A method to reduce the rejection rate in Monte Carlo Markov chains

A method to reduce the rejection rate in Monte Carlo Markov Chains

Carlo Baldassi[1,2]

[1] *Department of Applied Science and Technology, Politecnico di Torino, Corso Duca degli Abruzzi 24, I-10129 Torino, Italy*
[2] *Human Genetics Foundation-Torino, Via Nizza 52, I-10126 Torino, Italy*

We present a method for Monte Carlo sampling on systems with discrete variables (focusing in the Ising case), introducing a prior on the candidate moves in a Metropolis-Hastings scheme which can significantly reduce the rejection rate, called the reduced-rejection-rate (RRR) method. The method employs same probability distribution for the choice of the moves as rejection-free schemes such as the method proposed by Bortz, Kalos and Lebowitz (BKL) [1]; however, it uses it as a prior in an otherwise standard Metropolis scheme: it is thus not fully rejection-free, but in a wide range of scenarios it is nearly so. This allows to extend the method to cases for which rejection-free schemes become inefficient, in particular when the graph connectivity is not sparse, but the energy can nevertheless be expressed as a sum of two components, one of which is computed on a sparse graph and dominates the measure. As examples of such instances, we demonstrate that the method yields excellent results when performing Monte Carlo simulations of quantum spin models in presence of a transverse field in the Suzuki-Trotter formalism, and when exploring the so-called robust ensemble which was recently introduced in [2]. Our code for the Ising case is publicly available [3], and extensible to user-defined models: it provides efficient implementations of standard Metropolis, the RRR method, the BKL method (extended to the case of continuous energy specra), and the waiting time method [4].

CONTENTS

I. INTRODUCTION

Monte Carlo methods play a central role in the simulation and study of a variety of physical and mathematical problems. In particular, Monte Carlo Markov Chains (MCMC) are a class of very powerful and general algorithms to sample from arbitrary probability distributions [5]; furthermore, they are at the basis of some general optimization techniques such as simulated annealing [6] and quantum annealing [7]. Arguably, the most popular general MCMC scheme is based on the Metropolis-Hastings rule. Concrete implementations of the rule can however suffer, particularly in frustrated systems, from the problem of having a very slow dynamics; one typical manifestation of this is a very high rejection rate of the proposed moves. Indeed, a wealth of different techniques have been proposed in order to overcome this problem and improve the sampling efficiency for specific classes of systems, such as Kinetic Monte Carlo [8], Cluster Monte Carlo [9, 10], parallel tempering [11], and many others (see e.g. [12]).

In this paper, we propose yet another such variant, which we call "reduced-rejection-rate Monte Carlo", RRR for short, with a publicly available generic implementation [3], and present some exploratory numerical results that demonstrate its advantages. The method as described here is applicable to the case of Ising spin systems but could

be generalized to Potts-like models, and consists in introducing a prior on the choice of the candidate moves in an otherwise standard Metropolis-Hastings scheme, with the aim – as the name suggests – of reducing the rejection rate.

Consider a system of N interacting Ising spins $\sigma = \{\sigma_i\}_{i=1}^N \in \{-1,+1\}^N$ subject to some Hamiltonian $E(\sigma)$. For simplicity, in the following we will consider Hamiltonians restricted to at most pairwise interactions between the spins, but the treatment is general. Let us then call $\mathcal{L}$ the set of interacting pairs, and write:

$$E(\sigma) = -\sum_{(i,j)\in\mathcal{L}} J_{ij}\sigma_i\sigma_j + \sum_i h_i\sigma_i. \tag{1}$$

We want to sample from the Boltzmann probability distribution at inverse temperature β:

$$P(\sigma) = \frac{e^{-\beta E(\sigma)}}{Z}. \tag{2}$$

In the usual MCMC scheme, using the Metropolis-Hastings rule, one starts from a configuration σ, proposes a candidate move $\sigma \to \sigma'$ with some prior probability distribution $C(\sigma \to \sigma')$, and accepts the move with some acceptance rate $A(\sigma \to \sigma')$. Therefore, we write the transition probability of going from configuration σ to a different configuration σ' in the Markov Chain as:

$$P(\sigma \to \sigma') = C(\sigma \to \sigma')\, A(\sigma \to \sigma') \qquad \sigma' \neq \sigma. \tag{3}$$

Of course, $P(\sigma \to \sigma) = 1 - \sum_{\sigma'\neq\sigma} P(\sigma \to \sigma')$.

The rejection rate is determined by enforcing the detailed balance condition $P(\sigma)\,P(\sigma \to \sigma') = P(\sigma')\,P(\sigma' \to \sigma)$:

$$\frac{A(\sigma \to \sigma')}{A(\sigma' \to \sigma)} = e^{-\beta\left(E(\sigma')-E(\sigma)\right)}\frac{C(\sigma' \to \sigma)}{C(\sigma \to \sigma')} \tag{4}$$

which can be accomplished by the usual formula:

$$A(\sigma \to \sigma') = \min\left(1, e^{-\beta\left(E(\sigma')-E(\sigma)\right)}\frac{C(\sigma' \to \sigma)}{C(\sigma \to \sigma')}\right). \tag{5}$$

In a straightforward implementation, the proposed moves consist in choosing one spin uniformly at random and flipping it. Then, the C terms in the above equation simplify and one is left with the simple rule:

$$P(\sigma \to \sigma') = \min\left(1, e^{-\beta\left(E(\sigma')-E(\sigma)\right)}\right). \tag{6}$$

Throughout the paper, we refer to this choice as a "standard Metropolis scheme". In this work, we will instead focus on choosing the prior C in such a way to reduce the rejection rate, while still keeping its computation efficient. The rejection rate minimization is achieved, informally speaking, by making the quantity

$$R(\sigma \to \sigma') = e^{-\beta\left(E(\sigma')-E(\sigma)\right)}\frac{C(\sigma' \to \sigma)}{C(\sigma \to \sigma')} \tag{7}$$

as close as possible to 1. We will still only consider single flips as the candidate moves.

The basis of our method is (a generalization of) the rejection-free algorithm by Bortz, Kalos and Lebowitz [1], which we refer to as BKL throughout this paper (it is also known as "the n-fold way"). The BKL method goes under the general category of "Kinetic Monte Carlo", or "faster-than-the-clock" (FTTC) schemes, see e.g. [5]. The general FTTC technique is to pre-compute the probability distribution of the number of consecutive rejections before a move would be accepted, extract a number of MCMC iterations to skip from that distribution (thus "advancing the clock"), then choose a move without rejections. The resulting MCMC has the same statistics as a standard Metropolis scheme, but skipping the rejections may be computationally convenient, especially at low temperatures when the rejection rate is high. Of course, the additional computations involved only lead to an advantage in some cases. In particular, the BKL method was introduced as an efficient method for the case of Ising spin systems with small connectivity, and in which the energy shift $\Delta E(\sigma \to \sigma') = E(\sigma') - E(\sigma)$ induced by a spin flip can only belong to a small set of values. As an example of such system, consider an Edwards-Anderson model, i.e. a D-dimensional lattice with nearest neighbor interactions and periodic boundary conditions, with $J_{ij} \in \{-1,+1\}$ and $h_i = 0$, in which case there are only $2D+1$ possible values of ΔE: $\Delta E \in \{-4D, -4D+4, \ldots, 4D-4, 4D\}$. As it turns out, the requirement of small connectivity $K \ll N$ is the only crucial one in order to achieve an efficient implementation, while the requirement that the energy shifts are discrete is useful for further optimizing the method: for general energy shifts, an efficient

rejection-free method exists, the waiting time method (WTM for short) [4], but we will show that the BKL can also be extended rather straightforwardly and achieve comparable performances to WTM. We will shortly review the BKL method in section II, and show how to modify it in the case of general heterogeneous energy shifts.

The core of our RRR method is to use the same probability distribution for the choice of the move as in the BKL method, but this time as the prior C, without the initial skipping of the rejected moves; this leaves a residual rejection rate, given by eq. (5). Although this may appear to be a net loss, this is often not the case, and may even be slightly advantageous in a number of situations, as shown in section V A. The main advantage of the RRR method however is that it is more easily generalizable. Consider the case in which the energy can be written as the sum of two components:

$$E(\sigma) = E_s(\sigma) + E_d(\sigma) \tag{8}$$

where $E_s(\sigma)$ describes a model with low connectivity, while $E_d(\sigma)$ is a residual part of the energy for which the BKL or WTM methods are not convenient. With our approach, since – as we shall show – we can achieve

$$e^{-\beta\left(E_s(\sigma')-E_s(\sigma)\right)}\frac{C(\sigma'\to\sigma)}{C(\sigma\to\sigma')} \approx 1 \tag{9}$$

we are then only left with the acceptance rate relative to the non-sparse part of the system:

$$A(\sigma\to\sigma') \approx \min\left(1, e^{-\beta\left(E_d(\sigma')-E_d(\sigma)\right)}\right). \tag{10}$$

Then, with the RRR method, we can almost completely absorb the effect of the interaction E_s by including it in the prior. This basic idea was used, with the method we propose here in embryonic form, when studying the so-called robust ensemble (RE) introduced in ref. [2]; the RRR method is a generalization and an improvement of what was used in that paper. This case is discussed in section V C. An important case that also falls in this category is the simulation of quantum spin systems via the Suzuki-Trotter transformation [13], in which case E_s can be identified with the interactions between the replicated Suzuki-Trotter spins. This case is discussed in section V B, where indeed we will show that RRR is able to equilibrate much faster than the standard Metropolis scheme.

II. THE BKL METHOD

In this section, we briefly review the BKL method. In its original formulation, the authors called the method "the n-fold way", since they were considering the case in which the energy shifts arising from a spin flip could only take a value in a small discrete set. Here, however, we will start with the more general case, and recover the original formulation as a specialized case. We will thus introduce some definitions and notations which we will used for RRR as well.

Let us indicate by $\sigma^{(i)}$ a configuration of the spins obtained from another configuration σ by flipping the spin i:

$$\sigma_j^{(i)} = \begin{cases} \sigma_j & \text{if } j \neq i \\ -\sigma_j & \text{if } j = i. \end{cases} \tag{11}$$

We denote the effect of such a spin flip on the energy as:

$$\Delta E_\sigma^{(i)} = E\left(\sigma^{(i)}\right) - E(\sigma). \tag{12}$$

Then, for a given configuration σ and all possible spin flips, we define the following quantities:

$$p_\sigma^{(i)} = \min\left(1, e^{-\beta\Delta E_\sigma^{(i)}}\right) \tag{13}$$

$$z_\sigma = \sum_i p_\sigma^{(i)}. \tag{14}$$

With these, the probability of rejecting a move in a standard Metropolis scheme (eq. (6)) is $P(\sigma\to\sigma) = 1 - z_\sigma/N$, while $P\left(\sigma\to\sigma^{(i)}\right) = p_\sigma^{(i)}/N$ is the probability of transitioning to the new configuration $\sigma^{(i)}$. The BKL procedure at each step is then as follows:

1. Extract a number of iterations to skip as $\left\lfloor \frac{\log(1-r)}{\log(1-z_\sigma/N)} \right\rfloor$, where r is a random number extracted uniformly in $[0,1)$, and "advance the clock" accordingly.

2. Extract a spin i with probability $p_\sigma^{(i)}/z_\sigma$ and flip it, thereby changing the configuration to $\sigma^{(i)}$.

3. Compute the new values $p_{\sigma^{(i)}}^{(j)}$ for all j and the new $z_{\sigma^{(i)}}$.

As mentioned above, by this procedure one realizes the same MCMC transition matrix of standard Metropolis, but can save computational time by skipping the rejected iterations entirely (step 1), at the cost of requiring a specialized sampling procedure (step 2) and some additional bookkeeping (step 3). Note that in step 3 only the flipped spin and its neighbors may change their values of $p_\sigma^{(i)}$, therefore the bookkeeping operations can be relatively inexpensive for diluted graphs (the details on how to achieve this are given in section IV), which explains the reason for the requirement $K \ll N$ given in the introduction.

When expressed in this general form, the BKL method is almost equivalent to another closely related rejection-free method, the waiting time method (WTM) [4]. However, in the special case in which the set $\left\{\Delta E_\sigma^{(i)}\right\}$ has a small cardinality for all σ and all i (e.g. in the case of a regular lattice with ± 1 couplings mentioned in the introduction), a further optimization can provide additional computational advantages. In that case, which is the one originally considered by the authors in [1], the procedure is as follows: for a given configuration σ, we divide all the spins into classes, based on the energy shift induced by flipping them:

$$\mathcal{C}_\sigma\left(\Delta E\right) = \left\{i : E\left(\sigma^{(i)}\right) - E\left(\sigma\right) = \Delta E\right\}. \tag{15}$$

All the spins in the same class will thus have the same associated probability $p_\sigma^{(i)}$ (eq. (13)), thus we only need to keep track of one probability for each class, and of the class sizes. Let us then define:

$$n_\sigma\left(\Delta E\right) = \left|\mathcal{C}_\sigma\left(\Delta E\right)\right| \tag{16}$$

$$p_\sigma\left(\Delta E\right) = n_\sigma\left(\Delta E\right) \min\left(1, e^{-\beta \Delta E}\right). \tag{17}$$

With these, we can also express z_σ (eq. (14)) as:

$$z_\sigma = \sum_{\Delta E} p_\sigma\left(\Delta E\right). \tag{18}$$

In this specialized case, the step 2 of the BKL scheme above is performed in two separate steps:

2. (a) Extract a class $\mathcal{C}_\sigma\left(\Delta E\right)$ with probability $p_\sigma\left(\Delta E\right)/z_\sigma$
   (b) Extract a spin uniformly at random from the chosen class, and flip it.

The update step 3 is also simplified, since it can use a more efficient data structure. The details are given in section IV.

III. THE RRR METHOD

In the RRR method, when applied to sparse models, the proposal $C\left(\sigma \to \sigma'\right)$ simply follows the BKL step 2 of the previous section, without the skipping step 1 and without accepting the move right away, i.e.:

$$C\left(\sigma \to \sigma^{(i)}\right) = \frac{p_\sigma^{(i)}}{z_\sigma}. \tag{19}$$

With this choice, the expression involved in the acceptance rate $A\left(\sigma \to \sigma^{(i)}\right)$, eq. (7), takes the remarkably simple form:

$$R\left(\sigma \to \sigma^{(i)}\right) = \frac{z_\sigma}{z_{\sigma^{(i)}}} \tag{20}$$

where z_σ is the same as for the BKL algorithm, eq. (14), and $z_{\sigma^{(i)}}$ is the same quantity computed for the new candidate configuration $\sigma^{(i)}$. Indeed, this choice significantly reduces – and in many cases nearly eliminates – the rejection rate. The basic reason for this is that, at fixed (non-zero) temperature and in the thermodynamic limit $N \to \infty$, z_σ is an extensive quantity and the perturbation induced by the spin flip is at most of the order of the connectivity K, therefore $z_{\sigma^{(i)}} = z_\sigma + \mathcal{O}\left(K\right)$ and $R\left(\sigma \to \sigma^{(i)}\right) \approx 1$ to the leading order. Only at fixed N and very low temperatures the difference becomes more significant (and indeed, in the limit of $\beta \to \infty$ at fixed N the rejection rate must tend to 1 whenever the system is in a local minimum); in practice, though, this regime appears to be rather narrow, while the acceptance rate seems to be very close to 1 up to rather high values of β (see the numerical experiments of

section V A). Moreover, consider a system initialized in a random configuration in the initial transient regime, at an energy $E(\sigma)$ far above the equilibrium one:[1] the proposed move i will be more likely associated to a negative energy shift $\Delta E_\sigma^{(i)} < 0$, due to the form of $p_\sigma^{(i)}$ in eq. (13). In that case, the contribution of spin i to the normalization term z is 1 in z_σ and it is $e^{\beta \Delta E_\sigma^{(i)}}$ in $z_{\sigma^{(i)}}$, and thus $z_{\sigma^{(i)}} < z_\sigma$ unless the effect is overcome by the shifts on the neighbors, which means that in the initial stages there is a bias towards $R\left(\sigma \to \sigma^{(i)}\right) > 1$, further reducing the rejection rate.

The resulting MCMC transition matrix is thus no longer the same as that of standard Metropolis, and has in general better convergence properties (for very small systems, this can be assessed numerically by computing the eigenvalues of the transition matrices).

When the method is applied to the two-level systems mentioned in the introduction, i.e. systems such that the energy can be written as $E(\sigma) = E_s(\sigma) + E_d(\sigma)$ with only $E_s(\sigma)$ begin sparse, the RRR algorithm is straightforwardly modified as follows: first, a candidate spin i is chosen according to $E_s(\sigma)$ as described above, and $R\left(\sigma \to \sigma^{(i)}\right)$ is computed from eq. (20); then, the residual energy shift $\Delta E_{d\sigma}^{(i)} = E_d\left(\sigma^{(i)}\right) - E_d(\sigma)$ is computed; finally the move is accepted with probability

$$A\left(\sigma \to \sigma^{(i)}\right) = \min\left(1, R\left(\sigma \to \sigma^{(i)}\right) e^{-\beta\, \Delta E_{d\sigma}^{(i)}}\right). \tag{21}$$

When ΔE_s is typically much larger than ΔE_d, this method can provide very significant improvements. Experiments on this type of systems are shown in sections V B and V C.

In principle, it may also be possible to exploit the fact that discrete energy spectra allow for a specialized version of the algorithm, and apply the RRR method to sparse graphs with continuous spectra by discretizing the interactions and allowing for a small residual rejection rate (e.g. by writing the couplings as $J_{ij} = J_{ij}^0 + \delta J_{ij}$ with the J_{ij}^0 chosen from a small discrete set and such that $|\delta J_{ij}|$ is small). In our tests, however, this strategy led at best to the same performances as using the more general versions of BKL/RRR, or WTM, i.e. the slight advantage of the discretization was always at least counterbalanced by the disadvantage of having a slightly increased rejection rate, and we have thus abandoned this line of inquiry.

IV. IMPLEMENTATION DETAILS

During each step of the BKL or RRR algorithms, we need to have an efficient way to *a)* sample from the distribution $p_\sigma^{(i)}/z_\sigma$, and *b)* update the distribution under the assumption that only a small subset $K \ll N$ of the $p_\sigma^{(i)}$ will have changed. To achieve this in our implementation, we used a specialized binary tree which keeps "local cumulative distributions" in $\mathcal{O}(N)$ memory and can be accessed and updated in $\mathcal{O}(\log N)$ time [14]. More precisely, let us at first assume for simplicity that N is a power of 2, and define the table of arrays $\{a_l\}$, with $l = 1, \ldots, \log_2 N$, of variable length 2^{l-1}; in each entry of an array a_l, we store the sum of $N/2^l$ elements:

$$a_l(k) = \sum_{j=1}^{N/2^l} p_\sigma^{\left((k-1)N/2^{l-1}+j\right)} \qquad k = 1, \ldots, 2^{l-1}. \tag{22}$$

We also compute and store $z_\sigma = \sum_{i=1}^N p_\sigma^{(i)}$. To extract a random element i with probability $p_\sigma^{(i)}/z_\sigma$, we use this procedure:

1. extract a random number $x \in [0, z_\sigma)$, and initialize $k \leftarrow 0$;
2. looping over each $l = 1, \ldots, \log_2 N$, do:
   (a) if $x > a_l(k+1)$, set $k \leftarrow 2k+1$ and $x \leftarrow x - a_l(k+1)$
   (b) otherwise set $k \leftarrow 2k$;
3. return $i = k+1$.

This requires $\mathcal{O}(\log N)$ elementary operations. Updating the structure when one of the $p_\sigma^{(i)}$ is changed is also $\mathcal{O}(\log N)$ since each entry only appears in at most $\log_2 N$ entries in the $\{a_l\}$ table, and z_σ can be updated with a single operation.[2]

---

[1] And also far above that of any long-lived meta-stable state which may trap the dynamics, should they exists.

[2] Other schemes based on more sophisticated data structures such as [15] can achieve better asymptotic performances, but they involve more complex operations, and also consume more random numbers. We consider it unlikely that they could positively and significantly affect the results which we present here considering the system sizes involved (and in any case this would only affect the comparisons between BKL/RRR and WTM for the continuous cases), and we have thus left their use as a potential future improvement to the code.

In the general case in which N is not a power of 2, we simply pad the distribution with zeros. Also note that the WTM has the same complexity, since it uses a binary heap as its underlying data structure for performing the analogous sampling and update operations.

In the specialized case of discrete energy shifts, however, we use basically the same approach and data structures as described in the original BKL paper [1]: at each time, we keep track of the $\mathcal{C}(\Delta E)$ classes' compositions by using unsorted lists of variable size, and associated look-up tables to determine the position of each spin in the structure. Updating the position of a spin within the structure is an $\mathcal{O}(1)$ operation, since the structure is unsorted. For example, removing a spin from a list amounts at doing the following: locate its position using the look-up table, move the last spin of the list in its position, reduce the list size, update the look-up table. This data structure then allows to *a)* keep track of the values of n_σ (eq. (16)) for all classes, *b)* have an efficient way to choose a spin within a class, and *c)* determine how the neighbors' classes are affected by the move and perform the update. Note that when the number of classes is very small, the choice of the class (step 2a at the end of section II) is most efficiently performed by simply extracting a uniform random number in $[0, 1)$ and computing the cumulative distribution on the fly until it exceeds that number, without the need of specialized structures. This is for cache efficiency reasons. When instead the number of classes is not so small that more sophisticated methods are required, the discrete specialization is not particularly convenient over the more general method.

When determining the effect of the move on the energy shifts, a loop over the neighbors of the chosen spin is necessary. This introduces an $\mathcal{O}(K)$ cost, where K is the connectivity, as mentioned above. The difference between BKL and RRR is that, in the latter case, this computation is required before the move is accepted. For RRR, we have thus two possible modes of operation: we can either always accept the change at first and undo it if it happens to be rejected (we call this the *undo mode*), or we can keep in memory the results of the computation (the potential shifts induced by the move) and apply them later if the move is accepted (we call this the *staged mode*). Empirically, we found that the undo mode has better performance at high acceptance rate, while the staged mode becomes convenient at lower acceptance rates (the transition between the two regimes happening at about 0.5 or 0.8 acceptance rate on the discrete or continuous problems we tested, respectively). Therefore, our code determines which strategy to employ based on an ongoing estimate of the acceptance rate. When the acceptance rate of RRR is high, these additional computations are very rarely wasted with respect to the BKL method – it is rare that a move needs to be undone; on the other hand, the RRR method avoids the computation of the number of steps to skip (point 1 in section II), which, as it turns out, can also be relatively expensive. On balance, RRR may be slightly more convenient than BKL even on diluted models, up to some value of β where the acceptance rate of RRR starts decreasing.

Finally, we mention the fact that it is common practice in Monte Carlo simulations to boost performance by keeping some kind of memory cache (e.g. the local fields acting on a spin); the need, at least in the staged mode of operation, to perform tentative spin flips, then roll them back, and finally perhaps accepting them, introduces a slight complication in this regard. This however is solved quite easily by keeping an additional cache level which allows to undo the last performed change: in our tests, this proved to be both computationally cheap and sufficient to keep the advantages of the cache mechanisms. We refer the interested reader to the provided code [3] for concrete examples.

V. NUMERICAL EXPERIMENTS

We performed numerical experiments comparing the performance of standard Metropolis, BLK (where applicable), WTM (where applicable) and RRR. All tests were run under identical conditions on a 2.5 GHz Intel i7-4710HQ CPU with a single thread; the code is written in the Julia programming language and run with Julia version 0.5 on a Linux operating system.[3]

In all tests, we set a hard wall-clock-time limit (e.g. 60 seconds), and compared most results (e.g. the energy or the overlaps) as a function of the wall-clock time. The results were sampled at regular intervals in the simulated Monte Carlo time: we took one sample for every 10^4 attempted moves in the RRR case, and we scaled the sampling intervals for the other algorithms in order to make sure that a similar amount of samples was collected for each algorithm. We mostly explored the low-temperature regime, which is the one where RRR (as well as BKL and WTM) can be expected to offer an advantage over standard Metropolis.

[3] For reference, we benchmarked the performance of our generic Julia code against a very efficient specialized C implementation of BKL on a p-spin model which was kindly made available to us by G. Parisi, and found them to be identical.

A. Tests on sparse models

The first batch of experiments which we present here was performed on random regular graphs (RRG) with connectivity $K = 3$, with random couplings. We tested two cases, binary couplings $J_{ij} \in \{-1, +1\}$ and continuous normally distributed couplings $J_{ij} \sim \mathcal{N}(0, 1)$, extracted uniformly and independently, in both cases with no external fields ($h_i = 0$, see eq. (1)).[4] For each case, we tested at least 20 different random graphs, and ran 2 independent tests for each algorithm and each setting of the parameters. We performed all experiments at fixed β, starting from random initial conditions. By construction, we were not expecting the RRR method to outperform BKL or WTM on this type of graph. Thus, the main purpose of these experiments was to check the behavior of RRR, in particular as compared to BKL and WTM, exploring different values of β and N. As expected, at large N and not-too-large β, RRR has a very high acceptance rate (e.g. larger than 0.99 for $\beta = 2$ in both models with $N = 10^4$, see below), and thus it should behave almost identically to BKL. As it turns out, it is even slightly better in some regimes.

Figure 1 shows the energy as a function of time (for models with $N = 10^4$, with β ranging from 2 to 4), for 60 seconds of simulations (thus nowhere near equilibrium). Figure 2 shows the overlaps as a function of time (for models with $N = 10^4$ at $\beta = 2$) and the RRR acceptance rates at different values of β and N.

Following [16], we computed both the self-overlaps (the overlaps at different times, within a certain time interval, for the same run) and the cross-overlaps (the overlaps for different runs on the same graph at comparable times), and used their difference as a measure of the distance from equilibrium. In both cases, we divided the time in regular intervals in logarithmic scale (so that for example in figure 2 the points displayed at $t = 15s$ used all samples collected between $15s$ and $30s$, while the points displayed at $30s$ used the samples collected between $30s$ and $60s$). The results make clear that the tests are indeed very far from equilibrium, but they help discriminating better the different algorithms, and are qualitatively consistent with the picture that emerges from comparing the energies.

Figure 3 shows the number of moves performed on average per second for each algorithm, as a function of β, again confirming the overall picture.

As mentioned above, since on these models BKL is essentially identical to RRR in the very-high-acceptance-rate regime algorithmically, the slight advantage of RRR over BKL in the discrete case at $\beta = 2$ (which, although not clearly visible in figure 1a, is visible in figure 2c and even more clearly in figure 3a) can only be ascribed to the computation of the number of moves to skip (step 1 in section II).

Figures 1 and 3 show that only at $\beta = 4$ in the discrete case RRR performs slightly but measurably worse than BKL, since in that case the RRR acceptance rate dropped to 0.66 (figure 2e). For the continuous case, instead, even at these low temperatures the acceptance rate is still greater than 0.99 (figure 2f), and still better than BKL in terms of moves per second: this is due to the fact that a fraction of the variables have nearly 0 energy shift cost and are thus selected with higher probability, while in discrete RRG case no such cases exists since K is odd. In general, the bottom panels in figure 2 show that, for any given β, the RRR acceptance rate increases monotonically with N, as expected from the arguments given in section III.

B. Quantum Monte Carlo tests

In the second set of experiments, we used the Suzuki-Trotter transformation to study a quantum system of spins in a transverse magnetic field. We will only sketch the method here, see e.g. [17] for a nice and more thorough introduction. Consider the following Hamiltonian operator:

$$\hat{H} = -\sum_{ij} J_{ij} \hat{\sigma}_i^z \hat{\sigma}_j^z - \Gamma \sum_i \hat{\sigma}_i^x \tag{23}$$

where $\hat{\sigma}_i^z$ is the spin operator (Pauli matrix) in the longitudinal direction z, $\hat{\sigma}_i^x$ is the spin operator in the transverse direction x, and $\Gamma \geq 0$ is a magnetic field. The goal is to study the statistical mechanics properties of the system at inverse temperature β, i.e. the partition function $Z = \mathrm{Tr} e^{-\beta \hat{H}}$ and the average value of the observables $\left\langle \hat{O} \right\rangle = Z^{-1} \mathrm{Tr} \hat{O} e^{-\beta \hat{H}}$. The well-known Suzuki-Trotter transformation [13] allows to address this problem by using an effective classical Hamiltonian of Ising spins, with an additional dimension with periodic boundary conditions. The equivalence is realized when the size M of this additional ("imaginary", or "Trotter") dimension diverges. The effective classical Hamiltonian is written as:

$$H_{\text{eff}} = -\frac{1}{M} \sum_k \sum_{ij} J_{ij} \sigma_{ik} \sigma_{jk} - \gamma \sum_{ki} \sigma_{ik} \sigma_{i(k+1)} \tag{24}$$

[4] Qualitatively similar results as those shown here were obtained with Edwards-Anderson graphs and p-spin models with random couplings.

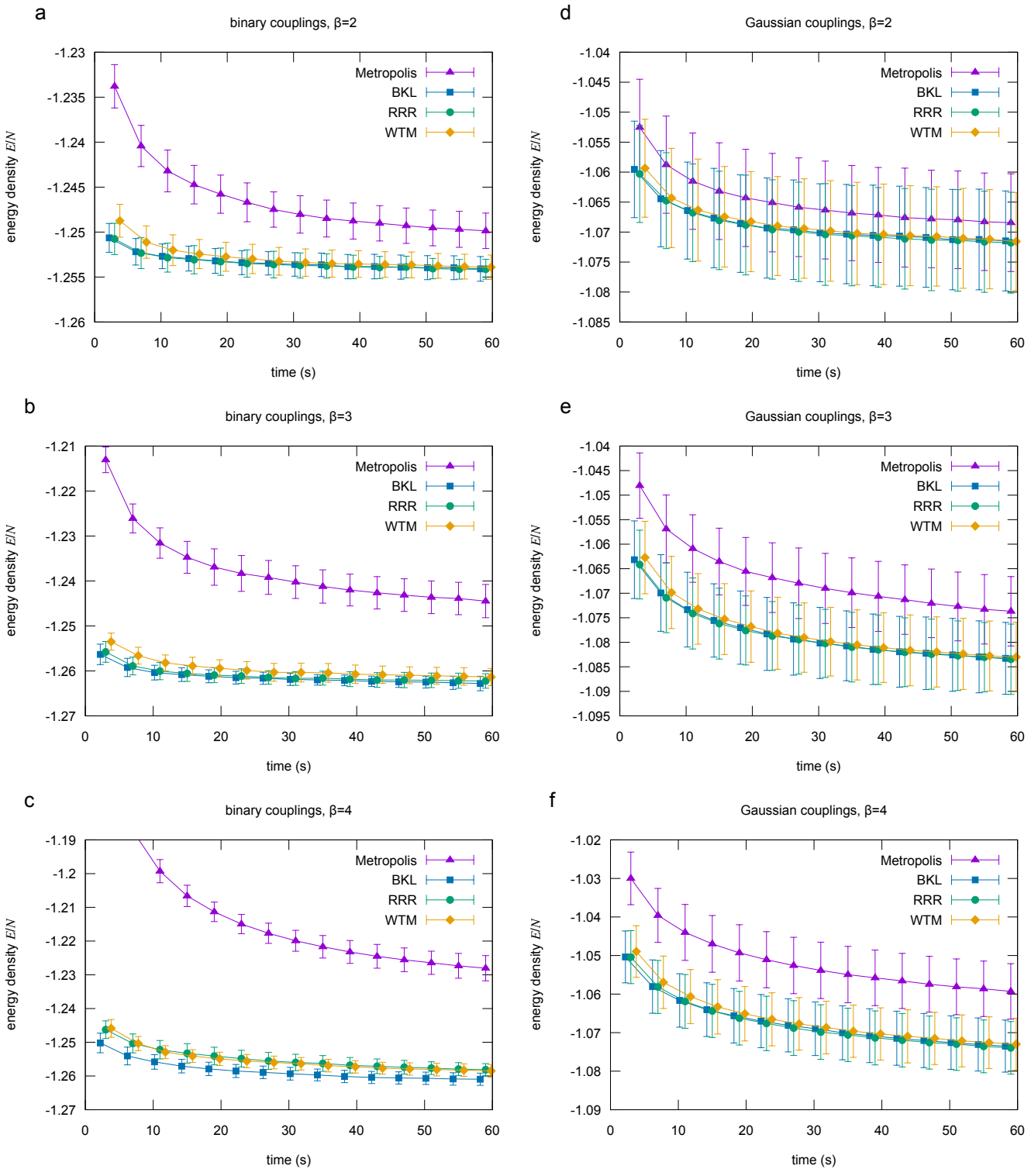

Figure 1. Energy density as a function of time for four different algorithms at different values of β on 20 random regular graphs (100 runs for panel **a**; 2 runs per algorithm per graph in all cases) with $N = 10^4$ and $K = 3$, with either binary couplings (left column) or Gaussian couplings (right column). See details in the text. The points and error bars show the means and standard deviations (computed on bins of regular size; they are slightly shifted relatively to each other for improved readability). For the discrete case, the RRR algorithm performs best at $\beta = 2$ (this is visible in figures 2c and 3a), RRR and BKL are about tied at $\beta = 3$, and BKL performs best at $\beta = 4$ when the acceptance rate of RRR drops (see also figures 2e and 3a); WTM is slightly less optimized for this case and performs worse than BKL. For the continuous case, BKL, RRR and WTM are all basically equivalent at all temperatures (WTM is slightly worse), see also figure 3b.

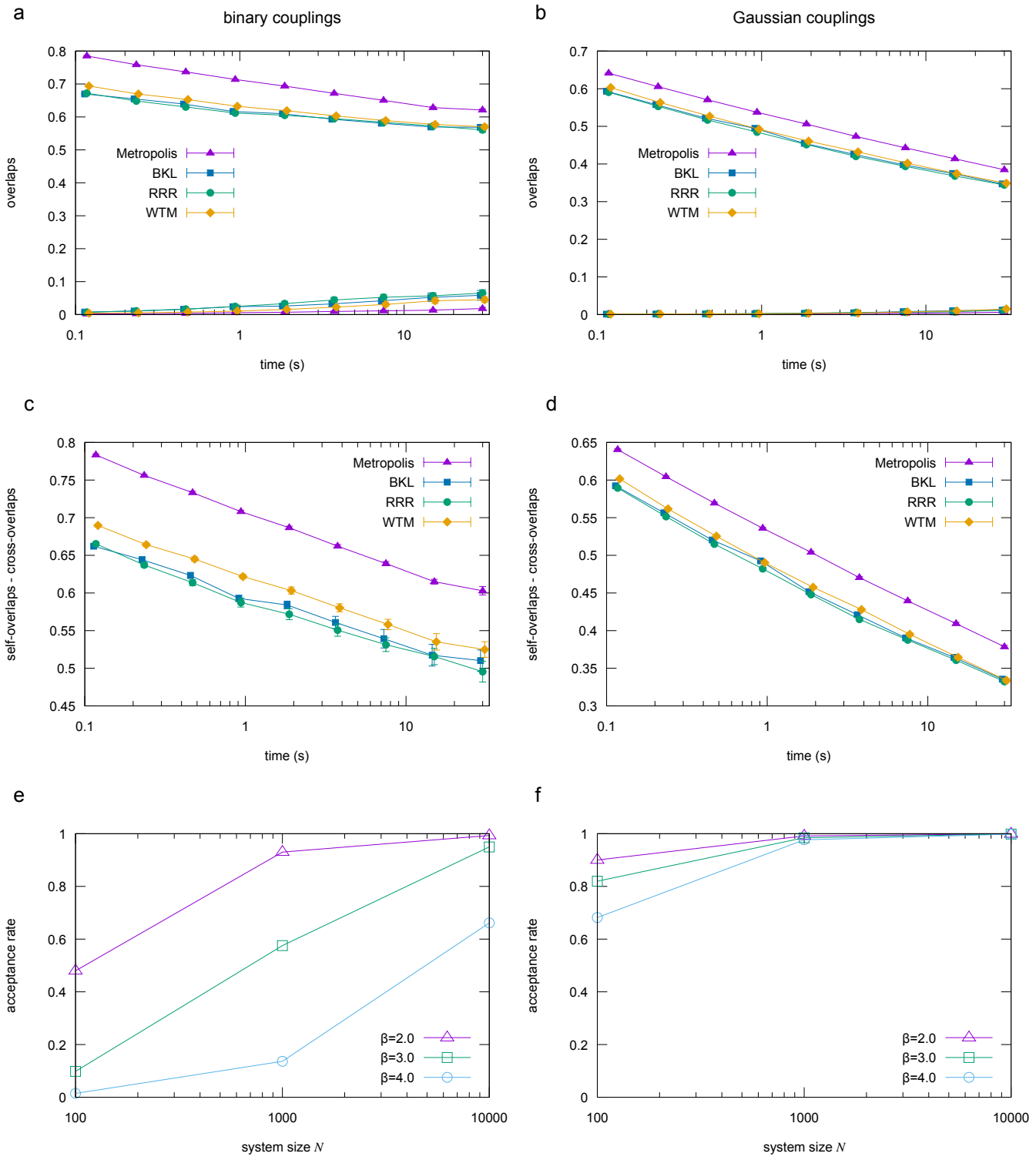

Figure 2. Overlaps vs time and RRR acceptance rates for random regular graphs models at $K = 3$ with either binary couplings (left column) or Gaussian couplings (right column). Top row: average self-overlaps (top curves) and cross-overlaps (bottom curves) for $N = 10^4$ and $\beta = 2$, for four different algorithms (same tests as panels **a** and **d** of figure 1). Points are slightly shifted relative to each other for improved readability. Middle row: average difference between self-overlaps and cross-overlaps, same data as the two top panels. This shows that BKL is slightly better than RRR in the discrete case at $\beta = 2$, and that RRR, BKL and WTM are equivalent in the continuous case; data for $\beta = 3$ and $\beta = 4$ (not shown) agrees with the qualitative picture emerging from figure 1. Bottom panels: acceptance rate of RRR as a function of N for different values of β (the values at $N = 10^4$ correspond to the plots in figure 1). Error bars are smaller than the data points. The rates are higher in the continuous case because the couplings can be small and some spins can be flipped almost freely.

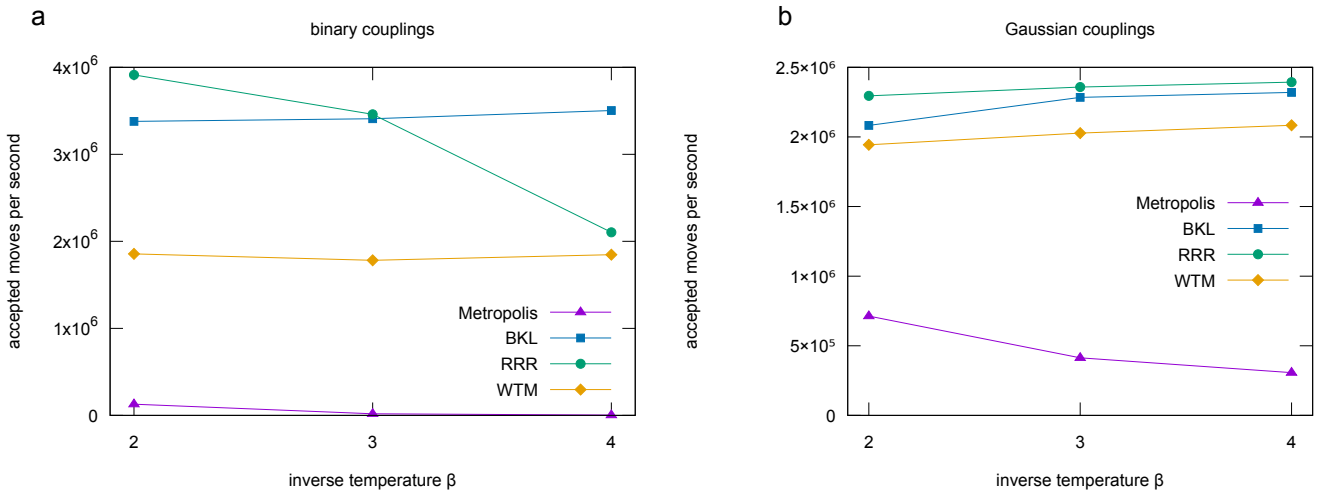

Figure 3. Average moves per second for the data of figure 1, as a function of the inverse temperature β. Error bars are smaller than the data points. RRR becomes worse than BKL only at low temperatures for the binary couplings case.

where $\sigma_{ik} \in \{-1, 1\}$ are classical spins, with $i \in \{1, \ldots, N\}$, $k \in \{1, \ldots, M\}$ and the periodic condition $\sigma_{i(M+1)} \equiv \sigma_{i1}$, and where $\gamma = \frac{1}{2\beta} \log\left(\coth\left(\frac{\beta\Gamma}{M}\right)\right)$. Therefore, the transformation consists in replicating M times the longitudinal part of the original system (the part of the Hamiltonian which commutes with the $\hat{\sigma}_i^z$), and adding pairwise nearest-neighbor ferromagnetic interactions along the Trotter dimension between the corresponding replicated spins. Hereafter, we refer to the system "slices" at fixed k as to the "Trotter replicas".

If such a system can be sampled efficiently, the average value of the observables can be easily computed. In particular, the average value of the energy density can be computed as:

$$\frac{\left\langle \hat{H} \right\rangle}{N} = \left\langle -\frac{1}{MN} \sum_k \sum_{ij} J_{ij}\sigma_{ik}\sigma_{jk} - \Gamma \left(\cosh(2\beta\gamma) - \left(\frac{1}{MN} \sum_{ki} \sigma_{ik}\sigma_{i(k+1)} \right) \sinh(2\beta\gamma) \right) \right\rangle . \tag{25}$$

The interactions along the Trotter dimension all have connectivity 2, and all couplings are equal: therefore, that part of the Hamiltonian H_{eff} describes a part of the model with small connectivity, and we can thus apply the RRR method (and since the energy levels are discrete, we can use the specialized discrete version for this case). It is interesting to notice that, at small values of Γ, the corresponding transverse interactions γ diverge (as $\Gamma \to 0$ the system becomes classical and the replicas collapse); therefore, in that regime, these interactions dominate, and accounting for them through a prior rather than the rejection rate seems particularly promising. The BKL and WTM methods, on the other hand, would need to take into account all the connections of each spin, which are extensive, and are thus those methods are not efficient in this case (we verified that they are indeed orders of magnitude slower than standard Metropolis in the settings we tested).

This formalism is not only used to study quantum physical systems (see e.g. [18, 19]), but it is also at the basis of some quantum annealing optimization techniques, in which the main idea is that the system can be made to escape local minima of the energy landscape via the tunneling effect by introducing a transverse field in an otherwise classical problem, rather than by thermal fluctuations. The general scheme is to simulate a system at low temperature but with an initially strong transverse field Γ, which is gradually lowered to 0 in order to recover the original classical system, hopefully in a low-energy configuration. The usage of Monte Carlo sampling with the Suzuki-Trotter scheme in this context is a well-known technique, see e.g. [7, 20, 21].

We tested this approach on a Sherrington-Kirkpatrick (SK) fully-connected model with random binary couplings $J_{ij} \in \left\{-1/\sqrt{N}, 1/\sqrt{N}\right\}$, $N = 1024$ at fixed $\beta = 2$ and $\Gamma = 0.3$, measuring the value of the "instantaneous Hamiltonian density estimator" (the term between angle brackets in eq. (25)) as a function of time. The results, shown in figure 4, clearly indicate that the RRR method can equilibrate much faster than standard Metropolis, and that the gain increases with larger values of the Trotter dimension M. The acceptance rate increases with M for RRR because of the scaling factor M^{-1} in the first term of eq. (24), while the second term is accounted for by the prior.

A more thorough exploration of the characteristics of the RRR method on this kind of models, and a comparison with alternative Monte Carlo algorithms specifically designed for this purpose like e.g. [22], is reserved for a future

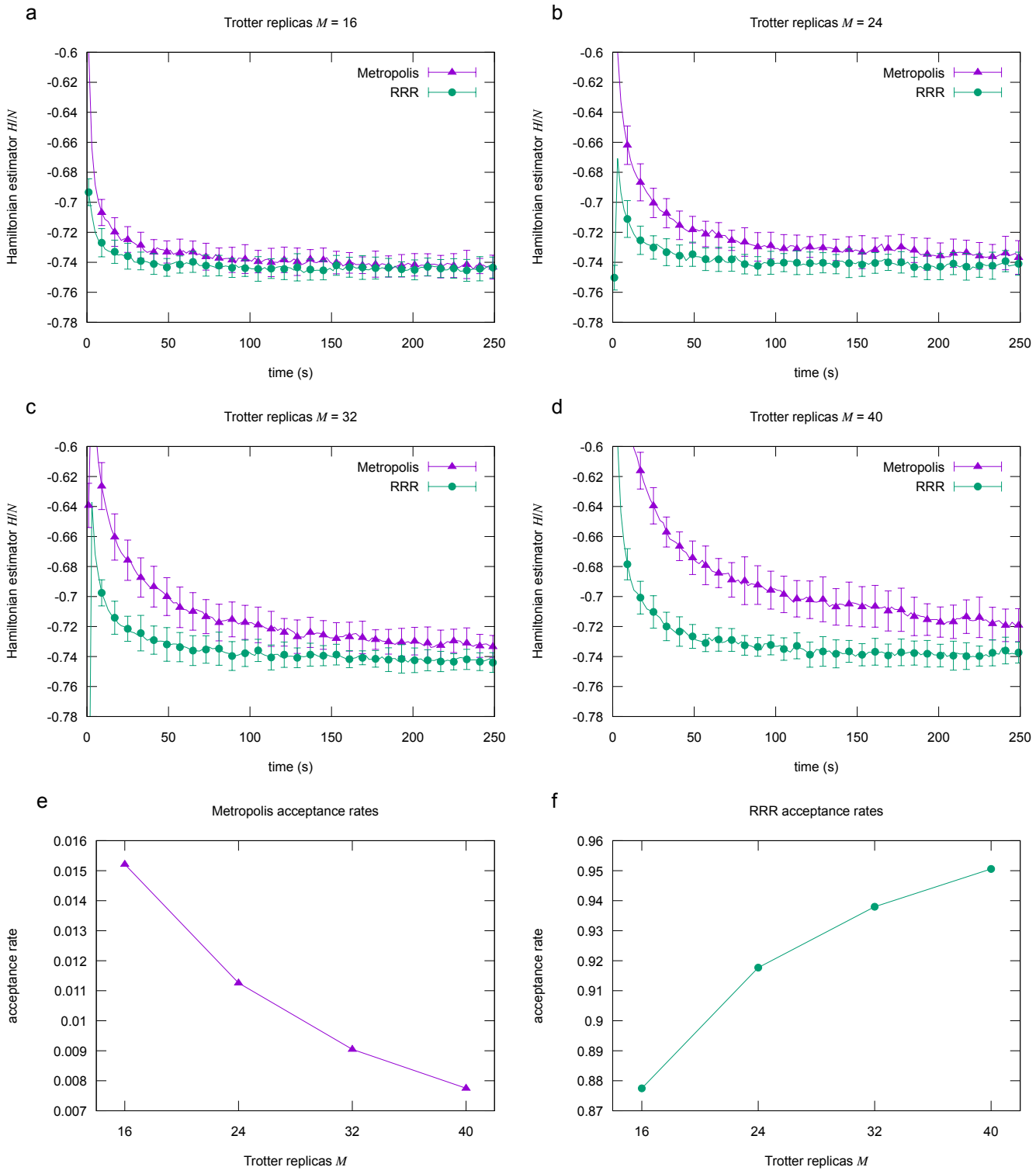

Figure 4. Panels **a-d**: Instantaneous Hamiltonian density estimation as a function of time for a fully connected quantum spin model with random binary couplings and a transverse magnetic field (see text for details), using different values of the Trotter replica. The simulations were performed on SK models with $N = 1024$ spins at fixed $\beta = 2$ and $\Gamma = 0.3$. The main plots compare standard Metropolis and RRR, showing means and standard deviations (after binning) on 20 models, one run per each algorithm. Note that the plotted quantity is the argument of the average in eq. (25), not the energy which governs the Monte Carlo process eq. (24), which accounts for the non-monotonic behavior observed in the initial iterations steps in some panels. RRR clearly converges to equilibrium much faster. Panels **e-f**: the acceptance rate for the two algorithms, as a function of the number of Trotter replicas M: not only RRR has a much higher acceptance rate (even in terms of accepted moves per second, RRR is at least a factor of 3 larger), but the behavior is opposite for the two algorithms.

work.

C. Robust Ensemble tests

In a final set of experiments, we applied the RRR method to the so-called robust ensemble (RE), recently introduced in [2]. For a given Ising model with energy $E(\sigma)$, its RE Hamiltonian at a given inverse temperature β is defined on a set of M interacting replicas of the original system:

$$H\left(\{\sigma^a\}_{a=1}^M;\beta,\gamma\right)=\sum_{a=1}^{M}E(\sigma^a)-\frac{1}{\beta}\sum_{i=1}^{N}\log\left(2\cosh\frac{\gamma}{2}\sum_{a=1}^{M}\sigma_i^a\right) \tag{26}$$

where γ is a control parameter. This Hamiltonian introduces a measure which, compared to the original Gibbs distribution on $E(\sigma)$, enhances the statistical weight of regions of the configuration space with a large free entropy (i.e., roughly speaking, extensive regions in which an exponential number of configurations have low energy), with the parameter γ indirectly controlling the scale of the region (larger γ values bring the focus to narrower regions). One typical use of this measure is to use it within a "scoping" procedure, in which γ is gradually increased. Therefore, the large-γ regime is of particular interest for this problem. In terms of observables, one of the main quantities of interest is the mean energy density across replicas, defined as:

$$\frac{\bar{E}}{N}=\left\langle\frac{1}{M}\sum_{a=1}^{M}E(\sigma^a)\right\rangle. \tag{27}$$

This system is formally similar to the Quantum Monte Carlo of the previous section,[5] and we can use the RRR method to almost entirely account for the effect of the interaction between replicas (the second term in eq. (26)). Note however that for this model the topology of this interaction is different (it's fully-connected rather than a loop), and methods based on flipping entire clusters of spins along a replicated dimensions such as [22] are not applicable.

As in the previous section, we tested this approach on a replicated Sherrington-Kirkpatrick fully-connected model with random binary couplings $J_{ij}\in\left\{-1/\sqrt{N},1/\sqrt{N}\right\}$, using $N=1024$, $M=5$ replicas at fixed $\beta=0.4$ and varying γ. The results are shown in figure 5. As expected, RRR is able to absorb most of the effect of the interaction in the prior, and therefore its acceptance rate is almost constant up to very large values of γ, while that of standard Metropolis drops dramatically. This is true both in terms of accepted moves per attempted spin flip (figure 5e) and of accepted moves per second (figure 5f). As a result, RRR is clearly advantageous in this large-γ regime.

VI. CONCLUSIONS

We have presented a Monte Carlo Markov Chain method, called reduced-rejection-rate Monte Carlo (RRR), which extends the realm of applicability of rejection-free methods: by transforming a kinetic Monte Carlo approach into a choice for a prior, we were able to show that it is possible to improve over the performance of a naïve Metropolis scheme by reducing the rejection rate on a class of models. While rejection-free methods such as BKL and WTM are effective (at low temperatures) for models with low connectivity, the RRR method can also be applied to models in which only a component of the Hamiltonian has that characteristic. For such models, the RRR method can, in many cases, almost remove the rejection rate associated with that component.

We demonstrated this by numerical experiments on Ising spin models, first by showing that RRR indeed nearly eliminates the rejection rate on sparse models (and is thus almost equivalent to BKL) in a wide range of regimes; then by applying it to two models (quantum models and robust-ensemble models) in which only a part of the Hamiltonian is sparse, and showing that the reduction of the rejection rates leads to improved dynamics with respect to a naïve Metropolis scheme when the sparse part dominates. The experiments were mostly exploratory and aimed at demonstrating the feasibility of the method, and are by no means exhaustive: further work is needed (and planned) to employ this method in practically relevant applications and as a component within more general algorithmic schemes (e.g. simulated annealing or parallel tempering). Since the code is generic and extensible, and publicly available, it will also easily allow for a more wide-range exploration of the effectiveness of the technique to other type of sampling and optimization problems.

---

[5] This remarkable fact will be the subject of a future separate work.

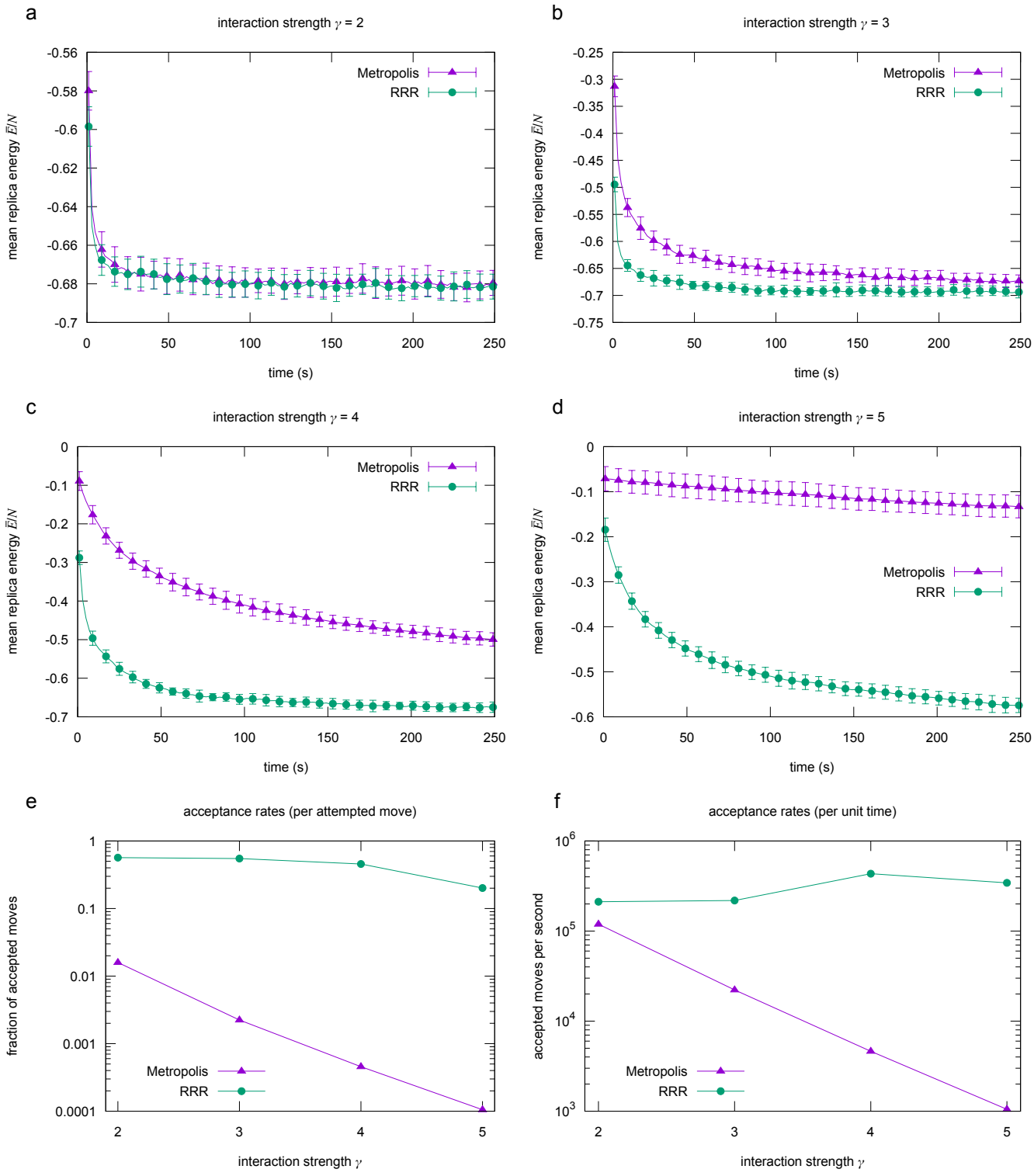

Figure 5. Panels **a-d**: Mean energy density across replicas (eq. (27)) as a function of time with robust ensemble distribution (see text for details), using different values of the interaction parameter γ. The simulations were performed on SK models with $N = 1024$ spins at fixed $\beta = 0.4$ and $M = 5$. The main plots compare standard Metropolis and RRR, showing means and standard deviations (after binning) on 20 models, one run per each algorithm. RRR clearly converges to equilibrium much faster at large γ. Panel **e**: the acceptance rates for the two algorithms, as a function of the interaction parameter γ, in logarithmic scale: RRR is much less affected by γ since it absorbs its effect in the prior. Panel **f**: the number of accepted moves per second for the two algorithms, in logarithmic scale.

The method itself is not restricted to Ising spin models: just like the rejection-free methods it is derived from, it could be straightforwardly applied to models with multiple states per variable (Potts-like models). Of course, this would come at an additional algorithmic cost. In general, the results presented here suggest that such scheme would certainly be convenient whenever the problem is dominated by a sparse component for which the kinetic Monte Carlo approach is better than standard Metropolis, but is also subject to additional dense interactions.

ACKNOWLEDGMENTS

I am deeply indebted to G. Parisi for some crucial observations on an early draft of this manuscript, to R. Zecchina for suggestions and discussions, and to the anonymous referees that reviewed this work for their precious comments. I also wish to thank M. Zamparo and A. Pagnani for helpful suggestions and discussions.

[1] A.B. Bortz, M.H. Kalos, and J.L. Lebowitz. A new algorithm for Monte Carlo simulation of Ising spin systems. *Journal of Computational Physics*, 17(1):10–18, January 1975. doi:10.1016/0021-9991(75)90060-1.

[2] Carlo Baldassi, Christian Borgs, Jennifer T. Chayes, Alessandro Ingrosso, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina. Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes. *Proceedings of the National Academy of Sciences*, 2016. doi:10.1073/pnas.1608103113.

[3] RRR Monte Carlo code. https://github.com/carlobaldassi/RRRMC.jl.

[4] Jesper Dall and Paolo Sibani. Faster Monte Carlo simulations at low temperatures. The waiting time method. *Computer Physics Communications*, 141(2):260–267, November 2001. doi:10.1016/S0010-4655(01)00412-X.

[5] Werner Krauth. *Statistical mechanics: algorithms and computations*. Number 13 in Oxford master series in physics Statistical, computational, and theoretical physics. Oxford Univ. Press, Oxford, 2006. OCLC: 254061236.

[6] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, May 1983. doi:10.1126/science.220.4598.671.

[7] G. E. Santoro. Theory of Quantum Annealing of an Ising Spin Glass. *Science*, 295(5564):2427–2430, March 2002. doi:10.1126/science.1068774.

[8] Arthur F. Voter. INTRODUCTION TO THE KINETIC MONTE CARLO METHOD. In Kurt E. Sickafus, Eugene A. Kotomin, and Blas P. Uberuaga, editors, *Radiation Effects in Solids*, volume 235, pages 1–23. Springer Netherlands, Dordrecht, 2007. doi:10.1007/978-1-4020-5295-8_1.

[9] Jian-Sheng Wang and Robert H. Swendsen. Cluster Monte Carlo algorithms. *Physica A: Statistical Mechanics and its Applications*, 167(3):565–579, September 1990. doi:10.1016/0378-4371(90)90275-W.

[10] Ulli Wolff. Collective Monte Carlo Updating for Spin Systems. *Physical Review Letters*, 62(4):361–364, January 1989. doi:10.1103/PhysRevLett.62.361.

[11] David J. Earl and Michael W. Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910, 2005. doi:10.1039/b509983h.

[12] David P. Landau and K. Binder. *A guide to Monte Carlo simulations in statistical physics*. Cambridge University Press, Cambridge; New York, 3rd ed edition, 2009. OCLC: ocn444371657.

[13] M. Suzuki. Relationship between d-Dimensional Quantal Spin Systems and (d+1)-Dimensional Ising Systems: Equivalence, Critical Exponents and Systematic Approximants of the Partition Function and Spin Correlations. *Progress of Theoretical Physics*, 56(5):1454–1469, November 1976. doi:10.1143/PTP.56.1454.

[14] C. K. Wong and M. C. Easton. An Efficient Method for Weighted Sampling without Replacement. *SIAM Journal on Computing*, 9(1):111–113, February 1980. doi:10.1137/0209009.

[15] Yossi Matias, Jeffrey Scott Vitter, and Wen-Chun Ni. Dynamic Generation of Discrete Random Variates. *Theory of Computing Systems*, 36(4):329–358, August 2003. doi:10.1007/s00224-003-1078-6.

[16] R. N. Bhatt and A. P. Young. Numerical studies of Ising spin glasses in two, three, and four dimensions. *Physical Review B*, 37(10):5606–5614, April 1988. doi:10.1103/PhysRevB.37.5606.

[17] Florent Krzakala, Alberto Rosso, Guilhem Semerjian, and Francesco Zamponi. Path-integral representation for quantum spin models: Application to the quantum cavity method and Monte Carlo simulations. *Physical Review B*, 78(13), October 2008. doi:10.1103/PhysRevB.78.134428.

[18] M. Suzuki, S. Miyashita, and A. Kuroda. Monte Carlo Simulation of Quantum Spin Systems. I. *Progress of Theoretical Physics*, 58(5):1377–1387, November 1977. doi:10.1143/PTP.58.1377.

[19] Tota Nakamura and Yoichi Nishiwaki. Spin-lattice model of magnetoelectric transitions in RbCoBr 3. *Physical Review B*, 78(10), September 2008. doi:10.1103/PhysRevB.78.104422.

[20] Roman Martoňák, Giuseppe E. Santoro, and Erio Tosatti. Quantum annealing of the traveling-salesman problem. *Physical Review E*, 70(5), November 2004. doi:10.1103/PhysRevE.70.057701.

[21] Demian A. Battaglia, Giuseppe E. Santoro, and Erio Tosatti. Optimization by quantum annealing: Lessons from hard satisfiability problems. *Phys. Rev. E*, 71:066707, Jun 2005. doi:10.1103/PhysRevE.71.066707.

[22] Tota Nakamura. Efficient Monte Carlo Algorithm in Quasi-One-Dimensional Ising Spin Systems. *Physical Review Letters*, 101(21), November 2008. doi:10.1103/PhysRevLett.101.210602.