Notes for 564 - Monte Carlo Methods

John Bulava

December 10, 2014

1 Introduction

These notes concern Monte Carlo integration methods, which approximate integrals stochastically. Perhaps the simplest illustration of the Monte Carlo method calculates the area of an irregular shape. Consider such a shape in two dimensions, which is contained in a square of side L. Now choose points uniformly and at random in the square and count the number of such points that lie within the irregular shape. Clearly,

$$\frac{A_{shape}}{L \times L} = \lim_{N_{point} \to \infty} \frac{N_{shape}}{N_{point}}, \tag{1}$$

where A_{shape} is the area of the shape, N_{shape} the number of points which fall in the shape, and N_{point} the total number of points. This method can be specialized calculate the area of a circle of particular radius, providing a determination of π.

Another illustrative historical example is **Buffon's Needle** (18th century), which may be the first Monte Carlo calculation ever performed. Conisder a floor composed of boards which form a pattern a parallel lines a distance t apart. Now repeated drop a needle of length ℓ uniformly on that floor. What is the probability that the needle crossed a line?

Call x the distance of the center of the needle to a line, and θ the angle that the needle makes with the horizontal axis. Clearly, the needle will cross a line if

$$x \leq \frac{\ell}{2} \cos\theta. \tag{2}$$

The probability P of this occurance is

$$\begin{aligned} P &= \frac{\int_0^{\frac{\pi}{2}} \int_0^{\frac{\ell}{2}\cos\theta} dx\, d\theta}{\int_0^{\frac{\pi}{2}} dx \int_0^{\frac{\pi}{2}} d\theta} \\ &= \frac{4}{\pi t} \int_0^{\frac{\pi}{2}} \frac{\ell}{2} \cos\theta\, d\theta = \frac{2\ell}{\pi t} \end{aligned} \tag{3}$$

therefore π can be obtained via

$$\pi = \frac{2\ell}{Pt} \tag{4}$$

2 Probability and Statistics Review

We call Ω the **sample space**, i.e. the set of all possible outcomes. The set of all **events**, or subsets of the sample space, is denoted $\mathcal{F}$. Finally we define a function $P : \mathcal{F} \to [0, 1]$, called the **probability**, such that

- General Properties:

$$\forall E \in \mathcal{F},\ P(E) \in \mathbb{R},\ P(E) \geq 0$$

- Unitarity:

$$P(\Omega) = 1 \tag{5}$$

- σ-additivity:

$$P(A \cup B) = P(A) + P(B),\ if\ A \cap B = \emptyset \tag{6}$$

These are known as the **Kolmogorov axioms** and can be used to derive many sensible properties about probabilites. For example, if the system in question is a (fair) 6-sided dice, the sample space consists of the sets $\{1\}, \{2\}, \ldots, \{6\}$ as well as e.g. the set $\{1, 3, 5\}$ which corresponds to the event that the dice roll is odd.

The **conditional** probability of A given B is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{7}$$

which leads to **Bayes' theorem**

$$P(A|B) = \frac{P(B|A)}{P(B)} P(A). \tag{8}$$

An important example involving conditional probabilities and Bayes' theorem is the calculation of probabilities of false positives in medical tests. Assume the test for a particular disease will give a positive result for 99.9% of patients who actually have the disease, but will give a 'false positive' for 1% of people with out the disease. If we denote the 'event' of a positive test result by B and the 'event' of the actual occurence of the disease by A, we have the conditional probabilities

$$P(B|A) = 0.999, \quad P(B|\bar{A}) = 0.01 \tag{9}$$

Now we consider the important situation when a positive test result is announced and doctor says 'the test is 99.9% accurate' (or even 99% sounds ominous enough). Furthermore, suppose the actual indicence of the disease in a relevant population is $P(A) = 0.0001$, or one in 10,000. We wish to compare the two probabilities $P(A|B)$, the probability of infection

given a positive test result, and $P(\bar{A}|B)$, the probability that the disease is not present given a positive test result. We have, applying Bayes' theorem,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{10}$$
$$P(\bar{A}|B) = \frac{P(B|\bar{A})P(\bar{A})}{P(B)}$$

so we must first calculate $P(B)$. Note that

$$P(B) = P(B \cap A) + P(B \cap \bar{A}) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A}) = .0100989 \tag{11}$$

so that $P(A|B) \approx 0.00989$ and $P(\bar{A}|B) \approx 0.99011$. Clearly, the false positive is the more likely outcome!

Next we discuss **random variables**, which are functions $X : \Omega \to E$, where E is some set. If the set E is countable (indexed by integers), then we call X a **discrete random variable**, otherwise it is a **continuous random variable**. For discrete random variables, we define a probability function $p_X : E \to [0, 1]$ such that

$$\forall x \in E, \quad p_X(x) = \sum_{\{\omega \in \Omega \mid X(\omega) = x\}} P(\omega), \tag{12}$$

where the symbol '$\forall$' means 'for all'. The quantity $p_X(x)$ is the probability that random variable X assumes value x.

For continuous random variables, the function p_X is called a probability density function (p.d.f. for short). In this case, the probability that X assumes a value in a small region dx around x is $p_X(x)dx$. Furthermore, the probability that x is found in an interval $[a, b]$ is

$$P(X \in [a, b]) = \int_a^b p_X(x)dx. \tag{13}$$

Also of interest is the **cumulative distribution function** (c.d.f.) defined as

$$F_X(x) = \int_{-\infty}^{x} p_X(s)ds. \tag{14}$$

These definitions of p.d.f's can be trivially extended to functions of multiple variables. Particularly interesting properties of p.d.f's are **mean**

$$\mu = \int_{-\infty}^{\infty} s\, p_X(s)ds \tag{15}$$

and **variance**

$$\sigma^2 = \int_{-\infty}^{\infty} (s - \mu)^2 p_X(s)ds \tag{16}$$

Other important results in probability theory are **Chebyshev's Inequality**, the (strong and weak) **Law of Large Numbers**, and the **Central Limit Theorem**: If $X_1, \ldots, X_N$

independent and identically distributed random variable taken from a sensible distribution with finite mean and variance, then the distribution of $(X_1 + \cdots + X_N - N\mu)/(\sigma\sqrt{N})$ approaches a normal distribution with mean zero and unit variance for asymptotically large N.

We now proceed to statistics. Questions in probability are often posed like 'Given the underlying distribution, what is the probability of a particular event?'. Statistics is concerned with the inverse problem, namely 'Presented with the existing data, what are the desired properties of the underlying distribution?'. Methods which attempt to estimate properties of the distribution without assuming a particular form, are called 'non-parametric'. For example, a common non-parametric estimator of the distribution itself is a histogram. Also, the standard estimators of the mean and variance are non-parametric. Often in Monte Carlo calculations we are concerned with estimating functions of the moments of probability distributions, or the moments themselves. For this we require estimates of the uncertainty as well.

Estimates of the uncertainty ask the question 'What is the probability that some quantity x is contained in an interval around its estimate μ'?, i.e. α is desired where

$$\alpha = P(|x - \mu| < b). \tag{17}$$

Given a range b, the **significance** α is defined using the above expression. Typically, significance is measured with regard to the Gaussian distribution. I.e. $1 - \sigma$ significance (or 'one standard deviation') means that the probability that the result is inside the error range is $\sim 68\%$, or the standard deviate of the normal distribution.

3 Markov Chain Monte Carlo

We call a **discrete random process** a sequence of random variables $\{X_t, t \in \mathbb{N}\}$, where $\mathbb{N}$ is the set of 'natural numbers', $\{0, 1, 2, 3, 4, 5, ...\}$. If the $\{X_t\}$ are independent, properties of the limiting distribution are not all that interesting. As the $\{X_t\}$ are independent and identically distributed (i.i.d.), this distribution is determined by the Laws of Large Numbers and the Central Limit Theorem.

Much more interesting are random processes satisfying the **Markov Property**:

$$P(X_t = x|X_0 = x_0, X_1 = x_1, ..., X_{t-1} = x_{t-1}) = P(X_t = x|X_{t-1} = x_{t-1}) \tag{18}$$

This property means that the value of the random variable at the current state of the chain depends only on the value assumed at the immediately previous step. We shall further restrict ourselves to **Homogeneous Markov Chains** (H.M.C.'s), for which the transition probability is is independent of t, i.e.

$$\forall t, t', \quad P(X_t = x|X_{t-1} = y) = P(X_{t'} = x|X_{t'-1} = y). \tag{19}$$

Therefore, it is sensible to define the **Markov matrix**

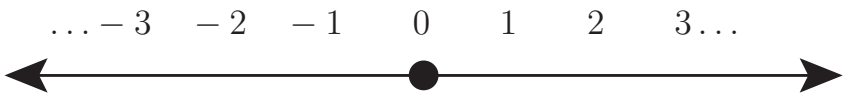
$$P(X_t = x|X_{t-1} = y) = p_{xy}. \tag{20}$$

A Markov chain is completely specified by its Markov matrix and all properties of the chain can be obtained from it.

Applications of H.M.C.'s fall into two broad categories:

- The use of H.M.C.'s as models of dynamical systems.
- The use of stationary H.M.C.'s to numerically solve large-dimensional sums and integrals. Stationary H.M.C.'s (rigorously defined later) are those which have a unique limiting long-time distribution.

A simple example of a dynamical system which can be modeled as an H.M.C. is the one-dimensional random walk. Here the sample space $\Omega = \mathbb{Z}$, which is the set of integers, which models an object (the black blob) moving on a line given some starting position, which we call zero.



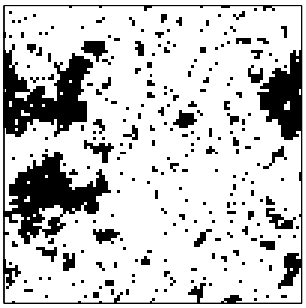
At each time the object moves either left or right with equal probability. We denote the probability that the object is at position i at time t by $(\pi_t)_i$. The probability distribution of the system at t is given by the vector $\boldsymbol{\pi}_t$. Based on the Markov Property, we have the matrix equation

$$\boldsymbol{\pi}_{t+1} = p\boldsymbol{\pi}_t, \tag{21}$$

where for this process

$$p_{ij} = \frac{1}{2}(\delta_{i,j-1} + \delta_{i,j+1}). \tag{22}$$

The use of stationary H.M.C.'s to solve sums and integrals is exemplified in the two-dimensional Ising model. Magnetic phenomena in metals is caused by the quantum-mechanical 'spin' of electrons, which here can take one of two values, conventionally called 'spin up' or 'spin down'. We approximate a two-dimensional metal by a lattice of such spins, and use black and white for spin up and down.



A possible configuration of the system is shown here for a 100-by-100 lattice of spins. Each site (indexed by i) contains a spin $\sigma_i \in \mathbb{Z}_2$, where $\mathbb{Z}_2$ is the set {-1,1}. The entire lattice is collectively denoted by σ.

The energy of a particular spin configuration is given by

$$E(\sigma) = -J\sum_{\langle i,j\rangle}\sigma_i\sigma_j - h\sum_j \sigma_j, \tag{23}$$

where J and h are free parameters and $\langle i, j \rangle$ is the set of all nearest nieghbor points. If the system is at a particular temperature T, statistical physics tells us that the probability of a given spin configuration is

$$P_\beta(\sigma) = \frac{\mathrm{e}^{-\beta E(\sigma)}}{Z_\beta}, \quad Z_\beta = \sum_{\{\sigma\}} \mathrm{e}^{-\beta E(\sigma)}, \quad \beta = \frac{1}{kT}, \tag{24}$$

where k is Boltzmann's constant and $\{\sigma\}$ denotes the set of all possible spin configurations.

One of the interesting quantities in the Ising model is the average magnetization

$$\langle m \rangle = \sum_\sigma m(\sigma)\ P_\beta(\sigma), \qquad m(\sigma) = \frac{1}{N_{spins}} \sum_i \sigma_i, \tag{25}$$

where N_{spins} is the total number of spins. Due to the Central Limit Theorem, such sums can evaluated by calculating $m(\sigma)$ on an ensemble of spin configurations which are distributed according to $P_\beta(\sigma)$ and averaging the result. As the number of configurations in the ensemble tends to infinity, we recover $\langle m \rangle$. However, unlike simple low-dimensional probability distributions, $P_\beta(\sigma)$ has a complicated dependence on many variables, in this case 10,000 spins! Therefore it is impossible generate spin configurations by performing a transformation on uniformly generated random numbers.

However, we shall see later that it is quite straightforward to construct a stationary H.M.C. such that

$$\lim_{t\to\infty} (\boldsymbol{\pi}_t)_\sigma = P_\beta(\sigma). \tag{26}$$

Then, the suitably distributed ensemble of spin configurations can be obtained by taking elements from the Markov Chain.

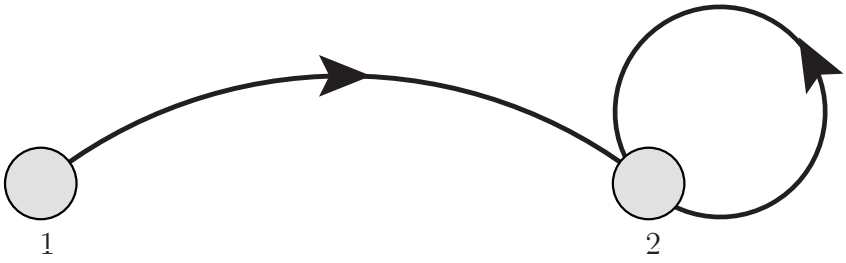
We recall that due to the defining Markov Property (Eq. 18), all elements of the Markov matrix are non-negative $p_{ij} \geq 0$ and

$$\sum_i p_{ij} = 1, \quad \forall j. \tag{27}$$

These properties make p a **stochastic matrix**. If in addition

$$\sum_j p_{ij} = 1, \quad \forall i, \tag{28}$$

p is called a **doubly-stochastic matrix**. Note that not all Markov matrices are doubly stochastic. Indeed, the simple example



is a Markov chain which is not doubly stochastic. The Markov matrix for this chain is

$$p = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} \tag{29}$$

for which the rows clearly do not sum to unity.

In addition to the Markov matrix we demand some reasonable constraints on the probability distributions $\boldsymbol{\pi}_t$ of the chain at time t. Namely,

$$\sum_i (\boldsymbol{\pi}_t)_i = 1, \quad \forall t \quad \text{and} \quad (\boldsymbol{\pi}_t)_i \geq 0, \quad \forall i, t \tag{30}$$

The Perron-Frobenius Theory of non-negative square matrices exhausts the analysis of their properties. We only examine a relatively simple property of stochastic matrices, namely that they all have at least one unit eigenvalue. This is interesting in the investigation of stationary states, as a **stationary state** is a probability distribution $\boldsymbol{\pi}$ such that

$$\boldsymbol{\pi} = p\boldsymbol{\pi}, \tag{31}$$

namely $\boldsymbol{\pi}$ is a right eigenvector of p with eigenvalue 1.

Since all $\det A = \det A^T$, any square matrix and its transpose have the same eigenvalues. Also, from the defining property of stochastic matrices (Eq. 27) we see that the vector $v_i = 1$ is a left eigenvector with unit eigenvalue:

$$\sum_i v_i p_{ij} = \sum_i p_{ij} = 1 = v_j. \tag{32}$$

The right eigenvector corresponding to this eigenvalue may not be unique however, and also may not be a valid probability distribution, i.e. it may have negative components.

Another quantity of interest is the **n-step transition probability** defined as the probability that a system in state i will be in state j after n steps and given by $(p^n)_{ji}$. The **n-step first visit probability**, denoted $f_{ji}^{(n)}$ is the probability that the system in state i visits state j *for the first time* after n steps. This is defined iteratively as

$$f_{ji}^{(n)} = 0, \quad f_{ji}^{(1)} = p_{ji}, \quad f_{ji}^{(2)} = \sum_{k \neq j} p_{jk} p_{ki}, \quad \ldots, \quad f_{ji}^{(n)} = \sum_{k \neq j} f_{jk}^{(n-1)} p_{ki}. \tag{33}$$

he **total visit probability**, denoted f_{ji}, is the probability that the chain will *ever* reach j starting from i. This is given by

$$f_{ji} = \sum_{n=1}^{\infty} f_{ji}^{(n)}, \tag{34}$$

while the **mean first passage time** is the expected number of steps it will take to reach j from i:

$$m_{ji} = \sum_{n=1}^{\infty} n f_{ji}^{(n)}. \tag{35}$$

As a special case, the **mean recurrence time** of a state i, denoted μ_i, is the mean first passage time from a state back to itself:

$$\mu_i = \sum_{n=1}^{\infty} n f_{ii}^{(n)}. \tag{36}$$

Recurrence plays an important role in the classification of states. Every state in a Markov Chain can be classified as

- **Positive Recurrent** (persistent): $f_{ii} = 1,\ \mu_i < \infty,\ \sum_n (p^n)_{ii} = \infty$
- **Null Recurrent**: $f_{ii} = 1,\ \mu_i = \infty,\ \sum_n (p^n)_{ii} = \infty$
- **Transient**: $f_{ii} < 1,\ \sum_n (p^n)_{ii} < \infty$

We say that state j is **accessible** from state i, denoted $i \to j$, if $\exists n,\ (p^n)_{ji} > 0$. If $i \to j$ and $j \to i$, then we say that states i and j **communicate**, and denote it $i \leftrightarrow j$. A set of states which all communicate with each other is called a **class**. Note that communication is transitive; if $i \leftrightarrow j$ and $j \leftrightarrow k$, then $i \leftrightarrow k$. Therefore, two classes C_1 and C_2 must be disjoint, i.e. $C_1 \cap C_2 = \emptyset$ or else they would be a single class. Our 1-D random walk contains a single communcating class, while the two state chain discussed previously in this section contains two classes, each containing a single state. Markov chains which contain a single communicating class are called **irreducible**.

We define the **period** of a state i, denoted d_i, as

$$d_i = \gcd\{n \geq 1 \mid (p^n)_{ii} > 0\}, \tag{37}$$

where 'gcd' means greatest common divisor. That is to say the period of a state is the greatest common divisor of all n such that the n-step return probability is non-zero. If a state i has $d_i > 1$ it is **periodic** (or cyclic), while if $d_i = 1$ it is called **aperiodic**. A Markov Chain is called aperiodic if all its states are aperiodic.

We now generalize the previous example to a random walk with transition matrix

$$p_{ij} = r\delta_{i,j-1} + (1-r)\delta_{i,j+1}, \quad 0 \leq r \leq 1. \tag{38}$$

In this example we examine the n-step return probabilities of the state $i = 0$, $(p^n)_{00}$. Clearly, an even number of steps are required to return, so

$$(p^{2n+1})_{00} = 0, \quad \forall n \geq 0. \tag{39}$$

For even n we may view the problem combinatorically. A trip that starts and ends at $i = 0$ in $2n$ steps must have an equal number of left and right moves, while the order in which they occur is irrelevant. Therefore, the number of such paths is

$$\binom{2n}{n} = \frac{(2n)!}{n!\ n!}\ . \tag{40}$$

The quantity $\binom{n}{k}$, for $n \geq k$ is called the binomial coefficient. Therfore the $2n$-step return probability is

$$(p^{2n})_{00} = \binom{2n}{n} r^n (1-r)^n. \tag{41}$$

We may put this in a more tractable form by employing Stirling's formula

$$\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi n}\left(\frac{n}{\mathrm{e}}\right)^n} = 1 \tag{42}$$

so that our $2n$-step return probability may be approximated as

$$(p^{2n})_{00} \approx \frac{[4r(1-r)]^n}{\sqrt{\pi n}}. \tag{43}$$

Now we know that if the series $\sum_n (p^{2n})_{00}$ converges, $i = 0$ is a transient state, while it is a recurrent state if the series diverges. The convergence of a series with a general term given by Eq. 43 depends on the magnitude of $4r(1-r)$. If $r \neq \frac{1}{2}$, $4r(1-r) < 1$ and $i = 0$ is a transient state. If $r = \frac{1}{2}$, $4r(1-r) = 1$ and the series diverges, so that $i = 0$ is a recurrent state. It can be demonstrated however (See Bremaud, Ch. 2, Ex. 7.6) that this is actually a null-recurrent state.

At long last, the **fundamental limit theorem for irreducible Markov chains** states that an irreducible, aperiodic Markov chain has a stationary distribution if and only if all its states are postive recurrent. Furthermore, this distribution is unique and given by

$$\pi_j = \lim_{n\to\infty} (p^n)_{ji}. \tag{44}$$

This distribution is also **universal**, as it is the limiting distribution independent of the initial state i. If a chain is irreducible, aperiodic, and positive recurrent, then we say it is **ergodic**. The above result says that an ergodic chain has a limiting distribution *independently* of the starting state.

The process of bringing a Markov chain from an arbitrary starting configuration to the limiting distribution is called **thermalization**. Once the probability distribution of a chain is equal (or asymptotically close) to the limiting stationary distribution, we say it is in **equlibrium**.

The idea of thermalization can be illustrated using a simple example. Consider a variant of the 1-D symmetric random walk which is constrained to move in a fixed range, here chosen to be $[-5, 5]$. The transition matrix for this chain is

$$p_{ij} = \begin{cases} \frac{1}{2}(\delta_{i,j+1} + \delta_{ij}), & j = -5 \\ \frac{1}{2}(\delta_{i,j-1} + \delta_{ij}), & j = 5 \\ \frac{1}{2}(\delta_{i,j+1} + \delta_{i,j-1}), & \text{otherwise.} \end{cases} \tag{45}$$

Note that although the unconstrained random walk is irreducible, periodic (period 2), and null-recurrent, because of the transition matrix defined here, this chain is ergodic. It is aperiodic because after 5 steps, there is a non-zero probability of staying at $x = -5, 5$ for

any number of steps. Therefore, the Fundamental Limit theorem can be applied, and the equilibrium distribution is the one where each site occurs with equal probability. This can be easily shown by noting that p_{ij} is a doubly-stochastic matrix.

This is implemented in the small C++ program `rw1.cc`. Here many independent chains of a fixed length are simulated, all started at $x = 0$. The resultant position of the chains after a fixed number of steps is the histogrammed, and the results shown in Fig. 1.

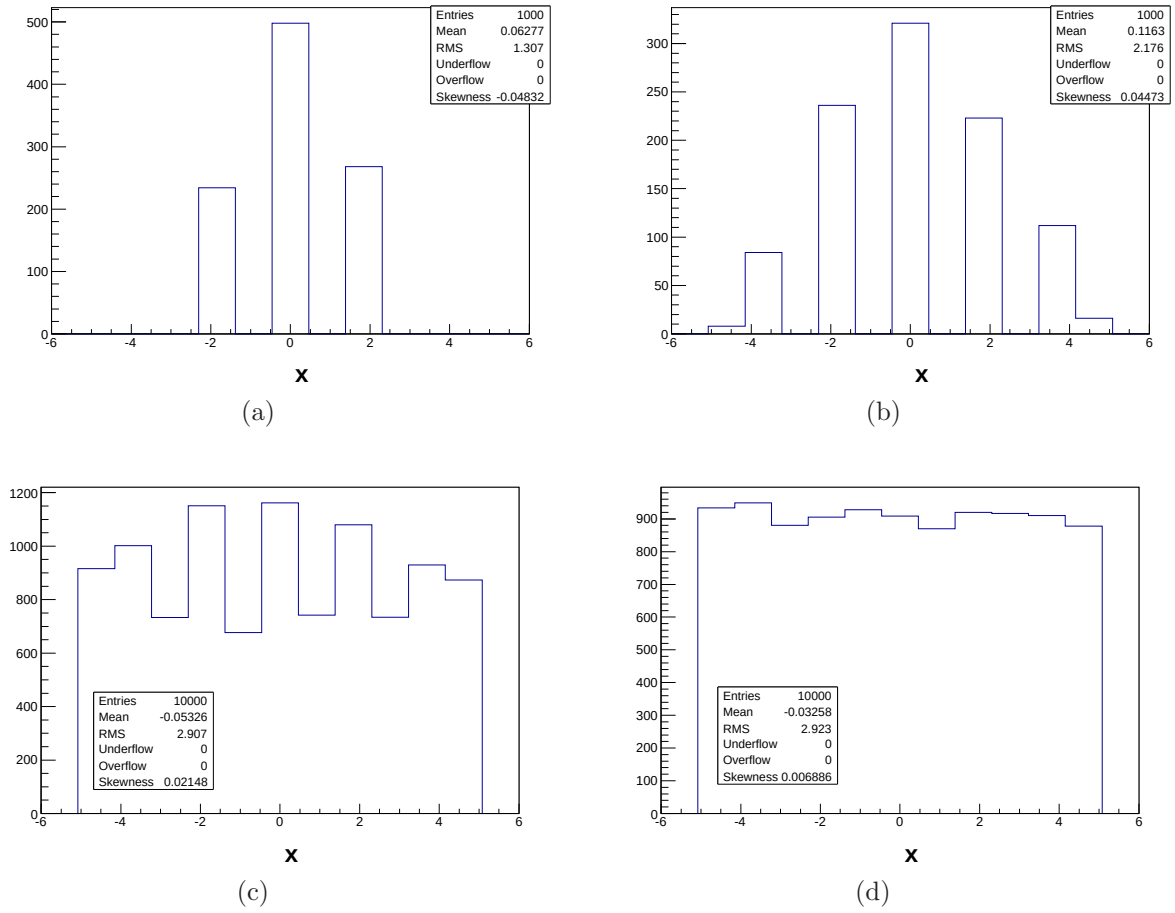


Figure 1: Thermalization for the constrained random walk described above. Many chains are run for a fixed number of steps, all starting at $x = 0$, and the resultant positions are histogrammed. Results are shown for (a) $n_{step} = 2$, (b) $n_{step} = 6$, (c) $n_{step} = 50$, and (d) $n_{step} = 10000$.

Given the Fundamental Limit Theorem we now know that an ergodic chain has a unique limiting distribution, independent of the starting state. However due to the Markov property, different elements of the chain are certianly not independent! Therefore, the standard Central Limit Theorem does not apply. In order to generalize it, we need to introduce several concepts. In what follows, we assume that all Markov chains have been thermalized.

The **autocovariance** of an (equilibrium) Markov chain $\{X_t\}$ with average value μ is

defined as

$$R(|t-s|) = E\left[(X_t-\mu)(X_s-\mu)\right], \tag{46}$$

where E is the standard expectation value. Note that $R(0) = \sigma^2$, where σ^2 is the variance. In practice it is convient to also work with the **autocorrelation** defined as

$$\rho(t) = \frac{R(t)}{R(0)}, \tag{47}$$

where $\rho(0) = 1$ and $-1 \leq \rho(t) \leq 1$. For ergodic chains, we can then recover the standard central limit theorem except with

$$\sigma^2 \to \sigma^2 + 2\sum_{t=1}^{\infty} R(t). \tag{48}$$

Recall that one of the conditions for the central limit theorem was a finite variance, $\sigma^2 < \infty$. In addition, we must add the condition that the autocorrelation function is absolutely summable,

$$\sum_{t=1}^{\infty} R(t) < \infty. \tag{49}$$

To summarize, to apply Markov Chain Monte Carlo to calculate an integral or weighted average, we have the following analogue of the Central Limit Theorem for ergodic chains:

$$\int D\vec{x}\, p(x)f(x) \approx \langle f\rangle + \sqrt{\frac{\sigma^2 + 2\sum_{t=1}^{\infty} R(t)}{N}}, \quad \langle f\rangle = \frac{1}{N}\sum_{i=1}^{N} f(x_i), \tag{50}$$

where the $\{x_i\}$ are drawn from a Markov chain in equilibrium with distribution $p(x)$. This formula becomes exact in the limit $N \to \infty$.

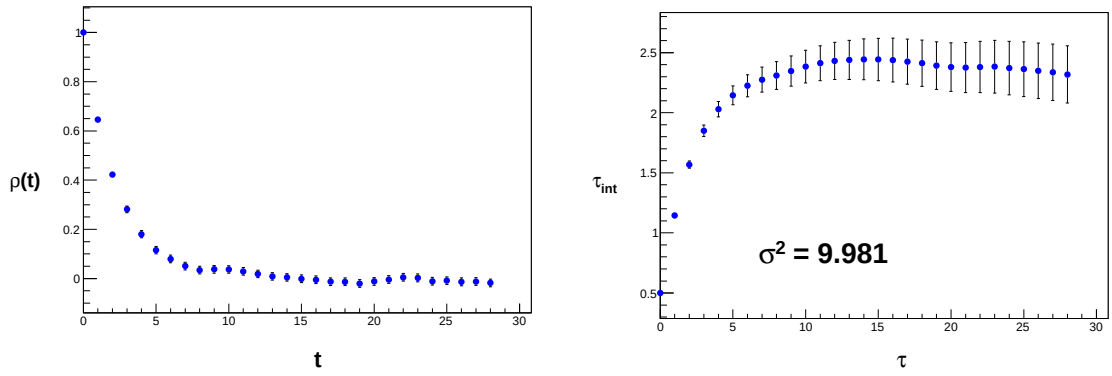
The ideas of autocorrelation and the modified Central Limit Theorem can be illustrated in the 1-D constrained symmetric random walk of the previous section. After thermalization, the chain is run for long time and the position is measured every 10 steps. This procedure is implemented in `rw2.cc`. The autocorrelation function of the position is shown in Fig. 2a. To get an estimate for the error, we must sum the autocorrelation function numerically.

In practice, when summing the autocorrlation function numerically it is useful to define the **integrated autocorrelation time**:

$$\tau_{\rm int}(\tau) = \frac{1}{2} + \sum_{t=0}^{\tau} \rho(t). \tag{51}$$

With this quantity the error that enters the modified central limit theorem can be expressed as

$$\tilde{\sigma}^2 = \sigma^2 + 2\sum_{t=1}^{\infty} R(t) = 2\tau_{\rm int}\sigma^2, \tag{52}$$



(a) (b)

Figure 2: The autocorrelation function (a) and integrated autocorrelation time (b) for the constrained random walk described above. After thermalization, the position of the chain is measured every 10 steps.

where $\tau_{\text{int}} = \lim_{\tau\to\infty} \tau_{\text{int}}(\tau)$. The quantity $\tau_{\text{int}}(\tau)$ is plotted in Fig. 2b. We see that $\tau_{\text{int}}/10 \sim 2.5$, as we are measuring every 10 steps for this example. Also shown in Fig. 2b is the variance $\sigma^2 \sim 10$ so that the modified variance which we can expect is $\tilde{\sigma}^2 \sim 500$. Note that this is substantially larger than σ^2!

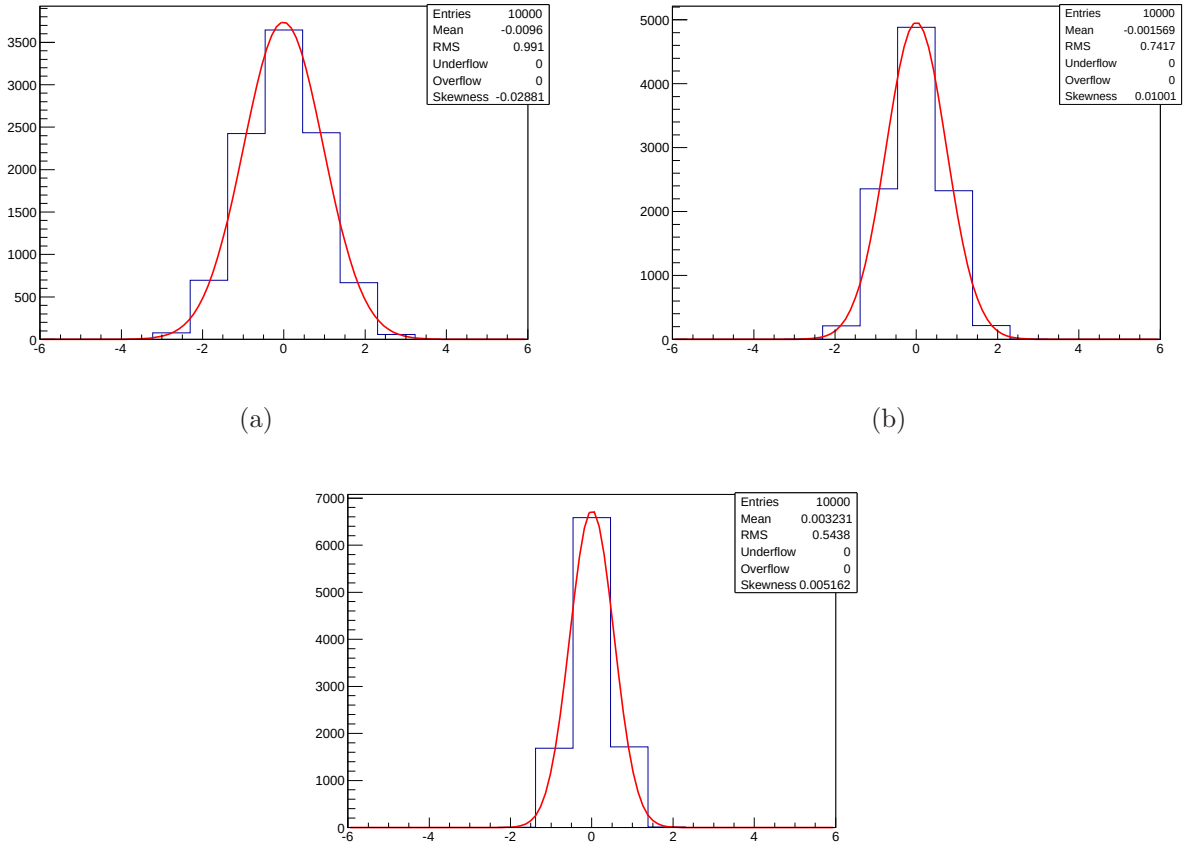
It should be noted that the errors shown in Fig. 2 come from the following approximations for the errors on the autocorrelation function and the integrated autocorrelation time.

$$(\delta\rho(t))^2 \approx \frac{1}{N}\sum_{m=1}^{\infty}\left(\rho(m+t)+\rho(m-t)-2\rho(m)\rho(t)\right)^2, \quad (\delta\tau_{int}(\tau))^2 \approx \frac{2(2\tau+1)}{N}\tau_{int}^2(\tau). \tag{53}$$

We can also numerically investigate the modified Central Limit Theorem, to see that indeed the modified variance is obtained. To this end we run many (thermalized) chains for a fixed number of steps, taking the average of the position at every step in each chain. This procedure is implemented in `rw3.cc`. The average value of the position for each chain is histogrammed in Fig. 3 for various values of the number of steps. We see that, as expected, the width of the histograms deceases roughly like $\tilde{\sigma}/\sqrt{n_{step}}$ and that the variance is about $\tilde{\sigma} \sim \sqrt{500} \sim 22$.

In theory, we now know that ergodic chains admit unique universal stationary distributions given by $\pi_j = \lim_{n\to\infty}(p^n)_{ji}$. However, for complicated chains, this limit is in practice difficult to calculate. Also, typically we know the distribution $\boldsymbol{\pi}$ and are interested in designing an ergodic chain which has it a limiting distribution. To this end we examine a few more desirable properies of chains which allow for the easy identification of the limiting distribution.

We first note that for an ergodic chain $\{X_n\}$ with transition matrix p^X and stationary distribution $\boldsymbol{\pi}$, the **time-reversed** process $\{Y_n\}$ also satisfies the Markov property. This



(a) (b)

(c)

Figure 3: Results for the average positions from many thermalized chains run for a for a fixed number of n_{step}, together with Gaussian fits. The width of the gaussian can be read from the 'RMS' entry in the info box. Shown are (a) $n_{step} = 500$, (b) $n_{step} = 1000$, and (c) $n_{step} = 2000$.

process is defined as

$$p_{ij}^{Y} = P(Y_{n+1} = i | Y_n = j) = \left(\frac{\pi_i}{\pi_j}\right) p_{ji}^{X}. \tag{54}$$

It is quite easy to show that $\{Y_n\}$ is a Markov Chain. Furthermore, we shall call an ergodic chain $\{X_n\}$ **reversible** if its transition matrix is equal to the transition matrix of the time-reversed chain. That is, if

$$p_{ij}^{Y} = p_{ij}^{X}, \quad \forall i, j. \tag{55}$$

Based on Eq. 54, this condition can also be expressed as

$$p_{ij}\pi_j = p_{ji}\pi_i. \tag{56}$$

This is a very important property of Markov Chains and is called **detailed balance**. Its utility lies in the following theorem. If a chain $\{X_n\}$ is ergodic, and there exists a distribution $\boldsymbol{\pi}$ satisfying the detailed balance condition, then $\boldsymbol{\pi}$ is the unique universal limiting distribution. This can be easily seen from the Fundamental Limit Theorem after it is noted that $\boldsymbol{\pi}$ is a right eigenvector of p

$$\sum_j p_{ij}\pi_j = \sum_j \pi_i p_{ji} = \pi_i \sum_j p_{ji} = \pi_i. \tag{57}$$

This represents an easier line of reasoning then using the Fundamental Limit Theorem alone. First, if the chain is ergodic and the limiting distribution is desired, simply solving the detailed balance equations for π_i is easier than taking the limit of products of the transition matrix. Secondly, if the distribution $\boldsymbol{\pi}$ is known and a chain is desired which has it as the limiting distribution, then it is often straightforward to construct an ergodic chain whose transition matrix satisfies the detailed balance condition.

We first detail a common strategy for constructing algorithms which satisfy Detailed Balance and converge to a desired probability distribution. This strategy is known as the **Hastings algorithm**. It is composed of two steps. First, a proposed change to a state j is made to the system (currently in state i) from a **proposal matrix** h_{ji}, where

$$P(Y = j \mid X_n = i) = h_{ji}, \tag{58}$$

and Y is the proposed change to the chain. Obviously we have that $h_{ji} \geq 0$, and that h is a stochastic matrix.

The second step of the Hastings algorithm is an acceptance step, which is given by the **acceptance matrix** a_{ji}. The matrix element a_{ji} represents the probability that the proposed change is accepted. Clearly $0 \leq a_{ji} \leq 1$. The Markov matrix for the Hastings algorithm is thus

$$p_{ji} = \begin{cases} a_{ji}h_{ji}, & i \neq j \\ 1 - \sum_{k \neq i} a_{jk}h_{ki}, & i = j \end{cases}. \tag{59}$$

The $i = j$ case can be obtained by noting that there are two possibilities for a system to stay in its current configuration (i.e. $i = j$). One is that the proposal step 'proposes' no move. If this is the case, the acceptance step is trivial, as acceptance and non-acceptance mean the same thing. Therefore, the probability of this happening is h_{ii}. The other way that no transition can occur is for the proposal step to propose some transition to a state $k \neq i$ and for that proposal to be rejected. The probability of this happening is $\sum_{k \neq i} (1 - a_{jk}) h_{ki}$. The addition of these two probabilites, plus the stochastic property of h gives the result above.

We can easily show that detailed balance is satisfied if

$$a_{ji} = \min\left(1,\ \frac{\pi_j h_{ij}}{\pi_i h_{ji}}\right) \tag{60}$$

To demonstrate this, we consider both the cases $i = j$ and $i \neq j$. The $i = j$ case is trivial, as the detailed balance relation is trivially satisfied if $i = j$. The $i \neq j$ proceeds by constructing the product $a_{ji} h_{ji}$. Since $h_{ji}, \pi_i, \pi_j \geq 0$, we can express the detailed balance condition as

$$\min\left(\pi_i h_{ji},\ \pi_j h_{ij}\right) = \min\left(\pi_j h_{ij},\ \pi_i h_{ji}\right). \tag{61}$$

As 'min' is a symmetric function of its two arguments, this expression holds and detailed balance is satisfied.

Although the Hastings algorithm is more general, for the remainder of this section we consider statistical systems composed of many 'sites', each site with some sample space S. Therefore, the sample space of the total system is $\Omega = S^N$ for a system with N sites. In this section we consider algorithms which update a single degree of freedom at a time. For simplicity, assume we proceed through the sites of the system in a regular manner.

3.1 The Gibbs Sampler

Here we change a single site v. Consider i the starting configuration with the old value v_0 and j the configuration with a new value for v, denoted v_1. We take as our proposal matrix

$$h_{ij} = P(v = v_1 \mid w = w_0, \forall w \neq v) = \frac{\pi_j}{s}, \tag{62}$$

where $s = \sum_{k \in \Theta_v} \pi_k$ and Θ_v is the set of all configurations with $\{w = w_0,\ \forall w \neq v\}$. This sum runs over all configurations of v in the fixed backgroud of the other sites.

We now proceed to calculate the acceptance matrix of Eq. 60. We see that

$$\frac{\pi_j \frac{\pi_i}{s}}{\pi_i \frac{\pi_j}{s}} = 1, \tag{63}$$

so that the proposed update is always accepted! Clearly this algorithm will only be feasible if the conditional probability in Eq. 62 can be easily sampled.

3.2 The Metropolis Algorithm

Whereas the Gibbs sampler has a (somewhat) complicated proposal step and a trivial acceptance step, the Metropolis algorithm has a (relatively) simple proposal step which is then

forced to follow the desired distribution by the acceptance step. Many variants are possible, but we will consider the proposal matrix

$$h_{ij} = \begin{cases} \frac{1}{\Delta}, & v_1 \in \left[v_0 - \frac{\Delta}{2}, v_0 + \frac{\Delta}{2}\right] \\ 0, & \text{otherwise.} \end{cases} \tag{64}$$

A suitable choice of Δ is required for an efficient algorithm. A Δ which is too large will result in a low acceptance rate, while a Δ which is too small will only slowly decorrelate successive elements in the chain. Generally a rule of thumb targets an acceptance rate of 60-80%. However, this can be tested by numerical experiment. Simply start with a large Δ (low acceptance) and slowly decrease Δ while measuring the autocorrelation time. At some point the autocorrelation time will stop decreasing and start increasing due to increasingly tiny moves. An optimal acceptance is somewhere near this minimum.

3.3 Metropolis vs. Gibbs Sampler

The effectiveness of the Metropolis algorithm against the Gibbs sampler will most likely vary depending on the application. Still we illustrate some general features with an example. We consider a two dimensional problem with joint p.d.f

$$P(x_1, x_2) \propto \exp\left[-\frac{a}{2}(x_1^2 + b)(x_2^2 + b)\right]. \tag{65}$$

Such a p.d.f is normalizable, while we consider the case $a = 100, b = 0.01$.

To implement the Gibbs sampler, we notice that for x_2 fixed, x_1 is drawn from a p.d.f

$$P(x_1 \mid x_2) \propto \exp\left[-\frac{x_1^2}{2\sigma^2}\right], \quad \sigma = \frac{1}{\sqrt{a(x_2^2 + b)}}, \tag{66}$$

which is a standard gaussian with an x_2-dependent width. The procedure is then repeated to update x_2. A single 'sweep' of the algorithm is defined as an update of both x_1 and x_2. This algorithm, with a large number of thermalization and measurement sweeps is implemented in `gibbs1.cc`.

For the Metropolis algorithm, we uniformly propose a change for x_1 in the interval $x_1' \in [x_1 - \frac{\Delta}{2}, x_1 + \frac{\Delta}{2}]$. This change is accepted with probability

$$p_{acc} = \min\left(1, \exp\left[-\frac{a}{2}(x_2^2 + b)(x_1'^2 - x_1^2)\right]\right) \tag{67}$$

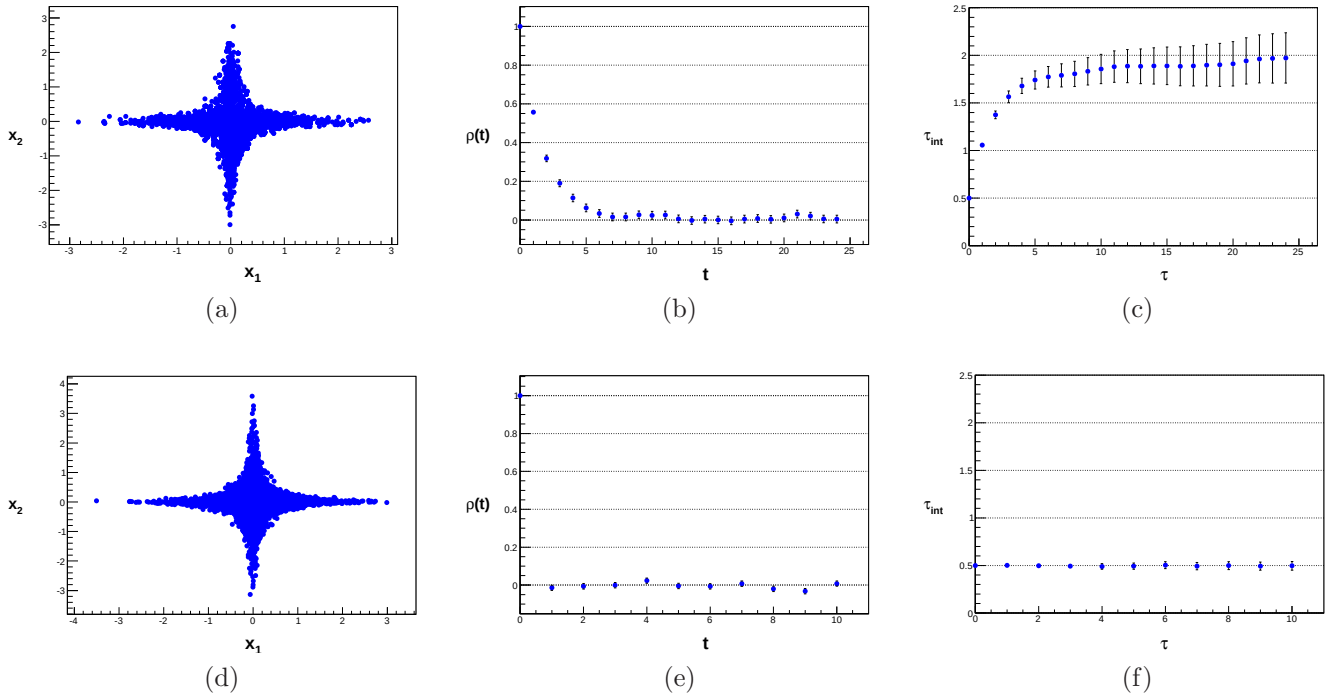
with an analogous update for x_2 together forming a single sweep. This algorithm, with a large number of thermalization and measurement sweeps is implemented in `metro1.cc`.

Of course, the parameter Δ should be tuned. The results of such a tuning are shown in Tab. 1. Once the Metroplis algorithm was sufficiently tuned, it was compared with the Gibbs sampler, the results of which are shown in Fig. 4.

While this example may be an unusual case, it seems that the Gibbs sampler has nearly no autocorrelation, while the Metropolis algorithm has a significant autocorrelation function and integrated autocorrelation time even at the optimally tuned value of Δ.

Δ	$p_{acc}(\%)$	τ_{int}
50.0	3.0	15.2(2.9)
15	9.9	4.83(93)
6.0	24.8	2.26(22)
2.0	52.0	5.72(74)
1.5	58.6	8.9(1.2)
1.0	68.0	19.9(4.6)

Table 1: Acceptance probability and integrated autocorrelation time for a variety of Δ values in the Metroplis algorithm. The situation described above was implemented with 1000 thermalization sweeps followed by 5000 measurement sweeps.



(a) (b) (c)

(d) (e) (f)

Figure 4: A comparison of the Metropolis algorithm and the Gibbs sampler for the p.d.f described above. In the top row a scatterplot of the elements obtained using the Metropolis algorithm with $\Delta = 6$ is shown (a) along with the autocorrelation function (b) and integrated autocorrelation time (c). The bottom row shows the scatterplot (d), autocorrelation function (e), and integrated autocorrelation time (f) for the Gibbs sampler.

4 The Ising Model

The local update algorithms discussed above can be illustrated in the Ising Model, first discussed in Sec. **??**. As discussed there, the two-dimensional variant of this model with $h = 0$ undergoes a second order phase transition from $\langle m \rangle = 0$ to $\langle m \rangle > 0$ at a finite temperature T_c. Onsager ('44) solved this model exactly to obtain the critical temperature

$$T_c = \frac{2J}{k_B \log(1+\sqrt{2})}, \quad \text{or} \quad \frac{k_B T_c}{J} = \frac{2}{\log(1+\sqrt{2})}. \tag{68}$$

One characteristic of second order phase transitions is a divergent **correlation length**. The correlation length is defined as ξ:

$$\lim_{|i-j|\to\infty} \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_i \rangle = C \exp\left[-\frac{|i-j|}{\xi} \right], \tag{69}$$

Where $|i-j|$ is the distance between spins i and j. As the temperature is decreased to the critical temperature, the correlation length diverges. In a finite system with length L on a side, at some point the correlation length saturates the size of the box and $\xi \sim L$.

The autocorrelation time is related to the correlation length

$$\tau \sim \xi^z, \tag{70}$$

where z is called the **dynamical critical exponent**. The value of z is very much dependent on the algorithm, but it can be shown that for the Ising model, algorithms that flip a single spin have $z \geq 1.75$.

5 Hybrid Monte Carlo

We now turn to an advanced topic in the Markov Chain Monte Carlo simulation of desired probability distributions. The examples we have looked at so far, such as the Ising Model, have had probability distributions which 'factorize', i.e. the contribution of a single site can be isolated and depends on its nearest neighbors. For such distributions which do not have this property, we must adopt a different approach. To this end, if our system is described by N continuous degrees of freedom $\phi_i \in \mathbb{R}$, $i = 1..N$, we first introduce N additional variables $\pi_i \in \mathbb{R}$.

If the target probability distribution of our system is $P(\phi) \propto \mathrm{e}^{-\beta E(\phi)}$, we now have the joint probability distribution

$$P(\phi, \pi) \propto \mathrm{e}^{-E(\phi)-T(\pi)}, \qquad T(\pi) = \sum_i \pi_i^2 \tag{71}$$

These additional degrees of freedom allow us to construct an update which efficiently changes the system in a global manner.

To implement this update, we use **Hamilton's Equations**

$$\dot{\phi}_i = \frac{\partial H}{\partial \pi_i}, \qquad \dot{\pi}_i = -\frac{\partial H}{\partial \phi_i} \tag{72}$$

where we take $H(\phi, \pi) = E(\phi) + T(\pi)$. It can be easily shown that these equations conserve H

$$\frac{\mathrm{d}H}{\mathrm{d}t} = \sum_i \left(\frac{\partial H}{\partial \phi_i} \dot{\phi}_i + \frac{\partial H}{\partial \pi_i} \dot{\pi}_i \right) = \sum_i \left(\frac{\partial H}{\partial \phi_i} \frac{\partial H}{\partial \pi_i} - \frac{\partial H}{\partial \pi_i} \frac{\partial H}{\partial \phi_i} \right) = 0. \tag{73}$$

However, in practice these equations will be integrated numerically, resulting in errors which depend on the integration stepsize. These errors will introduce a finite ΔH accumulated over a period of evolution (commonly termed a **trajectory**). With these considerations in mind, we devise the following algorithm:

1. Call the current configuration of the system ϕ.
2. Draw $\{\pi_i\}$ from the normal distribution with zero mean and unit variance.
3. Numerically integrate Hamiltion's equations over a distance τ in time, using a stepsize $\delta\tau$ and (ϕ, π) as the initial values. This will result in a new configuration (ϕ', π').
4. Accept this new configuration with probability

$$p_{acc} = \min(1, \mathrm{e}^{-\Delta H}). \tag{74}$$

This algorithm can be viewed as an implementation of the Hastings algorithm with a complicated (but still reversible!) proposal step. So far we have not discussed any possible numerical integration schemes. We first note that in order for the proposal step to be reversible, we need an integration scheme which is reversible and **symplectic**, i.e. area-preserving.

Such schemes are easily constructed, but we focus on a common scheme here, namely the **'leapfrog'** algorithm. Given the state of the system at time t, we specify a procedure to obtain the state of the system at a time $t + h$, where h is the stepsize of the integration.

1. $\phi_i(t + \frac{h}{2}) = \phi_i(t) + \frac{h}{2}\pi_i(t)$
2. $\pi_i(t + h) = \pi_i(t) - hF_i(t + \frac{h}{2})$
3. $\phi_i(t + h) = \phi_i(t + \frac{h}{2}) + \frac{h}{2}\pi_i(t + h)$

where $F_i = \frac{\partial H}{\partial \phi_i}$. Clearly, the algorithm gets its name by alternated half steps of ϕ with whole steps of π. Such a scheme can easily be shown to be reversible. The idea of reversibility is as follows. After proceeding for a trajectory of length h, after the transformation $\pi \to -\pi$ and another trajectory of length h, the system should go back to the starting point.

In other words, if we set $\phi'(t) = \phi(t + h)$ and $\pi'(t) = -\pi(t + h)$, we should have $\phi'(t + h) = \phi(t)$ and $\pi'(t + h) = -\pi(t)$. Therefore

$$\begin{aligned} \phi'_i(t + h) &= \phi'_i(t + \frac{h}{2}) + \frac{h}{2}\pi'_i(t + h) \\ &= \phi'_i(t) + h\pi'_i(t) - \frac{h^2}{2}F'_i(t + \frac{h}{2}) \\ &= \phi_i(t + \frac{h}{2}) - \frac{h}{2}\pi_i(t + h) - \frac{h^2}{2}F'_i(t + \frac{h}{2}) \\ &= \phi_i(t) + \frac{h^2}{2}F_i(t + \frac{h}{2}) - \frac{h^2}{2}F'_i(t + \frac{h}{2}) \end{aligned} \tag{75}$$

observing that

$$\phi_i(t+\frac{h}{2}) = \phi_i'(t) + \frac{h}{2}\pi'(t) = \phi_i(t+h) - \frac{h}{2}\pi_i(t+h) = \phi_i(t+\frac{h}{2}) \tag{76}$$

so that $F_i'(t+\frac{h}{2}) = F_i(t+\frac{h}{2})$ completes the result. The analagous proof for π is considerably simpler:

$$\begin{aligned}\pi_i'(t+h) &= \pi_i'(t) - hF_i'(t+\frac{h}{2}) \\ &= -\pi_i(t+h) - hF_i(t+\frac{h}{2}) \\ &= -\pi_i(t) + hF_i(t+\frac{h}{2}) - hF_i(t+\frac{h}{2}).\end{aligned} \tag{77}$$

In order to demonstrate that a particular transformation is symplectic, we must show that it preserves all infinitesimal oriented area elements. These are defined by the area of the parallelogram formed by two infinitesimal vectors $\mathrm{d}\mathbf{v}_1$ and $\mathrm{d}\mathbf{v}_2$

$$\text{oriented area} = \mathrm{d}\mathbf{v}_1^T\, J\, \mathrm{d}\mathbf{v}_2, \qquad J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \tag{78}$$

We wish to ensure that this oriented area is preserved under an integration step, i.e.

$$\mathrm{d}\mathbf{v}_1^T\, J\, \mathrm{d}\mathbf{v}_2 = \mathrm{d}\mathbf{v'}_1^T\, J\, \mathrm{d}\mathbf{v'}_2, \qquad \mathrm{d}\mathbf{v'}_i = \begin{pmatrix} \frac{\partial\pi'}{\partial\pi} & \frac{\partial\pi'}{\partial\phi} \\ \frac{\partial\phi'}{\partial\phi} & \frac{\partial\phi'}{\partial\pi} \end{pmatrix} \begin{pmatrix} \mathrm{d}\pi_i \\ \mathrm{d}\phi_i \end{pmatrix}. \tag{79}$$

We see that this condition is fufilled if

$$A^T\, J\, A = J, \qquad A = \begin{pmatrix} \frac{\partial\pi'}{\partial\pi} & \frac{\partial\pi'}{\partial\phi} \\ \frac{\partial\phi'}{\partial\phi} & \frac{\partial\phi'}{\partial\pi} \end{pmatrix}. \tag{80}$$

With some algebra, this can be demonstrated for the leapfrog integrator.

We illustrate the leapfrog algorithm in a simple example. This example consists of a single gaussian distributed variable so that $E(\phi) = \frac{1}{2}\phi^2$. We introduce π in the usual way and implement the leapfrog integrator in `hmc1.cc`. For a fixed trajectory length of $\tau = 100.0$, we examine the dependence of the acceptance rate on $\delta\tau$ for $1e5$ updates.

The results for the leapfrog algorithm are shown in Tab. 2. We see that the acceptance rate is resonable all the way up to $\delta\tau = 2.0$, at which point it drops sharply to zero. This somewhat unusual behavior can be understood by examining the leapfrog integrator in matrix form. Leapfrog is composed of two basic transformations

$$I_1(h) : (\pi, \phi) \to (\pi, \phi + h\pi) \tag{81}$$

$$I_2(h) : (\pi, \phi) \to (\pi - hF, \phi) \tag{82}$$

such that a complete iteration of the integrator can be expressed as

$$I(h) = I_1(\frac{h}{2}) I_2(h) I_1(\frac{h}{2}). \tag{83}$$

alg.	N_t	$\delta\tau$	acc
LPFRG	100	1.00	0.92
	80	1.25	0.85
	70	1.43	0.78
	60	1.66	0.66
	50	2.0	0.99
	49	2.04	0.0
OMF2	60	1.66	0.99
	50	2.0	0.96
	45	2.2	0.87
	40	2.5	0.58
	39	2.56	0.0

Table 2: Acceptance rates for the second order leapfrog (LPFRG) and Omelyan-Mryglod-Folk (OMF2) integrators for a single normally distributed degree of freedom.

When examined as a matrix in the $\phi - \pi$ plane, we see that the eigenvalues of $I(h)$ undergo a qualitative change at $h = 2.0$. For $h < 2.0$, they are 'complex' phases that simply move the system along the unit circle, while for $h > 2.0$ they are real-valued, so that the system is driven far from the starting point, resulting in a large ΔH.

The position of this integrator instability is problem and integrator dependent. We can consider an additional integrator, which, like leapfrog is second order, that is $\Delta H \sim \mathrm{O}(h^2)$. This Omelyan-Mryglod-Folk (OMF) integrator has the form

$$I(h) = I_1(\xi h) I_2(\frac{h}{2}) I_1\left((1-2\xi)h\right) I_2(\frac{h}{2}) I_1(\xi h), \qquad \xi \approx 0.1931833 \tag{84}$$

and is implemented for the same problem in `hmc2.cc`. Results for the acceptance rate for this integrator are also shown in Tab. 2. We see here that the instability happens at significantly larger values of $\delta\tau$.

Somewhat more generally, the strategy for choosing τ and $\delta\tau$ is as follows. A small τ will cause the system to move only a short distance, resulting in large autocorrelations. However, τ chosen too large will be computationally inefficient, as a large number of integration steps will be required to obtain a reasonable acceptance. In typical applications $\tau \sim 1 - 2$ is generally sufficient.