Markov Chain Monte Carlo: the method and applications in statistical physics

C. Bonati
claudio.bonati@unipi.it

Università di Pisa
Istituto Nazionale di Fisica Nucleare, sezione di Pisa.

Seminari di Analisi Numerica
Pisa, 18/03/2013

Outline

1. A premise
2. Statistical physics
3. The Monte Carlo method
4. Markov chains
5. Details of a simple case
6. Markov Chain Monte Carlo
7. An example in statistical physics: the $O(N)$ models
8. Bibliography

An important premise

The Monte Carlo method is the best method only when all the other methods are worst.

Basic example

Numerical integration of a continuous function $f : [0,1] \to \mathbb{R}$

- Monte Carlo: error $\sim 1/\sqrt{N}$ ($N \sim$ number of operations)
- Rectangle method: error $\sim 1/N$
- Trapezoidal rule: error $\sim 1/N^2$
- In general, non MC methods: error $\sim 1/N^\alpha$ with $\alpha \geq 1$

For $n-$dimensional integrals $\alpha \to \alpha/n$ so that the MC method is the best one to compute definite integrals of functions $f : [0,1]^n \to \mathbb{R}$ for $n \gg 1$ (perfect in the limit $n \to \infty$).

Natural question: why $n \to \infty$?

Statistical physics in a line

$$\text{Statistical physics} = \text{Boltzmann distribution} = e^{-\beta E},\ \beta = 1/(kT)$$

Denoting by x a "physical configuration" (positions of the particles, spin directions, momenta, ...) and with X the space of these configurations, the physical observables (energy, magnetization, ...) are functions $\mathcal{O} : X \to \mathbb{R}$ and their thermodynamical value is

$$\langle \mathcal{O} \rangle = \frac{\int_X \mathcal{O}(x) e^{-\beta E(x)} \mathrm{d}x}{Z} \qquad Z = \int_X e^{-\beta E(x)} \mathrm{d}x$$

The aim of statistical physics is, given a model for $E(x)$, to compute the thermodynamical values of some relevant observables.

Typical scales

Number of particles $\sim N_A \simeq 6.0 \times 10^{23}$
Very optimistic estimate of the space dimension: $\dim(X) \sim 2^{10^{23}}$

Simple sampling

Random configurations $\{x_i\}_{i\in[1,N]}$ are generated and the observables are estimated by

$$\langle \mathcal{O} \rangle \simeq \frac{\sum_i \mathcal{O}(x_i) e^{-\beta E(x_i)}}{\sum_i e^{-\beta E_{x_i}}}$$

This method has (at least) two fundamental problems:

1. since E is an extensive variable (i.e. it grows proportionally to the volume), we typically have $|E| \gg 1$, so that serious precision problems arise in finite arithmetics
2. the configurations which are more important in the average are the ones with smallest E and these configurations are typically very hard to produce by chance: all the random configuration generated are in the tails of the distribution and thus big statistical errors are present.

Solution: importance sampling

We generate the configuration with probability $e^{-\beta E}$ and the observables are estimated by

$$\langle \mathcal{O} \rangle \simeq \frac{\sum_i \mathcal{O}(x_i)}{N}$$

New problem: how to generate configurations according to a given statistical distribution?

Solution: the Metropolis(-Hastings) method

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller *Equations of State Calculations by Fast Computing Machines*. Journal of Chemical Physics **21**, 1087 (1953).

W. K. Hastings *Monte Carlo Sampling Methods Using Markov Chains and Their Applications* Biometrika **57** 97 (1970).

Markov chains

A Markov chain is a stochastic process at discrete times in which the transition probability of the system at time t_n depends on the state of the system at time t_n but not on the states at $t < t_n$.
By denoting the configuration space by S and the space of the time parameters by T, then a Markov chain is characterized by the function

$$P : T \times S \times S \to [0, 1]$$

$P(t, x, y)$ is the probability of going at time t from the state x to the state y.

If S is a discrete set, $\#S = N$, P can be represented by a matrix $P_{ij}(t)$, $i, j \in \{1, \ldots, N\}$.

In the following we will restrict only to the case of homogeneous Markov chains, for which the transition probability is independent of t.

Classification of the states

Let us denote by $F_{xy}^{(n)}$ the probability that, given $X_0 = x$, we have $X_n = y$ and $X_i \neq y$ for $i < n$ and by $P_{xy}^{(n)}$ the probability that, given $X_0 = x$, we have $X_n = y$ (it is simple to see that $P_{xy}^{(n)} = (P^n)_{xy}$).

1. the period of a state x is defined as
   $d(x) = \mathrm{MCD}\{n \geq 1 \quad t.c. \quad P_{xx}^{(n)} > 0\}$
   and x is said to be aperiodic if $d(x) = 1$
2. a state is said to be persistent if $\sum_{n=1}^{\infty} F_{xx}^{(n)} = 1$ and transient otherwise
3. the recurrence time of a persistent state x is defined by
   $\mu(x) = \sum_{n=1}^{\infty} n\, F_{xx}^{(n)}$
4. a persistent state x is said to be null if $\mu(x) = +\infty$
5. an aperiodic and persistent state x is said to be ergodic if $\mu(x) < +\infty$

Classification of the states

Theorem Let x be a state of a Markov chain, then

- x is transient if and only if $\sum_{n=1}^{\infty} P_{xx}^{(n)} < +\infty$, and in this case we have $\sum_{n=1}^{\infty} P_{yx}^{(n)} < +\infty$ for every initial state y. In particular $\lim_{n\to\infty} P_{yx}^{(n)} = 0$
- x is a persistent null state if and only if $\sum_{n=1}^{\infty} P_{xx}^{(n)} = +\infty$ and $\lim_{n\to\infty} P_{xx}^{(n)} = 0$, and in this case we have $\lim_{n\to\infty} P_{yx}^{(n)} = 0$ for every initial state y
- if x is an ergodic state we have

$$\lim_{n\to\infty} P_{yx}^{(n)} = \frac{1}{\mu(x)} F_{yx} \equiv \frac{1}{\mu(x)} \sum_{k=1}^{\infty} F_{yx}^{(k)}$$

Irreducible chains

A Markov chain is said to be irreducible if for every couple x, y of states an integer $n \geq 1$ exists such that $P_{xy}^{(n)} > 0$.

Theorem All the states of an irreducible Markov chain are in the same class.

Theorem In an irreducible ergodic chain the limit $\pi_x = \lim_{n\to\infty} P_{yx}^{(n)}$ exists and it is independent of the initial state y, moreover $\pi_x > 0$ and

$$\sum_x \pi_x = 1 \qquad \pi_y = \sum_x \pi_x P_{xy} \qquad (*)$$

Theorem Let's assume to have an irreducible and aperiodic chain for which numbers $\pi_x \geq 0$ exists that satisfy $(*)$, then all the states are ergodic and $\pi_x = \lim_{n\to\infty} P_{yx}^{(n)}$.

A probability distribution w_x that satisfies $w_y = \sum_x w_x P_{xy}$ is said to be an invariant distribution for the Markov chain.

Irreducible chains

Corollary An irreducible ergodic chain has one and only one invariant distribution.

Ergodic theorem Given an irreducible ergodic chain and a limited function $f : S \to \mathbb{R}$ then we have

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) \to \bar{f} \text{ per } n \to \infty\right) = 1$$

where X_i is the state of the chain at time i and $\bar{f}$ is the average with respect to the invariant distribution

$$\bar{f} = \sum_{x \in S} \pi_x f(x)$$

Proof in a simple case

Let us consider the space of the probability distributions on a compact set $O \subset \mathbb{R}^d$ (i.e. the spere of $\mathbb{L}^1(O)$), and the transition probability $P \in \mathbb{L}^2(O \times O)$ to be a function such that $P(x \leftarrow y) \geq \epsilon > 0$ for every x, y (so we have an irreducible aperiodic chain!) and let us define the action of P on the distribution $q(x)$ by

$$(Pq)(x) = \int P(x \leftarrow y) q(y) \mathrm{d}y$$

Theorem: $\|Pq_1 - Pq_2\| \leq (1 - \epsilon')\|q_1 - q_2\|$ with $\epsilon' > 0$.

By the Banach fixed point, by iterating the application of P to a general starting distribution we converge to the invariant distribution of P.

Proof in a simple case

$$\Delta q(x) = q_1(x) - q_2(x)$$

$$\|Pq_1 - Pq_2\| = \int \mathrm{d}x |Pq_1(x) - Pq_2(x)| = \int \mathrm{d}x \left| \int \mathrm{d}y P(x \leftarrow y) \Delta q(y) \right| =$$

$$= \int \mathrm{d}x \left| \int \mathrm{d}y P(x \leftarrow y) \Delta q(y) \left[\Theta(\Delta q(y)) + \Theta(-\Delta q(y)) \right] \right|$$

We now use $\big||a| - |b|\big| = |a| + |b| - 2\min(|a|, |b|)$ to arrive to

$$\leq \int \mathrm{d}x \int \mathrm{d}y P(x \leftarrow y) \, |\Delta q(y)| -$$

$$- 2 \int \mathrm{d}x \min_{\pm} \left| \int \mathrm{d}y P(x \leftarrow y) \Delta q(y) \Theta(\pm \Delta q(y)) \right| \leq$$

$$\leq \int \mathrm{d}y \, |\Delta q(y)| - 2 \int \mathrm{d}x \left[\inf_y P(x \leftarrow y) \right] \min_{\pm} \left| \int \mathrm{d}y \Delta q(y) \Theta(\pm \Delta q(y)) \right|$$

Proof in a simple case

By noting that

$$\int \mathrm{d}y \Delta q(y) \Theta(\Delta q(y)) + \int \mathrm{d}y \Delta q(y) \Theta(-\Delta q(y)) = \int \mathrm{d}y \Delta q(y) = 1 - 1 = 0$$

and thus

$$\int \mathrm{d}y |\Delta q(y)| = \int \mathrm{d}y \Delta q(y) \Theta(\Delta q(y)) - \int \mathrm{d}y \Delta q(y) \Theta(-\Delta q(y)) =$$
$$= 2 \left| \int \mathrm{d}y \Delta q(y) \Theta(\pm \Delta q(y)) \right|$$

we arrive to

$$\|Pq_1 - Pq_2\| \leq \int \mathrm{d}y \, |\Delta q(y)| - \epsilon |O| \int \mathrm{d}y \, |\Delta q(y)| = (1 - \epsilon') \|q_1 - q_2\|$$

Markov Chain Monte Carlo

Method to generate configurations according to a given probability distribution $\mathscr{P}$: we use an irreducible aperiodic Markov chain that has $\mathscr{P}$ as invariant distribution.
We thus have to find a transition probability $P(x \leftarrow y)$ such that

$$\mathscr{P}(x) = \int \mathrm{d}y P(x \leftarrow y) \mathscr{P}(y)$$

Sufficient condition: detailed balance

$$P(y \leftarrow x) \mathscr{P}(x) = P(x \leftarrow y) \mathscr{P}(y)$$

(it is sufficient to integrate and use $\int \mathrm{d}y P(y \leftarrow x) = 1$)

The Metropolis(-Hastings) method

Let G be a given transition matrix, associated to an irreducible aperiodic Markov chain, and let us build a new Markov chain in the following way:

- given the configuration $X_n = x$ we suggest the transition $y = Gx$
- the suggested transition is accepted with probability a_{xy}
- if the transition has been accepted $X_{n+1} = y$, otherwise $X_{n+1} = X_n$

The transition matrix of this process is

$$P_{xy} = a_{xy} G_{xy}$$
$$P_{xx} = a_{xx} G_{xx} + \sum_{z \neq x} (1 - a_{xz}) G_{xz}$$

and the detailed balance becomes

$$\frac{a_{xy}}{a_{yx}} = \frac{\mathscr{P}_y G_{yx}}{\mathscr{P}_x G_{xy}}$$

The Metropolis(-Hastings) method

It is not difficult to find a closed form for a_{ij} that satisfies the detailed balance: we can use for example

$$a_{xy} = F\left(\frac{\mathscr{P}_y G_{yx}}{\mathscr{P}_x G_{xy}}\right)$$

with $F : [0, \infty] \to [0, 1]$ a function that satisfies $F(z) = zF(1/z)$. Such functions are e.g.

$$F(z) = \min(1, z) \qquad F(z) = \frac{z}{1+z}$$

In applications we typically have $G_{xy} = G_{yx}$ and the previous formula reduces to

$$a_{xy} = F\left(\frac{\mathscr{P}_y}{\mathscr{P}_x}\right) \qquad (\text{if } G_{xy} = G_{yx})$$

The $O(N)$ models

Let us consider a cubic lattice of size L, whose sites are denoted by $\{x_i\}$. To every site a variable $\vec{s}(x_i) \in \mathbb{R}^N$ with $|\vec{s}| = 1$ is associated and the total energy is given by the expression

$$E = -\sum_{\langle x_i x_j \rangle} \vec{s}(x_i) \cdot \vec{s}(x_j)$$

where the sum is over all the neighbour site couples. This model is relevant in various physical applications:

- for $N = 2$ it describes the superfluid transition
- for $N = 3$ is the (classical) Heisenberg model of ferromagnetism

In simulations, in order to reduce the finite size effects, periodic boundary conditions are usually assumed, so that the cube becomes in fact an hyper-torus.

Metropolis algorithm for the $O(N)$ models

- we start from a configuration (whatever) $\vec{s}(x_i)$
- for every lattice site x_j
  1. we generate the new vector $\vec{s}_{prop}$ by means of a random rotation of $\vec{s}(x_j)$
  2. we compute the energy difference ΔE due to the eventual substitution $\vec{s}(x_j) \to \vec{s}_{prop}$
  3. we generate a random number $r \in [0, 1]$
  4. if $r \leq \min(1, \exp(-\beta \Delta E))$ we perform the substitution $\vec{s}(x_j) \to \vec{s}_{prop}$, otherwise we let $\vec{s}(x_j)$ unaltered
- we iterate the previous point as much as we can and after every iteration we measure the value of the relevant observables.

In the point (1) there is much freedom: for example the rotation can be chosen in such a way that $|\vec{s}_{prop} - \vec{s}(x_j)| < \epsilon$, in order to have a slighter change in the energy and to have an higher acceptance probability at point (4).

The data autocorrelation

Let's assume to have a series of measurements $\{O_i\}_{i\in[1,n]}$ of some observable. The expectation value of the statistical error is given by

$$\left\langle \left(\frac{1}{n}\sum_{i=1}^{n} O_i - \langle O \rangle \right)^2 \right\rangle =$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \langle (O_i - \langle O \rangle)^2 \rangle + \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=i+1}^{n} \left(\langle O_i O_j \rangle - \langle O \rangle^2 \right) =$$

$$= \frac{1}{n}\left[\langle O^2 \rangle - \langle O \rangle^2 + 2\sum_{i=1}^{n} \left(1 - \frac{i}{n}\right) \left(\langle O_0 O_i \rangle - \langle O \rangle^2 \right) \right]$$

We define the autocorrelation time by

$$\tau_{auto} = \frac{1}{2} + \sum_{i=1}^{\infty} \frac{\langle O_0 O_i \rangle - \langle O \rangle^2}{\langle O^2 \rangle - \langle O \rangle^2}$$

The data autocorrelation

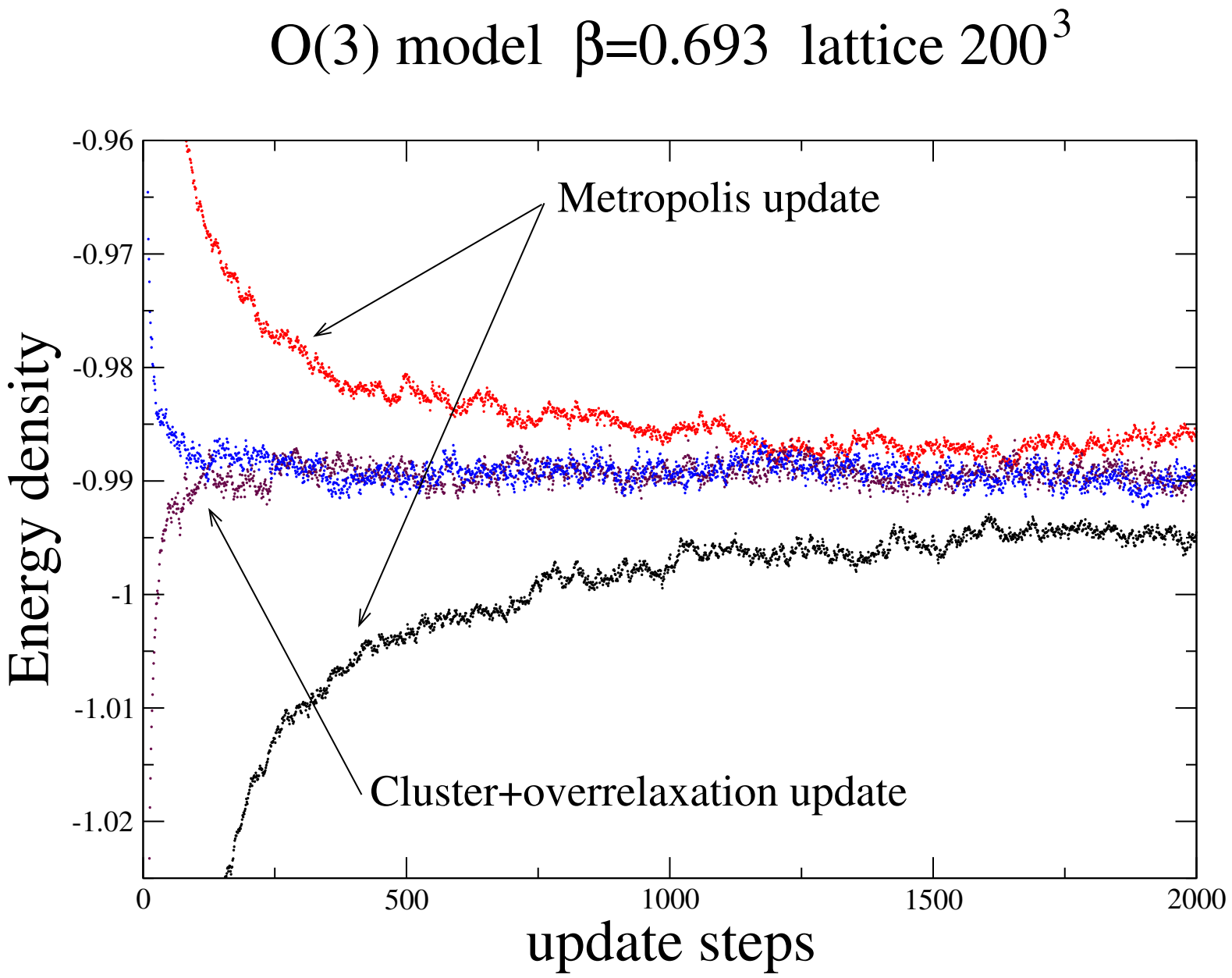
By using τ_{auto} we thus have

$$\langle (\delta O)^2 \rangle \approx \left(\langle O^2 \rangle - \langle O \rangle^2 \right) \frac{2\tau_{auto}}{n}$$

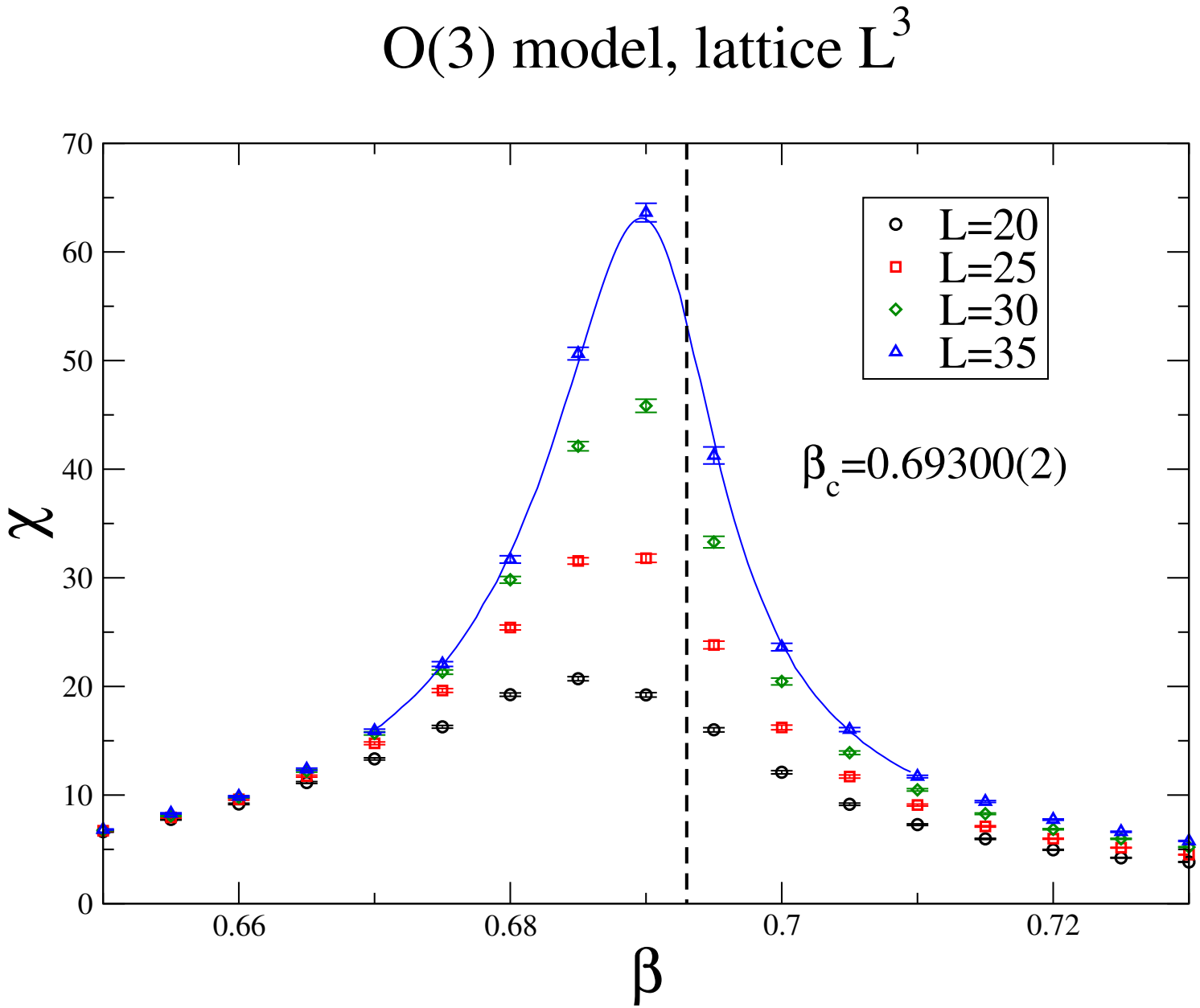
τ_{auto} is a carachteristic of the update algorithm: the smaller τ_{auto} the more the update is efficient.
The Metropolis algorithm is universal but typically not efficient.
Algorithms specific for some models but (much) more efficients are

1. heatbath
2. over-relaxation (also known as microcanonical update)
3. cluster updates
4. parallel tempering

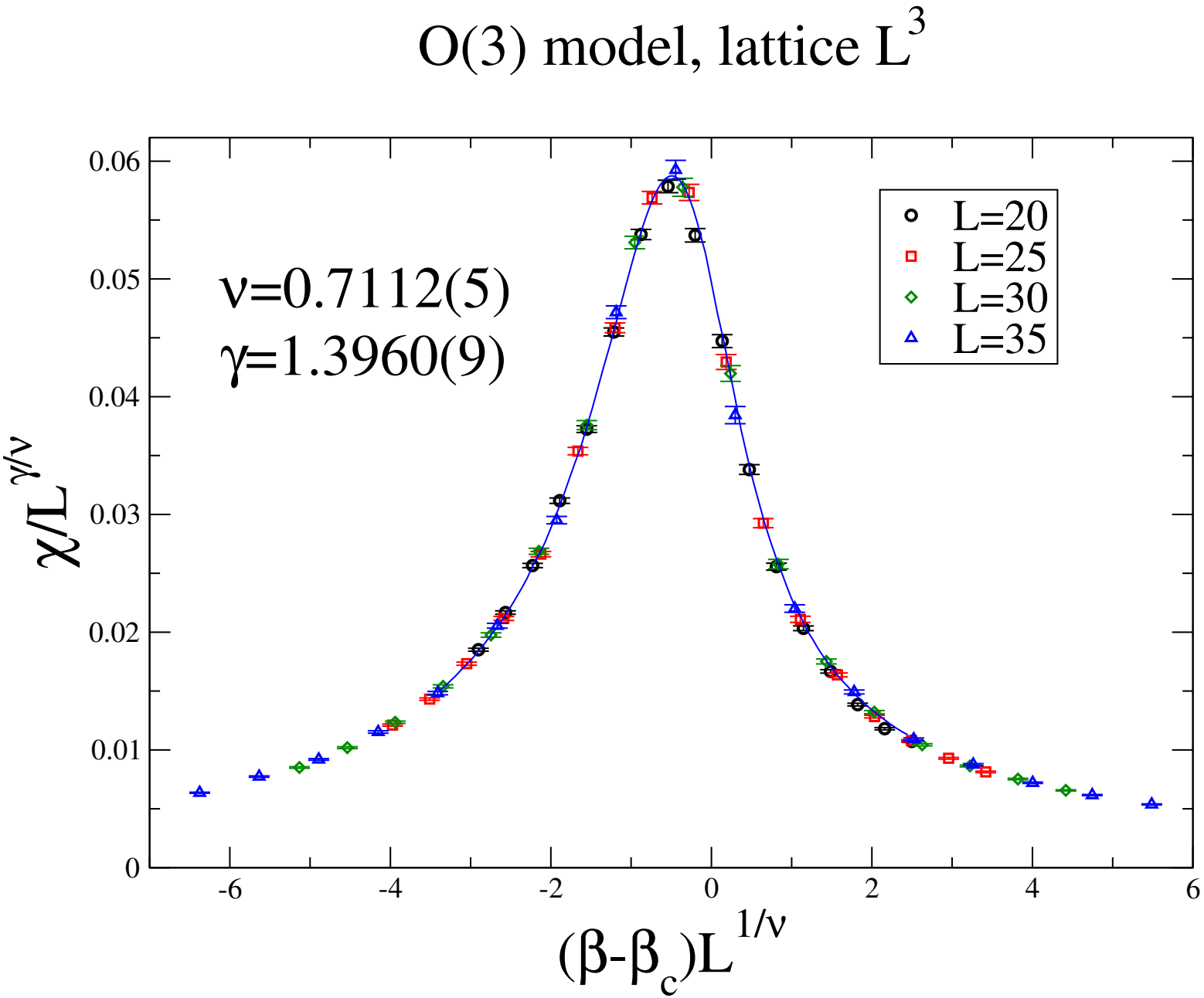
Real life comparison between algorithms



Real life physical example: Finite Size Scaling



Real life physical example: Finite Size Scaling



General bibliography

For statistical physics

- L. D. Landau, E. M. Lifsits "Statistical physics" Butterworth Heinemann
- D. Ruelle "Statistical Mechanics: Rigorous Results" World Scientific

For the theory of Markov chains

- W. Feller "An Introduction to Probability Theory and Its Applications" Wiley & Sons
- J. R. Norris "Markov Chains" Cambridge University Press

For statistical physics applications

- K. Binder, D. W. Heermann "Monte Carlo Simulation in Statistical Physics. An introduction" Springer Verlag
- D. L. Landau, K. Binder "A guide to Monte Carlo Simulations in Statistical Physics" Cambridge University Press
- M. E. J. Newman, G. T. Barkema "Monte Carlo Methods in Statistical Physics" Clarendon Press