Analysis and optimization of population annealing

Christopher Amey[*]

Department of Physics, University of Massachusetts, Amherst, Massachusetts 01003, USA

Jonathan Machta[†]

Department of Physics, University of Massachusetts, Amherst, Massachusetts 01003, USA
and Santa Fe Institute, Santa Fe, New Mexico 87501, USA

Population annealing is an easily parallelizable sequential Monte Carlo algorithm that is well suited for simulating the equilibrium properties of systems with rough free-energy landscapes. In this work we seek to understand and improve the performance of population annealing. We derive several useful relations between quantities that describe the performance of population annealing and use these relations to suggest methods to optimize the algorithm. These optimization methods were tested by performing large-scale simulations of the three-dimensional (3D) Edwards-Anderson (Ising) spin glass and measuring several observables. The optimization methods were found to substantially decrease the amount of computational work necessary as compared to previously used, unoptimized versions of population annealing. We also obtain more accurate values of several important observables for the 3D Edwards-Anderson model.

I. INTRODUCTION

Frustrated and disordered systems are ubiquitous in nature and, consequently, appear in a wide range of current research areas. Theoretical approaches to these systems are generally difficult to implement and have limited applicability and, as a result, computational methods are crucially important. Replica exchange Monte Carlo, also known as parallel tempering [1–4] and multicanonical algorithms such as the Wang-Landau algorithm [5,6] have become the standard methods in the field because they partially overcome the problem of rough free-energy landscapes. Perhaps unsurprisingly, the development of these algorithms is fundamentally intertwined with the frustrated systems they were invented to probe, and understanding the behavior of the algorithms contributes to understanding the system itself.

One algorithm that is under active development is the population annealing Monte Carlo method. Population annealing (PA) is an example of a sequential Monte Carlo algorithm [7] and was first introduced and applied to spin glasses by Hukushima and Iba [8]. Population annealing shares features with both parallel tempering and simulated annealing [9]; however, it has one major advantage: It is inherently parallel and can be easily implemented on large computing clusters. Population annealing is closely related to nested sampling [10,11].

Recently there has been a resurgence of interest in PA [12]. There are several recent works that use PA to simulate not only spin glasses [13–16] but also hard-sphere fluids [17] and the ferromagnetic Potts and Ising models using GPUs [18,19]. In this work we continue the analysis of population annealing and derive several analytic results relating to its implementation. These results naturally lead to several optimization ideas that we test in the context of the three-dimensional (3D) Edwards-Anderson (EA) spin glass. Related optimization ideas are explored in Ref. [20].

In Sec. II, we present the population annealing algorithm. We then introduce the Edwards-Anderson spin-glass model and the relevant observables of interest, which is our test bed to study the optimization of PA. In Sec. III we introduce several observables and measures of error specific to population annealing. Furthermore, we derive an important inequality between measures of systematic and statistical errors, and we derive lower bounds for these errors in PA. We also introduce a method to determine an optimized annealing schedule. These relations suggest several improvements to PA, which we describe in detail in Sec. IV. In Sec. V we show the results of a large-scale simulation of the 3D Edwards-Anderson spin glass, which demonstrate our analytic results and the efficacy of the optimizations as well as providing new benchmark values of several quantities.

II. BACKGROUND

Population annealing is a sequential Monte Carlo algorithm designed to sample equilibrium states in systems with rough free-energy landscapes. Population annealing is similar to a parallel version of simulated annealing in many respects: A population of R spin configurations, henceforth called "replicas," is initialized at high temperatures where equilibration is easy, and the population is slowly cooled. In addition, there is a resampling step at each temperature where individual replicas may be copied or eliminated according to their Boltzmann weights. Resampling ensures that the ensemble of replicas

---

[*]camey@physics.umass.edu
[†]machta@physics.umass.edu

is kept near thermal equilibrium. However, the resampled ensemble has degeneracies due to copied replicas and does not fully sample the low-energy spectrum at the lower temperature. In order to address these issues, a Markov Chain Monte Carlo (MCMC) procedure such as the Metropolis algorithm is used to decorrelate and additionally equilibrate the population.

A PA simulation traverses an annealing schedule with N_T inverse temperatures $\{\beta_{N_T-1}, \ldots, \beta_0\}$, where $\beta_{j+1} < \beta_j$, the initial temperature is typically infinite, $\beta_{N_T-1} = 0$, and the final β_0 is chosen as desired. Consider the resampling step from inverse temperature β to β'. Each replica, denoted by a subscript i, is given a resampling factor τ_i that is proportional to the ratio of the Boltzmann weights of that replica at the two temperatures,

$$\tau_i = \frac{R}{\tilde{R}_\beta} \frac{e^{-(\beta'-\beta)E_i}}{Q(\beta,\beta')}, \tag{1}$$

where E_i is the energy of replica i, Q is a normalization factor used to control the population size,

$$Q(\beta,\beta') = \frac{1}{\tilde{R}_\beta} \sum_{i=1}^{\tilde{R}_\beta} e^{-(\beta'-\beta)E_i}, \tag{2}$$

and $\tilde{R}_\beta$ is the population size at step β. The actual population size $\tilde{R}_\beta$ may fluctuate around the target population size R. The resampling factor is the expected number of copies of a replica, that is, $\tau_i = \langle n_i \rangle$, where n_i is the stochastically chosen integer number of copies of a replica i. There are several ways to implement resampling in PA. In this work we follow the method used in Ref. [13], which minimizes the variance of n_i by choosing $n_i(\beta,\beta') = \lfloor \tau_i(\beta,\beta') \rfloor$ with probability $\lceil \tau_i(\beta,\beta') \rceil - \tau_i(\beta,\beta')$ and $n_i(\beta,\beta') = \lceil \tau_i(\beta,\beta') \rceil$ otherwise. This method results in an average population equal to R, with fluctuations of order $\sqrt{R}$. The resampling step ensures that, for large R, if the population is an equilibrium ensemble at β, then it will also be an equilibrium ensemble at β'.

At the beginning of the simulation, each replica is statistically independent; however, after being copied, some replicas become statistically correlated with each other. Correlations are recorded via the "ancestry" of each replica. At the beginning of the simulation, each replica is labeled. During each resampling step, each replica passes its label to its offspring, and replicas that have the same ancestry label are said to be members of the same "family." Although the total number of families is always equal to the initial population size R, by the end of a simulation, most families will have no members. We shall see that the distribution of family sizes is closely related to both statistical and systematic errors.

A. Model and observables

We study and apply PA in the context of the 3D Edwards-Anderson (EA) spin glass, defined by the Hamiltonian

$$H = -\sum_{\langle n,m \rangle} J_{nm} s_n s_m, \tag{3}$$

where the summation is over nearest neighbors on a cubic lattice with periodic boundary conditions, $s_n = \pm 1$ are Ising

spins, and J_{nm} are bonds drawn from a Gaussian distribution with zero mean and unit standard deviation. In order to understand the thermodynamic properties of the EA model, it is necessary to do many simulations with different bond configurations and to then take an average afterwards. PA gives access to the ensemble average of observables for a fixed set of bonds; therefore many PA simulations must be conducted to take a bond average. An average of an observable O over bond configurations will be denoted as $[O]_{\mathcal{J}}$, whereas a thermal average will be denoted as $\langle O \rangle$.

There are several types of observables that are theoretically interesting in spin glasses. Observables that can be measured directly in a single spin configuration are the easiest to measure since PA effectively simulates the canonical ensemble at each temperature step. Thermodynamic quantities for a single bond configuration such as the average energy $E = \langle H \rangle$ or average magnetization are all straightforward to measure as a simple population average at each temperature step.

The order parameter for the EA model is obtained from the overlap between two spin configurations chosen independently from the canonical ensemble at a given temperature in the same bond configuration. The spin overlap q is defined as

$$q = \frac{1}{N} \sum_{n=1}^{N} s_n^1 s_n^2, \tag{4}$$

where s_n^1 is the nth spin of replica 1 and N is the total number of spins in the system. The thermal distribution of overlaps $P(q)$ for a given temperature and bond configuration has support on $[-1,1]$. The Edwards-Anderson order parameter, q_{EA}, is the thermal average of the absolute value of q in (any) single pure thermodynamic state. In order to measure $P(q)$, it is necessary to measure q many times from spin configurations 1 and 2 drawn independently from the equilibrium ensemble of replicas. This process is straightforward in PA as long as the family of each replica is recorded: PA gives access to the equilibrium ensemble, and replicas are guaranteed to be independent if they are from different families.

An important observable in spin-glass theory is the disorder-averaged integrated overlap weight around the origin,

$$I(q_0) = \left[\int_{-q_0}^{q_0} P(q) dq \right]_{\mathcal{J}}, \tag{5}$$

and following previous studies [13,21], we use $q_0 = 0.2$. The quantity $I(q_0)$ was introduced to distinguish between competing theories of the low-temperature spin-glass phase. Replica symmetry breaking theory [22–24] predicts that $I(q_0)$ goes to a nonzero constant as $N \to \infty$ for any $q_0 > 0$ while the droplet [25–28] and chaotic pairs pictures [29] predict that $I(q_0) \to 0$ as $N \to \infty$ for any $q_0 < q_{\text{EA}}$.

The link overlap q_l is another quantity that is defined from two independent spin configurations,

$$q_l = \frac{1}{N_b} \sum_{\langle n,m \rangle} s_n^1 s_m^1 s_n^2 s_m^2, \tag{6}$$

where $N_b = 3N$ is the number of bonds. Like the spin overlap, the link overlap is useful for distinguishing theories of the spin-glass phase [21]. It is also useful as a measure of equilibration using a relation found by Katzgraber and Young [21]. They

define

$$\Delta_{\text{KY}} = [e_l - e]_{\mathfrak{J}}, \quad (7)$$

where e is the thermally averaged energy per spin and e_l is an energylike quantity defined from the link overlap,

$$e_l = -\frac{1}{T}\frac{N_b}{N}(1 - \langle q_l \rangle). \quad (8)$$

For an individual bond configuration, it is not the case that $e_l = e$; however, for a disorder average over Gaussian bonds, $\Delta_{\text{KY}} = 0$. The requirement that $\Delta_{\text{KY}} \approx 0$ is a useful indicator of equilibration.

A free-energy estimator, $\tilde{F}$, can be obtained [12] at every step in the annealing schedule from the normalization factors [see Eq. (2)] of all previous annealing steps via

$$-\beta_k \tilde{F}(\beta_k) = \sum_{l=k+1}^{N_T-1} \ln Q(\beta_l, \beta_{l-1}) + \ln \Omega, \quad (9)$$

where $\Omega = 2^N$ is the number of microstates, N_T is the total number of β steps in the annealing schedule, and β_l is the current step.

If PA is run to a sufficiently low temperature, then we can obtain an estimator of the ground-state energy $\tilde{E}_0$ from the lowest energy encountered during the simulation. Population annealing also gives direct access to the estimated probability of being in the ground state at each temperature, $\tilde{g}_0$, which is simply the fraction of the population at the lowest energy found during the entire simulation. Alternatively, with the aid of the free-energy estimator, it is possible to obtain an indirect estimate of the probability of being in the ground state, $\bar{g}_0$, by calculating the Boltzmann weight,

$$\bar{g}_0 = 2\,e^{-\beta E_0 + \beta \tilde{F}}. \quad (10)$$

By comparing the measured $\tilde{g}_0$ to the calculated $\bar{g}_0$, we can asses systematic errors.

III. POPULATION ANNEALING THEORY

A. Error estimation

As shown in Ref. [13], the systematic error of an observable O is given by

$$\Delta O = \text{var}(\beta \tilde{F}) \left[\frac{\text{cov}(\tilde{O}, \beta \tilde{F})}{\text{var}(\beta \tilde{F})} \right], \quad (11)$$

where $\tilde{O}$ is the PA estimator of O and the (co)variances are taken with respect to independent runs of PA. The bracketed quantity is expected to converge to a finite limit as $R \to \infty$, meaning that for large R, the systematic error for any observable is proportional $\text{var}(\beta \tilde{F})$. Furthermore, the quantity $\text{var}(\beta \tilde{F})$ is expected to scale as $1/R$, and so it is natural to define an equilibration population size, ρ_f, as

$$\rho_f = \lim_{R \to \infty} R\,\text{var}(\beta \tilde{F}). \quad (12)$$

The equilibration population size sets a population scale for a given bond configuration, and by choosing the population such that $R \gg \rho_f$, systematic errors behave as ρ_f/R. One complication is that in order to measure ρ_f, many simulations of the same bond configuration must be made.

An analogous quantity for statistical errors, ρ_t, can be defined that corresponds to the population size required to minimize statistical errors. In PA, if no decorrelating Markov chain Monte Carlo were done, then the statistical errors would directly scale with the second moment of the family distribution, see Ref. [13]. Therefore we define ρ_t as

$$\rho_t = \lim_{R \to \infty} \frac{1}{R} \sum_{i=1}^{R} \eta_i^2, \quad (13)$$

where the summation is over families and η_i is the size of family i. Note that we can also express ρ_t in terms of the variance of the family size distribution. Since the average family size is 1,

$$\rho_t - 1 = \lim_{R \to \infty} \text{var}(\eta). \quad (14)$$

Because we do perform MCMC at each annealing step, the actual statistical errors will be lower and ρ_t can be used as a conservative estimate for the population size necessary to minimize statistical errors. Specifically, the statistical error δO in measuring an observable O in a PA run with population size R is bounded by

$$\delta O \leqslant \sqrt{\frac{\text{var}(O)\rho_t}{R}}, \quad (15)$$

where $\text{var}(O)$ here refers to the underlying variance of the observable in the Gibbs distribution.

Unlike ρ_f, ρ_t can be easily estimated from a single run and, as will be shown in the next section, there is a close relationship between ρ_f and ρ_t that can be used to our advantage.

B. Relation between ρ_t and ρ_f

In this section we provide an argument justifying the inequality

$$\rho_t - 1 > \rho_f. \quad (16)$$

The argument uses a modified version of PA where the exact free energy is known and is used for normalizing the weight of each spin configuration. This modified version of PA is similar to the idea of using blocks of a much larger total population to calculate ρ_f and ρ_t. In this version of PA, the weight of spin configuration i is given by

$$\tau_i = e^{-(\beta' - \beta)E_i + \beta' F' - \beta F}, \quad (17)$$

where F and F' are the exact free energies at β and β', respectively. The actual number of copies n_i of configuration i is a random non-negative integer, $n(\tau_i)$, whose mean is τ_i. Our implementation of $n(\tau)$ is given in Sec. II but the details are not important to the argument. Here we divide n_i into τ_i and a random remainder $z(\tau_i)$, whose mean is zero,

$$n_i = \tau_i + z(\tau_i). \quad (18)$$

The total population $\tilde{R}_{\beta'}$ at temperature step β' is given by

$$\tilde{R}_{\beta'} = \sum_{i=1}^{\tilde{R}_\beta} n_i, \quad (19)$$

which can also be expressed in terms of the family size distribution,

$$\tilde{R}_{\beta'} = \sum_{j=1}^{R} \eta_j, \tag{20}$$

where η_j is the size of family j at temperature step β'. In the exact free-energy version of PA, families are independent of each other, so the variance of the sum is the sum of variances and $\text{var}(\tilde{R}_{\beta'}) = R \, \text{var}(\eta)$. From the definition of ρ_t, Eq. (14),

$$\rho_t(\beta) - 1 = \lim_{R \to \infty} \frac{1}{R} \text{var}(\tilde{R}_\beta). \tag{21}$$

Thus, in the exact free-energy version of PA, there are relatively large fluctuations in the population size that scale as $\sqrt{\rho_t R}$ in contrast to the version we implement in the simulations, where the population size is nudged toward R at every step and population fluctuations scale as $\sqrt{aR}$ with $a \lesssim 1$.

We can also derive an expression for ρ_f that is related to population size fluctuations within the exact free-energy version of PA. Starting from the factor $Q(\beta, \beta')$ [see Eq. (2)] we have

$$Q(\beta, \beta') = \frac{1}{\tilde{R}_\beta} \sum_{i=1}^{\tilde{R}_\beta} e^{-(\beta' - \beta)E_i}$$

$$= \frac{1}{\tilde{R}_\beta} \left(\sum_{i=1}^{\tilde{R}_\beta} \tau_i \right) e^{-\beta' F' + \beta F}. \tag{22}$$

We also know that $Q(\beta, \beta')$ is an estimator of the ratio of the partition functions of two subsequent temperatures so that

$$Q(\beta, \beta') = e^{-\beta' \tilde{F}' + \beta \tilde{F}}, \tag{23}$$

where $\tilde{F}$ and $\tilde{F}'$ are the free-energy estimators at β and β', respectively. Combining the above two relations gives

$$e^{-\beta' \Delta \tilde{F}' + \beta \Delta \tilde{F}} = \frac{1}{\tilde{R}_\beta} \sum_{i=1}^{\tilde{R}_\beta} \tau_i$$

$$= \frac{1}{\tilde{R}_\beta} \sum_{i=1}^{\tilde{R}_\beta} [n_i - z(\tau_i)], \tag{24}$$

where $\Delta \tilde{F} = \tilde{F} - F$ is the deviation of the free-energy estimator from the exact free energy. If the population is large, then this deviation is small so we can expand the exponential of the free energy, and using Eq. (19), we obtain

$$1 - (\beta' \Delta \tilde{F}' - \beta \Delta \tilde{F}) = \frac{\tilde{R}_{\beta'}}{\tilde{R}_\beta} - \frac{\sum_{i=1}^{\tilde{R}_\beta} z(\tau_i)}{\tilde{R}_\beta}. \tag{25}$$

The population size at temperature β can be decomposed as $\tilde{R}_\beta = R + \delta \tilde{R}_\beta$, where R is the mean population and $\delta \tilde{R}_\beta$ is the deviation from the mean at temperature β. Expanding in $\delta R / R$ yields

$$1 - (\beta' \Delta \tilde{F}' - \beta \Delta \tilde{F}) = \left(1 + \frac{\delta \tilde{R}_{\beta'}}{R}\right)\left(1 - \frac{\delta \tilde{R}_\beta}{R}\right)$$

$$- \frac{1}{R} \sum_{i=1}^{\tilde{R}_\beta} z(\tau_i)\left(1 - \frac{\delta \tilde{R}_\beta}{R}\right). \tag{26}$$

From this point onwards, $\tilde{R}_j$ will denote the population at annealing step j with inverse temperature β_j, and τ_i^j will denote the weight of configuration i during the resampling step from β_{j+1} to β_j. Disregarding all $(\delta R / R)^2$ terms, summing Eq. (26) over all steps in the annealing schedule, and taking the variance of the result yields

$$\text{var}(\beta_k \Delta \tilde{F}_k) = \text{var}\left[\frac{\delta \tilde{R}_k}{R} - \frac{1}{R} \sum_{j=N_{T-1}}^{k+1} \sum_{i=1}^{\tilde{R}_{j+1}} z(\tau_i^j)\right]. \tag{27}$$

From the definition of ρ_f, Eq. (12), we have

$$\rho_f(\beta_k) = \lim_{R \to \infty} \frac{1}{R} \text{var}\left[\delta \tilde{R}_k - \sum_{j=N_{T-1}}^{k+1} \sum_{i=1}^{\tilde{R}_{j+1}} z(\tau_i^j)\right]. \tag{28}$$

Expanding the variance and using Eq. (21) yields a relation between ρ_f and ρ_t,

$$\rho_f(\beta_k) = \rho_t(\beta_k) - 1$$

$$- \lim_{R \to \infty} \frac{1}{R} \left\{ 2 \, \text{cov}\left[\delta \tilde{R}_k, \sum_{j=N_{T-1}}^{k+1} \sum_{i=1}^{\tilde{R}_{j+1}} z(\tau_i^j)\right]\right.$$

$$\left. - \text{var}\left[\sum_{j=N_{T-1}}^{k+1} \sum_{i=1}^{\tilde{R}_{j+1}} z(\tau_i^j)\right]\right\}. \tag{29}$$

In Appendix A we argue that the quantity in curly brackets is greater than zero, yielding the desired inequality, $\rho_t - 1 > \rho_f$.

There are two caveats concerning this inequality. First, the argument in Appendix A establishing the positivity of the term in square brackets is not rigorous. More importantly, the result applies to a version of PA that is normalized by the exact free energy and has large fluctuations in population size. We conjecture that an "equivalence of ensembles" result holds for the implemented and exact free-energy version of PA so that both ρ_f and ρ_t are the same for both algorithms but this question deserves further study.

We will see in the next two sections and in Appendix B that the inequality between ρ_f and $\rho_t - 1$ can be extended to an approximate equality, provided that the culling fraction is small at each step. This approximate equality and, by extension, the inequality are supported by numerical results shown in Sec. V A.

C. Temperature step size, culling fraction, and energy variance

As we shall see in Sec. IV B, a natural way to choose the β schedule for population annealing is to cull a fixed fraction of the population at each resampling step. In this section we derive a relation between the culling fraction, the variance of the energy distribution, and the size of the temperature step. To derive this relation, note that the expected number of copies of each configuration is τ_i and the actual number of copies is $\lfloor \tau_i \rfloor$ with probability $\lceil \tau_i \rceil - \tau_i$ or $\lceil \tau_i \rceil$ otherwise. Thus a configuration can be eliminated only if $\tau < 1$, and the expected

number of eliminated configurations in a resampling step is

$$\epsilon R = \sum_{\tau_i < 1} (1 - \tau_i), \tag{30}$$

where ϵ is the (expected) culling fraction. Let $\langle E \rangle$ and σ_E^2, respectively, be the thermal average energy and variance of the energy. Consider a resampling step from β to β' with $\Delta\beta = (\beta' - \beta)$. In the regime $\Delta\beta\sigma_E \ll 1$ we can expand the definition of τ_i, Eq. (1), to leading order in $\Delta\beta$ to obtain

$$\tau_i = 1 - \Delta\beta(E_i - \langle E \rangle), \tag{31}$$

meaning that τ is approximately a Gaussian random variable with mean 1 and standard deviation $\Delta\beta\sigma_E$. Within this Gaussian approximation and for large R, the sum defining ϵ in Eq. (30) can be replaced by an integral,

$$\epsilon \approx \int_{-\infty}^{1} (1 - \tau)\mathcal{N}(\tau; 1, \Delta\beta\sigma_E)\, d\tau$$
$$= \frac{\Delta\beta\sigma_E}{\sqrt{2\pi}}, \tag{32}$$

where $\mathcal{N}(x; \mu, \sigma)$ is the pdf of the normal distribution with mean μ and standard deviation σ. If we want to eliminate a fixed fraction of the population ϵ, then the β schedule must be chosen such that

$$\Delta\beta \approx \frac{\epsilon\sqrt{2\pi}}{\sigma_E}. \tag{33}$$

D. Growth of ρ_f in the MCMC-equilibrated regime

In the spin-glass phase where the MCMC procedure alone is unable to equilibrate the system, the growth of ρ_f depends, in a complicated way, on the structure and temperature dependence of the free-energy landscape. However, at high temperatures where the MCMC procedure is able to fully decorrelate replicas, which we refer to as the MCMC-equilibrated regime, we can show that ρ_f is simply proportional to the number of annealing steps times the culling fraction. To understand the behavior of ρ_f in the MCMC-equilibrated regime, note that ρ_f is defined in terms of the variance of βF [see Eq. (12)] and take the variance of both sides of Eq. (9),

$$\mathrm{var}[\beta_k \tilde{F}(\beta_k)] = \mathrm{var}\left[\sum_{l=k+1}^{N_T-1} \ln Q(\beta_l, \beta_{l-1})\right]$$
$$\approx \sum_{l=k+1}^{N_T-1} \mathrm{var}[\ln Q(\beta_l, \beta_{l-1})]. \tag{34}$$

The second approximate equality becomes exact when the population is equilibrated by the MCMC procedure during each annealing step. Using the definition of Q, Eq. (2), and assuming the variation of $\bar{R}$ is negligible, we expand $\mathrm{var}[\ln Q(\beta_l, \beta_{l-1})]$ to leading order in $(\Delta\beta_l)\sigma_E(\beta_l)$,

$$\mathrm{var}[\ln Q(\beta_l, \beta_{l-1})] \approx \frac{1}{R}(\Delta\beta_l)^2\sigma_E(\beta_l)^2, \tag{35}$$

where σ_E^2 is the variance of the energy distribution. Plugging Eq. (35) into Eq. (34) yields

$$\mathrm{var}[\beta_k \tilde{F}(\beta_k)] = \sum_{l=k+1}^{N_T-1} \frac{1}{R}(\Delta\beta_l)^2\sigma_E(\beta_l)^2. \tag{36}$$

From the relation between the size of the temperature step and the culling fraction, Eq. (32), we find

$$R\,\mathrm{var}[\beta_k \tilde{F}(\beta_k)] = \sum_{l=k+1}^{N_T-1} 2\pi\epsilon(l)^2, \tag{37}$$

where $\epsilon(l)$ is the culling fraction at the lth annealing step. Thus, for fixed culling fraction, ϵ, we find that ρ_f grows linearly in the number k of annealing steps,

$$\rho_f = 2\pi\epsilon^2 k. \tag{38}$$

This relation is valid if the culling fraction is small and enough MCMC sweeps are carried out in each annealing step that the replicas remain independent.

More generally, $\rho_f \geqslant 2\pi\epsilon^2 k$ and the inequality holds if the MCMC procedure is not able to keep the replicas fully decorrelated.

E. Growth of ρ_t in the MCMC-equilibrated regime

Similarly to the case of ρ_f, at high temperatures where the MCMC procedure is able to fully decorrelate the energy of the replicas at every annealing step, the growth of $\rho_t - 1$ is proportional to the number of annealing steps times the culling fraction. To derive this relation we note that ρ_t is equal to the variance of the family size distribution. In the MCMC-equilibrated regime, the size of a given family is described by a birth and death process. In an approximation where the annealing step k is taken to be a continuous "time" variable, the family size distribution, $P_\eta(k)$, is described by the master equation,

$$\dot{P}_\eta(k) = \epsilon[(\eta - 1)P_{\eta-1}(k) + (\eta + 1)P_{\eta+1}(k) - 2\eta P_\eta(k)], \tag{39}$$

where η is the family size and ϵ is the culling fraction. This is the classic birth–death process (see, for example, Ref. [30]) whose solution is

$$P_\eta(k) = \left(\frac{1}{1 + \epsilon k}\right)\left(\frac{\epsilon k}{1 + \epsilon k}\right)^{\eta-1} \text{ for } \eta \geqslant 1 \tag{40}$$

and

$$P_0(k) = \left(\frac{\epsilon k}{1 + \epsilon k}\right). \tag{41}$$

From this distribution it is easily seen that

$$\rho_t - 1 = \mathrm{var}(\eta) = 2\epsilon k. \tag{42}$$

This equation holds in the MCMC-equilibrated regime where the number sweeps in each annealing step is greater than or comparable to the integrated autocorrelation time of the energy, so that the energy of every replica is independent of its family designation.

TABLE I. Parameters of the simulations for each of the three system sizes.

	$L = 6$	$L = 8$	$L = 10$
Culling fraction, ϵ	0.103	0.168	0.272
Sweeps per replica	1005	1005	1009
Temperature steps, N_T	95	95	99
Initial population, R_0	2×10^3	2×10^3	2×10^4
Maximum population	2.64×10^5	1.5×10^7	1.5×10^7
Disorder samples	2×10^4	2×10^4	5×10^3
Unequilibrated samples	0	4	145

Comparing Eqs. (42) and (38), we see that in the MCMC-equilibrated regime and for small culling fraction,

$$\rho_f = \rho_t - 1 - (2\epsilon - 2\pi\epsilon^2)k. \tag{43}$$

To a high level of accuracy, this relation also holds outside of the MCMC-equilibrated regime. This is shown numerically in Sec. V A and supported analytically in Appendix B.

IV. OPTIMIZATION OF POPULATION ANNEALING

In this work we focus on three general improvements to the population annealing algorithm: a method to choose the population size for each individual bond configuration, a way to choose the β schedule to reduce statistical and systematic errors, and an ad hoc sweep schedule that improves equilibration.

A. Hardness-dependent population size

Previous authors have shown that the computational work required to equilibrate a specific bond configuration is approximately log-normally distributed for both population annealing and parallel tempering [13,31–33]. In population annealing, the work required to equilibrate a bond configuration is proportional to ρ_f, and so our first optimization takes advantage of the relation between ρ_f and ρ_t in order to tailor the population size necessary for each bond configuration, so we do not use too large a population on an easy system and spend resources inefficiently. In order to optimize the population for each bond configuration, it is usually necessary to do several simulations. An initial simulation is done with a small population, R_0. From this simulation, we obtain an estimate of ρ_t, called $\rho_t(R_0)$. If $R_0 > 100\rho_t(R_0)$, then we assume that $\rho_t(R_0) \approx \rho_t$ and the simulation was well equilibrated. Otherwise, we do another simulation with population size, $R_1 = 150\rho_t(R_0)$, which yields $\rho_t(R_1)$. This procedure is continued until the bond configuration is equilibrated according to the criterion of $R > 100\rho_t$. If R_0 is chosen wisely, then this method converges quickly and uses far fewer resources than choosing a single R adequate for all bond configurations. The values of R_0 used in our simulations are given in Table I.

B. Optimal annealing schedule

In the MCMC-equilibrated regime it is possible to derive an optimal annealing schedule, which is composed of both the β schedule and sweep schedule. This is done by minimizing ρ_f while keeping the total amount of computational work, W, fixed. We define the total work as

$$W = \sum_{l=k+1}^{N_T-1} S(\beta_l), \tag{44}$$

$$\approx \int_0^\beta \frac{S(\beta)}{\Delta\beta} d\beta, \tag{45}$$

where $S(\beta)$ is the sweep schedule, defined as the number of MCMC sweeps carried out per annealing step at inverse temperature β. In the MCMC-equilibrated regime we have an analytic expression for ρ_f from Eq. (36), which can also be approximated as an integral,

$$\rho_f = \sum_{l=k+1}^{N_T-1} (\Delta\beta_l)^2 \sigma_E(\beta_l)^2, \tag{46}$$

$$\approx \int_0^\beta \Delta\beta \sigma_E(\beta) d\beta. \tag{47}$$

We can use the method of Lagrange multipliers to minimize ρ_f while holding W constant by solving,

$$0 = \frac{\delta}{\delta\Delta\beta}(\rho_f + \lambda W), \tag{48}$$

which yields

$$\sigma_E^2 - \lambda \frac{S(\beta)}{\Delta\beta^2} = 0. \tag{49}$$

Equivalently,

$$\Delta\beta \sigma_E(\beta) \propto \sqrt{S(\beta)}, \tag{50}$$

and using Eq. (33),

$$\epsilon \propto \sqrt{S(\beta)}, \tag{51}$$

we see that in the MCMC-equilibrated regime, the optimal number of sweeps depends on the culling fraction at each step. In the case of a fixed culling fraction, a fixed sweep schedule is optimal.

Based on these ideas, in our simulations we employed a β schedule that holds the culling fraction roughly constant. We observed that the resulting schedule does not depend strongly on disorder realization so we chose a schedule based on a single run and tested that the culling fraction was similar for several disorder configurations of varying difficulty. This schedule was then employed without modification for production runs. Since, according to Eq. (33), $\Delta\beta = \epsilon\sqrt{2\pi}/\sigma_E$, the resulting schedule has many annealing steps at high temperatures, where the standard deviation of the energy is large, and few annealing steps at low temperatures. A similar uniform ϵ scheme was used in PA simulations of hard spheres [17]. Since the variance of the energy scales linearly with the system size (except near the critical point), the same β schedule can be used for many system sizes, though for larger sizes the uniform culling fraction will increase. If the range of sizes studied is too large, then interpolating temperatures can be added to the schedule to reduce the culling fraction. Although it is not theoretically well justified in the critical or the low-temperature regimes, we

continued to use a uniform culling fraction to determine the β schedule over the entire temperature range of our simulations.

In accordance with Eq. (51), we chose a sweep schedule that was fixed in the high-temperature regime. We found that focusing most of the computational work in the glass transition regime minimized ρ_t and that in the glassy phase it was favorable to rely on PA resampling and do very little MCMC work. Our ad hoc sweep schedule had three Monte Carlo sweeps performed for $\beta < 0.5$, 22 for $0.5 \leqslant \beta \leqslant 2.5$, and a single sweep for $\beta > 2.5$. Our annealing schedule is not directly comparable with the annealing schedules studied in Ref. [20]. While our β schedules is the same as their "std(E)" schedule, the computationally meaningful density of sweeps per unit β is not the same as for any of their annealing schedules because of our nonuniform sweep schedule. A deeper understanding of the optimal annealing schedule over the whole range of temperatures remains an open problem.

V. RESULTS

This section contains two types of numerical results. Sections VA–VD address the behavior of PA, validating theoretical predictions and testing the optimization ideas described in previous sections. Section V E reports measurements from large-scale simulations of several observables for the 3D Edwards-Anderson spin glass in order to compare with previous work and improve the state of the art. The parameters used in this work are shown in Table I.

A. Relationship of ρ_f and ρ_t

In Sec. III B we showed that $\rho_t - 1 \geqslant \rho_f$ and, in Eq. (43), gave an approximate relation between these two quantities. Here we test these relationships. Our proposal to optimize population size for each disorder realization relies on the easily measured ρ_t as a proxy for the more difficult to measure ρ_f, so it is important to determine the relationship between these two quantities.

In order to accurately measure $\rho_f = R\,\mathrm{var}(\beta \tilde{F})$, we ran population annealing 48 times for each configuration, with population sizes chosen such that $R \geqslant 100\rho_t$. This ensured that each simulation was well equilibrated and that we could measure $\mathrm{var}(\beta \tilde{F})$ with reasonable accuracy. We calculated ρ_f for 2000 $L = 6$ samples and for 300 $L = 8$ samples. Calculating the error of ρ_f is equivalent to calculating the error of a sample variance. As shown in previous work [13], $\beta \tilde{F}$ taken from a well-equilibrated bond configuration is normally distributed, which makes estimation of the error of $\mathrm{var}(\beta \tilde{F})$ particularly easy [34],

$$\mathrm{var}[\mathrm{var}(\beta \tilde{F})] = \frac{2}{M-1}\mathrm{var}(\beta \tilde{F})^2, \qquad (52)$$

where M is the number of trials. The corresponding error in ρ_f is

$$\delta\rho_f = \sqrt{\frac{2}{M-1}}\rho_f, \qquad (53)$$

which for 48 trials gives a relative error, $\delta\rho_f/\rho_f$, of about 21%. Since ρ_t is calculated from a single disorder realization, it is expected that $\delta\rho_t \ll \delta\rho_f$. We find this to be true empirically,
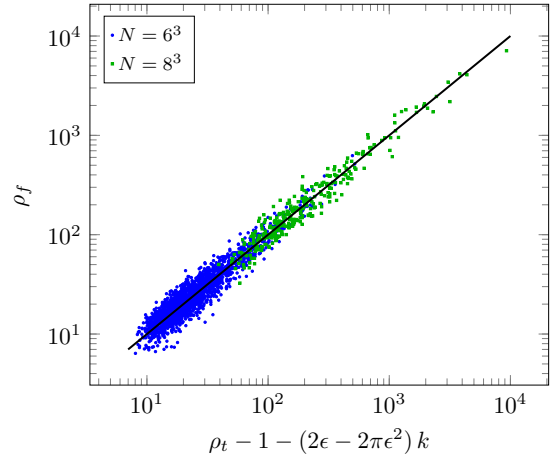

FIG. 1. Scatter plot of $\rho_t - 1 - (2\epsilon - 2\pi\epsilon^2)k$ vs. ρ_f at $\beta = 5$ ($k = 95$) for $L = 6$ and $L = 8$; each data point corresponds to a single bond configuration. The solid line corresponds to $\rho_t - 1 + (2\pi\epsilon^2 - 2\epsilon)k = \rho_f$, see Eq. (43).

with $\delta\rho_f \approx 20\delta\rho_t$ for $L = 6$ and $\delta\rho_f \approx 16\delta\rho_t$ for $L = 8$ simulations.

Figure 1 is a scatter plot of $\rho_t - 1 - (2\epsilon - 2\pi\epsilon^2)k$ vs. ρ_f at $\beta = 5$, where each point corresponds to one disorder realization and the solid line corresponds to $\rho_f = \rho_t - 1 - (2\epsilon - 2\pi\epsilon^2)k$. The results are consistent with $\rho_f \approx \rho_t - 1 - (2\epsilon - 2\pi\epsilon^2)k$ holding for all disorder realizations.

In the MCMC-equilibrated regime, $\rho_f = 2\pi\epsilon^2 k$ and $\rho_t = 1 + 2\epsilon k$, as shown in Secs. III D and III E, respectively. Figure 2 shows $[\rho_f]_{\mathcal{J}}$ and $[\rho_t]_{\mathcal{J}}$ for $L = 6$ with the dashed lines representing the theoretical linear dependence on number of annealing steps k. The solid line represents the estimated value of ρ_f calculated using $[\rho_t]_{\mathcal{J}} - 1 - (2\epsilon - 2\pi\epsilon^2)k$. The estimated value was found to be within 5% of the true value
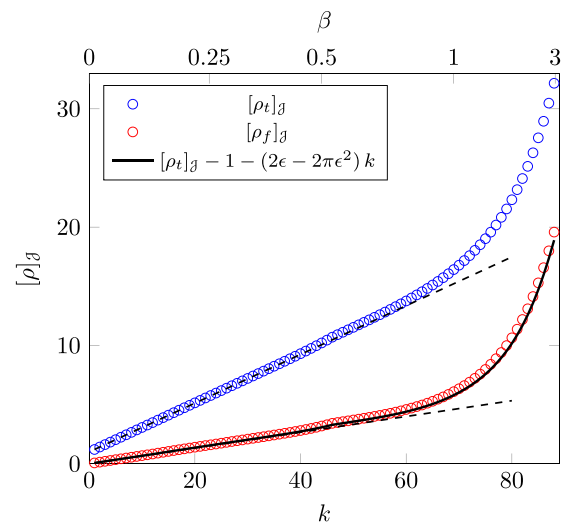

FIG. 2. $[\rho_f]_{\mathcal{J}}$ and $[\rho_t]_{\mathcal{J}}$ of $L = 6$ as a function of annealing step k, with the nonlinear β scale on the upper x axis. Dashed lines correspond to the theoretically predicted MCMC-equilibrated estimates, Eqs. (38) and (42). The solid line corresponds to the difference $[\rho_t]_{\mathcal{J}} - 1 - (2\epsilon - 2\pi\epsilon^2)k$, see Eq. (43).
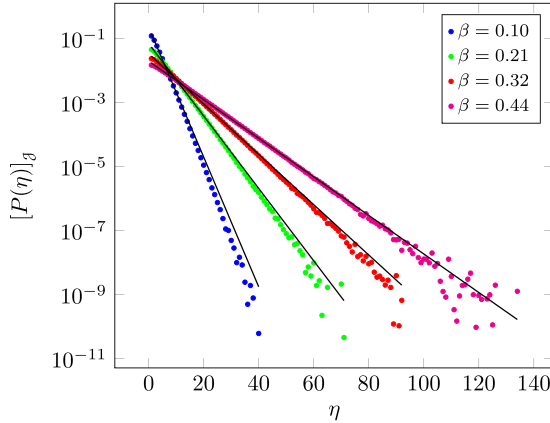
FIG. 3. The disordered-averaged family size distribution $[P(\eta)]_{\mathcal{J}}$ as a function of family size η, at several temperatures all in the high-temperature regime (the β of each distribution increases from left to right). The distributions are exponential and have shape parameters that match the predictions of Eq. (40). The value at $\eta = 0$ is not shown in this plot.

of ρ_f for all annealing steps for both $L = 6$ and $L = 8$ (not shown). Note that the sharp rise in both ρ_f and ρ_t occurs near the critical temperature, $\beta_c = 1.05$.

Figure 3 shows the disorder-averaged family size distribution for several temperatures in the MCMC-equilibrated regime for $L = 8$ and confirms the prediction of Sec. III E of an exponential family size distribution. The straight lines in the figure are obtained from Eq. (40) and show that there is good quantitative agreement except for tail of the distribution. The higher values of η are under-represented, especially for low values of β, due to the finite size of the population.

B. Distribution of ρ_t

It is known [13,31–33,35] that the computational hardness of simulating Ising spin glasses has a broad distribution with respect to disorder realizations. In the context of a MCMC algorithm such as parallel tempering the computational hardness is typically measured by the exponential or integrated autocorrelation times. These quantities have been found to be approximately lognormally distributed. For population annealing, we may use ρ_f or ρ_t for the purpose of measuring computational hardness. Figure 4 shows histograms of $\log_{10} \rho_t$ for the three system sizes simulated. We found that a log-inverse Gaussian distribution is an excellent fit to the ρ_t distributions. The fits are shown as solid lines in the figure. The three-parameter inverse Gaussian distribution is defined by

$$P(x; \mu, \lambda, l) = \sqrt{\frac{\lambda}{2\pi(x-l)^3}} \exp\left[\frac{-\lambda(x-\mu-l)^2}{2\mu^2(x-l)}\right] \quad (54)$$

with $x = \ln \rho_t$. The parameters of the fits are given in Table II. The shift parameter l, which shifts the support of the distribution from $(0, \infty)$ to (l, ∞), is necessary since ρ_t is bounded away from zero. A rough estimate of shift parameter can be obtained by assuming that the easiest bond realizations are MCMC equilibrated for all temperatures, resulting in $l = \log(\rho_t^{\min}) \approx$
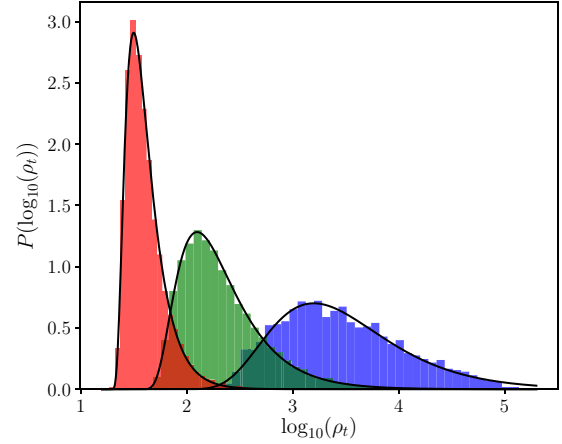

FIG. 4. The distribution of $\log_{10} \rho_t$ for system sizes $L = 6$ (left), 8 (middle), and 10 (right). The solid lines are inverse Gaussian fits with the parameters given in Table II.

$\log(1 + 2\epsilon k_{\max})$. The values for l obtained from this formula are 2.97, 3.50, and 4.00 for sizes $L = 6$, 8, and 10, respectively. These values are in reasonable agreement with the fitted values shown in Table II. The log inverse Gaussian also works well to fit the ρ_f distribution with a shift parameter predicted by $\rho_f^{\min} = 2\pi\epsilon^2 k$.

We can also compare the disorder averaged value of ρ_t with the predictions from the fit. The mean of $\rho_t = e^x$ is given by

$$[\rho_t]_{\mathcal{J}} = \exp\left[l + \frac{\lambda}{\mu}(1 - \sqrt{1 - 2\mu^2/\lambda})\right] \quad (55)$$

for $\lambda/2\mu^2 > 1$. The tail of the inverse Gaussian is exponential, so if $\lambda/2\mu^2 \leqslant 1$, then the mean of e^x is infinite. The fitted values of $[\rho_t]_{\mathcal{J}}$ obtained from this equation are shown in Table II. The values computed directly from the data are $[\rho_t]_{\mathcal{J}} = 46$, 385, and 12 700 for $L = 6$, 8, and 10, respectively. The large discrepancy between the $L = 10$ fitted and measured values may be due to several factors. First, the number of samples for $L = 10$ is smaller than for $L = 6$ and 8, so the tail of the ρ_t distribution may not be fully sampled. Second, a significant fraction of $L = 10$ samples were not equilibrated and, for these samples, we have most likely underestimated ρ_t. If the tail of the distribution is properly described by the log inverse Gaussian, then the value of $[\rho_t]_{\mathcal{J}}$ obtained from the fit may be more accurate than the average of the ρ_t data from a finite sample size. On the other hand, for $L = 10$, the ratio $\lambda/2\mu^2 = 1.04$, which is quite near the divergence at $\lambda/2\mu^2 = 1$ so results for $[\rho_t]_{\mathcal{J}}$ may be highly sensitive to errors in the fit. The near divergence of $[\rho_t]_{\mathcal{J}}$ for $L = 10$

TABLE II. Fits of the ρ_t distribution to a log inverse Gaussian distribution defined in Eq. (54).

	$L = 6$	$L = 8$	$L = 10$
μ	0.83(1)	1.97(2)	4.53(13)
λ	3.25(16)	9.28(35)	42.7(40)
l	2.88(1)	3.39(2)	3.50(13)
$[\rho_t]_{\mathcal{J}}$	46.4	491	66500

also suggests that the annealing schedule for this size should have either more temperature steps or more sweeps per step.

As seen in Table II, the computational effort required to reach equilibrium scales up rapidly with system size and is broadly distributed. How does this effort translate into wall clock time on a modern CPU? For $L = 10$ the algorithm's run time on a single CPU is approximately 0.03 s/replica. The typical value of ρ_t for $L = 10$, defined by $\exp([\log \rho_t]_{\mathcal{J}})$, is approximately 3000 and the equilibration criterion is that $R \geqslant 100\rho_t$, thus the typical running time is approximately 2.5 h. This relatively benign number is, however, misleading because of the exponential tail of the $\log \rho_t$ distribution. If one instead takes the average hardness, $[\rho_t]_{\mathcal{J}} \approx 66500$, predicted from the inverse Gaussian fit (see Table II), then the average running time to equilibrate every disorder realization in a very large $L = 10$ sample would be approximately 55 h per disorder realization. This number exceeds the computing time expended on our $L = 10$ simulations since we did not equilibrate all disorder realizations. Of course, the algorithm can be efficiently parallelized so the wall clock time per replica can be made much smaller than these numbers.

C. Optimized vs. unoptimized annealing schedule

In this section we compare the performance of an optimized and unoptimized annealing schedule used in PA. Our optimized annealing schedule has a β schedule that keeps the culling fraction ϵ nearly constant and an ad hoc sweep schedule that concentrates sweeps over a range of temperatures around the critical point, as described in Sec. IV B. The unoptimized annealing schedule has constant β steps, $\Delta\beta = 0.05$, with 10 sweeps per step, and is similar to the annealing schedule used in Ref. [13].

The figure of merit that we wish to minimize is size of the systematic errors, which scale as $\mathrm{var}(\beta F)$, times the total computational work, $W = RS$, where R is the population size and S is the total number of Monte Carlo sweeps per replica. Using the result $\rho_f \approx \rho_t$ we have that

$$W \, \mathrm{var}(\beta F) \approx S\rho_t. \qquad (56)$$

We have used the same number of sweeps in both the optimized and unoptimized algorithms so the comparison of the performance of the algorithms reduces to the comparison of ρ_t.

Figure 5 is a scatter plot of the values of ρ_t, the vertical position of each point is the optimized ρ_t^{opt}, and the horizontal position, ρ_t^{unopt}, the unoptimized value. Each point represents one of 300 disorder realizations for system size $N = 8^3$ at $\beta = 5$. The plot shows that the unoptimized algorithm is, on average, less efficient by a factor of 7.6. This means that, on average, the unoptimized algorithm requires a population 7.6 times larger to achieve the same quality of results as the optimized algorithm.

D. $L = 6, 8, 10$ results

To test the optimized algorithm and obtain state-of-the-art results for observables described in Sec. II A, we ran large-scale
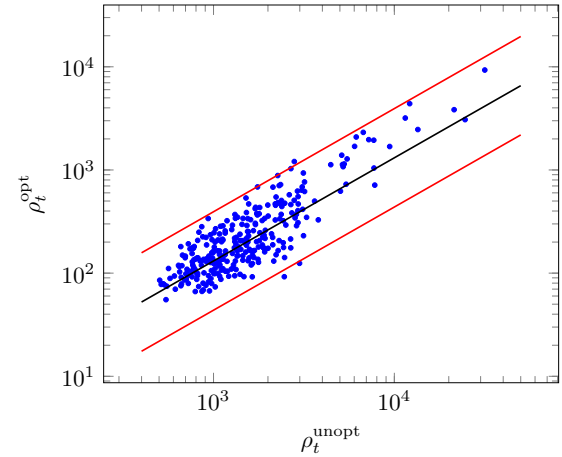

FIG. 5. Values of ρ_t at $\beta = 5$ for the optimized and unoptimized annealing schedules. Each point corresponds to one of 300 $L = 8$ bond configurations. The horizontal coordinate of each point is the unoptimized value and the vertical coordinate the optimized value of ρ_t. The central (black) line corresponds to an improvement of the optimized relative to the unoptimized annealing schedule by a factor of 7.6, and the upper and lower (red) lines correspond to factors of 7.6/3 and 7.6 × 3, respectively.

simulations of the 3D EA spin glass for three system sizes, with parameters provided in Table I. As seen in this table, the equilibration standard, $R \geqslant 100 \rho_t$, was met by nearly all configurations for $L = 6$ and $L = 8$, but approximately 3% of the configurations for $L = 10$ remained unequilibrated when the maximum population was restricted to $R = 1.5 \times 10^7$. The equilibration standard used was higher than previous papers employing PA to study spin glasses, and some systems that we rejected as unequilibrated would have been accepted previously.

It is worth emphasizing that the adaptive population scheme allowed us to sample more bond configurations than most previous studies, while ensuring that nearly all configurations
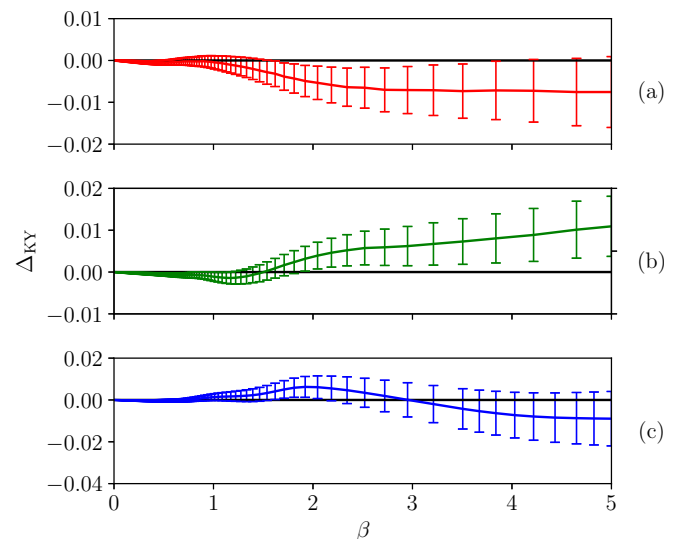

FIG. 6. Δ_{KY} as a function of inverse temperature β for $L = 6$ (a), $L = 8$ (b), and $L = 10$ (c).

TABLE III. The integrated overlap $I(0.2)$, Katzgraber-Young equilibration measure Δ_{KY}, and the bond-averaged ground-state energy per spin $[E_0/N]_{\mathcal{J}}$, all measured at $\beta = 5$.

	$L = 6$	$L = 8$	$L = 10$
$I(0.2)$	0.0188(5)	0.0185(5)	0.0185(10)
Δ_{KY}	$-0.0075(85)$	0.010(7)	$-0.009(13)$
$[E_0/N]_{\mathcal{J}}$	$-1.6891(4)$	$-1.6951(2)$	$-1.6976(3)$

were well equilibrated. As a result, errors associated with a finite number of bond configurations are especially low for $L = 6$ and 8, where we used 2×10^4 samples. The statistical errors reported in Table III and IV are obtained from the standard deviation of the observable with respect to disorder realization and do not include errors associated with individual disorder realizations. Despite the large number of disorder realizations, the error due to the finite sample size is substantially larger than the contribution from systematic and statistical errors of each disorder realization, as shown in Ref. [13].

Overall, the results in Table III are consistent with those found in previous works [13,14]. The average ground-state energy, $[E_0]_{\mathcal{J}}$, is within error bars of previous measurements, and Δ_{KY} is close to zero for all β, as shown in Fig. 6, which is consistent with a well-equilibrated set of samples. Our values of $I(0.2)$ are all within two standard deviations of those found in Ref. [13]; however, it is noteworthy that our values are consistently lower. Despite the slight difference in value, the trend that $I(0.2)$ remains constant over several system sizes is evident.

Table IV shows the disorder average of the logarithm of the directly measured fraction in the ground state $\tilde{g}_0$ and the indirectly measured quantity $\bar{g}_0$ calculated using the Gibbs distribution, Eq. (10). Although the two methods yield values that are within error bars, the computed $\bar{g}_0$ appears to be consistently larger than the measured g_0. To leading order, the free-energy estimator is systematically larger than the actual free energy by $F = \tilde{F} + \rho_f/2\beta R$ [13], so it is expected that $\bar{g}_0$ would be systematically larger than g_0. It should be noted that by definition g_0 cannot be zero because the ground-state energy is here defined as the lowest energy replica found, even if this is not the true ground state.

VI. CONCLUSIONS

This paper makes several contributions to understanding and improving the population annealing algorithm, especially as applied to spin glasses: We have studied the behavior of two important measures of equilibration for population

TABLE IV. Comparison of the disorder average of the logarithm of the fraction in the ground state, $[\log_{10}(\tilde{g}_0)]_{\mathcal{J}}$ at $\beta = 5$ to the indirect measure, $[\log_{10}(\bar{g}_0)]_{\mathcal{J}}$ based on the Boltzmann factor, see Eq. (10).

L	$[\log_{10} \tilde{g}_0]_{\mathcal{J}}$	$\log_{10} 2 - \beta[(\tilde{E}_0 - \tilde{F})]_{\mathcal{J}}/\ln(10)$
6	$-0.7577(23)$	$-0.7549(23)$
8	$-1.6957(34)$	$-1.6900(34)$
10	$-3.2251(93)$	$-3.2104(92)$

annealing, optimized the algorithm in several ways, and obtained state-of-the-art results for several important spin-glass observables. Our results help put population annealing on a firmer footing as an effective tool for highly parallelized simulations of disordered systems such as spin glasses that have rough free-energy landscapes. While this paper focuses on the three-dimensional Edwards-Anderson model, many of the theoretical results and optimization methods are applicable to population annealing simulations of a much broader class of systems.

The two equilibration measures, ρ_t and ρ_f, set the population size needed to control statistical and systematic errors, respectively. The equilibrium population size ρ_f is based on the variance of the free energy and is the more fundamental measure of systematic errors but more difficult to accurately measure in a single simulation. We have demonstrated that $\rho_f \leqslant \rho_t - 1$ and confirmed that these two quantities are close to being equal when both are large. We have also shown that in the MCMC-equilibrated regime, ρ_f and ρ_t each grow linearly in the number of annealing steps, a fact that can be used to design optimal annealing schedules. Finally, we have shown that for the 3D EA spin glass, the distribution of $\log \rho_t$ values is accurately described by an inverse Gaussian distribution.

We have shown that there are a number of simple modifications which improve the efficiency of population annealing. We have also shown that a β schedule that is chosen by fixing the culling fraction is optimal in the MCMC-equilibrated regime. Last, we have shown that the sweep schedule can be improved by increasing the number sweeps in the critical region. Annealing schedule optimizations alone have accounted for nearly an order of magnitude improvement over previous versions of the algorithm. In the context of spin glasses, where there is a broad distribution of computational hardness, tailoring the population size to the difficulty of the disorder realization yields the single largest improvement in efficiency. In addition, the sample-dependent population size results in simulations with both less work and higher overall accuracy.

We still lack a theoretical understanding of the best sweep schedule. Intuitively, we would like the Markov chain Monte Carlo subroutine to fully equilibrate each replica in the high-temperature regime. However, in the low-temperature regime, this is not feasible and the goal is to equilibrate replicas only within their free-energy minima, leaving resampling to properly redistribute replicas between distinct free-energy minima. We have not yet found a principled way to achieve this goal.

APPENDIX A: COVARIANCE INEQUALITY

In Sec. III B we showed that $\rho_t - 1 > \rho_f$, at least for the exact free-energy version of PA, if the following inequality

holds:

$$2\,\mathrm{cov}\left[\delta\tilde{R}_k, \sum_{j=N_{T-1}}^{k+1}\sum_{i=1}^{\tilde{R}_{j+1}} z(\tau_i^j)\right]$$
$$-\,\mathrm{var}\left[\sum_{j=N_{T-1}}^{k+1}\sum_{i=1}^{\tilde{R}_{j+1}} z(\tau_i^j)\right] > 0. \quad (A1)$$

To establish this inequality, we begin by noting that $\delta\tilde{R}_k$ is the sum of deviations from the initial population size R that have accumulated during each resampling step,

$$\delta\tilde{R}_k = \sum_{j=N_{T-1}}^{k+1}\sum_{i=1}^{\tilde{R}_{j+1}}\left[n(\tau_i^j)-1\right]$$
$$= \sum_{j=N_{T-1}}^{k+1}\sum_{i=1}^{\tilde{R}_{j+1}}\left[\tau_i^j + z(\tau_i^j) - 1\right]. \quad (A2)$$

Using this result to expand the covariance term shows that the desired inequality, Eq. (A1), can be rewritten as

$$2\,\mathrm{cov}\left[\sum_{j=N_{T-1}}^{k+1}\sum_{i=1}^{\tilde{R}_{j+1}}\tau_i^j, \sum_{j=N_{T-1}}^{k+1}\sum_{i=1}^{\tilde{R}_{j+1}} z(\tau_i^j)\right]$$
$$+\,\mathrm{var}\left[\sum_{j=N_{T-1}}^{k+1}\sum_{i=1}^{\tilde{R}_{j+1}} z(\tau_i^j)\right] > 0. \quad (A3)$$

The variance is obviously non-negative but we do not have a proof that the covariance term is also non-negative. However, we can motivate this assertion by noting that if the population at an earlier resampling step is stochastically increased ($z > 0$), then later population sizes will tend to be increased, i.e., $\mathrm{cov}[\tilde{R}_{j'}, z(\tau_i^j)] \geqslant 0$ for all i and all $j' < j$. Furthermore, if the population is stochastically increased at an early resampling step, since it is now larger, it will better explore the low-energy tail of the Gibbs distribution so that $\mathrm{cov}[\tau_{i'}^{j'}, z(\tau_i^j)] \geqslant 0$ for all i and i', and all $k < j$. These two mechanisms both cause the

covariance term in Eq. (A3) to be positive. It is worth noting that $\mathrm{cov}[\tau_{i'}^{j'}, z(\tau_i^j)] = 0$ for all i and i', and all $j' < j$.

APPENDIX B: VARIANCE EXPANSION

The variance term from Eq. (A3) can be further expanded to get to a form similar to that of Eq. (43). To do this, we make the approximation

$$\mathrm{var}\left[\sum_{j=N_{T-1}}^{k+1}\sum_{i=1}^{\tilde{R}_{j+1}} z(\tau_i^j)\right] \approx \sum_{j=N_{T-1}}^{k+1}\sum_{i=1}^{\tilde{R}_{j+1}}\mathrm{var}[z(\tau_i^j)]. \quad (B1)$$

It is possible to do this step exactly by noting that only z_i from the same family can be correlated and by including covariance terms between intrafamily $z(\tau_i)$; however, these terms have been found to be numerically insignificant.

The sum of the variances of $z(\tau_i)$ during an annealing step can be calculated using a method similar to that in Sec. III C. We begin by writing the values and probabilities of $z(\tau)$,

$$z(\tau) = \begin{cases} \tau - \lfloor\tau\rfloor & \text{w/prob. } \lceil\tau\rceil - \tau \\ \tau - \lceil\tau\rceil & \text{w/prob. } \tau - \lfloor\tau\rfloor, \end{cases} \quad (B2)$$

which means that we can write the variance explicitly,

$$\mathrm{var}[z(\tau)] = (\tau - \lfloor\tau\rfloor)^2(\lceil\tau\rceil - \tau) + (\tau - \lceil\tau\rceil)^2(\tau - \lfloor\tau\rfloor). \quad (B3)$$

For culling fraction small, we can make the approximation that within a single annealing step, τ is a Gaussian random variable. This allows us to replace the sum of variances with an integral that can be calculated explicitly. Using Eq. (B3),

$$\sum_{i=1}^{\tilde{R}_{j+1}}\mathrm{var}[z(\tau_i^j)] \approx \int_{-\infty}^{\infty}\mathrm{var}[z(\tau)]\mathcal{N}(\tau; 1, \Delta\beta\sigma_E)\,d\tau, \quad (B4)$$

$$= 2\epsilon - 2\pi\epsilon^2. \quad (B5)$$

Summing over all annealing steps and simplifying the expression assuming a constant culling fraction gives the desired result

$$\sum_{j=N_{T-1}}^{k+1}\sum_{i=1}^{\tilde{R}_{j+1}}\mathrm{var}[z(\tau_i^j)] \approx \sum_{j=N_{T-1}}^{k+1}\left[2\epsilon(l) - 2\pi\epsilon(l)^2\right], \quad (B6)$$

$$= (2\epsilon - 2\pi\epsilon^2)\,k. \quad (B7)$$

[1] R. H. Swendsen and J.-S. Wang, Replica Monte Carlo Simulations of Spin-Glasses, Phys. Rev. Lett. **57**, 2607 (1986).

[2] C. Geyer, Monte Carlo maximum likelihood for dependent data, in *Computing Science and Statistics: 23rd Symposium on the Interface*, edited by E. M. Keramidas (Interface, Fairfax Station, VA, 1991).

[3] E. Marinari and G. Parisi, Simulated tempering: A new Monte Carlo scheme, Europhys. Lett. **19**, 451 (1992).

[4] K. Hukushima and K. Nemoto, Exchange Monte Carlo method and application to spin glass simulations, J. Phys. Soc. Jpn. **65**, 1604 (1996).

[5] F. Wang and D. P. Landau, Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram, Phys. Rev. E **64**, 056101 (2001).

[6] F. Wang and D. P. Landau, Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States, Phys. Rev. Lett. **86**, 2050 (2001).

[7] A. Doucet, N. de Freitas, and N. Gordon, editors, *Sequential Monte Carlo Methods in Practice* (Springer-Verlag, New York, 2001).

[8] K. Hukushima and Y. Iba, Population annealing and its application to a spin glass, in *The Monte Carlo Method in the*

Physical Sciences: Celebrating the 50th Anniversary of the Metropolis Algorithm, Vol. 690, edited by J. E. Gubernatis (AIP, Washington, DC, 2003), p. 200.

[9] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, Optimization by simulated annealing, Science **220**, 671 (1983).

[10] J. Skilling, Nested sampling for general Bayesian computation, Bayesian Anal. **1**, 833 (2006).

[11] R. J. N. Baldock, N. Bernstein, K. M. Salerno, L. B. Partay, and G. Csanyi, Constant-pressure nested sampling with atomistic dynamics, Phys. Rev. E **96**, 043311 (2017).

[12] J. Machta, Population annealing with weighted averages: A Monte Carlo method for rough free-energy landscapes, Phys. Rev. E **82**, 026704 (2010).

[13] W. Wang, J. Machta, and H. Katzgraber, Population annealing: Theory and application in spin glasses, Phys. Rev. E **92**, 063307 (2015).

[14] W. Wang, J. Machta, and H. G. Katzgraber, Comparing Monte Carlo methods for finding ground states of Ising spin glasses: Population annealing, simulated annealing, and parallel tempering, Phys. Rev. E **92**, 013303 (2015).

[15] W. Wang, J. Machta, and H. G. Katzgraber, Bond chaos in spin glasses revealed through thermal boundary conditions, Phys. Rev. B **93**, 224414 (2016).

[16] W. Wang, Numerical simulations of Ising spin glasses with free boundary conditions: The role of droplet excitations and domain walls, Phys. Rev. E **95**, 032143 (2017).

[17] J. Callaham and J. Machta, Population annealing simulations of a binary hard-sphere mixture, Phys. Rev. E **95**, 063315 (2017).

[18] L. Y. Barash, M. Weigel, L. N. Shchur, and W. Janke, Exploring first-order phase transitions with population annealing, Eur. Phys. J. Spec. Top. **226**, 595 (2017).

[19] L. Y. Barash, M. Weigel, M. Borovsk, W. Janke, and L. N. Shchur, GPU accelerated population annealing algorithm, Comput. Phys. Commun. **220**, 341 (2017).

[20] A. Barzegar, C. Pattison, W. Wang, and H. G. Katzgraber, Optimization of population annealing Monte Carlo for large-scale spin-glass simulations, arXiv:1710.09025 (2017).

[21] H. G. Katzgraber, M. Palassini, and A. P. Young, Monte Carlo simulations of spin glasses at low temperatures, Phys. Rev. B **63**, 184422 (2001).

[22] G. Parisi, Infinite Number of Order Parameters for Spin-Glasses, Phys. Rev. Lett. **43**, 1754 (1979).

[23] G. Parisi, The order parameter for spin glasses: A function on the interval 0–1, J. Phys. A **13**, 1101 (1980).

[24] G. Parisi, Order Parameter for Spin-Glasses, Phys. Rev. Lett. **50**, 1946 (1983).

[25] W. L. McMillan, Domain-wall renormalization-group study of the three-dimensional random Ising model, Phys. Rev. B **30**, R476 (1984).

[26] D. S. Fisher and D. A. Huse, Ordered Phase of Short-Range Ising Spin-Glasses, Phys. Rev. Lett. **56**, 1601 (1986).

[27] D. S. Fisher and D. A. Huse, Absence of many states in realistic spin glasses, J. Phys. A **20**, L1005 (1987).

[28] A. J. Bray and M. A. Moore, Critical behavior of the three-dimensional Ising spin glass, Phys. Rev. B **31**, 631 (1985).

[29] C. M. Newman and D. L. Stein, Multiple states and thermodynamic limits in short-ranged Ising spin-glass models, Phys. Rev. B **46**, 973 (1992).

[30] P. L. Krapivsky, S. Redner, and E. Ben-Naim, *A Kinetic View of Statistical Physics* (Cambridge University Press, Cambridge, 2010).

[31] H. G. Katzgraber, M. Körner, and A. P. Young, Universality in three-dimensional Ising spin glasses: A Monte Carlo study, Phys. Rev. B **73**, 224432 (2006).

[32] R. A. Baños *et al.*, Nature of the spin-glass phase at experimental length scales, J. Stat. Mech. (2010) P06026.

[33] B. Yucesoy, J. Machta, and H. G. Katzgraber, Correlations between the dynamics of parallel tempering and the free-energy landscape in spin glasses, Phys. Rev. E **87**, 012104 (2013).

[34] S. S. Wilks, *Mathematical Statistics* (John Wiley & Sons, Inc., New York, 1962).

[35] S. Alder, S. Trebst, A. K. Hartmann, and M. Troyer, Dynamics of the Wang Landau algorithm and complexity of rare events for the three-dimensional bimodal Ising spin glass, J. Stat. Mech. (2004) P07008.