■Outline

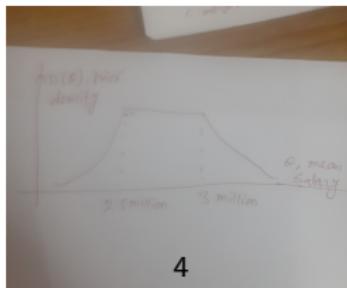
- Introduction
- Bayesian Inference
- Hierarchical modeling
- Model Assessment

# ■Introduction

- There is an approach - the Bayesian approach for statistical analysis. According to it, uncertainty is attributed not only to the data but also to the unknown parameter  $\theta$ . Some values of  $\theta$  are more likely than others. Then, as long as we talk about the likelihood, we can define a whole distribution of values of  $\theta$ . We call it a prior distribution, and it reflects our ideas, beliefs, and past experiences about the parameter before we collect and use the data.

# ■Introduction: An Example

- What do you think is the average starting annual salary of a Data Science graduate in Nepal? Is it Rs. 2,000,000 per year? Unlikely, that's too low. Perhaps, Rs. 4,000,000 per year? No, that's too high for a fresh graduate. Between Rs. 2.5 millions and three millions sounds like a reasonable range. We can certainly collect data on 100 recent graduates, compute their average salary and use it as an estimate, but before that, we already have our beliefs on what the mean salary may be. We can express it as some distribution with the most likely range between Rs. 2.5 millions and three millions

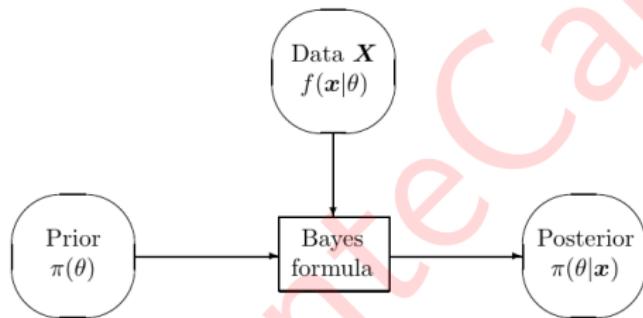


## ■Introduction: An Example

- One benefit of this approach is that we no longer have to explain our results in terms of a “long run”. Often we collect just one sample for our analysis and don’t experience any long runs of samples. Instead, with the Bayesian approach, we can state the result in terms of the distribution of parameter  $\theta$ . For example, we can clearly state the probability for a parameter to belong to a certain interval, or the probability that the hypothesis is true. This would have been impossible under the frequentist approach.
- Another benefit is that we can use both pieces of information, the data and the prior, to make better decisions. In Bayesian statistics, decisions are

$$\delta = \delta(\text{data}, \text{prior distribution})$$

# ■Introduction



Now we have two sources of information to use in our Bayesian inference:

- ① collected and observed data;
- ② prior distribution of the parameter.

These two pieces are combined via the Bayes formula

Bayes  
Rule

$$P\{B | A\} = \frac{P\{A | B\} P\{B\}}{P\{A\}}$$

## ■Introduction

Prior to the experiment, our knowledge about the parameter  $\theta$  is expressed in terms of the prior distribution (prior pmf or pdf)

$$\pi(\theta).$$

The observed sample of data  $\mathbf{X} = (X_1, \dots, X_n)$  has distribution (pmf or pdf)

$$f(\mathbf{x}|\theta) = f(x_1, x_2, \dots, x_n|\theta).$$

This distribution is conditional on  $\theta$ . That is, different values of the parameter  $\theta$  generate different distributions of data, and thus, conditional probabilities about  $\mathbf{X}$  generally depend on the condition,  $\theta$ .

Observed data add information about the parameter. The updated knowledge about  $\theta$  can be expressed as the posterior distribution.

## ■Introduction

$$\pi(\theta|\mathbf{x}) = \pi(\theta|\mathbf{X=x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})} \quad (1)$$

Posterior distribution of the parameter  $\theta$  is now conditioned on data  $\mathbf{X=x}$ . Naturally, conditional distributions  $f(\mathbf{x}-\theta)$  and  $\pi(\theta|\mathbf{x})$  are related via the Bayes Rule.

According to the Bayes Rule, the denominator of equation 1,  $m(\mathbf{x})$ , represents the unconditional distribution of data  $\mathbf{x}$ . This is the marginal distribution (pmf or pdf) of the sample  $\mathbf{x}$ . Being unconditional means that it is constant for different values of the parameter  $\theta$ . It can be computed by the Law of Total Probability or its continuous-case version.

# ■Introduction

## Marginal distribution of data

$$m(\mathbf{x}) = \sum_{\theta} f(\mathbf{x}|\theta)\pi(\theta)$$

for discrete prior distributions  $\pi$

$$m(\mathbf{x}) = \int_{\theta} f(\mathbf{x}|\theta)\pi(\theta)d\theta$$

for continuous prior distributions  $\pi$

## ■Introduction: An Example

- A manufacturer claims that the shipment contains only 5% of defective items, but the inspector feels that in fact it is 10%. We have to decide whether to accept or to reject the shipment based on  $\theta$ , the proportion of defective parts. Before we see the real data, let's assign a 50-50 chance to both suggested values of  $\theta$  i.e.

$$\pi(0.05) = \pi(0.10) = 0.5.$$

A random sample of 20 parts has 3 defective ones. Calculate the posterior distribution of  $\theta$

## ■Introduction: An Example

### Solution:

- Apply the Bayes formula. Given  $\theta$ , the distribution of the number of defective parts X is Binomial( $n = 20, \theta$ ). For  $x = 3$ , we have (from table of binomial distribution):

$$f(x|\theta = 0.05) = F(3|\theta = 0.05) - F(2|\theta = 0.05) = \\ 0.9841 - 0.9245 = 0.0596 \text{ and } f(x|\theta = 0.10) = F(3|\theta = 0.10) - F(2|\theta = 0.10) = 0.8670 - 0.6769 = 0.1901$$

## ■Introduction: An Example

The **marginal distribution** of  $X$  (for  $x = 3$ ) is

$$\begin{aligned}m(3) &= f(x \mid 0.05)\pi(0.05) + f(x \mid 0.10)\pi(0.10) \\&= (0.0596)(0.5) + (0.1901)(0.5) = 0.12485.\end{aligned}$$

Posterior probabilities of  $\theta = 0.05$  and  $\theta = 0.10$  are now computed as

$$\pi(0.05 \mid X = 3) = \frac{f(x \mid 0.05)\pi(0.05)}{m(3)} = \frac{(0.0596)(0.5)}{0.1248} = 0.2387;$$

$$\pi(0.10 \mid X = 3) = \frac{f(x \mid 0.10)\pi(0.10)}{m(3)} = \frac{(0.1901)(0.5)}{0.1248} = 0.7613.$$

## ■Introduction: An Example

- **Conclusion:** In the beginning, we had no preference between the two suggested values of  $\theta$ . Then we observed a rather high proportion of defective parts,  $3/20=15\%$ . Taking this into account,  $\theta = 0.10$  is now about three times as likely than  $\theta = 0.05$ .
- Bayesian approach presumes a prior distribution of the unknown parameter. Adding the observed data, the Bayes Theorem converts the prior distribution into the posterior which summarizes all we know about the parameter after seeing the data. Bayesian decisions are based on this posterior, and thus, they utilize both the data and the prior.

## ■Introduction: An Example

- Ultrasound tests done near the end of the first trimester of a pregnancy are often used to predict the sex of the baby. However, the errors made by radiologists in reading ultrasound results are not symmetric, in the following sense: girls are virtually always correctly identified as girls, while boys are sometimes misidentified as girls (in cases where the gender organ is not clearly visible, perhaps due to the child's position in the womb). More specifically, a leading radiologist states that

$$P(\text{test}+|G) = 1 \text{ and } P(\text{test}+|B) = 0.25$$

where “test +” denotes that the ultrasound test predicts the child is a girl. Thus, we have a 25% false positive rate for girl, but no false negatives.

## ■Introduction: An Example

- Suppose a particular woman's test comes back positive for girl, and we wish to know the probability she is actually carrying a girl. Assuming 48% of babies are girls, we can use Bayes Rules where "boy" and "girl" provide the  $J = 2$  mutually exclusive and exhaustive cases
- Here we need to find  $P(G|+)$

## ■Introduction: An Example

From Bayes Rule we have

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A)P(B)} \quad (2)$$

Here  $P(A) = P(G)$  and  $P(B) = P(B)$  i.e. Boys in this case). Further its given that  $P(A)$  i.e. probability of A girl is  $P(A) = P(G) = 0.48$  i.e. Boys is  $P(B) = P(\text{Boys}) = 0.52$ . (From previous observed data). Therefore

$$\begin{aligned} P(G|test+) &= \frac{P(test+|G)P(G)}{P(test+|G)P(G) + P(test+|B)P(B)} \\ &= \frac{(1) \times (0.48)}{(1) \times (0.48) + (0.25) \times (0.52)} \\ &= 0.787 \end{aligned} \quad (3)$$

This means there is only 78.7% chance that the baby is, in fact, a Girl.

## ■Introduction- Example: 3

Consider normal (Gaussian) likelihood :

$$f(y|\theta) = \frac{1}{\sqrt{(2\pi)\sigma}} \exp\left(-\frac{(y-\theta)^2}{2\sigma^2}\right) \quad (4)$$

Suppose we take

$$\pi(\theta|\eta) = N(\theta|\mu, \tau^2) \quad (5)$$

where  $\mu$  and  $\tau$  are hyperparameters so that  $\eta = (\mu, \tau)$ . Now using idea of Bayesian posterior from above prior and likelihood we get posterior distribution for  $\theta$  i.e.

$$p(\theta|y) = N\left(\theta \mid \frac{\sigma^2\mu + \tau^2y}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right) \quad (6)$$

## ■Introduction- Example: 3

Consider  $\mu = 2, \tau = 1, y = 6$ , and  $\sigma = 1$ . Plot prior (centered at  $\theta = 2$ ), the likelihood (centered at  $\theta = 6$ ) and posterior arising from above formula. Discuss your results.

## ■Prior distribution: Example 2.7 (Text)

**Example 2.7** Suppose that  $X$  is the number of pregnant women arriving at a particular hospital to deliver their babies during a given month. The discrete count nature of the data plus its natural interpretation as an arrival rate suggest adopting a Poisson likelihood,

$$f(x|\theta) = \frac{e^{-\theta}\theta^x}{x!}, \quad x \in \{0, 1, 2, \dots\}, \quad \theta > 0.$$

To effect a Bayesian analysis, we require a prior distribution for  $\theta$  having support on the positive real line. A reasonably flexible choice is provided

## ■Prior distribution: Example 2.7 (Text)

by the gamma distribution,

$$\pi(\theta) = \frac{\theta^{\alpha-1} e^{-\theta/\beta}}{\Gamma(\alpha)\beta^\alpha}, \quad \theta > 0, \alpha > 0, \beta > 0,$$

or  $\theta \sim G(\alpha, \beta)$  in distributional shorthand. Note that we have suppressed  $\pi$ 's dependence on  $\eta = (\alpha, \beta)$  since we assume it to be known. The gamma distribution has mean  $\alpha\beta$ , variance  $\alpha\beta^2$ , and can have a shape that is either one-tailed ( $\alpha \leq 1$ ) or two-tailed ( $\alpha > 1$ ); for large  $\alpha$  the distribution resembles a normal distribution. The  $\beta$  parameter is a *scale* parameter, stretching or shrinking the distribution relative to 0, but not changing its shape. Using Bayes' Theorem (2.1) to obtain the posterior density, we have

$$\begin{aligned} p(\theta|x) &\propto f(x|\theta)\pi(\theta) \\ &\propto (e^{-\theta}\theta^x) \left( \theta^{\alpha-1} e^{-\theta/\beta} \right) \\ &= \theta^{x+\alpha-1} e^{-\theta(1+1/\beta)}. \end{aligned} \tag{2.10}$$

## ■Prior distribution: Example 2.7 (Text)

Notice that since our intended result is a normalized function of  $\theta$ , we are able to drop any multiplicative functions that do not depend on  $\theta$ . (For example, in the first line we have dropped the marginal distribution  $m(x)$  in the denominator, since it is free of  $\theta$ .) But now looking at (2.10), we see that it is proportional to a gamma distribution with parameters  $\alpha' = x + \alpha$  and  $\beta' = (1 + 1/\beta)^{-1}$ . Note this is the *only* function proportional to (2.10) that still integrates to 1. Because density functions uniquely determine distributions, we know that the posterior distribution for  $\theta$  is indeed  $G(\alpha', \beta')$ , and that the gamma is the conjugate family for the Poisson likelihood.

## ■Prior distribution: Example 2.7 (Text)

As a concrete illustration, suppose we observe  $x = 42$  moms arriving at our hospital to deliver babies during December 2007. Suppose we adopt a  $G(5, 6)$  prior, which has mean  $5(6) = 30$  and variance  $5(6^2) = 180$  (see

**Now you write python code to above problem. Plot prior and posterior distribution for  $\theta = 0, 100$**

## ■Introduction: An Example

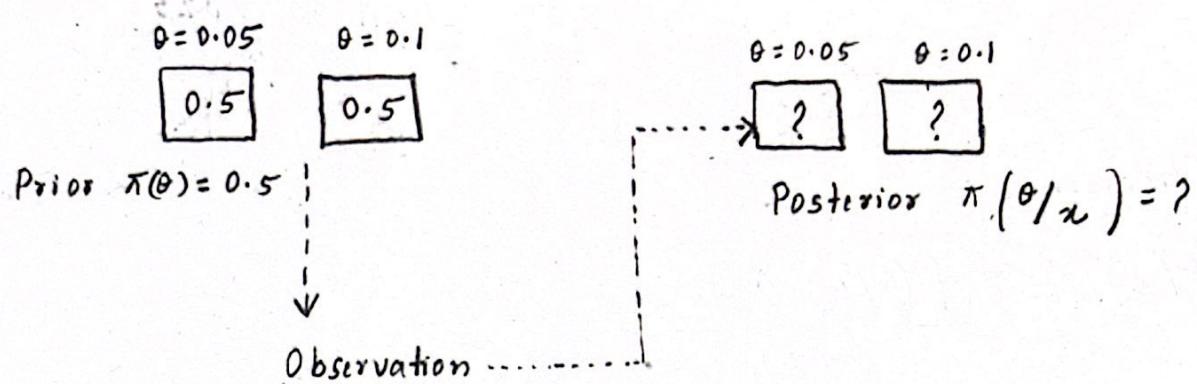
### Example 1

- A manufacturer claims that the shipment contains only 5% of defective items, but the inspector feels that in fact it is 10%. We have to decide whether to accept or to reject the shipment based on  $\theta$ , the proportion of defective parts. Before we see the real data, let's assign a 50-50 chance to both suggested values of  $\theta$  i.e.

$$\pi(0.05) = \pi(0.10) = 0.5.$$

A random sample of 20 parts has 3 defective ones. Calculate the posterior distribution of  $\theta$

### Solution



Let,  $\theta$  represent the proportion of defective items.  
We know,  $\pi(\theta = 0.05) = \pi(\theta = 0.10) = 0.5$

A random sample of 20 parts ( $n$ ) has 3 defective items ( $x$ ). This process follows binomial distribution i.e.  $X \sim B(n, \theta)$ .

Now,

for  $\theta = 0.05$ ,

$$\begin{aligned} \text{likelihood for data, } f(x=3/\theta=0.05) &= {}^n C_x \theta^x (1-\theta)^{n-x} \\ &= 20 C_3 (0.05)^3 (0.95)^{17} \\ &= 0.0595 \end{aligned}$$

for  $\theta = 0.10$ ,

$$\begin{aligned} \text{likelihood for data, } f(x=3/\theta=0.10) &= {}^n C_x \theta^x (1-\theta)^{n-x} \\ &= 20 C_3 (0.1)^3 (0.9)^{17} \end{aligned}$$

$$= 0.1901$$

Now,

Marginal probability of the data,  $m(x=3)$

$$= \sum_{\theta} f(x=3|\theta) \pi(\theta)$$

$$= f(x=3|\theta=0.05) \pi(\theta=0.05)$$

$$+ f(x=3|\theta=0.1) \pi(\theta=0.1)$$

$$= 0.0595 * 0.5 + 0.1901 * 0.5$$

$$= 0.1248$$

Thus, Posterior distribution of  $\theta$  is given as follows:

for  $\theta = 0.05$ ,

$$\pi(\theta=0.05|x=3) = \frac{f(x=3|\theta=0.05) \pi(\theta=0.05)}{m(x=3)}$$

$$= \frac{0.0595 * 0.5}{0.1248}$$

$$= 0.2383$$

for  $\theta = 0.10$ ,

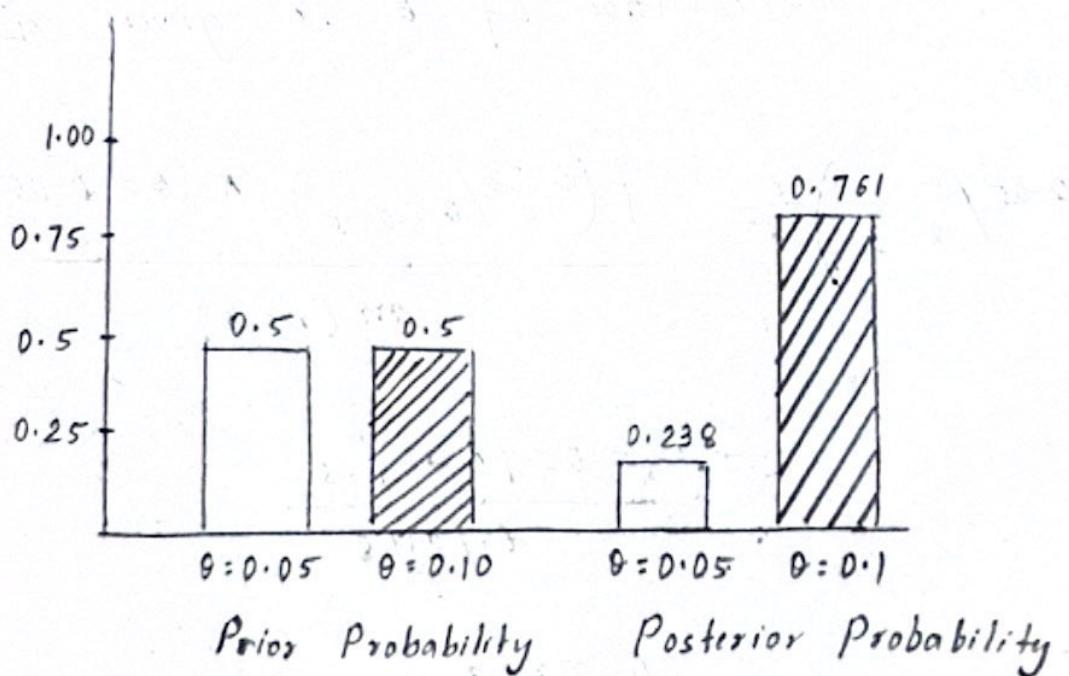
$$\pi(\theta = 0.10/x=3) = \frac{f(x=3/\theta=0.10) \pi(\theta=0.10)}{\pi(x=3)}$$

$$= \frac{0.1901 * 0.5}{0.1248}$$

$$= 0.7616$$

Posterior distribution of  $\theta$

$\theta$	0.05	0.10
$\pi(\theta/x)$	0.2383	0.7616



Interpretation: The manufacturer's claim that the defective item is 5% is ~~more than~~ almost three times less likely than inspector's

claim of 10%. Thus, it is better for the customer to  
reject the <sup>(O. 761)</sup> shipment.

- Ultrasound tests done near the end of the first trimester of a pregnancy are often used to predict the sex of the baby. However, the errors made by radiologists in reading ultrasound results are not symmetric, in the following sense: girls are virtually always correctly identified as girls, while boys are sometimes misidentified as girls (in cases where the gender organ is not clearly visible, perhaps due to the child's position in the womb). More specifically, a leading radiologist states that

$$P(\text{test}+|G) = 1 \text{ and } P(\text{test}+|B) = 0.25$$

where “test +” denotes that the ultrasound test predicts the child is a girl. Thus, we have a 25% false positive rate for girl, but no false negatives.

## Example 2

14

### ■Introduction: An Example

- Suppose a particular woman's test comes back positive for girl, and we wish to know the probability she is actually carrying a girl. Assuming 48% of babies are girls, we can use Bayes Rules where “boy” and “girl” provide the  $J = 2$  mutually exclusive and exhaustive cases
- Here we need to find  $P(G|+)$

Example 2

Solution

$$\pi(G_1) = 0.48 \quad \pi(B) = 0.52 \quad (\text{Prior})$$

Also given are the followings,

$$P(\text{tut+} | G_1) = 1$$

$$\text{and } P(\text{tut+} | B) = 0.25$$

$$\text{Now, } P(G_1 | \text{tut+}) = ?$$

$$\begin{aligned} \text{Marginal Probability for data } m(\text{tut+} = 1) &= P(\text{tut+} | G_1) \pi(G_1) \\ &\quad + P(\text{tut+} | B) \pi(B) \\ &= \frac{1 * 0.48 + 0.25 *}{0.52} \\ &= 0.61 \end{aligned}$$

$$\text{Now, } P(G_1 | \text{tut+}) = \frac{P(\text{tut+} | G_1) \pi(G_1)}{m(\text{tut+})}$$

$$= \frac{1 * 0.48}{0.61}$$

$$= 0.7868$$

Interpretation : For a positively tested result, the probability that the embryo is actually a girl is 0.7868.