Analysis and optimization of certain parallel Monte Carlo methods in the low temperature limit

Paul Dupuis^{*} and Guo-Jhen Wu[†]

November 12, 2020

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 multi-well setting we prove that a pure geometric sequence of temperature ratios is always nearly optimal, with a performance gap that decays geometrically in the number of temperatures.

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

^{*}Division of Applied Mathematics, Brown University, Providence, USA. Research supported in part by the National Science Foundation (DMS-1904992) and the AFOSR (FA-9550-18-1-0214).

[†]Department of Mathematics, KTH Royal Institute of Technology, Stockholm, Sweden. Research supported in part by the AFOSR (FA-9550-18-1-0214).

relatively long history of the use of large deviation ideas in the design of algorithms for estimating probabilities of single rare events [8, 13], 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 adverse overcome the impact of rare events on Markov chain Monte Carlo (MCMC).

Parallel tempering (PT) [23, 17], also known as *replica exchange*, and a scheme obtained as a suitable limit and known as *infinite swapping* (INS) [12], 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 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 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 on the mechanisms that produce variance reduction in the accelerated Monte Carlo methods. The proof of our main result, Theorem 4.12, is given in Section 5. Section 6 gives examples and discusses bounds on crucial parameters that appear in Theorem 4.12, and the Appendix sketches the proof of why the INS model satisfies a large deviation principle on path space.

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

$$\mu^{\varepsilon}(dx) \doteq \frac{1}{Z^{\varepsilon}_{\mu}} e^{-\frac{V(x)}{\varepsilon}} dx, \qquad (2.1)$$

where $V : \mathbb{R}^d \to \mathbb{R}$ 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^{ε}_{μ} 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 invariant distribution of the diffusion process $\{X^{\varepsilon}(t)\}_{t\geq 0}$ satisfying the stochastic differential equation

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

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

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

where δ_x is the Dirac measure at x. The ergodic theorem implies $\lambda^{\varepsilon,T}$ 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

$$\theta_{\mathrm{MC}}^{\varepsilon,T} \doteq \lambda^{\varepsilon,T} \left(A \right) = \frac{1}{T} \int_{0}^{T} \mathbf{1}_{A} \left(X^{\varepsilon} \left(t \right) \right) dt.$$
(2.4)

¹To 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.

We think of $\theta_{MC}^{\varepsilon,T}$ as the most straightforward MCMC estimator of $\mu^{\varepsilon}(A)$, and since we will later on 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, counting [20, 22]), 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 a *metastability*. 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) [14]. Following the convention of [16, Chapter 6], [14] 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 and with smooth boundary (precise regularity assumptions for M are given on [16, page 135]). This is also consistent with how MCMC algorithms for a process such as (2.2) are often implemented by using periodic boundary conditions that are far removed for the regions of interest. 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 [9] for the formulation of infinite swapping for discrete state 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

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

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

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

then by the ergodic theorem [5], $\theta^{\varepsilon,T} \to \mu^{\varepsilon}(A)$ 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 Parallel tempering

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 twotemperature 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

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

gives an approximation to the Gibbs measure with density $\psi^{\varepsilon}(x_1, x_2) \propto e^{-V(x_1)/\varepsilon}e^{-\alpha V(x_2)/\varepsilon}$. 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 $a (1 \wedge [\psi^{\varepsilon}(x_2, x_1)/\psi^{\varepsilon}(x_1, x_2)])$, then we have a *Markov jump-diffusion*. Moreover, it is straightforward to check this new 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 [12] and the asymptotic variance, can be optimized by letting $a \to \infty$. This suggests one should consider the limit as $a \to \infty$ (the infinite swapping limit). This cannot be done directly with the parallel tempering processes, since they will not be tight, and hence do not converge in a meaningful way. An alternative perspective is to consider a temperature swapping process and approximate $\psi^{\varepsilon}(x_1, x_2)dx_1dx_2$ by a corresponding weighted empirical measure instead (see [12] 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

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

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

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

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 $\rho^{\varepsilon,\alpha}(x_1, x_2) + \rho^{\varepsilon,\alpha}(x_2, x_1) = 1$). One can show that $\zeta^{\varepsilon,T}(dx)$ has precisely the same distribution as what one would obtain by forming the ordinary empirical measure of the parallel tempering process with swap rate a and letting $a \to \infty$.

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 [15].

Remark 3.2. Infinite swapping algorithms for continuous time reversible jump Markov processes and for discrete time reversible Markov processes are also discussed in [12, 9].

3.2 Infinite swapping

In this subsection we introduce the K-temperature INS estimator, 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 ; for any permutation $\sigma \in \Sigma_K$ and $\boldsymbol{x} \in M^K$, \boldsymbol{x}_σ denotes $(x_{\sigma(1)}, \ldots, x_{\sigma(K)})$;

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

 $\boldsymbol{\alpha} \doteq (\alpha_1, \dots, \alpha_K) \in \Delta$ denotes the K temperature multiplication factors appearing in the definition of the K-temperature INS estimator.

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

,

 M^K satisfying

$$\begin{cases} dX_{1}^{\varepsilon} = -\nabla V\left(X_{1}^{\varepsilon}\right)dt + \sqrt{2\varepsilon}\sqrt{\rho_{11}^{\varepsilon}/\alpha_{1} + \rho_{12}^{\varepsilon}/\alpha_{2} + \dots + \rho_{1K}^{\varepsilon}/\alpha_{K}}dW_{1} \\ dX_{2}^{\varepsilon} = -\nabla V\left(X_{2}^{\varepsilon}\right)dt + \sqrt{2\varepsilon}\sqrt{\rho_{21}^{\varepsilon}/\alpha_{1} + \rho_{22}^{\varepsilon}/\alpha_{2} + \dots + \rho_{2K}^{\varepsilon}/\alpha_{K}}dW_{2} \\ \vdots \\ dX_{K}^{\varepsilon} = -\nabla V\left(X_{K}^{\varepsilon}\right)dt + \sqrt{2\varepsilon}\sqrt{\rho_{K1}^{\varepsilon}/\alpha_{1} + \rho_{K2}^{\varepsilon}/\alpha_{2} + \dots + \rho_{KK}^{\varepsilon}/\alpha_{K}}dW_{K} \\ (3.3)$$

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

$$ho_{ij}^{arepsilon}\left(oldsymbol{x};oldsymbol{lpha}
ight)\doteq\sum_{\sigma:\sigma\left(j
ight)=i}w^{arepsilon}\left(oldsymbol{x}_{\sigma};oldsymbol{lpha}
ight),$$

and with

$$w^{\varepsilon}(\boldsymbol{x};\boldsymbol{\alpha}) \doteq \frac{\exp[-\frac{1}{\varepsilon}\sum_{\ell=1}^{K}\alpha_{\ell}V(x_{\ell})]}{\sum_{\sigma\in\Sigma_{K}}\exp[-\frac{1}{\varepsilon}\sum_{\ell=1}^{K}\alpha_{\ell}V(x_{\sigma(\ell)})]}.$$
(3.4)

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

$$\nu^{\varepsilon} \left(d\boldsymbol{x} \right) \doteq \frac{1}{K! Z_{\nu}^{\varepsilon}} \sum_{\sigma \in \Sigma_{K}} \exp \left[-\frac{1}{\varepsilon} \sum_{\ell=1}^{K} \alpha_{\ell} V \left(x_{\sigma(\ell)} \right) \right] d\boldsymbol{x}$$
(3.5)

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

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

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

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

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 [7]) 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 α over time T is therefore defined by

$$\theta_{\text{INS}}^{\varepsilon,T} \doteq \zeta^{\varepsilon,T} (A \times M^{K-1})$$

$$= \frac{1}{T} \int_0^T \sum_{\sigma \in \Sigma_K} w^{\varepsilon} \left(\boldsymbol{X}_{\sigma}^{\varepsilon}(t) ; \boldsymbol{\alpha} \right) \mathbf{1}_A \left(X_{\sigma(1)}^{\varepsilon}(t) \right) dt.$$
(3.6)

Remark 3.4. Besides $\mu^{\varepsilon}(A)$ for various choices of A, one is also interested in estimating risk sensitive functionals of the form

$$\int_{\mathbb{R}^d} e^{-\frac{1}{\varepsilon}F(x)} \mu^{\varepsilon} \left(dx \right),$$

as well as the analogous integrals with respect to some or all of the higher temperatures $\varepsilon/\alpha_{\ell}$. However, it is the lowest temperature which is most challenging, and thus we focus on the problem of estimating $\mu^{\varepsilon}(A)$ 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 $\theta_{\text{INS}}^{\varepsilon,T}$ to do better than $\theta_{\text{MC}}^{\varepsilon,T}$, we introduce the notion of *implied potential*.

Definition 3.5. Given a probability density ϕ^{ε} with respect to Lebesgue measure, we define the *implied potential* of ϕ^{ε} to be $-\varepsilon \log \phi^{\varepsilon}$.

Example 3.6. If μ^{ε} is a Gibbs measure as in (2.1), then the implied potential of μ^{ε} is V, the potential appearing in the dynamics (2.2).

From Example 3.6 we see that implied potential generalizes the notion of potential. By comparing the implied potential of ν^{ε} as in (3.5) and 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 $\ell \in \{2, \ldots, K\}$. Since as is well known the barrier heights determine the exponential time scale of transitions between neighborhoods of local minimum 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 α . In this paper we not only characterize the low temperature performance of a K-temperature INS estimator with a fixed set of temperature factors α , but we also provide optimal and nearly optimal temperatures for problems of interest in the same 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}^{ε} 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.

As noted in Remark 3.4, 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 3.7. 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 $x \in \overline{M}$

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

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 3.8. 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

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

If the lim inf is a limit that does not depend on x, then we call it the decay rate of the variance per unit time.

Remark 3.9. 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^{ε} , which may itself depend on the estimator. One may be concerned that improving the decay rate somehow lengthens the time till essential unbiasedness. However, as we discuss in a moment the selection of INS temperature parameters that lower the decay rate of the variance also reduce the growth rate of this time. Thus there is no conflict in using the decay rate of the variance as the sole performance measure.

Remark 3.10. 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 4.11 for a more detailed statement.

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 [6, 3], 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.

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 [14], 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 [24] for details.

The second main 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 get arbitrarily close to the benchmark articulated in Remark 3.10, with the "gap" between the two taking the form $(1/2)^{K-1}(V(A)+B)$ for some positive constant B, and therefore decaying geometrically in the number of temperatures. The proof of this result is also somewhat detailed, and is started in Section 5 and completed in Section 6. In particular, the results of Section 6 show how B depends in a natural and intuitive way on properties of the original potential V.

To apply the results of [14] we need to know that the INS process defined in (3.3) satisfies a large deviation principle 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 (3.4), which become discontinuous in \boldsymbol{x} as $\varepsilon \to 0$. Hence one is concerned with the large deviation properties of processes with discontinuous statistics [11, 10].

The sorts of discontinuities encountered are in fact analogous to those encountered in the large deviation analysis of stochastic networks, such as multiclass queueing networks. A general approach to proving that a large deviation principle holds for stochastic networks appears in [10], and can be adapted to the INS model (3.3). 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 [14] 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 [10] is in fact ideally suited to showing the existence of an LDP without necessarily having an expression for the rate function. We will assume the needed existence of the large deviation principle, and outline in the Appendix how one can adapt the argument of [10] to (3.3) for the case of the two well model, which features the main issues in the proof of the general case.

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 the following figure.



Figure 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 $\theta_{\text{INS}}^{\varepsilon,T^{\varepsilon}}$ 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 [14, Theorems 4.3 and 4.5]. The theorem, in particular, characterizes the decay rate of the variance for the INS estimator for a given α . For a set A let $V(A) \doteq \inf_{x \in A} V(x)$, and also define $\mathcal{K} \doteq \{1, 2, \ldots, K+1\}$.

Theorem 4.4. Assume Condition 4.1, and that the process defined by (3.3) satisfies a large deviation principle that is uniform with respect to initial conditions [6, Section 1.2]. Then for any closed interval $A \subset D$ with $x_L \notin A$ and $A = \overline{A}^\circ$,

$$\theta_{\text{INS}}^{\varepsilon,T^{\varepsilon}} = \frac{1}{T^{\varepsilon}} \int_{0}^{T^{\varepsilon}} \sum_{\sigma \in \Sigma_{K}} w^{\varepsilon} \left(\boldsymbol{X}_{\sigma}^{\varepsilon} \left(t \right); \boldsymbol{\alpha} \right) \mathbf{1}_{A} \left(\boldsymbol{X}_{\sigma(1)}^{\varepsilon} (t) \right) dt$$
(4.1)

is an essentially unbiased estimator of $\mu^{\varepsilon}(A)$, where $w^{\varepsilon}(\boldsymbol{x}; \boldsymbol{\alpha})$ is given by (3.4). Moreover, for any $\boldsymbol{\alpha} \in \Delta$ and $\boldsymbol{x} \in \mathbb{R}^{K}$, we have

$$\liminf_{\varepsilon \to 0} -\varepsilon \log \left(\operatorname{Var}_{\boldsymbol{x}} \left(\theta_{\operatorname{INS}}^{\varepsilon, T^{\varepsilon}} \right) T^{\varepsilon} \right) \geq \begin{cases} \hat{r}_{1} \left(\boldsymbol{\alpha} \right) \wedge \hat{r}_{3} \left(\boldsymbol{\alpha} \right), & \text{if } A \subset (-\infty, 0] \\ \hat{r}_{1} \left(\boldsymbol{\alpha} \right) \wedge \hat{r}_{2} \left(\boldsymbol{\alpha} \right), & \text{if } A \subset [0, \infty) \end{cases},$$

where

$$\hat{r}_{1}(\boldsymbol{\alpha}) \doteq \inf_{\boldsymbol{x} \in A \times \mathbb{R}^{K-1}} \left[2 \sum_{\ell=1}^{K} \alpha_{\ell} V(x_{\ell}) - \min_{\sigma \in \Sigma_{K}} \left\{ \sum_{\ell=1}^{K} \alpha_{\ell} V(x_{\sigma(\ell)}) \right\} \right],$$
$$\hat{r}_{2}(\boldsymbol{\alpha}) \doteq \min_{i \in \mathcal{K} \setminus \{1\}} \left\{ 2V(A) + \left[\sum_{\ell=1}^{i-2} \alpha_{K-\ell+1} - \alpha_{K-i+2} \right] (h_{L} - h_{R}) \right\} - \alpha_{K} h_{R},$$

and

$$\hat{r}_{3}\left(\boldsymbol{\alpha}\right) \doteq 2V\left(A\right) - \alpha_{K}h_{L}.$$

Remark 4.5. As mentioned in [14, Conjecture 4.10], we expect that the lower bound is tight. The proof of the conjecture for a special case is outlined in [14, 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 two results identify optimizers over α 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 to get within 2% of the maximum rate of 2V(A).

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

$$\sup_{\boldsymbol{\alpha}\in\Delta} \left[\hat{r}_{1}\left(\boldsymbol{\alpha}\right)\wedge\hat{r}_{3}\left(\boldsymbol{\alpha}\right)\right] = 2V\left(A\right) - \left(1/2\right)^{K-1}V\left(A\right)$$

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

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

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

Theorem 4.7. Assume the conditions of Theorem 4.4. For any closed set $A \subset [0, \infty)$ and if $h_L \ge 2h_R$ or $V(A) \ge h_L$, then

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

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

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

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

Remark 4.8. According to Theorems 4.4, 4.6 and 4.7, no matter what set A is considered, the optimal temperatures α^* form a geometric sequence with common ratio 1/2, except possibly the last and smallest value, which corresponds to the highest temperature.

Remark 4.9. By Theorems 4.4 and 4.7, if $A \subset [0,\infty)$, $h_L \leq 2h_R$ and $V(A) = h_L - h_R$, the last component of the optimal temperature α^* is 0. Of course the INS estimator is not well-defined with $\alpha_K^* = 0$. In fact, α^* is

not in Δ , though it is in the closure of Δ . However, since $\hat{r}_1(\alpha) \wedge \hat{r}_2(\alpha)$ is a continuous function of α , we can always approach the optimal performance by using α which is close to α^* , e.g., $\alpha = (1, 1/2, \ldots, (1/2)^{K-2}, \delta(1/2)^{K-2})$ for some $\delta \in (0, 1)$.

Remark 4.10. 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.

Remark 4.11. Let $\gamma_i > 0, i = 1, 2$ be given. Suppose that the lower bound on the variance decay rate is within γ_1 of the benchmark, here 2V(A), and that also $\alpha_K h_L \leq (1/2)^{K-1} h_L < \gamma_2$. When this is true, with the simulation time horizon $T^{\varepsilon} = e^{\frac{1}{\varepsilon}\gamma_2}$ (see Remark 4.3) we find that for small $\varepsilon > 0$

$$\operatorname{Var}_{x}\left(\theta_{\operatorname{INS}}^{\varepsilon,T^{\varepsilon}}\right) \leq \frac{1}{T^{\varepsilon}}e^{-\frac{2}{\varepsilon}\left(V(A)-\gamma_{1}\right)},$$

while the quantity being estimated is (approximately) of magnitude $\mu^{\varepsilon}(A) \approx e^{-\frac{1}{\varepsilon}V(A)}$. Therefore the ratio of the standard deviation of the estimator (recall that the bias will be negligible) to the quantity of interest satisfies

$$\frac{\mathrm{SD}_x\left(\theta_{\mathrm{INS}}^{\varepsilon,T^{\varepsilon}}\right)}{\mu^{\varepsilon}(A)} \le e^{\frac{1}{\varepsilon}(\gamma_1 - \gamma_2/2)},$$

with simulation time that scales like $e^{\frac{1}{\varepsilon}\gamma_2}$, and bounded relative error requires, in addition to the bound above, $\gamma_2 > 2\gamma_1$. Although the simulation time grows exponentially in $1/\varepsilon$, the constant gets small very quickly as Kincreases. Note also that the bound applies for arbitrary sets A. For comparison, let $\delta_1 > 0$ and $\delta_2 > h_L$. If we consider standard Monte Carlo with $T^{\varepsilon} = e^{\frac{1}{\varepsilon}\delta_2}$ we would have a lower bound of the form

$$\operatorname{Var}_{x}\left(\theta_{\mathrm{MC}}^{\varepsilon,T^{\varepsilon}}\right) \geq \frac{1}{T^{\varepsilon}}e^{-\frac{1}{\varepsilon}\left(V(A)+\delta_{1}\right)},$$

for small $\varepsilon > 0$, and

$$\frac{\mathrm{SD}_x\left(\theta_{\mathrm{MC}}^{\varepsilon,T^{\varepsilon}}\right)}{\mu^{\varepsilon}(A)} \ge e^{\frac{1}{\varepsilon}(V(A)/2 - \delta_1/2 - \delta_2/2)}.$$

In this case we cannot reduce T^{ε} below $e^{\frac{1}{\varepsilon}h_L}$. If $V(A) < h_L$ we can have bounded relative error, but if the set is moved further to the right so that $V(A) > h_L$ then we must increase the growth rate of T^{ε} for bounded relative error. In all cases, the time required grows exponentially in $1/\varepsilon$ and, unlike the INS case, we cannot make the constant small.

4.2 Sources of variance reduction

Here we make some remarks on the form of the optimal α 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 Vis 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 [15] 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, 4.6 and 4.7). 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 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 α , 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 LD rate function is what allows for the explicit computation of the analogues of the ρ weights. This produced 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 [15] does not involve any dynamics, the notion that any variance reduction is due to "increased mobility" is not possible. Indeed, as is discussed in [15] 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 Theorems 4.6 and 4.7 tells us how to do this in the low 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 Theorems 4.6 and 4.7, 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. We conjecture that the perturbation of α_K away from $(1/2)^{K-1}$ in the Markov case reflects that the optimization here benefits more from greater mobility than the variance reducing effects of the geometric sequence. There is even one case, where the optimal value of α_K is zero, which one can interpret as saying we should make the corresponding component as noisy as possible. It is also worth noting that the overall performance is not particularly sensitive to α_K having the optimal value, in that if we were to simply use the purely geometric sequence then we still have a decay rate that is within $(1/2)^{K-1}(V(A) \vee h_L)$ of the optimal 2V(A).

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 Theorems 4.6 and 4.7, 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). The parameter *B* that appears in the following theorem depends only on *V*, and is identified in Remark 6.4. In particular, it does not depend on ε . As will be illustrated by examples in Section 6, *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.

For the following theorem, we assume that $V : M \to \mathbb{R}$ is a smooth multi-well potential with a unique global minimum $y_1 \in M$ and without loss normalize so that V takes value 0 at y_1 (i.e., $V(y_1) = 0$ and V(x) > 0 for all $x \in M$). We assume that the gradient of V is Lipschitz continuous, and 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 $\bigcup_{i\in L} \{O_i\}$ coincides with the ω -limit set of the zero noise analogue of (3.3), so that $\bigcup_{i\in L} \{O_i\} = \{y_1, \ldots, y_H\}^K$. This imposes some additional structure on V, and in particular rules out open regions on which V is a constant.

Theorem 4.12. Assume that the process defined by (3.3) satisfies a large deviation principle that is uniform with respect to initial conditions. Then there exists $B < \infty$ such that the following hold. Consider the choice $\boldsymbol{\alpha}^* = (1, 1/2, \ldots, (1/2)^{K-1})$ and let $T^{\varepsilon} = e^{\frac{1}{\varepsilon}c}$ for some $c > B\alpha_K^* = B(1/2)^{K-1}$. Define $\theta_{\text{INS}}^{\varepsilon,T^{\varepsilon}}$ by (4.1). Then $\theta_{\text{INS}}^{\varepsilon,T^{\varepsilon}}$ is essentially unbiased, and

$$\liminf_{\varepsilon \to 0} -\varepsilon \log \left(\operatorname{Var}_x \left(\theta_{\operatorname{INS}}^{\varepsilon, T^{\varepsilon}} \right) T^{\varepsilon} \right) \ge \left(2 - (1/2)^{K-1} \right) V(A) - B(1/2)^{K-1}.$$

5 Proof of Theorem 4.12

We first recall notation from Subsection 3.2 and introduce additional notation. Given $K \in \mathbb{N}$, 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 (3.3), and denote $O_1 \doteq (y_1, \ldots, y_1)$. Figure 2 illustrates the points $\bigcup_{i \in L} \{O_i\}$ 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.



Figure 2: Franz potential $\theta = 0.85$ and equilibrium points of INS K = 2

To apply the results of [14] we need several quantities that are constructed in terms of the Freidlin-Wentzell quasipotential. The quasipotential for (3.3) is easy to identify because the system is reversible with $\nu^{\varepsilon} \in \mathcal{P}(M^K)$ defined by (3.5) as its unique stationary distribution. Thus if for $\boldsymbol{x} \in M^K$ we define

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

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

Since we assume that $\{\mathbf{X}^{\varepsilon}(t)\}_{0 \leq t \leq T}$ 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 $Q(\mathbf{x}, \mathbf{y})$ is defined for all $\mathbf{x}, \mathbf{y} \in M^K$ by

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

(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 $W \subset L = \{1, \ldots, l\}$, a directed graph consisting of arrows $i \to j$ $(i \in L \setminus W, j \in L, i \neq j)$ is called a W-graph on L if it satisfies the following conditions.



Figure 3: Symmetrized potential for K = 2

- 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 could 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

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

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

and

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

Before proceeding to the next subsection, we state and 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 $x, y \in M^{K}$, W(x) - W(y) = U(x) - U(y).

Proof. Since we know that the stationary distribution ν^{ε} of $\{X^{\varepsilon}(t)\}_{t\geq 0}$ is given by (3.5), we can apply [16, Theorem 4.3, Chapter 6] to find that for any $\eta > 0$ and for sufficiently small neighborhoods of x and y,

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

and

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

Thus

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

and

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

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

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

Therefore

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

where the last equality is 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 [14]. The result stated in [14] assumes a fixed function f, but the result as stated below follows from this and the uniform convergence $f_{\varepsilon} \to f$. The uniformity with respect to the initial condition is discussed on [14, page 12]. Let

$$h \doteq \min_{i \in L \setminus \{1\}} Q(O_1, O_i) \text{ and } w \doteq W(O_1) - \min_{i \in L \setminus \{1\}} W(O_1 \cup O_i).$$
 (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 Section 6.

Theorem 5.7. Assume that the process defined by (3.3) 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 $c > h \lor w$. 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 $\mathbf{x} \in M^K$,

$$\begin{split} \liminf_{\varepsilon \to 0} &-\varepsilon \log \left| E_{\boldsymbol{x}} \left(\frac{1}{T^{\varepsilon}} \int_{0}^{T^{\varepsilon}} e^{-\frac{1}{\varepsilon} f_{\varepsilon}(X_{t}^{\varepsilon})} 1_{A} \left(X_{t}^{\varepsilon} \right) dt \right) - \int_{M^{K}} e^{-\frac{1}{\varepsilon} f_{\varepsilon}(\boldsymbol{x})} 1_{A} \left(\boldsymbol{x} \right) \nu^{\varepsilon} \left(d\boldsymbol{x} \right) \\ &\geq \inf_{\boldsymbol{x} \in A} \left[f\left(\boldsymbol{x} \right) + W\left(\boldsymbol{x} \right) \right] - W\left(O_{1} \right) + c - (h \lor w), \end{split}$$

and

$$\liminf_{\varepsilon \to 0} -\varepsilon \log \left(T^{\varepsilon} \cdot \operatorname{Var}_{\boldsymbol{x}} \left(\frac{1}{T^{\varepsilon}} \int_{0}^{T^{\varepsilon}} e^{-\frac{1}{\varepsilon} f_{\varepsilon}(X_{t}^{\varepsilon})} 1_{A}(X_{t}^{\varepsilon}) dt \right) \right)$$
$$\geq \begin{cases} \min_{i \in L} \left(R_{i}^{(1)} \wedge R_{i}^{(2)} \right), & \text{if } h \geq w \\ \min_{i \in L} \left(R_{i}^{(1)} \wedge R_{i}^{(2)} \wedge R_{i}^{(3)} \right), & \text{otherwise} \end{cases},$$

where

$$R_i^{(1)} \doteq \inf_{\boldsymbol{x} \in A} [2f(\boldsymbol{x}) + Q(O_i, \boldsymbol{x})] + W(O_i) - W(O_1),$$
$$R_1^{(2)} \doteq 2 \inf_{\boldsymbol{x} \in A} [f(\boldsymbol{x}) + Q(O_1, \boldsymbol{x})] - h,$$

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 $i \in L$

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

To apply this theorem to the INS model, we note that the definition of $\theta_{\text{INS}}^{\varepsilon,T^{\varepsilon}}$ involves the sum of a finite number of integrals of the form

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

where σ is a permutation which for simplicity we take here to be the identity, and $w^{\varepsilon}(\boldsymbol{x}, \boldsymbol{\alpha})$ is defined in (3.4). Since V is bounded and continuous, it follows from standard features of the mollification used in the definition of w^{ε} in (3.4), 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 $\varepsilon \to 0$

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

uniformly in $\boldsymbol{x} \in M^K$ (see, e.g., [6, 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

$$\begin{split} \liminf_{\varepsilon \to 0} -\varepsilon \log \left| E_{\boldsymbol{x}} \left(\theta_{\text{INS}}^{\varepsilon, T^{\varepsilon}} \right) - \nu^{\varepsilon}(A) \right| \\ &\geq \inf_{\boldsymbol{x} \in A \times M^{K-1}} \left[f\left(\boldsymbol{x}, \boldsymbol{\alpha} \right) + W\left(\boldsymbol{x} \right) \right] - W\left(O_{1} \right) + c - (h \lor w) \\ &= \inf_{\boldsymbol{x} \in A \times M^{K-1}} \left[f\left(\boldsymbol{x}, \boldsymbol{\alpha} \right) + U\left(\boldsymbol{x} \right) \right] + c - (h \lor w). \end{split}$$

Since $f \ge 0, U \ge 0$ and $c > h \lor w$, this shows that $\theta_{\text{INS}}^{\varepsilon, T^{\varepsilon}}$ is essentially unbiased. Moreover, we find that $\liminf_{\varepsilon \to 0} -\varepsilon \log(T^{\varepsilon} \cdot \operatorname{Var}_{x}(\theta_{\text{INS}}^{\varepsilon, T^{\varepsilon}}))$ is bounded below by either $\min_{i \in L}(R_{i}^{(1)}(\boldsymbol{\alpha}) \land R_{i}^{(2)}(\boldsymbol{\alpha}))$ or $\min_{i \in L}(R_{i}^{(1)}(\boldsymbol{\alpha}) \land R_{i}^{(2)}(\boldsymbol{\alpha}) \land R_{i}^{(3)}(\boldsymbol{\alpha}))$, depending on whether $h \ge w$ or w > h.

In the next subsection, we will identify appropriate lower bounds for these two minima and then optimize the lower bounds over α .

Remark 5.8. As mentioned in Remark 3.4, we are also interested in estimating risk sensitive functionals of the form

$$\int_{\mathbb{R}^d} e^{-\frac{1}{\varepsilon}F(x)} \mu^{\varepsilon} \left(dx \right)$$

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 4.12 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

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

which is the same as $\inf_{\boldsymbol{x}\in A\times M^{K-1}} \{2f(\boldsymbol{x},\boldsymbol{\alpha}) + U(\boldsymbol{x})\}\)$, where $f(\boldsymbol{x},\boldsymbol{\alpha}) \doteq \sum_{\ell=1}^{K} \alpha_{\ell} V(x_{\ell}) - U(\boldsymbol{x})$. As the next lemma shows, this optimization problem, which plays a key role in the bounds we will derive, has an explicit solution. Although a proof appears in [15], we include a slightly simpler proof of the special case needed here owing to its central role.

Lemma 5.9. We have

$$\sup_{\boldsymbol{\alpha}\in\Delta}r(\boldsymbol{\alpha}) = \left(2 - (1/2)^{K-1}\right)V(A),$$

with the unique optimizer $\boldsymbol{\alpha}^* = (1, 1/2, \dots, (1/2)^{K-1}).$

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

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

For any $\tau \in \Sigma_K$ there exists $i \in \{1, \ldots, K\}$ which depends on τ such that $1 = \tau(i)$. We will use the rearrangement inequality [18, 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,

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

Let $\beta_{\ell} \doteq 2\alpha_{\tau(\ell)} - \alpha_{\ell}$, and for each $i \in \{1, \ldots, K\}$ define the sets

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

and

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

Note that for each τ (and using that *i* is the index such that τ (*i*) = 1)

$$\inf_{\boldsymbol{x}\in N_{\tau}} \left[\sum_{\ell=1}^{K} \beta_{\ell} V\left(\boldsymbol{x}_{\tau(\ell)}\right) \right] \\
= \inf_{(y_{1},\dots,y_{i})\in N_{\tau}^{i}} \left[\sum_{\ell=1}^{i-1} \beta_{\ell} V\left(y_{\ell}\right) + \beta_{i} V\left(y_{i}\right) \\
+ \inf_{(z_{i},\dots,z_{K})\in \bar{N}_{\tau}^{i}(y_{1},\dots,y_{i})} \left[\sum_{\ell=i+1}^{K} \beta_{\ell} V\left(z_{\ell}\right) \right] \right].$$

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

$$\inf_{(z_i,\dots,z_K)\in\bar{N}^i_\tau(y_1,\dots,y_i)} \left[\sum_{\ell=i+1}^K \beta_\ell V\left(z_\ell\right)\right] = \left(\sum_{\ell=i+1}^K \beta_\ell\right) V\left(y_i\right).$$
(5.7)

Recall that $\alpha_1 \geq \alpha_2 \cdots \geq \alpha_K > 0$. 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

$$\beta_{\ell} + \dots + \beta_{K} = 2\sum_{j=\ell}^{K} \alpha_{\tau(j)} - \sum_{j=\ell}^{K} \alpha_{j} \ge 2\sum_{j=\ell}^{K} \alpha_{j} - \sum_{j=\ell}^{K} \alpha_{j} > 0.$$

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

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

we can rewrite the infimum as

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

Iterating, we have (5.7). Letting $D \doteq \{V(x) : x \in A\}$,

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

The last equality holds because V is continuous.

We claim that the last display coincides with

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

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

$$\inf_{\substack{\{V_{\tau(i)} \in D\} \\ \{(V_{\tau(1)}, \dots, I_{\tau(K)}) : V_{\tau(1)} \leq V_{\tau(2)} \leq \dots \leq V_{\tau(K)} \leq V_{\tau(i)}\}}} \left[\sum_{\ell=1}^{K} \beta_{\ell} V_{\tau(\ell)} \right] \\
= \inf_{\substack{\{V_{\tau(i)} \in D\} \\ \{(V_{\tau(1)}, \dots, V_{\tau(i-1)}) : V_{\tau(1)} \leq V_{\tau(2)} \leq \dots \leq V_{\tau(i)}\}}} \left[\sum_{\ell=1}^{i-1} \beta_{\ell} V_{\tau(\ell)} + \left(\sum_{\ell=i}^{K} \beta_{\ell} \right) V_{\tau(i)} \right],$$

which establishes the claim.

To prove that $\sup_{\alpha} \bar{r}(\alpha) = \{(2 - (1/2)^{K-1})V(A)\}$, first rewrite $\bar{r}(\alpha)$ by noticing that since V_1 is the largest value in the set V,

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

obtains the minimum at some $\tau \in \Sigma_K$ with $\tau(K) = 1$. Therefore

$$\bar{r}(\boldsymbol{\alpha})$$

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

Suppose we are given any K-1 numbers and assign them to $\{V_\ell\}_{\ell=2,\ldots,K}$ 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 $\alpha_1 \geq \cdots \geq \alpha_K \geq 0$, by the rearrangement inequality, the smallest value of $\sum_{\ell=2}^{K} \alpha_\ell V_\ell$ is obtained by taking the V_ℓ , $\ell \geq 2$ in increasing order. By choosing this ordering of $\{V_\ell\}_{\ell=2,\ldots,K}$,

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

Thus,

$$\bar{r}(\boldsymbol{\alpha}) = \inf_{\substack{V_1 \in D\\ \{\boldsymbol{V}: 0 \le V_2 \le \cdots \le V_K \le V_1\}}} \left[(2\alpha_1 - \alpha_K) V_1 + 2\sum_{\ell=2}^K \alpha_\ell V_\ell - \sum_{\ell=2}^K \alpha_{\ell-1} V_\ell \right]$$
$$= \inf_{\substack{V_1 \in D\\ \{\boldsymbol{V}: 0 \le V_2 \le \cdots \le V_K \le V_1\}}} \left[(2\alpha_1 - \alpha_K) V_1 + \sum_{\ell=2}^K (2\alpha_\ell - \alpha_{\ell-1}) V_\ell \right]. \quad (5.8)$$

Using summation by parts and $\alpha_1 = 1$, we have

$$\bar{r}(\boldsymbol{\alpha})$$

$$= \inf_{\substack{V_1 \in D\\ \{\boldsymbol{V}: 0 \le V_2 \le \dots \le V_K \le V_1\}}} \left[(2\alpha_1 - \alpha_K) V_1 + \sum_{\ell=2}^{K-1} \alpha_\ell (2V_\ell - V_{\ell+1}) + 2\alpha_K V_K - V_2 \right]$$
(5.9)

Since V is continuous and bounded from below, there is $V_0 \in \overline{D}$ such that

$$\left(2 - (1/2)^{K-1}\right) V_0 = \left[\left(2 - (1/2)^{K-1}\right) V(A)\right].$$

Let $\boldsymbol{\alpha}^* \doteq (1, 1/2, \dots, 1/2^{K-1})$ and $\boldsymbol{V}^* = (V_1^*, \dots, V_K^*)$, with $V_1^* \doteq V_0$, $V_\ell^* \doteq (1/2)^{K-\ell+1} V_0$ for $\ell = 2, \dots, K$. We have the following inequalities, which are explained after the display:

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

The first equality follows from $2\alpha_{\ell}^* - \alpha_{\ell-1}^* = 0$ for $\ell = 2, \ldots, K$; the second equality from (5.8); the second inequality is from (5.9); the third equality uses $\alpha_1 = 1, 2V_{\ell}^* - V_{\ell+1}^* = 0$ for $\ell = 2, \ldots, K, -\alpha_K V_1^* + 2\alpha_K V_K^* = 0$ and $V_2^* = (1/2)^{K-1} V_0$. We therefore obtain

$$\sup_{\boldsymbol{\alpha}} \bar{r}(\boldsymbol{\alpha}) = \left\{ \left(2 - (1/2)^{K-1} \right) V(A) \right\}.$$

In the rest of the subsection, we will show that for any $\boldsymbol{\alpha} \in \Delta$, both $\min_{i \in L}(R_i^{(1)}(\boldsymbol{\alpha}) \wedge R_i^{(2)}(\boldsymbol{\alpha}))$ and $\min_{i \in L}(R_i^{(1)}(\boldsymbol{\alpha}) \wedge R_i^{(2)}(\boldsymbol{\alpha}) \wedge R_i^{(3)}(\boldsymbol{\alpha}))$ 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 5.10. For any $\boldsymbol{\alpha} \in \Delta$, we have $\min_{i \in L} R_i^{(1)}(\boldsymbol{\alpha}) = r(\boldsymbol{\alpha})$, $\min_{i \in L} R_i^{(2)}(\boldsymbol{\alpha}) \ge r(\boldsymbol{\alpha}) - h \lor w$ and $\min_{i \in L} R_i^{(3)}(\boldsymbol{\alpha}) \ge r(\boldsymbol{\alpha}) - w$.

Proof. First note that

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

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

$$\min_{i \in L \setminus \{1\}} R_i^{(2)}(\boldsymbol{\alpha})$$

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

$$\ge \inf_{\boldsymbol{x} \in A \times M^{K-1}} [2f(\boldsymbol{x}, \boldsymbol{\alpha}) + \min_{i \in L \setminus \{1\}} \{Q(O_i, \boldsymbol{x}) + W(O_i) - W(O_1)\}]$$

$$- W(O_1) + \min_{i \in L \setminus \{1\}} W(O_1 \cup O_i)$$

$$= \inf_{\boldsymbol{x} \in A \times M^{K-1}} [2f(\boldsymbol{x}, \boldsymbol{\alpha}) + \min_{i \in L \setminus \{1\}} \{Q(O_i, \boldsymbol{x}) + U(O_i)\}] - w,$$

using $U \ge 0$ we obtain

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

where the second equality is from Remark 5.6. Lastly,

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

6 Bounds on the error terms h and w

Lemma 5.10 shows that for any collection of temperature ratios $\boldsymbol{\alpha} \in \Delta$, $\liminf_{\varepsilon \to 0} -\varepsilon \log(T^{\varepsilon} \cdot \operatorname{Var}_{x}(\theta_{\mathrm{INS}}^{\varepsilon,T^{\varepsilon}}))$ is always bounded below by $r(\boldsymbol{\alpha}) - h \lor w$.

It remains to bound h and w for the INS model. Let H be the index set for equilibrium points of V and let $y_i \in M$ be the equilibrium corresponding to index $i \in H$. Recall that we assumed y_1 is the unique global minimum of V. Let b_1 be the minimum barrier height of y_1 , namely,

$$b_1 \doteq \min_{j \in H \setminus \{1\}} \hat{Q}(y_j, y_1),$$
 (6.1)

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

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 [16, Theorem 4.3, Chapter 4] and (5.1) to find

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

where $O_2 \doteq (y_1, \ldots, y_1, y_2) \in \partial D_1$ 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 [16, Theorem 4.3, Chapter 4] again to find

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

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 $\hat{G}_m(1)$ denote the subset of such graphs with the property that for every local maximum or saddle point y there is a local local minimum zsuch that $\hat{Q}(y, z) = 0$. We know that $\hat{G}_m(1)$ is nonempty since it contains the optimizing \hat{g} in the definition of $\hat{W}(y_1)$ [16, Lemma 4.3(a), Chapter 6]. Given $\hat{g} \in \hat{G}_m(1)$ 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 $i_1 = k, i_2, \ldots, i_m = 1$ in \hat{g} leading from k to 1:

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

Lemma 6.2. $w \leq K \alpha_K \min_{\hat{q} \in \hat{G}_m(1)} \max_{k \in H_{\hat{g}}} C_{\hat{g}}(k).$

Remark 6.3. Note that always $\min_{\hat{g}\in\hat{G}_{m}(1)} \max_{k\in H_{\hat{g}}} C_{\hat{g}}(k) \leq \hat{W}(y_{1})$, and that $\min_{\hat{g}\in\hat{G}_{m}(1)} \max_{k\in H_{\hat{g}}} C_{\hat{g}}(k)$ 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 is not tight.

Proof. We will show that for any $i \in L \setminus \{1\}$ and any $\hat{g} \in \hat{G}_{\mathrm{m}}(1), Q(O_i, O_1) \leq \alpha_K \max_{k \in H_{\hat{g}}} C_{\hat{g}}(k)$. 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 $W(O_1) \leq W(O_1 \cup O_i) + Q(O_i, O_1)$ 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 $\hat{g} \in \hat{G}_m(1)$, 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 known 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 $V(x_1) \geq V(x_\ell)$ for all $\ell \neq 1$, and $(x_1 \rightarrow z_1) \in \hat{g}$. 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 $(x_1, x_2 \dots, x_K)$ to $(z_1, x_2 \dots, x_K)$, and this gives

$$Q((x_1, x_2..., x_K), (z_1, x_2..., x_K)) = 0 = \alpha_K Q(x_1, z_1).$$

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 $\ell \neq 1$), and hence according to the definition of U

$$Q((x_1, x_2..., x_K), (z_1, x_2..., x_K)) = U(z_1, x_2..., x_K) - U(x_1, x_2..., x_K)$$

= $\alpha_K V(z_1) - \alpha_K V(x_1)$
= $\alpha_K \hat{Q}(x_1, z_1).$

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 $Q(O_i, O_1) \leq K \alpha_K \max_{k \in H_{\hat{g}}} C_{\hat{g}}(k)$. We then minimize on $\hat{g} \in \hat{G}_m(1)$.

Remark 6.4. A consequence of Lemmas 6.1 and 6.2 is that if we pick the temperature ratios to be $\boldsymbol{\alpha}^* = (1, 1/2, \dots, (1/2)^{K-1})$, then $\liminf_{\varepsilon \to 0} -\varepsilon \log(T^{\varepsilon} \cdot \operatorname{Var}_x(\theta_{\mathrm{INS}}^{\varepsilon,T^{\varepsilon}}))$ is bounded below by $2V(A) - (1/2)^{K-1}(V(A) + B)$, where $B \doteq b_1 \vee (K\min_{\hat{g} \in \hat{G}_{\mathrm{m}}(1)} \max_{k \in H_{\hat{g}}} C_{\hat{g}}(k))$. For fixed V, the gap between this value and the best possible 2V(A) decays geometrically in K.

6.1 Examples

Example 6.5. 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 the well containing y_1 is the hardest to escape from and also contains the global minimum.

Example 6.6. 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 $\min_{\hat{g}\in\hat{G}_m(1)} \max_{k\in H_{\hat{g}}} C_{\hat{g}}(k) = \hat{W}(y_1) = 7\alpha_2$, and $K\alpha_K \min_{\hat{g}\in\hat{G}_m(1)} \max_{k\in H_{\hat{g}}} C_{\hat{g}}(k) = 14\alpha_2$ is strictly larger then $w = 5\alpha_2$. Thus the bound for w from Lemma 6.2 is not tight, though it is still good enough to show the deviation from optimality decays geometrically in K.

Example 6.7. The last 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 .



Figure 4: A case with h > w

Let $\{y_i^n, i = 1, \ldots, 5, n = 1, \ldots, N\}$ with $y_1^n = y_1$ denote the critical 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 $\hat{W}(y_1) = \sum_{1 \le n \le N} C_{\hat{g}}(n)$, which in this case will be much larger that $\max_{1 \le n \le N} C_{\hat{g}}(n)$, a situation noted in 6.3.

7 Appendix

The results of [14] 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 holds with some rate function that is uniform with respect to initial conditions, and certain bounds on the rate function.



Figure 5: A case with h < w

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 non-degeneracy 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) 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.

Thus to apply the results of [14] two things need to be shown: an LDP holds that is uniform with respect to initial conditions, and that if $I_T(\phi)$ is the rate function for this LDP then the stability property for zero cost paths just mentioned is true. In this section we sketch how both of these can be shown for the INS model.

7.1 Properties of zero cost trajectories

A condition that is sufficient to show that the time spent away from δ -neighborhoods of the equilibrium points is the following.

1. There is a measurable function $\overline{L} : M^K \times (\mathbb{R}^d)^K \to [0,\infty)$ that is uniformly bounded on each compact subset, such that for all absolutely continuous $\psi \in C([0,T] : M^K)$, the rate function for the INS model discussed in the next section 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 and if $\psi(t)$ avoids the δ -neighborhoods of all the equilibrium points $\{\theta_i, i \in H\}^K$, then

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

Given that an LDP holds with rate function $I_T(\phi)$, it follows from the general large deviation upper bound proved in [11] 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

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

 $j = 1, \ldots, J$, 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 $\boldsymbol{\gamma} \neq \mathbf{0}$). 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

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

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 $\bigvee_{j=1}^{J} \bar{H}_{j}(\boldsymbol{x},\boldsymbol{\gamma})$ 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 $\bigvee_{j=1}^{J} \bar{H}_{j}(\boldsymbol{x},\boldsymbol{\gamma})$ at $\boldsymbol{\gamma} = \boldsymbol{0}$ is precisely $\{\boldsymbol{0}\}$, we get that $\bar{L}(\phi, \dot{\phi}) = 0$ if and only if each component of $\phi = (\phi_1, \phi_2, \dots, \phi_K)$ 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 [11] also proves that $J_T(\phi)$ has compact level sets. Since the equilibrium points of the combined system are just $\{\theta_i, i \in H\}^K$, the claimed property (7.1) follows from standard calculations (see, e.g., [16, Lemma 2.2, Chapter 4]).

7.2 Uniform LDP on path space

The second issue is more complicated. We want to argue the following:

• Let X_x^{ε} denote the solution to the INS dynamics (3.3) with initial condition $x \in M^K$. Fix any $T \in (0, \infty)$. Then $\{X_x^{\varepsilon}\}$ satisfies an LDP on $C([0,T]:M^K)$ with rate function I_T that is uniform in $x \in M^K$ [6, Section 1.2].

Owing to the discontinuities in the diffusion coefficient as $\varepsilon \to 0$, the INS model falls into what are called processes with "discontinuous statistics" in the large deviation literature. There are models with discontinuous statistics for which very explicit expressions for the rate function are possible, but there are also many examples where, although the existence of an LDP can be established, a precise characterization is difficult. The INS model falls into the latter category. We will describe in some detail one way to show the existence of an LDP for the INS model. To explain the main points we consider the particular case of an asymmetric two well model in dimension one, with K = 2. An example is the Franz potential with parameter θ depicted in Figure 2:

$$V(x) = V(x;\theta) = \frac{3x^4 - 4(\theta - 1)x^3 - 6\theta x^2}{2\theta + 1} + 1, \quad x \in \mathbb{R}.$$

For every $\theta \in [0, 1]$, $V(\cdot; \theta)$ has a fixed local minimum of zero at $x_L = -1$, another local minimum at $x_R = \theta$, and a fixed barrier of height 1 at x = 0. Taking $\theta = 1$ produces a symmetric two well potential and $\theta = 0$ gives a single well. As before, one should imagine that the potential has been extended in a periodic fashion while retaining this two local minimum structure. The symmetrized potential, which identifies the stationary distribution for the INS dynamics, is plotted in Figure 3. This potential has a global minimum with value 0 at (-1, -1), local minima at (-1, .85) and (.85, -1), and a highest local minimum at (.85, .85).

In Figure 6 we plot the regions in the pair of variables where the diffusion coefficients for the symmetrized dynamics converge to a discontinuous function. Away from these regions the $\rho_{ij}^{\varepsilon}(\boldsymbol{x};\boldsymbol{\alpha})$ converge uniformly to a constant, with limiting values 1 and 0.



Figure 6: Locations where limits of weights are discontinuous

Figure 2 superimposes the locations of the critical points of the symmetrized potential on the plat of the discontinuity region. Points O_1, O_3, O_7 , and O_9 are local minima with O_1 the global minimum. Points O_2, O_4, O_6 , and O_8 are saddle points, and O_5 is a local maximum.

To prove the LDP one can adapt the theory presented in [10], which was motivated by problems from queueing theory and hence focuses on continuous time processes that take values in a lattice, to deal with the diffusion models of INS. To do so one will want some regularity assumptions on the set D of discontinuities of the functions $\lim_{\varepsilon \to 0} w^{\varepsilon}(\boldsymbol{x}_{\sigma}, \boldsymbol{\alpha})$, which will impose conditions on V. These discontinuities occur when two or more $V(x_i)$ tie, and we will want that given any point in D there is a smooth change of variable so that in an open neighborhood of the point D can be mapped to a set consisting of the union of a finite set of hyperplanes of fixed dimension. These are mild conditions, imposing smoothness on V and ruling out sets of positive Lebesgue measure where V is a constant. When such conditions do not hold the local structure of D can be more complicated, and a more involved argument would be needed.

The method of [10] uses two steps to prove the LDP. One step is to show, using the Markov property, that it is sufficient to prove large deviation estimates of the following general form, rephrased for a continuous state model. We suppose for simplicity of terminology that the state space is $(\mathbb{R}^d)^K$ rather than M^K .

We consider the large deviation properties of increments of the process of the form

$$p^{\varepsilon}(z,\Delta;\beta,\eta) \doteq P\left(\sup_{s\in[0,\Delta]} \|X^{\varepsilon}(s) - s\beta\| < \eta \,\middle|\, X^{\varepsilon}(0) = z\right).$$

To establish an LDP on path space, it is sufficient to show the following. For each $y \in (\mathbb{R}^d)^K$ there is an affine space Γ_y with dimension strictly smaller than that of $(\mathbb{R}^d)^K$ and a lower semicontinuous function $L : (\mathbb{R}^d)^K \times \Gamma_y \to$ $[0, \infty)$, with the property that for each fixed y the map $\beta \to L(y, \beta)$ is convex, and such that

$$\lim_{\Delta \to 0} \frac{1}{\Delta} \lim_{\eta \to 0} \lim_{\delta \to 0} \lim_{\varepsilon \to 0} \inf_{\{z: \|z-y\| \le \delta\}} (-\varepsilon \log p^{\varepsilon}(z, \Delta; \beta, \eta))$$

$$= \lim_{\Delta \to 0} \frac{1}{\Delta} \lim_{\eta \to 0} \lim_{\delta \to 0} \lim_{\varepsilon \to 0} \sup_{\{z: \|z-y\| \le \delta\}} (-\varepsilon \log p^{\varepsilon}(z, \Delta; \beta, \eta))$$

$$= L(y, \beta).$$
(7.2)

The set Γ_y is a local approximation to the directions in which the dynamics of the process are in some sense uniformly (in ε) continuous, and it is in directions orthogonal to Γ_y that there are rapidly changing or perhaps even discontinuous behaviors. We illustrate the role of Γ_y through the two dimensional example. The definition of $L(y,\beta)$ for $\beta \notin \Gamma_y$ is unimportant when $\Gamma_y \neq (\mathbb{R}^d)^K$, since the Lebesgue measure of the times t where an absolutely continuous function $\psi : [0,T] \to (\mathbb{R}^d)^K$ lies on a hyperplane of dimension dK-1 and at the same time $\dot{\psi}$ is not on that plane is zero (i.e., $\dot{\psi}(s) \in \Gamma_{\psi(s)}$ a.s.).

Given the estimates of (7.2) and mild regularity properties of $L(y,\beta)$, in the second step [10] shows how to combine these estimates for increments using the Markov property to obtain a uniform LDP for $\{X^{\varepsilon}\}$ on path space. (There is an error in the proof of the LDP upper bound in [10] that was pointed out and corrected in [19].)

To connect to the INS model, we consider the two temperature two well model discussed earlier, and for which the discontinuity set D is depicted in Figure 6. There are qualitatively three types of points in this figure: (a) continuity points, (b) points y such that in a small neighborhood of ythe set D is smooth and one dimensional, and (c) points y such that in a small neighborhood of y the set D is the intersection of two smooth, one dimensional sets. For points of type (a) we can easily show (7.2) for $\Gamma_y = \mathbb{R}^2$ using many different methods and with an explicit expression for $L(y,\beta)$. For points of type (b) Γ_y is the one dimensional tangent space to D at y. Here we do not attempt to explicitly identify $L(y,\beta)$, and the argument to establish the existence of the limit in (7.2) uses a monotonicity argument, a method that allows existence of limits to be shown without their identification. For the last class of points of type (c) $\Gamma_y = \{0\}$.

We will describe how to prove the existence of the limits in each of the three cases mentioned above. We recall that the INS process model is given by the solution to

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

where $\alpha \in (0, 1)$ and

$$\rho^{\varepsilon,\alpha}(x_1, x_2) = \frac{e^{-\frac{1}{\varepsilon}[V(x_1) + \alpha V(x_2)]}}{e^{-\frac{1}{\varepsilon}[V(x_1) + \alpha V(x_2)]} + e^{-\frac{1}{\varepsilon}[\alpha V(x_1) + V(x_2)]}}.$$
 (7.3)

Recall also that D consists of points (x_1, x_2) such that $V(x_1) = V(x_2)$, and so if not in D then

$$[V(x_1) + \alpha V(x_2)] \neq [\alpha V(x_1) + V(x_2)].$$

7.2.1 $y \notin D$

In this case as $\varepsilon \to 0$ we have $\rho^{\varepsilon,\alpha}(x_1, x_2) \to 0$ or 1 uniformly in a neighborhood of y. Suppose that in fact the limit is 1. Then by standard large

deviation theory and elementary martingale bounds the large deviation limits are the same as those of the system

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

i.e., (7.2) holds with

$$L(y,\beta) = \frac{1}{4} \left[(\beta_1 + \nabla V(y_1))^2 + \alpha (\beta_2 + \nabla V(y_2))^2 \right].$$

The analogous result holds when $\rho^{\varepsilon,\alpha}(x_1, x_2) \to 0$.

7.2.2 $y \in D$ and locally D is a smooth 1-dimensional manifold

We can make a smooth change of variable to "flatten" D and also replace ∇V as it appears in the drift by $(\nabla V(y_1), \nabla V(y_2))$, and V as it appears in $\rho^{\varepsilon,\alpha}(x_1, x_2)$ by $(V(y_1) + (x_1 - y_1)\nabla V(y_1), V(y_2) + (x_2 - y_2)\nabla V(y_2))$. The reason such localization is relevant is because of the limit on Δ in (7.2). This can be justified by using comparison controls to bound the differences in optimal cost under the two sets of dynamics. For notational simplicity let $b = (\nabla V(y_1), \nabla V(y_2))$. To avoid degeneracy we will assume $b \neq 0$. (If b = 0 then the same arguments we use below to justify the replacement of ∇V by its affine approximation can be used to reduce to the case of $y \notin D$.)

One can check that if $b \neq 0$ then D is the line orthogonal to $(-b_1, b_2)$. Using that $V(y_1) = V(y_2)$ we find

$$\rho^{\varepsilon,\alpha}(x_1, x_2) = \frac{e^{-\frac{1}{\varepsilon}\langle (x-y), (b_1, \alpha b_2) \rangle}}{e^{-\frac{1}{\varepsilon}\langle (x-y), (b_1, \alpha b_2) \rangle} + e^{-\frac{1}{\varepsilon}\langle (x-y), (\alpha b_1, b_2) \rangle}}.$$
 (7.4)

In terms of the natural coordinates defined by $(g_1, g_2) = (-x_1b_1 + x_2b_2, x_1b_2 + x_2b_1) / ||b||^2$ we have

$$dG_1^{\varepsilon} = \frac{1}{\|b\|^2} \left(b_1^2 dt - b_1 \sqrt{\varepsilon} \sqrt{2\bar{\rho}(G_1^{\varepsilon}/\varepsilon) + 2\bar{\rho}(-G_1^{\varepsilon}/\varepsilon)/\alpha} dW_1 - b_2^2 dt + b_2 \sqrt{\varepsilon} \sqrt{2\bar{\rho}(G_1^{\varepsilon}/\varepsilon)/\alpha + 2\bar{\rho}(-G_1^{\varepsilon}/\varepsilon)} dW_2 \right)$$

and

$$dG_{2}^{\varepsilon} = \frac{1}{\|b\|^{2}} \left(-b_{2}b_{1}dt + b_{2}\sqrt{\varepsilon}\sqrt{2\bar{\rho}(G_{1}^{\varepsilon}/\varepsilon) + 2\bar{\rho}(-G_{1}^{\varepsilon}/\varepsilon)/\alpha}dW_{1} - b_{2}b_{1}dt + b_{1}\sqrt{\varepsilon}\sqrt{2\bar{\rho}(G_{1}^{\varepsilon}/\varepsilon)/\alpha + 2\bar{\rho}(-G_{1}^{\varepsilon}/\varepsilon)}dW_{2} \right),$$

where $\bar{\rho}(g_1) = e^{-g_1C}/(e^{g_1C} + e^{-g_1C})$ and $C = (1-\alpha) \left[b_1^2 + b_2^2\right] > 0$. To simplify notation we write this SDE as

$$dG^{\varepsilon} = \bar{B}dt + \sqrt{\varepsilon}C^{\varepsilon}(G_1^{\varepsilon})dW$$

where the diffusion matrix $C^1(g)$ is uniformly nondegenerate and can be written in terms of $\bar{\rho}(g/\varepsilon)$ and $\bar{\rho}(-g/\varepsilon)$. Note that the process depends smoothly on G_2^{ε} (in fact owing to the linearization it does not depend on G_2^{ε} at all), and the diffusion coefficient is discontinuous in G_1^{ε} in the limit $\varepsilon \to 0$. For each $\varepsilon > 0$ this SDE has a strong solution that is unique in the strong sense. A final modification that will ease the analysis and which is also justified by using comparison controls is to perturb $\bar{\rho}(g)$ slightly (with a controllable change in the cost by making Γ large), so that

$$\bar{\rho}(g) = 1$$
 for $g \ge \Gamma$ and $\bar{\rho}(g) = -1$ for $g \le -\Gamma$.

It will be enough to show that for any y and $\beta = (0, \beta_2)$ there is $L(y, \beta) \in [0, \infty)$ such that

$$L(y,\beta) = \lim_{\eta \to 0} \liminf_{\delta \to 0} \inf_{\varepsilon \to 0} \inf_{\{z: \|z-y\| \le \delta\}} (-\varepsilon \log p^{\varepsilon}(z,1;\beta,\eta))$$
(7.5)
$$= \lim_{\eta \to 0} \lim_{\delta \to 0} \limsup_{\varepsilon \to 0} \sup_{\{z: \|z-y\| \le \delta\}} (-\varepsilon \log p^{\varepsilon}(z,1;\beta,\eta)).$$

We prove (7.5) by using stochastic control arguments and quasistationary distributions. To be precise, we consider the controlled system

$$d\bar{G}^{\varepsilon} = \bar{B}dt + C^{\varepsilon}(\bar{G}_{1}^{\varepsilon})u^{\varepsilon}dt + \sqrt{\varepsilon}C^{\varepsilon}(\bar{G}_{1}^{\varepsilon})dW$$

where $u^{\varepsilon} = (u_1^{\varepsilon}, u_2^{\varepsilon})$ is any progressively measurable, square integrable control. We make one last elementary change, which is to absorb β_2 into \bar{B} . This can be done since C^{ε} does not depend on g_2 . Let

$$\tau^{\varepsilon} \doteq \inf\{s \ge 0 : \left\| \bar{G}^{\varepsilon}(s) \right\| \ge \eta\}.$$

Then we have the representation [4], [6, Chapter 3]

$$V^{\varepsilon}(g;\eta) \doteq -\varepsilon \log p^{\varepsilon}(y,1;\beta,\eta)$$

= $\inf_{u^{\varepsilon}} E\left[\frac{1}{2} \int_{0}^{1} \|u^{\varepsilon}(s)\|^{2} ds + \infty \mathbb{1}_{\{\tau^{\varepsilon} \leq 1\}} \middle| \bar{G}^{\varepsilon}(0) = g \right].$

We will need to show a type of uniform (in ε) continuity of $V^{\varepsilon}(g;\eta)$ in the neighborhood $B_{\delta}(0) \doteq \{y : ||y|| \leq \delta\}$ as described below. By a time change and scaling properties, we can relate $V^{\varepsilon}(g;\eta)$ to a control problem on the set $B_{1/\varepsilon}(0)$ over the time interval $[0, 1/\eta\varepsilon]$, and the dynamics

$$d\bar{G} = \bar{B}dt + C^1(\bar{G}_1)udt + C^1(\bar{G}_1)dW$$

and the same running cost and time averaged costs, but requiring no exit before $1/\eta\varepsilon$. If $\bar{V}^{\varepsilon}(g; 1/\eta\varepsilon)$ is the value function for this problem, then

$$V^{\varepsilon}(q;\eta) = \bar{V}^{\varepsilon}(q/\varepsilon;1/\eta\varepsilon),$$

so we want a uniformity of $\bar{V}^{\varepsilon}(y; 1/\eta \varepsilon)$ for y distance δ/ε from the origin.

Owing to the fact a limit $\eta \to 0$, it is natural to relate $\bar{V}^{\varepsilon}(g; 1/\eta \varepsilon)$ to an ergodic control problem. For $M \in (0, \infty)$ let λ^M be the minimal cost for the ergodic control problem when considered with these $\varepsilon = 1$ dynamics and which constrains the process to $B_M(0)$ with minimal cost per unit time. This ergodic control problem is closely related to the problem of existence of a quasistationary distribution (QSD) when the original dynamics are constrained to $B_M(0)$, with the ergodic cost equal to the decay rate under the QSD, and the QSD itself is the stationary distribution under the optimal ergodic control. This is proved by a verification argument when a classical sense solution to the HJB equation exists. The control problem is also related to the existence of suitable solutions to an eigenvalue problem [2]. The required existence holds in the present setting owing to the regularity of the boundary and smoothness and nondegeneracy of the dynamics [21].

By the use of comparison controls it is easy to see that λ^M is nonincreasing in M,

 $\lambda^M \downarrow \lambda^*,$

where $\lambda^* > 0$ if and only if $\overline{B} \neq 0$ (note that for the ergodic control problem we send $T \to \infty$ first). (In fact the ergodic cost is more generally monotone in that a larger set will correspond to a smaller cost, and hence the shape of the domain, a ball here, is not important.) It is easy to see that λ^* is finite. We outline why

$$\lim_{\eta \to 0} \lim_{\delta \to 0} \lim_{\varepsilon \to 0} \inf_{\|g\| \le \delta} \bar{V}^{\varepsilon}(g/\varepsilon; 1/\eta \varepsilon) = \lim_{\eta \to 0} \lim_{\delta \to 0} \lim_{\varepsilon \to 0} \sup_{\|g\| \le \delta} \bar{V}^{\varepsilon}(g/\varepsilon; 1/\eta \varepsilon) = \lambda^*$$

is valid.

To prove the upper bound, one would argue as follows. Fix $M < \infty$. Owing to the nondegeneracy, on an interval of the form $[0, \delta/\varepsilon\eta]$ we can drive the process from starting points within $\delta/\varepsilon\eta$ of zero to $B_{M/2}(0)$ with a cost of size (when averaged over the time interval $1/\varepsilon\eta$) of size δ . After this we can apply the optimal control for the λ^M problem. During the second interval of the form $[\delta/\varepsilon\eta, 1]$, ergodicity on the fixed compact set $B_M(0)$ gives a cost of the form $(\lambda^M + \delta)(1 - \delta)$. One then takes limits in the indicated order and then sends $M \to \infty$.

For the lower bound we will need to partition into cases, depending on what happens with the λ^M . It is convenient here to use $R_M(0) \doteq \{(g_1, g_2) : |g_1| \lor |g_2| < M\}$ rather than $B_M(0)$, which is possible due to monotonicity properties mentioned previously. Let μ_M be the stationary distribution under the optimal ergodic control. Suppose that for some sequence $M_i \to \infty$

$$\lim_{i \to \infty} \mu_{M_i} \{ (g_1, g_2) : -\Gamma < g_1 < \Gamma \} > 0.$$

Then the optimally controlled process must return to this set repeatedly. (When this is not the case then process will run off to $\pm \infty$ in the g_1 direction, and this case is handled with a simpler argument.) In this case the minimizing points of the cost potential $W^{M_i}(y)$ will be uniformly bounded in *i* (due to the need to return to $\{(g_1, g_2) : -\Gamma < g_1 < \Gamma\}$), and using comparison controls on any fixed compact set we will have uniform bounds on the Lipschitz constant of $W^{M_i}(g)$ for all large enough *i*. Hence we can pass to the limit

$$W^*(g) = \lim_{i \to \infty} W^{M_i}(g).$$

We claim that W^* will satisfy the limit HJB (see [1] for properties of exp $-W^*$ for special cases) and

$$W^*(g) \le \kappa \|g\| + K$$

for some $\kappa, K < \infty$ (in fact $W^*(g)$ will be independent of g_2).

If the lower bound is *not* true, then we know there is a > 0 and sequences $\eta_j \to 0$, $\delta_j \to 0$ with $\delta_j/\eta_j \to 0$, g_j with $\|g_j\| \leq \delta_j$ and $\varepsilon_j \to 0$ such that

$$\bar{V}^{\varepsilon_j}(g_j/\varepsilon_j; 1/\eta_j\varepsilon_j) \le \lambda^* - a \tag{7.6}$$

for all large enough j.

We use that $W^*(y)$ satisfies

$$\lambda^* = \left\langle DW^*(g), \bar{B} \right\rangle - \frac{1}{2} \left\| (C^1)^T(g_1) DW^*(g) \right\|^2 + \frac{1}{2} \operatorname{tr} \left[A(g_1) D^2 W^*(g) \right]$$

$$\leq \left\langle DW^*(g), \bar{B} + C^1(g_1) u \right\rangle + \frac{1}{2} \left\| u \right\|^2 + \frac{1}{2} \operatorname{tr} \left[A(g_1) D^2 W^*(g) \right],$$

where $A(g_1) = C^1(g_1)(C^1)^T(g_1)$. Also $\bar{V}^{\varepsilon_j}(g; 1/\eta_j \varepsilon_j)$ is equal to $U^{\varepsilon_j,\eta_j}(g,t)$ at t = 0, where U^{ε_j,η_j} satisfies

$$\partial_t U^{\varepsilon_j,\eta_j}(g,t) + \left\langle DU^{\varepsilon_j,\eta_j}(g,t), \bar{B} \right\rangle - \frac{1}{2} \left\| (C^1)^T(g_1) DU^{\varepsilon_j,\eta_j}(g,t) \right\|^2 + \frac{1}{2} \operatorname{tr} \left[A(g_1) D^2 U^{\varepsilon_j,\eta_j}(g,t) \right] = 0$$

plus a zero terminal condition at $t = 1/\varepsilon\eta$ for $g \in R_M(0)$ and $U^{\varepsilon_j,\eta_j}(g,t) = \infty$ for $g \in \partial [-1/\varepsilon, 1/\varepsilon]^2$. (The existence and uniqueness of a solution to this equation follows easily from the fact that $\exp -U^{\varepsilon_j,\eta_j}(g,t)$ satisfies a linear equation with zero boundary condition.)

Now suppose that

$$d\bar{G} = \bar{B}dt + C^1(\bar{G}_1)udt + C^1(\bar{G}_1)dW$$

is an optimally controlled process for $\overline{V}^{\varepsilon_j}(g; 1/\eta_j \varepsilon_j)$. Then

$$dW^*(\bar{G}) = \left\langle DW^*(\bar{G}), \bar{B} + C^1(\bar{G}_1)u \right\rangle dt + \frac{1}{2} \operatorname{tr} \left[A(\bar{G}_1)D^2W^*(\bar{G}) \right] dt + a \text{ martingale.}$$

If the lower bound does not hold, then by (7.6) there is a sequence of starting points g_j such that

$$\bar{V}^{\varepsilon_j}(g_j/\varepsilon_j; 1/\eta_j\varepsilon_j) \le \lambda^* - a$$

That means that since u is the corresponding optimal control

$$E_{y_j/\varepsilon_j,0}\left(\varepsilon_j\eta_j\int_0^{1/\varepsilon_j\eta_j}\frac{1}{2}\|u(t)\|^2\,dt\right)\leq\lambda^*-a.$$

With this control and starting point, by Itô's formula

$$\begin{split} E_{y_j/\varepsilon_j,0}W^*(\bar{G}(1/\varepsilon_j\eta_j)) &- W^*(g_j/\varepsilon_j) \\ &= E_{y_j/\varepsilon_j,0} \int_0^{1/\varepsilon_j\eta_j} \left\langle DW^*(\bar{G}(t)), \bar{B} + C^1(\bar{G}_1(t))u(t) \right\rangle dt + \frac{1}{2} \mathrm{tr} \left[A(\bar{G}_1(t))D^2W^*(\bar{G}(t)) \right] dt \\ &\geq \frac{1}{\varepsilon_j\eta_j} \lambda^* - E_{y_j/\varepsilon_j,0} \left(\int_0^{1/\varepsilon_j\eta_j} \frac{1}{2} \|u(t)\|^2 dt \right) \geq \frac{1}{\varepsilon_j\eta_j} \lambda^* - \frac{1}{\varepsilon_j\eta_j} (\lambda^* - a) \geq \frac{1}{\varepsilon_j\eta_j} a. \end{split}$$

Since we have normalized so that $W^* \ge 0$, $W^*(g_j/\varepsilon_j) \ge 0$. Using the upper bound $W^*(g) \le \kappa ||g|| + K$ and that $\overline{G}(1/\varepsilon_j \eta_j) \in R_{1/\varepsilon_j}(0)$ gives

$$E_{y_j/\varepsilon_j,0}W^*(\bar{G}(1/\varepsilon_j\eta_j)) \le \frac{\kappa}{\varepsilon_j} + K.$$

Since a > 0, since $\varepsilon_j \to 0$ and $\eta_j \to 0$ as $j \to \infty$ we get a contradiction to

$$\frac{1}{\varepsilon_j \eta_j} a \le \frac{\kappa}{\varepsilon_j} + K$$

7.2.3 D is the intersection of two smooth 1-dimensional manifolds, i.e., a point

The argument in this case is essentially the same as in the last case, except that the only velocity we need consider is $\beta = 0$, and so the centering around this velocity is no longer needed, and the linearization is done so as to make the $\rho^{\varepsilon,\alpha}(x_1, x_2)$ be of the form $e^{\frac{1}{\varepsilon}(g_1B_1+g_2B_2)}/(e^{\frac{1}{\varepsilon}(g_1B_1+g_2B_2)}+e^{-\frac{1}{\varepsilon}(g_1B_1+g_2B_2)})$.

References

- S. Agmon. Methods of Functional Analysis and Theory of Elliptic Equations, chapter On positivity and decay of solutions of second order elliptic equations on Riemannian manifolds, pages 19–52. Liguori, Naples, 1982. Ed. D. Greco.
- [2] H. Berestycki, L. Nirenberg, and S. R.S. Varadhan. The principal eigenvalue and maximum principle for second-order elliptic operators in general domains. *Communications on Pure and Applied Mathematics*, 47(1):47–92, January 1994.
- [3] J. Blanchet and H. Lam. State-dependent importance sampling for rare-event simulation: An overview and recent advances. Surveys in Operations Research and Management Science, 17(1):38 – 59, 2012.
- [4] M. Boué and P. Dupuis. A variational representation for certain functionals of Brownian motion. The Annals of Prob., 26:1641–1659, 1998.
- [5] L. Breiman. Probability Theory. Addison-Wesley, Reading, Mass., 1968.
- [6] A. Budhiraja and P. Dupuis. Analysis and Approximation of Rare Events: Representations and Weak Convergence Methods. Number 94 in Probability Theory and Stochastic Modelling. Springer-Verlag, New York, 2019.
- [7] A. de Acosta. On large deviations of empirical measures in the τ -topology. Journal of Applied Probability, 31:41–47, 1994.
- [8] T. Dean and P. Dupuis. Splitting for rare event simulation: A large deviations approach to design and analysis. *Stoch. Proc. Appl.*, 119:562– 587, 2009.

- [9] 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.*, pages 103–144, 2018.
- [10] P. Dupuis and R.S. Ellis. The large deviation principle for a general class of queueing systems, I. Trans. Amer. Math. Soc., 347:2689–2751, 1996.
- [11] P. Dupuis, R.S. Ellis, and A. Weiss. Large deviations for Markov processes with discontinuous statistics, I: General upper bounds. *Annals of Probability*, 19:1280–1297, 1991.
- [12] P. Dupuis, Y. Liu, N. Plattner, and J.D. Doll. On the infinite swapping limit for parallel tempering. SIAM J. Multiscale Model. Simul., 10:986– 1022, 2012.
- [13] P. Dupuis and H. Wang. Subsolutions of an Isaacs equation and efficient schemes for importance sampling. *Math. Oper. Res.*, 32:1–35, 2007.
- [14] P. Dupuis and G.-J. Wu. Large deviation properties of the empirical measure of a stochastic differential equation with small noise. page submitted, 2020.
- [15] P. Dupuis, G.-J. Wu, and M. Snarski. Infinite swapping using iid samples. TOMACS, 29:1–26, 2019.
- [16] M. I. Freidlin and A. D. Wentzell. Random Perturbations of Dynamical Systems. Springer-Verlag, New York, third edition, 2012.
- [17] C.J. Geyer. Markov chain Monte Carlo maximum likelihood. In Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface, pages 156–163, New York, 1991. American Statistical Association.
- [18] G.H. Hardy, J.E. Littlewood, and G. Pólya. *Inequalities*. Cambridge Mathematical Library. Cambridge University Press, 1952.
- [19] I. Ignatiouk-Robert. Large deviations for processes with discontinuous statistics. Ann. Probab., 33:1479–1508, 2005.
- [20] J.S. Liu. Monte Carlo Strategies in Scientific Computing. Springer, New York, 2004.

- [21] P.H. Rabinowitz. Théorie du degré topologique et applications à des prblèmes aux limits non linéaires. Lecture Notes Lab. Analyse Numerique. Université Paris VI, 1975.
- [22] R.Y. Rubinstein and D.P. Kroese. Simulation and the Monte Carlo Method. Wiley, New York, third edition, 2016.
- [23] R.H. Swendsen and J.S. Wang. Replica Monte Carlo simulation of spin glasses. *Phys. Rev. Lett.*, 57:2607–2609, 1986.
- [24] G.-J. Wu. Optimal temperature selection for infinite swapping in the low temperature limit. PhD thesis, Brown University, 2019.