Numerical integration
Grid based methods

Consider the 1D integration

$$F = \int_a^b f(x)dx \tag{1}$$

Rieman sum: Let a closed interval be partitioned by points $a < x_1 < x_2, \cdots < x_{N-1} < b$, where the lengths of the resulting intervals between the points are denoted by Δx_1, Δx_2, ..., Δx_k, ..., Δx_N. Let Δx_k^* be an arbitrary point in the kth subinterval. Then the quantity

$$\sum_{k=1}^{N} f(x_k^*)\Delta x_k \tag{2}$$

is called a Riemann sum for a given function and partition, and the value $max\ \Delta x_k$ is called the mesh size of the partition.

If the limit of the Riemann sums exists as $max\ \Delta x_k \to 0$, this limit is known as the Riemann integral of $f(x)$ over the interval $[a, b]$.

So the previous integral can be evaluated as

$$F = \lim_{N \to \infty} F_N \tag{3}$$

$$F_N = \frac{b-a}{N} \sum_{k=1}^{N} f(x_k) \tag{4}$$

Suppose a 1D integral is performed with n grid points.

A d-dim integral with n grid points in each dimension: time taken will scale as ??

Suppose the 1D integral takes 2 seconds.
A 10-dim integral will take ~ 1024 seconds.
A 100-dim integral will take $\sim 10^{30}$ seconds $\sim 4 \times 10^{22}$ years !
Age of the universe is ??
Hopeless situation !!!

Can we do something smarter?

Rather than evaluating the function at a fixed set of grid points, we can evaluate it at a random set of points $\mathbf{X}_k$ drawn from a given probability density function.

Question is: **what probability density function to use?**

Uniform sampling:
A set of N points $\mathbf{X}_k$ can be chosen from a uniform probability density function. Different set of N points will give different results for F_N. So there will be a *statistical error* associated with such *stochastic evaluation* of the integral.

A concrete example

$$F = erf(y) = \frac{2}{\sqrt{\pi}} \int_0^y e^{-x^2} dx \qquad (5)$$

F evaluated by randomly choosing 10,000 points from a uniform distribution in the interval $[0, y]$.

y	Uniform sampling MC	Exact	Error (%)
0.01	0.0113	0.0113	0
0.10	0.1125	0.1125	0
1.00	0.8420	0.8427	0.08
2.00	0.9900	0.9953	5
10.0	1.0297	1.0000	2.97
100.0	1.1272	1.0000	12.72

Why is the agreement getting worse with increasing y?

The function e^{-x^2} is peaked at $x = 0$, we are sampling points uniformly over the entire interval $[0, y]$.

So we are *wasting* our efforts in regions where the function contributes very little to the integral.

We need to devise a strategy in which we sample more points x_k where the function value is large.

Importance sampling

Suppose we have a function $w(x)$ defined over the interval $[a, b]$ so that $w(x) \approx f(x)$.

Also assume we can generate **(pseudo)random numbers** x_k from the normalized probability density function

$$p(x) = \frac{w(x)}{\int_a^b w(x)dx}, \tag{6}$$

where $p(x)$ is defined over $[a, b]$.

How to generate such **(pseudo)random numbers** will be discussed later.

To the extent that $w \approx f$, there will be more 'random grid points' in regions where $f(x)$ is large.

Define a function $g \equiv f/p$. Then $f = gp$, and the integral becomes

$$F = \int_a^b g(x)p(x)dx \approx \frac{1}{M}\sum_{k=1}^{M} g(x_k) \qquad (7)$$

The continuous integral is replaced by the average of g over a finite set of points x_i which are chosen from the probability density function $p(x)$.

What have we achieved using importance sampling?

In our example of $erf(y)$, let us use $w(x) = e^{-x}$.

y	US MC	IS MC	Exact
0.01	0.0113	0.0113	0.0113
0.10	0.1125	0.1125	0.1125
1.00	0.8420	0.8427	0.8427
2.00	0.9900	0.9919	0.9953
10.0	1.0297	0.9996	1.0000
100.0	1.1272	1.0019	1.0000

We will discuss the idea of statistical error later.

Examples from physics

In statistical mechanics, thermal averages of observables are calculated as

$$\langle O \rangle = \frac{\int O e^{-\beta E} dE}{\int e^{-\beta E} dE} \tag{8}$$

In quantum mechanics, expectation value of the Hamiltonian in a state $|\Psi\rangle$ is calculated as

$$\langle H \rangle_\Psi = \frac{\int \Psi^* H \Psi d\vec{R}}{\int |\Psi|^2 d\vec{R}} \tag{9}$$

If Ψ is an n-electron wave function, then the above integral is $3n$-dimensional.

The above two integrals are exactly like Eq. 7 with $p(x) = \frac{e^{-\beta E}}{\int e^{-\beta E} dE}$ in the first one.

Eq. 9 can be rewritten as

$$\langle H \rangle_\Psi = \frac{\int \frac{H\Psi}{\Psi} |\Psi|^2 d\vec{R}}{\int |\Psi|^2 d\vec{R}} \tag{10}$$

This is in the form of Eq. 7 with $p(x) = \frac{|\Psi|^2}{\int |\Psi|^2 d\vec{R}}$.

To evaluate such integrals:
For Eq. 8 generate E (micro-states of the system with energy E) with probability $\frac{e^{-\beta E}}{\int e^{-\beta E} dE}$ and use

$$\langle O \rangle = \frac{\int O e^{-\beta E} dE}{\int e^{-\beta E} dE} = \frac{1}{M} \sum_{k=1}^{M} O(E_k). \qquad (11)$$

For Eq. 9 use

$$\langle H \rangle_\Psi = \frac{\int \frac{H\Psi}{\Psi} |\Psi|^2 d\vec{R}}{\int |\Psi|^2 d\vec{R}} = \frac{1}{M} \sum_{k=1}^{M} \frac{H\Psi(\vec{R}_k)}{\Psi(\vec{R}_k)}. \quad (12)$$

electron configurations $\vec{R}$ with probability $\frac{|\Psi|^2}{\int |\Psi|^2 d\vec{R}}$

The task now is to be able to generate (pseudo)-random variables from arbitrary probability density functions.

Basics of probability and statistics

Set *A collection or an aggregate of well-defined objects is called a set.*

k-variate population *A set whose every element is characterized by k characteristics is called a $k-$variate population*, where k is a positive integer. If $k = 1$ then the population is univariate. Statistical population is the reference set based on which statistical hypotheses are tested and statistical decisions are made.

We will NOT study testing of statistical hypotheses in these lectures.

Example 1 Let s be the set of height measurements of all individuals in the campus. s is an example of a univariate finite population.

Example 2 Let S be the set of height and weight measurements of all individuals. Then S is an example of a bivariate finite population.

Sample *A subset of a statistical population can be called a sample.* A sample, in general, means only a subset of a given statistical population. It may or may not be useful statistically.

Example: Let us take *example 1* above. A subset of that population can be the set s_1 of the height measurements of all people below 12 years of age. However, clearly, the mean of the heights in this sample cannot be taken as an estimate of the population mean.

Representative Sample As we just discussed, any subset of a statistical population is called a sample. A sample can thus be selected in many different ways. Often what one aims to do is to draw inferences about the population from an analysis of the sample. It may be necessitated by the fact that the population is so large that it is impossible to study every member in it. Or that the inference one is trying to

draw is not worth the time, money and effort one needs to spend in order to study or survey every member of the population. However, in such situations, for the inference drawn to be of any relevance to the population, the sample has to be 'representative' of the population in some sense. Let us take an example, suppose 8 out of 10 steel balls produced by a machine in a particular 1 minute interval are found to be defective. It would wrong to infer that 80% of all bullets produced by the machine are defective unless it is known that the functioning of the machine in that 1-minute interval is typical of its behavior. How to draw samples that are representative of the population is a question of sampling technique, and we will not worry about that here. We will learn some definitions for now.

Simple Random Sample Consider a population of size N (distinct elements) and let a sample of size n be taken. *If a sample of size n is obtained in such a way that all possible samples of size n had an equal chance of being selected then the sample obtained is called a simple random sample from the population under consideration*.

This is the idea of simple random sample from a statistical population. You should be able to appreciate this intuitively. What will be more relevant for us in the discussion of Monte Carlo methods is simple random sample from a theoretical population. We now proceed to define that.

Stochastic variables In order to understand the definition of a stochastic variable, let us understand the idea of random experiments first.

Random experiments: *A random experiment is a procedure which results in some nondeterministic outcomes in a particular situation. An outcome is a single realization of a phenomenon under consideration.*

The Outcome Set: *The set of all possible outcomes of a random experiment is called an outcome set.*

Question: If two coins are tossed simultaneously, what is the outcome set?

Event: *Event is a subset of the outcome set, and elementary event is an event consisting of only one element.*

Stochastic Variable *A stochastic variable (s.v. also called random variable, variate etc.) is a variable which is obtained as the outcome of a set of random experiments and can take a set of predefined values.*

Probability Function *Probability that a random variable takes a particular value is denoted by a probability function (or probability density function)* $p(x)$. $p(x) \geq 0$.

Usually, a probability function is normalized to unity:

$$\sum_i p(x_i) = 1 \tag{13}$$

if x is a discrete variable and

$$\int p(x)dx = 1 \tag{14}$$

if x is a continuous variable.

Measures of central tendency

The Mean (arithmetic mean) is

$$\mu = \int xp(x)dx \tag{15}$$

Moments about the Origin *r-th moment about the origin* is

$$\mu'_{\mathsf{r}} = \int x^r p(x)dx. \tag{16}$$

Thus mean is the 1st moment about the origin.

Central Moments *The r-th central moment* is defined as

$$\mu_{\mathsf{r}} = \int (x-\mu)^r p(x)dx. \tag{17}$$

The variance is defined as the second central moment,

$$\mu_2 = \sigma^2 = \int (x-\mu)^2 p(x)dx. \tag{18}$$

Some probability density functions we have mentioned, or will find useful...

Uniform distribution

$$f(x) = \frac{1}{\beta - \alpha}, \quad \beta > \alpha. \tag{19}$$

Gaussian or normal distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}. \tag{20}$$

This is also denoted by $N(\mu, \sigma^2)$: mean $= \mu$. variance$= \sigma^2$

Standard normal distribution Is the case of normal distribution when $\mu = 0$ and $\sigma = 1$,

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{21}$$

Expected value: *If X is a s.v. with probability function $f(X)$ and if $\phi(X)$ is a function of X which is again a s.v. then the mathematical expectation of $\phi(X)$ is defined as,*

$$E(\phi(X)) = \int_{-\infty}^{\infty} \phi(X) f(X) dX \qquad (22)$$

Statistically independent: Two s.v.s X and Y are said to be statistically independent if their joint probability density function can be written as a product of probability functions of X and Y.

If $h(x, y)$ denotes the joint probability function and if $f(x)$ and $g(y)$ are the individual probability functions of X and Y,
then X and Y are independent if,

$$h(x, y) = f(x)g(y). \qquad (23)$$

Theoretical Population

We will now consider the idea of a *theoretical population*, a population defined by a stochastic variable and its probability density function. So we can have *Uniform population*, *Normal population*, *Binomial population* etc.

Simple random sample: t*A set of* n *s.v.'s* $X_1, X_2, \ldots X_n$ *which are independently and identically distributed is said to be a simple random sample of size* n *from* $f(X)$*, if* $f(X)$ *is the common probability function.*

Further if $(X_1, X_2, \ldots X_n)$ is a simple random sample of size n from a population designated by a probability function $f(X)$, the joint probability function of the sample values is

$$h(X_1, X_2, \ldots X_n) = f(X_1) \ldots f(X_n) = \Pi_{i=1}^{n} f(X_i). \tag{24}$$

Observed random sample: *A set of numbers* $x_1,\ x_2,\ \ldots x_n$ *is said to be an observed random sample of size* n *from a theoretical population if* $x_1,\ x_2,\ \ldots x_n$ *are one set of values assumed by* $X_1,\ X_2,\ \ldots X_n$ *where* $X_1,\ X_2,\ \ldots X_n$ *is a simple random sample of size* n *from the same theoretical population*.

Note: Simple random sample $\rightarrow$ a set of stochastic variables.
Observed random sample $\rightarrow$ a set of numbers.

Statistic: *Any function of a simple random sample* X_1, $X_2, \ldots, X_n$, *which is again a stochastic variable is called a statistic.*

Example: Sample mean $\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n}$.
Sample variance $S^2 = \sum_{i=1}^{n} \frac{(X_i-\bar{X})^2}{n}$.

Theorem: *If* $(X_1,\ X_2,\ \ldots X_n)$ *is a random sample from a population with finite mean value* μ, *then the mean of the sample means is also* μ.

Estimator: *If* T *is a statistic, and if* $E(T) = \theta$ *(*θ *is some parameter), then* T *is called an unbiased estimator for* θ .

$\bar{X}$ is an unbiased estimator for μ.

Covariance: *The covariance between two variates X and Y, denoted by $Cov(X,Y)$, is*

$$\begin{aligned} & Cov(X,Y) \\ &= \int\int (X-\bar{X})(Y-\bar{Y}) f(x)g(y)dxdy \\ &= \int\int XY f(x)g(y)dxdy - \bar{X}\bar{Y}. \end{aligned}$$

Other notations are also used: $C(x,y)$, $\sigma_{x,y}$ etc.

What is $Cov(X,Y)$ when X and Y are statistically independent?

Theorem: *Whenever the population variance σ^2 is finite, the variance of the sample mean is $\frac{\sigma^2}{n}$.*

Standard error: *The positive square root of the variance of any statistic is called the standard error of the statistic.*

Standard error of $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$.

Theorem: $E(S^2) = \frac{n-1}{n}\sigma^2$, *where* S^2 *is the sample variance, and* σ^2 *is the population variance.*

Thus $E\left(\sum \frac{(X_i-\bar{X})^2}{n-1}\right) = \sigma^2$.

$\sum \frac{(X_i-\bar{X})^2}{n-1}$ is an unbiased estimator for σ^2,

$\sum \frac{(X_i-\bar{X})^2}{n}$ is not unbiased for σ^2.

The Central Limit Theorem: There are several versions of the *Central Limit Theorem*. We will mention two here.

1. *If a s.v.* x *has a large number of independent degrees of freedom, in other words, if* x *is a sum of a large number of s.v.'s, then the distribution of* x *goes asymptotically to a normal distribution*.

2. *Let* $X_1, X_2, \ldots X_n$ *be a simple random sample from a population, continuous or discrete, with finite variance* σ^2 *and a mean value* μ*. Then the sample mean* $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *is distributed according to* $N(\mu, \frac{\sigma^2}{n})$ in the limit $n \to \infty$.

We are back to the problem of generating configurations according to arbitrary probability density functions:

According to $e^{-\beta E}$ in case of thermal average,

$|\Psi(\vec{R})|^2$ for calculation of expectation value.

Let us look at the general problem of generating random numbers.

What are random numbers? It is a collection of digits between 0 and 9, each chosen with probability 1/10.

Not rigorous, but will serve our purpose.

How to generate random numbers? Time sequence on decay of nuclei in a radioactive material.

May be random numbers, but not useful practically, particularly for computations.

For computations we use *pseudo-random numbers*.

Important differences between random numbers and pseudo-random numbers.

Pseudo-random numbers	**Random numbers**
Reproducible	Not reproducible
Finite period	Not periodic

Many pseudo-random number generators pass required statistical tests to be useful for practical applications

RNG suitable in one case may not be suitable in another case.

We may need RN from various distributions:

1. Uniform random numbers $[-W, W]$ for electron moving in a disordered medium.

2. RN from distribution $e^{-\beta E}$.

3. Gaussian RN

Many such distributions can be generated from a set random numbers uniformly distributed between $[0, 1]$, which in turn can be generated from a set of uniform random integers distributed in $[0, m]$.

So first we will try to generate uniform random integers.

Using the system clock on a computer Try the following command (works on SUSE, doesn't seem to work on Ubuntu):

```
> date +%N
```

We will only get 10^6 random numbers.

This number is too small. This algo cannot be used during a computation.

If the system clock is read at regular intervals the random numbers generated will be highly correlated.

In many practical RNG routines, the system clock is called once and that number is used as a 'seed' for the generator.

von Neumann Algo Suppose we take an integer of several digits and square it.

The initial and final digits are often predictable. What are not predictable are the intermediate digits.

By chopping off the leading and trailing digits we take the middle part as the random number. We square this number and repeat the process. *Example* $(123)^2 = 15129$; $(512)^2 = 262144$.

This can, on occasions, lead into unwanted directions. Suppose one of the numbers in the sequence turns out to be 0. Then all subsequent numbers will be 0.

Linear Congruence Method

One of the most popular methods for generating random integers is the linear congruence method, which incorporates ideas from systems clock and middle of the square methods.

In this algo, given a number X_n in the sequence the next one is generated by

$$X_{n+1} = (aX_n + c) \mod m. \qquad (25)$$

m is a positive integer.
$0 < m$; $0 < a < m$; $0 \leq c < m$.
To start the sequence, we need a value for X_0 called the 'seed' of the RNG.

LCM, von Neumann algo and the systems clock ideas: Instead of taking the square, the number X_n is multiplied by a in LCM.
To prevent a bad sequence (all zeros) developing, an increment c is added.
Somewhat in the spirit of the systems clock approach, the least significant part of the sum is retained.

The quality of random numbers generated by LCM critically depends on the values of a, c and m.
The range of integers generated by this method is $[0, m-1]$.

Some judicious set of values for a, c and m

In order to have a long period:

1. c and m must be relative prime to each other.

2. If p is a prime factor of m, then p should be a prime factor of $a - 1$.

3. If 4 divides m, then 4 divides $a - 1$.

Conversion to random FP numbers Random integers in the range $[0, m-1]$ as generated above can be converted to random numbers in the range $[0, 1]$ by simply dividing each number by $m-1$. Sometimes, a small variant of the algorithm given in eq. (25) is used.

$$X_{n+1} = \frac{aX_n + c}{d} \mod m, \qquad (26)$$

where d is another integer.

Park and Miller have shown that the use of increment c is not essential if the values of a and m are chosen very carefully (NR).

ran0 in NR is based on such an algo.

Improving the randomness

Improving Randomness by Shuffling

LCM is a very efficient way of generating random numbers, but it has certain known limitations. In order to improve upon it, several modifications have been suggested. We will mention a few here.

- Instead of using the numbers from a LCM on the fly, store L of them in an array of length L. While using, generate a random integer j in the range $[1, L]$, and use the number at the j-th location of the array. Once the number at the j-th location is used, generate another random number to be stored in the j-th location.

  ran1 in NR is based on this.

- In order to increase the period, two different sequences with different periods can be combined. The idea is to add the two sequences modulo the *modulus* of either one of them. If overflow in the intermediate stage is a concern, the sequences can be subtracted, and $m - 1$ can be added to any negative numbers. This way, the period becomes the lowest common multiple of the two periods.

  ran2 in the NR uses this algo.

- Another way to improve upon LCM is to use more than one previous members of the sequence. For example, define

$$X_{n+1} = (X_n + X_{n-k}) \mod m, \qquad (27)$$

where $k \geq 15$.

Another possibility is

$$X_{n+1} = (X_{n-24} + X_{n-55}) \mod m, \quad \text{for} \quad n \geq 55. \qquad (28)$$

The above two are called additive generators. One can also use a subtractive generator,

$$X_{n+1} = (X_{n-55} - X_{n-24}) \mod m. \qquad (29)$$

These RNG's are completely portable.

ran3 in NR is based on this algo.

Inversion method

Rejection Method

As we saw, random numbers with a given probability density function $p(y)$ can be generated from a uniform random variate x only when the indefinite integeral $P(y) = \int p(y)dy$ can be calculated, and the resulting function $P(y)$ can be inverted easily.

The last step may be particularly difficult in a general situation. Therefore, we look for other methods to generate random numbers with a given pdf.

Generate pairs of random numbers (X_1, X_2) in a unit square. Equate $X_1 = x$ and $X_2 = y$, which becomes coordinates of points on the 2D plane. Reject those points that lie above the curve $p(x)$, i.e., reject those points for which $y > p(x)$. Consider now the distribution of only the points remaining. The ratio of the number of points in two small (and equal) intervals at $x = a$ and $x = b$ will be $p(x = a)/p(x = b)$. In other words, the distribution of x values is $p(x)dx$.

A simple example will illustrate the point. Consider the semicircular distribution,

$$p(x) = \sqrt{1 - x^2} \text{ for } |x| < 1. \qquad (30)$$

For $x > 0$, the distribution covers only a quarter of a circle. We can now generate pairs of random numbers (X_1, X_2) between $[0, 1]$, and reject those for which $X_1^2 + X_2^2 \geq 1$.

Metropolis sampling

THE JOURNAL OF CHEMICAL PHYSICS VOLUME 21, NUMBER 6 JUNE, 1953

Equation of State Calculations by Fast Computing Machines

NICHOLAS METROPOLIS, ARIANNA W. ROSENBLUTH, MARSHALL N. ROSENBLUTH, AND AUGUSTA H. TELLER,
Los Alamos Scientific Laboratory, Los Alamos, New Mexico

AND

EDWARD TELLER,* *Department of Physics, University of Chicago, Chicago, Illinois*

(Received March 6, 1953)

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.

I. INTRODUCTION

THE purpose of this paper is to describe a general method, suitable for fast electronic computing machines, of calculating the properties of any substance which may be considered as composed of interacting individual molecules. Classical statistics is assumed, only two-body forces are considered, and the potential field of a molecule is assumed spherically symmetric. These are the usual assumptions made in theories of liquids. Subject to the above assumptions, the method is not restricted to any range of temperature or density. This paper will also present results of a preliminary two-dimensional calculation for the rigid-sphere system. Work on the two-dimensional case with a Lennard-Jones potential is in progress and will be reported in a later paper. Also, the problem in three dimensions is being investigated.

* Now at the Radiation Laboratory of the University of California, Livermore, California.

II. THE GENERAL METHOD FOR AN ARBITRARY POTENTIAL BETWEEN THE PARTICLES

In order to reduce the problem to a feasible size for numerical work, we can, of course, consider only a finite number of particles. This number N may be as high as several hundred. Our system consists of a square† containing N particles. In order to minimize the surface effects we suppose the complete substance to be periodic, consisting of many such squares, each square containing N particles in the same configuration. Thus we define d_{AB}, the minimum distance between particles A and B, as the shortest distance between A and any of the particles B, of which there is one in each of the squares which comprise the complete substance. If we have a potential which falls off rapidly with distance, there will be at most one of the distances AB which can make a substantial contribution; hence we need consider only the minimum distance d_{AB}.

† We will use the two-dimensional nomenclature here since it is easier to visualize. The extension to three dimensions is obvious.

One of the most powerful tools within MC methods is the *random walk*.

We will only use the idea of a (random) *walker* to develop the methods that we need. In fact, we will use random walk to generate a set of pseudorandom numbers distributed according to a desired probability density function.

This method is widely known in the literature as the *Metropolis algorithm or samplng*.

We define a mathematical entity called ‘walker’ whose attributes completely define the state of a system. Attributes of a walker can be completely general.

Two relevant examples from physics can be:
1. state of all spins (up or down) in an Ising model
2. The coordinates of all the particles in an interacting many-particle system.

The walker moves in an appropriate state space by a combination of deterministic and stochastic steps in general. This sequence of steps forms a *chain*.

If the system is in a state S_j at time j, then the sequence S_j from the beginning to the end of the random walk form a chain.

Markov chain: This sequence of states of the system is a *Markov chain* if the transition probabilities between any two states are independent of time and history.
That is they depend only on the current state of the system (and the one to which it attempts to make a transition), and not on when or how the system got there.

Let us define

$$P_{kj} = \left(S_k \leftarrow S_j\right) \tag{31}$$

as the probability of the system changing from state S_j at time j to the state S_k at time $j+1$. If the process is a Markov process then this probability depends on the state S_j, but on the time index j, or what other states the walker visited at earlier times.

Probability has to be normalized. the total transition probability from any state at a given time to all other states at the next point of time must be 1. Thus we have the normalization condition

$$\sum_{k=1}^{N} P_{kj} = 1. \tag{32}$$

Let the probability that the system is in a state S_k at a time i be $p_k^{(i)}$. The probability-space density may then be represented by a column vector

$$\mathbf{p} = \begin{bmatrix} p_1^{(i)} \\ \vdots \\ p_N^{(i)} \end{bmatrix}. \tag{33}$$

These probabilities also have to be normalized which gives

$$\sum_{k=1}^{N} p_k^{(i)} = 1. \tag{34}$$

At each point in time, the system may move from state S_j to S_k with a probability P_{kj}. The probability distribution will then evolve as

$$p_k^{(i+1)} = \sum_j P_{kj} p_j^{(i)}. \tag{35}$$

This can be written in a matrix form $\mathbf{p}^{(i+1)} = \mathbf{P}\mathbf{p}^{(i)}$. Using this rule, the system evolves the initial distribution as follows, $\mathbf{p}^{(1)} = \mathbf{P}\mathbf{p}^{(0)}$, $\mathbf{p}^{(2)} = \mathbf{P}\mathbf{p}^{(1)}$ *etc.* After m steps

$$\mathbf{p}^{(m)} = \mathbf{P}^m \mathbf{p}^{(0)}. \tag{36}$$

Suppose, after a sufficiently large number of steps M, $|p^{(M+1)} - p^{(M)}| \to 0$.

This implies the existence of an equilibrium probability distribution $\mathbf{p}^*$ such that

$$\mathbf{p}^* = \mathbf{P}\mathbf{p}^*. \tag{37}$$

The equilibrium distribution is a stationary state or fixed point distribution of the transition matrix $\mathbf{P}$.

Example Consider a Markov process with

$$\mathbf{P} = \begin{pmatrix} 1/4 & 1/8 & 2/3 \\ 3/4 & 5/8 & 0 \\ 0 & 1/4 & 1/3 \end{pmatrix} \tag{38}$$

The steady state fixed point can be obtained by solving the set of equations

$$p_1^* = 1/4p_1^* + 1/8p_2^* + 2/3p_3^* \tag{39}$$
$$p_2^* = 3/4p_1^* + 5/8p_2^* \tag{40}$$
$$p_3^* = 1/4p_2^* + 1/3p_3^* \tag{41}$$

And the answer is, $p_1^* = 4/15$, $p_2^* = 8/15$ and $p_3^* = 3/15$.

If a walker performs a random walk starting from $p_1^{(0)} = 1$, $p_2^{(0)} = p_3^{(0)} = 0$.

Iteration	p_1	p_2	p_3
0	1.00000	0.00000	0.00000
5	0.27213	0.53021	0.19766
10	0.26671	0.53332	0.19998
11	0.26666	0.53335	0.19999
12	0.26666	0.53334	0.20000
13	0.26667	0.53333	0.20000
$\mathbf{p}^*$	0.26667	0.53333	0.20000

Generalization to continuous variables (configuration space of N-particles). Here both space and time are taken as continuous variables.

The probability of moving a particle from $\mathbf{x}$ at time t to a point $\mathbf{y}$ at time $t + \Delta t$ is denoted by $G(\mathbf{y}, \mathbf{x}, \Delta t)$ – the continuous analog of the matrix P_{kj}.

Let the probability density of a particle at $\mathbf{x}$ at time t be $f(\mathbf{x}, t)$. Then we have

$$f(\mathbf{y}, t + \Delta t) = \int f(\mathbf{x}, t) G(\mathbf{y}, \mathbf{x}; \Delta t) d\mathbf{x}, \quad (42)$$

and

$$f(\mathbf{y}, t + m\Delta t) = \int f(\mathbf{x}, t) G(\mathbf{y}, \mathbf{x}; m\Delta t) d\mathbf{x}. \quad (43)$$

Again, there exists an equilibrium distribution that is independent of time

$$f^*(\mathbf{y}) = \int f^*(\mathbf{x}) G(\mathbf{y}, \mathbf{x}; t) d\mathbf{x}. \quad (44)$$

Random walk in state space So far we focused on the evolution of the probability density function along a Markov chain. Let us now focus on the evolution of the distribution in the *state space* S_k.

A single walker will be in a single state at a given time. *So what does it mean for a Markov chain to converge to an equilibrium density?*

For a single walker, equilibrium refers to the probability density with which the states are sampled in time. Therefore, during the walk, the states are sampled with probability $\mathbf{p}^*$. Apart from small fluctuations, all averages will be independent of time.

One of the conditions necessary for the random walk to reach equilibrium is *ergodicity*.

A process is ergodic if the spatial averages in the limit of infinite system is equal to temporal averages just discussed.

For a process to be ergodic it is necessary (though not sufficient) that all states have a non-zero probability of being visited.

To illustrate the point better, assume that a single walker visits points $\mathbf{X}^0$, $\mathbf{X}^1$, ..., $\mathbf{X}^m$ during the walk. The time average of the function $f(\mathbf{X})$ during this walk is given by

$$\langle f \rangle_t \equiv \frac{1}{m} \sum_{i=1}^{m} f(\mathbf{X}^i). \tag{45}$$

Once equilibrium has been achieved, $\langle f \rangle_t$ is independent of the starting point and time.

Now rather than following a single walker, consider an *ensemble of walkers* $\{\mathbf{X}\} = \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$, each performing independent random walks. The ensemble of spatial average is

$$\langle f \rangle_{\mathbf{X}} \equiv \frac{1}{N} \sum_{k=1}^{N} f(\mathbf{X}_k). \tag{46}$$

If the ensemble is drawn from the equilibrium distribution, then the two averages in eqs. 45 and 46 are equivalent. In such cases we may average over time and space in any combination to obtain,

$$\langle f \rangle_p = \frac{1}{Nm} \sum_{i=1}^{m} \sum_{k=1}^{N} f(\mathbf{X}_k^{(i)}). \qquad (47)$$

Taking either $m \to \infty$ or $N \to \infty$ then makes the average exact.

In our discussions so far the stationary distribution $\mathbf{p}^*$ is a consequence of $\mathbf{P}$.

If we wish to generate a distribution, we need to invert the above procedure to find a $\mathbf{P}$ for the desired $\mathbf{p}^*$.

This is accomplished by the *Metropolis method.*

Consider a discrete N-state system with equilibrium probabilities $\mathbf{p}^*$. Assume that S_i is the state of maximum probability, *i.e.*, $p_i^* = \max(\mathbf{p}^*)$.

We now use an acceptance/rejection step to arrive at the equilibrium probability distribution.

Acceptance probabilities (elements of $\mathbf{P}_{ki}$) are chosen such that the probability of moving to any state S_k from the state S_i of maximum probability is $A_{ki} = p_k^*/p_i^*$.

This acceptance ratio ensures that the relative probability of the states S_i and S_k are consistent with the probability density function.

Similarly, for the second most probable state S_j, the acceptance probability of moving to any other state S_k ($k \neq i$) is taken to be $A_{kj} = p_k^*/p_j^*$.

From these, $A_{kj}A_{ji} = A_{ki}$, so that the probability of moving from i to k is independent of the path.

The above argument can be continued till all the elements A_{ji} with $p_i^* \geq p_j^*$ are constructed. The remaining matrix elements of $\mathbf{A}$ correspond to moves to states with higher probability.

The correct probabilities are achieved by setting all these elements to 1 (including the diagonal elements A_{ii}).

Thus if we order the states in ascending probability ($p_1^* \geq p_2^* \cdots \geq p_N^*$), then $\mathbf{A}$ will have $A_{ji} = p_j^*/p_i^*$ as its lower triangle and 1 elsewhere.

How does this lead to equilibrium distribution? Let ν_i and ν_j be present populations in two states S_i and S_j in a large ensemble. Let $p_i^* > P_j^*$.

All ν_j walkers can move to S_i since $A_{ij} = 1$. Fractions of ν_i that can move to S_j is $\nu_i \frac{p_j^*}{p_i^*} = \nu_i A_{ji}$.

Net change of population in state S_j is

$$\delta\nu_j = \nu_i \frac{p_j^*}{p_i^*} - \nu_j \tag{48}$$

When $\nu_i/\nu_j > p_i^*/p_j^*$, $\delta\nu_j > 0$, and the population in S_j increases driving it towards equilibrium.

When $\nu_i/\nu_j < p_i^*/p_j^*$, $\delta\nu_j < 0$.

When $\nu_i/\nu_j = p_i^*/p_j^*$, $\delta\nu_j = 0$.

So the general form of acceptance probability is

$$A_{ji} = min\left(\frac{p_j^*}{p_i^*}, 1\right) \tag{49}$$

In Ising spin system: the acceptance probability is $e^{-\beta\Delta E}$.

Attempt a move $\rightarrow$ calculate $\Delta E \rightarrow$ calculate $e^{-\beta\Delta E}$.

Generate a uniform RN r.

If $r < e^{-\beta\Delta E}$ accept, else reject and retain the existing configuration.

Consider the two situations for three states S_i, S_j and S_k we discussed in the class. When $E_i < E_j < E_k$,

$$A_{ki} = \frac{p_k^*}{p_j^*} \times \frac{p_j^*}{p_i^*} = A_{kj} A_{ji}.$$

So the process is Markovian.

What happens to A_{ki} when $E_i < E_k < E_j$?

$A_{ji} = \frac{p_j^*}{p_i^*}$, $A_{kj} = 1$, $A_{ki} = \frac{p_k^*}{p_j^*}$.

So

$$A_{ki} \neq A_{kj} A_{ji}. \tag{50}$$

The process is apparently non-Markovian.

Suppose we have a condition

$$A_{kj}p_j^* = A_{jk}p_k^*. \tag{51}$$

This will turn eq 50 into

$$A_{ki} = \frac{p_k^*}{p_j^*} \times \frac{p_j^*}{p_i^*} = \frac{A_{kj}}{A_{jk}} A_{ji}.$$

But $A_{jk} = 1$, so

$$A_{ki} = A_{kj}A_{ji} \tag{52}$$

Eq 51 is called detail balance. The way we have constructed the transition matrix $\mathbf{A}$ for Metropolis, detail balance is satisfied (convince yourself if it is not already clear to you).

How did we get the detail balance condition?

Rate of change of probability of occupation of a certain state S_i is given by the rate equation

$$\frac{dp_i}{dt} = \sum_j \Big(A_{ij} p_j - A_{ji} p_i \Big) \tag{53}$$

This has to be equal to zero at equilibrium. Thus $\sum_j \Big(A_{ij} p_j - A_{ji} p_i \Big) = 0$.

As further constraint, we can demand that terms for each j be zero

$$A_{ij} p_j^* = A_{ji} p_i^*.$$

We get the detail balance condition.

Imposition of detail balance also ensures that we reach desired equilibrium, and not a *limit cycle* of the transition matrix $\mathbf{A}$.

I cheated a bit in the Ising simulation problem.

What I did is a little more complicated, and involves *generalized Metropolis*.

In problem 2 of the assignment, each walker attempts to move to all the three states with equal probability.

In the Ising problem, we flipped only one spin at a time. So given the system is in a state S_i, *not all other states are accessible*.

This idea can be formalized by stating that the **probability of attempting a move from state S_i to S_j is T_{ji}**. Then the proposed moves are accepted with probabilities B_{ji}.

In analogy with simple Metropolis, detailed balance will be satisfied if we take

$$\frac{B_{ji}}{B_{ij}} = \frac{p_j^*}{p_i^*}\frac{T_{ij}}{T_{ji}} \tag{54}$$

And this leads to

$$B_{ji} = min\left(\frac{p_j^*}{p_i^*}\frac{T_{ij}}{T_{ji}}, 1\right) \tag{55}$$

What are the thse probabilities T_{ij}s in the way we did MC for the Ising model?

State $S_i \rightarrow$ a definite configuration of spins.
State $S_j \rightarrow$ a new configuration of spins.
Can be generated by flipping any 1, any 2, any 3 *etc.* spins.

We considered *single-spin-flip* dynamics.
Notice that single-spin-flip dynamics is distinct from Metropolis algorithm. It is the specific acceptance probabilities which constitutes Metropolis algorithm.

In *single-spin-flip* if S_i and S_j differ by more than one spin flips $T_{ji} = 0$.
If there are n states S_j that differ by a single spin flip from S_i, $T_{ji} = \frac{1}{n}$ for each one of them.