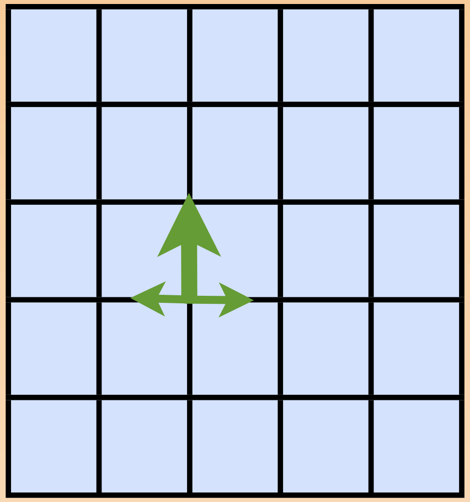
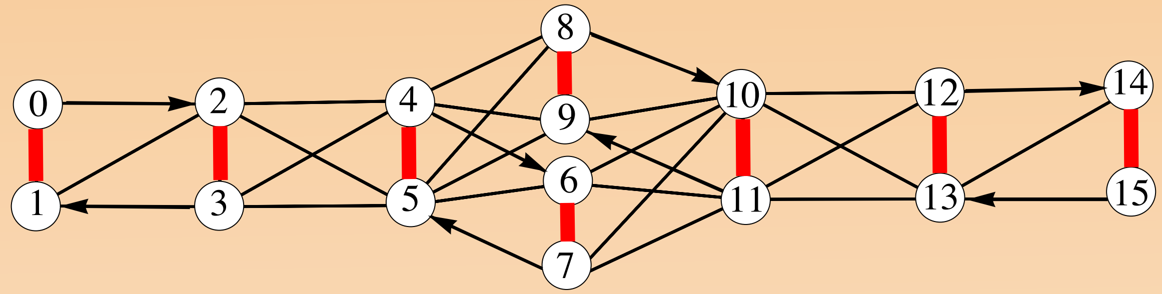
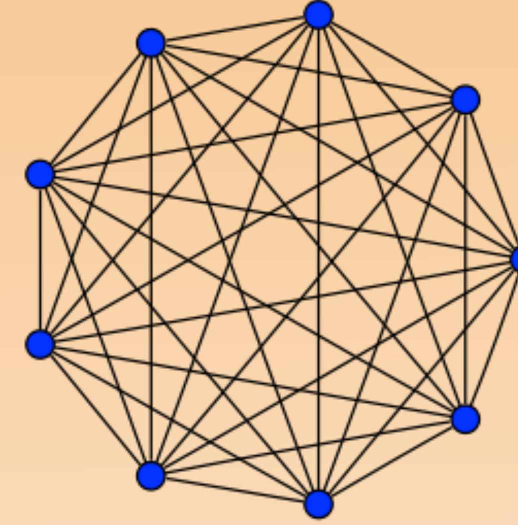






Sampling with non-reversible Markov Chains

Marija Vucelja
New York University

"Deep Computation in Statistical Physics", Santa Fe Institute, Aug 2013

MCMC sampling outline

- goal: create samples of a system at steady state
- reversible and non-reversible - physical intuition
- theorems: Peskun (reversible MC), multi-commodity flow
- **Examples**
    - torus $\mathcal{O}(n^2) \to \mathcal{O}(n)$
    - mixing on a permutation group $\mathcal{O}(n^3 \log n) \to \mathcal{O}(n^2)$
    - n-point path $\mathcal{O}(n^2) \to \mathcal{O}(n)$
    - mean field Ising model (Curie-Weiss) $\mathcal{O}(n^{3/2}) \to \mathcal{O}(n^{3/4})$
    - 1d Ising (**Koji Hukushima**)
    - Rejection free algorithms (**Werner Krauth**)
    - 2d Ising caveats
    - more 2d and 3d Ising (see **Koji Hukushima**'s talk)
    - spin-glasses caveats (parallel tempering, work with Jon Machta)

Detailed balance

Ω set of states

$T(x,y)$ transition matrix

$$\pi^t(y) = \sum_{x\in\Omega} \pi^t(x) T(x,y)$$

stochastic matrix

$$\sum_{y\in\Omega} T(x,y) = 1 \quad \forall x \in \Omega$$

If $T(x,y)$ is irreducible the a steady state exists and it is unique $\pi_s(y) = \sum_{x\in\Omega} \pi_s(x) T(x,y)$

Balance condition

$$\sum_{x\in\Omega} [\pi_s(x)T(x,y) - \pi_s(y)T(y,x)] = 0$$

$$\pi_s(x)T(x,y) = \pi_s(y)T(y,x)$$

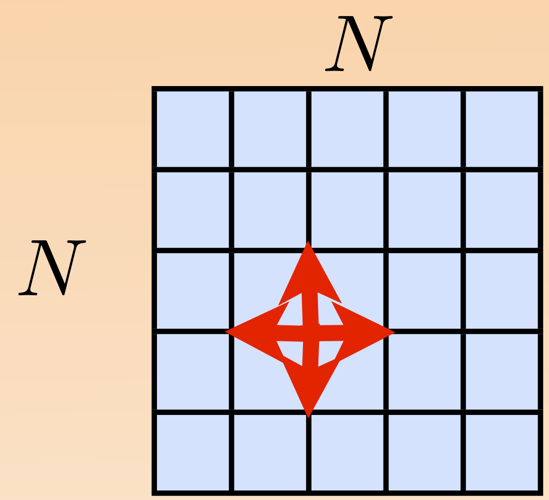
Detailed balance (reversibility):

Detailed balance is sufficient, but not necessary!

Lifting on a torus *Chen, Lovasz, Pak 1999*

goal: sample with uniform probability from a torus $N \times N$

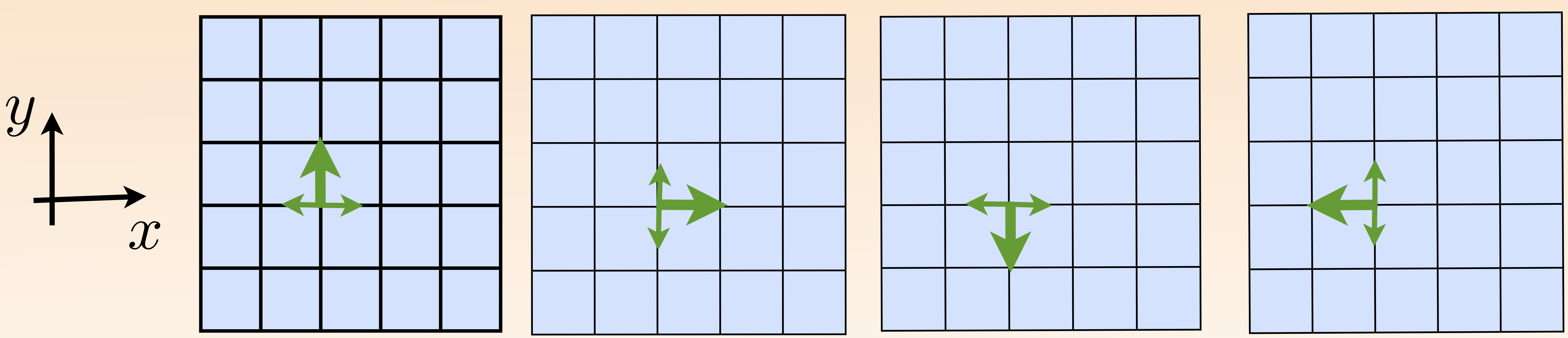
Diffusion



random walker on a torus
$p_N = p_W = p_E = p_S = 1/4$

mixing time on a torus $\mathcal{O}(N^2)$

Lifting added advection



north east south west

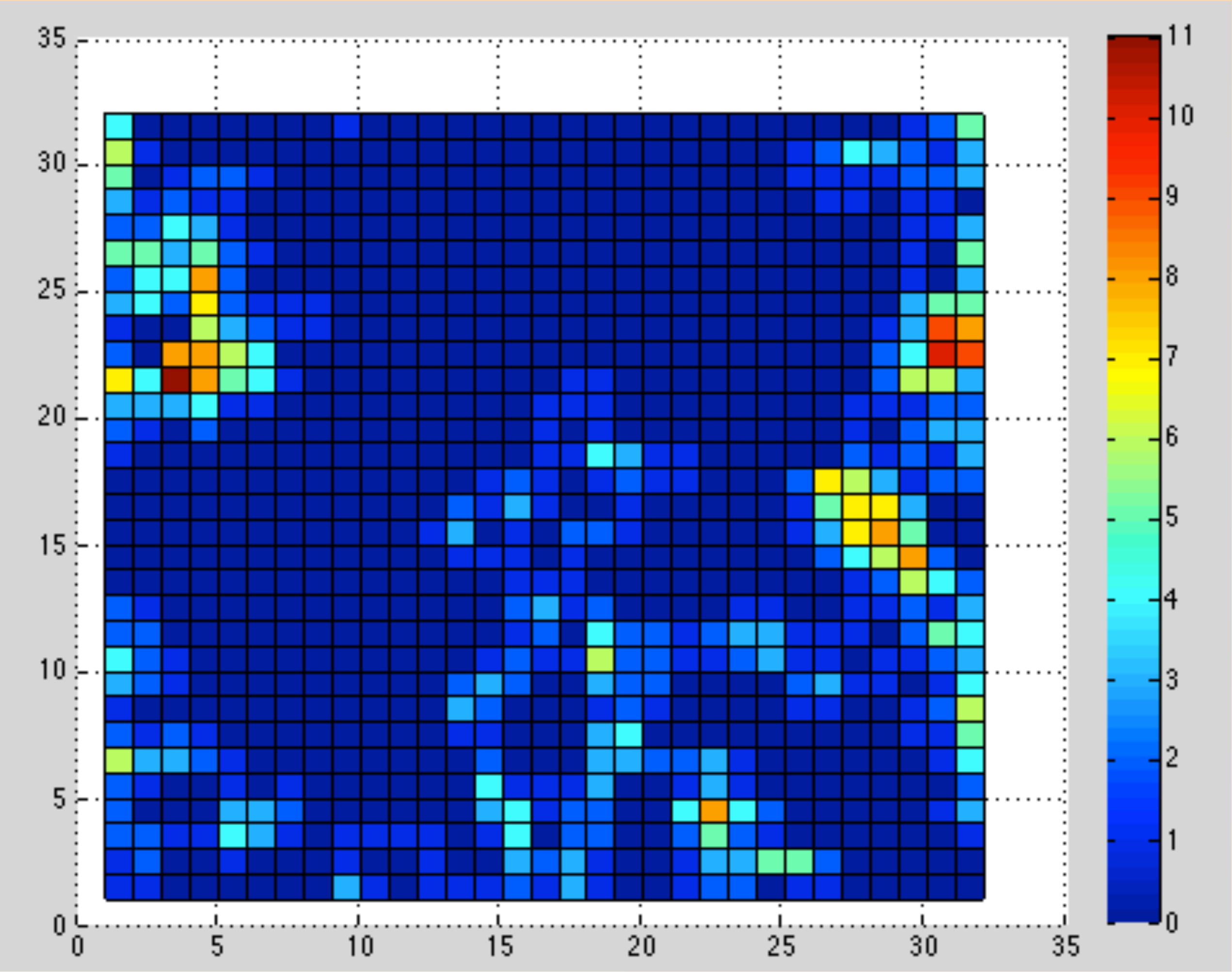
$p_N = 1 - N^{-1}$
$p_E = p_W = (2N)^{-1}$
$p_S = 0$

mixing time on a torus $\mathcal{O}(N)$

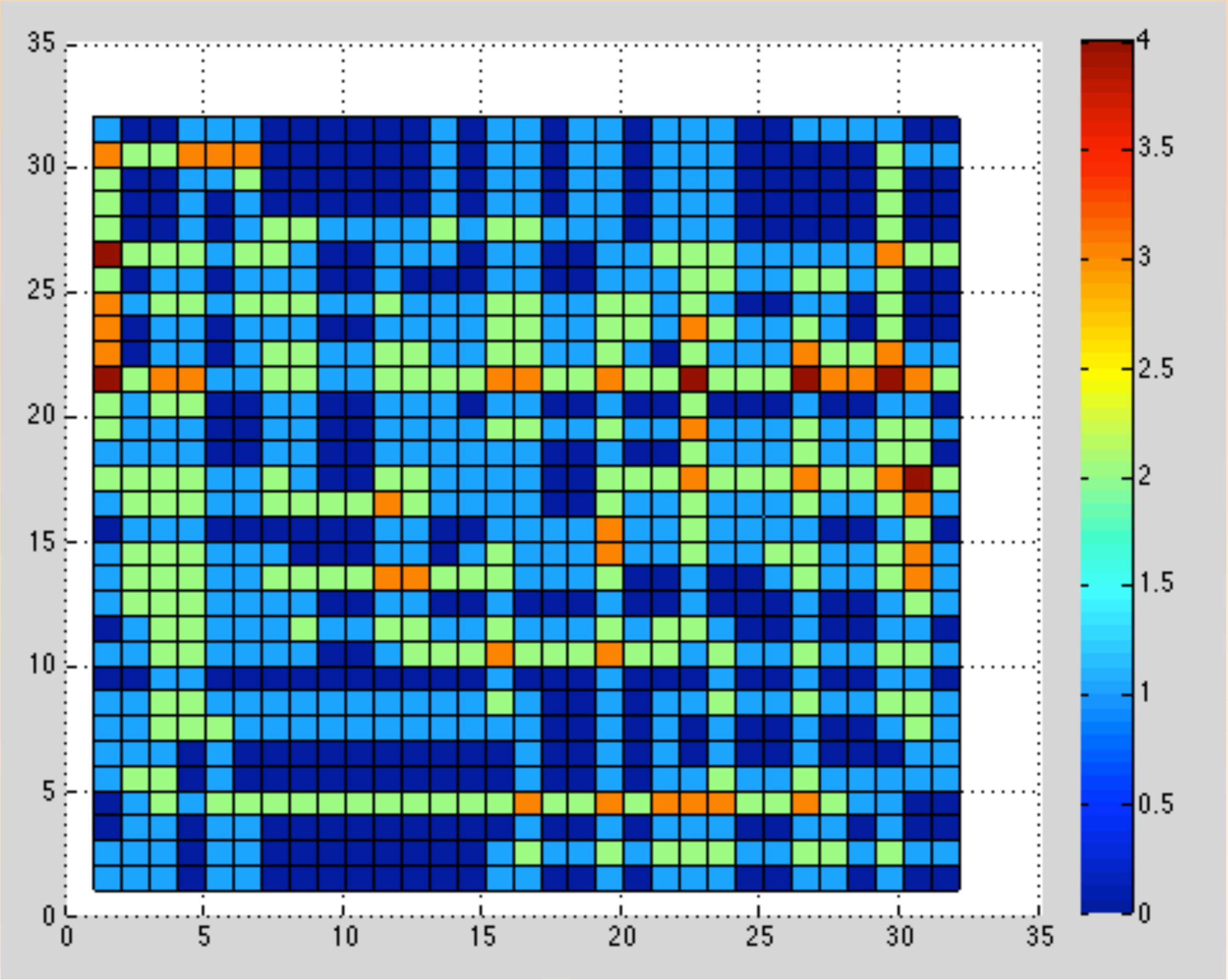
$\mathcal{O}(N)$ randomize along y-axis
$\mathcal{O}(N)$ randomize along x-axis

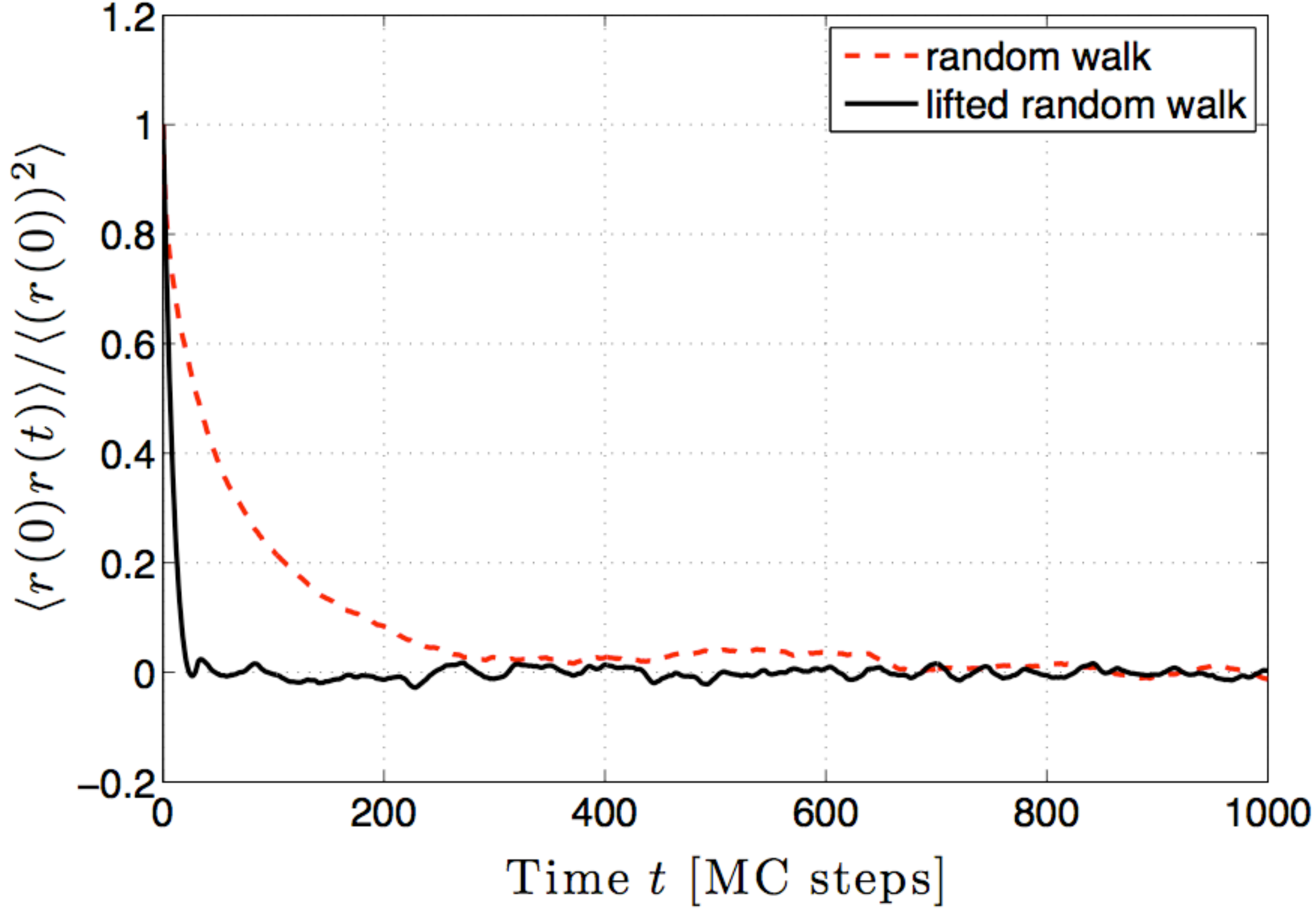
Density of visited sites on a torus of 1024 sites, after 1024 steps

Random walk

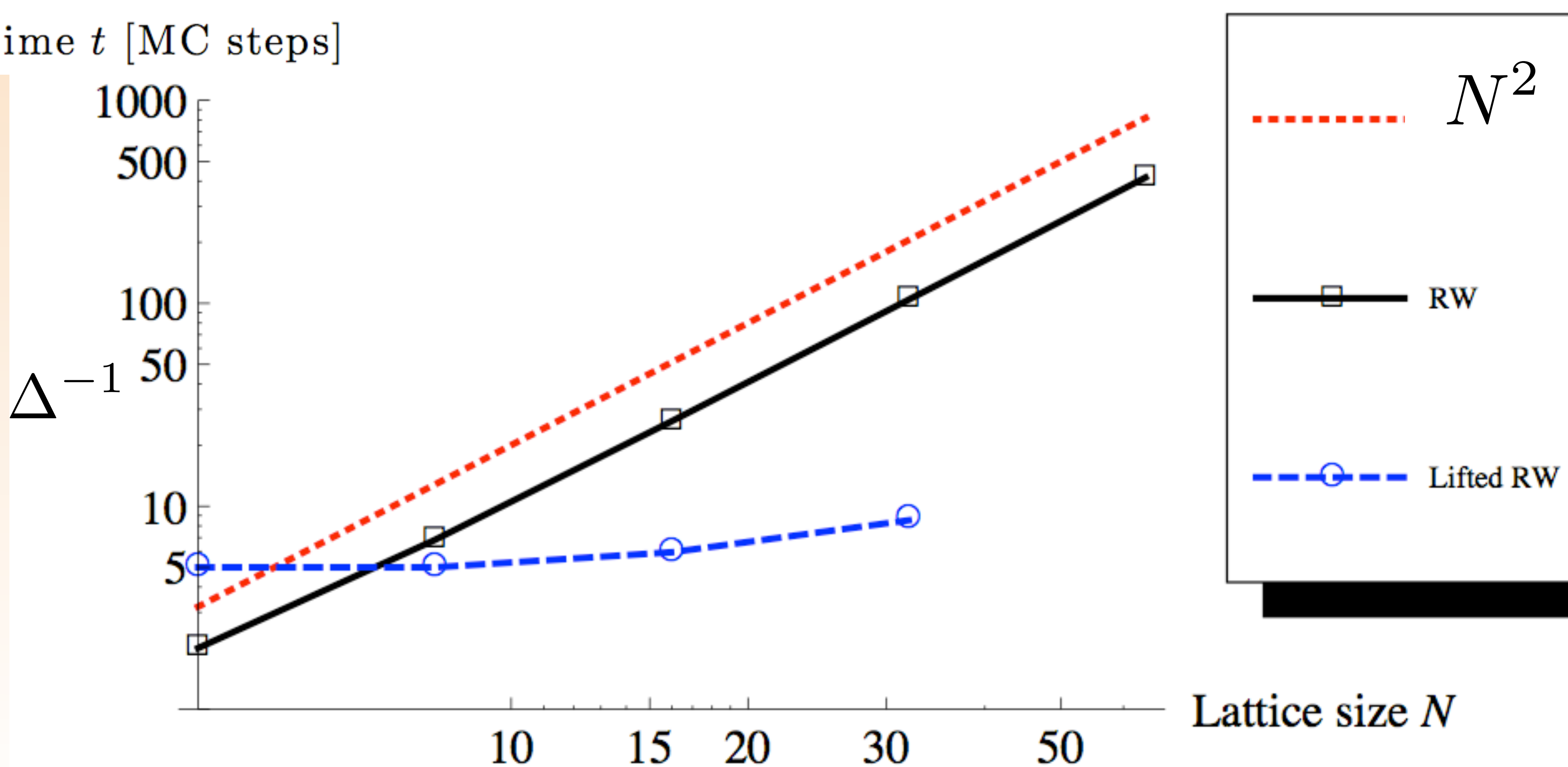


Lifted random walk





Autocorrelation of distance from origin $\boldsymbol{r}(t)$

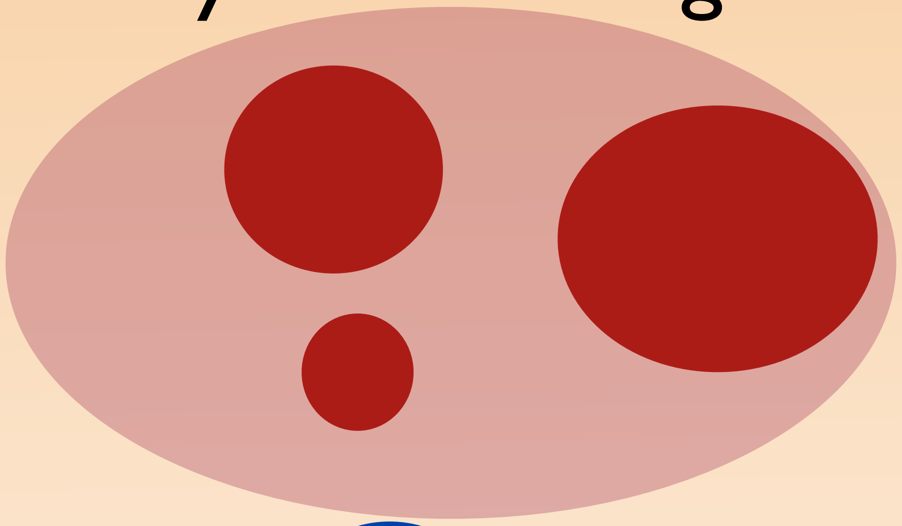


Inverse spectral gap

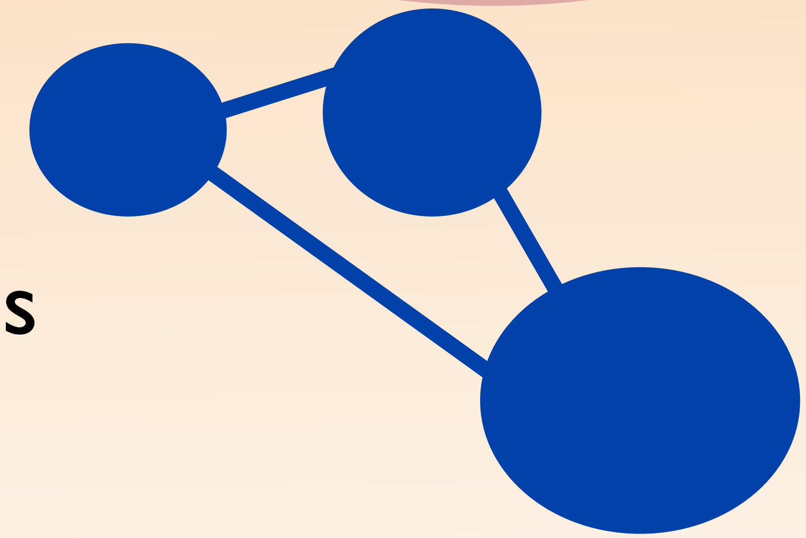
Slow convergence

Several types of distributions are characterized by slow mixing:

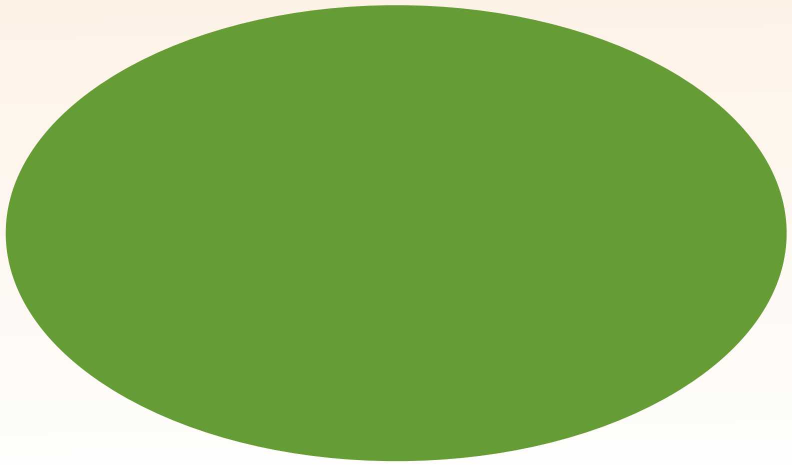
- Glassy landscapes: Regions that dominate the partition function are separated by "energy barriers"



- Entropy barriers: Regions of high probability are separated by narrow paths (high probability but small entropy)



- Single region with high probability of large size (entropy)



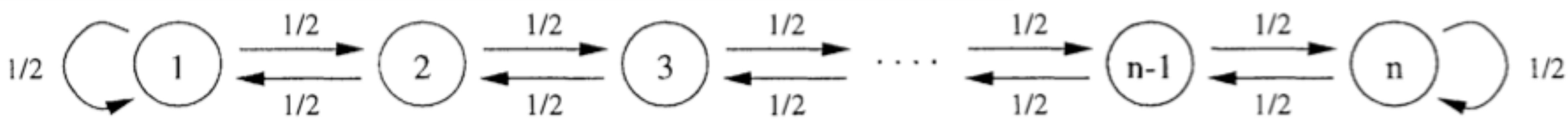
The Annals of Applied Probability
2000, Vol. 10, No. 3, 726–752

ANALYSIS OF A NONREVERSIBLE MARKOV CHAIN SAMPLER

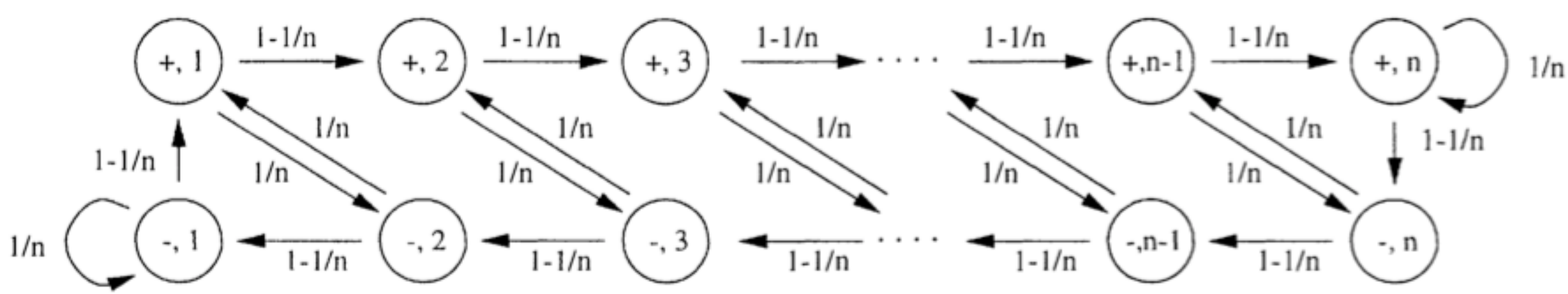
BY PERSI DIACONIS,[1] SUSAN HOLMES AND RADFORD M. NEAL[2]

Stanford University, Stanford University and INRA and University of Toronto

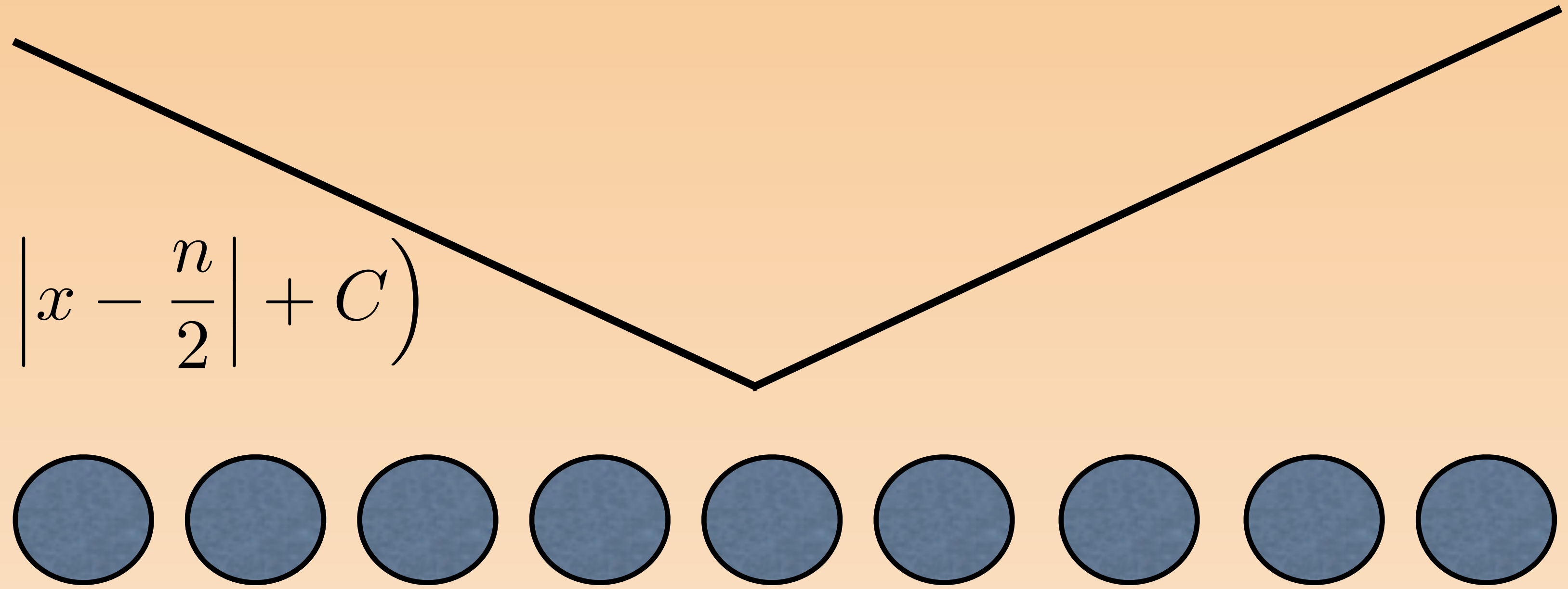
n-point walk



lifted n-point walk



$$\pi(x) = \frac{1}{Z}\left(1\left|x - \frac{n}{2}\right| + C\right)$$



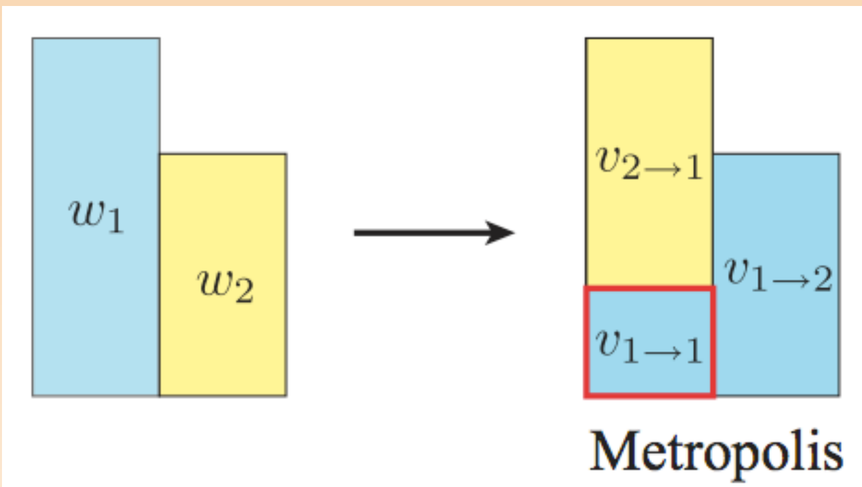
	Ideal	*Directed*	*Metropolis*	*Min. Prob.*
$C = 1,\ n = 50:$	0.00308	0.00151	0.000347	0.000769
$C = 1,\ n = 100:$	0.000785	0.000386	0.0000763	0.000196
$C = 1,\ n = 200:$	0.000198	0.0000979	0.0000170	0.0000495
$C = 2,\ n = 50:$	0.00593	0.00295	0.000479	0.00148
$C = 2,\ n = 100:$	0.00154	0.000758	0.000102	0.000385
$C = 2,\ n = 200:$	0.000392	0.000193	0.0000220	0.0000980

FIG. 3. *Convergence rates of the three methods, for various V-shaped distributions. The rate is the value of r for which total variation distance goes down with t in proportion to e^{-rt}, asymptotically. The last column is the minimum probability in the distribution (at the bottom of the V).*

PRL **105**, 120603 (2010) PHYSICAL REVIEW LETTERS week ending 17 SEPTEMBER 2010

Markov Chain Monte Carlo Method without Detailed Balance

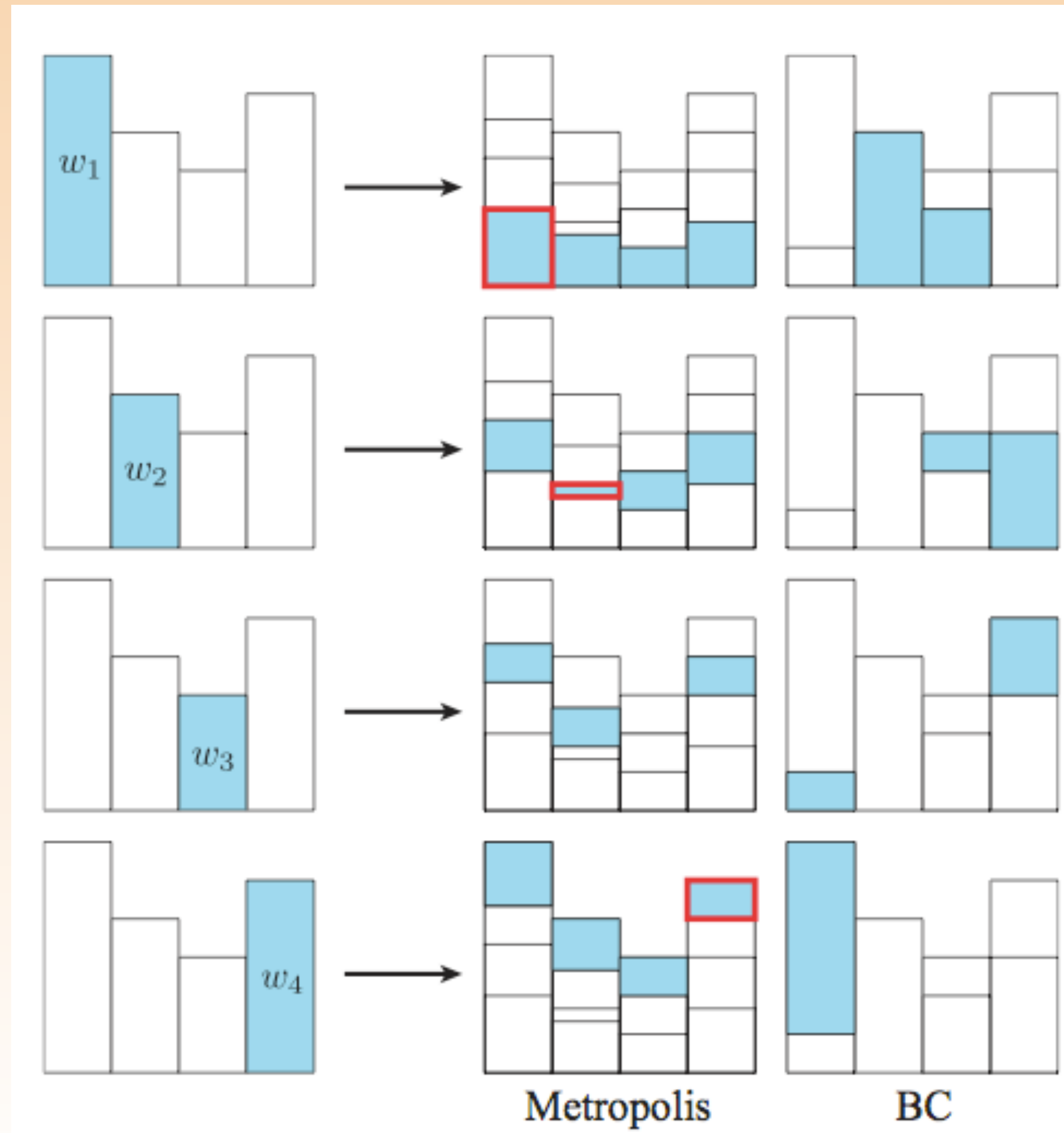
Hidemaro Suwa[1] and Synge Todo[1,2]
[1]*Department of Applied Physics, University of Tokyo, Tokyo 113-8656, Japan*
[2]*CREST, Japan Science and Technology Agency, Kawaguchi 332-0012, Japan*



$$v_{i\to j} = w_i T_{i\to j}$$

BC: first the maximum weight (w1) is allocated to the second box. It saturates the second box, and the remainder is all put into the third one (first row). Next, w2 is allocated to the partially filled box and the subsequent box (second row). The same procedure is repeated for w3 and w4.

Potts, worm algorithm for quantum spins



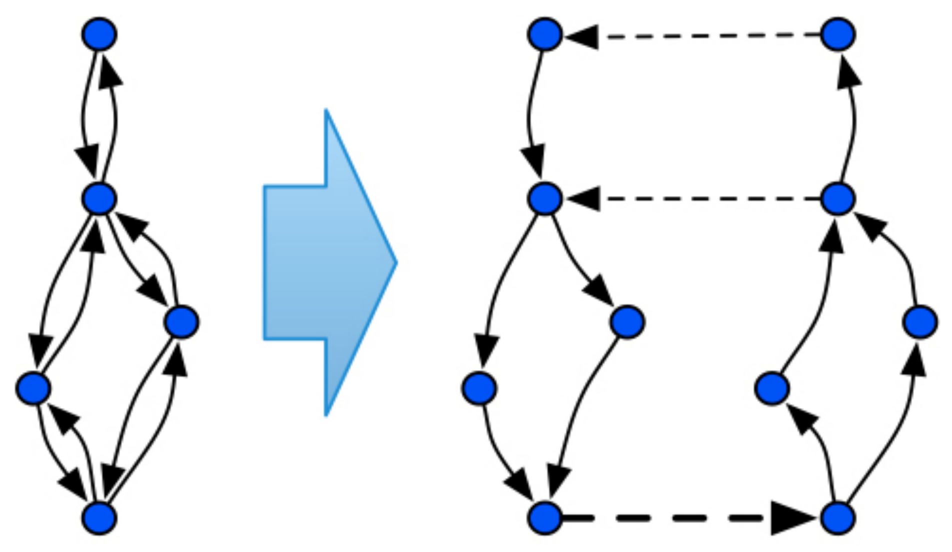
Skewed detailed balance

K. S. Turitsyn, M. Chertkov, MV (2008)

- Create two copies of the system ('+' and '-')
- Decompose transition probabilities as

$$T = T^{(+)} + T^{(-)}$$

$$\pi(x)T^{(+)}(x,y) = \pi(y)T^{(-)}(y,x)$$



- Compensate the compressibility by introducing transition between copies

$$\Lambda^{(\pm,\mp)}(x,x) = \max\left\{\sum_{y\in\Omega}\left(T^{(\mp)}(x,y) - T^{(\pm)}(x,y)\right), 0\right\}$$

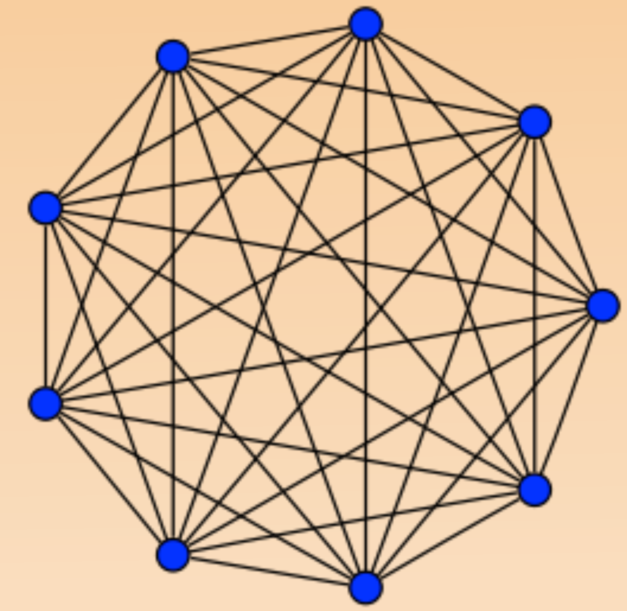
Skewed detailed balance continued

- Extended matrix satisfies balance condition and corresponds to irreversible process:

$$\mathcal{T} = \begin{pmatrix} T^{(+)} & \Lambda^{(+,-)} \\ \Lambda^{(-,+)} & T^{(-)} \end{pmatrix}$$

- Random walk becomes non-Markovian in the original space.
- System copy index is analogous to momentum in physics: diffusive motion turns into ballistic/ super-diffusive.
- No complexity overhead for Glauber and other dynamics.

Curie-Weiss Ising model



N-spins ferromagnetic cluster

Ising model on a complete graph

Stationary distribution

$$J > 0 \qquad \pi_{s_1,\ldots,s_N} = Z^{-1} \exp\left[-\frac{J}{N}\sum_{k,k'} s_k s_{k'}\right]$$

A state of the system is completely characterized by its global spin (magnetization)

$$S = \sum_k s_k$$

probability distribution of global spin

$$P(S) \sim \frac{N!}{N_+! N_-!} \exp\left(-\frac{JS^2}{2N}\right)$$

$$N_\pm = \frac{N \pm S}{2}$$

Physics of the spin-cluster continued

In the thermodynamic limit $N \to \infty$

the system undergoes a phase transition at $J = 1$

Away from the transition in the paramagnetic phase $J < 1$

$P(S)$ is centered around $S = 0$

and the width of the distribution is estimated by $\delta S \sim \sqrt{N/J}$

At the critical point (J=1) the width is $\delta S \sim N^{3/4}$

One important consequence of the distribution broadening is a slowdown observed at the critical point for reversible MH–Glauber sampling.

Correlation time of S reversible case

characteristic correlation time of S (measured in the number of Markov chain steps) is estimated as

$$T_{rev} \propto (\delta S)^2$$

the computational overhead associated with the critical slowdown is

$$\sim \sqrt{N}$$

Advantage of using irreversibility

The irreversible modification of the MH–Glauber algorithm applied to the spin cluster problem achieves complete removal of the critical slowdown.

Correlation time of S irreversible case

switching from one replica to another the system always go through the S = 0 state, since

$$\Lambda_{ii}^{(+,-)} = 0 \quad \text{if} \quad S > 0$$ (+) to (-) switching + spins in (+) replica

$$\Lambda_{ii}^{(-,+)} = 0 \quad \text{if} \quad S < 0$$ (-) to (+) switching - spins in (-) replica

The Markovian nature of the algorithm implies that all the trajectories connecting two consequent S = 0 swipes are statistically independent, therefore the correlation time roughly the number of steps in each of these trajectories.

Recalling that inside a replica (i.e. in between two consecutive swipes) dynamics of S is strictly monotonous, one estimates

$$T_{irr} \sim \delta S$$

$$T_{irr} \sim \sqrt{T_{rev}} \ll T_{rev}$$

Numerical verification

Analyzed decay of the pair correlation function, <S(0)S(t)>, with time.

Correlation time was reconstructed by fitting the large time asymptotics with exponential function

$$T \sim \exp(-t/T_{rev})$$
$$T \sim \exp(-t/T_{irr})\cos(\omega t - \phi)$$

for both MH and IMH algorithms we constructed transition matrix corresponding to the random walk in S, calculated spectral gap, Δ, related to the correlation time as,

$$T = 1/Re\Delta$$

In both tests we analyzed critical point J = 1 and used different values of N ranging from 16 to 4096.

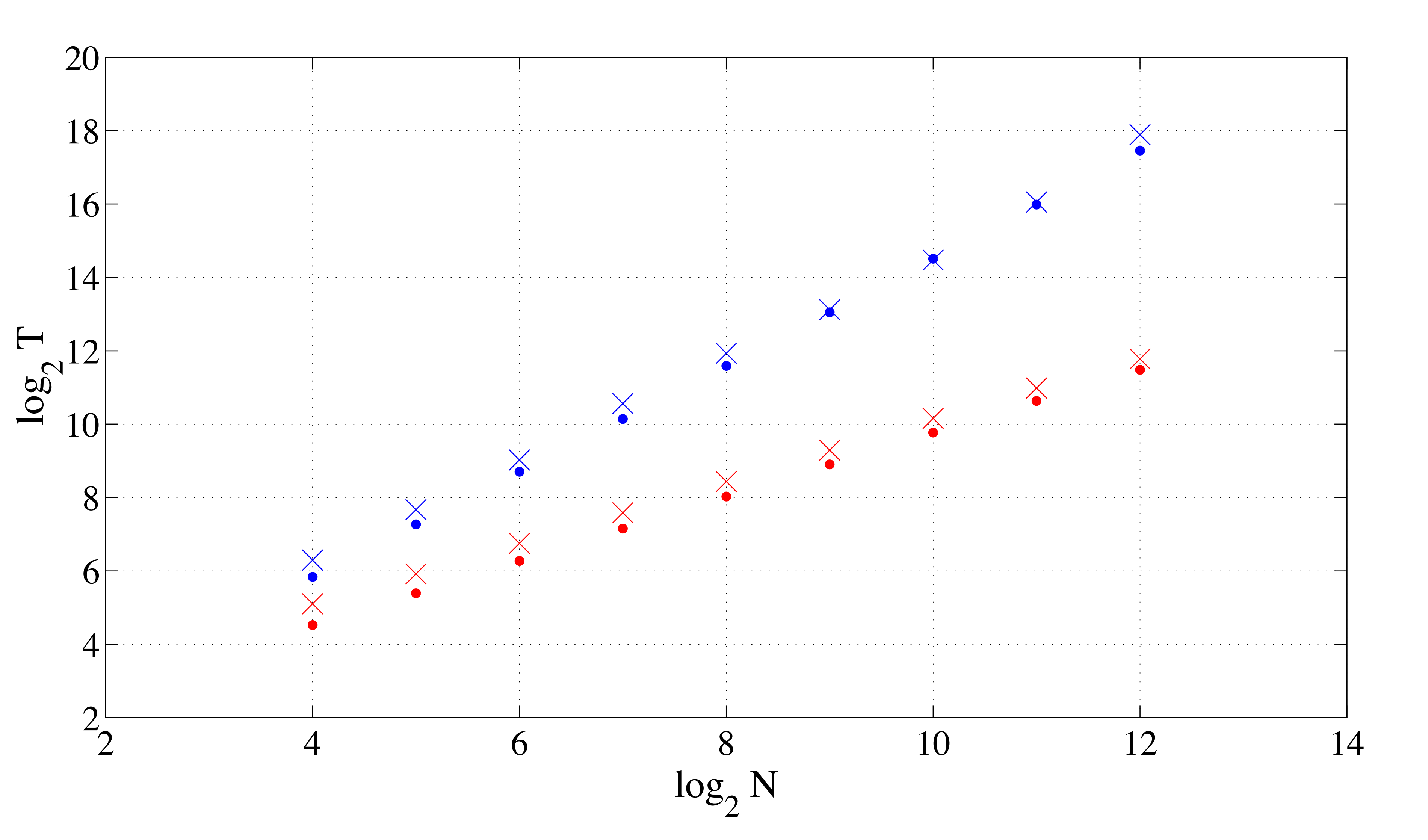
Correlation time of $\langle S(0)S(t)\rangle$ *(dots)*
Inverse spectral gap *(crosses)*

Reversible

$T \sim N^{1.43}$

Irreversible

$T \sim N^{0.85}$



A square root improvement: $T \sim N^{3/2} \rightarrow T \sim N^{3/4}$

Best case scenario: square root improvement *Chen, Lovasz, Pak etc.*

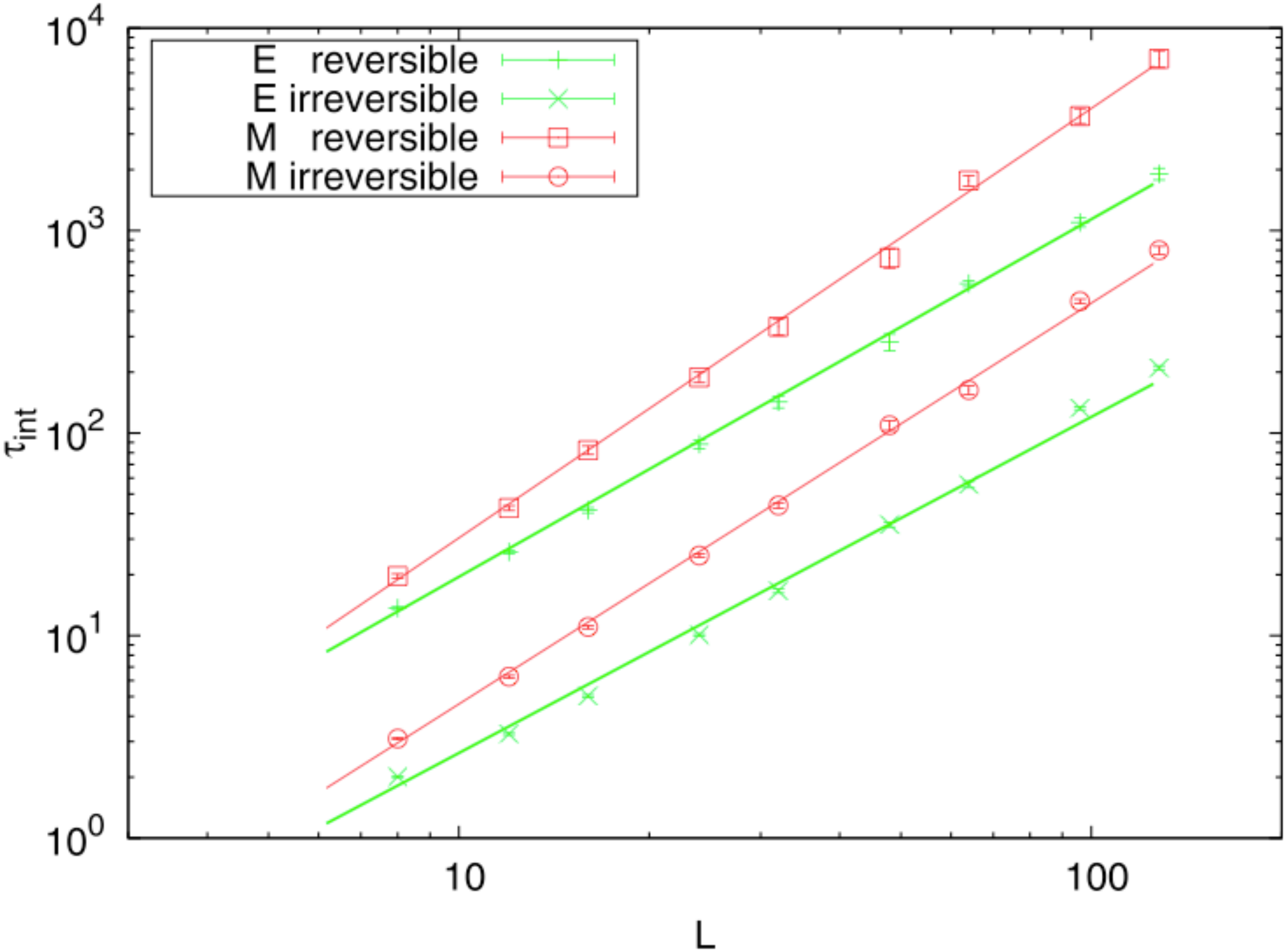
H.C.M. Fernandes, M. Weigel / Computer Physics Communications 182 (2011) 1856–1859

(a) reversibly update $(E, y) \mapsto (E + y|\Delta E|, -y)$ with the Metropolis acceptance probability,

$$p_{\text{acc}} = \min\left[1, \frac{N_{\Delta E = y|\Delta E|}}{N'_{\Delta E = -y|\Delta E|}} e^{-\beta \Delta E}\right], \tag{10}$$

(b) unconditionally negate $y \mapsto -y$,

(c) with probability θ, randomly choose a new step size $|\Delta E|$.

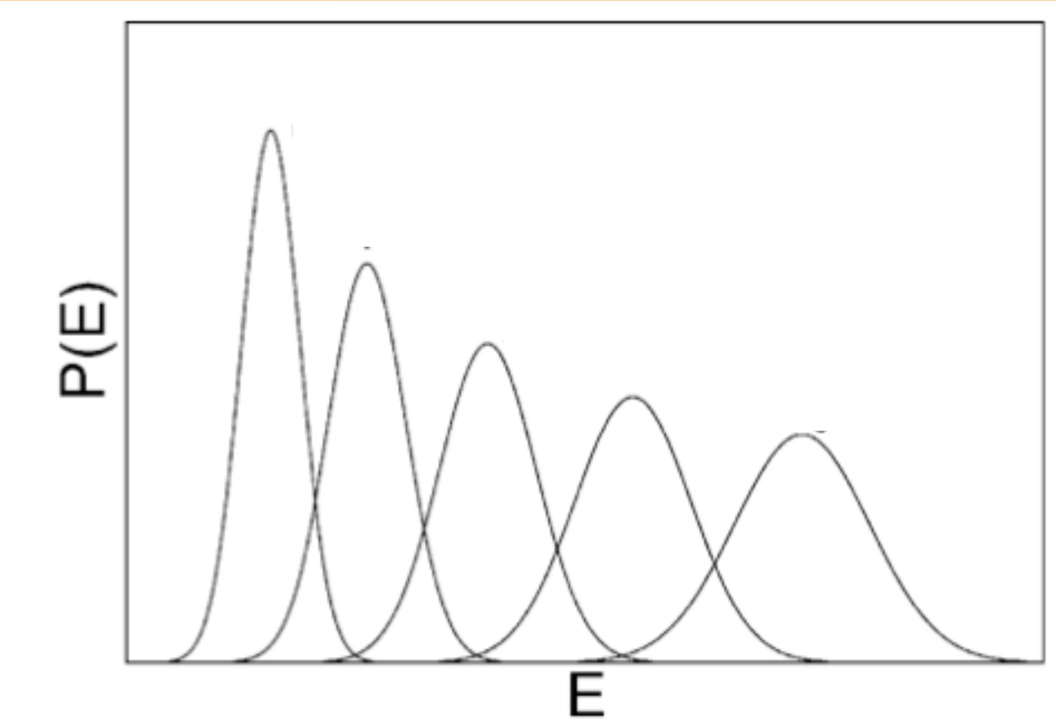


2d Ising

Parallel tempering (Replica Exchange Monte Carlo)

- Independently introduced several groups in order to study spin glasses: *Swendsen and Wang, Geyer, Hukushima and Nemoto, Parisi ...*
- Replica exchange MC (Monte Carlo) is an important tool in many areas of computational physics where the free energy landscape has many metastable minima separated by barriers, such as:
    - spin glasses
    - protein folding
    - lattice gauge theory
    - ...

R replicas of the system at different inverse temperatures $\beta_0 > \ldots > \beta_{R-1}$



$\beta_0 > \ldots > \beta_{R-1}$

Many replicas of the system are simulated in parallel using a standard MC technique for sampling the Gibbs distribution (*such as Metropolis-Hastings algorithm*). The replicas have different temperatures: starting from low T where equilibration takes a long time to high T where the equilibration is rapid.

probability of accepting replica exchange move (temperature swap) between (E, β) and (E', β')

$$p_{\text{swap}} = \min\left[1, \exp(\beta - \beta')(E - E')\right]$$

Optimizing replica exchange MC

- by choosing the set of replica temperatures
- or choosing other parameters to minimize the round-trip time.

Katzgraber et al (2006), Trebst et al (2006), Bittner et al (2008), Ballard and Jarzinsky (2009, 2012)

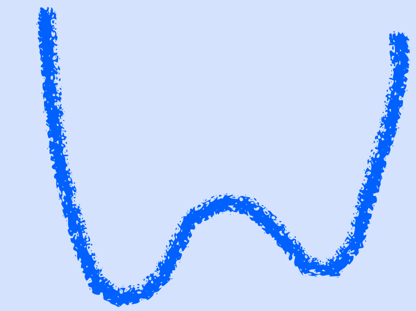
Replica exchange MC is closely related to simulated annealing and various ensemble methods.

PHYSICAL REVIEW E **80**, 056706 (2009)

Strengths and weaknesses of parallel tempering

J. Machta*
Physics Department, University of Massachusetts, Amherst, Massachusetts 01003, USA

Double-well potential work with Jon Machta



We discuss efficiency of replica exchange MC in the context of free energy landscape with two minima separated by a barrier, such as occurs in the ϕ^4 theory. Free energy F

$$\beta F_\sigma(\beta) = -\frac{1}{2}(\beta - \beta_c)^2(K + H\sigma)$$

well parameter $$\sigma = \begin{cases} 0 & \text{shallow well} \\ 1 & \text{deep well} \end{cases}$$

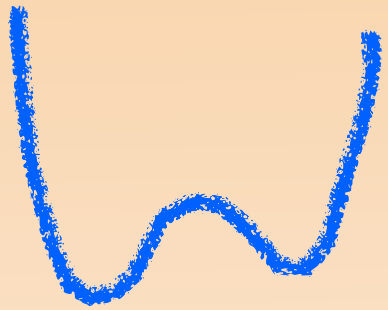
We assume that the free energy at the saddle point between the wells is 0 (so that F is the free-barrier between for transitions between the wells).

Internal energy $$U_\sigma(\beta) = -(\beta - \beta_c)(K + H\sigma)$$

Variance of the energy $$\Delta_\sigma^2 = (K + H\sigma)$$

Double-well potential

The free-energy difference $\beta\delta F(\beta)$ between the wells is controlled by H $(H \geq 0)$ and is given by



$$\frac{1}{2}(\beta - \beta_c)^2 H$$

The probability $c(\beta)$ of being in the deep well at inverse temperature β is

$$c(\beta) = \mathbb{E}(\sigma) = -\frac{1}{1 + e^{-\beta\delta F(\beta)}}$$

We assume that the energy distribution in each well is a normal distribution with mean $U_\sigma(\beta)$ and variance Δ_σ^2

GOAL: To understand the time scale for reaching the equilibrium well distribution.

- **Assumption:** Each replica is equipped with single temperature dynamics that is much faster than the rate of replica exchange attempts.

- Time scale for transitions between wells by single temperature dynamics for $\beta > \beta_c$ is $\propto \exp(-\beta F)$. **Simplification:** In analysis and simulations we do not permit well changes except at the highest temperature β_c

- At $\beta = \beta_c$ there is no barrier between the wells (they are equally likely).

- The described replica exchange dynamics satisfies detailed balance. The normal distributions of E are kept by fiat and $c(\beta)$ is obtained from replica exchange.

Average rate of replica exchange

$$\mathcal{W}_{\sigma,\sigma'}(\beta,\beta') = \mathbb{E}\left(\min\left[1, e^{(\beta-\beta')(E-E')}\right]\right)$$

$\mathbb{E}(\cdot)$ average over energies E, E'

Degenerate wells $H = 0$

(recall: $\beta F_\sigma(\beta) = -\frac{1}{2}(\beta-\beta_c)^2(K + H\sigma)$)

$$\mathcal{W}_{\sigma,\sigma'}(\beta,\beta') = \mathrm{erfc}\left(\frac{(\beta-\beta')\sqrt{K}}{2}\right)$$

Equilibration time for degenerate wells

Suppose there are R equally space replicas with $\beta_0 > \ldots > \beta_c$

Equilibration requires that a replica in one well at lowest temperature β_0 diffuses to β_c where the well is randomized.

Equilibration time $\tau(R)$ for R replicas scales like the mean first passage time for a random walk between the ends of a chain of R sites with hopping rate $\mathcal{W}$ with a reflecting boundary at β_0 and absorbing boundary at β_c :

$$\tau(R) \propto (R-1)^2/\mathrm{erfc}\left(\frac{(\beta-\beta')\sqrt{K}}{2(R-1)}\right)$$

Equilibration time - degenerate wells

$$\tau(R) \propto (R-1)^2 / \mathrm{erfc}\left(\frac{(\beta-\beta')\sqrt{K}}{2(R-1)}\right)$$

Optimal number of replicas $\quad R_{\rm opt} \propto (\beta_0 - \beta_c)\sqrt{K}$

$$\tau^D \propto (R_{\rm opt} - 1)^2$$

Asymmetric (non-degenerate) wells $\quad H > 0$

The wells are asymmetric and the motion is a biased diffusion. Replicas in deep wells move towards lower temperatures and replicas in shallow wells move towards higher temperatures.

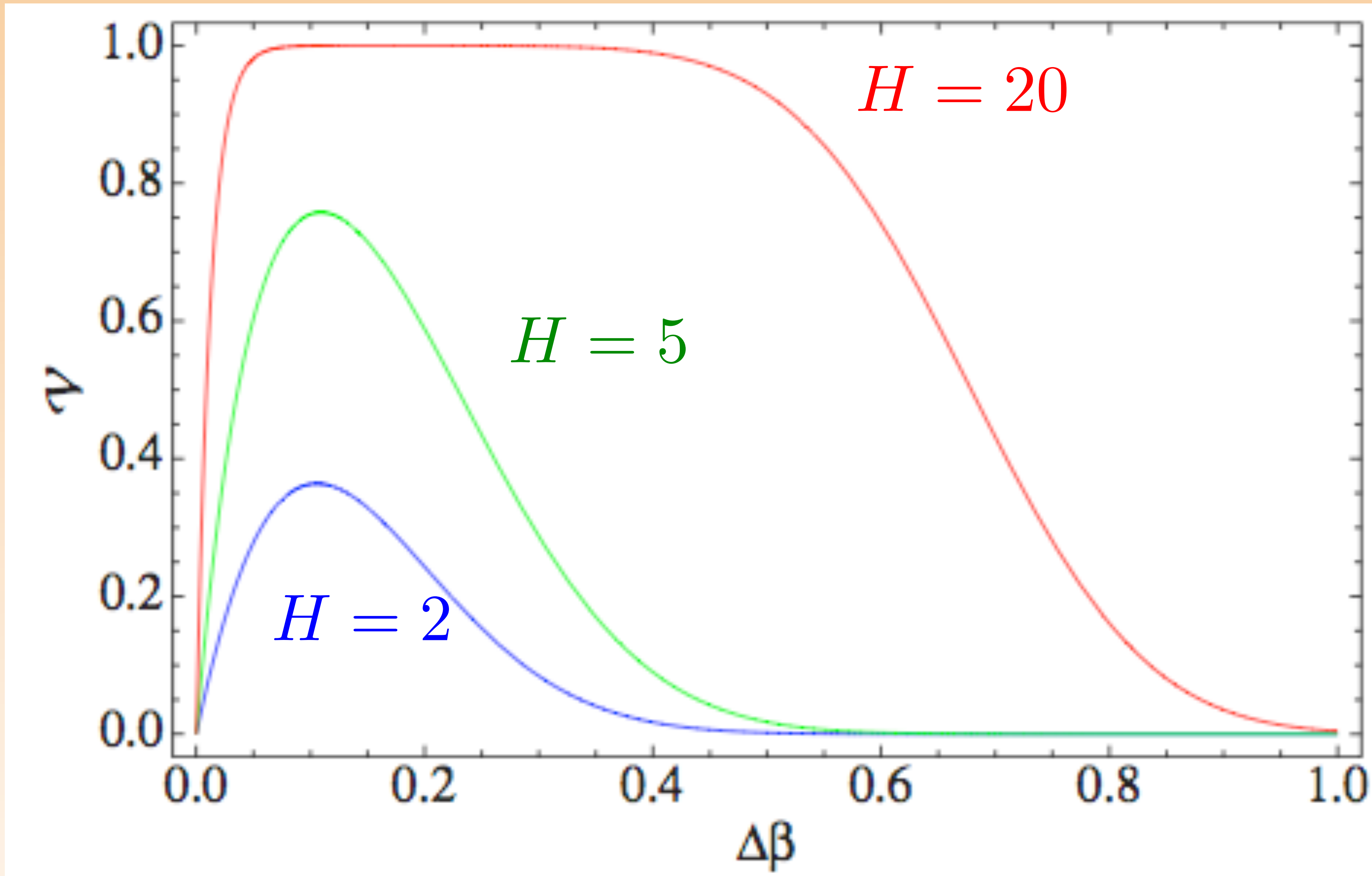
$\mathcal{W}_{0,1}(\beta,\beta') > \mathcal{W}_{1,0}(\beta,\beta') \qquad \beta \geq \beta' \qquad$ (0 shallow, 1 deep)

Asymmetric (non-degenerate) wells $H > 0$

$K = 16$

$\beta_0 = 5$

$\beta_c = 1$



$\mathcal{V}(\beta, \beta') = \mathcal{W}_{0,1} - \mathcal{W}_{1,0}$ velocity of deep (shallow)well replicas toward lower (higher) temperatures

$\mathcal{V} = 0$ diffusion, $\mathcal{V} = 1$ ballistic motion in temperature space

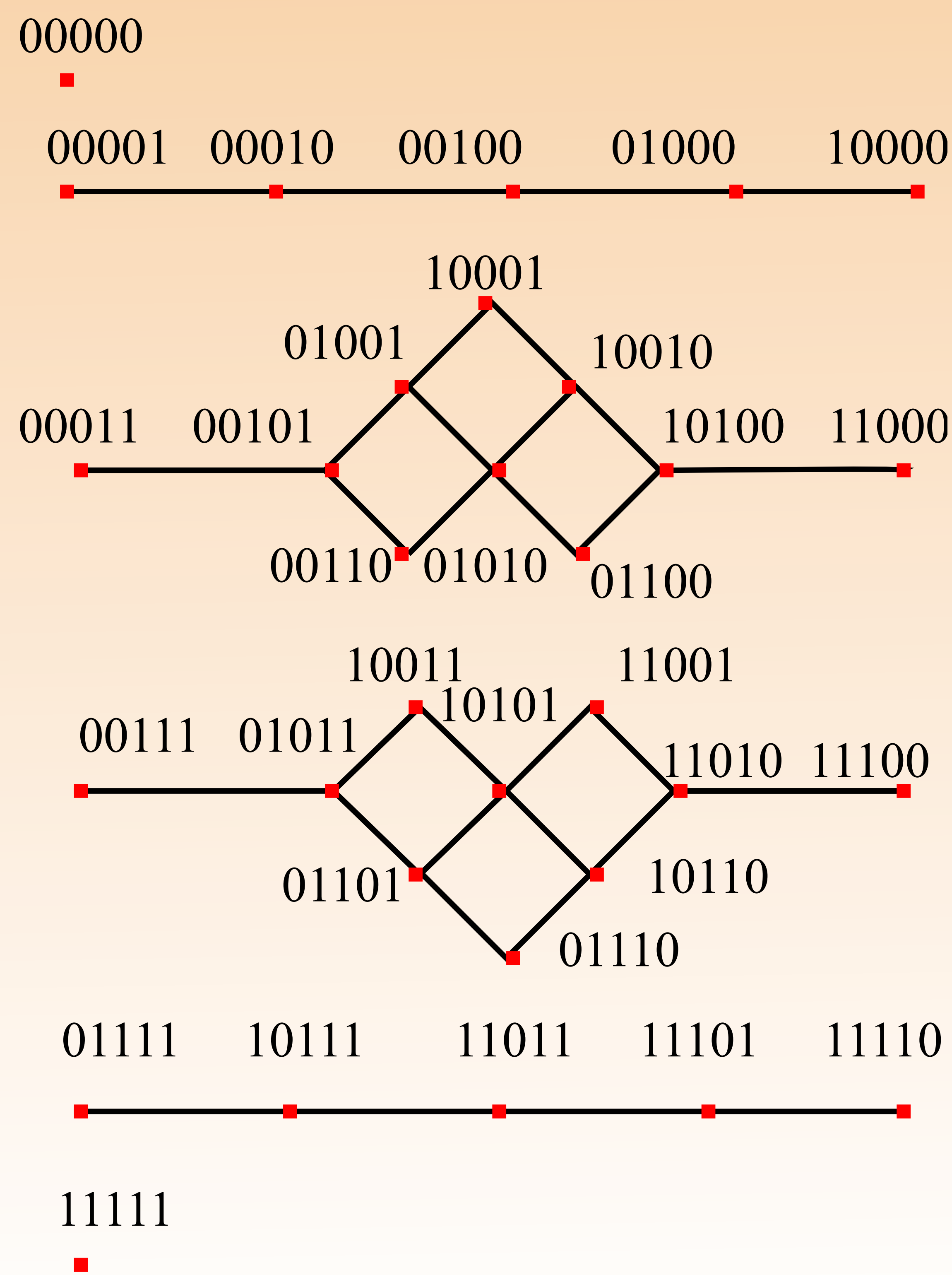
Parallel tempering is slow in the case of many degenerate minima (often the case in spin glasses)

We need somehow to bias the diffusion in temperature space

- How about breaking Detailed balance?

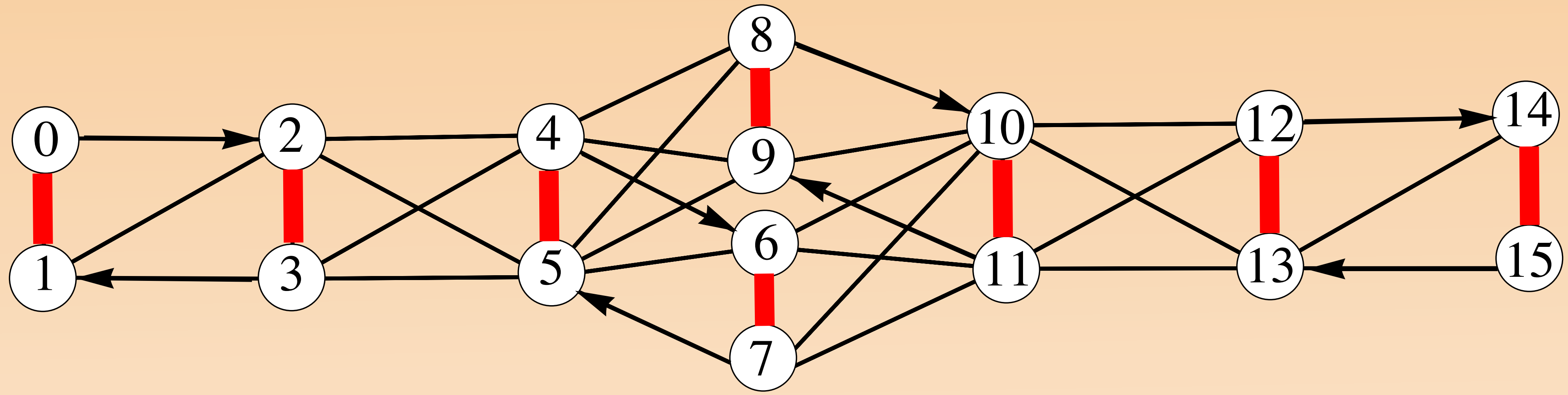
Parallel tempering: Double-well potential

Binary notation $R = 5$ case
00000 all replicas in state 0
00001 last replica in state 1
replicas ordered from left to right $\beta_0 > \ldots > \beta_c$



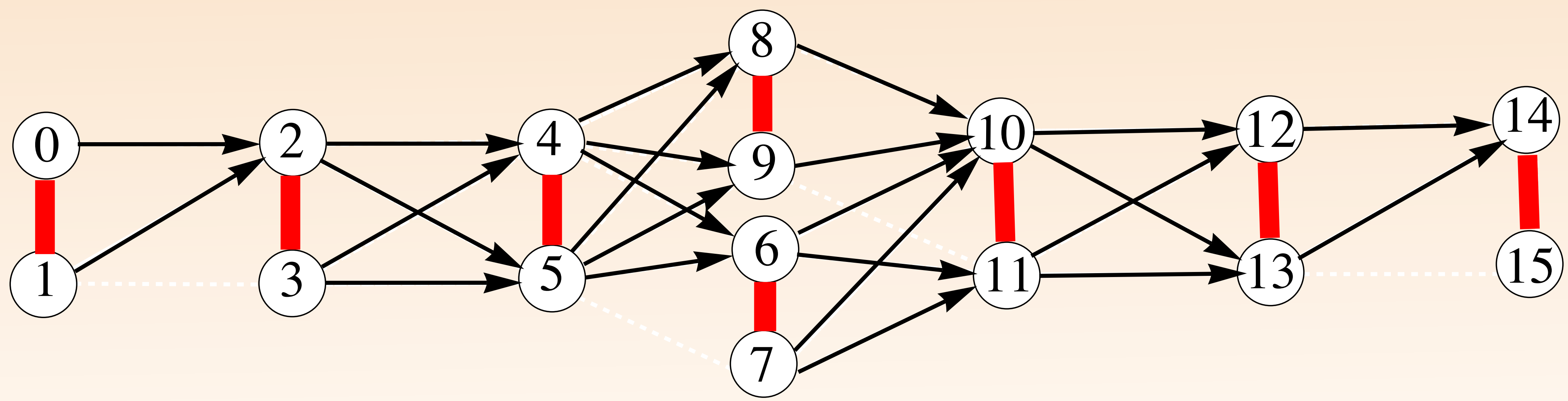
There are two types of moves: $T = P\,A$
P - randomizes the well at highest temperature replica
A - replica exchange of states at neighboring temperatures

With detailed balance



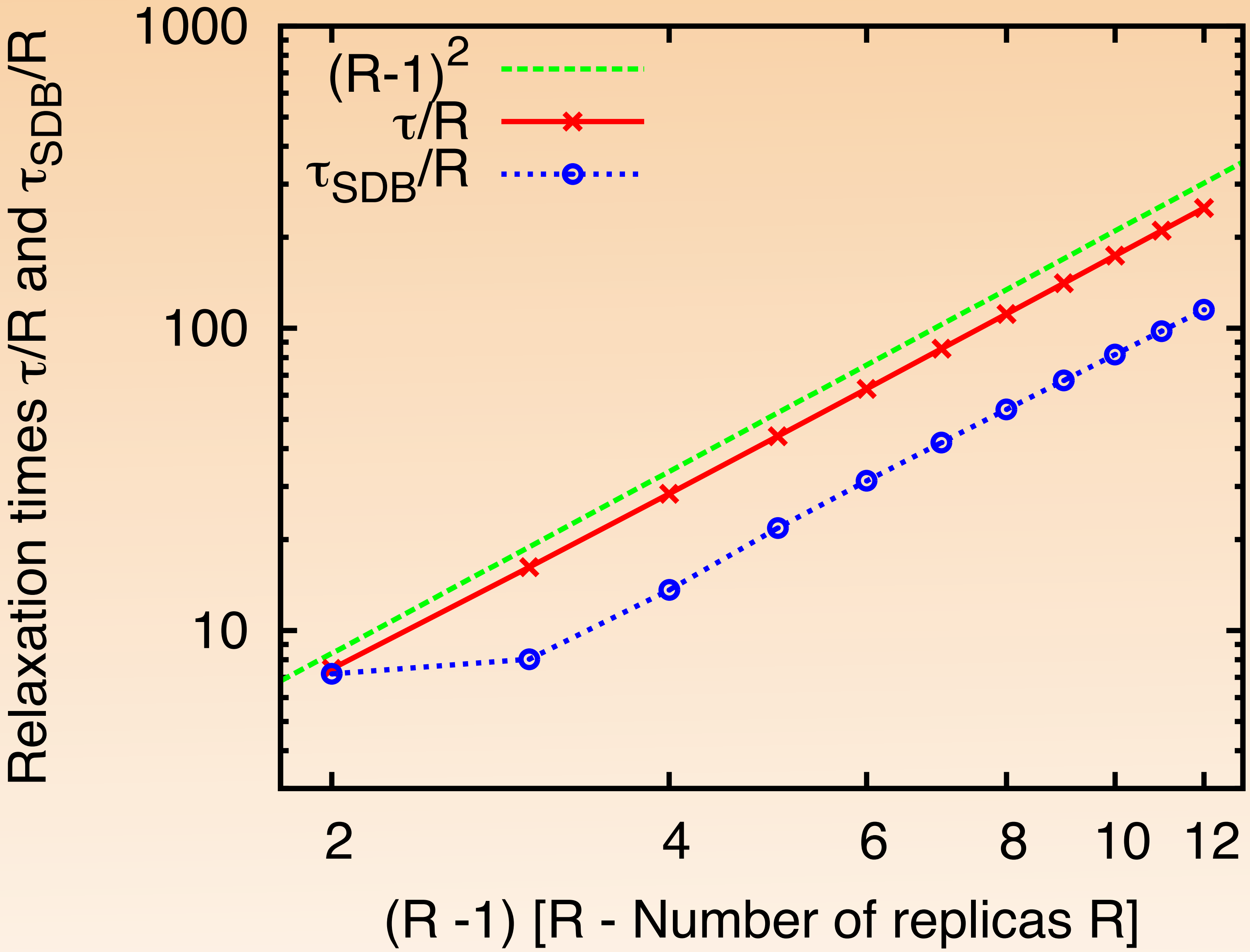
Without detailed balance

$T^{(+)}$ (one of the 2 copies of the system)



$$\mathcal{T} = \begin{pmatrix} T^{(+)} & \Lambda^{(+,-)} \\ \Lambda^{(-,+)} & T^{(-)} \end{pmatrix}$$

Twice faster (?)



R	$(\beta_0 - \beta_c)\sqrt{K}$	$\tau_{\rm SDB}/\tau$
3	3.367	0.963
4	5.05051	0.497
5	6.73401	0.480
6	8.41751	0.496
7	10.101	0.495
8	11.7845	0.490
9	13.468	0.483
10	15.1515	0.477
11	16.835	0.471
12	18.5185	0.465
13	20.202	0.460
14	21.8855	

$\sqrt{\text{well depth}}$

ratio of relaxation times

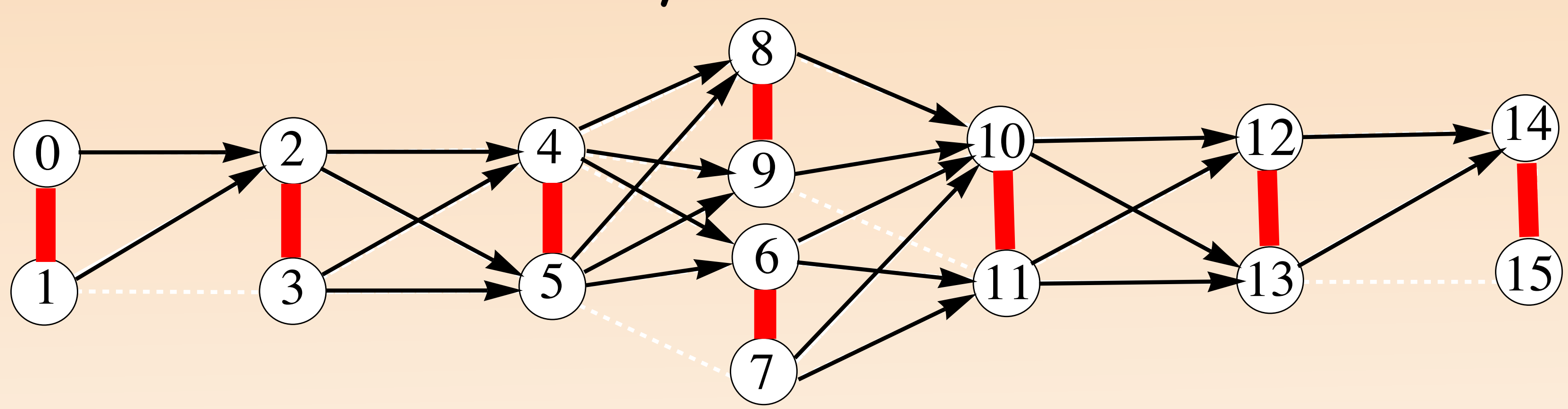
Relaxation times in the case with detailed balance τ and without detailed balance $\tau_{\rm SDB}$

Fluctuations matter!

(they make the graph bellow directionaless)

The simulations with energy fluctuations show

$$\tau_{SDB}/\tau = 0.7 \div 0.8$$



New examples one dimensional spin glass... Similar results.

Summary

- adaptive algorithms
- more convergence theorems for irreversible MC chains
- lifting for 1D chains with energy barriers