Generalized Replica Exchange Method

Jaegil Kim, a) Thomas Keyes, and John E. Straub

Department of Chemistry, Boston University, Boston, Massachusetts 02215, USA

(Received 24 March 2010; accepted 30 April 2010; published online 11 June 2010)

We present a powerful replica exchange method, particularly suited to first-order phase transitions associated with the backbending in the statistical temperature, by merging an optimally designed generalized ensemble sampling with replica exchanges. The key ingredients of our method are parametrized effective sampling weights, smoothly joining ordered and disordered phases with a succession of unimodal energy distributions by transforming unstable or metastable energy states of canonical ensembles into stable ones. The inverse mapping between the sampling weight and the effective temperature provides a systematic way to design the effective sampling weights and determine a dynamic range of relevant parameters. Illustrative simulations on Potts spins with varying system size and simulation conditions demonstrate a comprehensive sampling for phase-coexistent states with a dramatic acceleration of tunneling transitions. A significant improvement over the power-law slowing down of mean tunneling times with increasing system size is obtained, and the underlying mechanism for accelerated tunneling is discussed. © 2010 American Institute of Physics. [doi:10.1063/1.3432176]

I. INTRODUCTION

Recently, the temperature replica exchange method (rREM) [also called parallel tempering (PT)] has become a key workhorse for equilibrium sampling of a variety of complex systems across multidisciplinary fields. Simulating a set of replicas of the same system at a distribution of temperatures or microcanonical temperatures exhibits a negative slope region in Fig. 1(b), the so-called backbending or S-loop. The existence of the backbending has been verified in recent experiments on nuclear fragmentation and cluster melting, and its physical origin has been attributed to avoiding a “static” phase coexistence due to the free energy cost forming interfaces.

The backbending in $T_E$ manifests a bimodal structure in $P_T(E) \propto e^{-\beta F(E,T)}$, $F(E,T)=E-TS(E)$ being the Helmholtz free energy density and $\beta = [k_B T]^{-1}$ (k_B=1). With the uncertainty principle, $E_i^*$, of $F(E,T)$ determined by an extremum condition, $T_E(E_i^*)=T$, with $E_1^*<E_2^*<E_3^*$ for $T_2<T<T_1$. $P_T(E)$ becomes double peaked at $E_1^*$ and $E_3^*$ with a minimum at $E_2^*$. The stability condition, $\beta F^{\prime\prime}(E) = \beta^3 T^2_E(E)$, the prime being a differentiation with respect to $E$, reveals that the intermediate energy states between $E_1^*$ and $E_3^*$ in Fig. 1(b) corresponding to the backbending region in $T_E$ are intrinsically unstable for the canonical ensemble.

For a small system size, $L$, the canonical ensemble can sample both free energy minima, at $E_1^*$ and $E_3^*$, across the free energy barrier at $E_2^*$. However, as $L$ increases, the backbending energy region becomes inaccessible due to a high free energy barrier, implying that tunneling transitions between the two macroscopic phases become unlikely, and $P_T(E)$ becomes localized around $E_1^*$ or $E_3^*$, depending on whether $T<T_1$ or not. Accordingly, the acceptance of replica exchanges for a pair of inverse temperatures, $\beta$ and $\beta^*$, close to $\beta_1=1/T_1$, becomes exponentially suppressed as

$$A(\beta E; \beta^* E^*) = \min[1, e^{A(\beta E^* - E^*)}] = e^{-|\Delta E^*|},$$

where $\Delta E = E - E^*$ and $\Delta E^* = E_3^* - E_1^*$. Notice that $\Delta E = E - E^* = \Delta E^*$ for $\beta > \beta^*$, and $\Delta E = \Delta E^*$ for $\beta < \beta^*$, since $P_T(E)$ is centered at $E_1^*$ for $\beta > \beta_c$ and $E_3^*$ for $\beta < \beta_c$. We conclude that the instability of the canonical ensemble to the negative slope region in $T_E$ is the main cause of the poor

a)Electronic mail: jaegil@bu.edu.
acceptance of replica exchanges in the standard REM, with increasing system size.

An obvious way to restore the full power of replica exchanges is to utilize a noncanonical ensemble, avoiding the instability to the negative slope region in $T_E(E)$, and allowing a unimodal energy distribution. The Gaussian ensemble approach and its parallel version accomplish this goal by multiplying the Boltzmann factor by a Gaussian in energy.

In this paper, we propose a general framework to systematically build up optimized noncanonical ensembles, transforming unstable or metastable energy states of canonical ensembles into stable states in the presence of the back-bending in $T_E(E)$. Exploiting the one-to-one correspondence between the sampling weight and the effective temperatures, we develop an inverse mapping strategy, which determines a set of optimal generalized ensemble weights from the parametrized effective temperatures tailored to naturally bridge between ordered and disordered phases via a succession of unimodal energy distributions.

Simulations on Potts spins with varying system size, $L$, demonstrate that optimally designed generalized ensembles combined with replica exchanges (gREM) yield a dramatic acceleration of tunneling transitions and enable a comprehensive sampling of the phase transition region. A detailed mechanism for an order of magnitude acceleration in tunneling is discussed, with a quantitative performance comparison to the Wang–Landau (WL) method. Finite size scaling analysis reveals that the mean tunneling time, $\tau_\text{T}$, can be as favorably as $\propto L^{2\Delta}$ characteristic of a perfect random walk in flat histogram methods.

The paper is organized as follows: In Sec. II, the theory of the gREM is presented, with detailed simulation protocols. In Sec. III, the performance of the gREM is examined and quantitatively compared to that of WL sampling, for Potts spin systems under various simulation conditions. Conclusions and a brief summary are presented in Sec. IV.

II. GENERALIZED REPLICA EXCHANGE METHOD

A. Inverse mapping strategy

The key idea of the gREM is to construct a set of generalized ensemble weights, $W_\alpha(E;\lambda_\alpha)$ ($\alpha=1,2,\ldots,M$), which, as the parameter $\lambda_\alpha$ varies, successively access the unstable energy region between $E^*_\alpha$ and $E^*_{\alpha+1}$ in Fig. 1(b). Here $\alpha$ and $M$ are the replica index and the number of replicas, respectively. An important relation in the design of optimal weights is the inverse mapping between the $W_\alpha$ and the effective temperature

$$T_\alpha(E;\lambda_\alpha) = [\partial W_\alpha/\partial E]^{-1},$$

$w_\alpha = -\ln W_\alpha$ being the generalized effective potential. Notice that the definition of the effective temperature in Eq. (3) is analogous to that of the statistical temperature in Eq. (1). Based on the one-to-one correspondence in Eq. (3), the effective temperature completely determines the sampling weight up to a constant through the inverse mapping,

$$w_\alpha(E;\lambda_\alpha) = \int E^{T_\alpha(E;\lambda_\alpha)} 1/T_\alpha(z;\lambda_\alpha)dz.$$

A necessary and sufficient condition on $T_\alpha(E;\lambda_\alpha)$, such that unstable or metastable energy states of the canonical ensemble between $E^*_\alpha$ and $E^*_{\alpha+1}$ are transformed into stable ones with a unimodal PDF, is derived by identifying an extremum, $E^*_\alpha$, of a generalized free energy density, $\beta F_\alpha(E) = w_\alpha(E)$

$$T_\alpha(E^*_\alpha;\lambda_\alpha) = T_\alpha(E^*_\alpha),$$

and a stability condition

$$\beta F_\alpha(E^*_\alpha) = (\gamma_5 - \gamma_\alpha)/T^2_\alpha,$$

where $\gamma_5 = T_5(E^*_\alpha)$ and $\gamma_\alpha = T^\alpha_\alpha(E^*_\alpha;\lambda_\alpha)$. Based on Eqs. (5) and (6) we find that a unimodal distribution in the generalized PDF (GPD), i.e., $P_\alpha(E) = e^{-\beta F_\alpha}$, can arise from forming the unique crossing point, $E^*_\alpha$, between $T_\alpha(E;\lambda_\alpha)$, subject to $\gamma_\alpha(E^*_\alpha) < \gamma_5(E^*_\alpha)$.

More quantitatively, expanding the GPD up to second order yields

$$P_\alpha(E;\lambda_\alpha) \approx \exp\left(- (E - E^*_\alpha)^2/2\sigma_\alpha^2\right),$$

with $\sigma_\alpha \approx T^\alpha_\alpha/(\gamma_5 - \gamma_\alpha)$, illuminating that the “stable” in Fig. 2, in which $\gamma_\alpha < \gamma_5$, generates a Gaussian PDF centered at $E^*_\alpha$ with the positive $\sigma_\alpha$ even for a negative $\gamma_\alpha$. As $\gamma_\alpha$ further decreases to $-\infty$, the Gaussian in Eq. (7) approaches $\delta(E - E^*_\alpha)$, corresponding to the microcanonical ensemble case. On the other hand, $P_\alpha(E)$ becomes locally flat around $E^*_\alpha$, with $\gamma_\alpha = \gamma_5$ (“marginally” in Fig. 2). The crossing point, $E^*_\alpha$, becomes unstable for $\gamma_\alpha > \gamma_5$ as in the canonical ensemble.
FIG. 2. Statistical temperature $T_s(E)$ and effective temperatures $T_s(E) = T_c + \gamma_0(E-E_0)$ with varying $\gamma_0$. Unstable energy states around $E_2$ in the canonical ensemble ($\gamma_0=0$) become stable in the generalized ensemble of Eq. (10) with $\gamma_0<\gamma_S^{min}$. Marginal corresponds to $\gamma_0 = \gamma_S^{min}$.

B. Linear effective temperature: Tsallis weight

The simplest parametrization scheme for forming stable crossing points between $T_s(E;\lambda_\alpha)$ and $T_s(E)$ is to align linear effective temperatures in parallel with the constant slope, $\lambda_0$ [see Fig. 3(a)], as

$$T_s(E;\lambda_\alpha) = \lambda_0 + \gamma_0(E-E_0),$$

(8)

the control parameter $\lambda_\alpha$ being the $T$-intercept at an arbitrarily chosen $E_0$. To form the unique stable crossing point $E_\alpha$ in each replica, $\gamma_0$ must be less than the minimum slope $\gamma_S^{\min}$, $\gamma_S^{\min} = \min[T_s'(E)]$ being the minimum slope of $T_s(E)$ for the sampled energy region. Since $T_s(E)$ is monotonically increasing except for the transition region, in most cases a proper $\gamma_0$ is easily guessed from the approximate $T_s(E)$ by connecting a few points of $[\tilde{U}(T), T]$, $\tilde{U}(T)$ being an average energy of a short canonical run at $T$. For example, $\gamma_0$ can be simply chosen as $\gamma_L = (T_{M} - T_1)/(\tilde{U}_{M} - \tilde{U}_1)$, $T_1$ and $T_M$ being the lowest and highest temperature, and $\tilde{U}_{\alpha} = \tilde{U}(T_\alpha)$.

Once $\gamma_0$ is fixed the dynamic range of $\lambda_\alpha$ is determined to cover the interesting temperature range between $T_1$ and $T_M$ as $\lambda_1 = T_1$ and $\lambda_M = T_M - \lambda_0(\tilde{U}_M - \tilde{U}_1)$, with $E_0 = \tilde{U}_1$. The first and $M$th effective temperatures are chosen to cross $[\tilde{U}_1, T_1]$ and $[\tilde{U}_M, T_M]$, respectively. Then, the intermediate values of $\lambda_\alpha (1 < \alpha < M)$ are determined by equally dividing the parameter space as

$$\lambda_\alpha = \lambda_1 + (\alpha - 1)\Delta \lambda$$

and $\Delta \lambda = (\lambda_M - \lambda_1)/(M - 1)$.

Interestingly, the linear effective temperature of Eq. (8) produces a generic form of the Tsallis weight:

$$W_\alpha(E;\lambda_\alpha) \sim [\lambda_\alpha + \gamma_0(E-E_0)]^{-1/\gamma_0}.$$ 

(10)

Identifying $\gamma_0$ and $\lambda_\alpha$ by $(q-1)$ and $\beta^{-1}_\alpha$, respectively, $q$ being the nonextensivity parameter, recovers the original form of the Tsallis weight proposed in nonextensive statistical mechanics. Thus the gREM with the linear effective temperature of Eq. (8) is basically equivalent to the Tsallis-weight based REM, previously presented in the form of the generalized PT, $q$-REM, and Tsallis-REM.

However, it should be noted that the parametrized Tsallis weights in the gREM are targeted to transform unstable and metastable energy states of the canonical ensemble into stable ones, resulting in much narrower energy distributions than the canonical PDFs. On the other hand, other variants of the Tsallis-weight based REM are mainly focused on producing more delocalized energy distributions to increase an acceptance of replica exchanges. This is why the gREM utilizes $\lambda_\alpha$ as a control parameter rather than $q$ in the conventional implementation.

Another advantage of the gREM is that the inverse mapping enables a systematic selection of relevant parameters, which is particularly beneficial for the negative slope region of $T_s(E)$. The dynamic range of $\lambda_\alpha$ lies between $\lambda_1$ and $\lambda_M$ determined by two short canonical runs at $T_1$ and $T_M$, respectively. The inverse mapping is a general framework for the combination of any generalized ensemble sampling with replica exchanges. Depending on the profile of $T_s(E)$, characteristic of the phase transition, different types of effective temperatures can be designed to produce optimal weights. In the limit, $\gamma_0 \rightarrow 0$, the gREM recovers the REM running on temperatures $\lambda_\alpha$, denoting $W_\alpha(E;\lambda_\alpha) \sim e^{-(E-E_0)/\lambda_\alpha}$.

C. Simulation protocols of the gREM

Detailed simulation protocols of the gREM are outlined as follows.
TABLE I. Simulation parameters and mean tunneling times \( \tau_g \) for the gREM and the WL simulations for Potts spins with \( L=64 \). \( \tau_g \) represents a total simulation time in the gREM and a time spent for refining \( \bar{S}(E) \) up to \( f_g=10^{-5} \) in the WL. The gREM\(^*\) simulations are associated with Eq. (13) based on \( [E_{\alpha}^*,T_{\alpha}^*] \) determined by the gREM\(^\dagger\).

<table>
<thead>
<tr>
<th>Methods</th>
<th>( L )</th>
<th>( M )</th>
<th>( \gamma_0 )</th>
<th>( E_0 )</th>
<th>( \tau_g ) (MCS)</th>
<th>( t_s ) (MCS) ( \times 10^6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>gREM(_1)</td>
<td>64</td>
<td>30</td>
<td>-0.000010</td>
<td>-1.83</td>
<td>193.0</td>
<td>1.5</td>
</tr>
<tr>
<td>gREM(_2)</td>
<td>64</td>
<td>30</td>
<td>-0.000025</td>
<td>-1.83</td>
<td>388.5</td>
<td>1.5</td>
</tr>
<tr>
<td>gREM(_3)</td>
<td>64</td>
<td>30</td>
<td>-0.000075</td>
<td>-1.83</td>
<td>1511.5</td>
<td>1.5</td>
</tr>
<tr>
<td>gREM(_4)</td>
<td>64</td>
<td>30</td>
<td>-0.00015</td>
<td>-1.83</td>
<td>3950.8</td>
<td>1.5</td>
</tr>
<tr>
<td>gREM(_5)</td>
<td>64</td>
<td>30</td>
<td>-0.000075</td>
<td>-1.83</td>
<td>390.6</td>
<td>1.5</td>
</tr>
<tr>
<td>gREM(_6)</td>
<td>64</td>
<td>30</td>
<td>-0.000015</td>
<td>-1.83</td>
<td>739.4</td>
<td>1.5</td>
</tr>
<tr>
<td>gREM(_7)</td>
<td>64</td>
<td>30</td>
<td>-0.000075</td>
<td>-1.83</td>
<td>1553.7</td>
<td>1.5</td>
</tr>
<tr>
<td>WL</td>
<td>2.64</td>
<td></td>
<td></td>
<td></td>
<td>2.34 \times 10^5</td>
<td>1.53</td>
</tr>
</tbody>
</table>

(i) Perform short canonical runs at several temperatures between \( T_1 \) and \( T_M \) to determine the data set, \( \{ \bar{U}_{\alpha},T_{\alpha} \} \). Select a proper \( \gamma_0 \) to be less than \( \gamma_{\alpha}^{\text{max}} \) and determine \( \lambda_\alpha \) by employing Eq. (9) between \( \lambda_1=T_1 \) and \( \lambda_M=T_M-\gamma_0(\bar{U}_M-\bar{U}_1) \), with \( E_0=\bar{U}_1 \).

(ii) Run the gREM simulation in each replica by making trial moves in configuration space with the acceptance,

\[
A_{\text{trial}}(\mathbf{x} \rightarrow \mathbf{x}') = \min[1,e^{w_\alpha(E)-w_\alpha(E')}].
\]

Every fixed time step, attempt a replica exchange between neighboring replicas with the acceptance

\[
A_{\text{inter}}(\alpha;\mathbf{x}\mathbf{x}') = \min[1,\exp(\Delta_\alpha)],
\]

\[
\Delta_\alpha = w_{\alpha+1}(E') - w_{\alpha+1}(E) + w_\alpha(E) - w_\alpha(E').
\]

(iii) Once a sufficiently long production run has been performed, calculate the entropy estimate \( \bar{S}(E) \) by joining multiple generalized ensemble runs via the weighted histogram analysis method (WHAM).\(^{55}\)

### III. NUMERICAL SIMULATIONS AND RESULTS

To illustrate how effectively the gREM works around a typical first-order transition, we have chosen eight-state Potts spins as a benchmark, with the energy \( E=-\sum_{\langle ij \rangle} \delta(S_i,S_j) \), in which the sum runs over the nearest-neighbor spins on an \( L \times L \) square lattice and Kronecker \( \delta \) takes the value one if \( S_i=S_j \) and zero otherwise. Here \( S_i=1,2,\ldots, Q=8 \) are spin variables.

#### A. Characteristic features of the gREM

To determine the dynamic range of \( \lambda_\alpha \) and the optimal value of \( \gamma_0 \) we first performed short canonical Monte Carlo (MC) simulations for \( 2 \times 10^4 \) MC sweeps (MCSs) at \( T_1=0.7 \) and \( T_M=0.8 \), which determine \( \bar{U}_1=-7507.8 \) and \( \bar{U}_M=-3438.7 \). Here one MCS means \( L^2 \) MC trial moves of all spins. Based on \( \gamma_L=(T_M-T_1)/(\bar{U}_M-\bar{U}_1)=-2.5 \times 10^{-5} \) we performed various gREM simulations with varying \( \gamma_0 \) for \( L=64 \), as summarized in Table I. Replica exchanges were attempted every \( L \) MC moves and the ground state was used as the initial configuration in all replicas.

Setting \( E_0=\bar{U}_1 \) in Eq. (8), the dynamic range of \( \lambda_\alpha \), between \( \lambda_1=T_1 \) and \( \lambda_M=T_M-\gamma_0(\bar{U}_M-\bar{U}_1) \), depends on \( \gamma_0 \). For \( \gamma_0=0.000075 \) corresponding to the gREM\(_1\) in Table I, \( \lambda_1=0.7 \) and \( \lambda_M=1.1 \). Resulting effective temperatures (solid lines) in Fig. 3(a) are chosen to fully cover the phase transition region, including the backbending region between \( E_0^*=\pm 1.5L^2 \) and \( E_0^*=\pm 1.15L^2 \) in \( T_3(E) \). For comparison, we also plot the exact \( T_3(E) \), which was determined by the statistical temperature Monte Carlo (STMC) algorithm\(^{49}\) for \( 2 \times 10^8 \) MCS. All relevant parameters in the gREM\(_1\) have been chosen based on short canonical runs at \( T_1 \) and \( T_M \) and full knowledge of \( T_3(E) \) is not necessary.

Since \( T_3(E;\lambda_\alpha) \) was designed to form a unique, stable crossing point, \( E_\alpha^* \) with \( T_3(E) \), the resulting GPDFs in Fig. 3(b) are rapidly localized around \( E_\alpha^* \) with a Gaussian shape, and naturally bridge between ordered and disordered phases with unimodal energy distributions across the transition region. Since \( P_\alpha(E) \) is sharply peaked at \( E_\alpha^* \), \( T(E_\alpha^*,\lambda_\alpha) \approx T_3(E_\alpha^*) \), the set of most probable energies, \( [E_\alpha^*,T_{\alpha}^*] \), asymptotically converge toward a locus of \( T_3(E) \). Indeed, the profile of \( [E_\alpha^*,T_{\alpha}^*] \) shows a perfect coincidence with \( T_3(E) \) determined by STMC, and exactly correspond to crossing points between \( T_3(E) \) and \( T_3(E;\lambda_\alpha) \) in Fig. 3(a). For convenience, the most probable energy \( E_\alpha^* \) was approximated by the average energy summed over the orth replica.

The superimposed energy distribution, \( P_1(E) = (1/M) \sum_\alpha P_\alpha(E) \), shows an almost uniform sampling across the backbending region with a characteristic structure, resulting from narrowly peaked \( P_\alpha(E) \). Each energy trajectory sampled by \( W_\alpha(E;\lambda_\alpha) \) in Fig. 3(c) is constantly fluctuating around \( E_\alpha^* \) explicitly illustrating that the gREM achieves a comprehensive sampling for phase-coexistent states by transforming unstable energy states of the canonical ensembles into stable ones.

Actual replica trajectories initiated from \( \alpha=1 \) and 30 in Figs. 4(a) and 4(b) start to sample the whole dynamic energy range after \( 3 \times 10^4 \) MCS and exhibit very frequent tunneling transitions in both energy and replica spaces via the localized GPDFs across the transition region. For comparison we also performed a rREM simulation with \( M=30 \), employing an equidistant temperature allocation for the same temperature range between \( T_1 \) and \( T_M \). Two effectively disjointed sampling domains with no replica exchanges around \( \alpha=14 \) are
apparent in Figs. 4(c) and 4(d) due to a vanishing energy overlap between far-separated energy distributions of neighboring replicas at $T_c$. 

B. $\gamma_0$ dependence of the gREM

Varying $\gamma_0$ directly affects the gREM simulation by changing the width of GPDFs and the distribution of $E^*_\alpha$. These effects are entangled in Eq. (8) since decreasing $\gamma_0$ produces a more dense population of $E^*_\alpha$ for the transition region, as demonstrated in Fig. 5(a), but a narrower GPDF in each replica due to the increased $\sigma_\gamma = T^\gamma_{\alpha} / (\gamma_0 - \gamma_0)$. The collapse of $[E^*_\alpha, T^\gamma_{\alpha}]$ of both $gREM_1$ and $gREM_4$ into $T_\gamma(E)$ determined by STMC reveals that the most probable energy set indeed traces a locus of $T_\gamma(E)$. 

Due to the opposing effects of $\gamma_0$ for energy overlaps between neighboring replicas, the acceptance probability, $p_{acc}(\alpha)$, for replica exchanges between $\alpha$ and $(\alpha + 1)$ shows a nonmonotonic $\gamma_0$-dependence in Fig. 5(b). The rREM shows no replica exchanges around $T_\gamma = 0.74546$, with a minimum in $p_{acc}(\alpha)$ at $\alpha = 14$ ($T_{14} = 0.7448$ and $T_{15} = 0.7482$). On the other hand, a systematic enhancement of $p_{acc}(\alpha)$ shows up around $\alpha = 14$ in the gREM as the effective temperatures form stable crossing points in the transition region, upon decreasing $\gamma_0$ to $-0.000 025$. A further decrease to $\gamma_0 = -0.000 15$ yields an almost uniform acceptance, but an overall reduction in $p_{acc}(\alpha)$ due to significantly narrowed GPDFs.

The $\gamma_0$-dependence of the gREM is also clearly captured in the superimposed energy distributions, $P_T(E)$, in Fig. 5(c), illustrating that phase-coexistent energy states between $-1.5L^2$ and $-1.1L^2$ are only accessible for $\gamma_0 < \gamma_0^{\text{min}} = -0.000 005$. The existence of the threshold value $\gamma_0^{\text{min}}$ reveals that the success of the gREM relies on the transformation of unstable energy states of canonical ensembles into stable ones with optimized noncanonical ensembles. As steeper effective temperatures produce a denser population of $E^*_\alpha$ in the transition region, $P_T(E)$ becomes more flattened for the entire energy range, with characteristic oscillatory structures, stemming from sharply peaked $P_{ac}(E)$ as in the gREM$_4$.

Forming more stable crossing points, $E^*_\alpha$, in the transition region is crucial for sampling phase-coexistent states and smoothly joining ordered and disordered phases in the gREM. However, narrowed GPDFs with decreasing $\gamma_0$ is not desirable for $p_{acc}(\alpha)$, as shown in Fig. 5(a). To resolve this
is mainly due to more delocalized GPDFs in the region is more dramatic with a lower $g_{11002}$.

The systematic increase in $p_{\text{acc}}(\alpha)$, with an overall uniform acceptance, is obtained for gREM$_3^*$, in comparison with gREM$_3$, even with the same $\gamma_0=-0.000\,075$. The increase in $p_{\text{acc}}(\alpha)$ for the transition region is more dramatic with a lower $\gamma_0$, as in gREM$_1^*$ and gREM$_2^*$, corresponding to $\gamma_0=-0.000\,007\,5$ and $-0.000\,015$, respectively. The systematic increase in $p_{\text{acc}}(\alpha)$ is mainly due to more delocalized GPDFs in the gREM$^*$ simulations with increasing $\gamma_0$, while maintaining centers of Gaussians at nearly the same $E_{a^*}$, as illustrated in Fig. 6(b). More delocalized energy distributions maximize energy overlaps among replicas, leading to a significant acceleration of replica exchanges. As a result, the uniform energy sampling in $P_T(E)$ remains unchanged in Fig. 6(c) regardless of $\gamma_0$ and oscillatory structures are smoothed out. This implies that an optimal performance of the gREM is achieved by first determining densely populated $[E_{a^*}, T_{a^*}]$ with a steeper $\gamma_0$ and then switching to a production run with Eq. (13) at $\gamma_0 \approx \gamma_5^{\min}$.

### C. Accelerated tunneling transitions

To systematically quantify the performance of the gREM simulations we compute the number of tunneling transitions in energy, denoted by $N_E$, measuring how often all replicas make transitions from one phase to the other.$^{25,47,53,57,58}$ We also calculate the mean tunneling time, $\tau_T$, by determining the inverse of a linear slope of $N_E$ in production phases of the gREM simulations. Tunneling transitions were counted between the two boundary energies of $-1.80\,E_L^2$ and $-0.85L^2$.

Except for an equilibration stage, $N_E$ in Fig. 7(a) is linearly increasing as a function of the total simulation time. This implies that barrier crossing rates, which are proportional to the slope of $N_E$, are almost constant throughout the simulations. The most frequent tunneling transitions were observed in gREM$_1$ with $\gamma_0=-0.000\,01$, and $N_E$ systematically decreases as $\gamma_0$ decreases from $\gamma_5^{\min}$, even with a more comprehensive sampling and a better $p_{\text{acc}}(\alpha)$ for the phase transition region in Figs. 5(a) and 5(b).

The attenuation of $N_E$ with decreasing $\gamma_0$ is mainly due to the increased number of intermediate replicas in the transition region, causing tunneling to require more consecutive replica exchanges. On the other hand, tunneling in the gREM...
with $\gamma_0$ close to $\gamma_S^{\min}$ is rather rapid since a couple of replica exchanges among sparsely distributed intermediate replicas leads to a tunneling even with a relatively poor $p_{\text{acc}}(\alpha)$. However, the accelerated $N_E$ with $\gamma_0 = \gamma_S^{\min}$ does not necessarily lead to an enhanced convergence because the transition region is poorly sampled due to a lack of replicas, as seen in $P_f(E)$ of Fig. 5(c).

The effect of $\gamma_0$ on $N_E$ is more apparent in the gREM simulations in Fig. 7(b), in which the location of intermediate replicas around the transitions region is almost frozen. As expected, increasing $\gamma_0$, yielding more delocalized GPDFs centered at the same $E_\alpha$ of the gREM, results in a monotonic acceleration of $N_E$, while retaining a faithful sampling in the transition region even with $\gamma_0$ close to $\gamma_S^{\min}$. Approximately a tenfold speedup of $N_E$ is observed in gREM$^4$ compared to gREM$^3$.

For a comparison we also performed the entropic version of the WL sampling$^{48}$ for the energy range between $-1.83L^2$ and $-0.83L^2$. The entropy estimate $\tilde{S}(E)$ has been refined up to $f_\alpha=f=10^{-9}$, employing the dynamic update scheme $\tilde{S}(E) = S(E) + \ln f$. Following the original convention, the modification factor $\sqrt{f}$ constantly biases the system to move away from visited energy regions. The reduction in $f$ is very rapid initially, and $f_\alpha$ reaches $10^{-6}$ after $3.3 \times 10^7$ MCS, but significantly slows down as $f_\alpha$ becomes smaller, resulting in a total of $1.5 \times 10^8$ MCS up to $f_\alpha = 10^{-9}$.

The most remarkable point in Fig. 7(a) is the dramatic acceleration of $N_E$ in the gREM simulations over the WL. Examining the $\tau_E$ in Table I, reveals a speedup of tunneling transitions in the poorest case of our algorithm (gREM$^4$) of almost 60 times, increasing up to about 150 times for gREM$^3$. The acceleration of $N_E$ is more significant for the gREM simulations, in which replicas are more densely populated for the backbending regime regardless of $\gamma_0$ with Eq. (13). The ratio of tunneling times, $R_E = \tau_E^{\text{REM}} / \tau_E^{\text{WL}}$, is about 600 in the gREM$^4$. Here we only considered the simulations having a comprehensive sampling for the transition region, as in WL sampling.

What is the underlying mechanism for an order of magnitude acceleration of $N_E$ in the gREM over the WL? The bottleneck of barrier crossings in Potts spins is the formation of phase-coexistent states with interfacial regions, which becomes exponentially suppressed in the canonical ensemble. Flat histogram methods such as the multicanonical (MUCA) sampling$^{47}$, WL$^{48}$, and STMC (Ref. 49) alleviate this exponential slowing down by enhancing mixed-phase configurations, employing a sampling weight inversely proportional to the density of states. However, in a single replica simulation of the flat histogram type, forming mixed-phase configurations starting from ordered or disordered phases is intrinsically sequential and cumulative, and $\tau_E$ is mainly limited by the diffusion rate in energy space.

On the other hand, some intermediate replicas located at the transition region in the gREM always retain phase-coexistent states and naturally bridge the ordered and disordered phases through replica exchanges, as illustrated in Figs. 8(a)–8(f), in which fractions of spin states, $f_q = \sum f_i \delta(S_i, q) / L^2$, in the gREM, were plotted as a function of MCS for different $\alpha$. In the ordered phase ($\alpha=1$), most spins occupy a single spin state, while all spin states are equally probable in the disordered phase ($\alpha=30$). Intermediate replicas of $\alpha=13$, 15, and 17 exclusively sample mixed-phase configurations characterized by a majority state with interfacial regions, which become a single spin state, while all other states are preserved throughout the simulation. We conclude that the main limiting factor for tunneling in the gREM is the efficacy of replica exchanges among intermediate replicas sampling phase-coexistent states.

**D. Scaling behavior of $\tau_E$ and $T_{S}(E)$**

In a completely unbiased random walk in energy, corresponding to an ideal limit of flat histogram methods, $\tau_E$
scales as $L^{2d}$, with $d$ being the dimension of the system ($d=2$ in this paper), since the dynamic energy range expands as $L^d$. In an actual simulation, $\tau_E$ scales like $L^{2d+\varepsilon}$ due to a deviation from a perfect random walk.\textsuperscript{47,57} The value of $\varepsilon$ depends on the system and is known to be 0.65 for Potts spins and 0.73 for Ising spins, from studies using the MUCA (Ref. 47) and the flat histogram sampling exploiting the exact $S(E)$,\textsuperscript{57} respectively.

As summarized in Table II, we performed several WL simulations, varying $L$ from 16 to 128. The dynamic energy range was restricted between $E_1^*$ and $E_M^*$ determined by the gREM simulation at the same $L$. A similar scaling behavior $\tau_E^{WL} = 2.44 \times L^{2d+\varepsilon}$\textsuperscript{47,57} appears in the WL simulation for Potts spins, as seen in Fig. 9(a), in which the log-log plot of $\tau_E$ and $L$ shows a clear linear relationship. A slightly larger exponent of $\varepsilon=0.77$, in comparison with that of the MUCA,\textsuperscript{47} might be due to the difference in energy boundaries for tunneling. In our study, transitions were counted between $E_1^*$ and $E_M^*$, but tunneling transitions in the MUCA were counted between two free energy maxima at $T_c$, which have a much smaller separation.

To examine scaling properties of the $g$REM we performed two different sets of simulations with varying $L$, as summarized in Table III. Simulations in set I denoted by $g$REM\textsuperscript{I} are associated with the effective temperature of Eq. (8) with $M=30$ at the fixed $\gamma_0^I=-0.000 075$. Set II denoted by the $g$REM\textsuperscript{II} was performed with the scaled $\gamma_0^I = -0.000 075 \times (64/L)^2$ and $M=50$. Since the backbending region in $T_3(E)$ becomes flat with increasing $L$ [see Fig. 9(b)], $\gamma_0^I$ in the $g$REM\textsuperscript{II} is systematically lowered by the scaling factor $(64/L)^2$, producing more delocalized GPDFs and enhancing $p_{acc}(d)$ for the transition region. Since the simulation with $L=128$ in the $g$REM\textsuperscript{I} did not show statistically meaningful transitions, it was excluded in the scaling analysis.

The comparison of $\tau_E$ in Table III reveals that the speedup in $N_T$ with increasing $L$ is even more remarkable in both $g$REM\textsuperscript{I} and $g$REM\textsuperscript{II}. The ratio $R_E = \tau_E^{WL}/\tau_E^{g$REM$}$ increases from about $10^2$ order at $L=32$ to $10^4$ order at $L=128$. An apparent linear behavior in the log-log plot of $L$ and $\tau_E$ is seen in the $g$REM\textsuperscript{I} of Fig. 9(a). The scaling behavior follows $\tau_E^{g$REM$} = 0.047 \times L^{2d+0.5}$, implying that the acceleration of $N_T$ in the $g$REM\textsuperscript{I} is about $10^2$ order up to $L=100$. When more crossing points are created in the phase transition region with $M=50$, the growth of $\tau_E$ as a function of $L$ is weaker, with a lower slope in the log-log plot in the $g$REM\textsuperscript{II}, yielding $\tau_E^{g$REM$} = 0.156 \times L^{2d+0.0477}$. The scaling behavior of the $g$REM approaches that of a random walk in energy as more replicas are built up in the phase transition region.\textsuperscript{57}

The profiles of $T_3(E)$ represented by the most probable energy set $[E_n^*, T_n^*]$ show a clear backbending for $L=16$ and $L=32$ in Fig. 9(b), but the extent of the backbending gradually declines with increasing $L$ and becomes almost flat at $L=128$. In contrast to the backbending in the microcanonical calorlic curve, $T_3(E)$, the inverse of the internal energy, $T = U^{-1}(E)$, which corresponds to a calorlic curve in the canonical ensemble, monotonically increases across the phase transition region, implying that statistical ensembles are not equivalent in finite size systems.\textsuperscript{29,30}

### Table III. Simulation parameters and mean tunneling times $\tau_E$ for the $g$REM\textsuperscript{I} and $g$REM\textsuperscript{II} simulations with varying $L$. The $g$REM\textsuperscript{I} simulations are associated with $\gamma_0^I = -0.000 075$ and $M=30$. The $g$REM\textsuperscript{II} was performed with Eq. (13) based on $[E_n^*, T_n^*]$ of the $g$REM\textsuperscript{I}. $\tau_E^I$ and $\tau_E^II$ correspond to the $g$REM\textsuperscript{I} and $g$REM\textsuperscript{II}, respectively.

<table>
<thead>
<tr>
<th>Methods</th>
<th>$L$</th>
<th>$\gamma_0^I(M=30)$</th>
<th>$\tau_E^I$ (MCS)</th>
<th>$\tau_E^II$ (MCS)</th>
<th>$t_5$ (MCS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>gREM\textsuperscript{I}</td>
<td>16</td>
<td>-0.00012</td>
<td>48.1</td>
<td>43.7</td>
<td>$1.0 \times 10^7$</td>
</tr>
<tr>
<td>gREM\textsuperscript{I}</td>
<td>32</td>
<td>-0.00003</td>
<td>274.7</td>
<td>224.2</td>
<td>$7.0 \times 10^7$</td>
</tr>
<tr>
<td>gREM\textsuperscript{I}</td>
<td>50</td>
<td>-0.000012</td>
<td>851.0</td>
<td>487.1</td>
<td>$1.5 \times 10^8$</td>
</tr>
<tr>
<td>gREM\textsuperscript{I}</td>
<td>64</td>
<td>-0.0000075</td>
<td>1511.7</td>
<td>755.3</td>
<td>$1.5 \times 10^8$</td>
</tr>
<tr>
<td>gREM\textsuperscript{I}</td>
<td>80</td>
<td>-0.0000048</td>
<td>2710.2</td>
<td>1371.1</td>
<td>$1.5 \times 10^8$</td>
</tr>
<tr>
<td>gREM\textsuperscript{I}</td>
<td>100</td>
<td>-0.0000031</td>
<td>4739.1</td>
<td>1920.2</td>
<td>$1.5 \times 10^8$</td>
</tr>
<tr>
<td>gREM\textsuperscript{I}</td>
<td>128</td>
<td>-0.00000187</td>
<td>2998.3</td>
<td>$1.5 \times 10^8$</td>
<td></td>
</tr>
<tr>
<td>gREM\textsuperscript{II}</td>
<td>128</td>
<td>-0.0000038</td>
<td>139.5</td>
<td>$1.5 \times 10^8$</td>
<td></td>
</tr>
</tbody>
</table>
and the size dependent critical temperature, $\beta_c(L) = 1/T_c(L)$, approaches the thermodynamic transition temperature, $\beta_c^\infty = 1/T_c^\infty$. This asymptotic behavior can be predicted by finite size scaling analysis,61 as

$$\beta_c(L) = \beta_c^\infty + 1/L^d,$$

in which the critical temperature, $\beta_c(L)$, was determined as the temperature at the peak in $C_v(T)$.

As seen in Fig. 10(b), the critical temperatures determined by the gREM60 and the WL-MUCA are in quantitative agreement. Furthermore, plotting $\beta_c(L)$ versus $L^{-2}$ yields a good straight line, extrapolating to $\beta_c^\infty = 1.332 447$, close to the exact $\ln(1+\sqrt{q}=8)=1.342 454$.66 On the other hand, a noticeable deviation from a linear fit shows up in the scaling behavior of $\beta_c(L)$ in the original WL. To eliminate finite size effects at a smaller $L$, we used the data set between $L=50$ and $L=120$ for a linear fit.

### IV. Conclusions

The generalized Replica Exchange Method (gREM) has been developed to facilitate effective configurational sampling in systems exhibiting backbending, or an S-loop, in the statistical temperature, $T_S(E)$, characteristic of first-order phase transitions in finite systems. By combining optimally parametrized, generalized ensemble samplings with replica exchanges, our method enables a comprehensive sampling for phase transition regions with successive unimodal energy distributions, by transforming metastable or unstable energy states of canonical ensembles into stable ones. Exploiting the one-to-one correspondence between the sampling weight and the effective temperature, we also present an inverse mapping strategy, which determines optimal sampling weights from tailored effective temperatures, avoiding an intrinsic instability of the canonical ensemble to the negative slope region of $T_S(E)$.

The effectiveness of our method has been explicitly demonstrated for Potts spin systems, for various simulations conditions, as a function of the system size, $L$. The quantitative comparison between the gREM and WL sampling reveals that the gREM provides an order of magnitude acceleration of tunneling transitions over the WL, while maintaining a faithful sampling for the phase transition region as in flat histogram methods. The underlying mechanism for accelerated tunneling transitions is the capability of the gREM to preserve mixed-phase configurations in intermediate replicas located at the transition region. Finite size scaling analysis shows that in the optimum case of the gREM, the scaling of the mean tunneling time, $\tau_E$, approaches $\approx L^{2d}$, corresponding to the ideal limit of flat histogram methods. It is also shown that the gREM provides a correct canonical thermodynamics via the reweighting with a much smaller simulation time.

Finally, we would like to emphasize that the inverse mapping strategy is a general framework that can be applied to the combination of any generalized ensemble sampling and the replica exchange method. We anticipate that a hybrid extension augmented by the inverse mapping, applying the gREM to selective transition regions while retaining the
rREM for other energy regions, will allow the widespread use of our method to various systems with first-order transitions.

ACKNOWLEDGMENTS

We are grateful to the National Science Foundation (Grant Nos. CHE-0750309 and CHE-0848427) and the National Institutes of Health (Grant No. R01 GM076688) for the generous support of our research.