Numerical methods for lattice field theory

Mike Peardon

Trinity College Dublin

August 9, 2007

Numerical methods - references

- Good introduction to the concepts: “Simulation”, Sheldon M. Ross, Academic Press ISBN 0-12-598053-1
- Detail on the theory of Markov chains: “Markov chains, Gibbs fields, Monte Carlo simulations and queues”, Pierre Brémaud, Springer ISBN 0-387-98509-3
- A classic: “The Art of Computer Programming, Volume 2” Donald E. Knuth, Addison-Wesley ISBN 0-201-48541-9.
- And another: “Numerical Recipes: The Art of Scientific Computing (3rd Edition)”, Press, Teukolsky, Vetterling and Flannery, CUP ISBN 0-521-88068-8
- Applications to field theory: “The Monte Carlo method in quantum field theory”, Colin Morningstar `arXiv:hep-lat/0702020`

Markov Processes

- A more general method for generating points in configuration space is provided by considering Markov processes.
- In 1906, Markov was interested in demonstrating that independence was not necessary to prove the (weak) law of large numbers.
- He analysed the alternating patterns of vowels and consonants in Pushkin's novel "Eugene Onegin".

А. А. Марков (1886).

- In a Markov process, a system makes stochastic transitions such that the probability of a transition occuring depends only on the start and end states. The system retains no memory of how it came to be in the current state. The resulting sequence of states of the system is called a Markov chain.

Markov Processes (2)

A Markov Chain

Let $\{\psi_i\}$ for $i = 0..n+1$ be a sequence of states generated from a finite state space Ω by a stochastic process. Let $\chi_k \in \Omega$ be a state, such that $\chi_i \cup \chi_j = \emptyset$ if $i \neq j$ and $\Omega = \chi_1 \cup \chi_2 \cup \chi_3 \cup \dots \chi_m$. If the conditional probability obeys

$$P(\psi_{n+1} = \chi_i | \psi_n = \chi_j, \psi_{n-1} = \chi_{j_{n-1}}, \dots, \psi_0 = \chi_{j_0}) = P(\psi_{n+1} = \chi_i | \psi_n = \chi_j)$$

then the sequence is called a Markov Chain

- Moreover, if $P(\psi_{n+1} = \chi_i | \psi_n = \chi_j)$ is independent of n, the sequence is a *homogenous Markov chain*.
- From now on, most of the Markov chains we will consider will be homogenous, so we'll drop the label.

Markov Processes (3)

- The transition probabilites fully describe the system. They can be written in matrix form (usually called the Markov matrix);

$$M_{ij} = P(\psi_{n+1} = \chi_i | \psi_n = \chi_j)$$

- The probability the system is in a given state after one application of the process is then

$$P(\psi_{n+1} = \chi_i) = \sum_{j=1}^{m} P(\psi_{n+1} = \chi_i | \psi_n = \chi_j) P(\psi_n = \chi_j)$$

- Writing the probabilistic state of the system as a vector, application of the process looks like linear algebra

$$p_i(n+1) = P(\psi_{n+1} = \chi_i) = M_{ij} p_j(n)$$

Markov Processes (4)

- An example: Seattle's weather.
- It is noticed by a resident that on a rainy day in Seattle, the probability tomorrow is rainy is 80%. Similarly, on a sunny day the probability tomorrow is sunny is 40%.
- This suggests Seattle's weather can be described by a (homogenous) Markov process. From this data, can we compute the probability any given day is sunny or rainy?
- For this system, the Markov matrix is

$$\begin{array}{ccc} & \text{Sunny} & \text{Rainy} \\ \text{Sunny} & 0.4 & 0.2 \\ \text{Rainy} & 0.6 & 0.8 \end{array}$$

Markov Processes (5)

- If today is sunny, then $\psi_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, the state vector for tomorrow is then $\psi_1 = \begin{pmatrix} 0.4 \\ 0.6 \end{pmatrix}$, and $\psi_2 = \begin{pmatrix} 0.28 \\ 0.72 \end{pmatrix}$, $\psi_3 = \begin{pmatrix} 0.256 \\ 0.744 \end{pmatrix}$, ...
- If today is rainy, then $\psi_0 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, the state vector for tomorrow is then $\psi_1 = \begin{pmatrix} 0.2 \\ 0.8 \end{pmatrix}$, and $\psi_2 = \begin{pmatrix} 0.24 \\ 0.76 \end{pmatrix}$, $\psi_3 = \begin{pmatrix} 0.248 \\ 0.752 \end{pmatrix}$,
- The vector ψ quickly collapses to a fixed-point, which must be π, the eigenvector of M with eigenvalue 1, normalised such that $\sum_{i=1}^{2} \pi_i = 1$.
- We find $\pi = \begin{pmatrix} 0.25 \\ 0.75 \end{pmatrix}$. This is the invariant probability distribution of the process; with no prior information these are the probabilities any given day is sunny (25%) or rainy (75%).

The Markov Matrix (1)

- The Markov matrix has some elementary properties
  1. Since all elements are probabilities,

$$0 \leq M_{ij} \leq 1$$

  2. Since the system always ends in Ω,

$$\sum_{i=1}^{N} M_{ij} = 1$$

- From these properties alone, **the eigenvalues of M must be in the unit disk;** $|\lambda| \leq 1$, since if v is an eigenvector,

$$\sum_j M_{ij} v_j = \lambda v_i \implies |\sum_j M_{ij} v_j| = |\lambda||v_i| \implies \sum_j M_{ij}|v_j| \geq |\lambda||v_i|$$

$$\sum_j \left(|v_j| \sum_i M_{ij} \right) \geq |\lambda| \sum_i |v_i| \implies 1 \geq |\lambda|$$

The Markov Matrix (2)

- Also, **a Markov matrix must have at least one eigenvalue equal to unity.** Considering the vector $v_i = 1, \ \forall i$, we see

  $$\sum_i v_i M_{ij} = \sum_i M_{ij} = 1, \ \forall j$$

  and thus v is a left-eigenvector, with eigenvalue 1.

- Similarly, for the right-eigenvectors,

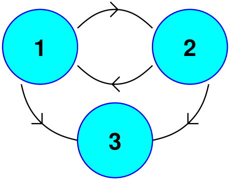
  $$\sum_j M_{ij} v_j = \lambda v_i \implies \sum_j v_j \sum_i M_{ij} = \lambda \sum_i v_i \implies \sum_j v_j = \lambda \sum_i v_i$$

  and so either $\lambda = 1$ or if $\lambda \neq 1$ then $\sum_i v_i = 0$

Topology of the Markov matrix

- To proceed, we need to consider the topology of the Markov matrix. By topology, we mean the properties of the transition graph, a directed graph with all states as nodes, and all non-zero transitions as links between nodes. For example, the diagram below shows a Markov matrix (with $0 < \kappa_{1,2,3,4} < 1$) and the corresponding transition graph.

$$\begin{pmatrix} 1-\kappa_1-\kappa_2 & \kappa_3 & 0 \\ \kappa_1 & \kappa_4 & 0 \\ \kappa_2 & 1-\kappa_3-\kappa_4 & 1 \end{pmatrix}$$



- A state χ_i is **accessible** from χ_j if there is some M such that $[M^N]_{ij} > 0$.
- States χ_i and χ_j **communicate** if χ_i is accessible from χ_j and χ_j is accessible from χ_i
- States that communicate fall into classes; for the example above the communication classes would be $\{1,2\},\{3\}$

Topology of the Markov matrix (2)

- If the chain has only one communication class, then it is said to be **irreducible**.
- Now define the **return time**, T_i for state χ_i. This is the biggest possible interval between successive occurances of χ_i in the chain.
- A state χ_i is called **recurrent** if $P(T_i < \infty) = 1$, otherwise it is called **transient**.
- A recurrent state is called **positive recurrent** if $E[T_i] < \infty$, otherwise it is **null recurrent** (positive recurrence means "the chain will return if you wait").
- Recurrence is a communication class property; if χ_i and χ_j communicate, they are both positive recurrent, both null recurrent or both transient.
- Since an irreducible chain has just one class, all states are have the same nature.

The important results...

Unique stationary state

An irreducible homogenous Markov chain has a unique stationary distribution if and only if it is positive recurrent.

- This stationary distribution is the only eigenvector of the Markov matrix that can be interpreted as a probability, since if $|\lambda| < 1$ then $\sum_i v_i = 0$, so the entries in v must be both positive and negative.

The ergodic theorem

Let $X_n, n \geq 0$ be an irreducible, positive recurrent Markov chain with stationary distribution π and let $f : \Omega \to \mathbb{R}$ such that

$$\sum_{\chi \in \Omega} |f(\chi)| \, \pi(\chi) < \infty$$

then for any initial state,

$$\lim_{N \to \infty} \sum_{k=1}^{N} f(X_k) = \sum_{\chi \in \Omega} f(\chi) \, \pi(\chi)$$

Detailed Balance

- So how can we construct Markov processes with our chosen fixed point, π?
- One very useful technique is to build methods that obey the **detailed balance** condition for π.

  $$M_{ij}\pi_j = M_{ji}\pi_i \qquad \text{(no sum)}$$

  and for a system described by continuous variables, this would read

  $$\mathcal{P}(\phi_i \leftarrow \phi_j)\pi(\phi_j) = \mathcal{P}(\phi_j \leftarrow \phi_i)\pi(\phi_i)$$

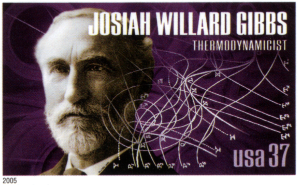
  where $\mathcal{P}$ now defines a probability density for making transitions from two states.
- If the Markov process obeys detailed balance for π, then

  $$M_{ij}\pi_j = M_{ji}\pi_i \implies \sum_i M_{ij}\pi_j = \sum_i M_{ji}\pi_i \implies \pi_j = \sum_i M_{ji}\pi_i$$

  and so π is the stationary distribution for M. If the chain is irreducible and positive recurrent then this is unique.

The Gibbs sampler

- A commonly used Markov process to generate importance sampling ensembles is the **Gibbs sampler**
- Use single-variable methods (transformation, rejection, etc.) and **update each degree of freedom in turn.** The update order only changes performance (as we shall see).



"A mathematician may say anything he pleases, but a physicist must be at least partially sane"

- Let the current state of a system with m degrees of freedom be

$$\Phi = \{\phi^{(1)}, \phi^{(2)}, \dots, \phi^{(q)}, \dots \phi^{(m)}\}$$

- and select a site, q to update and make the next entry in the chain

$$\Phi' = \{\phi^{(1)}, \phi^{(2)}, \dots, \phi'^{(q)}, \dots \phi^{(m)}\}$$

- by drawing $\phi'^{(q)}$ from the conditional probability

$$\pi^{(q)}(\phi^{(q)}) = \pi(\phi^{(q)} | \phi^{(p\neq q)})$$

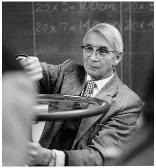
The Gibbs sampler (2)

- The theorem of conditional probability $P(A|B) = P(A \cap B)/P(B)$ tells us
$$\pi(\phi^{(q)}|\phi^{(p\neq q)}) = \frac{\pi(\Phi)}{\pi(\phi^{(p\neq q)})}$$
and
$$\pi(\phi'^{(q)}|\phi^{(p\neq q)}) = \frac{\pi(\Phi')}{\pi(\phi^{(p\neq q)})}$$
so
$$\frac{\pi(\phi'^{(q)}|\phi^{(p\neq q)})}{\pi(\phi^{(q)}|\phi^{(p\neq q)})} = \frac{\pi(\Phi')}{\pi(\Phi)}$$
- so if the single-site update obeys detailed balance for the local conditional probability (lhs) then it must also obey detailed balance for the full system's fixed point probability (rhs).
- If all sites can be updated, then the chain is usually irreducible and positive recurrent. Pathological cases can be constructed, however!

Example 1 - the Ising model in 1d

- A simple example: 1d Ising model with m spins.
- $\Sigma = \{\sigma^{(i)} \in \{-1,+1\}, i = 1..m\}$ with $\sigma^{(0)} \equiv \sigma^{(m)}$

$$S(\Sigma) = \sum_{i=1}^{m} (1 - \sigma^{(i)}\sigma^{(i+1)})$$



Ernst Ising

- We want to generate configurations with stationary probability

$$\pi(\Sigma) = \frac{1}{Z(\beta)} e^{-\beta S(\Sigma)}$$

(with β the inverse temperature) in order to perform importance sampling. Use the Gibbs sampler to define a Markov chain.

- At each step in the chain, replace spin $\sigma^{(q)}$ in Σ. Need the conditional probability for this spin in the stationary state. It is

$$\pi(\sigma^{(q)} | \{\sigma^{(p \neq q)}\}) = \frac{\pi(\Sigma)}{\sum_{\sigma^{(q)}=\pm 1} \pi(\Sigma)}$$

Example 1 - the Ising model in 1d (2)

- Split off the part of S which depends on $\sigma^{(q)}$, (This is a nearest-neighbour interaction, so computing this is cheap).

$$S = -\sigma^{(q)}\sigma^{(q+1)} - \sigma^{(q-1)}\sigma^{(q)} + \tilde{S}^{(q)}$$

  so $\tilde{S}^{(q)}$ is independent of $\sigma^{(q)}$

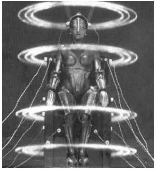
- The conditional probability becomes

$$\pi(\sigma^{(q)}|\{\sigma^{(p\neq q)}\}) = \frac{e^{\beta\sigma^{(q)}\mu^{(q)} - \beta\tilde{S}^{(q)}}}{\sum_{\sigma^{(q)}=\pm 1} e^{\beta\sigma^{(q)}\mu^{(q)} - \beta\tilde{S}^{(q)}}} = \frac{e^{\beta\sigma^{(q)}\mu^{(q)}}}{2\cosh \beta\sigma^{(q)}\mu^{(q)}}$$

  with $\mu^{(q)} = \sigma^{(q-1)} + \sigma^{(q+1)}$

- Now generating the Markov chain becomes
  1. Choose a site q, throw away the current value of $\sigma_i^{(q)}$
  2. Compute $\mu^{(q)}$ (a local computation).
  3. Draw a new value for $\sigma_{i+1}^{(q)}$ from the conditional probability

The Metropolis-Hastings method (1)



- Gibbs samplers are useful for local theories, but inefficient when all degrees of freedom interact with one-another.
- More general methods are needed. . .
- The Metropolis-Hastings algorithm is composed of two parts
  1. A **reversible proposal** step, which suggests a new state Φ'
  2. The **accept/reject** test.

$$\mathcal{P}_{\text{acc}} = \min\left[1, \frac{\pi(\Phi')}{\pi(\Phi)}\right]$$

- For a discrete system with states $\{\chi_1, \chi_2, \dots\}$, the Markov transition probability for moving from state $\chi_j \rightarrow \chi_i$ is then

$$M_{ij} = \min\left[1, \frac{\pi_i}{\pi_j}\right] R_{ij} \text{ when } i \neq j$$

where $R_{ij} = P(\chi_i \leftarrow \chi_j)$ is the conditional probability state χ_i is proposed (in step 1) given the current state is χ_j.

The Metropolis-Hastings method (2)

$$\frac{M_{ij}}{M_{ji}} = \frac{\min\left[1, \frac{\pi_i}{\pi_j}\right] R_{ij}}{\min\left[1, \frac{\pi_j}{\pi_i}\right] R_{ji}}$$

- Reversibility of the proposal step implies $R_{ij} = R_{ji}$ and if

  $$\frac{M_{ij}}{M_{ji}} = \frac{\pi_i}{\pi_j} \text{ if } \pi_i \leq \pi_j \text{ or } \frac{M_{ij}}{M_{ji}} = \frac{1}{\pi_j/\pi_i} = \frac{\pi_i}{\pi_j} \text{ if } \pi_i > \pi_j$$

  so detailed balance holds and if the proposal step is chosen carefully, the Markov chain is irreducible and positive recurrent so it is a suitable ensemble for importance sampling.
- For a system described by continuous random numbers, the proposal step must be an **area-preserving mapping**,

  $$\Phi' : \Omega \to \Omega \text{ such that } \left|\frac{\mathcal{D}\Phi'}{\mathcal{D}\Phi}\right| = 1$$

Autocorrelations

- While the Markov process is called memoryless, this does not mean nearby entries in the chain are independent random variables.
- Statistical analysis of Markov Chain Monte Carlo data must thus be done carefully.
- Markov chains must be “burnt in” to remove dependence on the initial state (often a very unlikely one!) before sampling is done.
- To see this, consider a two-state system with homogenous Markov matrix

$$M = \begin{pmatrix} 1-\kappa_1 & \kappa_2 \\ \kappa_1 & 1-\kappa_2 \end{pmatrix}, \quad 0 < \kappa_{1,2} < 1$$

- The conditional probability is

$$P(\psi_t = \chi_i | \psi_{t-n} = \chi_j) = [M^n]_{ij}$$

- The eigenvalues of M are $1, \lambda_2 = 1 - \kappa_1 - \kappa_2$ and

$$M^n = \frac{1}{\kappa_1 + \kappa_2} \begin{pmatrix} \kappa_1 + \lambda_2^n \kappa_2 & \kappa_1(1-\lambda_2^n) \\ \kappa_2(1-\lambda_2^n) & \kappa_2 + \lambda_2^n \kappa_1 \end{pmatrix}$$

Autocorrelations (2)

- M has two parts; one constant and the other $\propto \lambda_2^n$.
- Since the chain is irreducible and positive recurrent, $|\lambda_2| < 1$ so the correlation term **falls exponentially**; $\lambda_2^n = e^{-n/\tau}$ where $\tau = 1/\ln(1-\kappa_1-\kappa_2)$ is the (exponential) autocorrelation time.
- Suppose we measure some observable $F_i = f(\psi_i)$ on the states in our chain. The **autocorrelation function** of f on the chain is defined as

$$C_f(t) = E[(F_{i+t} - \mu_F)(F_i - \mu_F)] = E[F_t F_0] - \mu_F^2$$

so by definition, $C_f(0) = \sigma_F^2$ and the previous analysis suggests

$$\lim_{t\to\infty} C_f(t) \propto e^{-t/\tau}$$

where τ is related to the second-largest eigenvalues of the Markov matrix

Autocorrelations (3)

- We use our MC data to estimate an integral, and the sample variance is also used to provided an uncertainty in our determination (via the central limit theorem). For correlated data, more care is needed.
- The variance of the sample mean, $\bar{F} = \sum_{i=1}^{N} F_i$ for correlated data is

$$\sigma^2_{\bar{F}} = E[\bar{F}^2] - E[\bar{F}]^2 = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(E[F_i F_j] - \mu_F^2 \right)$$

then using the definition of the autocorrelation time

$$\sigma^2_{\bar{F}} = \frac{\sigma^2_F}{N} \left(1 + 2 \sum_{t=1}^{N-1} (1 - \frac{t}{N}) \frac{C_F(t)}{C_F(0)} \right)$$

- The quantity in the bracket is often called the **integrated autocorrelation time** of F. It is the ratio of the true variance of $\bar{F}$ to the "naive" variance (σ^2_F/N).

Autocorrelations (4)

- If measurements are expensive, make them on a subset of well-separated configurations.
- Another alternative is to **bin** data from the chain. Take the original sequence and average over adjacent pairs

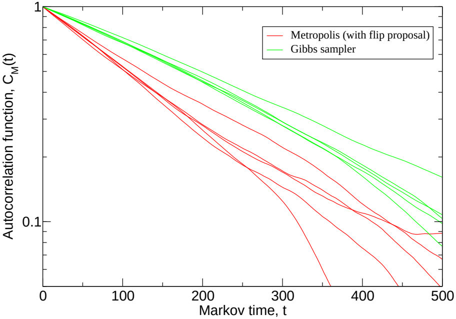
$$\underbrace{F_1, F_2,}_{\bar{F}_1^{(2)},} \underbrace{F_3, F_4,}_{\bar{F}_2^{(2)},} \underbrace{F_5, F_6,}_{\bar{F}_3^{(2)},}$$

  and compute the sample variance of $\bar{F}^{(2)}$. Since the autocorrelations in this new variable are weaker, the "naive" error estimate will be more reliable.
- NB: The "naive" variance of the mean (assuming N independent) measurements is $\sigma^2_{\bar{F}} = \sigma^2_F / N$
- This binning is then repeated until a stable error estimate is found.

For more detail, see "Markov chain Monte Carlo simulations and their statistical analysis" Bernd Berg (World Scientific ISBN 981-238-935-0).

Autocorrelations (5)

- Different algorithms (simulating the same physics) will have different autocorrelation behaviour.
- Want independent samples: **smaller autocorrelations are better**.
- Example - 1d Ising model comparing the Gibbs sampler to a Metropolis update where the proposal is a single spin-flip. Measure the autocorrelation of the total magnetisation.
- Metropolis is better.


Metropolis (with flip proposal)
Gibbs sampler
Autocorrelation function, $C_M(t)$
Markov time, t