J Stat Phys (2011) 144:541–553
DOI 10.1007/s10955-011-0249-0

Monte Carlo Methods for Rough Free Energy Landscapes: Population Annealing and Parallel Tempering

J. Machta · R.S. Ellis

Received: 7 April 2011 / Accepted: 12 June 2011 / Published online: 30 June 2011
© Springer Science+Business Media, LLC 2011

Abstract Parallel tempering and population annealing are both effective methods for simulating equilibrium systems with rough free energy landscapes. Parallel tempering, also known as replica exchange Monte Carlo, is a Markov chain Monte Carlo method while population annealing is a sequential Monte Carlo method. Both methods overcome the exponential slowing associated with high free energy barriers. The convergence properties and efficiencies of the two methods are compared. For large systems, population annealing is closer to equilibrium than parallel tempering for short simulations. However, with respect to the amount of computation, parallel tempering converges exponentially while population annealing converges only inversely. As a result, for sufficiently long simulations parallel tempering approaches equilibrium more quickly than population annealing.

Keywords Monte Carlo methods · Parallel tempering · Population annealing

1 Introduction

Equilibrium systems with rough free energy landscapes, such as spin glasses, configurational glasses and proteins, are difficult to simulate using conventional Monte Carlo methods because the simulation tends to be trapped in metastable states and fails to explore the full configuration space. A number of techniques have been proposed to reduce this problem. Some of these techniques involve simulating an extended state space that includes many temperatures [1, 2]. Multicanonical simulations, umbrella sampling, the Wang Landau method, simulated tempering and parallel tempering all fall into this class. Parallel tempering [3–6], also known as replica exchange Monte Carlo, is perhaps the most widely used of

J. Machta (✉)
Physics Department, University of Massachusetts, Amherst, MA 01003, USA
e-mail: machta@physics.umass.edu

R.S. Ellis
Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003, USA
e-mail: rsellis@math.umass.edu

these methods because it is simple to program and performs well in many settings. It is the standard method for simulating spin glasses [7, 8] and is used for protein folding [9, 10] and lattice gauge theory [11].

Parallel tempering and the other members of its class are all Markov chain Monte Carlo methods. In Markov chain Monte Carlo, the target distribution is approached via repeated application of an elementary process that typically satisfies detailed balance with respect to the target distribution. In the case of parallel tempering, the target distribution is a joint distribution whose marginals are equilibrium distributions for a set of temperatures. In sequential Monte Carlo, by contrast, the target distribution is the last member of a sequence of distributions, each of which is visited once. The initial distribution is easy to equilibrate and a resampling step transforms one distribution to the next in the sequence. Population annealing [12, 13] is a sequential Monte Carlo algorithm in which the sequence of distributions are equilibrium distributions of decreasing temperature.

In this paper we will describe both parallel tempering and population annealing and compare their efficiencies and convergence properties in the context of a simple, tractable free energy landscape comprised of two wells separated by a high barrier. Although this free energy landscape is highly simplified compared to the landscapes of more realistic models, we believe that it captures some of the essential features of rough free energy landscapes and that the lessons learned from this analysis will be useful in understanding and improving the performance of both parallel tempering and population annealing in realistic settings. In Ref. [14] we analyzed the performance of parallel tempering for this landscape.

Although they are based on quite different strategies, parallel tempering (PT) and population annealing (PA) share a number of common features. Both are methods that build on a conventional Markov chain Monte Carlo procedure whose stationary distribution is a fixed temperature equilibrium ensemble, such as the Metropolis algorithm or Glauber dynamics. We refer to this procedure as the *equilibrating subroutine*. At sufficiently high temperature the equilibrating subroutine converges rapidly to the equilibrium ensemble. Both PT and PA take advantage of this rapid equilibration at high temperature to accelerate the convergence to equilibrium at lower temperatures. Both PT and PA attempt to transform equilibrium high temperature configurations into equilibrium low temperature configurations through a sequence of temperature steps such that the system remains close to equilibrium. In PT there is a single replica of the system at each temperature in the sequence, and replicas are allowed to move between temperatures via replica exchange. These replica exchange moves are carried out with acceptance probabilities that satisfy detailed balance so that the entire set of replicas tends toward equilibrium at their respective temperatures.

Population annealing is closely related to simulated annealing [15, 16]. In simulated annealing a single realization of the system is cooled from high to low temperature following an *annealing schedule*. After each temperature step the system is out of equilibrium and the equilibrating subroutine is used to move it back toward equilibrium. However, at low temperatures, the equilibrating subroutine is unable to facilitate transitions between different minima of the free energy landscape, and simulated annealing falls out of equilibrium if the weights associated with the free energy minima vary with temperature, as is typically the case. Thus simulated annealing cannot be used to sample equilibrium ensembles, and its primary use is to find ground states. Population annealing solves this problem by simultaneously cooling a population of replicas of the system through a sequence of temperatures. Each temperature step is associated with a resampling of the population so that some replicas are copied and other replicas are destroyed in such a way that the replicas are correctly weighted in the colder ensemble. In this way, at least for large populations, the population remains close to equilibrium as the system is cooled. The resampling step in population

annealing is similar to methods used in diffusion Monte Carlo [17] and the "go with the winner" strategy [18].

Sequential Monte Carlo methods [19], of which population annealing is an example, are not well known in statistical physics but have been widely applied in statistics and finance. One purpose of this paper is to bring this general method to the attention of computational statistical physicists. We argue that PA may have an important role to play in simulations of systems with rough free energy landscapes and has some advantages over parallel tempering, especially in situations where a moderately accurate result is required quickly and parallel computing resources are available.

The outline of the paper is as follows. We describe population annealing in Sect. 2 and parallel tempering in Sect. 3. Section 4 introduces the two-well free energy landscape, and Sect. 5 analyzes the performance of population annealing in this landscape. Section 6 compares the performance of population annealing and parallel tempering, and the conclusions of this section are supported by numerical results presented in Sect. 7. Section 8 concludes the paper with a discussion.

2 Population Annealing

The population annealing algorithm operates on a population of R replicas of the system. For disordered spin systems, each replica has the same set of couplings. The algorithm consists of cooling the population of replicas through an annealing schedule from a high temperature, where equilibrium states are easily sampled, to a low target temperature, where the equilibrating subroutine cannot by itself feasibly equilibrate the system. The annealing schedule is defined by a set of S inverse temperatures,

$$\beta_0 > \beta_1 > \cdots > \beta_{S-1}. \tag{1}$$

The highest temperature, $1/\beta_{S-1}$, is chosen to be a temperature for which it is easy to equilibrate the system. It is often convenient to choose $\beta_{S-1} = 0$ as this facilitates the calculation of the absolute free energy of the system at each temperature in the annealing schedule.

In each temperature step the population is resampled and the equilibrating subroutine is applied to every replica at the new temperature. The first part of a temperature step from β to β' is resampling the population so that lower energy replicas are multiplied while higher energy replicas are eliminated from the population. Suppose that the population is in equilibrium at β; the relative weight of a replica j with energy E_j at inverse temperature β' is given by $\exp[-(\beta' - \beta)E_j]$. Thus, the expected number of copies of replica j that appear in the resampled population at β' is

$$\rho_j(\beta, \beta') = \frac{\exp[-(\beta' - \beta)E_j]}{Q(\beta, \beta')}, \tag{2}$$

where Q is the normalization given by

$$Q(\beta, \beta') = \frac{\sum_{j=1}^{R} \exp[-(\beta' - \beta)E_j]}{R}. \tag{3}$$

The new population of replicas is generated by resampling the original population such that the expected number of copies of replica j is ρ_j. The actual number of copies $n_1, n_2, \ldots, n_R$ of each replica in the new population is given by the multinomial distribution

$p[R; n_1, \ldots, n_R; \rho_1/R, \ldots, \rho_R/R]$ for R trials. In this implementation the population size is fixed. Other valid resampling methods are available. For example, the number of copies of replica j can be chosen as a Poisson random variable with mean proportional to $\rho_j(\beta, \beta')$, in which case the population size fluctuates [13].

For large R and small $(\beta' - \beta)$, the resampled distribution is close to an equilibrium ensemble at the new temperature, β'. However, the regions of the equilibrium distribution for β' that differ significantly from the equilibrium distribution for β are not well sampled, leading to biases in the population at β'. In addition, due to resampling, the replicas are no longer independent. To mitigate both of these problem, the equilibrating subroutine is now applied. Finally, observables are measured by averaging over the population.

The entire algorithm consists of $S - 1$ steps: in step k the temperature is lowered from β_{S-k} to β_{S-k-1} via resampling followed by the application of the equilibrating subroutine and data collection at temperature β_{S-k-1}.

Population annealing permits one to estimate free energy differences. If the annealing schedule begins at infinite temperature corresponding to $\beta_{S-1} = 0$, then it yields an estimate of the absolute free energy $\tilde{F}(\beta_k)$ at every temperature in the annealing schedule. The following calculation shows that the normalization factor $Q(\beta, \beta')$ is an estimator of the ratio of the partition functions at the two temperatures:

$$\begin{aligned}
\frac{Z(\beta')}{Z(\beta)} &= \frac{\sum_\gamma e^{-\beta' E_\gamma}}{Z(\beta)} \\
&= \sum_\gamma e^{-(\beta'-\beta)E_\gamma} \left(\frac{e^{-\beta E_\gamma}}{Z(\beta)} \right) \\
&= \langle e^{-(\beta'-\beta)E_\gamma} \rangle_\beta \\
&\approx \frac{1}{R} \sum_{j=1}^{R} e^{-(\beta'-\beta)E_j} = Q(\beta, \beta'). \qquad (4)
\end{aligned}$$

The summation over γ is a sum over the microstates of the system while the sum over j is a sum over the population of replicas in PA. The last approximate equality becomes exact in the limit $R \to \infty$. From (4) the estimated free energy difference from β to β' is found to be

$$-\beta' \tilde{F}(\beta') = -\beta F(\beta) + \log Q(\beta, \beta'), \qquad (5)$$

where $F(\beta)$ is the free energy at β and $\tilde{F}$ is the estimated free energy at β'. Given these free energy differences, if $\beta_S = 0$, then the PA estimator of the absolute free energy at each simulated temperature is

$$-\beta_k \tilde{F}(\beta_k) = \sum_{\ell=k+1}^{S-1} \log Q(\beta_\ell, \beta_{\ell-1}) + \log \Omega, \qquad (6)$$

where $\Omega = \sum_\gamma 1$ is the total number of microstates of the system; i.e., $k_B \log \Omega$ is the infinite temperature entropy.

3 Parallel Tempering

Parallel tempering, also known as replica exchange Monte Carlo, simultaneously equilibrates a set of R replicas of a system at S inverse temperatures

$$\beta_0 > \beta_1 > \cdots > \beta_{S-1}. \tag{7}$$

There is one replica at each temperature so that $R = S$ in contrast to population annealing, where typically the number R of replicas greatly exceeds the number S of temperatures; i.e., $R \gg S$. The equilibrating subroutine operates on each replica at its respective temperature. Replica exchange moves are implemented that allow replicas to diffuse in temperature space. The first step in a replica exchange move is to propose a pair of replicas $(k, k-1)$ at neighboring temperatures $\beta = \beta_k$ and $\beta' = \beta_{k-1}$. The probability for accepting the replica exchange move is

$$p_{\text{swap}} = \min[1, e^{(\beta-\beta')(E-E')}]. \tag{8}$$

Here E and E' are the respective energies of the replicas that were originally at β and β'. If the move is accepted, the replica equilibrating at β is now set to equilibrate at β' and vice versa. Equation (8) insures detailed balance so that the Markov chain defined by parallel tempering converges to a joint distribution whose marginals are equilibrium distributions at the S temperatures of (7). Diffusion of replicas in temperature space allows round trips from low to high temperature and back. The benefit of these roundtrips is that free energy barriers are crossed in a time that grows as a power of the barrier height [14] rather than exponentially with respect to the barrier height as is the case for most single temperature dynamics. Optimization schemes for PT depend in part on adjusting parameters to maximize the rate of making roundtrips [20–22].

4 Two-Well Model Free Energy Landscape

In this section we describe a simple free energy landscape with two minima such as occurs, for example, in the low temperature phase of the Ising model or ϕ^4 field theories. This free energy was introduced in Ref. [14] in the context of analyzing the efficiency of parallel tempering. For $\beta \geq \beta_c$ and β_c a critical temperature, the free energy $F_\sigma(\beta)$ associated with each well is defined by

$$\beta F_\sigma(\beta) = -\frac{1}{2}(\beta - \beta_c)^2 \Delta_\sigma^2, \tag{9}$$

where

$$\Delta_\sigma^2 = \begin{cases} K + H/2 & \text{if } \sigma = 1, \\ K - H/2 & \text{if } \sigma = 0, \end{cases} \tag{10}$$

σ labels the wells, K controls the well depth and H controls the asymmetry between the wells. The deep well corresponds to $\sigma = 1$ and the shallow well to $\sigma = 0$. The well index σ is the only macroscopic parameter in the model and the "landscape" is zero dimensional. However, we also assume that the free energy at the saddle point between the wells is zero so that F_1 is the (larger) free energy barrier between the wells. The landscape is flat at $\beta = \beta_c$. The parameter K is a proxy for system size. In more realistic systems, barrier heights typically grow as a power of the number of degrees of freedom N of the system.

The statistics of the energy of microstates in each well follows from this free energy using thermodynamics. The internal energy $U_\sigma(\beta)$ is the average of the energy distribution in well σ and is obtained from

$$U_\sigma(\beta) = \frac{\partial \beta F_\sigma}{\partial \beta} = -(\beta - \beta_c)\Delta_\sigma^2. \tag{11}$$

Using the relationship between specific heat and energy fluctuations, we find that the variance of the energy in well σ is simply Δ_σ^2.

The free energy also determines the probability $c(\beta)$ of being in the deep well according to

$$c(\beta) = \mathbf{E}(\sigma) = \frac{1}{1 + e^{-(\beta-\beta_c)^2 H}}, \tag{12}$$

where $\mathbf{E}$ indicates expectation. Microstates of the model are specified by an energy and a well index. We assume that in equilibrium the distribution of energies, conditioned on the well index σ, is a normal distribution with mean $U_\sigma(\beta)$ and variance Δ_σ^2. Thus, the well index σ is a Bernoulli random variable such that $\sigma = 1$ with probability $c(\beta)$ and $\sigma = 0$ with probability $1 - c(\beta)$. The energy E is given by

$$E = \begin{cases} N(U_1(\beta), \Delta_1^2) & \text{if } \sigma = 1, \\ N(U_0(\beta), \Delta_0^2) & \text{if } \sigma = 0, \end{cases} \tag{13}$$

where $N(\mu, \Delta^2)$ is a normal random variable with mean μ and variance Δ^2.

The dynamics of the model under the equilibrating subroutine for $\beta \le \beta_c$ is assumed to have the following properties. The well index is conserved except at the critical temperature, β_c. That is, there are no transitions between the wells except for $\beta = \beta_c$. On the other hand, for $\beta > \beta_c$ the equilibrating subroutine is assumed to equilibrate the system within each well in a single time unit. Thus, the sequence of energies produced by successive steps of the equilibrating subroutine will be i.i.d. normal random variables $N(U_\sigma(\beta), \Delta_\sigma^2)$, where σ is the well index. For $\beta = \beta_c$ the equilibrating subroutine first chooses one of the wells with equal probability and then chooses the energy from the associated normal distribution.

5 Convergence of Population Annealing

We first consider a single step of population annealing from inverse temperature β to inverse temperature $\beta' > \beta$. We will compute the error made by population annealing in the free energy and the fraction of the population in the deep well as a function of the number of replicas, the size of the temperature step $\beta' - \beta$, and the parameters β_c, K and H of the two-well model. Let

$$Y_j = \exp[-(\beta' - \beta)E_j + \lambda\sigma_j], \tag{14}$$

where E_j is the energy and σ_j is the well index of replica j. Setting $\lambda = 0$ yields the unnormalized re-weighting factor (see (2)) of replica j from inverse temperature β to β'. The extra term $\lambda\sigma_j$ in the exponent will be used to calculate the probability of being in the deep well.

To obtain the error in the free energy and the fraction of replicas in the deep well at temperature β' assuming the correct equilibrium distribution at β, we compute $I(\beta, \beta', \lambda)$,

which is defined by:

$$I(\beta, \beta', \lambda) = \mathbf{E} \log\left(\frac{1}{R}\sum_{j=1}^{R} Y_j\right). \tag{15}$$

Using (5), we obtain the PA estimate of the free energy difference by setting $\lambda = 0$ in $I(\beta, \beta', \lambda)$:

$$\beta' \tilde{F}(\beta') - \beta F(\beta) \approx -I(\beta, \beta', 0) = -\mathbf{E}(\log Q(\beta', \beta)). \tag{16}$$

In this equation, $\beta' \tilde{F}(\beta')$ is the estimate of the free energy at β' given the exact value at β. The estimate of the fraction in the deep well at the lower temperature $\tilde{c}(\beta')$ is obtained by differentiation with respect to λ:

$$\tilde{c}(\beta') \approx \left.\frac{dI(\beta, \beta', \lambda)}{d\lambda}\right|_{\lambda=0}. \tag{17}$$

Our goal is to determine how much these estimates deviate from the corresponding exact values.

Let S_n be a sum of n independent, identically distributed random variables X_j. Generically, one can use Taylor's Theorem to prove that the leading terms in an asymptotic expansion of the expectation of a function $f(S_n/n)$ have the form

$$\mathbf{E} f(S_n/n) = \mathbf{E} f\left(\sum_{j=1}^{n} X_j/n\right) = f(\mathbf{E}X) + \frac{1}{2n} f''(\mathbf{E}X)\mathbf{Var}\, X + \mathrm{O}\left(\frac{1}{n^{3/2}}\right), \tag{18}$$

where $\mathbf{E}(X)$ and $\mathbf{Var}(X)$ denote the expectation and variance of X_j, respectively. Thus, for our case,

$$I(\beta, \beta', \lambda) = \log(\mathbf{E}Y) - \frac{1}{2R}\frac{\mathbf{Var}\, Y}{(\mathbf{E}Y)^2} + \mathrm{O}\left(\frac{1}{R^{3/2}}\right). \tag{19}$$

The first term is the exact result and the second term is the leading order systematic error in the population annealing estimate due to a finite population size. Setting $\lambda = 0$, we have

$$\beta' \tilde{F}(\beta') - \beta' F(\beta') = \frac{1}{2R}\frac{\mathbf{Var}\, Y}{(\mathbf{E}Y)^2} + \mathrm{O}\left(\frac{1}{R^{3/2}}\right). \tag{20}$$

This result shows that the systematic error decreases as the inverse of the population size and that the free energy approaches the exact value from above as the number of replicas increases.

The variance of the free energy estimator was observed to be a useful measure of the convergence of the algorithm [13]. Here we formalize that observation by computing the variance of $\beta' \tilde{F}(\beta')$ presuming that $\beta F(\beta)$ is exactly known and thus has no variance. The variance of the free energy estimator is given by,

$$\mathbf{Var}(\beta' \tilde{F}(\beta')) = \mathbf{E}\left(\log^2 \frac{1}{R}\sum_{j=1}^{R} e^{-(\beta'-\beta)E_j}\right) - \left(\mathbf{E} \log \frac{1}{R}\sum_{j=1}^{R} e^{-(\beta'-\beta)E_j}\right)^2. \tag{21}$$

Applying (18) to both log and $\log^2$ and substituting the results into (21) yields

$$\mathbf{Var}(\beta' \tilde{F}(\beta')) = \frac{1}{R}\frac{\mathbf{Var}\, Y}{(\mathbf{E}Y)^2} + \mathrm{O}\left(\frac{1}{R^{3/2}}\right), \tag{22}$$

where λ is set to zero on the RHS of this equation. Comparing (22) and (20), we find that

$$\beta'\tilde{F}(\beta') - \beta' F(\beta') = \frac{1}{2}\mathbf{Var}(\beta'\tilde{F}(\beta')) + \mathrm{O}\left(\frac{1}{R^{3/2}}\right). \tag{23}$$

This equation is useful because $\mathbf{Var}(\beta'\tilde{F}(\beta'))$ can be directly estimated from multiple runs of PA, and thus the accuracy of the algorithm as applied to a specific system can be estimated from the algorithm itself. Although (23) was derived for a single step of PA for the two-well model, the calculation does not rely on specific features of the model. Furthermore, since the variance is additive, we conjecture that (23) is a good approximation for the full PA algorithm applied to any statistical mechanical system. In support of this conjecture, we note that (23) is a good approximation for the PA estimate of the low temperature free energy of the one-dimensional Ising spin glass studied in [14], for which the exact free energy can be calculated using transfer matrix methods.

For the case of the two-well model, we can evaluate the Gaussian integrals exactly for both the mean and variance of the weight factor Y. The general result is

$$\begin{aligned} I(\beta,\beta',\lambda) = {} & \frac{K}{2}[(\beta'-\beta_c)^2 - (\beta-\beta_c)^2] \\ & + \log\cosh\left[(\beta'-\beta_c)^2\frac{H}{4} + \lambda/2\right] - \log\cosh\left[(\beta-\beta_c)^2\frac{H}{4}\right] + \lambda/2 \\ & + \frac{1}{2R} - \frac{1}{2R}\exp[(\beta'-\beta)^2 K] \\ & \times \left(\frac{\cosh[(\beta-\beta_c)^2 H/4]\cosh[(2\beta'-\beta-\beta_c)^2 H/4 + \lambda]}{\cosh[(\beta'-\beta_c)^2 H/4 + \lambda/2]^2}\right) \\ & + \mathrm{O}\left(\frac{1}{R^{3/2}}\right). \end{aligned} \tag{24}$$

From this general result it is instructive to consider expansions to first order in H. The estimate of the free energy difference is

$$\begin{aligned} \beta'\tilde{F}(\beta') - \beta F(\beta) = {} & -\frac{K}{2}[(\beta'-\beta_c)^2 - (\beta-\beta_c)^2] \\ & + \frac{1}{2R}[\exp((\beta'-\beta)^2 K) - 1] + \mathrm{O}\left(\frac{1}{R^{3/2}}\right) + \mathrm{O}(H^2). \end{aligned} \tag{25}$$

The first term on the RHS of this expression is the exact free energy difference in the two-well model for symmetric wells. The second term is the error made by population annealing, which decreases inversely in R. The form of the error term also reveals that the size of the temperature steps should be $(\beta'-\beta) \lesssim 1/\sqrt{K}$ to keep the error under control as the barrier height increases.

The error in the free energy estimate at β_0 for small H can be obtained by summing the errors made in each temperature step of the algorithm. Since the errors depend only on the size of the temperature step, the algorithm is optimized with constant size steps in β. Thus the error estimate at β_0 is

$$\beta_0\tilde{F}(\beta_0) - \beta_0 F(\beta_0) = \frac{S}{2R}(\exp[(\beta_0-\beta_c)^2 K/S^2] - 1) + \mathrm{O}\left(\frac{1}{R^{3/2}}\right) + \mathrm{O}(H^2). \tag{26}$$

Next we consider the fraction of the population in the deep well at temperature β' given the equilibrium value at β. Combining (17) and (24) and expanding to leading order in H, we obtain

$$\tilde{c}(\beta') = \frac{1}{2} + \frac{H(\beta'-\beta_c)^2}{8} - \frac{H}{8R}(\beta'-\beta)(3\beta'-\beta-2\beta_c)\exp[(\beta'-\beta)^2 K] + \mathrm{O}\left(\frac{1}{R^{3/2}}\right) + \mathrm{O}(H^2). \tag{27}$$

The first two terms on the RHS of this expression are the leading order in H expansion of the exact value, (12). The correction term shows that again the temperature steps should satisfy $(\beta'-\beta) \lesssim 1/\sqrt{K}$.

From (27) we could obtain an estimate of the overall error in the fraction in the deep well by summing over the S temperature steps. Unfortunately the result significantly underestimates the true error. The reason is that the resampling step introduces correlations between replicas so that the probability distribution for the number of replicas in the deep well has a variance that is broader than that of the binomial distribution assumed in the above analysis. Nonetheless, we conjecture that the leading term in the error made by the full algorithm in the fraction in the deep well, $\tilde{c}(\beta_0) - c(\beta_0)$, behaves as $1/R$ and can be minimized when $S \sim \sqrt{K}$. The effect of correlations is much less important for the free energy, as evidenced by the absence of a term that is order H in (25), and we conjecture that (26) is exact to leading order. We are currently studying these questions.

The two main conclusion from this analysis are that (1) the error decreases inversely with the number of replicas and (2) the error can be made small only if the temperature step size satisfies $(\beta'-\beta) \lesssim 1/\sqrt{K}$.

6 Parallel Tempering vs. Population Annealing

How do PT and PA compare in the efficiency with which they converge to equilibrium? Here we estimate the amount of computational work, defined below, needed to make the deviation from equilibrium small. The quantities that we increase, holding other parameters fixed, are the number of replicas R for PA and the number of sweeps t for PT. Within the stylized two-well model, we define computational work $\mathcal{W}$ as the total number of times replicas are acted on by the equilibrating subroutine. For PA with R replicas and S temperature steps the work is $\mathcal{W} = RS$. For PT with R replicas the computational work is given by $\mathcal{W} = Rt$ where t is the number of PT sweeps.

Our measure of computational work ignores the time required to resample the population in PA or implement replica exchange in PT. For large systems, this time is negligible compared to the time spent in the equilibrating subroutine. The quantity $\mathcal{W}$ assigns one unit of time for one sweep of the equilibrating subroutine for a single replica; thus the computational work measured in elementary operations rather than sweeps of the equilibrating subroutine is $N\mathcal{W}$, where N is the number of degrees of freedom of the system. Since N is the same for PA and PT and not explicitly defined in the two-well model, we do not consider this factor explicitly.

Suppose we carry out the simulations on a massively parallel computer and consider parallel time instead of sequential time (work). Since each replica can be independently acted on by the equilibrating subroutine, one parallel time unit is required for one sweep of all the replicas. Thus, for the highly parallel PA, the parallel time is the number of temperature steps S whereas for the less parallel PT, the parallel time is the number of PT sweeps t.

In [14] we analyzed the efficiency of PT for the two-well model. As is generally the case for Markov chain Monte Carlo methods, convergence to equilibrium is asymptotically exponential. The deviation from equilibrium is controlled by an exponential autocorrelation time $\tau_{\exp}$, and its leading behavior is proportional to $\exp(-t/\tau_{\exp})$, where t is the number of Monte Carlo sweeps. In the two-well model, for small asymmetries between the wells (i.e., $(\beta_0 - \beta_c)^2 H \leq 1$), $\tau_{\exp}$ is controlled by the diffusive time scale for a replica to diffuse between the lowest and highest temperature. If the number of temperatures is sufficiently large, then the acceptance fraction for replica exchange is not small and the elementary time step in this diffusive process is order unity so that $\tau_{\exp} \sim R^2$. The optimum number of replicas was shown to scale as $R = S + 1 \sim \sqrt{K}$, and given this choice, $\tau_{\exp} \sim R^2 \sim K$. When the asymmetry becomes large, the optimum number of temperatures remains approximately the same but there is a crossover to a ballistic regime, and $\tau_{\exp} \sim R \sim \sqrt{K}$. Convergence to equilibrium occurs on a characteristic time scale $\tau_{\exp}$ so that, in the small-asymmetry diffusive regime, the characteristic computational work scale required for equilibration, $\mathcal{W}_0$, is given by $\mathcal{W}_0 \sim R\tau_{\exp} \sim K^{3/2}$.

For PA we found in (24) that the error term depends on K as $\exp[(\beta' - \beta)^2 K]$ so that the optimum number of temperature steps scales as $S \sim \sqrt{K}$, just as is the case for PT. Based on the form of (27), we conjecture that the overall error behaves as S^a/R with a an exponent less than or equal to unity. Thus, the error decreases as $S^{1+a}/\mathcal{W} \sim K^{(1+a)/2}/\mathcal{W}$, and the computational work $\mathcal{W}_0$ required to begin to achieve moderately accurate results behaves as $\mathcal{W}_0 \sim K^{(1+a)/2}$. For large systems (large K) and nearly degenerate free energy minima, population annealing is expected to be more efficient initially than parallel tempering by a factor of a power of the barrier height, $K^{1-a/2}$. However, for large amounts of computational work (i.e., $\mathcal{W} \gg R\tau_{\exp} \sim K^{3/2}$) PT is much closer to equilibrium than PA because PT converges exponentially in t while PA converges inversely in the comparable variable R.

7 Numerical Results

We have carried out simulations of the two-well model using both PT and PA to compare the efficiency of the algorithms and test the conjectures of the previous section. In these simulations, the equilibrating subroutine samples a Gaussian random number with mean and variance appropriate to the temperature and well-index of the replica, (13). The well index is a conserved quantity except at β_c. At β_c the equilibrating subroutine first chooses the well index with equal probability and then chooses the energy according to (13) with $\beta = \beta_c$ and the given value of σ.

Let γ denote the deviation from equilibrium of the probability of being in the deep well at the lowest temperature given by

$$\gamma = \tilde{c}(\beta_0) - c(\beta_0). \tag{28}$$

Figures 1(a), 1(b) and 1(c) show the convergence to equilibrium of γ, as a function of the number of sweeps t for PT (small points—blue online) or population size R for PA (large points—red online). The horizontal axis therefore measures computational work for both algorithms in the same units. The figures differ according to the value of the well depth K and number of temperatures S with $K = 16$, $S = 11$ for Fig. 1(a), $K = 64$, $S = 23$ for Fig. 1(b) and $K = 256$, $S = 47$ for Fig. 1(c). In each case, there is a small asymmetry, $H = 0.1$ and at the lowest temperature $c(\beta_0) = 0.68997$. The highest and lowest temperatures are $\beta_c = 1$ and $\beta_0 = 5$, respectively, and the number of temperatures $S + 1$ is chosen to be close to

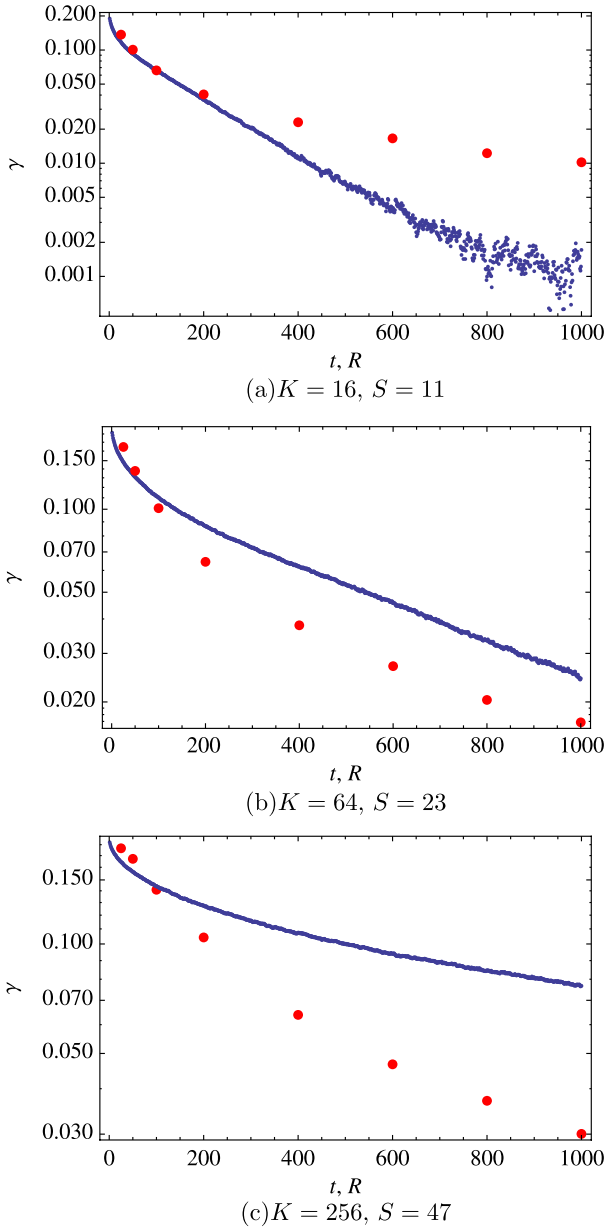

Fig. 1 The deviation of the fraction in the deep well from the equilibrium value γ vs. Monte Carlo sweeps t for PT (*small points—blue online*) or population size R for PA (*large points—red online*) for the values of K and S shown in the subcaptions. For each case $H = 0.1$

the optimum value for PT and to scale as $\sqrt{K}$. The simulations confirm the conclusions of Sect. 6 and show that for larger systems (larger K) PA is closer to equilibrium than PT for short simulations.

We have also tested two conjectures concerning the convergence of PA for the two-well model. The first conjecture is that the error in the fraction in the deep well decreases inversely in the population size. Figure 2 shows $R\gamma$ as a function of R for the case $H = 0.1$, $K = 64$, and $S = 23$. It is clear that within the error bars γ is behaving as $1/R$ over the range

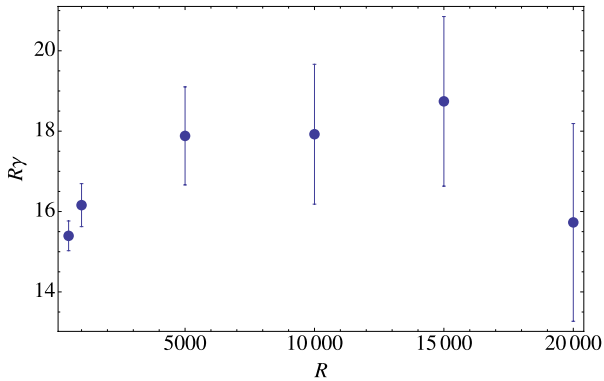

Fig. 2 $R\gamma$ vs. R for $K = 64$, $S = 23$ and $H = 0.1$

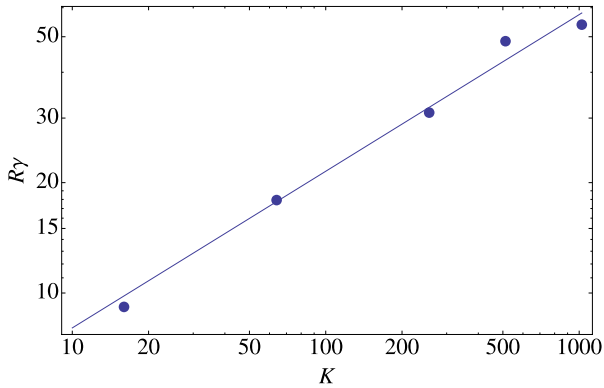

Fig. 3 $R\gamma$ vs. K for $R = 10000$ and $H = 0.1$. The *solid line* is the best power-law fit, $R\gamma = K^{0.43}$

of R studied. The averages and error bars are obtained from 10^5 independent runs of PA for each population size.

The second conjecture is that for large R and $S \sim \sqrt{K}$

$$R\gamma \sim K^{a/2}. \tag{29}$$

Figure 3 is a log-log plot of $R\gamma$ vs. K. For each value of K, PA is run with a large population size $R = 10000$, $H = 0.1$, and the values of S given above satisfying $S \sim \sqrt{K}$. Averages and errors are obtained from 10^5 independent runs. The best fit to (29) is shown as the solid line and yields $a = 0.85 \pm 0.08$. Note that $a \leq 1$ as conjectured, supporting the hypothesis that for modest amounts of computational work, PA yields a smaller error in γ than PT and that this advantage increases with well-depth parameter K.

8 Discussion

We have seen that both parallel tempering and population annealing are able to solve the problem of sampling equilibrium states of systems having several minima in the free energy landscape separated by high barriers. Parallel tempering is a Markov chain Monte Carlo method and thus converges to equilibrium exponentially in the number of sweeps, whereas population annealing converges inversely in the population size. Their relative efficiencies

are described qualitatively by the parable of the tortoise and the hare. For a given amount of computational work, the hare (population annealing) is initially closer to equilibrium but ultimately the tortoise (parallel tempering) catches up and gets ahead.

The problem of high barriers between different free energy minima is only one of the difficulties encountered in simulating systems with rough free energy landscapes. A second, generic problem is that the relevant free energy minima may have small basins of attraction so that when the system is annealed starting from high temperature, it is very unlikely that the relevant low temperature states will be found. Small basins of attraction occur for first order transitions and for NP-hard combinatorial optimization problems and are almost certainly a feature of the low temperature phase of spin glasses. In this situation a very large number of sweeps in parallel tempering or a very large population size in population annealing are required simply to find the relevant states.

Parallel tempering is widely used in several areas of computational physics while population annealing is not well known. One of the conclusions of this paper is that population annealing is an attractive alternative to parallel tempering for studies where moderately accurate answers are required quickly. This is especially the case if massively parallel computing resources are available since population annealing is well suited to parallelization.

Acknowledgements Jon Machta was supported in part from NSF grant DMR-0907235 and Richard Ellis by NSF grant DMS-0604071.

References

1. Okamoto, Y.: J. Mol. Graph. Model. **22**, 425 (2004)
2. Nadler, W., Hansmann, U.H.E.: Phys. Rev. E **75**, 026109 (2007)
3. Swendsen, R.H., Wang, J.-S.: Phys. Rev. Lett. **57**, 2607 (1986)
4. Geyer, C.: In: Keramidas, E.M. (ed.) Computing Science and Statistics: 23rd Symposium on the Interface, p. 156. Interface Foundation, Fairfax Station (1991)
5. Hukushima, K., Nemoto, K.: J. Phys. Soc. Jpn. **65**, 1604 (1996)
6. Earl, D.J., Deem, M.W.: Phys. Chem. Chem. Phys. **7**, 3910 (2005)
7. Katzgraber, H.G., Körner, M., Young, A.P.: Phys. Rev. B **73**, 224432 (2006)
8. Banos, R.A., Cruz, A., Fernandez, L.A., Gil-Narvion, J.M., Gordillo-Guerrero, A., Guidetti, M., Maiorano, A., Mantovani, F., Marinari, E., Martin-Mayor, V., et al.: J. Stat. Mech. Theory Exp. **2010**, P06026 (2010)
9. Hansmann, U.H.E.: Chem. Phys. Lett. **281**, 140 (1997)
10. Schug, A., Herges, T., Verma, A., Wenzel, W.: J. Phys., Condens. Matter **17**, S1641 (2005)
11. Burgio, G., Fuhrmann, M., Kerler, W., Muller-Preussker, M.: Phys. Rev. D **75**, 014504 (2007)
12. Hukushima, K., Iba, Y.: In: Gubernatis, J.E. (ed.) The Monte Carlo Method in the Physical Sciences: Celebrating the 50th Anniversary of the Metropolis Algorithm. AIP Conf. Proc., vol. 690, pp. 200–206. AIP, New York (2003)
13. Machta, J.: Phys. Rev. E **82**, 026704 (2010)
14. Machta, J.: Phys. Rev. E **80**, 056706 (2009)
15. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Science **220**, 671 (1983)
16. van Laarhoven, P., Aarts, E.: Simulated Annealing: Theory and Applications. Reidel, Dordrecht (1987)
17. Anderson, J.B.: J. Chem. Phys. **63**, 1499 (1975)
18. Grassberger, P.: Comput. Phys. Commun. **147**, 64 (2002)
19. Doucet, A., de Freitas, N., Gordon, N. (eds.): Sequential Monte Carlo Methods in Practice. Springer, New York (2001)
20. Katzgraber, H.G., Trebst, S., Huse, D.A., Troyer, M.: J. Stat. Mech. 03018 (2006)
21. Trebst, S., Troyer, M., Hansmann, U.H.E.: J. Chem. Phys. **124**, 174903 (2006)
22. Bittner, E., Nußbaumer, A., Janke, W.: Phys. Rev. Lett. **101**, 130603 (2008)