Monte Carlo
A general overview of methods, theory and practice

Alexandros Sopasakis
Department of Mathematics
UNC Charlotte

Lectures

- Lecture 1- Background and theoretical/analytical development of the Monte Carlo method
- Lecture 2- Numerical simulation practices and common techniques used in modern modeling applications
- Lecture 3- A research project perspective: application to traffic flow

Historical background and general development

Monte Carlo simulation idea is older than the computer

MC method originally used to estimate pathological integrals as well as π.

Comte de Buffon showed in 1777 that:

$$\frac{M}{N} \xrightarrow{N\to\infty} \frac{2l}{\pi d} \text{ for } d > l$$

Wolf 1850 (3.1596)

Smith 1855 (3.1553)

Lazzarini 1901 (3.14159292)

William Thomson also describes an early 1901 MC method for the calculation of the motion of molecules undergoing collision in a gas (credit William Anderson).

It is believed that the first real application of the statistical sampling method was undertaken by Enrico Fermi in the 1930s.

Ulam, Metropolis and von Neumann reinvented Fermi's statistical sampling methods around 1947.

Nicolas Metropolis named the *statistical sampling* method used at the time *Monte Carlo* in a paper published in 1949

Dorrie in 1965 solves the equivalent of "Buffon's needle" example using MC methods

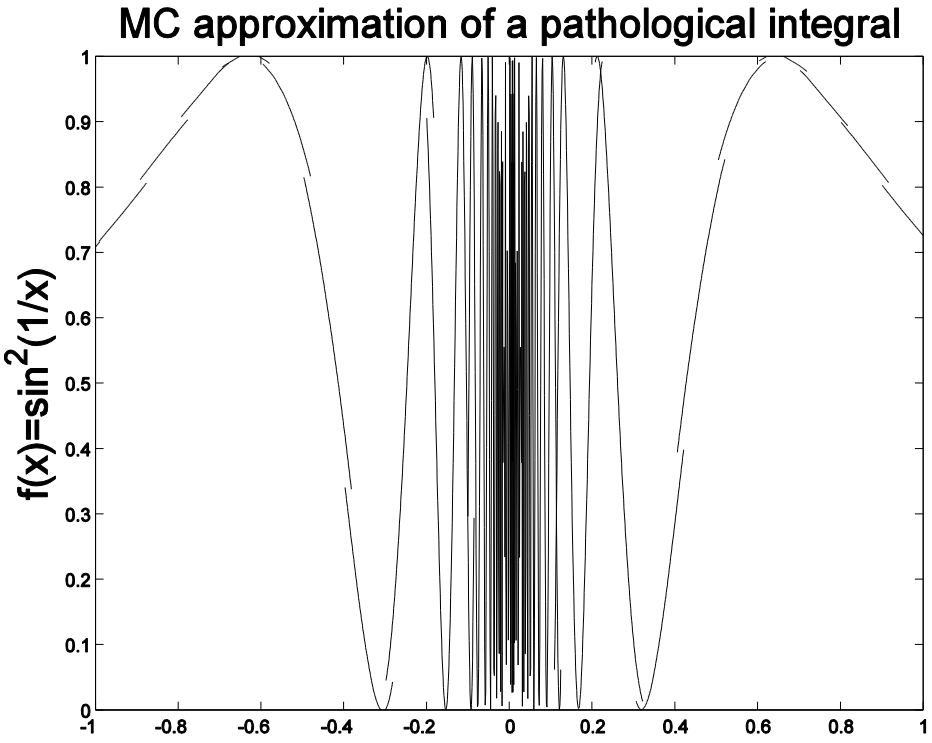
Beliefs and facts about Monte Carlo

"The only good Monte Carlo is a dead Monte Carlo" Trotter and Tukey 1954

"Anyone who considers arithmetic methods of producing random digits is, of course, in a state of sin" John Von Neumann, 1951

- Simple Monte Carlo – The direct modeling of a random process (queuing problems)
- Sophisticated Monte Carlo – Methods which recast deterministic problems in probabilistic terms

Monte Carlo Simulation: an archetypical example



Integrating pathological functions:

$$f(x) = \sin^2(1/x)$$

The integral $I(x) = \int_0^x \sin^2(1/y)dy$

is bounded by $0 \le I(x) \le x$

To use MC we choose

- random number u distributed between 0 and x and another
- random number v between 0 and 1.

These will represent our horizontal and vertical coordinates on the plot of f(x).

The probability that this point (u,v) will be below the line of f(x) is I(x)/x.

Thus for a large number N of these random points we count the number M of those bellow the line f(x) and obtain,

$$I(x) = \lim_{N\to\infty} \frac{Mx}{N}$$

Some facts to put things into perspective

- One liter of air at standard temperature and pressure contains about 3x10^22 molecules (oxygen, nitrogen, carbon dioxide, etc…).
- The atmosphere of the earth contains 4x10^21 liters of air or about 1x10^44 molecules (all moving around and colliding).
- Clearly it is not feasible to solve Hamilton's equations for these systems
  Too many equations
- Surprisingly however the macroscopic properties of air or gas are well-behaved and many times predictable.
- We can conclude that there must be something special about the behavior of the solutions of these many equations which "averages" out and gives us a predictable behavior for the system.

  This is where **statistical mechanics** is employed!

Facts About Simulating Large Systems

Let's consider a system which we wish to solve using a computational method.

- The most straightforward approach is to put this system into a lattice.
- Let's assume a small system of 10x10 (2 dimensional) lattice arrangement
- Suppose that each spin on every lattice node is allowed to take only two values: +1, -1
- Thus this small system has a total 2^100=33,554,432 possible states
- Fact: *even for a small problem it would be impossible to visit all states of the system!*
- Imagine the situation even if we only wish to model just one liter of oxygen=10^22 molecules. Furthermore, molecules change states at a rate of 10^9 collisions per second.
- At this rate it would take (10^10)^23 times the lifetime of the universe for our liter of gas to move through all its possible states.

It is therefore *impossible to just visit all states* in order to solve such a system!

What do we do then?

A general modeling paradigm for large interacting systems

- Our Hamiltonian system is fed by a thermal reservoir
- The thermal reservoir is an external system which acts as a source and sink of heat constantly exchanging energy with our system
- In general we think of the thermal reservoir as a weak perturbation of our Hamiltonian system which we ignore when calculating the energy levels of our system.
- Effects of the reservoir can be incorporated in our calculations by assigning the system a rule by which it can change from one state to another. This rule gives the system its dynamics
- There are several different types of dynamics – to be listed later- which can describe the physics for out system

Outline of what will follow

1. Markov Chain Monte Carlo (MCMC)

   Task: sample from a given probability distribution $\pi = (\pi(\mu))_{\mu \in \Sigma}$
   where Σ denotes our state space

   Idea: construct a discrete-time Markov Chain with a known probability matrix

   P= $(P(\mu \to \nu))_{\mu,\nu \in \Sigma}$ having π as a stationary (invariant) distribution.

We will:

- Define the transition probabilites $P(\mu \to \nu) = g(\mu \to \nu)\, A(\mu \to \nu)$
- Pick the selection probabilities $g(\mu \to \nu)$ for a proposed move from μ to ν
- Discuss optimal acceptance ratios $A(\mu \to \nu)$
- Use ergodicity to ensure that it is the invariant distribution π we sample
- Condition the Markov Chain to satisfy detailed balance.

2. Continuous Time Monte Carlo (CTMC) – Kinetic Monte Carlo

We similarly can construct a continuous-time Monte Carlo Chain with invariant measure π

Note that

- The random jump time is known (exponentially distributed, etc.) and defines the time step dt of the simulator
- No rejected moves!
- Drawback: algorithmically difficult to implement and computer memory intensive.
- CTMC is “real” dynamics

Some useful definitions

1. We say that states $\nu, \mu \in \Sigma$ **communicate** if

$$P(\nu \to \mu) > 0, \quad P(\mu \to \nu) > 0$$

If all states communicate the Markov chain is called **irreducible**.

2. Suppose, $P(X_{t+k} = \nu \mid X_t = \nu) > 0$

we define the **period** of a state ν to be the greatest common divisor of all k greater than or equal to 1.

A Markov chain is called **aperiodic** if each state has period 1.

3. If state $\nu \in \Sigma$ is revisited with probability 1 at some finite time then it is called **recurrent**; otherwise the state is called **transient**.

Markov Processes-building the Markov Chain

For purposes of MC methods a **Markov Process** is a mechanism by which a system is taken from one state to another in a random fashion.

The probability of changing from state μ to state ν is called the **transition probability** and is denoted by $P(\mu \rightarrow \nu)$. *Transition probabilities are responsible for building the Markov Process.* Transition probabilities must satisfy three conditions:

- Stay constant over time
- Depend only on the current states μ, ν (memoryless property), using the Markov property: supppose a Markov sequence Xn takes the vales a1…aN:

$$P(x_n = a_{i_n} \mid x_{n-1} = a_{i_{n-1}}, \ldots, x_1 = a_{i_1}) = P(x_n = a_{i_n} \mid x_{n-1} = a_{i_{n-1}})$$

- $$\sum_{\nu} P(\mu \rightarrow \nu) = 1$$

Note that it is possible to go from state μ back to the same state μ with a non-zero probability.

During a MC simulation we will use this mechanism to generate a **Markov Chain**.

Main idea will be to run this mechanism long enough so that new states will appear with probabilities given by their corresponding Boltzmann distribution.
When that happens we say that the system has **equilibrated**.

System description

We assume that our systems starts at a given state μ and **define the rates P(μ->ν).** Thus P(μ→ν)Δt is the probability that is will be in state ν at time Δt later.

We also define a set of **weights Wμ(t)** which represent the probability that the system will be in state μ at time t.

The system is described by the following **Master Equation** which is a rule for the evolution of the weights Wμ(t) in terms of the rates P(μ→ν):

$$\frac{dw_\mu}{dt} = \sum_\nu [w_\nu(t)P(\nu \to \mu) - w_\mu(t)P(\mu \to \nu)]$$

subject to $\sum_\mu w_\mu(t) = 1 \quad \text{for all } t$

Statistical mechanics deals with these weights and represent our entire knowledge of the state of the system.

Some statistical mechanics concepts. Equilibrium:

Define the **equilibrium probabilities**:

$$p_\mu = \lim_{t\to\infty} w_\mu(t)$$

Boltzmann distribution – [Gibbs, 1902]:

$$p_\mu = \frac{1}{Z} e^{-E_\mu / kT}$$

Z is known as the **partition function**.

$$Z = \sum_\mu e^{-E_\mu / kT} = \sum_\mu e^{-\beta E_\mu}$$

Obtaining statistical information for quantities of interest

In terms of measuring quantities of interest, once we obtain the solution for $W_\mu(t)$ we can also obtain information for quantities of interest, say Q,

$$\langle Q \rangle = \sum_\mu Q_\mu w_\mu(t)$$

At equilibrium $\lim_{t\to\infty} w_\mu(t) = p_\mu = \frac{1}{Z} e^{-\beta E_\mu}$ this becomes,

$$\langle Q \rangle = \frac{\sum_\mu Q_\mu e^{-\beta E_\mu}}{\sum_\mu e^{-\beta E_\mu}}$$

for ALL states

Question: Can we really calculate this quantity?

Answer: Not really!

A toy example

Example: suppose a small 3D system 10x10x10. This system is so small that is actually of no use towards realistic predictions. Still, this system would have a total of 2^1000= 10^300 states.

What is currently possible? It is possible, in a very fast computer, to sample 10^10 states given a few hours. In other words we would only sample 1 in every 10^290 states.

Problem: in cases of low temperature however very few states are responsible for the majority of the behavior of this system.

Clearly not possible to sample over all states that a system may visit…

How Monte Carlo Works:

Monte Carlo techniques work by **sampling only a subset** of all possible states….

Question: *is it enough to just sample a finite number, M, of those states*?

$$Q_M := \frac{\sum_{\mu=1..M} Q_\mu e^{-\beta E_\mu}}{\sum_{\mu=1..M} e^{-\beta E_\mu}}$$ ← **BAD IDEA**

Answer: *Yes, but not exactly as given in the formula above!*

Questions:

But which M states should we choose?
Do any M states work?
*H*ow do we choose the important few states which are responsible for the majority of the behavior for this system?

Answer: **Importance Sampling**! That is how Monte Carlo works.

Importance Sampling
Which states are the most important?

What is importance sampling?

Pick each state μ based on its Boltzmann probability $p_\mu = e^{-\beta E_\mu} / Z$

$$Q_M = \frac{\sum_{\mu=1..M} Q_\mu p_\mu^{-1} e^{-\beta E_\mu}}{\sum_{\mu=1..M} p_\mu^{-1} e^{-\beta E_\mu}} = \frac{1}{M} \sum_{\mu=1}^{M} Q_\mu$$

Note that Q_M is a good estimator for <Q> since in fact

$Q_M \rightarrow$ <Q> as M → all states

Question: still… how do we really pick these states with their correct Boltzmann probabilities?

Ergodicity-Reaching all states

Note that states do not all need to have a non-zero probability assigned to them. In practice most transition probabilities are set to zero as long as there are exist paths which connect any state to any other one.

Our Markov Chain must be able to reach all states of the system – otherwise we will not be able to produce new states with their correct Boltzmann probabilities. Reaching all states from any starting state is the property of **ergodicity.**

Detailed Balance-ensuring equilibrium distribution

Detailed balance is responsible for ensuring that ***it is*** the Boltzmann probability which we generate, instead of any other distribution. This is equivalent in fact to saying that the system is in equilibrium.

We can achieve this by simply allowing the rates by which the system transitions into and out of state μ to be equal:

$$\sum_{\nu} p_{\mu} P(\mu \to \nu) = \sum_{\nu} p_{\nu} P(\nu \to \mu)$$

which, since $\sum_{\nu} P(\mu \to \nu) = 1$ simplifies to $p_{\mu} = \sum_{\nu} p_{\nu} P(\nu \to \mu)$

The probability Wν(t+1) of being in state ν at time t+1 is given by,

$$w_\nu(t+1) = \sum_\mu P(\mu \to \nu) w_\mu(t)$$

In matrix notation this becomes

$$w(t+1) = P \cdot w(t)$$

Simple Equilibrium: $w(\infty) = P \cdot w(\infty)$

Dynamic Equilibrium (limit cycle): $w(\infty) = P^n \cdot w(\infty)$

Guaranteeing the desired probability distribution

$$p_\mu P(\mu \to \nu) = p_\nu P(\nu \to \mu)$$

This also enforces **Detailed Balance**!

Note that detailed balance inherently forbids the appearance of limit cycles.

Once the limit cycles are removed we can prove that

as t→infinity the w(t) → exponentially towards the eigenvector corresponding to the largest eigenvalue of the stochastic matrix P.

Note also that from $p_\mu P(\mu \to \nu) = p_\nu P(\nu \to \mu)$ we can get

$$\frac{P(\mu \to \nu)}{P(\nu \to \mu)} = \frac{p_\nu}{p_\mu}$$

Which, given that $p_\mu = Z^{-1} e^{-\beta E_\mu}$ gives

$$\frac{P(\mu \to \nu)}{P(\nu \to \mu)} = e^{-\beta(E_\mu - E_\nu)}$$

As a result, one possible choice for the transition probabilities would be

$$P(\mu \to \nu) \propto e^{-\beta(E_\mu - E_\nu)}$$

This however is not really a very good choice….

Acceptance Ratios or

How to construct an efficient MC algorithm

Let's start by defining the transition probability as follows,

$$P(\mu \to \nu) = g(\mu \to \nu) A(\mu \to \nu)$$

where

g(μ→ν) represents the **selection probability** (0<=g<=1)

A(μ→ν) is the **acceptance ratio** (0<=A<=1)

So now we have the freedom to choose higher acceptance rations:

$$\frac{P(\mu \to \nu)}{P(\nu \to \mu)} = \frac{g(\mu \to \nu) A(\mu \to \nu)}{g(\nu \to \mu) A(\nu \to \mu)}$$

Note that the ratio $0 \le \frac{A(\mu \to \nu)}{A(\nu \to \mu)} < \infty$

while $g(\mu \to \nu), g(\nu \to \mu)$ can take any values

Good MC algorithm practices

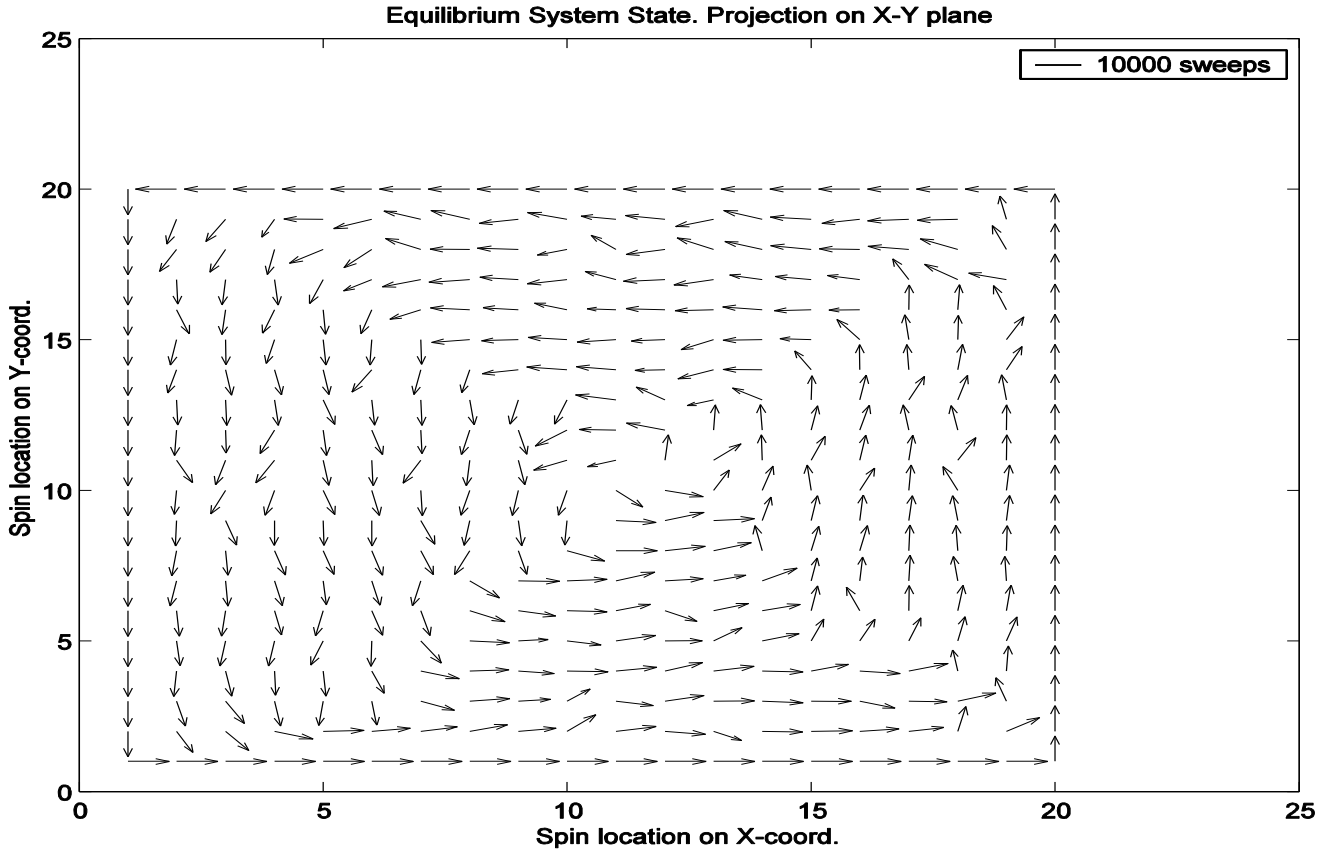
We wish to maximize our chance of accepting a new state $A(\mu \rightarrow \nu)$ (near 1)while at the same time satisfying the previous equation. This will ensure that at least in one direction the acceptance ratio will be as high as possible.

Ideal algorithm: new states selected with the correct transition probabilities while the acceptance ratio is always 1.

Good algorithm: again correct transition probabilities and acceptance ratios close to 1.

An example on picking acceptance ratios and selections probabilities: The Ising Model

An Ising model consists of micromagnets (dipoles) on a lattice. The magnetic spins are only allowed to take two values 0,1 which represent the magnetic charge of the dipole. If the lattice has N nodes then the system can be in any of 2^N possible states at any given time.



The Ising Model

An Ising model consists of micromagnets (dipoles) on a lattice. The magnetic spins are only allowed to take two values 0,1 which represent the magnetic charge of the dipole. If the lattice has N nodes then the system can be in any of 2^N possible states at any given time.

The energy of this system is given by its Hamiltonian:

$$H = -J \sum_{<i,j>} s_i s_j - h \sum_i s_i$$

where J is the interaction energy

<i,j> represent nearest neighbour spins

h is a given external field (for now we will assume h=0)

We typically wish to examine the average magnetization m as well as the specific heat c for this system

The Metropolis algorithm for single spin-flip Ising Model dynamics

In the Metropolis algorithm all selection probabilities are chosen to be equal. In fact one such natural choice is to let them be 1/N where N represents the total number of nodes in the lattice

$$g(\mu \to \nu) = \frac{1}{N}$$

Revisiting now our ratio of transition probabilities we obtain,

$$\frac{P(\mu \to \nu)}{P(\nu \to \mu)} = \frac{g(\mu \to \nu)A(\mu \to \nu)}{g(\nu \to \mu)A(\nu \to \mu)} = \frac{A(\mu \to \nu)}{A(\nu \to \mu)}$$

Recall that $\frac{P(\mu \to \nu)}{P(\nu \to \mu)} = e^{-\beta(E_\mu - E_\nu)}$ and $A(\mu \to \nu) = A_0 e^{-\frac{1}{2}\beta(E_\mu - E_\nu)}$

Question: how big a constant Ao can we choose to maximize the acceptance ratio?

The Metropolis algorithm (continuous)

Thus for a lattice node with four neighbors the difference of Eμ-Eν can be as big as 8J.

Thus the acceptance ratio $A(\mu \to \nu) = A_0 e^{-\frac{1}{2}\beta(E_\mu - E_\nu)}$

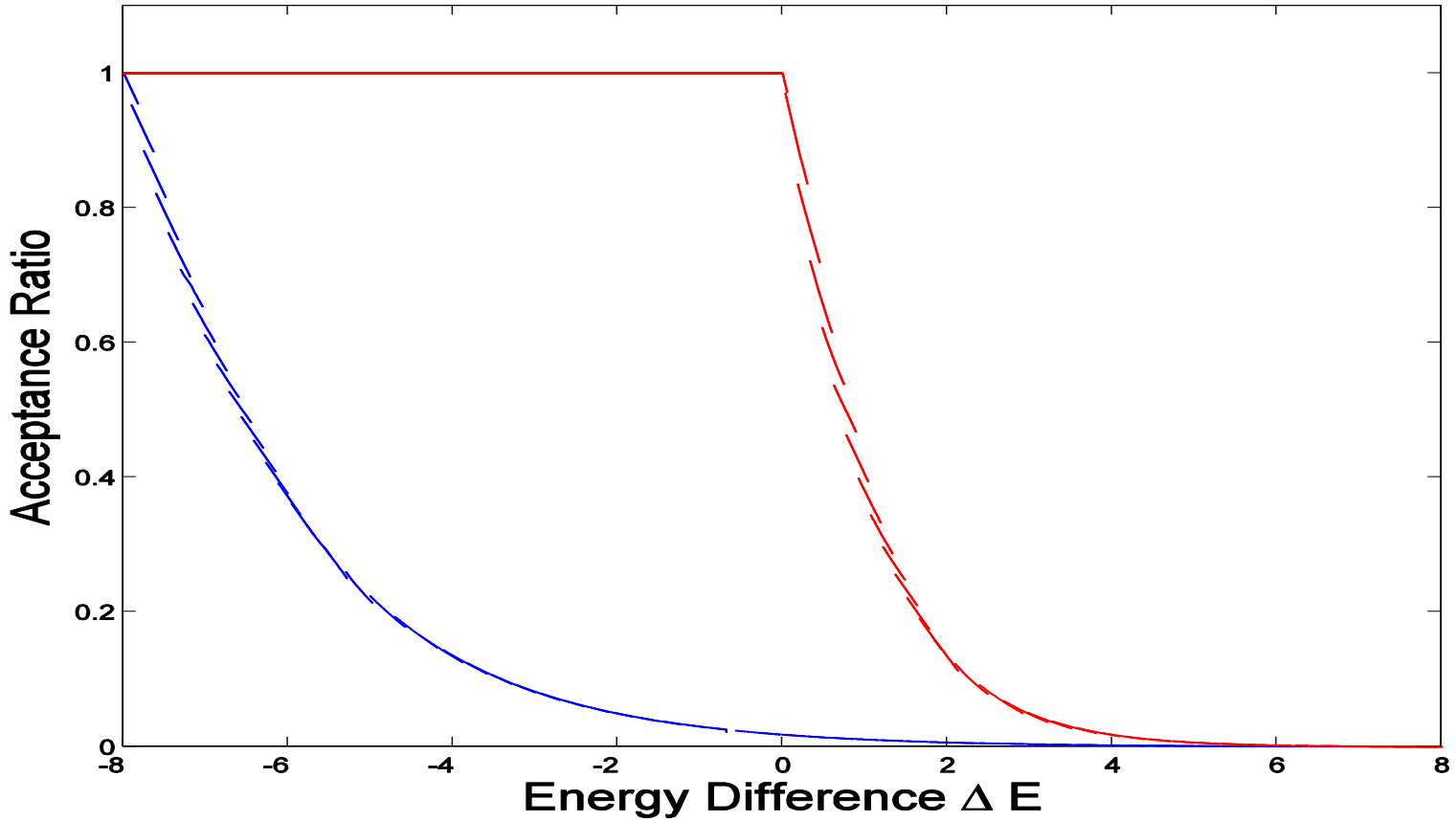
can be as big as $A(\mu \to \nu) \le A_0 e^{4\beta J}$

Therefore in order to maximize the acceptance ratio A(μ→ν) (close to 1) we choose

$$A_0 \le e^{-4\beta J}$$

Thus the best possible acceptance ratio is

$$A(\mu \to \nu) = e^{-\frac{1}{2}\beta(E_\mu - E_\nu + 8J)}$$



To make this more efficient we choose a different strategy:

$$A(\mu \to \nu) = \begin{cases} e^{-\beta(E_\nu - E_\mu)} & \text{if} \quad E_\nu > E_\mu \\ 1 & \text{otherwise} \end{cases}$$

This is the Metropolis algorithm!

Equilibrium and Autocorrelation

One way to obtain reliable information, such as calculating an expected value, about quantities of interest is to obtain

$$Corr(Q)(t^*) = \int (Q(t+t^*) - < Q >)(Q(t) - < Q >) dt$$

This also reveals the correct MC time τ necessary for equilibration.

As a rule of thumb we can use the resulting MC simulation and collect independent statistics for our parameters of interest every 2τ.

Books and other useful references

Books:

M. E. Kalos and P. A. Whitlock, 1986, *Monte Carlo Methods, Volume 1: Basics*, Wiley, New York.

K. Binder and D. W. Heermann, 1992, Monte Carlo simulation in statistical physics, Springer-Verlag, Berlin.

M. E. J. Newman and G. T. Bakerma, 1999, *Monte Carlo Methods in Statistical Physics*, Clarendon Press-Oxford.

More General Books:

Gibbs, 1902, Elementary principles in statistical mechanics(reprinted: 1981, Ox. Bow Press, Woodridge)

H. Dorrie, 1965, *One hundred great problems of elementary mathematics*, Dover, New York.

E. Segre, 1980, *From X-Rays to Quarks*, Freeman, San Fransisco

Most Important Scientific Papers:

N. Metropolis and S. Ulam, 1949, J. Am. Stat. Assoc., 44, 335.

Bortz, Kalos and Lebowitch, 1975, J. Comput. Phys. 17, 10.

Swendsen and Wang, 1987, Phys. Rev. Lett., 58, 86.

Kawasaki, 1965, Phys. Rev. 145, 224.

Sweeny, 1983, Phys. Rev. B, 27, 4445.

Wolff, 1989, Phys. Rev. Lett. 62, 361.