Lifting – A Nonreversible Markov Chain Monte Carlo Algorithm

Marija Vucelja*
Center for Studies in Physics and Biology, The Rockefeller University,
1230 York Avenue, New York, NY, 10065, USA
(Dated: February 3, 2015)

Markov Chain Monte Carlo algorithms are invaluable numerical tools for exploring stationary properties of physical systems – in particular when direct sampling is not feasible. They are widely used in many areas of physics and other sciences. Most common implementations are done with reversible Markov chains – Markov chains that obey detailed balance. Reversible Markov chains are sufficient in order for the physical system to relax to equilibrium, but it is not necessary. Here we review several works that use "lifted" or nonreversible Markov chains, which violate detailed balance, yet still converge to the correct stationary distribution (they obey the global balance condition). In certain cases, the acceleration is a square root improvement at most, to the conventional reversible Markov chains. We introduce the problem in a way that makes it accessible to non-specialists. We illustrate the method on several representative examples (sampling on a ring, sampling on a torus, an Ising model on a complete graph and 1d Ising model). We also provide a pseudocode implementation, review related works and discuss the applicability of such methods.

MONTE CARLO – AN INVALUABLE METHOD

One resorts to Monte Carlo algorithms when faster-converging numerical approaches are not applicable; such is usually the case in statistical physics and quantum field theory – especially when estimating high-dimensional integrals with sums. A Riemann sum of $n$-terms evaluated, for example, by the Simpsons rule deviates from the exact value of the integral by an amount proportional to $n^{-3/4}$. In comparison to that, for most Monte Carlo methods, the error scales as $n^{-1/2}$ (square root of the computational budget) and is independent on the dimensionality of the integral. Thus, already for $d > 8$ Monte Carlo outperforms the Simpsons rule. This point is stressed in Sokal’s lecture notes [1] and is accompanied by cautionary words: "Monte Carlo is an extremely bad method; it should be used only when all alternative methods are worse." Such is precisely the case in statistical physics, where Monte Carlo algorithms are often irreplaceable.

In a dynamical Monte Carlo method, one defines a stochastic process on the configuration space of the considered physical system, such that as time goes to infinity the system relaxes to its equilibrium. A typical stochastic process used is a Markov chain. In this review, we will focus on dynamical Monte Carlo methods utilizing Markov chains, the so-called Markov chain Monte Carlo (MCMC) methods. At sufficiently large times, when the system is close enough to equilibrium – the “numerical experiment” can begin. At that point we have means of generating samples from the equilibrium distribution, measuring observables, correlations between observables, estimating the partition function, etc. This stochastic time evolution is only an auxiliary fictitious dynamics, and a lot of work is being done in order to speed up its convergence [2–8]. The focus of our article is to show how to modify the stochastic time evolution in order to accelerate the convergence to equilibrium. Our intuition stems from hydrodynamics and the mixing in non-equilibrium physics. Such analogies will be made below, when we introduce our nonreversible MCMC [2]. In mathematics and computer science literature our kind of nonreversible Markov chains are called "lifted" Markov chains, and we will adopt this jargon as well [7, 9].

In the next section we will introduce some mathematical definitions and notation. Afterward we will describe a very famous variant of the MCMC algorithm family – the Metropolis-Hastings algorithm, followed by measures of relaxation to equilibrium. Finally we will introduce the lifted (nonreversible) MCMC, illustrate it on pertinent examples and conclude with a discussion.

MATHEMATICAL PREREQUISITS AND NOTATION

Suppose we have a physical system that we initially prepared in a state $x$ from the finite set of possible configurations $\Omega$. A Markov chain is a sequence of states where the probability of the next state is determined by the previous state. It is specified by a transition matrix $P$ between the states of the physical system and the initial probability distribution over the states of the system $\rho(t = 0)$. In a Markov chain the conditional probability of switching from state $x$ to state $y$ at the next time step, $P(x, y) = \text{Prob}[x \rightarrow y]$, is independent of the path that preceded to getting to state $x$. Precisely this "independence from the past" is the reason that $P$ a matrix of size $|\Omega| \times |\Omega|$ and a row-vector $\rho(t = 0)$ of length $|\Omega|$ are sufficient to specify evolution of this realization.

The $x$–th row of the transition matrix $P$ is a distribution itself. As a corollary of the conservation of proba-
bility $P$ has non-negative elements and is stochastic, i.e.:

$$\sum_{y \in \Omega} P(x, y) = 1,$$  \hspace{1cm} (1)

holds for all $x \in \Omega$. The probability distribution $\rho(t)$ evolves according to

$$\rho(t; x) = \sum_{y \in \Omega} \rho(t-1; y) P(y, x) = \ldots = \sum_{y \in \Omega} \rho(0; y) P^t(y, x),$$  \hspace{1cm} (2)
or in vectorially

$$\rho(t) = \rho(t-1) P = \ldots = \rho(0) P^t,$$  \hspace{1cm} (3)

where the time $t \geq 0$ is assumed to be discrete and $P^t$ is the matrix $P$ raised to the power $t$. Note that we are multiplying a matrix by a vector on the left. One can also define continuous time Markov chains, but this is beyond the scope of our review. A distribution is called stationary if it does not change with time, i.e.

$$\pi = \pi P.$$  \hspace{1cm} (4)

A transition matrix $P$ is called irreducible if a path can be found between any two states on $\Omega$, that is for any two states $x, y \in \Omega$ there exists at integer $t$ such that $P^t(x, y) > 0$. Irreducibility of $P$ means, that it is possible to get from any state to any other state using only transitions of positive probability; physicists usually call this property ergodicity. A Markov chain ($\rho(t = 0), P$) will converge to a stationary distribution $\pi$ if the Global Balance (GB) condition

$$\sum_{y \in \Omega} [\pi(x) P(x, y) - \pi(y) P(y, x)] = 0, \quad \forall x \in \Omega,$$  \hspace{1cm} (5)

holds and $P$ is irreducible. Furthermore, one can show that the stationary distribution $\pi$ is unique. The book of Levin, Peres, and Wilmer [10] provides a very comprehensible source of knowledge on Markov chains, stochastic processes, and mixing. We strongly recommend this book for the mathematically inclined novice.

The GB condition, Eq. (5), signifies that the total influx to a state is equal to the total efflux from this state. To employ a hydrodynamics analogy – GB amounts to the incompressibility of the phase space. A special case of Eq. (5) is the pairwise cancelation of the terms in the sum – the so-called Detailed Balance (DB) condition:

$$\pi(x) P(x, y) = \pi(y) P(y, x), \quad \forall x, y \in \Omega.$$  \hspace{1cm} (6)

Note that as opposed to GB, DB is a local, microscopic reversibility property. A hydrodynamic analogy would be that of an irrotational flow. DB is a special case of GB, likewise as all irrotational flows are incompressible. Markov chains obeying DB are called reversible Markov chains. Usually, it is much easier to implement and numerically enforce DB since it is a local condition. Below we will show several examples of MCMC with GB, which are faster than their conventional counterparts with DB. Prior to that, we will show how to implement MCMC with DB.

**METROPOLIS-HASTINGS (MH) ALGORITHM**

One of the most famous MCMC algorithms is due to Metropolis et al. [11]. It is called the Metropolis-Hastings (MH) algorithm, since it has been later generalized by Hastings [12]. We will use the MH algorithm in all of our examples that follow, so let us first explain how it works. We start off with an arbitrary initial distribution $\rho(t = 0)$ and a transition matrix $Q$. We want a Markov chain that will converge to a specified equilibrium distribution $\pi$. This can be achieved by introducing the "acceptance probabilities" $a(x, y)$ in such a way that the new transition matrix $P$, has off-diagonal elements ($x \neq y$)

$$P(x, y) = a(x, y) Q(x, y),$$  \hspace{1cm} (7)

while the diagonal elements are set by conservation of total probability: $P(x, x) = 1 - \sum_{y \in \Omega \setminus x} P(x, y)$. As we have seen in the previous section, sufficient condition for an arbitrary initial distribution $\rho(t = 0)$ to converge at large times to the equilibrium distribution $\pi$ is that $P$ obeys DB, Eq. (6). For this to be fulfilled, the acceptance probability should satisfy

$$a(x, y) = \frac{\min \left[ \frac{\pi(x) Q(x, y)}{\pi(y) Q(y, x)} \right]}{\pi(y) Q(y, x) + \pi(x) Q(x, y)},$$  \hspace{1cm} (8)

Solutions of Eq. (8) are the heat bath acceptance probability

$$a(x, y) = \frac{\pi(y) Q(y, x)}{\pi(y) Q(y, x) + \pi(x) Q(x, y)},$$  \hspace{1cm} (9)

and the Metropolis-Hastings (MH) acceptance probability

$$a(x, y) = \min \left[ \frac{\pi(y) Q(y, x)}{\pi(x) Q(x, y)} \right].$$  \hspace{1cm} (10)

Both solutions are widely used. In the following sections we will use the Metropolis-Hastings (MH) acceptance probability.

Now that we have defined a correct MCMC process, in that it converges to the target equilibrium distribution, we also need to know how fast it converges and this is the topic of the next section.

**CONVERGENCE MEASURES**

Eventually once equilibrium is achieved we will be interested in the average values and correlations between observables. In this section we will describe various measure of convergence following the lecture notes of Sokal [1]. Suppose $f$ is an observable, it is a function of the possible system states – for example in a magnetic system, magnetization is a function of the spin configuration. We define a Markov process with a transition matrix $P$ and
we start the system with the initial distribution \( \rho(t = 0) \). The mean \( \mu_f(t) \) and variance \( \sigma^2_f(t) \) of the observable \( f \) are time dependent and equal to

\[
\mu_f(t) \equiv \langle f(t) \rangle \equiv \sum_{x \in \Omega} \rho(t; x) f(x),
\]

\[
\sigma^2_f(t) \equiv \langle f^2(t) \rangle - \langle f(t) \rangle^2 = \sum_{x \in \Omega} \rho(t; x) [f(x) - \mu_f(t)]^2.
\]

For the types of Markov processes that we consider here, as the time goes to infinity, the \( \rho(t; x) \) converges to the equilibrium distribution \( \pi \) and the average properties become stationary (time-independent)

\[
\mu_{f,\pi} \equiv \langle f \rangle \equiv \sum_{x \in \Omega} \pi(x) f(x),
\]

\[
\sigma^2_{f,\pi} \equiv \langle f^2 \rangle - \langle f \rangle^2 \equiv \sum_{x \in \Omega} \pi(x) [f(x) - \mu_{f,\pi}]^2.
\]

Above we omitted the dependence on time to stress that these averages are time-independent and added a subscript \( \pi \) to also emphasize that these are equilibrium averages. A good measure of how close the system dynamics is to equilibrium is an autocorrelation function, that describes the correlations between the stochastic observable at different times. An autocorrelation function for the observable \( f \), \( R_{ff}(t_1, t_2) \), is defined as

\[
R_{ff}(t_1, t_2) = \frac{(\langle f(t_1) \rangle - \mu_{f,\pi}) (\langle f(t_2) \rangle - \mu_{f,\pi})}{\sigma^2_{f,\pi}}.
\]

For a second-order stationary stochastic process (a process where the first and the second moment do not vary with respect to time) the autocorrelation function only depends on the time-lag between the observables \( t = t_2 - t_1 \) but not on their position in time. In this case above formula simplifies to

\[
R_{ff}(t) = \frac{\langle f(0) f(t) \rangle - \langle f(0) \rangle^2}{\langle f(0)^2 \rangle - \langle f(0) \rangle^2} = \frac{\langle f(0) f(t) \rangle - \mu^2_{f,\pi}}{\sigma^2_{f,\pi}},
\]

where \( \langle f(0) f(t) \rangle \equiv \sum_{x,y \in \Omega} f(x) \pi(x) P^t(x, y) f(y) \).

As measures of convergence to equilibrium it is customary to define the exponential autocorrelation, the integrated autocorrelation time, and the inverse spectral gap. The exponential autocorrelation time of the observable \( f \), \( \tau_{exp,f} \), is

\[
\tau_{exp,f} = \lim_{t \to \infty} \sup_{t \to \infty} \frac{t}{-\ln |R_{ff}(t)|},
\]

that is the eventual least upper bound of \( t/(-\ln |R_{ff}(t)|) \) as \( t \to \infty \). We define exponential autocorrelation time \( \tau_{exp} \) to be the relaxation of the slowest observable of the system

\[
\tau_{exp} = \sup_f \tau_{exp,f}.
\]

For infinite systems \( \tau_{exp} \) may be infinite. In a nutshell, \( \tau_{exp} \) places an upper bound on the number of iterations that should be discarded at a beginning of a run, before the system relaxed to the equilibrium for all practical purposes.

The inverse spectral gap is also a frequently used measure of convergence. The transition probability matrix has the following spectral properties. The eigenvalues of the matrix \( P \) lie in a unit disk: \( 1 = \lambda_1 \leq \lambda_2 \leq ... \leq -1 \), where \( \lambda_1 = 1 \) is the largest eigenvalue and it is a simple eigenvalue (non-degenerate). The corresponding eigenvector \( \varphi_1 \) is a constant function with \( \varphi_1(x) = 1 \) for all \( x \in \Omega \). This follows from stochasticity, Eq. (1). The spectral gap is defined as the difference between the two largest eigenvalues, here

\[
\Delta = 1 - \lambda_2.
\]

At a finite time \( t \) we have

\[
\frac{P^t(x, y)}{\pi(y)} = 1 + \varphi_2(x) \varphi_2(y) \lambda_2^t + O(\lambda_2^t),
\]

where \( \varphi_2 \) is the eigenvector of the second largest eigenvalue. This expression can be written as

\[
P^t(x, y) \approx \pi(y) (1 + \varphi_2(x) \varphi_2(y)(1 - \Delta)^t)
\]

\[
\approx \pi(y) (1 + \varphi_2(x) \varphi_2(y) e^{-\Delta t}),
\]

thus we see that \( e^{-\Delta t} \) is a measure of how fast the Markov chain converges to \( \pi \). Also if \( P \) obeys DB, then one can show that it has real eigenvalues. To see that we define another matrix \( A(x, y) = \pi(x)^{1/2} \pi(y)^{-1/2} P(x, y) \), which has the same eigenvalues as \( P \), and observe that \( A \) is symmetric and thus has real eigenvalues. On the other hand if \( P \) obeys GB, the eigenvalues are in general complex and it makes more sense to use the "absolute spectral gap" \( \Delta^* \equiv 1 - |\lambda_2| \) or the real part of the spectral gap \( Re(\Delta) \) as measures of convergence. In fact from the definition of \( \tau_{exp} \) it follows

\[
\tau_{exp} = \frac{1}{\Delta^*}.
\]

In this case a system relaxes to equilibrium with damped oscillations.

Another useful measure of convergence is the so-called integrated autocorrelation time \( \tau_{int} \) which is defined as

\[
\tau_{int} = \sup_f \tau_{int,f} = \sup_f \left( \frac{1}{2} + \sum_{t=1}^\infty R_{ff}(t) \right).
\]

Note that if \( R_{ff}(t) \sim e^{-t/\tau} \) and \( \tau \gg 1 \) we have

\[
\tau_{int,f} \approx \tau_{exp,f},
\]

which one can check by direct substitution. The integrated autocorrelation time controls the statistical error
in the Monte Carlo measurements of the equilibrium averages, such as \( \mu_{f, \pi} \). For further details on MCMC methods and convergence see the excellent lecture notes by Sokal [1]. Also another fantastic source on Monte Carlo algorithms and their applications to statistical physics is the book of Krauth [13].

**LIFTING**

MCMC methods that obey detailed balance use equilibrium dynamics to sample a phase space. For example when visiting a lattice with uniform steady state distribution \( \pi \), MH moves are unbiased hops to nearest neighbor lattice sites which occur with acceptance rate 1 – that is an ordinary random walk on a lattice. In this case the MH moves perform a diffusion like motion in the phase space. We use the term "diffusion" for a motion that requires \( \propto N^2 \) steps to travel distance \( N \) from its point of origin. Now, what if this is too slow? One can imagine that sometimes it would be beneficial to have some "inertia" or "momentum", when performing auxiliary fictitious Markov chain hops in the phase space. Very much like using a spoon to stir a coffee helps to spread the sugar in the whole cup faster. Another example is odor from the dinner table – if it were to diffuse it would reach us in a few hours instead of minutes. We can smell our dinner in a timely way thanks to air turbulence. Certain MCMC algorithms, like MH, are particularly slow close to phase transitions, where the dynamics suffers from critical slowing down due to large fluctuations of the observables. When sampling close to phase transitions, it would be beneficial to introduce some "inertia" and "bias".

"Lifting" is an idea that originated in computer science – we increase the phase space in order to create more biases and explore the new enlarged phase space more efficiently than we would explore the original space [6, 7]. Here we show how to implement lifting for discrete Markov processes. Lifting will alter the convergence time. It is an open question whether and when it will make it faster. The method we will introduce is potentially good for overcoming entropic, but not high energetic barriers (see the discussion).

One can create a lifting in an uncontrolled way by adding many cycles, since cycles in phase space do not change the steady state, but the practical caveat is how many and what cycles to add to the already existing transitions. In [2] we introduced a controlled way to create a nonreversible Markov Chain. Suppose that \( \pi \) is a stationary distribution: \( \pi = \pi P \), where \( P \) is a stochastic matrix (Eq. (1) holds). We define a larger space \( \tilde{\Omega} = \Omega \times \{1, -1\} \) and denote a state in this space as \( x_\xi = x \in \Omega, \xi \in \{1, -1\} \). Next we impose skew-detailed balance

\[
\tilde{\pi}(x_\xi) \tilde{P}(x_\xi, y_\xi) = \tilde{\pi}(y_{\xi-}) \tilde{P}(y_{\xi-}, x_{\xi-}), \tag{26}
\]

for

\[
\tilde{\pi} = \frac{1}{2}(\pi, \pi). \tag{27}
\]

Recall that \( \tilde{\pi} \) and \( \pi \) are vectors. We enforce that the lifted transition matrix \( \tilde{P} \) is stochastic

\[
\sum_{y_\eta \in \tilde{\Omega}} \tilde{P}(x_\xi, y_\eta) = 1, \forall x_\xi \in \tilde{\Omega}, \tag{28}
\]

by adjusting the diagonal elements \( \tilde{P}(x_\xi, x_\xi) = 1 - \sum_{y_\eta \in \tilde{\Omega} \setminus \{x_\xi\}} \tilde{P}(x_\xi, y_\eta) \). The matrix \( \tilde{P} \) has the following block structure: two diagonal blocks describe transitions inside \( \Omega \times \{1\} \) and \( \Omega \times \{-1\} \) spaces respectively, while the off-diagonal blocks are describe transitions between \( x_\xi \) and \( y_{\xi-} \) states. For simplicity, we assume that the off-diagonal blocks are diagonal matrices of the following form

\[
\tilde{P}(x_\xi, y_{\xi-}) = \delta_{x_\xi} \tilde{P}(x_\xi, y_{\xi-}). \tag{29}
\]

A distribution \( \tilde{\pi} \) satisfying the skew-detailed balance Eq. (26) is stationary with respect to \( \tilde{P} \), i.e. \( \tilde{\pi} = \tilde{\pi} \tilde{P} \). One can prove this as follows:

\[
\sum_{y_\eta \in \tilde{\Omega}} \tilde{\pi}(y_\eta) \tilde{P}(y_\eta, x_\xi) = \sum_{y_\eta \in \tilde{\Omega}} \tilde{\pi}(y_\eta) \tilde{P}(y_\eta, x_\xi)(\delta_{\xi_\eta} + 1 - \delta_{\xi_\eta}) = \sum_{y_\eta \in \tilde{\Omega}} \left( \tilde{\pi}(y_\eta) \tilde{P}(y_\eta, x_\xi) + \tilde{\pi}(y_{\xi-}) \tilde{P}(y_{\xi-}, x_\xi) \right) \tag{30}
\]

Next using skew-detailed balance Eq. (26) and Eq. (29) we have

\[
\sum_{y_\eta \in \tilde{\Omega}} \tilde{\pi}(y_\eta) \tilde{P}(y_\eta, x_\xi) = \tilde{\pi}(x_{\xi-}) \left( \sum_{y_\eta \in \tilde{\Omega}} \tilde{P}(x_{\xi-}, y_{\xi-}) + \tilde{P}(x_{\xi-}, x_\xi) \right) \tag{31}
\]

\[
= \tilde{\pi}(x_{\xi-}) \sum_{y_\eta \in \tilde{\Omega}} \tilde{P}(x_{\xi-}, y_\eta) = \tilde{\pi}(x_{\xi-}), \tag{32}
\]

where the last equality follows from Eq. (28). Finally using Eq. (27) we get

\[
\sum_{y_\eta \in \tilde{\Omega}} \tilde{\pi}(y_\eta) \tilde{P}(y_\eta, x_\xi) = \tilde{\pi}(x_\xi), \tag{33}
\]

which is the definition of stationarity and thus concludes our proof.

We should also determine the off-diagonal elements. From stochasticity and Eq. (29) we have

\[
\tilde{P}(x_\xi, x_{\xi-}) = 1 - \sum_{y_\eta \in \tilde{\Omega}} \tilde{P}(x_\xi, y_\eta). \tag{34}
\]
or
\[ \tilde{P}(x_{-\xi}, x_{\xi}) = 1 - \sum_{y_{\xi} \in \Omega} \tilde{P}(x_{-\xi}, y_{-\xi}). \] (36)

Subtracting the two equations we get
\[ \tilde{P}(x_{\xi}, x_{-\xi}) - \tilde{P}(x_{-\xi}, x_{\xi}) = \sum_{y_{\xi} \in \Omega} \left( \tilde{P}(x_{-\xi}, y_{-\xi}) - \tilde{P}(x_{\xi}, y_{\xi}) \right). \] (37)

From all possible solutions we want to choose the minimal rates, since the higher rates will impede the relaxation to equilibrium, by fostering too many transitions between the two copies of the same state \((x_{\xi} \text{ and } x_{-\xi})\). The rates are the smallest when one of the terms is 0. This leads to
\[ \tilde{P}(x_{\xi}, x_{-\xi}) = \max \left[ 0, \sum_{y_{\xi} \in \Omega} \left( \tilde{P}(x_{-\xi}, y_{-\xi}) - \tilde{P}(x_{\xi}, y_{\xi}) \right) \right]. \] (38)

Note that there is still freedom in adjusting \(\tilde{P}(x_{\xi}, y_{\xi})\) transition rates, even when the skew-detailed balance Eq. (26) is imposed. The choice ultimately has to depend on the physics of the system and measures of how much DB is violated [2, 3].

**APPLICATIONS OF LIFTING**

**Ring with a uniform stationary distribution**

As a first example let us look at a Markov chain on a ring of \(N\) states converging to a uniform distribution \(\pi(x) = N^{-1}\) for all \(x \in \{1, \ldots, N\}\). The idea is illustrated in Fig. 1. A random walker would cover every state along the ring in time that scales as diffusion timescale \(t \propto N^2\) (see Fig. 1A). Lifting here can improve the convergence to the stationary distribution \(\pi\). To apply lifting we create two rings of \(N\) states: one on which the transitions are made only in the counter clockwise direction and the other where the transitions are made only in a clockwise direction. The bias is set by choosing \(\varepsilon\): with probability \(1 - \varepsilon\) the walker continues to hop in the same direction, otherwise the walker jumps to the same state, but on the other replica of the system. This system converges to a steady state \(\pi\) after \(\propto O(N)\) steps [6, 7].

**Torus with uniform stationary distribution**

A reversible Markov chain on a torus of \(N^2\) sites with a uniform stationary distribution \(\pi(x) = N^{-2}\) (for all \(x \in \{1, \ldots, N^2\}\)) performs a random walk, whose convergence toward \(\pi\) scales as \(t \propto O(N^2)\) (see e.g. [7] or the inset on Fig. 3). Now let us give the random walker some inertia.

More exactly, we define a walk as follows: if we enter a node, we are most likely to exit over the opposite edge \((1 - \varepsilon)\), then we turn left or right with probability \(\varepsilon/2\) and we never turn back. If we start the walk in a given direction, say towards north, we are likely to continue going north until we circle the torus about a \(2N/\varepsilon\) times. Afterward, we will likely turn; the node where we turn is essentially uniformly distributed over the cycle we started on. Now we circle the torus for another approximately \(2N/\varepsilon\) times, going either east or west. When we turn again, our east-west coordinate will also be essentially uniformly distributed. In other words, the second turning point will be essentially uniformly distributed over all nodes of our graph. To get to this second turning point, we only needed is \(O(N)\) steps. This idea described in the work [7] and we simulated it here (see Fig. 2 and Fig. 3). The decay of the pair correlation function and the scaling of the real part of the inverse gap with the system size is presented on Fig. 3.

**The Ising Model on a complete graph**

Suppose we have \(N\) Ising spins on a complete graph (every pair of distinct vertices is connected by a unique edge). This system exhibits a phase transition, symmetry breaking and emergence of spontaneous magnetization, at a finite positive temperature in the limit infinite number of spins \(N \rightarrow \infty\). Let each vertex carry a spin \(s_i \in \{1, -1\}\). The energy of a spin configuration \(\mathcal{C} = \{s_1, \ldots, s_N\}\) is
\[ E(\mathcal{C}) = -\frac{J}{N} \sum_{i<j} s_i s_j, \] (39)
the energy effects will dominate and the spins will tend to align with magnetization \( M = \sum_{i=1}^{N} s_i \) (different from zero (though the average magnetization will still remain zero). As infinitesimal perturbation will determine which ground state is selected. This is an example of "symmetry breaking" in statistical mechanics. Somewhere at \( T \) positive and non-zero there will be a phase transition, where the energy and entropy are of the same order of magnitude.

A key observation in this model is that the energy depends solely on the magnetization: \( E = -\frac{J}{2} (M^2 - 1) \). This means that instead of summing over configurations, the partition function can be written down with \( N + 1 \) terms, since this is how many different values of magnetization are there. Each magnetization occurs with multiplicity \( D(M) = \binom{N}{N_+} \), where \( N_+ = (N + M)/2 \) is the number of positive spins. Therefore the partition sum is

\[
Z(T) = \sum_{c} e^{-\beta E(c)} = \sum_{m=-1,\ldots,1} D(m) e^{\beta E(m)},
\]

where \( m \equiv M/N \) is magnetization density. The entropy density \( s \equiv S/M \) at fixed magnetization, in the limit of large \( N \), is

\[
s(m) = -\frac{1 + m}{2} \ln \left( \frac{1 + m}{2} \right) - \frac{1 - m}{2} \ln \left( \frac{1 - m}{2} \right).
\]

Using \( s(m) = N \ln D(m) \), in the limit of large \( N \) we have

\[
Z = \sum_{m=-1,\ldots,1} e^{-\beta N f(m)},
\]

\[
f(m) = \frac{J}{2} m^2 - Ts(m),
\]

where \( f(m) \) is the free energy functional. As we see on Fig. 4, there is a critical temperature \( T_c = J \), below which there are two free energy minima at \( m \propto O(1) \) and above which it is only one free energy minima at zero magnetization. The degeneracy between the two free energy functional minima is lifted with a small perturbation (such as an external magnetic field) and remains lifted even after the perturbation has vanished. At \( N \to \infty \) we have a phase transition, but at finite \( N \) we have fluctuations of magnetization proportional to \( O(N^{3/4}) \). Expanding the free energy functional Eq. (43) for \( T \) close to \( T_c \) we have

\[
f(m) \approx \frac{\tau}{2} m^2 + \frac{m^4}{12} - T \ln 2 + O(m^6, \tau m^4)
\]

where \( \tau = 1 - T/T_c \). For \( \tau > 0 \) the fluctuations of \( Nm \) are proportional to \( N^{-1/2} \), which gives \( \delta = 1/2 \). At the critical temperature the quadratic term and the average magnetization vanish, but the fluctuations are of the order \( N^{-1/4} \) and thus the magnetization \( Nm \) has a distribution of width \( N^{3/4} \). One can learn more about the Ising model on a complete graph from the lecture notes.
FIG. 4. The free energy functional for $J = 1$ and three different temperatures. We notice that $T_c = J$ is the critical temperature — above this temperature the probability distribution of magnetization is centered around 0, below there are two non-zero minima. The degeneracy between the two is lifted with a small perturbation, such as an external magnetic field and remains lifted even after the perturbation has vanished. At $T_c = J$ we notice that the curvature of the saddle point $m = 0$ vanishes, which signifies the second order phase transition.

of Monasson [14], which we followed here to explain this phase transition. Also [15] gives an excellent introduction to magnetic spin system and their simulations.

The time it takes for a reversible MCMC algorithm to decorrelate $R_{mm}(t)$ is proportional to the variance of $m$, i.e. $\propto N^{3/2}$ close to the phase transition. On the other-hand, the proposed lifting algorithm converges as $N^{3/4}$. This we argued and confirmed numerically in [2], by measuring both the inverse spectral gap $\Delta^{-1}$ and the decay of the autocorrelation function of magnetization $R_{mm}(t)$. Here we reproduced these results on Fig. 5.

We created two copies of the system, where in one copy we would be just increasing magnetization and in the other we would be just decreasing magnetization. This resulted in an effective magnetic field, which was dependent on the state of the system and allowed the system to linger longer at states of very low and very high magnetization [2]. Similar observations were made for 1d Ising in [3]. Ultimately the lifted MCMC converged faster to the equilibrium [2] and a pseudocode is given on Fig. 6.

1d Ising Model

Next example is a 1d Ising model of $N$ spins with periodic boundary conditions and energy

$$E(C) = -J \sum_{i=1}^{N} s_i s_{i+1}. \quad (45)$$

Sakai and Hukushima implemented in [3] three different nonreversible MCMC algorithms. All three cases fulfill Eq. (37) and one of them was exactly Eq. (38). In all three cases the nonreversible MCMC converged faster than the reversible MCMC. In the excellent and very intuitively written work the authors explain analytically why in some of the cases the MCMC convergence was faster. The lifted MCMC 1d Ising had Glauber-like transition rates as of a reversible MCMC algorithm for a 1d Ising in a magnetic field that depends on the state of the system [3]. The skew-detailed balance Eq. (26) condition insured that the nonreversible Markov Chain converges to the equilibrium distribution without the magnetic field [3].

FIG. 5. This figure is reproduced from [2] it represents the correlation time of the magnetization autocorrelation function in the Ising model on complete graph. The blue labels represent the reversible MH algorithm and the red labels are the nonreversible MH algorithm (color online). The dots represent $T = 1/(\text{Re}(\Delta))$, where $\Delta$ is the spectral gap, obtained by exact diagonalization of respective transition matrices. The crosses represent correlation times reconstructed by fitting the large time asymptotics with an exponential function, $\exp(-t/\tau_{\text{rev}})$, and exponential-oscillatory function, $\exp(-t/\tau_{\text{nonrev}}) \cos(\omega t - \phi)$, in the reversible and nonreversible cases, respectively. Best slope fits are $\tau_{\text{rev}}(N) \propto N^{1.43}$ and $\tau_{\text{nonrev}}(N) \propto N^{0.85}$.

FIG. 6. Implementation of "lifting" with a MH MCMC algorithm.

{\small
\begin{verbatim}
{initially: system at state $x_+$}
\begin{align*}
&\quad s_i \leftarrow \text{randomly selected spin from $x_+$} \\
&\quad a(x_+, y_+) = \min \left[ 1, \frac{\pi(y_+) / \pi(x_+)}{\pi(x_+) / \pi(y_+)} \right] \{\text{MH proposal (flipping $s_i$ in $x_+$ gives $y_+$)}\} \\
&\quad p \leftarrow \text{random number in (0, 1)} \\
&\begin{cases}
\text{if } p \leq a & \text{spin is flipped and the new state is } y_+ \{\text{MH move}\} \\
\text{else } & \text{attempt to change from } x_+ \text{ to } x_- \\
& \quad q \leftarrow \text{random number in (0, 1)} \\
&\quad \text{if } q \leq \frac{P(x_+, x_-)}{(1 - \sum_{x_+ \neq x_-} P(x_+, x_-)) P(x_+, x_-)} \text{ then} \\
&\quad \quad \text{the new state is } x_- \\
&\quad \text{end if} \\
&\text{end if}
\end{cases}
\end{align*}
\end{verbatim}
}
Applying lifting in the same way as in the previous two examples – by controlling magnetization, did not yield a significant speed up on the 2d Ising model at the asymptotic critical point [5]. Also creating replicas where we allow only specific/biased energy changes, but obey global balance, did not lead to a significant speed up [5]. This speed up seemed to have only changed a prefactor at most, but not the scaling with the system size [3, 5]. Further investigation is needed on how to make lifting adaptive to the obstacles of the configuration space. Thus, although we have successfully created a correct algorithm, that obeys GB, and converges to the proper equilibrium distribution, in the case of 2d Ising we have not yet been able to find the lifting that leads to significantly faster convergence. The lifting that provides the fastest convergence has to utilize the physics of the system at hand. For example in the mean field Ising model – the slowest observable to converge is magnetization and the equilibrium distribution can be written as a function magnetization only. Thus it was very much natural to choose a lifting that "tempers" the magnetization.

DISCUSSION AND COMPARISON TO OTHER METHODS

Besides the above mentioned implementations of lifting [3, 5], there were several other similar ideas, such as e.g. [16–21]. Suwa and Todo have proposed another MCMC method that violates DB and significantly reduces the convergence time of the Potts model, compared to the respective reversible MCMC method [17]. A new non-reversible algorithm was developed for hard-sphere systems [18, 19], with substantial acceleration compared to related variants obeying detailed balance. For a more detailed comparison of the proposed algorithms that violated DB we refer the reader to [2].

On the more rigorous side some of the results that were pertinent to our work can be found in these papers [6–8, 22–25]. Still much less is known about non-reversible Markov Chains, compared to vast knowledge on reversible Markov Chains (see e.g. [10]). For example for the reversible Markov Chains, the Peskun theorem holds. It states that the asymptotic variance of any observable is reduced by increasing the acceptance probability of the Markov Chain (\(\alpha(x,y)\) in our notation for the MH Markov Chain). Chen, Lovasz and Pak show in [7] that lifting can at most introduce a square root improvement of the convergence time. Such an acceleration is still quite impressive for large convergence times.

We have reviewed several works that show how to controllably transform a reversible MCMC algorithm into a nonreversible MCMC algorithm. The main idea is to enlarge the phase space and by that "escape" entropic bottleneck and traps. This method was not designed to cover rugged energy landscapes, there the convergence time to equilibrium is determined by rare events of escaping deep energy wells. Lifting is potentially useful where we have "entropic barriers" – such as a vast energetically almost flat configuration space or a maze with paths of low entropy. Lifting does not require any particular symmetry of the configuration space (for example it does not rely on the \(Z_2\) symmetry of the Ising model). The illustrated simple examples show that lifting can lead to a dramatic reduction of the convergence time. Methods using non-equilibrium mixing (methods that violated detailed balance) might prove useful in studies of phase transitions, soft matter dynamics, protein structures and granular media, etc. An interesting direction for future research is to explore if this additional nonreversibility can improve the convergence properties of well known reversible algorithms [26].

Part of this work was completed in Aspen Center for Physics. MV acknowledges the Aspen Center for Physics and NSF grant #1066293 for support. MV thanks J. Mactha, K. S. Turitsyn, M. Chertkov, C. Moore, K. Hukushima, W. Krauth, C. Godrèche and T. Hayes for illuminating discussions.

* vmarija@rockefeller.edu

[1] A. D. Sokal (1996), Lectures at the Cargése Summer School on "Functional Integration: Basics and Applications".
[2] K. Turitsyn, M. Chertkov, and M. Vucelja, Physica D Nonlinear Phenomena 240, 410 (2011).
[3] Y. Sakai and K. Hukushima, Journal of the Physical Society of Japan 82, 064003 (2013), 1301.4801.
[4] J. A. Fill, The Annals of Applied Probability 1, 62 (1991).
[5] H. C. Fernandes and M. Weigel, Computer Physics Communications 182, 1856 (2011), ISSN 0010-4655, computer Physics Communications Special Edition for Conference on Computational Physics Trondheim, Norway, June 23-26, 2010, URL http://www.sciencedirect.com/science/article/pii/S0010465510004637.
[6] P. Diaconis, S. Holmes, and R. M. Niel, The Annals of Applied Probability 10, 726 (2000).
[7] F. Chen, L. Lovasz, and I. Pak, Proceedings of the ACM symposium on Theory of Computing pp. 275–281 (1999).
[8] J. Bierkens, ArXiv e-prints (2014), 1401.8087.
[9] Liftings of Tree-Structured Markov Chains (2010).
[10] D. A. Levin, Y. Peres, and E. L. Wilmer, Markov Chains and Mixing Times (American Mathematical Society, 2009).
[11] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, J. of Chem. Phys. 21, 1087 (1953).
[12] W. K. Hastings, Biometrika 57, 97 (1970).
[13] W. Krauth, Statistical Mechanics: Algorithms and Computations (Oxford University Press, Oxford, 2006).
[14] R. Monasson (2005), lecture notes.
[15] H. Gould and J. Tobochnik, Statistical and Thermal Physics with Computer Applications (Princeton University Press, 41 William Street, Princeton, New Jersey 08540, 2010).
[16] A. Ichiki and M. Ohzeki, Phys. Rev. E 88, 020101 (2013), URL http://link.aps.org/doi/10.1103/PhysRevE.88.020101.
[17] H. Suwa and S. Todo, Phys. Rev. Lett. 105, 120603 (2010), URL http://link.aps.org/doi/10.1103/PhysRevLett.105.120603.
[18] E. P. Bernard, W. Krauth, and D. B. Wilson, Phys. Rev. E 80, 056704 (2009), URL http://link.aps.org/doi/10.1103/PhysRevE.80.056704.
[19] E. P. Bernard and W. Krauth, Phys. Rev. E 86, 017701 (2012), URL http://link.aps.org/doi/10.1103/PhysRevE.86.017701.
[20] R. M. Neal, ArXiv Mathematics e-prints (2004), math/0407281.
[21] R. D. Schram and G. T. Barkema, Physica A: Statistical Mechanics and its Applications 418, 88 (2015), ISSN 0378-4371, proceedings of the 13th International Summer School on Fundamental Problems in Statistical Physics, URL http://www.sciencedirect.com/science/article/pii/S0378437114004804.
[22] T. Hayes, private communication.
[23] C. Godrèche and A. J. Bray, Journal of Statistical Mechanics: Theory and Experiment 2009, P12016 (2009).
[24] T. Lelièvre, F. Nier, and G. Pavliotis, Journal of Statistical Physics 152, 237 (2013), ISSN 0022-4715, URL http://dx.doi.org/10.1007/s10955-013-0769-x.
[25] L. Rey-Bellet and K. Spiliopoulos, ArXiv e-prints (2014), 1404.0105.
[26] M. Vucelja and J. Machta (2014), in preparation.