MULTISCALE MODEL. SIMUL.
Vol. 20, No. 1, pp. 220–249

ANALYSIS AND OPTIMIZATION OF CERTAIN PARALLEL MONTE CARLO METHODS IN THE LOW TEMPERATURE LIMIT*

PAUL DUPUIS† AND GUO-JHEN WU‡

Abstract. Metastability is a formidable challenge to Markov chain Monte Carlo methods. In this paper we present methods for algorithm design to meet this challenge. The design problem we consider is temperature selection for the infinite swapping scheme, which is the limit of the widely used parallel tempering scheme obtained when the swap rate tends to infinity. We use a recently developed tool for the analysis of the empirical measure of a small noise diffusion to transform the variance reduction problem into an explicit optimization problem. Our first analysis of the optimization problem is in the setting of a double-well model, and it shows that the optimal selection of temperature ratios is a geometric sequence except possibly the highest temperature. In the same setting we identify two different sources of variance reduction and show how their competition determines the optimal highest temperature. In the general multiwell setting we prove that the same geometric sequence of temperature ratios as in the two-well case is always nearly optimal, with a performance gap that decays geometrically in the number of temperatures.

Key words. parallel tempering, infinite swapping, Monte Carlo, large deviations, Gibbs measures, variance reduction

AMS subject classifications. 60F10, 65C05

DOI. 10.1137/21M1402029

1. Introduction.

Monte Carlo methods are among the most general purpose stochastic simulation methods currently available. However, rare events present a particular challenge for the design of efficient Monte Carlo methods. There is a relatively long history of the use of large deviation ideas in the design of algorithms for estimating probabilities of single rare events [6, 12], since large deviation results can be used to determine how the rare events are most likely to occur. But less is known on how to overcome the impact of rare events on Markov chain Monte Carlo (MCMC).

Parallel tempering (PT) [21, 16], also known as *replica exchange*, and a scheme obtained as a suitable limit and known as *infinite swapping* (INS) [11] are methods for accelerating MCMC. They work by coupling reversible Markov chains with different "temperatures" to enhance the sampling properties of the ensemble. An important question that remains to be answered is how to choose the temperatures in these algorithms.

In this paper, we apply recently developed methods for the analysis of the empirical measure of a small noise diffusion to characterize the optimal temperatures in the low temperature limit, which is the setting where the difficulties caused by rare events and related metastable behaviors are most severe. The analysis is done for the

*Received by the editors March 2, 2021; accepted for publication (in revised form) December 13, 2021; published electronically February 24, 2022.

https://doi.org/10.1137/21M1402029

Funding: This research was supported in part by the AFOSR (FA-9550-18-1-0214). The first author was supported in part by the National Science Foundation (DMS-1904992). The second author was supported in part by the Swedish e-Science Research Centre through the Data Science MCP.

†Division of Applied Mathematics, Brown University, Providence, RI 02912 USA (paul_dupuis@brown.edu).

‡Department of Mathematics, KTH Royal Institute of Technology, Stockholm, 11428 (gjwu@kth.se).

INS scheme, which is itself an optimized limit of parallel tempering, in part because of this optimality, and also in part because the large deviation properties needed for the analysis take a simpler form for INS than for PT. However, the conclusions regarding optimal temperature placements will also be at least approximately valid for parallel tempering if the swap rate is high enough that it approximates infinite swapping.

In the course of the analysis we are able to identify mechanisms that produce variance reduction, and we find that it has two sources. As will be discussed in detail later, one source of improved sampling is the increased mobility obtained by lowering the maximum energy barriers. A second and less obvious source of variance reduction is due to certain weights appearing in INS, which play a role reminiscent of the likelihood ratios that appear in importance sampling (see section 4.2). As it turns out, it is the weights that are responsible for most of the variance reduction, and which ultimately determine the proper placement of the temperatures in the low temperature limit.

The paper is organized as follows. The problem of interest is described in section 2. Various Monte Carlo methods, including PT and INS, are discussed in section 3, as are the performance measure we will use to characterize good performance. Section 4 states the main theoretical results of the paper and also includes a discussion of the mechanisms that produce variance reduction in the accelerated Monte Carlo methods. The proofs of our main results, Theorems 4.10 and 4.11 are given in sections 5 and 6, respectively. Moreover, a subsection in section 5 (subsection 5.2) gives examples and discusses bounds on crucial parameters that appear in Theorem 4.10. The appendix proves properties of certain zero cost trajectories associated with the INS process.

2. Problem formulation. We are concerned with computing integrals with respect to a Gibbs measure on the state space $\mathbb{R}^d$. The measure takes the form

$$\mu^{\varepsilon}(dx) \doteq \frac{1}{Z_{\mu}^{\varepsilon}} e^{-\frac{V(x)}{\varepsilon}} dx, \tag{2.1}$$

where the notation "$\doteq$" is understood as "is defined as" throughout the paper, $V : \mathbb{R}^d \to \mathbb{R}$ is the potential of a complex physical system, $\varepsilon > 0$ is proportional to a parameter that is interpreted as temperature in physical systems, and the normalization constant Z_{μ}^{ε} is typically unknown.[1] As an elementary example, one would like to estimate $\mu^{\varepsilon}(A)$ for a set $A \subset \mathbb{R}^d$ which does not contain the global minimum of V, with ∂A regular. Problems of this general sort occur in chemistry, physics, statistics, Bayesian statistics and elsewhere.

Under proper conditions on V, one can check using detailed balance that μ^{ε} is the unique stationary distribution of the diffusion process $\{X^{\varepsilon}(t)\}_{t\geq 0}$ satisfying the stochastic differential equation (SDE)

$$dX^{\varepsilon}(t) = -\nabla V\left(X^{\varepsilon}(t)\right) dt + \sqrt{2\varepsilon} dW(t), \tag{2.2}$$

where W is a d-dimensional standard Wiener process.

The empirical measure of $\{X^{\varepsilon}(t)\}_{t\geq 0}$ over the time interval $[0, T]$ is defined by

$$\lambda^{\varepsilon, T}(dx) \doteq \frac{1}{T} \int_0^T \delta_{X^{\varepsilon}(t)}(dx)\, dt, \tag{2.3}$$

[1] To be precise, in a physical system one would have $\varepsilon = k_B T$, where T is the temperature and k_B is Boltzmann's constant, but we abuse terminology and simplify notation by referring to ε as a temperature.

where δ_x is the Dirac measure at x. The ergodic theorem [3] implies that $\lambda^{\varepsilon,T}$ gives an approximation to μ^ε, and strictly speaking it is the use of discrete time analogues in this context that is known as MCMC, though we will also use the term for the continuous time model. For the particular problem of approximating $\mu^\varepsilon(A)$, we have the estimator

$$\theta^{\varepsilon,T}_{\mathrm{MC}} \doteq \lambda^{\varepsilon,T}(A) = \frac{1}{T}\int_0^T 1_A\left(X^\varepsilon(t)\right) dt. \tag{2.4}$$

We think of $\theta^{\varepsilon,T}_{\mathrm{MC}}$ as the most straightforward MCMC estimator of $\mu^\varepsilon(A)$, and since we will later introduce more complicated estimators, a subscript (e.g., MC) will be used to distinguish the different estimators.

In many applications (e.g., chemistry, physics, Bayesian inference, and counting [18, 20]), $V(x)$ is a complicated surface which contains multiple local minima of varying depths. The diffusion $\{X^\varepsilon(t)\}_{t\geq 0}$ can be trapped within these deep local minima for a long time before moving out to other parts of the state space, a phenomena sometimes referred to as *metastability* [2]. As a result, it requires a very long (exponential in $1/\varepsilon$) simulation time for $\lambda^{\varepsilon,T}$ to approximate the equilibrium μ^ε when ε is small.

Our analysis of the performance of computational approximations for μ^ε will be based on recently derived large deviation approximations for variances associated with empirical measures such as (2.4) [13]. Following the convention of [15, Chapter 6], [13] considers in place of, say, (2.2) a small noise diffusion that takes values in a compact and connected manifold $M \subset \mathbb{R}^d$ of dimension $r < d$ and with smooth boundary (precise regularity assumptions for M are given on [15, p. 135]). This is also consistent with how MCMC algorithms for a process such as (2.2) are often implemented. To be precise one uses periodic boundary conditions with the boundary far removed from the regions of interest and the potential V taking a large value on the boundary. For purposes of mathematics it is more convenient to identify the periodic domain with a smooth manifold in a space of larger dimension, such as a circle in $\mathbb{R}^2$ in place of a one-dimensional periodic domain or a torus in $\mathbb{R}^3$ for a square with periodic boundary conditions. However, for ease of discussion we will keep the notation of the SDE model, but with the understanding that we mean a diffusion process with the same local characteristics that takes values in the compact space M, with M locally equivalent to a Euclidean space.

Remark 2.1. In this paper we focus on the problem of computing integrals with respect to a Gibbs measure on a continuous state space. However, analogous results for discrete state systems are expected. See [7] for the formulation of infinite swapping for discrete state Markov process models.

3. Accelerated MCMC. In this section we introduce various alternative estimators of $\mu^\varepsilon(A)$ as in (2.1). Consider an ergodic Markov process $\{\bar{X}^\varepsilon(t)\}_t \subset \bar{M}$ and suppose that $\nu^\varepsilon \in \mathcal{P}(\bar{M})$ is the unique stationary distribution of $\{\bar{X}^\varepsilon(t)\}_t$. As an example, $\bar{M}$ could be $K \in \mathbb{N}$ products of the M just introduced. If we define $\theta^{\varepsilon,T}$ by

$$\theta^{\varepsilon,T} \doteq \frac{1}{T}\int_0^T f^\varepsilon\left(\bar{X}^\varepsilon(t)\right) dt \tag{3.1}$$

for a bounded and measurable function $f^\varepsilon : \bar{M} \to \mathbb{R}$ such that

$$\int_{\bar{M}} f^\varepsilon(\bar{x})\, \nu^\varepsilon(d\bar{x}) = \mu^\varepsilon(A),$$

then by the ergodic theorem [3], $\theta^{\varepsilon,T} \to \mu^\varepsilon(A)$ with probability 1 (w.p.1) as $T \to \infty$, which means one can also consider $\theta^{\varepsilon,T}$ as an approximation to $\mu^\varepsilon(A)$. We will consider several classes of estimators that are of the general form (3.1).

3.1. Passage from parallel tempering to infinite swapping. Parallel tempering is an algorithm used to speed up the sampling of a "slowly converging" Markov process, i.e., one for which the empirical measure converges slowly to the stationary distribution. Specifically, the idea of two-temperature parallel tempering is to introduce a *higher temperature* ε/α in addition to ε with $\alpha \in (0,1)$. If W_1 and W_2 are independent Wiener processes, then the empirical measure of the pair

$$\begin{cases} dX_1^\varepsilon = -\nabla V(X_1^\varepsilon)dt + \sqrt{2\varepsilon}dW_1, \\ dX_2^\varepsilon = -\nabla V(X_2^\varepsilon)dt + \sqrt{2\varepsilon/\alpha}dW_2 \end{cases} \tag{3.2}$$

gives an approximation to the Gibbs measure on $\mathbb{R}^d \times \mathbb{R}^d$ with density $\psi^\varepsilon(x_1,x_2) \propto e^{-V(x_1)/\varepsilon}e^{-\alpha V(x_2)/\varepsilon}$ for all $x_1, x_2 \in \mathbb{R}^d$, where "$\propto$" means "is proportional to." We will often suppress the argument t of $X_1^\varepsilon, X_2^\varepsilon$, especially when it appears in a wide display, and later on we will sometimes suppress t for other processes as well. If we allow *swaps* between X_1^ε and X_2^ε, i.e., X_1^ε and X_2^ε *exchange locations* with the state-dependent intensity $a\,(1 \wedge [\psi^\varepsilon(x_2,x_1)/\psi^\varepsilon(x_1,x_2)])$, where a is a positive constant and known as the swap rate, then we have a *Markov jump-diffusion* $(X_1^{\varepsilon,a}, X_2^{\varepsilon,a})$. Moreover, it is straightforward to check whether this new particle swapped process still satisfies detailed balance with respect to $\psi^\varepsilon(x_1,x_2)$ if this swapping intensity is used, and so can be used for numerical approximations.

It has been shown that various rates of convergence, such as the large deviation empirical measure rate [11] and the asymptotic variance, can be optimized by letting $a \to \infty$. This suggests one should consider the limit of $(X_1^{\varepsilon,a}, X_2^{\varepsilon,a})$ as $a \to \infty$ (the infinite swapping limit). This cannot be done directly with the particle swapped process $(X_1^{\varepsilon,a}, X_2^{\varepsilon,a})$ since, as discussed in [11], this process is not tight and hence does not converge in a meaningful way. An alternative perspective is to consider a temperature swapped process and approximate $\psi^\varepsilon(x_1,x_2)dx_1dx_2$ by a corresponding weighted empirical measure instead (see [11] for details). The advantage of doing so is that we have a well-defined weak limit process as $a \to \infty$, though as noted the empirical measure is replaced by a weighted analogue. The limit model is as follows. We define $(Y_1^\varepsilon, Y_2^\varepsilon)$ as the solution to

$$\begin{cases} dY_1^\varepsilon = -\nabla V(Y_1^\varepsilon)dt + \sqrt{2\varepsilon\rho^{\varepsilon,\alpha}(Y_1^\varepsilon,Y_2^\varepsilon) + 2\varepsilon\rho^{\varepsilon,\alpha}(Y_2^\varepsilon,Y_1^\varepsilon)/\alpha}dW_1, \\ dY_2^\varepsilon = -\nabla V(Y_2^\varepsilon)dt + \sqrt{2\varepsilon\rho^{\varepsilon,\alpha}(Y_1^\varepsilon,Y_2^\varepsilon)/\alpha + 2\varepsilon\rho^{\varepsilon,\alpha}(Y_2^\varepsilon,Y_1^\varepsilon)}dW_2 \end{cases} \tag{3.3}$$

and then define the weighted empirical measure of $(Y_1^\varepsilon, Y_2^\varepsilon)$ and its permutation $(Y_2^\varepsilon, Y_1^\varepsilon)$ by

$$\zeta^{\varepsilon,T}(dx_1dx_2) \doteq \frac{1}{T}\int_0^T \left[\rho^{\varepsilon,\alpha}(Y_1^\varepsilon,Y_2^\varepsilon)\delta_{(Y_1^\varepsilon,Y_2^\varepsilon)}(dx_1dx_2) + \rho^{\varepsilon,\alpha}(Y_2^\varepsilon,Y_1^\varepsilon)\delta_{(Y_2^\varepsilon,Y_1^\varepsilon)}(dx_1dx_2)\right]dt,$$

where

$$\rho^{\varepsilon,\alpha}(x_1,x_2) = \frac{e^{-\frac{1}{\varepsilon}[V(x_1)+\alpha V(x_2)]}}{e^{-\frac{1}{\varepsilon}[V(x_1)+\alpha V(x_2)]} + e^{-\frac{1}{\varepsilon}[V(x_2)+\alpha V(x_1)]}}.$$

(Note that $\rho^{\varepsilon,\alpha}(x_1,x_2) + \rho^{\varepsilon,\alpha}(x_2,x_1) = 1$, and that since we have passed to the limit we do not interpret $(Y_1^\varepsilon, Y_2^\varepsilon)$ as corresponding to any particular swap rate.) One can show that $\zeta^{\varepsilon,T}(dx_1dx_2)$ has precisely the same distribution as what one would obtain

by forming the ordinary empirical measure of the particle swapped process with swap rate a and by letting $a \to \infty$.

Furthermore, as shown in [11], one can consider the parallel tempering algorithm with more than two temperatures, and then by applying an analogous reasoning, there exists a corresponding limit process and a weighted empirical measure. These are presented in the next subsection.

Remark 3.1. We see that the infinite swapping scheme uses a *symmetrized* version of the original dynamics together with a *weighted empirical measure* to construct approximations to $\mu^{\varepsilon}(dx_1)\mu^{\varepsilon/\alpha}(dx_2)$. As noted previously, the weights $\rho^{\varepsilon,\alpha}$ will play an important role in the reduction of variance and are in some sense analogous to the likelihood ratio appearing in importance sampling [14].

3.2. Infinite swapping for K temperatures. In this subsection we introduce the K-temperature INS estimator for $K-1 \in \mathbb{N}$, which is the main object of study. We use the following notation: $\boldsymbol{x} \doteq (x_1, \ldots, x_K)$ denotes an element in M^K; $d\boldsymbol{x}$ denotes $dx_1 \cdots dx_K$; Σ_K is the collection of all permutations on $\{1, \ldots, K\}$; for any permutation $\sigma \in \Sigma_K$ and $\boldsymbol{x} \in M^K$, $\boldsymbol{x}_\sigma$ denotes $(x_{\sigma(1)}, \ldots, x_{\sigma(K)})$;

$$\Delta \doteq \left\{(x_1, \ldots, x_K) \in \mathbb{R}^K : 1 = x_1 \geq x_2 \geq \cdots \geq x_K > 0\right\};$$

$\boldsymbol{\alpha} \doteq (\alpha_1, \ldots, \alpha_K) \in \Delta$ denotes the K-temperature multiplication factors appearing in the definition of the K-temperature INS estimator. We note that $\boldsymbol{x}$ is only used for an element in the space M^K, and an element in any other space, such as $\mathbb{R}^d, M$, and $\bar{M}$, is denoted by x.

To define the K-temperature INS estimator for a given $\boldsymbol{\alpha}$, we consider the (symmetric) diffusion process $\{\boldsymbol{X}^{\varepsilon}(t)\}_{t\geq 0} = \{(X_1^{\varepsilon}(t), \ldots, X_K^{\varepsilon}(t))\}_{t\geq 0}$ on M^K satisfying

$$\begin{cases} dX_1^{\varepsilon} = -\nabla V\left(X_1^{\varepsilon}\right) dt + \sqrt{2\varepsilon}\sqrt{\rho_{11}^{\varepsilon}/\alpha_1 + \rho_{12}^{\varepsilon}/\alpha_2 + \cdots + \rho_{1K}^{\varepsilon}/\alpha_K}\, dW_1, \\ dX_2^{\varepsilon} = -\nabla V\left(X_2^{\varepsilon}\right) dt + \sqrt{2\varepsilon}\sqrt{\rho_{21}^{\varepsilon}/\alpha_1 + \rho_{22}^{\varepsilon}/\alpha_2 + \cdots + \rho_{2K}^{\varepsilon}/\alpha_K}\, dW_2, \\ \qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ dX_K^{\varepsilon} = -\nabla V\left(X_K^{\varepsilon}\right) dt + \sqrt{2\varepsilon}\sqrt{\rho_{K1}^{\varepsilon}/\alpha_1 + \rho_{K2}^{\varepsilon}/\alpha_2 + \cdots + \rho_{KK}^{\varepsilon}/\alpha_K}\, dW_K, \end{cases} \tag{3.4}$$

where $W_1, \ldots, W_K$ are independent Wiener processes and, for any $i, j \in \{1, \ldots, K\}$ and $\sigma \in \Sigma_K$, ρ_{ij}^{ε} denotes $\rho_{ij}^{\varepsilon}(\boldsymbol{X}^{\varepsilon}(t); \boldsymbol{\alpha})$ with

$$\rho_{ij}^{\varepsilon}\left(\boldsymbol{x}; \boldsymbol{\alpha}\right) \doteq \sum_{\sigma:\sigma(j)=i} w^{\varepsilon}\left(\boldsymbol{x}_\sigma; \boldsymbol{\alpha}\right),$$

and with

$$w^{\varepsilon}\left(\boldsymbol{x}; \boldsymbol{\alpha}\right) \doteq \frac{\exp[-\frac{1}{\varepsilon}\sum_{\ell=1}^{K} \alpha_\ell V\left(x_\ell\right)]}{\sum_{\sigma\in\Sigma_K} \exp[-\frac{1}{\varepsilon}\sum_{\ell=1}^{K} \alpha_\ell V(x_{\sigma(\ell)})]}. \tag{3.5}$$

Notice that when $K = 2$, the SDE (3.4) is the same as (3.3).

Using detailed balance, one can show that for each $\varepsilon \in (0, \infty)$, ν^{ε} is the unique stationary distribution of $\{\boldsymbol{X}^{\varepsilon}(t)\}_{t\geq 0}$, where

$$\nu^{\varepsilon}\left(d\boldsymbol{x}\right) \doteq \frac{1}{K! Z_\nu^{\varepsilon}} \sum_{\sigma\in\Sigma_K} \exp\left[-\frac{1}{\varepsilon}\sum_{\ell=1}^{K} \alpha_\ell V(x_{\sigma(\ell)})\right] d\boldsymbol{x} \tag{3.6}$$

with

$$Z_\nu^{\varepsilon} \doteq \int_{M^K} \exp\left[-\frac{1}{\varepsilon}\sum_{\ell=1}^{K} \alpha_\ell V\left(x_\ell\right)\right] d\boldsymbol{x}.$$

Remark 3.2. For any $\sigma \in \Sigma_K$, we also have

$$Z_\nu^\varepsilon = \int_{M^K} \exp\left[-\frac{1}{\varepsilon}\sum\nolimits_{\ell=1}^{K} \alpha_\ell V(x_{\sigma(\ell)})\right] d\boldsymbol{x}.$$

Let $\zeta^{\varepsilon,T}(d\boldsymbol{x})$ be the weighted empirical measure of $\{\boldsymbol{X}^\varepsilon(t)\}_{t\geq 0}$ over the time interval of length T given by

$$\zeta^{\varepsilon,T}(d\boldsymbol{x}) \doteq \frac{1}{T}\int_0^T \sum_{\sigma\in\Sigma_K} w^\varepsilon\left(\boldsymbol{X}_\sigma^\varepsilon(t);\boldsymbol{\alpha}\right)\delta_{\boldsymbol{X}_\sigma^\varepsilon(t)}(d\boldsymbol{x})dt.$$

It then follows from the ergodic theorem that $\zeta^{\varepsilon,T}$ converges in the topology of weak convergence of probability measures (and in fact in the stronger τ-topology [5]) to $\mu^{\varepsilon/\alpha_1}\times\mu^{\varepsilon/\alpha_2}\times\cdots\times\mu^{\varepsilon/\alpha_K}$ w.p.1 as $T\to\infty$. The K-temperature INS estimator of $\mu^\varepsilon(A)$ with parameter $\boldsymbol{\alpha}$ over time T is therefore defined by

$$\theta_{\mathrm{INS}}^{\varepsilon,T} \doteq \zeta^{\varepsilon,T}(A\times M^{K-1}) = \frac{1}{T}\int_0^T \sum_{\sigma\in\Sigma_K} w^\varepsilon\left(\boldsymbol{X}_\sigma^\varepsilon(t);\boldsymbol{\alpha}\right)1_A(X_{\sigma(1)}^\varepsilon(t))dt. \tag{3.7}$$

Remark 3.3. Besides $\mu^\varepsilon(A)$ for various choices of A, one is also interested in estimating risk-sensitive functionals of the form

$$\int_M e^{-\frac{1}{\varepsilon}F(x)}\mu^\varepsilon(dx)$$

for some nice (e.g., bounded and continuous) function $F: M\to\mathbb{R}$ as well as the analogous integrals with respect to some or all of the higher temperatures ε/α_ℓ. However, it is the lowest temperature ε/α_1 which is most challenging, and thus we focus on the problem of estimating $\mu^\varepsilon(A)=\mu^{\varepsilon/\alpha_1}(A)$ (recall that $\alpha_1=1$) but seek rates of decay for the relative error that are in some sense uniform in A.

Before discussing a property which makes it heuristically clear why one would expect $\theta_{\mathrm{INS}}^{\varepsilon,T}$ to do better than $\theta_{\mathrm{MC}}^{\varepsilon,T}$, we introduce the notion of implied potential.

DEFINITION 3.4. *Given a probability measure $\mu^\varepsilon(dx)=\phi^\varepsilon(x)dx$ on $\mathbb{R}^d$, we define the* implied potential *of μ^ε to be $-\varepsilon\log\phi^\varepsilon$.*

Example 3.5. If μ^ε is a Gibbs measure as in (2.1), then up to an additive constant the implied potential of μ^ε is V, the potential appearing in the dynamics (2.2).

From Example 3.5 we see that implied potential generalizes the notion of potential. By comparing the implied potential of ν^ε as in (3.6) and the implied potential of the product measure $\mu^{\varepsilon/\alpha_1}\times\cdots\times\mu^{\varepsilon/\alpha_K}$ with μ^ε as in (2.1), one can show that the maximum barrier of the implied potential of the former is smaller than that of the latter, provided that $\alpha_\ell<1$ for some $\ell\in\{2,\dots,K\}$ [19]. Since as is well known the barrier heights determine the exponential time scale of transitions between neighborhoods of local minima of the implied potential, this lowering of the energy barriers is expected to enhance the sampling of the entire space.

While it is intuitive that lowering energy barriers is helpful, it does not by itself lead to schemes that are in any sense optimal at low temperatures. A more important and open question in the design of the K-temperature INS estimator is how to select the ensemble of multiplicative factors $\boldsymbol{\alpha}$. In this paper we not only characterize the low temperature performance of a K-temperature INS estimator with a fixed set of

temperature factors $\boldsymbol{\alpha}$, but we also provide optimal and nearly optimal temperatures for problems of interest in the same low temperature limit. As we will see, the optimal temperature schedule is dominated by a geometric relation, and moreover is fairly insensitive to the particular numerical quantity of interest.

3.3. Performance measure. In this subsection we discuss the performance measure that will be used to characterize good performance of an estimator. Let $\{\bar{X}^\varepsilon\}_{\varepsilon\in(0,\infty)} \subset C([0,T] : \bar{M})$ be a sequence of stochastic processes that will be used to define an estimator. For complicated potentials V we expect these processes to exhibit metastability, which means that the time required for $\bar{X}^\varepsilon$ to visit the various parts of the state space that are needed for good estimation scales like $T^\varepsilon = e^{\frac{1}{\varepsilon}c}$ for some $c > 0$. As a consequence, if we wish to compare algorithms after they have become reasonably accurate we should assume the simulation interval scales in this way. Moreover, we will assume $T^\varepsilon = e^{\frac{1}{\varepsilon}c}$ for some $c \in (0,\infty)$ throughout this paper, though the value of c will depend on the particular context.

As noted in Remark 3.3, we focus on the problem of estimating $\mu^\varepsilon(A)$ for some set $A \subset M$, and assume there is a large deviation limit (i.e., $\lim_{\varepsilon\to 0} \varepsilon \log \mu^\varepsilon(A)$ exists).

DEFINITION 3.6. *An estimator* $\theta^{\varepsilon,T^\varepsilon}$ *of* $\mu^\varepsilon(A)$ *is called* essentially unbiased *if there is* $c_0 \in (0,\infty)$ *such that for any* $x \in \bar{M}$

$$\liminf_{\varepsilon\to 0} -\varepsilon \log\left(\left|E_x\theta^{\varepsilon,T^\varepsilon} - \mu^\varepsilon(A)\right|\right) \geq \lim_{\varepsilon\to 0} -\varepsilon\log\mu^\varepsilon(A) + c_0,$$

where E_x *is the conditional expectation given* $\bar{X}^\varepsilon(0) = x$.

This says that the bias of $\theta^{\varepsilon,T^\varepsilon}$ (i.e., the difference between $E_x\theta^{\varepsilon,T^\varepsilon}$ and $\mu^\varepsilon(A)$) decays strictly faster than $\mu^\varepsilon(A)$ as $\varepsilon \to 0$.

DEFINITION 3.7. *Given an estimator* $\theta^{\varepsilon,T^\varepsilon}$, *the* lower bound on the decay rate of the variance per unit time *of* $\theta^{\varepsilon,T^\varepsilon}$ *is defined as*

$$\inf_{x\in\bar{M}} \liminf_{\varepsilon\to 0} -\varepsilon\log\left(\mathrm{Var}_x\left(\theta^{\varepsilon,T^\varepsilon}\right)T^\varepsilon\right),$$

where Var_x *is the conditional variance given* $\bar{X}^\varepsilon(0) = x$. *If the* $\liminf$ *is a limit that does not depend on* x, *then we call it the* decay rate of the variance per unit time.

Remark 3.8. In this paper, we seek to optimize the decay rate of the variance per unit time (often referred to simply as the decay rate of the variance), but only among estimators that are essentially unbiased. A criticism is that essential unbiasedness depends on the time scaling $T^\varepsilon = e^{\frac{1}{\varepsilon}c}$, which may itself depend on the estimator. One may be concerned that improving the decay rate somehow lengthens the time till essential unbiasedness, namely, requiring larger c. While this is in fact possible, as we discuss in detail in Remark 4.12, this potential competition affects only the selection of the highest temperature, i.e., the choice of α_K, and is in fact not of great consequence at all.

Remark 3.9. We will take as our ideal performance benchmark a decay rate of the variance exactly *twice* $\lim_{\varepsilon\to 0} -\varepsilon\log\mu^\varepsilon(A)$. The reason is as follows. Suppose that we measure errors by the standard deviation (and assume essential unbiasedness). If we achieve this best possible decay rate, then the amount of time needed for the numerical error $\theta^{\varepsilon,T^\varepsilon} - \mu^\varepsilon(A)$ to be comparable to $\mu^\varepsilon(A)$ itself becomes subexponential in ε. See Remark 4.12 for more details.

Strictly speaking, $2\lim_{\varepsilon\to 0} -\varepsilon\log\mu^\varepsilon(A)$ is not the best possible decay rate of the variance, but rather the best practically achievable decay rate. Indeed, in analogy

with the zero variance estimator that one can define when using importance sampling for rare event estimation [4, 1], it is possible to define estimators with a larger decay rate. But these are not useful since they require information that is not typically available, such as knowing $\mu^\varepsilon(A)$. *Hence the aim in the design of an INS algorithm is to obtain a lower bound on the decay rate of the variance that is close to this maximum practical value, while at the same time reducing the growth rate of T^ε that is needed for essential unbiasedness.*

4. Statement of the main results. In this section we state the main results on the performance and optimal design of the INS scheme in the low temperature limit. The proofs involve applying the results of [13] and then simplifying the variational problem that characterizes the decay rate of the variance.

We present two main results. The first considers the restricted setting of a simple two-well model. In this case we can obtain a very precise reduction of the variational problem. Using this simplified expression, we can then probe in some detail the question of how INS achieves variance reduction. Our interest in this model is twofold. One reason is that with an exact expression (rather than a tight bound) for the solution to the variational problem we can explore issues relating to how variance reduction is obtained through swapping. The second is that it properly suggests very useful bounds for the general model. (While exact simplifications are possible there as well, the number of cases quickly becomes unwieldy as the number of local minima increases.) Since the proof of the reduction is long, we refer the reader to [22] for details.

The second and more important result is concerned with temperature selection when there are an arbitrary number of wells. Owing to this generality, we do not attempt to find the exact optimizer, but rather show that the geometric relation for temperatures suggested by the two-well model allows one to meet the design goal stated in Remark 3.9 at the end of the last section. In particular, there is a choice so that the rate of decay is arbitrarily close to the benchmark stated there, with the "gap" no larger than $(1/2)^{K-2}V(A)$, where $V(A) \doteq \inf_{x\in A} V(x)$, and the parameter c in the assertion of essential unbiasedness can be made small geometrically in K.

To apply the results of [13] we need to know that the INS process defined in (3.4) satisfies a large deviation principle (LDP) on $C([0,T] : M^K)$ for arbitrary $T \in (0,\infty)$. This is not straightforward, owing to the fact that the diffusion coefficients involve $w^\varepsilon(\boldsymbol{x};\boldsymbol{\alpha})$ defined in (3.5), which become discontinuous in $\boldsymbol{x}$ as $\varepsilon \to 0$. Hence one is concerned with the large deviation properties of processes with *discontinuous statistics* [10, 9].

The sorts of discontinuities encountered are in fact analogous to those encountered in the large deviation analysis of stochastic networks, such as multiclass queuing networks. A general approach to proving that a large deviation principle holds for stochastic networks appears in [9] and can be adapted to the INS model (3.4). It is important to note that we do not need the precise form of the rate function, but only that the LDP holds with some rate function and basic qualitative properties. This is because with the INS model we already have an expression for the stationary distribution. Various quantities are defined in [13] using the rate function that allow the identification of the Freidlin–Wentzell quasipotential and related objects. For the INS model the explicit formula for the stationary distribution directly identifies the quasipotential, thereby eliminating the need for the explicit form of the rate function. The technique of [9] is in fact ideally suited to showing the existence of an LDP without necessarily having an expression for the rate function, and in this paper we will simply assume that an LDP holds with some rate function.

4.1. Two-well model. Our first result considers the setting of a double-well potential. Let $V : \mathbb{R} \to \mathbb{R}$ ($d = 1$) be as in Figure 1.

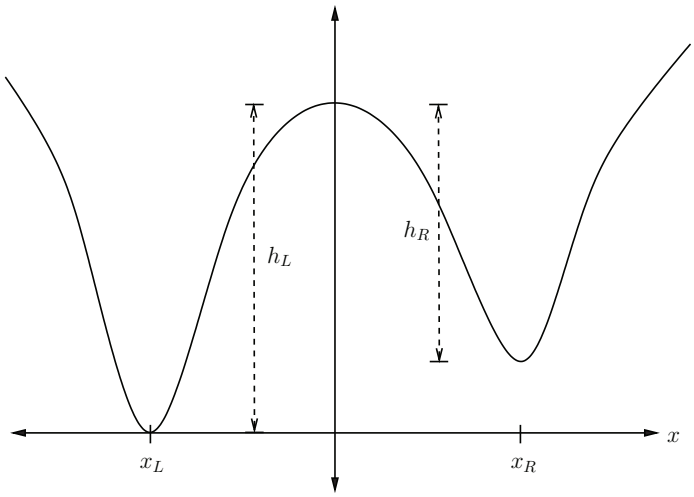


FIG. 1. *Asymmetric two-well model.*

Assume V satisfies the following condition.

CONDITION 4.1. *V is a bounded C^2 function and*

- *V is defined on a compact interval $D \subset \mathbb{R}$ and extended periodically as a C^2 function;*
- *V has only two local minima at x_L and x_R with values $V(x_L) < V(x_R)$;*
- *V has only one local maximum at $0 \in (x_L, x_R)$;*
- *$V(x_L) = 0$, $V(0) = h_L$ and $V(x_R) = h_L - h_R > 0$;*
- *$\inf_{x\in\partial D} V(x) > h_L$.*

Remark 4.2. As noted previously, the use of periodic boundary conditions is common in numerical implementation. It is assumed that the boundary is away from the neighborhoods of the equilibrium points of interest, and that the potential at the boundary is high enough that transitions across the boundary are unimportant. For our purposes, this means that the relevant large deviation calculations involve only paths that remain in D.

Remark 4.3. In the analysis of $\theta^{\varepsilon,T^\varepsilon}_{\mathrm{INS}}$ we will assume T^ε satisfies $T^\varepsilon = e^{\frac{1}{\varepsilon}c}$ with $c > \alpha_K h_L$. Recall that α_K is the smallest of the α_ℓ, and hence determines the highest temperature. As we will see, this condition ensures asymptotic unbiasedness.

The next result follows from [13, Theorems 4.3 and 4.5]. The theorem, in particular, characterizes the decay rate of the variance for the INS estimator for a given $\boldsymbol{\alpha}$. The proof of the theorem is analogous to the proof of Theorem 4.10, and so is omitted.

THEOREM 4.4. *Assume Condition* 4.1 *and that the process defined by* (3.4) *satisfies a large deviation principle that is uniform with respect to initial conditions* [4, section 1.2]. *Then for any closed interval $A \subset D$ with $x_L \notin A$ and $A = \bar{A}^\circ$,*

$$\theta^{\varepsilon,T^\varepsilon}_{\mathrm{INS}} = \frac{1}{T^\varepsilon}\int_0^{T^\varepsilon} \sum_{\sigma\in\Sigma_K} w^\varepsilon\left(\boldsymbol{X}^\varepsilon_\sigma(t)\,;\boldsymbol{\alpha}\right) 1_A(X^\varepsilon_{\sigma(1)}(t))dt \tag{4.1}$$

is an essentially unbiased estimator of $\mu^\varepsilon(A)$, *where* $w^\varepsilon(\boldsymbol{x};\boldsymbol{\alpha})$ *is given by* (3.5). *Moreover, for any* $\boldsymbol{\alpha}\in\Delta$ *and* $\boldsymbol{x}\in\mathbb{R}^K$, *we have*

$$\liminf_{\varepsilon\to 0} -\varepsilon\log\left(\mathrm{Var}_{\boldsymbol{x}}\left(\theta^{\varepsilon,T^\varepsilon}_{\mathrm{INS}}\right)T^\varepsilon\right)\geq\begin{cases}\hat{r}_1(\boldsymbol{\alpha})\wedge\hat{r}_3(\boldsymbol{\alpha}) & \text{if } A\subset(-\infty,0],\\ \hat{r}_1(\boldsymbol{\alpha})\wedge\hat{r}_2(\boldsymbol{\alpha}) & \text{if } A\subset[0,\infty),\end{cases}$$

where

$$\hat{r}_1(\boldsymbol{\alpha})\doteq\inf_{\boldsymbol{x}\in A\times\mathbb{R}^{K-1}}\left[2\sum_{\ell=1}^K\alpha_\ell V(x_\ell)-\min_{\sigma\in\Sigma_K}\left\{\sum_{\ell=1}^K\alpha_\ell V(x_{\sigma(\ell)})\right\}\right],$$

$$\hat{r}_2(\boldsymbol{\alpha})\doteq\min_{i\in\{2,\dots,K+1\}}\left\{2V(A)+\left[\sum_{\ell=1}^{i-2}\alpha_{K-\ell+1}-\alpha_{K-i+2}\right](h_L-h_R)\right\}-\alpha_K h_R,$$

and

$$\hat{r}_3(\boldsymbol{\alpha})\doteq 2V(A)-\alpha_K h_L.$$

Remark 4.5. As mentioned in [13, Conjecture 4.11], we expect that the lower bound is tight. The proof of the conjecture for a special case is outlined in [13, section 11].

Recall that the optimal decay rate of the variance per unit time is twice the large deviation decay rate of $\mu^\varepsilon(A)$, which is $V(A)$. The next result identifies optimizers over $\boldsymbol{\alpha}$ for the relevant variational problems. Note that in all cases we can get close to the best possible decay rate by choosing K appropriately, and in fact the gap goes to zero geometrically in K. For example, $K=7$ will get within 2% of the maximum rate of $2V(A)$. As we mentioned in the beginning of this section, since the proof of the variance reduction is long, we refer the reader to [22] for details.

THEOREM 4.6. *Assume the conditions of Theorem* 4.4. *For any closed set* $A\subset(-\infty,0]$ *with* $x_L\notin A$, *if* $V(A)\geq h_L$, *then*

$$\sup_{\boldsymbol{\alpha}\in\Delta}\left[\hat{r}_1(\boldsymbol{\alpha})\wedge\hat{r}_3(\boldsymbol{\alpha})\right]=2V(A)-(1/2)^{K-1}V(A)$$

with the optimal $\boldsymbol{\alpha}^*=(1,1/2,\dots,(1/2)^{K-2},(1/2)^{K-1})\in\Delta$. *If* $V(A)\leq h_L$, *then*

$$\sup_{\boldsymbol{\alpha}\in\Delta}\left[\hat{r}_1(\boldsymbol{\alpha})\wedge\hat{r}_3(\boldsymbol{\alpha})\right]=2V(A)-(1/2)^{K-2}\left(\frac{h_L}{V(A)+h_L}\right)V(A)$$

with the optimal $\boldsymbol{\alpha}^*=(1,1/2,\dots,(1/2)^{K-2},\frac{V(A)}{V(A)+h_L}(1/2)^{K-2})\in\Delta^{cl}$, *where* Δ^{cl} *is the closure of* Δ.

For any closed set $A\subset[0,\infty)$ *and if* $h_L\geq 2h_R$ *or* $V(A)\geq h_L$, *then*

$$\sup_{\boldsymbol{\alpha}\in\Delta}\left[\hat{r}_1(\boldsymbol{\alpha})\wedge\hat{r}_2(\boldsymbol{\alpha})\right]=2V(A)-(1/2)^{K-1}\left(V(A)\vee h_L\right)$$

with the optimal $\boldsymbol{\alpha}^*=(1,1/2,\dots,(1/2)^{K-2},(1/2)^{K-1})\in\Delta$. *If* $h_L\leq 2h_R$ *and* $V(A)\in[h_L-h_R,h_L]$, *then*

$$\sup_{\boldsymbol{\alpha}\in\Delta}\left[\hat{r}_1(\boldsymbol{\alpha})\wedge\hat{r}_2(\boldsymbol{\alpha})\right]=2V(A)-(1/2)^{K-2}\left(\frac{h_R}{V(A)-(h_L-2h_R)}\right)V(A)$$

with the optimal $\boldsymbol{\alpha}^*=(1,1/2,\dots,(1/2)^{K-2},\frac{V(A)-(h_L-h_R)}{V(A)-(h_L-2h_R)}(1/2)^{K-2})\in\Delta^{cl}$.

Remark 4.7. According to Theorems 4.4 and 4.6, no matter what set A is considered, the optimal temperatures $\boldsymbol{\alpha}^*$ form a geometric sequence with common ratio $1/2$, except possibly the last and smallest value, which corresponds to the highest temperature.

Remark 4.8. By Theorem 4.6, if $A \subset [0,\infty)$, $h_L \leq 2h_R$, and $V(A) = h_L - h_R$, the last component of the optimal temperature $\boldsymbol{\alpha}^*$ is 0. Of course the INS estimator is not well-defined with $\alpha_K^* = 0$. However, since $\hat{r}_1(\boldsymbol{\alpha}) \wedge \hat{r}_2(\boldsymbol{\alpha})$ is a continuous function of $\boldsymbol{\alpha}$, we can always approach the optimal performance by using $\boldsymbol{\alpha}$, which is close to $\boldsymbol{\alpha}^*$, e.g., $\boldsymbol{\alpha} = (1, 1/2, \ldots, (1/2)^{K-2}, \delta(1/2)^{K-2})$ for some $\delta \in (0,1)$.

Remark 4.9. Analogous results hold for a high-dimensional double-well potential $V : \mathbb{R}^d \to \mathbb{R}$, where x_L and x_R are the two local minima (and the former is the unique global minimum) and 0 is the unique local maximum. Moreover, one should interpret $(-\infty, 0]$ and $[0,\infty)$ as the closure of the domain of attraction of x_L and that of x_R, respectively.

4.2. Sources of variance reduction. Here we make some remarks on the form of the optimal $\boldsymbol{\alpha}$ and its interpretation regarding how variance reduction is achieved by INS. The remarks will also apply to parallel tempering to some extent if the swap rate is sufficiently high, though in this case the weights ρ used in INS are then implicitly computed by the algorithm, giving another sense in which INS is an optimized version of PT.

To begin, we note that the most obvious qualitative change when adding a higher temperature particle to one or more particles with lower temperature is that the "mobility," by which we mean the ease with which it crosses energy barriers, of the new particle is greater than that of all other particles. (What this means for INS is that the particle with the currently highest value of V is essentially given this temperature, with a slightly modified interpretation when two or more particles share the highest V value.)

Hence it is tempting to explain the improved sampling of INS, especially with respect to functionals that correspond to integration with respect to the lowest temperature, as a consequence of this greater mobility being passed between higher temperatures and lower temperatures. The mobility is passed via the swap mechanism with PT, and by the ρ weights with INS. For example, with PT the argument would be that the sharing of mobility between different temperatures obtained via swapping makes it easier for the low temperature particle to overcome potential barriers, and hence the empirical measure will converge more quickly. While plausible in a qualitative way, it is not clear, for example, how to relate the claim of faster convergence of the empirical measure to the properties of the variance. In fact, the situation is more complex.

In order to understand the role played by "mobility," in a previous paper [14] we introduced and studied what we call INS for IID, which stands for *infinite swapping* for *independent and identically distributed random variables*. The setting of that paper considers the integral of a distribution with respect to some risk-sensitive functional (including as a special case probabilities of sets with a positive large deviation rate, as is the case of Theorems 4.4 and 4.6). Because straightforward Monte Carlo will not work well, the paper follows the logic of parallel tempering but within the context of INS. It is assumed that the distribution (say μ^ε) is indexed by a parameter ε that corresponds to temperature here, and that a large deviation principle holds for $\{\mu^\varepsilon\}$ with a known rate function. This measure is then coupled with measures indexed by

higher values of the temperature using a parameter exactly analogous to $\boldsymbol{\alpha}$, and using symmetrization in the same way as INS one can define an estimator for integrals with respect to the lowest temperature using ρ weights in the way (suitable for the static setting) that is exactly analogous to what is done in the present paper for the Markov setting. Knowledge of the large deviation rate function is what allows for the explicit computation of the analogues of the ρ weights. This produces unbiased estimators analogous to those of the Markov setting, but for this purely static setting.

A key observation is the following. Since the setting of [14] does not involve any dynamics, the notion that any variance reduction is due to "increased mobility" is not possible. Indeed, as is discussed in [14] the ρ weights act in a way similar to the likelihood ratio in a well-designed importance sampling scheme, helping to cluster the values of the unbiased estimate around the true value, thereby reducing variance. We argue that the analogous property holds here, and that the primary role of the higher temperatures (except possibly the highest temperature) is to provides this variance reduction, and that solving the variational problems as in Theorem 4.6 tells us how to do this in the low temperature limit. Indeed, we obtain exactly the same geometric spacing of all temperatures (save the highest) in the low temperature limit in the Markovian setting as was obtained in the static setting. An analogous claim could be made regarding PT in the high swap rate setting, though as noted for PT the computation of the weights is carried out implicitly via the swaps and averaging in time.

While this motivates the form of the lower temperatures, it leaves out the highest temperature. Here we find a variety of behaviors that depend on the particular quantity that is being estimated, and one might argue that it is here that the mobility of a particle plays a role in determining the value of α_K. In all the cases of Theorem 4.6, we find that the optimal α_K is less than or equal to $(1/2)^{K-1}$, which is the value one finds in the static setting. However, Theorem 4.6 is concerned only with the decay rate of the variance per unit time. As noted in Remark 4.3, to reduce the constant c appearing in essential unbiasedness we would want to make α_K as small as possible. Given the relatively small impact that α_K has on the decay rate of the variance, one may prefer in practice to select its value so that the highest temperature particle is not impacted greatly by metastability, if such a temperature can be estimated or guessed.

4.3. Multiple-well model. The second main result considers a finite but otherwise arbitrary number of wells. While it is possible that one could derive results analogous to Theorem 4.6 which identify the optimizer appearing in the lower bound of Theorem 4.4, we will instead settle for showing that the geometric spacing suggested by the two-well model leads to a variance decay rate that can be made close to the optimum of $2V(A)$.

For the following theorem, we assume that $V : M \to \mathbb{R}$ is a smooth multiwell potential with a unique global minimum $y_1 \in M$, and without loss of generality we normalize V so that V takes value 0 at y_1 (i.e., $V(y_1) = 0$ and $V(x) > 0$ for all $x \in M$). Let H be the index set for equilibrium points of V, and let $y_i \in M$ be the equilibrium corresponding to index $i \in H$. We assume that the gradient of V is Lipschitz continuous, and we also assume that there exists a finite collection of points $\{O_i\}_{i\in L} \subset M^K$ with $L \doteq \{1, 2, \ldots, l\}$ for some $l \in \mathbb{N}$, such that $\cup_{i\in L}\{O_i\}$ coincides with the ω-limit set of the zero noise analogue of (3.4), so that $\cup_{i\in L}\{O_i\} = \{y_1, \ldots, y_{|H|}\}^K$, where $|H|$ is the total number of elements in H. This imposes some additional structure on V, and in particular rules out open regions on which V is a constant.

THEOREM 4.10. *Assume that the process defined by* (3.4) *satisfies a large deviation principle that is uniform with respect to initial conditions. Then there exists* $B \in (0,\infty)$ *such that the following hold. Consider any* $\boldsymbol{\alpha} \in \Delta$ *and* $\boldsymbol{x} \in M^K$*, and let* $T^\varepsilon = e^{\frac{1}{\varepsilon}c}$ *for some* $c > \alpha_K B$*. For any closed set* $A \subset M$*, define* $\theta^{\varepsilon,T^\varepsilon}_{\mathrm{INS}}$ *by* (4.1) *with this* $\boldsymbol{\alpha}$*. Then* $\theta^{\varepsilon,T^\varepsilon}_{\mathrm{INS}}$ *is essentially unbiased, and*

$$\liminf_{\varepsilon\to 0} -\varepsilon \log\left(\mathrm{Var}_{\boldsymbol{x}}(\theta^{\varepsilon,T^\varepsilon}_{\mathrm{INS}})T^\varepsilon\right) \geq r(\boldsymbol{\alpha}) - \alpha_K B,$$

where

$$r(\boldsymbol{\alpha}) \doteq \inf_{\boldsymbol{x}\in A\times M^{K-1}} \left\{ 2\sum_{\ell=1}^K \alpha_\ell V(x_\ell) - \min_{\sigma\in\Sigma_K}\left\{\sum_{\ell=1}^K \alpha_\ell V(x_{\sigma(\ell)})\right\}\right\}.$$

The parameter B that appears in the Theorem 4.10 depends only on V and K and is identified in Remark 5.13. In particular, it does not depend on ε. As will be illustrated by examples in subsection 5.3, B contains interesting information on how the geometry and other properties of the original potential V affect the rate of decay of the variance. For example, if the well that corresponds to the global minimum O_1 is also the most difficult well to escape from, then the situation of the multiple-well model is very similar to that of the two-well model. However, when this is not the case one can have $B > V(A)$, and B will depend on how the local minima are interconnected.

The next result identifies the optimizer of $r(\boldsymbol{\alpha})$ over $(\alpha_2,\dots,\alpha_{K-1})$ with a fixed α_K (recall that $\alpha_1 = 1$).

THEOREM 4.11. *For any closed set* $A \subset M$*,* $K-1 \in \mathbb{N}$ *and any* $\alpha_K \in (0,1]$*,*

$$\sup_{(\alpha_2,\dots,\alpha_{K-1})\in[\alpha_K,1]^{K-2}} r(\alpha_1,\alpha_2,\dots,\alpha_{K-1},\alpha_K)$$
$$= \begin{cases} (2-\alpha_K)V(A) \text{ if } \alpha_K \in [(1/2)^{K-1},1], \\ (2+\alpha_K-(1/2)^{K-2})V(A) \text{ if } \alpha_K \in (0,(1/2)^{K-1}]. \end{cases}$$

If $\alpha_K \in (0,(1/2)^{K-1}]$*, then the supremum is achieved at* $(\alpha_2^*,\dots,\alpha_{K-1}^*)$ *with* $\alpha_\ell^* = (1/2)^{\ell-1}$ *for* $\ell \in \{2,\dots,K-1\}$*. If* $\alpha_K \in ((1/2)^m,(1/2)^{m-1}]$ *for some* $m \in \{1,\dots,K-1\}$*, then the supremum is achieved at* $(\alpha_2^*,\dots,\alpha_{K-1}^*)$ *with*

$$\alpha_\ell^* = \begin{cases} (1/2)^{\ell-1} \text{ if } 2 \leq \ell \leq m, \\ \alpha_K \text{ if } m+1 \leq \ell \leq K-1. \end{cases}$$

Remark 4.12. Theorem 4.10 provides a lower bound for the decay of variance per unit time of an INS estimator with arbitrary $\boldsymbol{\alpha}$. Combining this with Theorem 4.11, for any given α_K one finds an associated optimal sequence of lower temperatures and the performance with such optimal temperatures. It remains to decide the value of α_K. As in the two-well case there is some conflict, in that by Theorem 4.10 we reduce the exponential time horizon required for essential unbiasedness by taking α_K small, while maximizing the decay rate for the variance requires $\alpha_K = (1/2)^{K-1}$. Since in all cases we end up with $\alpha_K \in (0,(1/2)^{K-1}]$, the selection of the lower temperatures as decided by Theorem 4.11 is unambiguous. However, given the geometric dependence in K of the coefficient of $V(A)$ in Theorem 4.11, one may choose to make α_K small to reduce the time required for essential unbiasedness. Also as in the two-well case, a natural interpretation of $\alpha_K = 0$ is that ε/α_K should be large enough that the corresponding single temperature process easily moves between different important local minima,

and one might choose to make this ratio independent of ε (so that α_K tends to zero as $\varepsilon \to 0$). Lastly we note the effect that increasing K has on c, where c is chosen to satisfy $c > \alpha_K B$ (for essential unbiasedness) and $c > ((1/2)^{K-2} - \alpha_K)V(A) + \alpha_K B$ (for bounded relative error, see Remark 3.9). It is easily checked that regardless of the choice of $\alpha_K \in (0, (1/2)^{K-1}]$, c decays exponentially in K.

Remark 4.13. From Remark 4.12 we know that regardless of the complexity of an energy landscape, in the low temperature regime the INS estimator with a geometric sequence of temperatures (save the highest temperature, for which we require $\alpha_K \in (0, (1/2)^{K-1}]$) performs well and reaches the optimal decay rate exponentially fast as $K \to \infty$. This suggests that the process defined by (3.4) with a geometric sequence of temperatures explores the landscape in an organized and meaningful way, and therefore could be useful in finding the global minimum of V. The use of INS for global optimization was first suggested in [8]. We conjecture here that INS and related processes in the low temperature regime with the geometric sequence of temperatures given in Theorem 4.11 will perform especially well in function minimization problems, and we will consider such issues in future work.

5. Proof of Theorem 4.10. We first recall notation from subsection 3.2 and introduce additional notation. Given $K - 1 \in \mathbb{N}$, for any $\boldsymbol{\alpha} \in \Delta$ we consider the diffusion process $\{\boldsymbol{X}^\varepsilon(t)\}_{t\geq 0} = \{(X_1^\varepsilon(t), \ldots, X_K^\varepsilon(t))\}_{t\geq 0}$ on M^K satisfying (3.4), and we denote $O_1 \doteq (y_1, \ldots, y_1)$, where y_1 is the unique global minimum of V. Figure 2 illustrates the points $\cup_{i\in L}\{O_i\}$ when V is the Franz potential and $K = 2$, with O_1, O_3, O_7, and O_9 local minima in the multidimensional potential defined in (5.1), O_2, O_4, O_6, and O_8 saddle points, and O_5 a local maximum.

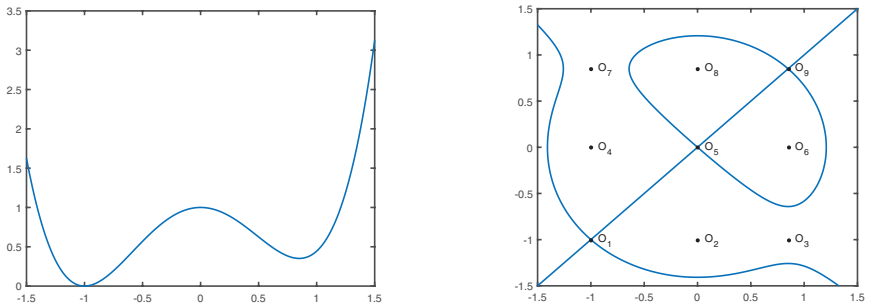


FIG. 2. *Franz potential* $\theta = 0.85$ *and equilibrium points of INS* $K = 2$.

In order to prove Theorem 4.10, we will utilize the large deviation results developed in [13], and to apply those results, we need to introduce several quantities that are constructed in terms of the Freidlin–Wentzell quasipotential. The quasipotential for (3.4) is easy to identify because the system is reversible with $\nu^\varepsilon \in \mathcal{P}(M^K)$ defined by (3.6) as its unique stationary distribution. Thus if for $\boldsymbol{x} \in M^K$ we define

$$U(\boldsymbol{x}) \doteq \min_{\sigma\in\Sigma_K} \left\{ \sum_{\ell=1}^{K} \alpha_\ell V(x_{\sigma(\ell)}) \right\}, \tag{5.1}$$

then U corresponds to a potential, and it is easy to see that $U(O_1) = 0$. Figure 3 depicts U for the Franz potential when $K = 2$.

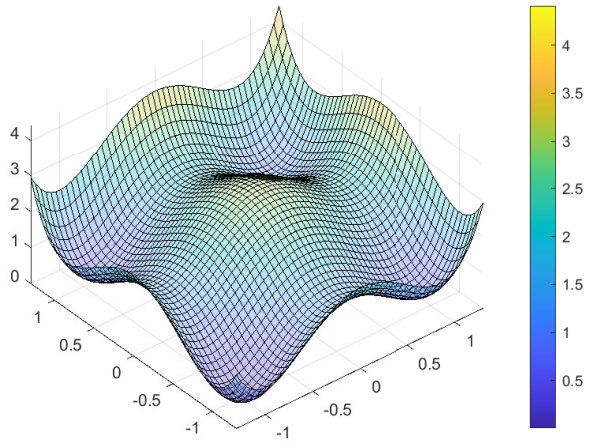


FIG. 3. *Symmetrized potential for* $K = 2$.

Since we assume that $\{\boldsymbol{X}^\varepsilon(t)\}_{0\le t\le T}$ satisfies a large deviation principle on $C([0,T] : M^K)$ with rate function $I_T : C([0,T] : M^K) \to [0,\infty]$ for arbitrary $T \in (0,\infty)$, the quasipotential $Q(\boldsymbol{x},\boldsymbol{y})$ is defined for all $\boldsymbol{x},\boldsymbol{y} \in M^K$ by

$$Q(\boldsymbol{x},\boldsymbol{y}) \doteq \inf\{I_T(\phi) : \phi(0) = \boldsymbol{x}, \phi(T) = \boldsymbol{y}, T < \infty\}.$$

(In fact the specific form of the quasipotential is already known since we know the rate function for the stationary distributions $\{\nu^\varepsilon\}$.)

Next we give a definition from graph theory which will be used in the proofs of the main results.

DEFINITION 5.1. *Given a subset* $W \subset L = \{1,\dots,l\}$, *a directed graph consisting of arrows* $i \to j$ $(i \in L \setminus W, j \in L, i \ne j)$ *is called a* W-graph on L *if it satisfies the following conditions:*

1. *Every point* $i \in L \setminus W$ *is the initial point of exactly one arrow.*
2. *For any point* $i \in L \setminus W$, *there exists a sequence of arrows leading from* i *to some point in* W.

We note that we can replace the second condition by the requirement that there are no closed cycles in the graph. We denote by $G(W)$ the set of W-graphs; we shall use the letter g to denote graphs.

Remark 5.2. We use $G(i)$ to denote $G(\{i\})$, and $G(i,j)$ to denote $G(\{i,j\})$.

DEFINITION 5.3. *For all* $j \in L$, *define*

$$W(O_j) \doteq \min_{g\in G(j)}\left[\sum_{(m\to n)\in g} V(O_m,O_n)\right], \tag{5.2}$$

$$W(O_1 \cup O_j) \doteq \min_{g\in G(1,j)}\left[\sum_{(m\to n)\in g} V(O_m,O_n)\right], \tag{5.3}$$

and

$$W(\boldsymbol{x}) \doteq \min_{i \in L} \left[W(O_i) + Q(O_i, \boldsymbol{x}) \right]. \tag{5.4}$$

Remark 5.4. Heuristically, if we interpret $V(O_m, O_n)$ as the "cost" of moving from O_m to O_n, then $W(O_j)$ is the "least total cost" of reaching O_j from every O_i with $i \in L \setminus \{j\}$.

We next prove a lemma that ties up the relation between W and U. The relation will also be used later on for solving the optimization problem

LEMMA 5.5. *For any* $\boldsymbol{x}, \boldsymbol{y} \in M^K$, $W(\boldsymbol{x}) - W(\boldsymbol{y}) = U(\boldsymbol{x}) - U(\boldsymbol{y})$.

Proof. Since we know that the stationary distribution ν^ε of $\{\boldsymbol{X}^\varepsilon(t)\}_{t \geq 0}$ is given by (3.6), we can apply [15, Theorem 4.3, Chapter 6] to find that for any $\eta > 0$ and for sufficiently small neighborhoods of $\boldsymbol{x}$ and $\boldsymbol{y}$,

$$\frac{\nu^\varepsilon(B_\delta(\boldsymbol{x}))}{\nu^\varepsilon(B_\delta(\boldsymbol{y}))} \leq \frac{\exp\left\{-\frac{1}{\varepsilon}\left(W(\boldsymbol{x}) - \min_{i \in L} W(O_i) - \eta\right)\right\}}{\exp\left\{-\frac{1}{\varepsilon}\left(W(\boldsymbol{y}) - \min_{i \in L} W(O_i) + \eta\right)\right\}} = e^{-\frac{1}{\varepsilon}(W(\boldsymbol{x}) - W(\boldsymbol{y}) - 2\eta)}$$

and

$$\frac{\nu^\varepsilon(B_\delta(\boldsymbol{x}))}{\nu^\varepsilon(B_\delta(\boldsymbol{y}))} \geq \frac{\exp\left\{-\frac{1}{\varepsilon}\left(W(\boldsymbol{x}) - \min_{i \in L} W(O_i) + \eta\right)\right\}}{\exp\left\{-\frac{1}{\varepsilon}\left(W(\boldsymbol{y}) - \min_{i \in L} W(O_i) - \eta\right)\right\}} = e^{-\frac{1}{\varepsilon}(W(\boldsymbol{x}) - W(\boldsymbol{y}) + 2\eta)}.$$

Thus

$$\limsup_{\varepsilon \to 0} -\varepsilon \log\left(\frac{\nu^\varepsilon(B_\delta(\boldsymbol{x}))}{\nu^\varepsilon(B_\delta(\boldsymbol{y}))}\right) \leq W(\boldsymbol{x}) - W(\boldsymbol{y}) + 2\eta$$

and

$$\liminf_{\varepsilon \to 0} -\varepsilon \log\left(\frac{\nu^\varepsilon(B_\delta(\boldsymbol{x}))}{\nu^\varepsilon(B_\delta(\boldsymbol{y}))}\right) \geq W(\boldsymbol{x}) - W(\boldsymbol{y}) - 2\eta.$$

On the other hand, for $\boldsymbol{w} = \boldsymbol{x}, \boldsymbol{y}$ the definition of U implies

$$\begin{aligned}\int_{B_\delta(\boldsymbol{w})} \exp\left\{-\frac{1}{\varepsilon} U(\boldsymbol{z})\right\} d\boldsymbol{z} &\leq \int_{B_\delta(\boldsymbol{w})} \left[\sum_{\sigma \in \Sigma_K} \exp\left\{-\frac{1}{\varepsilon} \sum_{\ell=1}^K \alpha_\ell V\left(z_{\sigma(\ell)}\right)\right\}\right] d\boldsymbol{z} \\ &\leq K! \cdot \int_{B_\delta(\boldsymbol{w})} \exp\left\{-\frac{1}{\varepsilon} U(\boldsymbol{z})\right\} d\boldsymbol{z}.\end{aligned}$$

Therefore

$$\begin{aligned}&\lim_{\varepsilon \to 0} -\varepsilon \log\left(\frac{\nu^\varepsilon(B_\delta(\boldsymbol{x}))}{\nu^\varepsilon(B_\delta(\boldsymbol{y}))}\right) \\ &= \lim_{\varepsilon \to 0} -\varepsilon \log\left(\frac{\int_{B_\delta(\boldsymbol{x})} \left[\sum_{\sigma \in \Sigma_K} \exp\left\{-\frac{1}{\varepsilon} \sum_{\ell=1}^K \alpha_\ell V\left(z_{\sigma(\ell)}\right)\right\}\right] d\boldsymbol{z}}{\int_{B_\delta(\boldsymbol{y})} \left[\sum_{\sigma \in \Sigma_K} \exp\left\{-\frac{1}{\varepsilon} \sum_{\ell=1}^K \alpha_\ell V\left(z_{\sigma(\ell)}\right)\right\}\right] d\boldsymbol{z}}\right) \\ &= \lim_{\varepsilon \to 0} -\varepsilon \log\left(\frac{\int_{B_\delta(\boldsymbol{x})} \exp\left\{-\frac{1}{\varepsilon} U(\boldsymbol{z})\right\} d\boldsymbol{z}}{\int_{B_\delta(\boldsymbol{y})} \exp\left\{-\frac{1}{\varepsilon} U(\boldsymbol{z})\right\} d\boldsymbol{z}}\right) \\ &= \min_{\boldsymbol{u} \in B_\delta(\boldsymbol{x})} U(\boldsymbol{u}) - \min_{\boldsymbol{u} \in B_\delta(\boldsymbol{y})} U(\boldsymbol{u}),\end{aligned}$$

where we obtain the last equality from Laplace's principle. Hence $\min_{\boldsymbol{u} \in B_\delta(\boldsymbol{x})} U(\boldsymbol{u}) - \min_{\boldsymbol{u} \in B_\delta(\boldsymbol{y})} U(\boldsymbol{u})$ is between $W(\boldsymbol{x}) - W(\boldsymbol{y}) \pm 2\eta$. Sending $\eta \to 0$ (and thus $\delta \to 0$), we find $W(\boldsymbol{x}) - W(\boldsymbol{y}) = U(\boldsymbol{x}) - U(\boldsymbol{y})$. ∎

Remark 5.6. By (5.4) and Lemma 5.5,

$$U(\boldsymbol{x}) = \min_{i \in L} \left[U(O_i) + Q(O_i, \boldsymbol{x}) \right].$$

We can now state the main result of [13]. The result stated in [13] assumes a fixed function f, but the result as stated below follows from this and the uniform convergence $f_\varepsilon \to f$. The uniformity of a large deviation principle with respect to the initial condition is discussed in [4, section 1.2]. Let

$$h \doteq \min_{i \in L \setminus \{1\}} Q(O_1, O_i) \qquad \text{and} \qquad w \doteq W(O_1) - \min_{i \in L \setminus \{1\}} W(O_1 \cup O_i). \tag{5.5}$$

The quantity h is related to the time that it takes for the process to leave a neighborhood of O_1, and $W(O_1) - W(O_1 \cup O_i)$ is related to the transition time from a neighborhood of O_i to one of O_1. The roles of h and w will be further explained in subsection 5.2.

THEOREM 5.7. *Assume that the process defined by* (3.4) *satisfies a large deviation principle that is uniform with respect to initial conditions, and let* ν^ε *be its unique stationary distribution and let* $T^\varepsilon = e^{\frac{1}{\varepsilon} c}$ *for some* $c > h \vee w$. *Suppose that for each* $\varepsilon > 0$, $f_\varepsilon : M^K \to \mathbb{R}$, *and that for a continuous function* $f : M^K \to \mathbb{R}$ *we have* $f_\varepsilon \to f$ *uniformly on* M^K. *Then for any compact set* $A \subset M^K$ *and* $\boldsymbol{x} \in M^K$,

$$\begin{aligned} \liminf_{\varepsilon \to 0} -\varepsilon \log & \left| E_{\boldsymbol{x}} \left(\frac{1}{T^\varepsilon} \int_0^{T^\varepsilon} e^{-\frac{1}{\varepsilon} f_\varepsilon (X_t^\varepsilon)} 1_A \left(X_t^\varepsilon \right) dt \right) - \int_{M^K} e^{-\frac{1}{\varepsilon} f_\varepsilon (\boldsymbol{x})} 1_A \left(\boldsymbol{x} \right) \nu^\varepsilon \left(d\boldsymbol{x} \right) \right| \\ & \geq \inf_{\boldsymbol{x} \in A} \left[f \left(\boldsymbol{x} \right) + W \left(\boldsymbol{x} \right) \right] - W \left(O_1 \right) + c - (h \vee w) \end{aligned}$$

and

$$\begin{aligned} \liminf_{\varepsilon \to 0} -\varepsilon \log & \left(T^\varepsilon \cdot \mathrm{Var}_{\boldsymbol{x}} \left(\frac{1}{T^\varepsilon} \int_0^{T^\varepsilon} e^{-\frac{1}{\varepsilon} f_\varepsilon (X_t^\varepsilon)} 1_A (X_t^\varepsilon) dt \right) \right) \\ & \geq \begin{cases} \min_{i \in L} (R_i^{(1)} \wedge R_i^{(2)}) & \text{if } h \geq w, \\ \min_{i \in L} (R_i^{(1)} \wedge R_i^{(2)} \wedge R_i^{(3)}) & \text{otherwise,} \end{cases} \end{aligned}$$

where for $i \in L$

$$R_i^{(1)} \doteq \inf_{\boldsymbol{x} \in A} [2 f(\boldsymbol{x}) + Q(O_i, \boldsymbol{x})] + W(O_i) - W(O_1),$$

$$R_1^{(2)} \doteq 2 \inf_{\boldsymbol{x} \in A} [f(\boldsymbol{x}) + Q(O_1, \boldsymbol{x})] - h,$$

for $i \in L \setminus \{1\}$

$$R_i^{(2)} \doteq 2 \inf_{\boldsymbol{x} \in A} [f(\boldsymbol{x}) + Q(O_i, \boldsymbol{x})] + W(O_i) - 2W(O_1) + W(O_1 \cup O_i),$$

and for $i \in L$

$$R_i^{(3)} \doteq 2 \inf_{\boldsymbol{x} \in A} [f(\boldsymbol{x}) + Q(O_i, \boldsymbol{x})] + 2W(O_i) - 2W(O_1) - w.$$

Proof of Theorem 4.10. After introducing all the necessary notation and results, we can now start the proof of Theorem 4.10, which consists of three steps:

- apply Theorem 5.7 to the INS model;
- bound $R_i^{(1)}$, $R_i^{(2)}$, and $R_i^{(3)}$ for every $i \in L$ in subsection 5.1;
- bound h and w in subsection 5.2.

For the first step, we note that the definition of $\theta_{\mathrm{INS}}^{\varepsilon,T^\varepsilon}$ involves the sum of a finite number of integrals of the form

$$\frac{1}{T^\varepsilon}\int_0^{T^\varepsilon} w^\varepsilon(\boldsymbol{X}_\sigma^\varepsilon(t);\boldsymbol{\alpha})1_A(X_{\sigma(1)}^\varepsilon(t))dt,$$

where $w^\varepsilon(\boldsymbol{x};\boldsymbol{\alpha})$ is defined in (3.5). In addition, we note that $\mu^\varepsilon(A)$ has an analogous decomposition:

$$\begin{aligned}\int_{M^K}\left(\sum_{\sigma\in\Sigma_K} w^\varepsilon\left(\boldsymbol{x}_\sigma;\boldsymbol{\alpha}\right)1_A(x_{\sigma(1)})\right)\nu^\varepsilon(d\boldsymbol{x}) &= K!\int_{M^K} w^\varepsilon\left(\boldsymbol{x};\boldsymbol{\alpha}\right)1_A(x_1)\nu^\varepsilon(d\boldsymbol{x})\\ &= \frac{1}{Z_\nu^\varepsilon}\int_{M^K}\exp\left[-\frac{1}{\varepsilon}\sum\nolimits_{\ell=1}^K \alpha_\ell V(x_\ell)\right]1_A(x_1)d\boldsymbol{x}\\ &= \frac{1}{Z_\mu^\varepsilon}\int_M e^{-\frac{1}{\varepsilon}V(x_1)}1_A(x_1)dx_1 = \mu^\varepsilon(A),\end{aligned}$$

where the first equality comes from the fact that ν^ε is permutation-invariant and the third equality holds since $Z_\nu^\varepsilon = Z_\mu^\varepsilon \times Z_\mu^{\varepsilon/\alpha_2}\times\cdots\times Z_\mu^{\varepsilon/\alpha_K}$. Thus it will be enough to obtain bounds on the differences of the corresponding terms. Also, since the difference is independent of the permutation, for simplicity of notation we take σ to be the identity.

Since V is bounded and continuous, it follows from standard features of the mollification used in the definition of w^ε in (3.5) that if we write $w^\varepsilon(\boldsymbol{x};\boldsymbol{\alpha})$ in the form $e^{-\frac{1}{\varepsilon}\sum_{\ell=1}^K \alpha_\ell V(x_\ell)+\frac{1}{\varepsilon}g_\varepsilon(\boldsymbol{x},\boldsymbol{\alpha})}$, then as $\varepsilon\to 0$

$$g_\varepsilon(\boldsymbol{x},\boldsymbol{\alpha})\to U(\boldsymbol{x}) \doteq \min_{\sigma\in\Sigma_K}\left[\sum_{\ell=1}^K \alpha_\ell V(x_{\sigma(\ell)})\right] \tag{5.6}$$

uniformly in $\boldsymbol{x}\in M^K$ (see, e.g., [4, Lemma 14.7]). Define

$$f(\boldsymbol{x},\boldsymbol{\alpha}) = \sum_{\ell=1}^K \alpha_\ell V(x_\ell) - U(\boldsymbol{x}).$$

We can then apply Theorem 5.7 with the function $f_\varepsilon(\boldsymbol{x},\boldsymbol{\alpha}) = \sum_{\ell=1}^K \alpha_\ell V(x_\ell) - g_\varepsilon(\boldsymbol{x},\boldsymbol{\alpha})$ and the compact set $A\times M^{K-1}\subset M^K$, to find that

$$\begin{aligned}\liminf_{\varepsilon\to 0} -\varepsilon\log\left|E_{\boldsymbol{x}}\frac{1}{T^\varepsilon}\int_0^{T^\varepsilon} w^\varepsilon(\boldsymbol{X}^\varepsilon(t);\boldsymbol{\alpha})1_A(X_1^\varepsilon(t))dt - \int_{M^K} w^\varepsilon\left(\boldsymbol{x};\boldsymbol{\alpha}\right)1_A(x_1)\nu^\varepsilon(d\boldsymbol{x})\right|&\\ \geq \inf_{\boldsymbol{x}\in A\times M^{K-1}}\left[f\left(\boldsymbol{x},\boldsymbol{\alpha}\right)+W\left(\boldsymbol{x}\right)\right] - W\left(O_1\right) + c - (h\vee w)&\\ = \inf_{\boldsymbol{x}\in A\times M^{K-1}}\left[f\left(\boldsymbol{x},\boldsymbol{\alpha}\right)+U\left(\boldsymbol{x}\right)\right] + c - (h\vee w)&\\ = \inf_{\boldsymbol{x}\in A\times M^{K-1}}\left[\sum_{\ell=1}^K \alpha_\ell V(x_\ell)\right] + c - (h\vee w) \geq V(A) + c - (h\vee w).&\end{aligned} \tag{5.7}$$

Combining (5.7) with the facts that $-\varepsilon \log \mu^\varepsilon(A) \to V(A)$ and $c > h \vee w$ shows that $\theta^{\varepsilon,T^\varepsilon}_{\mathrm{INS}}$ is essentially unbiased. Moreover, we find that $\liminf_{\varepsilon\to 0} -\varepsilon \log(T^\varepsilon \cdot \mathrm{Var}_{\boldsymbol{x}}(\theta^{\varepsilon,T^\varepsilon}_{\mathrm{INS}}))$ is bounded below by either $\min_{i\in L}(R_i^{(1)}(\boldsymbol{\alpha}) \wedge R_i^{(2)}(\boldsymbol{\alpha}))$ or $\min_{i\in L}(R_i^{(1)}(\boldsymbol{\alpha}) \wedge R_i^{(2)}(\boldsymbol{\alpha}) \wedge R_i^{(3)}(\boldsymbol{\alpha}))$, depending on whether $h \geq w$ or $w > h$.

We can now complete the proof assuming bounds proved in the next two sections. Specifically, by Lemma 5.9 we find that both minima are bounded below by $r(\boldsymbol{\alpha}) - h \vee w$ with

$$r(\boldsymbol{\alpha}) \doteq \inf_{\boldsymbol{x}\in A\times M^{K-1}} \left\{ 2\sum_{\ell=1}^K \alpha_\ell V(x_\ell) - \min_{\sigma\in\Sigma_K} \left\{ \sum_{\ell=1}^K \alpha_\ell V(x_{\sigma(\ell)}) \right\} \right\}. \tag{5.8}$$

The lower bound on the variance as stated in the theorem can then be obtained by using Lemmas 5.10 and 5.11 to bound h and w, respectively, and the constant B can then be identified easily and is displayed in Remark 5.13. ∎

In the next subsection, we will establish the aforementioned lower bound for these two minima.

Remark 5.8. As mentioned in Remark 3.3, we are also interested in estimating risk-sensitive functionals of the form

$$\int_M e^{-\frac{1}{\varepsilon}F(x)} \mu^\varepsilon(dx).$$

We can apply Theorem 5.7 to the associated INS estimator in this case as well by using the function $f_\varepsilon(\boldsymbol{x},\boldsymbol{\alpha}) = F(x_1) + \sum_{\ell=1}^K \alpha_\ell V(x_\ell) - g_\varepsilon(\boldsymbol{x},\boldsymbol{\alpha})$ and the compact set M^K. Moreover, one can modify the arguments in subsection 5.1 to derive an analogous version of Theorem 4.10 for the risk-sensitive functional case.

5.1. Bounds for the optimization problem. In this subsection we provide suitable lower bounds for $\min_{i\in L}(R_i^{(1)}(\boldsymbol{\alpha}) \wedge R_i^{(2)}(\boldsymbol{\alpha}))$ and $\min_{i\in L}(R_i^{(1)}(\boldsymbol{\alpha}) \wedge R_i^{(2)}(\boldsymbol{\alpha}) \wedge R_i^{(3)}(\boldsymbol{\alpha}))$. Define $r(\boldsymbol{\alpha})$ by (5.8), which is the same as $\inf_{\boldsymbol{x}\in A\times M^{K-1}} \{2f(\boldsymbol{x},\boldsymbol{\alpha}) + U(\boldsymbol{x})\}$, where $f(\boldsymbol{x},\boldsymbol{\alpha}) \doteq \sum_{\ell=1}^K \alpha_\ell V(x_\ell) - U(\boldsymbol{x})$. We will show that both minima are bounded below by quantities slightly smaller than $r(\boldsymbol{\alpha})$. Actually, we will find lower bounds for $\min_{i\in L} R_i^{(k)}(\boldsymbol{\alpha})$ for $k = 1, 2$, and 3, individually. The precise statement is given in the following lemma.

LEMMA 5.9. *For any $\boldsymbol{\alpha} \in \Delta$, we have $\min_{i\in L} R_i^{(1)}(\boldsymbol{\alpha}) = r(\boldsymbol{\alpha})$, $\min_{i\in L} R_i^{(2)}(\boldsymbol{\alpha}) \geq r(\boldsymbol{\alpha}) - h \vee w$, and $\min_{i\in L} R_i^{(3)}(\boldsymbol{\alpha}) \geq r(\boldsymbol{\alpha}) - w$.*

Proof. First note that

$$\begin{aligned}
\min_{i\in L} R_i^{(1)}(\boldsymbol{\alpha}) &= \min_{i\in L} \left(\inf_{\boldsymbol{x}\in A\times M^{K-1}} \{2f(\boldsymbol{x},\boldsymbol{\alpha}) + Q(O_i,\boldsymbol{x})\} + W(O_i) - W(O_1) \right) \\
&= \inf_{\boldsymbol{x}\in A\times M^{K-1}} \left\{ 2f(\boldsymbol{x},\boldsymbol{\alpha}) + \min_{i\in L} [Q(O_i,\boldsymbol{x}) + W(O_i)] - W(O_1) \right\} \\
&= \inf_{\boldsymbol{x}\in A\times M^{K-1}} \{2f(\boldsymbol{x},\boldsymbol{\alpha}) + W(\boldsymbol{x}) - W(O_1)\} \\
&= \inf_{\boldsymbol{x}\in A\times M^{K-1}} \{2f(\boldsymbol{x},\boldsymbol{\alpha}) + U(\boldsymbol{x})\} = r(\boldsymbol{\alpha}),
\end{aligned}$$

where we use (5.4) for the third equality and Lemma 5.5 for the fourth equality.

Moreover, since

$$\begin{aligned}
&\min_{i\in L\setminus\{1\}} R_i^{(2)}(\boldsymbol{\alpha})\\
&= \min_{i\in L\setminus\{1\}}\left[2\inf_{\boldsymbol{x}\in A\times M^{K-1}}[f(\boldsymbol{x},\boldsymbol{\alpha})+Q(O_i,\boldsymbol{x})]+W(O_i)-2W(O_1)+W(O_1\cup O_i)\right]\\
&\geq \inf_{\boldsymbol{x}\in A\times M^{K-1}}[2f(\boldsymbol{x},\boldsymbol{\alpha})+\min_{i\in L\setminus\{1\}}\{Q(O_i,\boldsymbol{x})+W(O_i)-W(O_1)\}]\\
&\quad -W(O_1)+\min_{i\in L\setminus\{1\}} W(O_1\cup O_i)\\
&= \inf_{\boldsymbol{x}\in A\times M^{K-1}}[2f(\boldsymbol{x},\boldsymbol{\alpha})+\min_{i\in L\setminus\{1\}}\{Q(O_i,\boldsymbol{x})+U(O_i)\}]-w,
\end{aligned}$$

using $U(O_1)=0$ and $Q\geq 0$ we obtain

$$\begin{aligned}
\min_{i\in L} R_i^{(2)}(\boldsymbol{\alpha}) &= R_1^{(2)}(\boldsymbol{\alpha})\wedge\left(\min_{i\in L\setminus\{1\}} R_i^{(2)}(\boldsymbol{\alpha})\right)\\
&\geq \left(\inf_{\boldsymbol{x}\in A\times M^{K-1}}[2f(\boldsymbol{x},\boldsymbol{\alpha})+Q(O_1,\boldsymbol{x})]-h\right)\\
&\qquad\wedge\left(\inf_{\boldsymbol{x}\in A\times M^{K-1}}[2f(\boldsymbol{x},\boldsymbol{\alpha})+\min_{i\in L\setminus\{1\}}\{Q(O_i,\boldsymbol{x})+U(O_i)\}]-w\right)\\
&\geq \inf_{\boldsymbol{x}\in A\times M^{K-1}}[2f(\boldsymbol{x},\boldsymbol{\alpha})+\min_{i\in L}\{Q(O_i,\boldsymbol{x})+U(O_i)\}]-h\vee w\\
&= \inf_{\boldsymbol{x}\in A\times M^{K-1}}[2f(\boldsymbol{x},\boldsymbol{\alpha})+U(\boldsymbol{x})]-h\vee w\\
&= r(\boldsymbol{\alpha})-h\vee w,
\end{aligned}$$

where the second equality is from Remark 5.6. Lastly,

$$\begin{aligned}
\min_{i\in L} R_i^{(3)}(\boldsymbol{\alpha}) &= \min_{i\in L}\left\{2\inf_{\boldsymbol{x}\in A\times M^{K-1}}[f(\boldsymbol{x},\boldsymbol{\alpha})+Q(O_i,\boldsymbol{x})]+2W(O_i)-2W(O_1)-w\right\}\\
&= \min_{i\in L}\left\{2\inf_{\boldsymbol{x}\in A\times M^{K-1}}[f(\boldsymbol{x},\boldsymbol{\alpha})+Q(O_i,\boldsymbol{x})]+2U(O_i)\right\}-w\\
&= 2\inf_{\boldsymbol{x}\in A\times M^{K-1}}[f(\boldsymbol{x},\boldsymbol{\alpha})+\min_{i\in L}\{Q(O_i,\boldsymbol{x})+U(O_i)\}]-w\\
&= 2\inf_{\boldsymbol{x}\in A\times M^{K-1}}[f(\boldsymbol{x},\boldsymbol{\alpha})+U(\boldsymbol{x})]-w\\
&\geq \inf_{\boldsymbol{x}\in A\times M^{K-1}}[2f(\boldsymbol{x},\boldsymbol{\alpha})+U(\boldsymbol{x})]-w\\
&= r(\boldsymbol{\alpha})-w.
\end{aligned}$$

∎

5.2. Bounds on the error terms h and w. Lemma 5.9 shows that for any collection of temperature ratios $\boldsymbol{\alpha}\in\Delta$, $\liminf_{\varepsilon\to 0}-\varepsilon\log(T^\varepsilon\cdot\mathrm{Var}_x(\theta_{\mathrm{INS}}^{\varepsilon,T^\varepsilon}))$ is always bounded below by $r(\boldsymbol{\alpha})-h\vee w$. In this subsection, it will be shown that we can bound h and w for the INS model by quantities depending only on α_K. This will identify the constant B appearing in Theorem 4.10.

Recall that H is the index set for equilibrium points of V and $y_i\in M$ is the equilibrium point corresponding to index $i\in H$. Additionally, we assumed y_1 is the unique global minimum of V. Let b_1 be the minimum barrier height of y_1, namely,

$$b_1 \doteq \min_{j\in H\setminus\{1\}}\hat{Q}(y_j,y_1), \tag{5.9}$$

where $\hat{Q}$ is the quasipotential associated with the original diffusion (2.2), and $\hat{W}$ is defined analogously to W but for this process.

LEMMA 5.10. $h \doteq \min_{i\in L\setminus\{1\}} Q(O_1, O_i) = \alpha_K b_1$.

Proof. Letting D_1 be the domain of attraction of O_1, we define

$$Q_{D_1}(\boldsymbol{x}, \boldsymbol{y}) \doteq \inf\{I_T(\phi) : \phi(0) = \boldsymbol{x}, \phi(T) = \boldsymbol{y}, \phi(t) \in D_1 \text{ for all } 0 \le t \le T, T < \infty\}.$$

Recall that $Q(\boldsymbol{x}, \boldsymbol{y})$ is defined by

$$Q(\boldsymbol{x}, \boldsymbol{y}) \doteq \inf\{I_T(\phi) : \phi(0) = \boldsymbol{x}, \phi(T) = \boldsymbol{y}, T < \infty\}.$$

Now since O_1 is the only equilibrium point in D_1, this implies that

$$h \doteq \min_{i\in L\setminus\{1\}} Q(O_1, O_i) \ge \inf_{\boldsymbol{x}\in\partial D_1} Q_{D_1}(O_1, \boldsymbol{x}).$$

Moreover, we can apply [15, Theorem 4.3, Chapter 4] and (5.1) to find

$$\begin{aligned}\inf_{\boldsymbol{x}\in\partial D_1} Q_{D_1}(O_1, \boldsymbol{x}) &= -\lim_{\varepsilon\to 0} \varepsilon \log\left(\frac{\nu^\varepsilon(\partial D_1)}{\nu^\varepsilon(D_1)}\right) = \inf_{\boldsymbol{x}\in\partial D_1} U(\boldsymbol{x}) - \inf_{\boldsymbol{x}\in D_1} U(\boldsymbol{x}) \\ &= U(O_2) - U(O_1) = U(O_2) = \alpha_K V(y_2) = \alpha_K b_1,\end{aligned}$$

where $O_2 \doteq (y_1, \dots, y_1, y_2) \in \partial D_1$ with y_2 being an unstable equilibrium point such that $b_1 = \hat{Q}(y_1, y_2) = V(y_2)$. Thus, we have $h \ge \alpha_K b_1$. For the other direction, we use the definitions of Q_{D_1} and Q, and we apply [15, Theorem 4.3, Chapter 4] again to find

$$h \le Q(O_1, O_2) \le Q_{D_1}(O_1, O_2) = U(O_2) - U(O_1) = \alpha_K b_1. \qquad \square$$

Recall that $w \doteq W(O_1) - \min_{i\in L\setminus\{1\}} W(O_1 \cup O_i)$. We provide an upper bound for w in the next lemma. To state the lemma, we need some more definitions. Let $\hat{G}(1)$ denote the collection of graphs on $\{y_i\}_{i\in H}$ that end at y_1. Let $\hat{G}_{\mathrm{m}}(1)$ denote the subset of such graphs with the property that for every local maximum or saddle point y there is a local minimum z such that $\hat{Q}(y, z) = 0$. We know that $\hat{G}_{\mathrm{m}}(1)$ is nonempty since it contains the optimizing $\hat{g}$ in the definition of $\hat{W}(y_1)$ [15, Lemma 4.3(a), Chapter 6]. Given $\hat{g} \in \hat{G}_{\mathrm{m}}(1)$, let $H_{\hat{g}} \subset H\setminus\{1\}$ be the indices which are starting points, i.e., $k \in H_{\hat{g}}$ means that there is no arrow in the graph that leads to y_k. Given $k \in H_{\hat{g}}$, let $C_{\hat{g}}(k)$ be the cost along the path $i_1 = k, i_2, \dots, i_m = 1$ in $\hat{g}$ leading from k to 1:

$$C_{\hat{g}}(k) = \sum_{j=1}^{m-1} \hat{Q}\left(y_{i_j}, y_{i_{j+1}}\right).$$

LEMMA 5.11. $w \le K\alpha_K \min_{\hat{g}\in\hat{G}_m(1)} \max_{k\in H_{\hat{g}}} C_{\hat{g}}(k)$.

Remark 5.12. Note that always $\min_{\hat{g}\in\hat{G}_{\mathrm{m}}(1)} \max_{k\in H_{\hat{g}}} C_{\hat{g}}(k) \le \hat{W}(y_1)$, and that $\min_{\hat{g}\in\hat{G}_{\mathrm{m}}(1)} \max_{k\in H_{\hat{g}}} C_{\hat{g}}(k)$ can in some cases be much smaller than $\hat{W}(y_1)$. For example, this is often the case when H is large but all equilibrium points of V can reach y_1 while passing through only a few intermediate equilibrium points. The lemma is useful owing to the scaling in K that is obtained, but unlike the expression for h it is not tight.

Proof. We will show that for any $i \in L \setminus \{1\}$ and any $\hat{g} \in \hat{G}_{\mathrm{m}}(1)$, $Q(O_i, O_1) \le K\alpha_K \max_{k\in H_{\hat{g}}} C_{\hat{g}}(k)$. If this is true, then from the definition of $W(O_1 \cup O_i)$ we can

construct a graph to use in the definition of $W(O_1)$ that gives $W(O_1) \le W(O_1 \cup O_i) + Q(O_i, O_1)$ for any $i \in L \setminus \{1\}$. Combining these two inequalities with the definition of w in (5.5) complete the proof.

To prove the upper bound for $Q(O_i, O_1)$ we fix a graph $\hat{g} \in \hat{G}_{\mathrm{m}}(1)$ and note that for any y_ℓ with $\ell \in H_{\hat{g}}$, there is a unique sequence of arrows (containing no loop) that leads from y_ℓ to y_1 with cost $C_{\hat{g}}(\ell)$. Furthermore, we know that in this $\hat{g}$, every local maximum or saddle point will lead to a local minimum with zero $\hat{Q}$-cost. Using these facts, we design a route from O_i to O_1 through points from $(\{y_i\}_{i\in H})^K$ in the following way:

- We change only one component at a time.
- We change the component with the largest V-value and replace it by the next equilibrium point suggested by the graph $\hat{g}$. If there is more than one component with the largest V-value, then we can move any one of them.
- Then repeat the process until all the components reach y_1, i.e., O_i reaches O_1.

Next we analyze the Q-cost for each single step. For notational convenience, suppose without lose of generality that it is the first component that takes the largest V-value. Then we will move from $(x_1, x_2 \ldots, x_K)$ to some $(z_1, x_2, \ldots, x_K)$, with $V(x_1) \ge V(x_\ell)$ for all $\ell \ne 1$, and $(x_1 \to z_1) \in \hat{g}$. We claim that $Q((x_1, x_2 \ldots, x_K), (z_1, x_2 \ldots, x_K))$ is always equal to $\alpha_K \hat{Q}(x_1, z_1)$.

We first consider the case when x_1 is a saddle point or a local maximum of V. In this case then we know that z_1 must be a local minimum of V such that $\hat{Q}(x_1, z_1) = 0$, so it is easy to see that we can construct a zero Q-cost trajectory from $(x_1, x_2 \ldots, x_K)$ to $(z_1, x_2 \ldots, x_K)$, and this gives

$$Q((x_1, x_2 \ldots, x_K), (z_1, x_2 \ldots, x_K)) = 0 = \alpha_K \hat{Q}(x_1, z_1).$$

On the other hand, if x_1 is a local minimum of V, then $V(z_1)$ must be larger than $V(x_1)$ (which is larger than $V(x_\ell)$ for all $\ell \ne 1$), and hence according to the definition of U

$$\begin{aligned} Q((x_1, x_2 \ldots, x_K), (z_1, x_2 \ldots, x_K)) &= U(z_1, x_2 \ldots, x_K) - U(x_1, x_2 \ldots, x_K) \\ &= \alpha_K V(z_1) - \alpha_K V(x_1) \\ &= \alpha_K \hat{Q}(x_1, z_1). \end{aligned}$$

As a result, the overall cost for each component to reach y_1 is not larger than $\alpha_K \max_{k\in H_{\hat{g}}} C_{\hat{g}}(k)$, and because there are K components in total, we conclude that $Q(O_i, O_1) \le K\alpha_K \max_{k\in H_{\hat{g}}} C_{\hat{g}}(k)$. We then minimize on $\hat{g} \in \hat{G}_{\mathrm{m}}(1)$. ∎

Remark 5.13. A consequence of Lemmas 5.10 and 5.11 is that for any temperature ratios $\boldsymbol{\alpha} \in \Delta$, if $T^\varepsilon = e^{c/\varepsilon}$ with $c > \alpha_K B$, then $\liminf_{\varepsilon\to 0} -\varepsilon \log(T^\varepsilon \cdot \mathrm{Var}_{\boldsymbol{x}}(\theta^{\varepsilon,T^\varepsilon}_{\mathrm{INS}}))$ is bounded below by $r(\boldsymbol{\alpha}) - \alpha_K B$, where

$$B \doteq b_1 \vee (K \min_{\hat{g}\in\hat{G}_{\mathrm{m}}(1)} \max_{k\in H_{\hat{g}}} C_{\hat{g}}(k)).$$

5.3. Examples.

Example 5.14. We first consider the situation depicted in Figure 4. If we use INS with two temperatures, i.e., $K = 2$ and $1 = \alpha_1 \ge \alpha_2 > 0$, then some algebra shows $h = \alpha_2 b_1 = 4\alpha_2$ and $w = W(O_1) - \min_{i\ne 1} W(O_1 \cup O_i) = 3\alpha_2$, and therefore $h > w$. The outcome $h > w$ reflects the fact that the well containing y_1 is the hardest to escape from and also contains the global minimum.

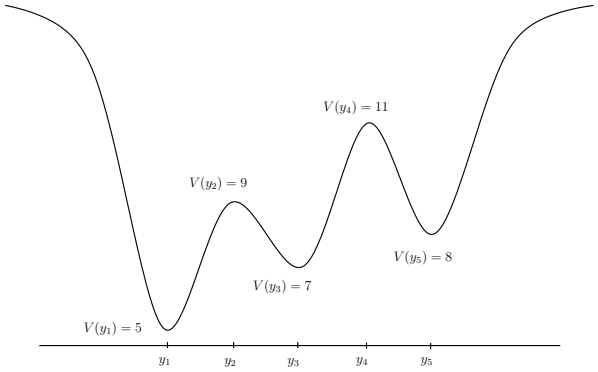


FIG. 4. *A case with $h > w$.*

Example 5.15. In this example, we consider the situation depicted in Figure 5. With the same two-temperature setting as in the last example, one finds $h = \alpha_2 b_1 = 4\alpha_2$ and $w = W(O_1) - \min_{i\neq 1} W(O_1 \cup O_i) = 5\alpha_2$, which gives $w > h$. Here we see that there is a secondary well from which escape is harder than from that which contains y_1. Moreover, in this case $\min_{\hat{g}\in\hat{G}_{\mathrm{m}}(1)} \max_{k\in H_{\hat{g}}} C_{\hat{g}}(k) = \hat{W}(y_1) = 7$, and $K\alpha_K \min_{\hat{g}\in\hat{G}_{\mathrm{m}}(1)} \max_{k\in H_{\hat{g}}} C_{\hat{g}}(k) = 14\alpha_2$ is strictly larger than $w = 5\alpha_2$. Thus the bound for w from Lemma 5.11 is not tight, though it is still good enough to show the deviation from optimality decays geometrically in K.

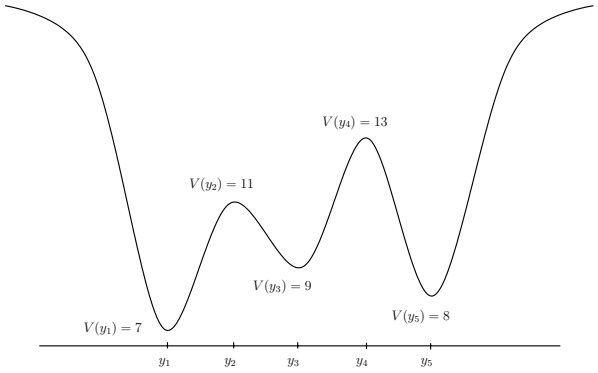
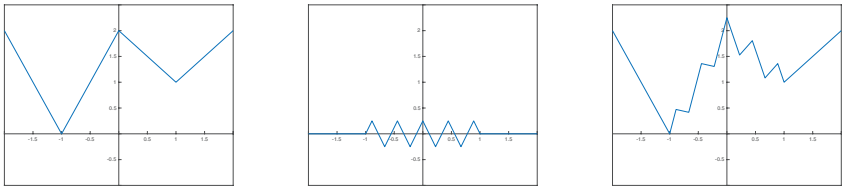


FIG. 5. *A case with $h < w$.*

Example 5.16. The next example we consider is a potential V with a unique global minimum y_1 in the deepest well which is surrounded by N collections of wells of the same form as depicted in Figure 5, with y_1 common to all collections, and each collection arranged in a radial direction out from y_1. Let $\{y_i^n, i = 1, \ldots, 5, n = 1, \ldots, N\}$ with $y_1^n = y_1$ denote the equilibrium points of V. Let $\hat{g}$ be the graph with all arrows pointing in along the radial direction. In this case $H_{\hat{g}}$ has N vertices, and with n indexing such a vertex let $C_{\hat{g}}(n) = V(y_4^n) - V(y_5^n) + V(y_2^n) - V(y_3^n)$. With this example, so long as we have a uniform bound on $C_{\hat{g}}(n)$ there is a bound on w that is independent of N. Note that if there are large barriers between the radial collections, then we will also have $\hat{W}(y_1) = \sum_{1\leq n\leq N} C_{\hat{g}}(n)$, which in this case will be much larger than $\max_{1\leq n\leq N} C_{\hat{g}}(n)$, a situation noted in Remark 5.12.



(a) (b) (c)

FIG. 6. (a) $\bar{V}$, (b) S *with* $\delta = 1/9$, $k = 2$, *and* $h = 0.25$, (c) V.

Example 5.17. In this example we show that the definition of w is in some sense robust with respect to the appearance of insignificant secondary wells on top of V. For simplicity of statement the function we construct will be piecewise affine (a smooth analogue is easily constructed), as depicted in Figure 6. We let $\bar{V}$ be piecewise affine with a local minimum of 0 at -1, a local minimum of 1 at 1, and a maximum of 2 at 0. If we denote w associated to $\bar{V}$ by $w_{\bar{V}}$, then it is straightforward to check that $w_{\bar{V}} = 1$.

We next add to $\bar{V}$ a continuous triangle wave function S that is equal to zero outside $[-1, 1]$, with maxima and minima of $\pm h$ for some $h \in (0, 1)$, and a space between successive maxima (and minima) of $4\delta > 0$ so that the derivative of S equals h/δ in absolute value, a.e. Finally we assume δ is such that $S(0) = S(-1+\delta) = S(1-\delta) = h$. If $h/\delta < 6$, then $V = \bar{V} + S$ has a local minimum of 0 at -1, a local minimum of 1 at 1, and a global maximum of $2+h$ at 0. This requires $2k \cdot 4\delta + 2\delta = 2$ for some $k \in \mathbb{N}$ (where the 2δ are from the ends of the interval). Moreover, depending on the value of h/δ, V could have many secondary local minima which appear due to the perturbation by S. These local minima are at $1-(3+4m)\delta$ for $m \in \{0, 1, \ldots, 2k-1\}$. If we construct the analogue of w for the function V, denoted by w_V, we find that $w_V = 1 + h$, since one can show that $w_V = W(\{-1\}) - W(\{-1\} \cup \{1\})$ and

$$\begin{aligned} W(\{-1\}) &= k[2h + 2\delta] + k[2h - 4\delta] + [h + \delta], \\ W(\{-1\} \cup \{1\}) &= k[2h - 4\delta] + k[2h - 2\delta]. \end{aligned}$$

Thus $w_V - w_{\bar{V}} = h$.

From these calculations we see that the difference between $w_{\bar{V}}$ and w_V is minor as long as the perturbation magnitude h is small, and also that the difference is independent of the number of local minima introduced, which is controlled by δ.

6. Proof of Theorem 4.11. Recall that

$$r(\boldsymbol{\alpha}) \doteq \inf_{\boldsymbol{x} \in A \times M^{K-1}} \left\{ 2 \sum_{\ell=1}^{K} \alpha_\ell V(x_\ell) - \min_{\sigma \in \Sigma_K} \left\{ \sum_{\ell=1}^{K} \alpha_\ell V(x_{\sigma(\ell)}) \right\} \right\}.$$

In this section we provide the proof of Theorem 4.11. An important step of the proof is to reformulate $r(\boldsymbol{\alpha})$ as in the next lemma. Although a version of the lemma appears in [14], we include a somewhat simpler proof of a special case owing to its central role.

LEMMA 6.1. *For any* $\boldsymbol{\alpha} \in \Delta$, *we have* $r(\boldsymbol{\alpha}) = \bar{r}(\boldsymbol{\alpha})$, *where*

$$\bar{r}\left(\boldsymbol{\alpha}\right) \doteq \inf_{\substack{V_1 \in D \\ \{(V_1,\ldots,V_K): V_\ell \in [0,V_1] \text{ for } \ell \geq 2\}}} \left[2\sum_{\ell=1}^{K} \alpha_\ell V_\ell - \min_{\sigma \in \Sigma_K} \left\{ \sum_{\ell=1}^{K} \alpha V_{\sigma(\ell)} \right\} \right],$$

with $D \doteq \{V(x) : x \in A\}$. *Moreover,*

$$\bar{r}\left(\boldsymbol{\alpha}\right) = \inf_{\substack{V_1 \in D \\ \{\boldsymbol{V}: 0 \leq V_2 \leq \cdots \leq V_K \leq V_1\}}} \left[\left(2\alpha_1 - \alpha_K\right) V_1 + \sum_{\ell=2}^{K} \left(2\alpha_\ell - \alpha_{\ell-1}\right) V_\ell \right] \tag{6.1}$$

and

$$\bar{r}\left(\boldsymbol{\alpha}\right) = \inf_{\substack{V_1 \in D \\ \{\boldsymbol{V}: 0 \leq V_2 \leq \cdots \leq V_K \leq V_1\}}} \left[\left(2\alpha_1 - \alpha_K\right) V_1 + \sum_{\ell=2}^{K-1} \alpha_\ell \left(2V_\ell - V_{\ell+1}\right) + 2\alpha_K V_K - V_2 \right]. \tag{6.2}$$

Proof. The first step is to decompose $A \times M^{K-1}$ as $\cup_{\tau \in \Sigma_K} N_\tau$, where

$$N_\tau \doteq \left\{ \boldsymbol{x} \in A \times M^{K-1} : V(x_{\tau(1)}) \leq V(x_{\tau(2)}) \leq \cdots \leq V(x_{\tau(K)}) \right\}.$$

For any $\tau \in \Sigma_K$ there exists $i \in \{1,\ldots,K\}$ which depends on τ such that $1 = \tau\left(i\right)$. We will use the rearrangement inequality [17, section 10.2, Theorem 368], which says that if $\boldsymbol{x} \in N_\tau$, then since α_ℓ is nonincreasing in ℓ the minimum in $U(\boldsymbol{x}) \doteq \min_{\sigma \in \Sigma_K} \{\sum_{\ell=1}^{K} \alpha_\ell V(x_{\sigma(\ell)})\}$ is at $\sigma = \tau$. Thus,

$$\begin{aligned}
&\inf_{\boldsymbol{x} \in A \times M^{K-1}} \left[2\sum_{\ell=1}^{K} \alpha_\ell V\left(x_{_\ell}\right) - U(\boldsymbol{x}) \right] \\
&= \min_{\tau \in \Sigma_K} \left\{ \inf_{\boldsymbol{x} \in N_\tau} \left[2\sum_{\ell=1}^{K} \alpha_\ell V\left(x_{_\ell}\right) - \min_{\sigma \in \Sigma_K} \left\{ \sum_{\ell=1}^{K} \alpha_\ell V(x_{\sigma(\ell)}) \right\} \right] \right\} \\
&= \min_{\tau \in \Sigma_K} \left\{ \inf_{\boldsymbol{x} \in N_\tau} \left[\sum_{\ell=1}^{K} \left(2\alpha_{\tau(\ell)} - \alpha_\ell\right) V(x_{\tau(\ell)}) \right] \right\}.
\end{aligned}$$

Let $\beta_\ell \doteq 2\alpha_{\tau(\ell)} - \alpha_\ell$, and for each τ (and using that i is the index such that $\tau\left(i\right) = 1$), we define the sets

$$N_\tau^i \doteq \left\{ \left(x_{\tau(1)},\ldots,x_{\tau(i)}\right) : \boldsymbol{x} \in N_\tau \right\}$$

and

$$\bar{N}_\tau^i\left(\boldsymbol{y}\right) \doteq \left\{ \left(x_{\tau(i)},\ldots,x_{\tau(K)}\right) : \boldsymbol{x} \in N_\tau \text{ and } \left(x_{\tau(1)},\ldots,x_{\tau(i)}\right) = \boldsymbol{y} \right\}.$$

Then

$$\inf_{\boldsymbol{x} \in N_\tau} \left[\sum_{\ell=1}^{K} \beta_\ell V(x_{\tau(\ell)}) \right] = \inf_{(y_1,\ldots,y_i) \in N_\tau^i} \left[\begin{array}{c} \sum_{\ell=1}^{i-1} \beta_\ell V\left(y_\ell\right) + \beta_i V\left(y_i\right) \\ + \inf_{(z_i,\ldots,z_K) \in \bar{N}_\tau^i(y_1,\ldots,y_i)} \left[\sum_{\ell=i+1}^{K} \beta_\ell V\left(z_\ell\right) \right] \end{array} \right].$$

Next we show that given $(y_1,\ldots,y_i)$ (and noting that by definition $z_i = y_i$),

$$\inf_{(z_i,\ldots,z_K) \in \bar{N}_\tau^i(y_1,\ldots,y_i)} \left[\sum_{\ell=i+1}^{K} \beta_\ell V\left(z_\ell\right) \right] = \left(\sum_{\ell=i+1}^{K} \beta_\ell \right) V\left(y_i\right). \tag{6.3}$$

Recall that $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_K > 0$. Therefore, $\beta_K = 2\alpha_{\tau(K)} - \alpha_K \geq 2\alpha_K - \alpha_K = \alpha_K > 0$. More generally, since $\tau(\ell), \ldots, \tau(K)$ are distinct values drawn from $\{1, \ldots, K\}$, for each ℓ

$$\beta_\ell + \cdots + \beta_K = 2\sum_{j=\ell}^{K} \alpha_{\tau(j)} - \sum_{j=\ell}^{K} \alpha_j \geq 2\sum_{j=\ell}^{K} \alpha_j - \sum_{j=\ell}^{K} \alpha_j > 0.$$

Using $\beta_K \geq 0$ and the fact that $(z_i, \ldots, z_K) \in \bar{N}_\tau^i(y_1, \ldots, y_i)$ implies the restriction

$$V(z_i) \leq V(z_{i+1}) \leq \cdots \leq V(z_K),$$

we can rewrite the infimum as

$$\inf_{(z_i,\ldots,z_K)\in \bar{N}_\tau^i(y_1,\ldots,y_i)} \left[\sum_{\ell=i+1}^{K-2} \beta_\ell V(z_\ell) + (\beta_{K-1} + \beta_K) V(z_{K-1})\right].$$

Iterating, we have (6.3). Recalling $D \doteq \{V(x) : x \in A\}$,

$$\begin{aligned}
&\inf_{\boldsymbol{x}\in A\times M^{K-1}} \left[2\sum_{\ell=1}^{K} \alpha_\ell V(x_\ell) - \min_{\sigma\in\Sigma_K} \left\{\sum_{\ell=1}^{K} \alpha_\ell V(x_{\sigma(\ell)})\right\}\right] \\
&= \min_{\tau\in\Sigma_K} \left\{\inf_{\boldsymbol{x}\in N_\tau} \left[\sum_{\ell=1}^{K} \left(2\alpha_{\tau(\ell)} - \alpha_\ell\right) V(x_{\tau(\ell)})\right]\right\} \\
&= \min_{\tau\in\Sigma_K} \left\{\inf_{(x_{\tau(1)},\ldots,x_{\tau(i)})\in N_\tau^i} \left[\sum_{\ell=1}^{i-1} \beta_\ell V(x_{\tau(\ell)}) + \left(\sum_{\ell=i}^{K} \beta_\ell\right) V(x_{\tau(i)})\right]\right\} \\
&= \min_{\tau\in\Sigma_K} \left\{\inf_{\substack{V_{\tau(i)}\in D \\ \{(V_{\tau(1)},\ldots,V_{\tau(i-1)}):V_{\tau(1)}\leq V_{\tau(2)}\leq\cdots\leq V_{\tau(i)}\}}} \left[\sum_{\ell=1}^{i-1} \beta_\ell V_{\tau(\ell)} + \left(\sum_{\ell=i}^{K} \beta_\ell\right) V_{\tau(i)}\right]\right\}.
\end{aligned}$$

The last equality holds because V is continuous.

We claim that the last display coincides with

$$\begin{aligned}
\bar{r}(\boldsymbol{\alpha}) &\doteq \inf_{\substack{V_1\in D \\ \{\boldsymbol{V}:V_\ell\in[0,V_1] \text{ for } \ell\geq 2\}}} \left[2\sum_{\ell=1}^{K} \alpha_\ell V_\ell - \min_{\sigma\in\Sigma_K} \left\{\sum_{\ell=1}^{K} \alpha_\ell V_{\sigma(\ell)}\right\}\right] \\
&= \min_{\tau\in\Sigma_K} \left\{\inf_{\substack{V_{\tau(i)}\in D \\ \{(V_{\tau(1)},\cdots,V_{\tau(K)}):V_{\tau(1)}\leq V_{\tau(2)}\leq\cdots\leq V_{\tau(K)}\leq V_{\tau(i)}\}}} \left[\sum_{\ell=1}^{K} \left(2\alpha_{\tau(\ell)} - \alpha_\ell\right) V_{\tau(\ell)}\right]\right\}.
\end{aligned}$$

Since $\boldsymbol{V} \in N_\tau$ implies $V_{\tau(\ell)} \geq V_{\tau(i)}$ and hence $V_{\tau(\ell)} = V_{\tau(i)}$ for $i < \ell \leq K$,

$$\begin{aligned}
&\inf_{\substack{V_{\tau(i)}\in D \\ \{(V_{\tau(1)},\ldots,V_{\tau(K)}):V_{\tau(1)}\leq V_{\tau(2)}\leq\cdots\leq V_{\tau(K)}\leq V_{\tau(i)}\}}} \left[\sum_{\ell=1}^{K} \beta_\ell V_{\tau(\ell)}\right] \\
&\quad = \inf_{\substack{V_{\tau(i)}\in D \\ \{(V_{\tau(1)},\ldots,V_{\tau(i-1)}):V_{\tau(1)}\leq V_{\tau(2)}\leq\cdots\leq V_{\tau(i)}\}}} \left[\sum_{\ell=1}^{i-1} \beta_\ell V_{\tau(\ell)} + \left(\sum_{\ell=i}^{K} \beta_\ell\right) V_{\tau(i)}\right],
\end{aligned}$$

which establishes the claim.

Now rewrite $\bar{r}(\boldsymbol{\alpha})$ by noticing that since V_1 is the largest value in the set $\boldsymbol{V}$, $\min_{\tau\in\Sigma_K}\{\sum_{\ell=1}^K \alpha_\ell V_{\tau(\ell)}\}$ obtains the minimum at some $\tau \in \Sigma_K$ with $\tau(K) = 1$. Therefore

$$\bar{r}(\boldsymbol{\alpha}) = \inf_{\substack{V_1\in D \\ \{\boldsymbol{V}:V_\ell\le V_1 \text{ for } \ell\ge 2\}}} \left[(2\alpha_1 - \alpha_K) V_1 + 2\sum_{\ell=2}^K \alpha_\ell V_\ell - \min_{\tau\in\Sigma_K,\tau(K)=1} \left\{ \sum_{\ell=1}^{K-1} \alpha_\ell V_{\tau(\ell)} \right\} \right].$$

Suppose we are given any $K-1$ numbers and assign them to $\{V_\ell\}_{\ell=2,\dots,K}$ in a certain order. Then the value of

$$\min_{\tau\in\Sigma_K,\tau(K)=1} \left\{ \sum_{\ell=1}^{K-1} \alpha_\ell V_{\tau(\ell)} \right\}$$

is independent of the order. But since $\alpha_1 \ge \cdots \ge \alpha_K > 0$, by the rearrangement inequality, the smallest value of $\sum_{\ell=2}^K \alpha_\ell V_\ell$ is obtained by taking the V_ℓ, $\ell \ge 2$, in increasing order. By choosing this ordering of $\{V_\ell\}_{\ell=2,\dots,K}$,

$$\min_{\tau\in\Sigma_K,\tau(K)=1} \left\{ \sum_{\ell=1}^{K-1} \alpha_\ell V_{\tau(\ell)} \right\} = \sum_{\ell=2}^K \alpha_{\ell-1} V_\ell.$$

Thus we obtain (6.1):

$$\begin{aligned}\bar{r}(\boldsymbol{\alpha}) &= \inf_{\substack{V_1\in D \\ \{\boldsymbol{V}:0\le V_2\le\cdots\le V_K\le V_1\}}} \left[(2\alpha_1 - \alpha_K) V_1 + 2\sum_{\ell=2}^K \alpha_\ell V_\ell - \sum_{\ell=2}^K \alpha_{\ell-1} V_\ell \right] \\ &= \inf_{\substack{V_1\in D \\ \{\boldsymbol{V}:0\le V_2\le\cdots\le V_K\le V_1\}}} \left[(2\alpha_1 - \alpha_K) V_1 + \sum_{\ell=2}^K (2\alpha_\ell - \alpha_{\ell-1}) V_\ell \right].\end{aligned}$$

Using summation by parts and $\alpha_1 = 1$ gives (6.2). ∎

Proof of Theorem 4.11. Let $\alpha_K \in (0, (1/2)^{K-1}]$. Since V is continuous and bounded from below, there is $V_0 \in D$ such that $V_0 = V(A)$. Let $\boldsymbol{\alpha}^* \doteq \left(1, 1/2, \dots, (1/2)^{K-2}, \alpha_K\right)$ and $\boldsymbol{V}^* = (V_1^*, \dots, V_K^*)$, with $V_1^* \doteq V_0$, $V_\ell^* \doteq (1/2)^{K-\ell} V_0$ for $\ell = 2, \dots, K$. We have the following inequalities, which are explained after the display:

$$\begin{aligned}&\left(2 + \alpha_K - (1/2)^{K-2}\right) V_0 \\ &= \inf_{\substack{V_1\in D \\ \{\boldsymbol{V}:0\le V_2\le\cdots\le V_K\le V_1\}}} \left[(2 - \alpha_K) V_1 + \sum_{\ell=2}^K \left(2\alpha_\ell^* - \alpha_{\ell-1}^*\right) V_\ell \right] \\ &= \bar{r}(\boldsymbol{\alpha}^*) \\ &\le \sup_{(\alpha_2,\dots,\alpha_{K-1})\in[\alpha_K,1]^{K-2}} \bar{r}(1, \alpha_2, \dots, \alpha_{K-1}, \alpha_k) \\ &\le \sup_{(\alpha_2,\dots,\alpha_{K-1})\in[\alpha_K,1]^{K-2}} \left[(2\alpha_1 - \alpha_K) V_1^* + \sum_{\ell=2}^{K-1} \alpha_\ell \left(2V_\ell^* - V_{\ell+1}^*\right) + 2\alpha_K V_K^* - V_2^* \right] \\ &= \left(2 + \alpha_K - (1/2)^{K-2}\right) V_0.\end{aligned}$$

The first equality follows from $2\alpha_\ell^* - \alpha_{\ell-1}^* = 0$ for $\ell = 2, \dots, K-1$ and $\alpha_K \in (0, (1/2)^{K-1}]$; the second equality comes from (6.1); the second inequality is from

(6.2); the third equality uses $\alpha_1 = 1$, $2V_\ell^* - V_{\ell+1}^* = 0$ for $\ell = 2, \ldots, K-1$, $V_1^* = V_0 = V_K^*$, and $V_2^* = (1/2)^{K-2} V_0$. In addition, since $r(\boldsymbol{\alpha}) = \bar{r}(\boldsymbol{\alpha})$ for any $\boldsymbol{\alpha} \in \Delta$ from Lemma 6.1, we therefore obtain

$$\sup_{(\alpha_2,\ldots,\alpha_{K-1}) \in [\alpha_K, 1]^{K-2}} r(1, \alpha_2, \ldots, \alpha_{K-1}, \alpha_k) = \left(2 + \alpha_K - (1/2)^{K-2}\right) V(A).$$

If $\alpha_K \in ((1/2)^{K-1}, 1]$, then $\alpha_K \in ((1/2)^m, (1/2)^{m-1}]$ for some $m \in \{1, \ldots, K-1\}$. We can apply an analogous argument for each m with $\boldsymbol{\alpha}^* = (1, \alpha_2^*, \ldots, \alpha_{K-1}^*, \alpha_K)$ and $\boldsymbol{V}^* = (V_0, 0, \ldots, 0)$, where

$$\alpha_\ell^* = \begin{cases} (1/2)^{\ell-1} \text{ if } 2 \le \ell \le m, \\ \alpha_K \text{ if } m+1 \le \ell \le K-1, \end{cases}$$

to show that

$$\sup_{(\alpha_2,\ldots,\alpha_{K-1}) \in [\alpha_K, 1]^{K-2}} r(1, \alpha_2, \ldots, \alpha_{K-1}, \alpha_k) = (2 - \alpha_K) V(A). \qquad \square$$

7. Appendix. The results of [13] use the large deviation principle for a small noise diffusion process to characterize large deviation properties of the variance of the empirical measure in the limit as the time horizon tends to infinity and the strength of the noise tends to zero. One use of the rate function on path space is to determine probabilities of transitions between equilibrium points of the noiseless system. As noted previously for the INS model this is not needed, in that the known form of the stationary distribution hands us this information directly. Because of this, all that is needed is that the LDP hold with some rate function that is uniform with respect to initial conditions, which we assume, and certain bounds on the rate function.

One bound that is needed is an upper bound on the cost to go from any point $\boldsymbol{x}$ to any nearby point $\boldsymbol{y}$, i.e., $\inf\{I_T(\phi) : \phi(0) = \boldsymbol{x}, \phi(T) = \boldsymbol{y}, T \in (0, \infty)\}$, which shows that this cost can be made small by making the distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ small (a controllability type condition). Such a bound follows easily from the nondegeneracy of the noise and boundedness of ∇V by making comparison with the case of Brownian motion.

The other bound needed is used to show that for many calculations what happens away from neighborhoods of the equilibrium points is not so important, in that the process spends very little time (in a relative sense) on any place but in the union of these neighborhoods. For this, the key property of the rate function is a result that shows that if $\delta > 0$, then all zero cost trajectories (i.e., paths ϕ such that $I_T(\phi) = 0$ for all $T \in (0, \infty)$) that start outside the union of the δ-neighborhoods of the equilibrium points must reach that set in a time that is uniformly bounded over all initial conditions and paths. Two conditions that are sufficient to show that the time spent away from δ-neighborhoods of the equilibrium points are the following:

1. There is a measurable function $\bar{L} : M^K \times (\mathbb{R}^d)^K \to [0, \infty)$ that is uniformly bounded on each compact subset, such that for all absolutely continuous $\psi \in C([0, T] : M^K)$, the rate function I_T for the INS model discussed in the next subsection of the appendix satisfies

$$\int_0^T \bar{L}(\psi, \dot{\psi}) ds \le I_T(\psi),$$

and in all other cases $I_T(\psi) = \infty$.

2. For each $\delta > 0$ there is $f : [0,\infty) \to [0,\infty)$ that satisfies $f(t) \to \infty$ as $t \to \infty$, and if $\psi : [0,\infty) \to M^K$ is absolutely continuous and if $\psi(t)$ avoids the δ-neighborhoods of all the equilibrium points $\{y_i, i \in H\}^K$, then

$$\int_0^T \bar{L}(\psi, \dot{\psi}) ds \geq f(T). \tag{7.1}$$

Given that an LDP holds with rate function $I_T(\phi)$, it follows from the general large deviation upper bound proved in [10] that $I_T(\phi) \geq J_T(\phi)$, with $J_T(\phi)$ giving the upper bound rate and with $J_T(\phi) = \int_0^T \bar{L}(\phi, \dot{\phi}) ds$ of the following form. For each point $\boldsymbol{x} \in M^K$ there is a finite collection of functions

$$H_j(\boldsymbol{x}, \boldsymbol{\gamma}) \doteq \sum_{k=1}^K \left[\langle -\nabla V(x_k), \gamma_k \rangle + c_k^j \left\| \gamma_k \right\|^2 \right] = \sum_{k=1}^K \langle -\nabla V(x_k), \gamma_k \rangle + \bar{H}_j(\boldsymbol{x}, \boldsymbol{\gamma}),$$

$j = 1, \ldots, J$, where each $\gamma_k \in \mathbb{R}^d$ and for each j the c_k^j take distinct values from $\{\alpha_1^{-1}, \ldots, \alpha_K^{-1}\}$, and the equality defines $\bar{H}_j(\boldsymbol{x}, \boldsymbol{\gamma})$. Note that each $\bar{H}_j(\boldsymbol{x}, \boldsymbol{\gamma})$ is quadratic and positive definite (i.e., greater than zero if $\boldsymbol{\gamma} \neq \boldsymbol{0}$). For $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ with each β_k in the tangent space to M at x_k (the only values where $\bar{L}(\boldsymbol{x}, \boldsymbol{\beta})$ will be finite), we then have that

$$\begin{aligned} \bar{L}(\boldsymbol{x}, \boldsymbol{\beta}) &= \sup_{\{\boldsymbol{\gamma}_k\}} \left[\sum_{k=1}^K \langle \beta_k, \gamma_k \rangle + \sum_{k=1}^K \langle \nabla V(x_k), \gamma_k \rangle - \vee_{j=1}^J \bar{H}_j(\boldsymbol{x}, \boldsymbol{\gamma}) \right] \\ &= \sup_{\{\boldsymbol{\gamma}_k\}} \left[\sum_{k=1}^K \langle (\beta_k + \nabla V(x_k)), \gamma_k \rangle - \vee_{j=1}^J \bar{H}_j(\boldsymbol{x}, \boldsymbol{\gamma}) \right]. \end{aligned}$$

From standard theory of the Legendre–Fenchel transform, $\bar{L}(\boldsymbol{x}, \boldsymbol{\beta}) \geq 0$ with equality if and only if $\boldsymbol{\beta} + \boldsymbol{v}$ is in the set of subdifferentials of $\vee_{j=1}^J \bar{H}_j(\boldsymbol{x}, \boldsymbol{\gamma})$ in the $\boldsymbol{\gamma}$ variable at $\boldsymbol{\gamma} = \boldsymbol{0}$, with $\boldsymbol{v}$ being the vector of components $\nabla V(x_k)$. Since the subdifferentials of $\vee_{j=1}^J \bar{H}_j(\boldsymbol{x}, \boldsymbol{\gamma})$ at $\boldsymbol{\gamma} = \boldsymbol{0}$ are precisely $\{\boldsymbol{0}\}$, we get that $\bar{L}(\phi, \dot{\phi}) = 0$ if and only if each component of $\phi = (\phi_1, \phi_2, \ldots, \phi_K)$ satisfies $\dot{\phi}_k = -\nabla V(\phi_k)$. Since we assume there are only finitely many equilibrium points of V it must be true that each component reaches the δ-neighborhood of one of the equilibrium points in finite time. The reference [10] also proves that $J_T(\phi)$ has compact level sets. Since the equilibrium points of the combined system are just $\{y_i, i \in H\}^K$, the claimed property (7.1) follows from standard calculations (see, e.g., [15, Lemma 2.2, Chapter 4]).

REFERENCES

[1] J. Blanchet and H. Lam, *State-dependent importance sampling for rare-event simulation: An overview and recent advances*, Surv. Oper. Res. Manag. Sci., 17 (2012), pp. 38–59, https://doi.org/https://doi.org/10.1016/j.sorms.2011.09.002.

[2] A. Bovier and F. den Hollander, *Metastability: A Potential-Theoretic Approach*, Grundlehren Math. Wiss., Springer, Cham, 2015.

[3] L. Breiman, *Probability Theory*, Addison-Wesley, Reading, MA, 1968.

[4] A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events: Representations and Weak Convergence Methods*, Probab. Theory Stoch. Model. 94, Springer-Verlag, New York, 2019.

[5] A. de Acosta, *On large deviations of empirical measures in the τ-topology*, J. Appl. Probab., 31 (1994), pp. 41–47.

[6] T. Dean and P. Dupuis, *Splitting for rare event simulation: A large deviations approach to design and analysis*, Stochastic Process. Appl., 119 (2009), pp. 562–587.
[7] J. Doll, P. Dupuis, and P. Nyquist, *A large deviations analysis of certain qualitative properties of parallel tempering and infinite swapping algorithms*, Appl. Math. Optim., 78 (2018), pp. 103–144.
[8] J. Doll, N. Plattner, D. L. Freeman, Y. Liu, and P. Dupuis, *Rare-event sampling: Occupation-based performance measures for parallel tempering and infinite swapping Monte Carlo methods*, J. Chem. Phys, 137 (2012), 204112.
[9] P. Dupuis and R. Ellis, *The large deviation principle for a general class of queueing systems,* I, Trans. Amer. Math. Soc., 347 (1996), pp. 2689–2751.
[10] P. Dupuis, R. Ellis, and A. Weiss, *Large deviations for Markov processes with discontinuous statistics,* I: *General upper bounds*, Ann. Probab., 19 (1991), pp. 1280–1297.
[11] P. Dupuis, Y. Liu, N. Plattner, and J. D. Doll, *On the infinite swapping limit for parallel tempering*, Multiscale Model. Simul., 10 (2012), pp. 986–1022, https://doi.org/10.1137/110853145.
[12] P. Dupuis and H. Wang, *Subsolutions of an Isaacs equation and efficient schemes for importance sampling*, Math. Oper. Res., 32 (2007), pp. 1–35.
[13] P. Dupuis and G.-J. Wu, *Large deviation properties of the empirical measure of a metastable small noise diffusion*, J. Theoret. Probab., (2021), https://doi.org/10.1007/s10959-020-01072-3.
[14] P. Dupuis, G.-J. Wu, and M. Snarski, *Infinite swapping using IID samples*, ACM Trans. Model. Comput. Simul., 29 (2019), pp. 1–26.
[15] M. I. Freidlin and A. D. Wentzell, *Random Perturbations of Dynamical Systems*, 3rd ed., Springer-Verlag, New York, 2012.
[16] C. Geyer, *Markov chain Monte Carlo maximum likelihood*, in Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface, American Statistical Association, New York, 1991, pp. 156–163.
[17] G. Hardy, J. Littlewood, and G. Pólya, *Inequalities*, Cambridge Mathematical Library, Cambridge University Press, 1952, https://books.google.com/books?id=t1RCSP8YKt8C.
[18] J. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer, New York, 2004.
[19] N. Plattner, J. Doll, P. Dupuis, H. Wang, Y. Liu, and J. Gubernatis, *An infinite swapping approach to the rare-event sampling problem*, J. Chem. Phys., 135 (2011), 134111.
[20] R. Rubinstein and D. Kroese, *Simulation and the Monte Carlo Method*, 3rd ed., Wiley, New York, 2016.
[21] R. Swendsen and J. Wang, *Replica Monte Carlo simulation of spin glasses*, Phys. Rev. Lett., 57 (1986), pp. 2607–2609.
[22] G.-J. Wu, *Optimal Temperature Selection for Infinite Swapping in the Low Temperature Limit*, Ph.D. thesis, Brown University, 2019.