Large Deviations Rate Functions for the Empirical Measure: Explicit Formulas and an Application to Monte Carlo

by

Yufei Liu

B. S., Tsinghua University, 2007

Sc. M., Brown University, 2009

A Dissertation submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in the Division of Applied Mathematics at Brown University

Providence, Rhode Island

May 2013

This dissertation by Yufei Liu is accepted in its present form
by the Division of Applied Mathematics as satisfying the
dissertation requirement for the degree of Doctor of Philosophy.

Date ______________ ______________________________________

Paul Dupuis, Director

Recommended to the Graduate Council

Date ______________ ______________________________________

Jimmie Doll, Reader

Date ______________ ______________________________________

Kavita Ramanan, Reader

Approved by the Graduate Council

Date ______________ ______________________________________

Peter M. Weber, Dean of the Graduate School

Vitæ

Born on May 17 1985, in the city of Tianjin, China. Completed the B.Sc. in Mathematics at Tsinghua University, Beijing, China in July 2007. Started the Ph.D. program in the Division of Applied Mathematics at Brown University in September, 2007.

Dedicated to My Teachers

Acknowledgements

I would like to thank my advisor Professor Paul Dupuis for his extensive mentor and support in the past six years. He introduced me into these intriguing problems, directed me into the appropriate way of approaching these problems, and provided me with abundant advice and guidance for the difficult steps in solving these problems. He also taught me that expressing ideas brilliantly is as important as coming up with brilliant ideas, and helped me in improving my presentation skills by showing me how to write papers and how to prepare for conference talks.

I am also indebted to Professor Jimmie Doll, Professor Hui Wang and Professor Kavita Ramanan for all their generous instructions and enlightenments.

Finally, I am grateful to my family and friends for their encouragement that helped me through the tough times.

Preface

The large deviations theory is a useful tool that characterizes the asymptotic behaviors of a sequence of random variables. As we know, the ergodic theory asserts that the empirical measures of Markov processes converge to the stationary distribution. The large deviations theory for the empirical measures [**5, 6**] gives an estimation of this rate of convergence, which is linked to the large deviations rate function. The hindrance with this approach lies in the fact that the rate function often appears in the form of a variational formula which is difficult to evaluate.

It is thus useful to obtain explicit formulas for the large deviations rate functions. Donsker-Varadhan [**5**] provided such a result for Markov processes with a "diffusive"term. However, this result is not applicable to one of the simplest Markov models: the Markov jump processes, in which the transitions of states only happen at jump times. In fact, the general Donsker-Varadhan large deviations theory of empirical measures does not cover this type of models. In this thesis, we establish a large deviations theory of the empirical measures as well as an explicit formula of the large deviations rate function for Markov jump processes.

An application of such an explicit formula leads to the study of the properties of a Markov chain Monte Carlo algorithm: the parallel tempering algorithm (first introduced in [**9, 13, 20, 21**]). By examining the large deviations rate function as a proxy for the rate of convergence of the Markov chain, we discover a new set of algorithms: the infinite swapping algorithm (first introduced in [**16, 8**] and further discussed in [**4**]), which is the distributional limit of the parallel tempering algorithm with "infinite"swap rate. In this thesis, we review the parallel tempering algorithm and give an introduction to the infinite swapping algorithm as well as its more implementable derivative: the partial infinite swapping algorithm. We also summarize some of the theoretical and numerical results regarding infinite swapping and partial infinite swapping algorithm in this thesis.

Yufei Liu

Providence, Rhode Island

Contents

List of Tables

List of Figures

Abstract of "Large Deviations Rate Functions for the Empirical Measure: Explicit Formulas and an Application to Monte Carlo"

by Yufei Liu, Ph.D., Brown University, May 2013

In this thesis, we use the large deviations principle to characterize the rate of convergence of the empirical measures of Markov processes. An explicit formula of the large deviations rate function is very useful for this purpose. The main result in this aspect is the Donsker-Varadhan formula for the reversible, continuous time Markov processes. However, it is not applicable to the Markov jump processes, and establishing such a formula is our goal in the first part of this thesis. In the second part, we apply such a formula to a Markov chain Monte Carlo algorithm, the parallel tempering algorithm. By examining the rate function, we arrive at a new class of MCMC algorithm, the infinite swapping algorithm. We then devote the rest of this thesis into the study of infinite swapping algorithm, for both of its theoretical properties and its implementation challenges and solutions.

CHAPTER 1

Introduction

1. Empirical measures and Markov chain Monte Carlo

Let S be a Polish space. There are a lot of problems in practice that require the estimation of the integration of some functional V against some probability distribution π, i.e. the estimation of

$$E_{\pi}\left[V\right] \doteq \int_{S} V\left(x\right) \pi\left(dx\right), \tag{1}$$

we assume V is integrable with respect to π throughout this thesis. When π is difficult to sample directly from, a standard numerical method to compute this integral is the Markov chain Monte Carlo scheme. The idea of MCMC is very simple: if one is unable to generate independent samples from the distribution π, it is almost as useful to sample an irreducible Markov chain having π as the invariant distribution. Then one can apply the ergodic theory to link the empirical measures of the Markov chain to its invariant distribution.

The following notations will make this idea explicit. Let $\{X_n, n \in \mathbb{N}\}$ be a Markov chain defined on a probability space $(\Omega, \mathcal{F}, \Pr)$ taking values in S. We denote the transition probability function by $P\left(x, dy\right)$, i.e.

$$\Pr\left(X_{n+1} \in dy | X_n = x\right) = P\left(x, dy\right).$$

We assume that π is an invariant distribution of the Markov chain, i.e. for any Borel subset $A \subset S$,

$$\pi\left(A\right) = \int_{S} P\left(x, A\right) \pi\left(dx\right).$$

Then under suitable recurrence and transitivity conditions on P, π is the unique invariant distribution of P and we have

$$\lim_{N \to \infty} \Pr\left(X_N \in dy | X_0 = x\right) = \pi\left(dy\right).$$

We now turn to the definition of empirical measures. For each $N \in \mathbb{N}$, $\omega \in \Omega$, and Borel subset $A \subset S$, we define the empirical measure, or the normalized occupation measure up to time $N-1$, by

$$\eta_N(\omega, A) = \eta_N(A) \doteq \frac{1}{N} \sum_{n=0}^{N-1} \delta_{X_n(\omega)}(A). \tag{2}$$

Then under suitable ergodicity conditions the ergodic theory guarantees that η_N converges weakly to π as $N \to \infty$ and the **ergodic average**

$$\bar{V}_N \doteq \frac{1}{N} \sum_{n=0}^{N-1} V(X_n) \tag{3}$$

converges to $E_\pi[V]$ defined in (1).

This is the gist of Markov chain Monte Carlo that is implemented in practice. In order to derive some theoretical results, however, it is more convenient to work with the continuous time MCMC. Specifically, instead of constructing a Markov chain, one constructs a time homogeneous Markov process $\{X(t), t \in [0, \infty)\}$. Let $\mathcal{F}_t^X \doteq \sigma(X(s) : s \leq t)$ and let $P(t, x, \Gamma)$ denote the time homogeneous transition function [**11**, pp161] defined on $[0, \infty) \times S \times \mathcal{B}(S)$, i.e.

$$\Pr\left(X(t+s) \in \Gamma | \mathcal{F}_t^X\right) = P(s, X(t), \Gamma).$$

Let $M(S)$ be the space of Borel measurable functions on S. We use T to denote the measurable contraction semigroup associated with $P(t, x, \Gamma)$ [**11**, pp161], in other words, for all $s, t \geq 0$, and all function $f \in M(S)$, we have

$$E\left[f(X(t+s)) | \mathcal{F}_t^X\right] = T(s) f(X(t)). \tag{4}$$

Let $\mathcal{L}$ be the infinitesimal generator of the semigroup $\{T(t)\}$ [**11**, pp8], hence by definition for f that belongs to the domain of $\mathcal{L}$ denoted by $\mathcal{D}(\mathcal{L})$ [**11**, pp8]

$$\mathcal{L}f(x) = \lim_{t \to 0} \frac{1}{t} E\left[f(X(t)) - f(x) | X(0) = x\right]. \tag{5}$$

We say that π is a stationary distribution of $X(t)$ if

$$\int_S (\mathcal{L}f(x)) \pi(dx) = 0$$

for all $f \in \mathcal{D}(\mathcal{L})$. For this Markov process, the empirical measure up to time T is defined as

$$\eta_T(A) \doteq \frac{1}{T}\int_0^T \delta_{X(t)}(A)\,dt. \tag{6}$$

A different version of the ergodic theory asserts the weak convergence of η_T to π and the convergence of

$$\frac{1}{T}\int_0^T V(X(t))\,dt \to \int_S V(x)\,\pi(dx).$$

2. Rate of convergence: second eigenvalue value criteria

The question is, how do we characterize the rate of convergence of the MCMC scheme?

Among the extensive studies regarding this issue, a standard rule of thumb is the second eigenvalue criteria, which asserts a geometric bound on the rate of convergence. There are numerous results that relate the rate of convergence of a Markov chain to its second largest (absolute value) eigenvalue. For example, the following is proved by P. Diaconis and D. Stroock in [**3**]. We assume the state space S is finite with m different states, the transition function can thus be denoted by $P(x,y)$ and the invariant distribution be denoted by $\pi(x)$ for $x, y \in S$. We assume $P(x,y)$ is reversible relative to π, that is

$$P(x,y)\,\pi(x) = P(y,x)\,\pi(y). \tag{7}$$

Clearly a distribution that satisfy (7) is a stationary distribution of P. Reversibility also guarantees that the m eigenvalues of the stochastic matrix P are real. We denote them by

$$1 = \beta_0 \geq \beta_1 \geq \cdots \geq \beta_{m-1} \geq -1.$$

If we define $\beta_* \doteq \max(|\beta_1|, |\beta_{m-1}|)$, then for all $n \in \mathbb{N}$ and $x \in S$

$$4\,\|p^n(x,\cdot) - \pi(\cdot)\|_{\mathrm{Var}} \leq \frac{1-\pi(x)}{\pi(x)}\beta_*^{2n},$$

where the variation distance between two probability distributions is defined as

$$\|\mu-\nu\|_{\mathrm{Var}} \doteq \max_{A\subset S}|\mu(A)-\nu(A)| = \frac{1}{2}\sum_{x\in S}|\mu(x)-\nu(x)|.$$

For a general state space S, one can apply the spectral theory to obtain a similar result. There are a lot of proofs for the following well known result, for one see e.g. [**23**]. Once again we assume the reversibility condition

$$P(x,dy)\,\pi(dx) = P(y,dx)\,\pi(dy), \tag{8}$$

which means that P is a self-adjoint operator on $L^2(\pi)$. We extend the L^2 norm for signed measures. Given a signed measure μ, define $\|\mu\|_2$ by

$$\|\mu\|_2^2 \doteq \begin{cases} \int_S \left|\frac{d\mu}{d\pi}\right|^2 d\pi & \text{if } \mu \ll \pi \\ \infty & \text{otherwise} \end{cases}.$$

The self-adjointness of P guarantees that the spectrum is real and we can define [**19**]

$$\begin{aligned} \lambda_1 &\doteq \inf\operatorname{spec}(P|_{\mathbf{1}^\perp}) \\ \lambda_2 &\doteq \sup\operatorname{spec}(P|_{\mathbf{1}^\perp}) \end{aligned}$$

and $\rho \doteq \max(|\lambda_0|,|\lambda_1|)$. Then for any $\mu \in L^2(\pi)$ and $n\in\mathbb{N}$, we have

$$\|\mu P^n - \pi\|_2 \le \|\mu-\pi\|_2\,\rho^n.$$

Similar results can also be established when P is not self-adjoint, in which case the eigenvalues are not necessarily real and one considers instead the magnitude of the eigenvalues, for example see [**17**]. Note that the second eigenvalue criteria gives geometric bound on the difference between the n-step transitional distribution and the invariant distribution, whereas what we are really looking for is the rate of convergence of the empirical measures to the target distribution π.

Another rule that is widely used to characterize the efficiency of MCMC is the asymptotic variance criteria. It is based on the following central limit theorem (Kipnis and Varadhan, 1986). Recall the definition in (1) and (3). If X_n is an irreducible, aperiodic, and reversible Markov chain, then there exists some positive constant σ, such that $N^{1/2}(\bar{V}_N - E_\pi[V])$ converges in distribution to a normal distribution $N(0,\sigma^2)$.

If we use $\mathrm{Var}_\pi [V]$ to denote the variance of V under the distribution π, then the ratio $\mathrm{Var}_\pi [V] / \sigma^2$ is a measure of the efficiency of the MCMC for estimating $E_\pi [V]$. There are two difficulties with this approach. The first one is the estimation of σ^2, there are a lot of literatures that discuss methods for estimating σ^2, for a summary see [**14**]. The second one is that this approach analyzes the MCMC error when the Markov chain has already reached equilibrium, however, our main concern lies in how many Monte Carlo steps are needed in order for the Markov chain to reach equilibrium.

Also, there are a wide range of articles that propose diagnostic schemes on determining whether the Markov chain has reached equilibrium, see [**2**] for a brief review. In this type of methods, various random variables that are functions of the collective data from MCMC sampling are constructed, and the Markov chain is claimed to have reached equilibrium when these random variables hit some cutoff threshold values. Again, these kind of approach did not answer our question directly since we are interested in knowing up-front how many Monte Carlo steps is needed to get accurate estimation.

3. Rate of convergence: large deviations theory

In order to establish a direct relationship between the number of Monte Carlo steps and the distance from the empirical measure to the target distribution, we use the large deviations theory as a performance measure for MCMC.

We first review the definition of large deviations principle. Let $\mathcal{X}$ be a complete separable metric space and $\{R_n\}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, \Pr)$ taking values in $\mathcal{X}$.

DEFINITION 1. *We say that* $\{R_n\}$ *obeys the* ***large deviations principle*** *with a* ***rate function*** I *if there exists a function* I *from* $\mathcal{X}$ *into* $[0, \infty]$ *satisfying:*

(i) $I(\cdot)$ *is lower semicontinuous*

(ii) For each $l < \infty$ *the set* $\{x : I(x) \le l\}$ *is a compact set in* $\mathcal{X}$*.*

(iii) For each closed set $C \subset \mathcal{X}$

$$\limsup_{n\to\infty} \frac{1}{n} \log \Pr (R_n \in C) \leq - \inf_{x\in C} I(x). \tag{9}$$

(iv) For each open set $G \subset \mathcal{X}$

$$\liminf_{n\to\infty} \frac{1}{n} \log \Pr (R_n \in G) \geq - \inf_{x\in G} I(x). \tag{10}$$

We will refer to (9) *as* ***large deviations upper bound*** *and* (10) *as* ***large deviations lower bound****.*

We will heavily use the fact that large deviations principle is equivalent to the following Laplace principle (see [**7**, Theorem 12.3]).

Definition 2. *Let I be a rate function from $\mathcal{X}$ into $[0,\infty]$ (so I is lower semicontinuous and has compact level sets). The sequence $\{R_n\}$ is said to satisfy the* ***Laplace principle*** *on $\mathcal{X}$ with rate function I if for all bounded continuous functions h on $\mathcal{X}$*

$$\lim_{n\to\infty} \frac{1}{n} \log E\{\exp[-nh(R_n)]\} = - \inf_{x\in\mathcal{X}} \{h(x) + I(x)\}.$$

The term ***Laplace principle upper bound*** *refers to the validity of*

$$\limsup_{n\to\infty} \frac{1}{n} \log E\{\exp[-nh(R_n)]\} \leq - \inf_{x\in\mathcal{X}} \{h(x) + I(x)\}$$

for all bounded continuous functions h while the term ***Laplace principle lower bound*** *refers to the validity of*

$$\liminf_{n\to\infty} \frac{1}{n} \log E\{\exp[-nh(R_n)]\} \geq - \inf_{x\in\mathcal{X}} \{h(x) + I(x)\}$$

for all bounded continuous functions h.

We now look at the large deviations principle for empirical measures. The following two theorems are proved by Donsker and Varadhan in [**5**]. Note that one can obtain a large deviations principle for empirical measures under weaker conditions.

Recall that S is a Polish space. We further assume S is compact. Let $\lambda(dx)$ be a probability measure on S.

First consider the empirical measure for a discrete time Markov chain. Recall that $\{X_n, n \in \mathbb{N}\}$ is a Markov chain defined on a probability space $(\Omega, \mathcal{F}, \Pr)$ taking values in S. Assume:

1. $P(x, dy) = p(x, y)\,\lambda(dy)$,

2. there exists constants a and A such that $0 < a \le p(x, y) \le A < \infty$ for all $x \in S$ and almost all (λ-measure) $y \in S$,

3. for any function $u \in L^1(\lambda)$,

$$\int_S p(x, y)\, u(y)\, \lambda(dy)$$

is a continuous function of x.

Let $\mathcal{U}_1$ be the set of all strictly positive and continuous functions u on S with

$$Pu(x) \doteq \int_S u(y)\, P(x, dy),$$

we define, for any probability measure μ on S

$$I(\mu) = -\inf_{u \in \mathcal{U}_1} \int_S \log\left(\frac{Pu}{u}(x)\right) \mu(dx).$$

Let $\mathcal{P}(S)$ be the space of all probability measures on S equipped with the topology of weak convergence. Recall the definition of empirical measure in (2), we have the following:

Theorem 1. *For any closed set $C \subset \mathcal{P}(S)$,*

$$\limsup_{N \to \infty} \frac{1}{N} \log \Pr(\eta_N \in C) \le -\inf_{\mu \in C} I(\mu).$$

For any open set $G \subset \mathcal{P}(S)$,

$$\liminf_{N \to \infty} \frac{1}{N} \log \Pr(\eta_N \in G) \ge -\inf_{\mu \in G} I(\mu).$$

There is an analogous large deviations principle for the empirical measures of a continuous Markov process $\{X(t), t \in [0, \infty)\}$. Assume transition probabilities $P(t, x, dy)$ have densities relative to $\lambda(dy)$, i.e.

$$P(t, x, dy) = p(t, x, y)\, \lambda(dy), \tag{11}$$

where for each $t>0$

$$0<a(t)\leq p(t,x,y)\leq A(t)<\infty.$$

Recall the semigroup $T(t)$ defined in (4). Let $C(S)$ be the space of all continuous functions on S. Assume $T(t)$ is strongly continuous [**11**, pp6] on $C(S)$. As before, we use $\mathcal{L}$ to denote the infinitesimal generator of $T(t)$, recall the definition of $\mathcal{L}$ in (5) and recall that $\mathcal{D}(\mathcal{L})$ is the domain of $\mathcal{L}$. For any $\mu\in\mathcal{P}(S)$ we define

$$I(\mu)\doteq-\inf_{\substack{u>0,\\ u\in\mathcal{D}(\mathcal{L})}}\int_S\left(\frac{\mathcal{L}u}{u}\right)(x)\,\mu(dx)\,. \tag{12}$$

Recall the definition of empirical measure in (6), we have the following:

THEOREM 2. *For any closed set $C\subset\mathcal{P}(S)$,*

$$\limsup_{T\to\infty}\frac{1}{T}\log\Pr(\eta_T\in C)\leq-\inf_{\mu\in C}I(\mu)\,.$$

For any open set $G\subset\mathcal{P}(S)$,

$$\liminf_{T\to\infty}\frac{1}{T}\log\Pr(\eta_T\in G)\geq-\inf_{\mu\in G}I(\mu)\,.$$

We point out that the compactness of S is assumed for both theorems to hold. We impose compactness in order to obtain the large deviations principle under minimum assumptions on transition functions. In fact, for a general Polish state space, similar large deviations theorems can be proved under additional conditions for transition functions, see [**6**]. In order to focus on the problems that we are interested in without getting into too many technical details, we will always assume compactness on S throughout this thesis. We remind our readers that in fact compactness is not necessary to obtain these results and we refer our readers to [**6**] for the generalization to arbitrary Polish space.

4. Outline of the thesis

The rest of the thesis is organized as follows.

In Chapter 2, we first take a look at a special case of the empirical measures: the empirical measures of a reversible Markov jump processes. We recall the Donsker-Varadhan explicit formula for the rate function, point out that it is not applicable in the pure jump framework. Then we will formally state the large deviations principle and establish the explicit form of the rate function.

In Chapter 3, we will take a weak convergence approach in proving the large deviations principle stated in Chapter 2, although the part of dealing with the randomness in the number of discrete jump steps is not standard at all.

In Chapter 4 we turn to the discussion of an MCMC application. We will look at a special technique in MCMC, namely the parallel tempering algorithm. We will use the large deviations principle to analyze the rate of convergence of parallel tempering schemes. By studying the explicit formula of the large deviations rate function, we discover a new set of Monte Carlo schemes: the infinite swapping algorithm. We examine infinite swapping algorithms in detail, and then move on to its more implementable derivative, the partial infinite swapping scheme.

Finally, in Chapter 5 we provide the implementation manual of infinite swapping and partial infinite swapping and exhibit some numerical results.

CHAPTER 2

An explicit formula of large deviations rate function for the empirical measures of reversible Markov processes

1. Donsker-Varadhan formula for self-adjoint case

In [**5**], Donsker and Varadhan proposed an explicit formula for the rate function defined in (12) for self-adjoint case. We state it as follows:

Let $B(S)$ be the space of bounded measurable functions on S. Recall that $T(t)$ is a semigroup on $B(S)$ associated with the transition function $P(t,x,dy)$ defined by (4), $\mathcal{L}$ is the infinitesimal generator of T defined by (5), and $\mathcal{D}(\mathcal{L})$ is the domain of $\mathcal{L}$. Let $C(S)$ be the space of continuous functions on S. Let B_0 be the strongly continuous center of $T(t)$, i.e.

$$B_0 \doteq \left\{ f \in C(S) : \lim_{t\to 0} \sup_{x\in S} |T(t) f(x) - f(x)| = 0 \right\}.$$

Assume that

1. There exists a σ-finite measure $\lambda(dx)$ such that $P(t,x,dy)$ has a density $p(t,x,y)$ with respect to λ for each $t>0$; the density $p(t,x,y)$ is jointly measurable in x and y and satisfies

$$p(t,x,y) = p(t,y,x) \text{ almost everywhere } (\lambda\times\lambda).$$

2. B_0 is large enough in the sense that $B_0 \cap L^2(\lambda)$ is dense in $L^2(\lambda)$.

3. For any probability measure μ on S and any $f \in B(S)$ there is a sequence of $\{f_n\} \in B_0$ such that $\|f_n\| \le \|f\|$ and $f_n \to f$ almost everywhere with respect to λ.

Under the assumptions above, $\mathcal{D}(\mathcal{L})$ is dense in B_0. The closure [**11**, pp16] of $\mathcal{L}$ which we denote by $\bar{\mathcal{L}}$, has a domain which we denote by $\mathcal{D}(\bar{\mathcal{L}})$, dense in $L^2(\lambda)$. Since $T(t)$ is self-adjoint and contracted, $\bar{\mathcal{L}}$ is self-adjoint and negative semidefinite. We denote by $(-\bar{\mathcal{L}})^{1/2}$ the canonical positive semidefinite square root of $-\bar{\mathcal{L}}$ see e.g.

[**19**, Chapter 12]. Let $\bar{\mathcal{D}}_{1/2}$ be the domain of $\left(-\bar{\mathcal{L}}\right)^{1/2}$. With all these notations defined and $I(\mu)$ as in (12) we have following:

THEOREM 3. *$I(\mu) < \infty$ if and only if $\mu \ll \lambda$ and $\left(d\mu/d\lambda\right)^{1/2} \in \bar{\mathcal{D}}_{1/2}$. Moreover, letting $f \doteq d\mu/d\lambda$ and $g \doteq f^{1/2}$, we have*

$$I(\mu) = \left\| \left(-\bar{\mathcal{L}}\right)^{1/2} g \right\|^2 . \tag{13}$$

2. Donsker-Varadhan formula and reversible Markov jump processes

The simplest continuous time Markov process is a Markov jump process with a bounded infinitesimal generator. Let $\mathcal{B}(S)$ be the Borel σ-algebra on S and let $\alpha(x, \Gamma)$ be a transition function on $S \times \mathcal{B}(S)$. Let $B(S)$ denote the space of bounded Borel measurable functions on S and let $q \in B(S)$ be nonnegative. Then

$$\mathcal{L}f(x) \doteq q(x) \int_S \left(f(y) - f(x)\right) \alpha(x, dy) \tag{14}$$

defines a bounded linear operator on $B(S)$ and $\mathcal{L}$ is the generator for a Markov process in S that can be constructed as follows.

Let $\{X_n, n \in \mathbb{N}\}$ be a Markov chain in S with transition function $\alpha(x, \Gamma)$, i.e.

$$\Pr\left(X_{n+1} \in \Gamma | X_0, X_1, \ldots, X_n\right) = \alpha(X_n, \Gamma)$$

for all $\Gamma \in \mathcal{B}(S)$ and $n \in \mathbb{N}$. Let $\tau_1, \tau_2, \ldots$ be independent and exponentially distributed with mean 1, and they are all independent with $\{X_n, n \in \mathbb{N}\}$. Define sojourn time s_i for each $i = 1, 2, \ldots$ by

$$q(X_{i-1}) s_i = \tau_i. \tag{15}$$

then

$$X(t) = X_n \text{ for } \sum_{i=1}^{n} s_i \leq t < \sum_{i=1}^{n+1} s_i$$

(with the convention of $\sum_{i=1}^{0} s_i = 0$) defines a Markov process $\{X(t), t \in [0, \infty)\}$ in S with infinitesimal generator $\mathcal{L}$.

We now point out an extremely simple special case of Markov jump processes for which Theorem 3 does not hold. Using the notations above, assume $S = [0, 1]$, $q \equiv 1$

and for each $x \in [0,1]$, $\alpha(x,\cdot)$ is the uniform distribution on $[0,1]$. The infinitesimal generator $\mathcal{L}$ reduces to

$$\mathcal{L}f(x) = \int_0^1 f(y)\,dy - f(x),$$

which is clearly a self-adjoint operator. Now let C be the collection of all Dirac measures on S, then C is closed under the topology of weak convergence on $\mathcal{P}(S)$. Hence the large deviations upper bound

$$\limsup_{T\to\infty} \frac{1}{T} \log \Pr(\eta_T \in C) \le -\inf_{\mu \in C} I(\mu) \tag{16}$$

applies. On the other hand, the probability for the very first exponential holding time to be bigger than T is exactly $\exp\{-T\}$, and that when this happens, the empirical measure is a Dirac measure located at some point that is uniformly random on $[0,1]$. Hence

$$\liminf_{T\to\infty} \frac{1}{T} \log P(\eta_T(\cdot) \in C) \ge \liminf_{T\to\infty} \frac{1}{T} \log P(\tau_1 > T) = -1.$$

In fact, we will prove later that the rate function for the empirical measures of this Markov jump process never exceeds 1. However, if Theorem 3 were to hold under this case, one would have $I(\delta_a) = \infty$ for any $a \in [0,1]$ and by (16) we have

$$\limsup_{T\to\infty} \frac{1}{T} \log P(\eta_T(\cdot) \in D) = -\infty,$$

which is a contradiction.

This example implies that this type of Markov jump processes are not covered by Theorem 3. In fact, for any $t > 0$ by considering the case where the exponential holding time is greater than t and the case where it is smaller than t, the transition function $P(t,x,dy)$ takes the following form

$$P(t,x,dy) = e^{-t}\delta_x(dy) + \left(1 - e^{-t}\right) 1_{[0,1]}(y)\,dy,$$

which means that we cannot find a reference measure λ on S such that $P(t,x,\cdot)$ has a density with respect to $\lambda(\cdot)$ for all almost all $x \in S$ and $t > 0$, which is a violation to the first assumption in Theorem 2, let alone the first assumption in Theorem 3

We will prove below that the Markov jump processes with bounded infinitesimal generators always have bounded rate functions. In later sections, we will uncover the special features about the Markov jump processes that justify this boundedness of rate functions.

3. A large deviations principle for reversible Markov jump processes I: statement of assumptions

In the rest of this chapter we will state a large deviations principle for Markov jump processes and provide the explicit formula for the rate function.

We first specify our assumptions that will be expected for the rest of this chapter and Chapter 3. We have already defined Markov jump processes in the last section, and below we will use the same notations as we used there.

Our first assumption is that the Polish state space S is compact. As we have mentioned before, compactness is not needed to derive a large deviations principle. We impose it here in order to obtain the large deviations principle under minimum assumptions on transition functions, and to focus on the problem of interest without getting into too many technical details. To generalize the result from compact space to all Polish space, see e.g. [**6**]

We also assume the jump intensity q in (14) is a continuous function on S, and there exist $0 < K_1 \leq K_2 < \infty$ such that

$$K_1 \leq q(x) \leq K_2. \tag{17}$$

Next, to ensure ergodicity of $X(t)$, we need several conditions on the transition function α. Recall that $\mathcal{P}(S)$ is the metric space of probability measures on S equipped with Levy-Prohorov metric, which is compatible with the topology of weak convergence [**11**, pp96]. We assume the following:

1. α satisfies Feller property. That is, $\alpha(x, \cdot) : S \longmapsto \mathcal{P}(S)$ is continuous in x.

2. α satisfies transitivity condition. That is, there exist positive integers l_0 and n_0 such that for all x and ζ in S

$$\sum_{i=l_0}^{\infty} \frac{1}{2^i} \alpha^{(i)}(x, dy) \ll \sum_{j=n_0}^{\infty} \frac{1}{2^j} \alpha^{(j)}(\zeta, dy).$$

Here we use $\alpha^{(k)}$ to denote the k-step transition probability.

Remark 1. *The Feller property and the compactness of S guarantee that α has an invariant distribution* [**7**, Proposition 8.3.4], *which we denote by $\widetilde{\pi}$. The boundedness of q enables us to define another probability measure π by*

$$\pi(A) \doteq \frac{\int_A \frac{1}{q(x)} \widetilde{\pi}(dx)}{\int_S \frac{1}{q(x)} \widetilde{\pi}(dx)}. \tag{18}$$

Since $\widetilde{\pi}$ is invariant under α, i.e., $\widetilde{\pi}(\cdot) = \int_S \alpha(x, \cdot) \widetilde{\pi}(dx)$, we have

$$\int_S (\mathcal{L} f(x)) \pi(dx) = \frac{1}{\int_S \frac{1}{q(x)} \widetilde{\pi}(dx)} \int_S \int_S [f(y) - f(x)] \alpha(x, dy) \widetilde{\pi}(dx) = 0.$$

Therefore by Echeverria's Theorem [**11**, Theorem 4.9.17], *π is an invariant distribution of the Markov process $X(t)$.*

Remark 2. *Under transitivity condition, $\widetilde{\pi}$ is the unique invariant distribution of α* [**7**, Lemma 8.6.2]. *Thus π defined by (18) is the unique probability measure satisfying $\int_S (\mathcal{L} f(x)) \pi(dx) = 0$, and hence by* [**11**, Theorem 4.9.17] *π is the unique invariant distribution of $X(t)$.*

We also impose some conditions on the invariant distributions $\widetilde{\pi}$ and π:

1. The support of π is S.
2. There exists an integer N and a positive real number c such that

$$\alpha^N(x, \cdot) \leq c \widetilde{\pi}(\cdot).$$

for all $x \in S$.

Remark 3. *The first condition guarantees that any measure $\eta \in \mathcal{P}(S)$ can be approximated by probability measures that are absolutely continuous with respect to π. Indeed, the Lebesgue Decomposition Theorem implies that for any such η there exist $\theta \in L^1(\pi)$ and $\eta^{\perp}$ such that $\theta \geq 0$, $\eta^{\perp} \perp \pi$ and for any $A \subset S$,*

$$\eta(A) = \int_A \theta(x) \pi(dx) + \eta^{\perp}(A).$$

If $\eta^{\perp}(S) > 0$, then one can find a subset $S_1 \subset S$ such that $\pi(S_1) = 0$, $\eta^{\perp}(S_1) > 0$ and $\eta^{\perp}(S \backslash S_1) = 0$. For any $x \in S_1$ and any open neighborhood N_x of x, since the support of π is S we have $\pi(N_x) > 0$ which implies that $N_x \cap S_1 = \{x\}$. Hence S_1

only contains isolated points, i.e., $\eta^{\perp}$ is a discrete measure. A discrete measure can be approximated by measures that are absolutely continuous with respect to π since the support of π is S.

Remark 4. *The first condition excludes the existence of transient states. Although one can obtain an LDP for $X(t)$ that has transient states, one would end up with a rate function that depends on the initial state. We stay away from this subtlety in this thesis in order to focus on our main point of interest.*

Remark 5. *Assumptions like the second condition in the above are not uncommon in the large deviations analysis of empirical measures, see e.g.* [**10**, Hypothesis 1.1].

Reversibility is always needed in order to obtain an explicit formula for the rate function. Recall $\mathcal{D}(\mathcal{L})$ is the domain of $\mathcal{L}$. In this thesis, we assume $\mathcal{L}$ is self-adjoint (or reversible) under π in the following sense: for any $f, g \in \mathcal{D}(\mathcal{L})$

$$\int_S (\mathcal{L}f(x))\, g(x)\, \pi(dx) = \int_S (\mathcal{L}g(x))\, f(x)\, \pi(dx). \tag{19}$$

An obvious equivalent condition for (19) to hold is the "detailed balance" condition, i.e., for π-a.e. $x, y \in S$

$$q(x)\, \alpha(x, dy)\, \pi(dx) = q(y)\, \alpha(y, dx)\, \pi(dy). \tag{20}$$

Note that (20) directly implies $\int_S (\mathcal{L}f(x))\, \pi(dx) = 0$ for all $f \in \mathcal{D}(\mathcal{L})$.

4. A large deviations principle for reversible Markov jump processes II: definition of rate function

We are now ready to give the definition of the rate function I.

We wish to study the large deviations principle for the empirical measure $\eta_T \in \mathcal{P}(S)$ defined by (6). Under compactness of S and Feller condition on α, η_T converges in distribution to an invariant distribution of $X(t)$ ([**11**, Theorem 4.9.3]). As pointed out by Remark 2, π is the unique invariant distribution of $X(t)$, hence η_T converges in distribution to π. Let H be the collection of all distributions that are absolutely continuous with respect to π, i.e.

$$H \doteq \{\eta \in \mathcal{P}(S) : \eta \ll \pi\}. \tag{21}$$

For $\eta \in H$, define

$$I(\eta) \doteq -\int_S \theta^{1/2}(x)\, \mathcal{L}\left(\theta^{1/2}(x)\right) \pi(dx)$$

where $\theta = d\eta/d\pi$. This is a rewriting of $\left\|(-\bar{\mathcal{L}})^{1/2} g\right\|^2$ in (13). By inserting the form of $\mathcal{L}$ from (14), we obtain

$$I(\eta) = \int_S q(x)\,\eta(dx) - \int_{S\times S} \theta^{1/2}(x)\,\theta^{1/2}(y)\,q(x)\,\alpha(x,dy)\,\pi(dx). \tag{22}$$

Note that by applying (20) and using the Cauchy-Schwartz inequality, one can prove that I defined by (22) is nonnegative. Recall that K_2 is the upper bound of q as in (17), hence I is bounded above by K_2. Furthermore, it is convex and continuous.

We want to define I for all measures in $\mathcal{P}(S)$. As pointed out in Remark 3, H is dense in $\mathcal{P}(S)$ under the topology of weak convergence. Hence we can extend the definition of I to all of $\mathcal{P}(S)$ via lower semicontinuous regularization with respect to the topology of weak convergence. Thus if $\eta_n \to \eta$ weakly and $\{\eta_n\} \subset H$, $\liminf_{n\to\infty} I(\eta_n) \geq I(\eta)$; and equality holds for at least one such sequence. This extension guarantees that the extended I is convex, lower semicontinuous and bounded above by K_2 on all of $\mathcal{P}(S)$. The compactness of S and the lower semicontinuity of I ensure that I has compact level sets. Being a nonnegative, lower semicontinuous function with compact level sets, I indeed is a valid rate function.

We have finished the definition of the rate function I, now we will state the large deviations principle.

5. A large deviations principle for reversible Markov jump processes III: statement of LDP

We dedicate the whole next chapter into proving the following theorem.

Theorem 4. *Let $X(t)$ be a Markov jump process satisfying all the assumptions in Chapter 2 Section 3 and let I be defined as in Chapter 2 Section 4. Recall that $\mathcal{P}(S)$ is the metric space of probability measures on S equipped with Levy-Prohorov metric, which is compatible with the topology of weak convergence. Then for any open set $G \subset \mathcal{P}(S)$*

$$\liminf_{T\to\infty} \frac{1}{T} \log P(\eta_T(\cdot) \in G) \geq -\inf_{\eta\in G} I(\eta),$$

and for any closed set $C \subset \mathcal{P}(S)$

$$\limsup_{T\to\infty} \frac{1}{T} \log P(\eta_T(\cdot) \in C) \leq -\inf_{\eta\in C} I(\eta).$$

6. Basic elements in the weak convergence approach

In the next chapter, we will use a weak convergence approach to prove Theorem 4. We will use this section to review the concept of relative entropy and the some of its most frequently used properties, for details see [**7**].

Definition 3. *Let $(\mathcal{V}, \mathcal{A})$ be a measurable space. For $\theta \in \mathcal{P}(\mathcal{V})$, the* ***relative entropy*** *$R(\cdot \| \theta)$ is a mapping from $\mathcal{P}(\mathcal{V})$ into the extended real numbers. It is defined by*

$$R(\gamma \| \theta) \doteq \int_{\mathcal{V}} \left(\log \frac{d\gamma}{d\theta} \right) d\gamma$$

when $\gamma \in \mathcal{P}(\mathcal{V})$ is absolutely continuous with respect to θ and $\log d\gamma/d\theta$ is integrable with respect to γ. Otherwise we set $R(\gamma \| \theta) \doteq \infty$.

Recall that relative entropy $R(\cdot \| \cdot)$ is nonnegative, convex and lower semicontinuous in both variables (with respect to the weak topology on space $\mathcal{P}(\mathcal{V})$). We state the following two properties of relative entropy.

Lemma 1. *(Variational formula) Let $(\mathcal{V}, \mathcal{A})$ be a measurable space, k a bounded measurable function mapping $\mathcal{V}$ into $\mathbb{R}$, and θ a probability measure on $\mathcal{V}$. The following conclusions hold.*

(a) We have the variational formula

$$-\log \int_{\mathcal{V}} e^{-k} d\theta = \inf_{\gamma\in\mathcal{P}(\mathcal{V})} \left\{ R(\gamma \| \theta) + \int_{\mathcal{V}} k d\gamma \right\}. \tag{23}$$

(b) The infimum in (23) is attained uniquely at γ_0 *defined by*

$$\frac{d\gamma_0}{d\theta}(x) \doteq e^{-k(x)} / \int_{\mathcal{V}} e^{-k} d\theta.$$

THEOREM 5. *(Chain rule) Let* $\mathcal{X}$ *and* $\mathcal{Y}$ *be Polish spaces and* β *and* γ *probability measures on* $\mathcal{X} \times \mathcal{Y}$. *We denote by* β_1 *and* γ_1 *the first marginals of* β *and* γ *and by* $\beta(dy|x)$ *and* $\gamma(dy|x)$ *the stochastic kernels on* $\mathcal{Y}$ *given* $\mathcal{X}$ *for which we have the decompositions*

$$\beta(dx \times dy) = \beta_1(dx) \otimes \beta(dy|dx) \quad \text{and} \quad \gamma(dx \times dy) = \gamma_1(dx) \otimes \gamma(dy|dx).$$

Then the function mapping $x \in \mathcal{X} \to R(\beta(\cdot|x) \| \gamma(\cdot|x))$ *is measurable and*

$$R(\beta \| \gamma) = R(\beta_1 \| \gamma_1) + \int_{\mathcal{X}} R(\beta(\cdot|x) \| \gamma(\cdot|x)) \beta_1(dx).$$

CHAPTER 3

Proof of Theorem 4 and a discussion on the boundedness of the rate function

1. The Laplace principle

In Definition 2 we have defined the Laplace principle, we also pointed out there that the Laplace principle is equivalent to large deviations principle. In this chapter, we will make use of this equivalence. In order to establish Theorem 4, we prove the Laplace principle under all the assumptions in Chapter 2 Section 3 and for I defined as in Chapter 2 Section 4. Specifically, we show that for any bounded continuous function $F : \mathcal{P}(S) \to \mathbb{R}$

$$\lim_{T\to\infty} -\frac{1}{T} \log E\left[\exp\left\{-TF(\eta_T)\right\}\right] = \inf_{\eta\in\mathcal{P}(S)} \left[F(\eta) + I(\eta)\right]. \tag{24}$$

By adding a constant to both sides of (24) we can assume $F \geq 0$, and we do that for the rest of the this chapter. The proof is based on a weak convergence approach and is split into two parts: a Laplace upper bound and a Laplace lower bound.

2. Proof of the Laplace upper bound

In this section, we prove the Laplace upper bound part of (24), i.e.

$$\liminf_{T\to\infty} -\frac{1}{T} \log E\left[\exp\left\{-TF(\eta_T)\right\}\right] \geq \inf_{\eta\in\mathcal{P}(S)} \left[F(\eta) + I(\eta)\right]. \tag{25}$$

Recall all the notations used in Chapter 2 Section 2. We first define a random integer R_T as the index when the total "holding time" first exceeds T, i.e.

$$\sum_{i=1}^{R_T-1} s_i \leq T < \sum_{i=1}^{R_T} s_i. \tag{26}$$

The empirical measure η_T can thus be written as

$$\begin{aligned}
\eta_T(\cdot) &= \frac{1}{T}\int_0^T \delta_{X(t)}(\cdot)\,dt \\
&= \frac{1}{T}\left[\sum_{i=1}^{R_T-1} \delta_{X_{i-1}}(\cdot)\, s_i + \delta_{X_{R_T-1}}(\cdot)\left(T - \sum_{i=1}^{R_T-1} s_i\right)\right] \\
&= \frac{1}{T}\left[\sum_{i=1}^{R_T-1} \delta_{X_{i-1}}(\cdot)\frac{\tau_i}{q(X_{i-1})} + \delta_{X_{R_T-1}}(\cdot)\left(T - \sum_{i=1}^{R_T-1} \frac{\tau_i}{q(X_{i-1})}\right)\right]. \qquad (27)
\end{aligned}$$

Fix some constant $C > 0$. The proof of (25) will be partitioned into two cases: $R_T/T > C$ and $0 \le R_T/T \le C$. We will later send C to ∞ after sending $T \to \infty$.

2.1. The case $R_T/T > C$. As we have discussed in Chapter 2 Section 1, let $F : \mathcal{P}(S) \to \mathbb{R}$ be nonnegative and continuous. Then

$$\begin{aligned}
-\frac{1}{T}\log E\left[1_{(C,\infty)}(R_T/T)\cdot \exp\{-TF(\eta_T)\}\right] &\ge -\frac{1}{T}\log P\left\{\sum_{i=1}^{\lfloor TC\rfloor+1} s_i \le T\right\} \\
&= -\frac{1}{T}\log P\left\{\sum_{i=1}^{\lfloor TC\rfloor+1} \frac{\tau_i}{q(X_{i-1})} \le T\right\} \\
&\ge -\frac{1}{T}\log P\left\{\sum_{i=1}^{\lfloor TC\rfloor+1} \tau_i \le K_2 T\right\}.
\end{aligned}$$

Using Chebyshev's inequality, for any $\alpha \in (0,\infty)$

$$\begin{aligned}
P\left\{\sum_{i=1}^{\lfloor TC\rfloor+1} \tau_i \le K_2 T\right\} &= P\left\{e^{-\alpha\sum_{i=1}^{\lfloor TC\rfloor+1}\tau_i} \ge e^{-\alpha K_2 T}\right\} \\
&\le e^{\alpha K_2 T} E\left[e^{-\alpha\sum_{i=1}^{\lfloor TC\rfloor+1}\tau_i}\right] \\
&= e^{\alpha K_2 T} e^{(\lfloor TC\rfloor+1)\log\frac{1}{1+\alpha}}.
\end{aligned}$$

For the last equality we have used the fact that if τ is distributed according to σ then $Ee^{a\tau} = 1/(1-a)$ for any $a \in (-\infty, 1)$. Combining the last two inequalities, we have

$$\begin{aligned}
\liminf_{T\to\infty} -\frac{1}{T}\log E\left[1_{(C,\infty)}(R_T/T)\cdot\exp\{-TF(\eta_T)\}\right] \\
&\geq \sup_{\alpha\in(0,\infty)}\left[-K_2\alpha + C\log(1+\alpha)\right] \\
&= C + C\log C - K_2 - C\log K_2.
\end{aligned}$$

Note that $C + C\log C - K_2 - C\log K_2 \to \infty$ as $C \to \infty$.

2.2. The case $0 \leq R_T/T \leq C$.

2.2.1. *A stochastic control representation.* In this case we adapt a standard weak convergence argument, see [**7**] for details. Specifically, we first establish a stochastic control representation for the left hand side of (24) and then obtain a lower bound for the limit as $T \to \infty$. For the representation, all distributions can be perturbed from their original form, but such a perturbation pays a relative entropy cost. We distinguish the new distributions and random variables by using an overbar. In the following, the barred quantities are constructed analogously to their unbarred counterparts. Hence $\bar{\tau}_i$ and $\bar{X}_i$ are chosen recursively according to stochastic kernels $\bar{\sigma}_i(\cdot)$ and $\bar{\alpha}_i(\cdot)$, i.e., $\bar{\sigma}_i(\cdot)$ and $\bar{\alpha}_i(\cdot)$ are conditional distributions that can depend on the whole past. Specifically, $\bar{\sigma}_i(\cdot)$ depends on $\left\{\bar{X}_0, \bar{\tau}_1, \bar{X}_1, \bar{\tau}_2, \ldots, \bar{X}_{i-1}\right\}$ and $\bar{\alpha}_i(\cdot)$ depends on $\left\{\bar{X}_0, \bar{\tau}_1, \bar{X}_1, \bar{\tau}_2 \ldots, \bar{X}_{i-1}, \bar{\tau}_i\right\}$; $\bar{s}_i$ is defined by (15) using $\bar{X}_i$ and $\bar{\tau}_i$; $\bar{R}_T$ is defined by (26) using $\bar{s}_i$; and $\bar{\eta}_T$ is defined by (27) using $\bar{X}_i$, $\bar{\tau}_i$ and $\bar{R}_T$. It will be sufficient to consider any deterministic sequence $\{r_T\}$ such that $0 \leq r_T/T \leq C$, and $r_T/T \to A$ for some $A \in [0, C]$ as $T \to \infty$. We restrict our consideration to controlled processes such that $\bar{R}_T = r_T$; we place an infinite cost penalty on controls which lead to any other outcome with positive probability. Let $\mathbf{1}(A)$ to denote the indicator function of a set A. By applying [**7**, Proposition 4.5.1] and Theorem 5 the following is valid:

(28)

$$
\begin{aligned}
&-\frac{1}{T}\log E\left[\exp\left\{-TF(\eta_T)-T\cdot\infty\cdot\left(1_{\{r_T/T\}^c}\left(R_T/T\right)\right)\right\}\right]\\
&=-\frac{1}{T}\log E\left[\exp\left\{-TF(\eta_T)-T\cdot\infty\cdot\mathbf{1}\left(\sum_{i=1}^{r_T-1}s_i\leq T<\sum_{i=1}^{r_T}s_i\right)\right\}\right]
\end{aligned}
$$

(29)

$$
=\inf E\left[F\left(\bar{\eta}_T\right)+\infty\cdot\mathbf{1}\left(\sum_{i=1}^{r_T-1}\bar{s}_i\leq T<\sum_{i=1}^{r_T}\bar{s}_i\right)+\frac{1}{T}\sum_{i=1}^{r_T}\left[R\left(\bar{\alpha}_{i-1}\left\|\alpha\right.\right)+R\left(\bar{\sigma}_i\left\|\sigma\right.\right)\right]\right]
$$

where the infimum is taken over all control measures $\{\bar{\alpha}_i,\bar{\sigma}_i\}$. Since in Chapter 3 Section 3 we will show a similar but more involved representation formula-Lemma 7-we omit the proof of this representation. Due to the restriction $\bar{R}_T=r_T$, one can rewrite $\bar{\eta}_T$ as

$$
\bar{\eta}_T\left(\cdot\right)=\frac{1}{T}\left[\sum_{i=1}^{r_T-1}\delta_{\bar{X}_{i-1}}\left(\cdot\right)\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)}+\delta_{\bar{X}_{r_T-1}}\left(\cdot\right)\left(T-\sum_{i=1}^{r_T-1}\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)}\right)\right]. \tag{30}
$$

In the following proof, we repeatedly extract further subsequences of T . To keep the notations concise, we will use T to denote all subsequences. Note also that in proving a lower bound for (28) it suffices to consider a subsequence of T such that

$$
\sup_T -\frac{1}{T}\log E\left[\exp\left\{-TF(\eta_T)-T\cdot\infty\cdot\left(1_{\{r_T/T\}^c}\left(R_T/T\right)\right)\right\}\right]<\infty. \tag{31}
$$

We assume this condition for the rest of this section.

The relative entropy cost in (29) includes two parts, $RE_T^1\doteq\frac{1}{T}\sum_{i=1}^{r_T}R\left(\bar{\alpha}_{i-1}\left\|\alpha\right.\right)$ and $RE_T^2\doteq\frac{1}{T}\sum_{i=1}^{r_T}R\left(\bar{\sigma}_i\left\|\sigma\right.\right)$. We will prove that for any sequence of controls $\{\bar{\alpha}_i,\bar{\sigma}_i\}$ in (29)

$$
\liminf_{T\to\infty}E\left[F\left(\bar{\eta}_T\right)+RE_T^1+RE_T^2\right]\geq\inf_{\eta\in\mathcal{P}(S)}\left[F\left(\eta\right)+I\left(\eta\right)\right]. \tag{32}
$$

Toward this end, it is enough to show that along any subsequence of T such that $r_T/T\to A$, we can extract a further subsequence along which (32) holds. In addition, it suffices to consider only functions F that besides being nonnegative, are also lower semicontinuous and convex. This reduction will be justified at the end of this section, see Remark 6.

In light of (29) and (31) we can assume

$$\sup_T E\left[F(\bar{\eta}_T) + RE_T^1 + RE_T^2\right] < \infty \tag{33}$$

Since the proof of (32) is lengthy, we will analyze each term on the left hand side of (32) separately.

2.2.2. *The term RE_T^1.* The cost RE_T^1 comes from distorting the dynamics of the embedded Markov chain, and indeed the analysis results in a very similar conclusion to that of an ordinary Markov chain ([**7**, Chapter 8]). For any probability measure ν on $S \times S$ we will use notations $[\nu]_1$ and $[\nu]_2$ to denote the first and second marginals of ν. We have the following result for RE_T^1.

LEMMA 2. *Consider any sequence of controls $\{\bar{\alpha}_i, \bar{\sigma}_i\}$ in (29) such that (33) holds. Along any subsequence of T satisfying $r_T/T \to A$, define a sequence of probability measures on $S \times S$ via*

$$\mu_T(dx, dy) \doteq \frac{1}{r_T} \sum_{i=1}^{r_T} \delta_{\bar{X}_{i-1}}(dx)\, \bar{\alpha}_{i-1}(dy).$$

Then one can extract a further subsequence such that $E\mu_T$ converge in distribution to a probability measure $\tilde{\mu}$ on $S \times S$, and

$$\liminf_{T\to\infty} E\left[RE_T^1\right] \geq AR\left(\tilde{\mu} \,\|[\tilde{\mu}]_1 \otimes \alpha\right).$$

Furthermore, if $A > 0$ then $\tilde{\mu}$ satisfies

$$[\tilde{\mu}]_1 = [\tilde{\mu}]_2. \tag{34}$$

PROOF. By the chain rule (Theorem 5) and the convexity of relative entropy

$$
\begin{aligned}
E\left[RE_T^1\right] &= E\left[\frac{1}{T}\sum_{i=1}^{r_T} R\left(\bar{\alpha}_{i-1}\,\|\alpha\right)\right] \\
&= E\left[\frac{1}{T}\sum_{i=1}^{r_T} R\left(\delta_{\bar{X}_{i-1}}\otimes\bar{\alpha}_{i-1}\,\middle\|\delta_{\bar{X}_{i-1}}\otimes\alpha\right)\right] \\
&\geq E\left[\frac{r_T}{T} R\left(\mu_T\,\|[\mu_T]_1\otimes\alpha\right)\right] \\
&\geq \frac{r_T}{T} R\left(E\mu_T\,\|[E\mu_T]_1\otimes\alpha\right).
\end{aligned}
$$

Since $S\times S$ is compact, for any subsequence of T there exits a further subsequence along which $E\mu_T$ converges weakly to a probability measure $\tilde{\mu}$. Under the Feller property of α, $[E\mu_T]_1\otimes\alpha$ converges weakly to $[\tilde{\mu}]_1\otimes\alpha$. The lower semicontinuity of relative entropy then implies

$$
\liminf_{T\to\infty} E\left[RE_T^1\right] \geq \liminf_{T\to\infty}\frac{r_T}{T} R\left(E\mu_T\,\|[E\mu_T]_1\otimes\alpha\right) \geq AR\left(\tilde{\mu}\,\|[\tilde{\mu}]_1\otimes\alpha\right).
$$

This finishes the first part of Lemma 2. For the second part, we employ a standard martingale argument. Let $\mathcal{F}_i$ be the σ-algebra generated by the random variables $\left\{\left(\bar{X}_0,\ldots,\bar{X}_i\right),\left(\bar{\tau}_1,\ldots,\bar{\tau}_i\right)\right\}$. Thus $\mathcal{F}_i$ is a sequence of increasing σ-algebra's and, since $\bar{\alpha}_i$ selects the conditional distribution of $\bar{X}_i$, for any bounded continuous function f on S

$$
E\left[\left(f\left(\bar{X}_i\right)-\int_S f(y)\,\bar{\alpha}_i(dy)\right)|\mathcal{F}_{i-1}\right]=0.
$$

Hence for integers $0\leq i<k\leq r_T-1$

$$
E\left[\left(f\left(\bar{X}_i\right)-\int_S f(y)\,\bar{\alpha}_i(dy)\right)\left(f\left(\bar{X}_k\right)-\int_S f(y)\,\bar{\alpha}_k(dy)\right)\right]=0,
$$

and thus for any bounded continuous function f on S

$$
\begin{aligned}
& E\left[\int_{S\times S} f\left(x\right)\mu_T\left(dx,dy\right) - \int_{S\times S} f\left(y\right)\mu_T\left(dx,dy\right)\right]^2 \\
& = E\left[\frac{1}{r_T}\sum_{i=1}^{r_T} f\left(\bar{X}_{i-1}\right) - \frac{1}{r_T}\sum_{i=1}^{r_T}\int_S f\left(y\right)\bar{\alpha}_{i-1}\left(dy\right)\right]^2 \\
& \leq \frac{1}{r_T^2}\sum_{i=1}^{r_T} E\left[f\left(\bar{X}_{i-1}\right) - \int_S f\left(y\right)\bar{\alpha}_{i-1}\left(dy\right)\right]^2 \\
& \leq \frac{4}{r_T}\left\|f\right\|_\infty^2 .
\end{aligned}
$$

Since $0 < A = \lim_{T\to\infty} r_T/T$, we have $r_T/T \geq A/2$ for all T large enough. Using Chebyshev's inequality and the last display we conclude that $\left[\mu_T\right]_1 - \left[\mu_T\right]_2$ converges to 0 in probability as $T \to \infty$, and therefore $\left[\tilde{\mu}\right]_1 = \left[\tilde{\mu}\right]_2$ with probability 1. This concludes the second part of Lemma 2. $\square$

2.2.3. *The term RE_T^2.* We now turn to the second cost RE_T^2. This cost comes from distorting the exponential sojourn times. We introduce a function ℓ which is closely related to the relative entropy of exponential distributions: $\ell\left(x\right) \doteq x\log x - x + 1$ for any $x \geq 0$.

LEMMA 3. *Given any sequence of controls $\{\bar{\alpha}_i, \bar{\sigma}_i\}$, fix a subsequence of T for which the conclusions in Lemma 2 holds. Then we can further extract a subsequence along which*

$$
\liminf_{T\to\infty} E\left[RE_T^2\right] \geq \int_{S\times\mathbb{R}_+} \ell\left(u\right)\tilde{\xi}\left(dx,du\right).
$$

Here $\tilde{\xi}$ is a finite measure on $S\times\mathbb{R}_+$ and is related to $\tilde{\mu}$ in Lemma 2 by

$$
\int_{\mathbb{R}_+} u\tilde{\xi}\left(dx,du\right) = A\left[\tilde{\mu}\right]_1\left(dx\right). \tag{35}
$$

Before proving this lemma, we need to define $g : \mathbb{R}_+ \to \mathbb{R}$ by $g\left(b\right) \doteq -\log b + b - 1$. The functions g and ℓ are related by

$$
g\left(x\right) = x\ell\left(1/x\right),
$$

and g has the following property.

LEMMA 4. *Let σ be an exponential distribution with mean 1. Then*

$$\inf\left\{R(\gamma\,\|\sigma) : \int_{\mathbb{R}_+} u\gamma(du) = b\right\} = g(b). \tag{36}$$

PROOF. Let σ_b be the exponential distribution with mean b, i.e.,

$$\sigma_b(du) = \frac{1}{b}e^{-\frac{u}{b}}du.$$

Then $\frac{d\sigma_b}{d\sigma}(u) = \frac{1}{b}e^{\left(1-\frac{1}{b}\right)u}$ for $u > 0$. Picking any γ such that $R(\gamma\,\|\sigma) < \infty$,

$$\begin{aligned}
R(\gamma\,\|\sigma) &= \int_{\mathbb{R}_+} \log\left(\frac{d\gamma}{d\sigma}\right)\gamma(du) \\
&= \int_{\mathbb{R}_+} \log\left(\frac{d\gamma}{d\sigma_b}\right)\gamma(du) + \int_{\mathbb{R}_+} \log\left(\frac{d\sigma_b}{d\sigma}\right)\gamma(du) \\
&= R(\gamma\,\|\sigma_b) + \int_{\mathbb{R}_+}\left[-\log b + \left(1-\frac{1}{b}\right)u\right]\gamma(du) \\
&= R(\gamma\,\|\sigma_b) + g(b) \\
&\geq g(b)
\end{aligned}$$

and the infimum in (36) is achieved when $R(\gamma\,\|\sigma_b) = 0$, i.e., $\gamma = \sigma_b$. □

PROOF OF LEMMA 3. Lemma 4 guarantees that

$$RE_T^2 \geq \frac{1}{T}\sum_{i=1}^{r_T} g\left(\int u\bar{\sigma}_i(du)\right). \tag{37}$$

Recall the definition of $\mathcal{F}_i$ as the σ-algebra generated by the controlled process up to time i. Since $\bar{\sigma}_i$ selects the conditional distribution of $\bar{\tau}_i$,

$$E\left[\bar{\tau}_i|\mathcal{F}_{i-1}\right] = \int u\bar{\sigma}_i(du).$$

Define $\bar{m}_i \doteq \int u\bar{\sigma}_i(du)$, for $i = 1, \ldots, r_T - 1$. The definition of $\bar{m}_{r_T}$ requires more work. Recalling the definition of $\bar{R}_T$ by the equation analogous to (26) and the restriction that $\bar{R}_T = r_T$,

$$T - \sum_{i=1}^{r_T-1}\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)} \leq \frac{\bar{\tau}_{r_T}}{q\left(\bar{X}_{r_T-1}\right)}.$$

Multiplying both sides by $q\left(\bar{X}_{r_T-1}\right)$ and taking expectations conditioned on $\mathcal{F}_{r_T-1}$,

$$q\left(\bar{X}_{r_T-1}\right)\left(T-\sum_{i=1}^{r_T-1}\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)}\right)\leq E\left[\bar{\tau}_{r_T}|\mathcal{F}_{r_T-1}\right]=\int u\bar{\sigma}_{r_T}\left(du\right).$$

Define

$$\bar{\Delta}_T\doteq q\left(\bar{X}_{r_T-1}\right)\left(T-\sum_{i=1}^{r_T-1}\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)}\right), \tag{38}$$

and define $\bar{m}_{r_T}$ by

$$\bar{m}_{r_T}\doteq\begin{cases}\int u\bar{\sigma}_{r_T}\left(du\right) & \text{if } \bar{\Delta}_T\leq\int u\bar{\sigma}_{r_T}\left(du\right)<1\\ 1 & \text{if } \bar{\Delta}_T\leq 1\leq\int u\bar{\sigma}_{r_T}\left(du\right)\\ \bar{\Delta}_T & \text{if } 1<\bar{\Delta}_T\leq\int u\bar{\sigma}_{r_T}\left(du\right)\end{cases} \tag{39}$$

i.e., $\bar{m}_{r_T}$ is the median of the triplet $\left(\bar{\Delta}_T,\int u\bar{\sigma}_{r_T}\left(du\right),1\right)$. Since g is increasing on $(1,\infty)$, we have $g\left(\int u\bar{\sigma}_{r_T}\left(du\right)\right)\geq g\left(\bar{m}_{r_T}\right)$ in all three cases. Thus by (37),

$$RE_T^2\geq\frac{1}{T}\sum_{i=1}^{r_T}g\left(\int u\bar{\sigma}_i\left(du\right)\right)\geq\frac{1}{T}\sum_{i=1}^{r_T}g\left(\bar{m}_i\right). \tag{40}$$

Next consider the measure on $S\times\mathbb{R}_+$ defined by

$$\xi_T\left(dx,du\right)\doteq\frac{1}{T}\sum_{i=1}^{r_T}\delta_{\bar{X}_{i-1}}\left(dx\right)\delta_{\left(\bar{m}_i\right)^{-1}}\left(du\right)\bar{m}_i. \tag{41}$$

The total mass of $E\xi_T$ is

$$E\xi_T\left(S\times\mathbb{R}_+\right)=\frac{1}{T}\sum_{i=1}^{r_T}E\left[\bar{m}_i\right].$$

According to (33) and the assumption that $F\geq 0$, we have

$$\sup_T E\left[RE_T^2\right]<\infty. \tag{42}$$

By (40) $\sup_T E\left[\sum_{i=1}^{r_T}g\left(\bar{m}_i\right)/T\right]<\infty$. We also have by a straightforward calculation that $x\leq\max\left\{50,10g\left(x\right)/9\right\}$. Using this and the fact that $r_T/T\leq C$ we have $\sup_T E\left[\sum_{i=1}^{r_T}\bar{m}_i/T\right]<\infty$, i.e., the total mass of $E\xi_T$ has uniform bound in T. Thus when viewed as a sequence of measures on the compact space $S\times[0,\infty]$, $E\xi_T$ is tight due to the uniform boundedness of the total mass. We denote the weak limit by $\tilde{\xi}$, which is a finite measure. Since the function ℓ is nonnegative and continuous,

$$\begin{aligned}
\liminf_{T\to\infty} E\left[RE_T^2\right] &\geq \liminf_{T\to\infty} E\left[\frac{1}{T}\sum_{i=1}^{r_T} g\left(\bar{m}_i\right)\right] \\
&= \liminf_{T\to\infty} E\left[\int_{S\times\mathbb{R}_+} \ell\left(u\right)\xi_T\left(dx,du\right)\right] \\
&= \liminf_{T\to\infty} \int_{S\times\mathbb{R}_+} \ell\left(u\right) E\xi_T\left(dx,du\right) \qquad (43)\\
&\geq \int_{S\times\mathbb{R}_+} \ell\left(u\right)\tilde{\xi}\left(dx,du\right).
\end{aligned}$$

We next explore the relation between $\tilde{\xi}$ and $\tilde{\mu}$. In order to establish (35), it suffices to show that for any bounded continuous function f on S

$$\int_{S\times\mathbb{R}_+} uf\left(x\right)\tilde{\xi}\left(dx,du\right) = A\int_S f\left(x\right)\left[\tilde{\mu}\right]_1\left(dx\right).$$

By the definitions of ξ_T and μ_T

$$\int_{S\times\mathbb{R}_+} uf\left(x\right)E\xi_T\left(dx,du\right) = \frac{r_T}{T}\int_S f\left(x\right)\left[E\mu_T\right]_1\left(dx\right). \qquad (44)$$

Then (42) and (43) imply there is a uniform upper bound on

$$\int_{\mathbb{R}_+} \ell\left(u\right)\int_S f\left(x\right)E\xi_T\left(dx,du\right).$$

If we consider $\int_S f\left(x\right)E\xi_T\left(dx,du\right)$ as a sequence of measures on $\mathbb{R}_+$ with bounded total mass, then $\int_S f\left(x\right)E\xi_T\left(dx,du\right)$ converges weakly to $\int_S f\left(x\right)\tilde{\xi}\left(dx,du\right)$. Since ℓ is superlinear, [**7**, Theorem A.3.19] implies that

$$\lim_{T\to\infty}\int_{S\times\mathbb{R}_+} uf\left(x\right)E\xi_T\left(dx,du\right) = \int_{S\times\mathbb{R}_+} uf\left(x\right)\tilde{\xi}\left(dx,du\right).$$

Using

$$\lim_{T\to\infty}\frac{r_T}{T}\int_S f\left(x\right)\left[E\mu_T\right]_1\left(dx\right) = A\int_S f\left(x\right)\left[\tilde{\mu}\right]_1\left(dx\right)$$

and (44) we arrive at (35). □

2.2.4. *The term $E\bar{\eta}_T$.*

LEMMA 5. *Given any sequence of controls $\{\bar{\alpha}_i, \bar{\sigma}_i\}$, fix a subsequence of T for which the conclusions in Lemma 3 hold. Then we can further extract a subsequence along which*

$$\liminf_{T\to\infty} E\left[F\left(\bar{\eta}_T\right)\right] \geq F\left(\tilde{\eta}\right)$$

for some probability measure $\tilde{\eta}$ on S, which is related to $\tilde{\xi}$ in Lemma 3 by

$$q\left(x\right)\tilde{\eta}\left(dx\right) = [\tilde{\xi}]_1\left(dx\right). \tag{45}$$

PROOF. As a sequence of probability measures on the compact space S, we can always extract a subsequence of T such that $E\bar{\eta}_T$ converges weakly to a probability measure on S which we denote by $\tilde{\eta}$. The convexity and lower semicontinuity of F imply that

$$\liminf_{T\to\infty} E\left[F\left(\bar{\eta}_T\right)\right] \geq \liminf_{T\to\infty} F\left(E\bar{\eta}_T\right) \geq F\left(\tilde{\eta}\right).$$

By the definitions of $\bar{\eta}_T$ in (30) and $\bar{\Delta}_T$ in (38)

$$\begin{aligned}
&q\left(x\right)E\bar{\eta}_T\left(dx\right)\\
&= \frac{q\left(x\right)}{T}E\left[\sum_{i=1}^{r_T-1}\delta_{\bar{X}_{i-1}}\left(dx\right)\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)} + \delta_{\bar{X}_{r_T-1}}\left(dx\right)\left(T - \sum_{i=1}^{r_T-1}\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)}\right)\right]\\
&= \frac{1}{T}E\left[\sum_{i=1}^{r_T-1}\delta_{\bar{X}_{i-1}}\left(dx\right)\bar{\tau}_i + \delta_{\bar{X}_{r_T-1}}\left(dx\right)\bar{\Delta}_T\right]\\
&= \frac{1}{T}\left(\sum_{i=1}^{r_T-1}E\left[E\left[\delta_{\bar{X}_{i-1}}\left(dx\right)\bar{\tau}_i|\mathcal{F}_{i-1}\right]\right] + E\left[\delta_{\bar{X}_{r_T-1}}\left(dx\right)\bar{\Delta}_T\right]\right)\\
&= \frac{1}{T}\left(\sum_{i=1}^{r_T-1}E\left[\delta_{\bar{X}_{i-1}}\left(dx\right)\bar{m}_i\right] + E\left[\delta_{\bar{X}_{r_T-1}}\left(dx\right)\bar{\Delta}_T\right]\right).
\end{aligned}$$

Recalling the definition of ξ_T in (41), we have

$$\left[E\xi_T\right]_1\left(dx\right) = \frac{1}{T}\sum_{i=1}^{r_T}E\left[\delta_{\bar{X}_{i-1}}\left(dx\right)\bar{m}_i\right].$$

This implies the total variation bound

$$\|q(x)\,E\bar{\eta}_T(dx) - [E\xi_T]_1(dx)\|_{\mathrm{TV}} \le \frac{1}{T} E\left|\bar{\Delta}_T - \bar{m}_{r_T}\right|.$$

Recalling the definition of $\bar{m}_{r_T}$ in (39) we conclude that

$$\|q(x)\,E\bar{\eta}_T(dx) - [E\xi_T]_1(dx)\|_{\mathrm{TV}} \le \frac{1}{T}.$$

By taking limits we arrive at (45). □

Now Lemma 2, Lemma 3 and Lemma 5 together imply for a sequence of controls $\{\bar{\alpha}_i, \bar{\sigma}_i\}$ satisfying (29), along any subsequence of T such that $r_T/T \to A$, we can extract a further subsequence along which

(46)

$$\liminf_{T\to\infty} E\left[F(\bar{\eta}_T) + RE_T^1 + RE_T^2\right] \ge F(\eta) + AR(\mu\,\|[\mu]_1 \otimes \alpha) + \int_{S\times\mathbb{R}_+} \ell(u)\,\xi(dx,du)$$

where η, μ and ξ satisfy the constraints (35), (45), and (34) if $A > 0$.

Recall that our goal is to prove (32). Hence we need to establish the relationship between the right hand side of (46) and the rate function I defined in Chapter 2 Section 4.

2.2.5. *Properties of the rate function I.* We prove the following lemma, for which we adopt the convention $0 \cdot \infty \doteq 0$. This is in fact the key link, showing that the rate function that is naturally obtained by the weak convergence analysis used to prove the upper bound in fact equals I for suitable measures, and also indicating how to construct controls to prove the lower bound for this same collection of measures.

Lemma 6. *Let $I(\eta)$ be defined by (22). Suppose that $\eta \ll \pi$, that μ and ξ satisfy the constraints*

$$q(x)\,\eta(dx) = [\xi]_1(dx) \ \ \text{and} \ \int_{\mathbb{R}_+} u\xi(dx,du) = A[\mu]_1(dx), \tag{47}$$

and that when $A > 0$ the constraint $[\mu]_1 = [\mu]_2$ is also true. Then

$$I(\eta) \le AR(\mu\,\|[\mu]_1 \otimes \alpha) + \int_{S\times\mathbb{R}_+} \ell(u)\,\xi(dx,du). \tag{48}$$

Moreover,

$$I(\eta) = \inf \left[AR\left(\mu \,\|\, [\mu]_1 \otimes \alpha\right) + \int_{S\times\mathbb{R}_+} \ell(u)\,\xi(dx,du) \right]$$

where the infimum is over all possible choices of $A \geq 0$, μ and ξ satisfying these constraints.

The proof of this lemma is detailed. The reason we present it here instead of in an appendix is the previously mentioned fact that the construction of A, μ and ξ that minimize the right hand side of (48) indicates how to hit target measures η that are absolutely continuous with respect to the invariant measure in the proof of the Laplace lower bound.

Proof. We first prove the inequality (48). If the right hand side of (48) is ∞, there is nothing to prove. Hence we assume it is finite. First assume $A > 0$, in which case $R\left(\mu \,\|\, [\mu]_1 \otimes \alpha\right) < \infty$. Define

$$Q \doteq \int_S q(x)\,\pi(dx), \tag{49}$$

so that by (18)

$$\widetilde{\pi}(dx) = q(x)\,\pi(dx)/Q. \tag{50}$$

Since $\widetilde{\pi}$ is invariant under α, by [**7**, Lemma 8.6.2] $[\mu]_1 \ll \widetilde{\pi}$. By (17) q is bounded from below, and hence $[\mu]_1 \ll \pi$. Recall that the definition of I in (22) uses $\theta = d\eta/d\pi$. Define $\Theta \doteq \{x \in S : \theta(x) = 0\}$. By (47)

$$\begin{aligned}
[\mu]_1(dx) &= \frac{\int_{\mathbb{R}_+} u\xi_{2|1}(du|x)}{A}[\xi]_1(dx) \\
&= \frac{\int_{\mathbb{R}_+} u\xi_{2|1}(du|x)}{A} q(x)\,\eta(dx) \\
&= \frac{\int_{\mathbb{R}_+} u\xi_{2|1}(du|x)}{A} q(x)\,\theta(x)\,\pi(dx) \qquad (51)
\end{aligned}$$

where for a measure ν on $S\times\mathbb{R}_+$, $\nu_{2|1}$ denotes the regular conditional distribution on the second argument given the first. Thus $[\mu]_1(\Theta) = 0$. Now suppose that

$$\int_{S\times S} \theta^{1/2}(x)\,\theta^{1/2}(y)\,q(x)\,\alpha(x,dy)\,\pi(dx) = 0.$$

Then for π-a.e. $x \in S \backslash \Theta$, $\alpha(x, \Theta) = 1$, and hence $(\mu_1 \otimes \alpha)[(S\backslash\Theta) \times \Theta] = 1$. On the other hand, $\mu((S\backslash\Theta) \times \Theta) = 0$ due to $[\mu]_1 = [\mu]_2$. This violates the fact that $R(\mu \| [\mu]_1 \otimes \alpha) < \infty$. We conclude that

$$\int_{S\times S} \theta^{1/2}(x)\, \theta^{1/2}(y)\, q(x)\, \alpha(x, dy)\, \pi(dx) > 0.$$

Lemma 1 implies that

$$\begin{aligned} &-\log \int_{S\times S} \theta^{1/2}(x)\, \theta^{1/2}(y)\, \alpha(x, dy)\, \widetilde{\pi}(dx) \\ &\quad = -\log \int_{S\times S} e^{\frac{1}{2}[\log\theta(x) + \log\theta(y)]} \alpha(x, dy)\, \widetilde{\pi}(dx) \\ &\quad \leq R(\mu \| \widetilde{\pi} \otimes \alpha) - \frac{1}{2} \int_{S\times S} [\log\theta(x) + \log\theta(y)]\, \mu(dx, dy). \end{aligned} \tag{52}$$

Strictly speaking, the inequality above does not fall into the framework of Lemma 1 because $\log\theta$ is not bounded. However, if one goes through the proof of this lemma [**7**, Proposition 1.4.2], then the above inequality is true as long as the right hand side is not of the form $\infty - \infty$. Towards this end, it suffices to prove

$$\frac{1}{2} \int_{S\times S} [\log\theta(x) + \log\theta(y)]\, \mu(dx, dy) = \int_S \log\theta(x)\, [\mu]_1(dx) < \infty. \tag{53}$$

In the appendix we will prove

$$R([\mu]_1 \| \widetilde{\pi}) < \infty. \tag{54}$$

For now, we assume this is true. Using (49), (50), and (51) to evaluate the relative entropy,

(55)

$$\infty > R([\mu]_1 \| \widetilde{\pi}) = \int_S \log\left(\int_{\mathbb{R}_+} u \xi_{2|1}(du|x) \right) [\mu]_1(dx) + \int_S \log\theta(x)\, [\mu]_1(dx) + \log\frac{Q}{A}.$$

We know from (17) that $Q \geq K_1$. Also, by (47) and the nonnegativity of ℓ

$$
\begin{aligned}
&\int_S \log\left(\int_{\mathbb{R}_+} u\xi_{2|1}(du|x)\right)[\mu]_1(dx) \\
&= \frac{1}{A}\int_S \left(\int_{\mathbb{R}_+} u\xi_{2|1}(du|x)\right)\log\left(\int_{\mathbb{R}_+} u\xi_{2|1}(du|x)\right)[\xi]_1(dx) \\
&= \frac{1}{A}\int_S \ell\left(\int_{\mathbb{R}_+} u\xi_{2|1}(du|x)\right)[\xi]_1(dx) + \frac{1}{A}\int_{S\times\mathbb{R}_+} u\xi(dx,du) - \frac{1}{A}\int_S [\xi]_1(dx) \\
&= \frac{1}{A}\int_S \ell\left(\int_{\mathbb{R}_+} u\xi_{2|1}(du|x)\right)[\xi]_1(dx) + \int_S [\mu]_1(dx) - \frac{1}{A}\int_S q(x)\eta(dx) \qquad (56)\\
&\geq 1 - \frac{1}{A}K_2.
\end{aligned}
$$

the second constraint in (47) is used for the first equality; the definition of ℓ gives the second equality; both parts of (47) assure the third equality; finally the nonnegativity of ℓ is used. Thus (55) implies (53). The chain rule of relative entropy gives

$$
\begin{aligned}
&R(\mu\,\|\widetilde{\pi}\otimes\alpha) - \frac{1}{2}\int_{S\times S}[\log\theta(x) + \log\theta(y)]\,\mu(dx,dy) \\
&= R([\mu]_1\,\|\widetilde{\pi}) + \int_S R\left(\mu_{2|1}\,\|\alpha\right)[\mu]_1(dx) - \int_S \log\theta(x)\,[\mu]_1(dx) \\
&= R([\mu]_1\,\|\widetilde{\pi}) + R(\mu\,\|[\mu]_1\otimes\alpha) - \int_S \log\theta(x)\,[\mu]_1(dx). \qquad (57)
\end{aligned}
$$

By (55) and (56) and the convexity of ℓ

$$
\begin{aligned}
&R([\mu]_1\,\|\widetilde{\pi}) - \int_S \log\theta(x)\,[\mu]_1(dx) \\
&= \int_S \log\left(\int_{\mathbb{R}_+} u\xi_{2|1}(du|x)\right)[\mu]_1(dx) + \log\frac{Q}{A} \\
&= \frac{1}{A}\int_S \ell\left(\int_{\mathbb{R}_+} u\xi_{2|1}(du|x)\right)[\xi]_1(dx) + \int_S [\mu]_1(dx) - \frac{1}{A}\int_S q(x)\eta(dx) + \log\frac{Q}{A} \qquad (58)\\
&\leq \frac{1}{A}\int_{S\times\mathbb{R}_+} \ell(u)\,\xi(dx,du) + 1 - \frac{1}{A}\int_S q(x)\eta(dx) + \log\frac{Q}{A}.
\end{aligned}
$$

In summary (52), (57) and (58) imply

$$
\begin{aligned}
&-\log \int_{S\times S} \theta^{1/2}(x)\,\theta^{1/2}(y)\,q(x)\,\alpha(x,dy)\,\pi(dx) \\
&\qquad = -\log \int_{S\times S} \theta^{1/2}(x)\,\theta^{1/2}(y)\,\alpha(x,dy)\,\widetilde{\pi}(dx) - \log Q \\
&\qquad \le R\left(\mu \,\|[\mu]_1 \otimes \alpha\right) + \frac{1}{A}\int_{S\times\mathbb{R}_+} \ell(u)\,\xi(dx,du) + 1 - \frac{1}{A}\int_S q(x)\,\eta(dx) + \log\frac{1}{A}.
\end{aligned}
$$

Thus

$$
\begin{aligned}
&-\int_{S\times S} \theta^{1/2}(x)\,\theta^{1/2}(y)\,q(x)\,\alpha(x,dy)\,\pi(dx) \\
&\le -\exp\left\{-\left(R\left(\mu \,\|[\mu]_1 \otimes \alpha\right) + \frac{1}{A}\int_{S\times\mathbb{R}_+} \ell(u)\,\xi(dx,du) + 1 - \frac{1}{A}\int_S q(x)\,\eta(dx) + \log\frac{1}{A}\right)\right\}.
\end{aligned}
$$

(48) then follows by the fact that $-e^{-r} \le ar + a\log a - a$ for any $r \in \mathbb{R}$ and $a \in \mathbb{R}_+$ by taking $a = A$ and

$$
r = R\left(\mu \,\|[\mu]_1 \otimes \alpha\right) + \frac{1}{A}\int_{S\times\mathbb{R}_+} \ell(u)\,\xi(dx,du) + 1 - \frac{1}{A}\int_S q(x)\,\eta(dx) + \log\frac{1}{A}.
$$

For the case when $A = 0$, (47) implies that $\int_{\mathbb{R}_+} u\xi(dx,du) = 0$, which means that $\int_{\mathbb{R}_+} u\xi_{2|1}(du|x) = 0$ $[\xi]_1$-a.e. Hence by the convexity of ℓ and $q(x)\,\eta(dx) = [\xi]_1(dx)$

$$
\begin{aligned}
\int_{S\times\mathbb{R}_+} \ell(u)\,\xi(dx,du) &\ge \int_S \ell\left(\int_{\mathbb{R}_+} u\xi_{2|1}(du|x)\right)[\xi]_1(dx) \\
&= \int_S [\xi]_1(dx) \\
&= \int_S q(x)\,\eta(dx) \\
&\ge \int_S q(x)\,\eta(dx) - \int_{S\times S} \theta^{1/2}(x)\,\theta^{1/2}(y)\,q(x)\,\alpha(x,dy)\,\pi(dx) \\
&= I(\eta).
\end{aligned}
$$

Thus (48) also holds in this case, and completes the proof of the first part of Lemma 6.

We now turn to the second part of Lemma 6. In light of the second part of Lemma (1), we define μ by

$$
\text{(59)} \qquad \frac{d\mu}{d(\widetilde{\pi}\otimes\alpha)}(x,y) \doteq \theta^{1/2}(x)\,\theta^{1/2}(y) \Big/ \int_{S\times S} \theta^{1/2}(x)\,\theta^{1/2}(y)\,(\widetilde{\pi}\otimes\alpha)(dx,dy)\,.
$$

Note that by the Cauchy-Schwartz inequality, the detailed balance condition (20) implies

$$\int_{S\times S}\theta^{1/2}(x)\,\theta^{1/2}(y)\,(\widetilde{\pi}\otimes\alpha)(dx,dy)\leq\int_{S\times S}\theta(x)\,\alpha(x,dy)\,\widetilde{\pi}(dx)\leq\frac{K_2}{Q}.$$

Hence μ is well defined and $[\mu]_1=[\mu]_2$. Then Lemma 1 implies that

(60)

$$-\log\int_{S\times S}\theta^{1/2}(x)\,\theta^{1/2}(y)\,\alpha(x,dy)\,\widetilde{\pi}(dx)=R(\mu\,\|\widetilde{\pi}\otimes\alpha)-\int_S\log\theta(x)\,[\mu]_1(dx)\,.$$

If $R(\mu\,\|\widetilde{\pi}\otimes\alpha)=\infty$ or $-\int_S\log\theta(x)\,[\mu]_1(dx)=\infty$, the last display implies

$$\int_{S\times S}\theta^{1/2}(x)\,\theta^{1/2}(y)\,q(x)\,\alpha(x,dy)\,\pi(dx)=0.$$

By letting $A\doteq 0$ and $\xi(dx,du)\doteq q(x)\,\eta(dx)\,\delta_0(du)$, then ξ and μ satisfy (47) and

$$AR(\mu\,\|[\mu]_1\otimes\alpha)+\int_{S\times\mathbb{R}_+}\ell(u)\,\xi(dx,du)=\int_S q(x)\,\eta(dx)=I(\eta)\,.$$

Next assume $R(\mu\,\|\widetilde{\pi}\otimes\alpha)<\infty$ and $-\int_S\log\theta(x)\,[\mu]_1(dx)<\infty$. Define A by

$$A\doteq\exp\left\{-\left[R(\mu\,\|\widetilde{\pi}\otimes\alpha)-\int_S\log\theta(x)\,[\mu]_1(dx)-\log Q\right]\right\}. \tag{61}$$

Define measure

$$\rho(dx)\doteq q(x)\,\theta(x)\,\pi(dx)\,, \tag{62}$$

and

$$\kappa\doteq d\,[\mu]_1\,/d\rho. \tag{63}$$

Then for any $x\in S\backslash\Theta$ (recall $\Theta=\{x\in S:\theta(x)=0\}$)

$$\kappa(x)=\frac{d\,[\mu]_1}{d\rho}(x)=\frac{1}{Q\theta(x)}\frac{d\,[\mu]_1}{d\widetilde{\pi}}(x)\,.$$

In addition

$$\begin{aligned}\int_S\kappa(x)\log\kappa(x)\,\rho(dx)&=\int_S\log\kappa(x)\,[\mu]_1(dx)\\&=R([\mu]_1\,\|\widetilde{\pi})-\int_S\log\theta(x)\,[\mu]_1(dx)-\log Q.\end{aligned} \tag{64}$$

Define

$$b(x) \doteq \begin{cases} 0 & \text{for } x \in \Theta \\ A\kappa(x) & \text{for } x \notin \Theta \end{cases} \tag{65}$$

and

$$\xi(dx, du) \doteq q(x)\,\eta(dx)\,\delta_{b(x)}(du). \tag{66}$$

Then ξ satisfies the second part of (47). To see that the first part of (47) is satisfied, note that

$$[\mu]_1(\Theta) = 0 = \int_{\Theta\times\mathbb{R}_+} u\xi(dx, du),$$

and for $x \in S\backslash\Theta$,

$$\begin{aligned} \int_{\mathbb{R}_+} u\xi(dx, du) &= b(x)\,q(x)\,\eta(dx) \\ &= A\kappa(x)\,q(x)\,\theta(x)\,\pi(dx) \\ &= A[\mu]_1(dx). \end{aligned}$$

By using the definitions we arrive at the following, each line of which is explained below:

$$\begin{aligned} & AR(\mu \,\|[\mu]_1 \otimes \alpha) + \int_S \ell(u)\,\xi(dx, du) \\ &= AR(\mu \,\|[\mu]_1 \otimes \alpha) + \int_S \ell(b(x))\,q(x)\,\eta(dx) \\ &= AR(\mu \,\|[\mu]_1 \otimes \alpha) + \int_\Theta q(x)\,\eta(dx) + \int_{S\backslash\Theta} \ell(b(x))\,\rho(dx) \\ &= AR(\mu \,\|[\mu]_1 \otimes \alpha) + \int_S q(x)\,\eta(dx) + A\log A - A + A\int_S \kappa(x)\log\kappa(x)\,\rho(dx) \\ &= \int_S q(x)\,\eta(dx) + A\log A - A + A\left(R(\mu \,\|\widetilde{\pi} \otimes \alpha) - \int_S \log\theta(x)\,[\mu]_1(dx) - \log Q\right) \\ &= \int_S q(x)\,\eta(dx) - A. \end{aligned}$$

The first equality uses (66) and the second uses (65). The third uses (65) again, expands ℓ, and uses $\kappa \doteq d[\mu]_1/d\rho$ and $\eta(\Theta) = \rho(\Theta) = 0$. Equality four then uses

(64) and the fifth follows from (61). Note that (60) and (18) imply

$$A=\int_{S\times S}\theta^{1/2}\left(x\right)\theta^{1/2}\left(y\right)q\left(x\right)\alpha\left(x,dy\right)\pi\left(dx\right). \tag{67}$$

Hence we obtain

$$AR\left(\mu\left\|\left[\mu\right]_1\otimes\alpha\right.\right)+\int_S\ell\left(u\right)\xi\left(dx,du\right)=I\left(\eta\right).$$

□

The representation formula (28), the lower bound (46) and Lemma 6 together give

$$\liminf_{T\to\infty}-\frac{1}{T}\log E\left[\exp\left\{-TF(\eta_T)-T\cdot\infty\cdot\left(1_{\{r_T/T\}^c}\left(R_T/T\right)\right)\right\}\right]\geq\inf_{\eta\in\mathcal{P}(S)}\left[F\left(\eta\right)+I\left(\eta\right)\right].$$

2.3. Combining the cases. In the last section, we showed that for any sequence $\{r_T\}$ such that $r_T/T\to A\in[0,C]$

$$\liminf_{T\to\infty}-\frac{1}{T}\log E\left[\exp\left\{-TF(\eta_T)-T\cdot\infty\cdot\left(1_{\{r_T/T\}^c}\left(R_T/T\right)\right)\right\}\right]\geq\inf_{\eta\in\mathcal{P}(S)}\left[F\left(\eta\right)+I\left(\eta\right)\right].$$

An argument by contradiction shows that the bound is uniform in A. Thus

$$\begin{aligned}&\liminf_{T\to\infty}-\frac{1}{T}\log\left\{\sum_{r_T=1}^{\lfloor TC\rfloor}E\left[\exp\left\{-TF(\eta_T)-T\cdot\infty\cdot\left(1_{\{r_T/T\}^c}\left(R_T/T\right)\right)\right\}\right]\right\}\\&\geq\liminf_{T\to\infty}-\frac{1}{T}\log\left\{TC\cdot\bigvee_{r_T=1}^{\lfloor TC\rfloor}E\left[\exp\left\{-TF(\eta_T)-T\cdot\infty\cdot\left(1_{\{r_T/T\}^c}\left(R_T/T\right)\right)\right\}\right]\right\}\\&\geq\inf_{\eta\in\mathcal{P}(S)}\left[F\left(\eta\right)+I\left(\eta\right)\right].\end{aligned}$$

We now partition $E\left[\exp\left\{-TF(\eta_T)\right\}\right]$ according to this two cases to obtain the overall lower bound

$$\begin{aligned}&\liminf_{T\to\infty}-\frac{1}{T}\log E\left[\exp\left\{-TF(\eta_T)\right\}\right]\\&\qquad\geq\min\left\{\inf_{\eta\in\mathcal{P}(S)}\left[F\left(\eta\right)+I\left(\eta\right)\right],\left[C+C\log C-K_2-C\log K_2\right]\right\}\end{aligned}$$

Letting $C\to\infty$ we have the desired Laplace upper bound

$$\liminf_{T\to\infty}-\frac{1}{T}\log E\left[\exp\left\{-TF(\eta_T)\right\}\right]\geq\inf_{\eta\in\mathcal{P}(S)}\left[F\left(\eta\right)+I\left(\eta\right)\right]. \tag{68}$$

Remark 6. *We now explain why it suffices to consider function F that is nonnegative, lower semicontinuous and convex. Although the argument is well known in the literature, we include it for completeness.* [**7**, Theorem 1.2.3] *asserts that the Laplace upper bound implies the large deviation upper bound. Thus the large deviation upper bound holds if (68) is true for all such F. For any closed convex set $D \subset \mathcal{P}(S)$, define F by*

$$F(\eta) \doteq \begin{cases} 0 & \text{if } \eta \in D \\ \infty & \text{if } \eta \in D^c \end{cases}.$$

Then

$$\begin{aligned}\limsup_{T\to\infty} \frac{1}{T} \log P(\eta_T \in D) &= \limsup_{T\to\infty} \frac{1}{T} \log E\left[\exp\{-TF(\eta_T)\}\right] \\ &\leq - \inf_{\eta\in\mathcal{P}(S)} [F(\eta) + I(\eta)] \\ &= - \inf_{\eta\in D} I(\eta).\end{aligned}$$

Hence the large deviation upper bound holds for any convex set. Now consider any closed set D. Pick any $\varepsilon > 0$. Given any $\nu \in D$, the lower semicontinuity of I implies there exists $\delta_v > 0$ such that

$$\inf_{\eta\in\overline{N_{\delta_v}(\nu)}} I(\eta) \geq I(\nu) - \varepsilon$$

where $N_{\delta_v}(\nu)$ is the δ_v open neighborhood of ν and $\overline{N_{\delta_v}(\nu)}$ is the closure of $N_{\delta_v}(\nu)$. Since $\mathcal{P}(S)$ is compact and D is closed, D is also compact. Thus there is a finite set $\{\nu_1, \ldots, \nu_J\} \subset D$ such that

$$D \subset \cup_{j=1}^{J} N_{\delta_{v_j}}(\nu_j).$$

The large deviation upper bound for convex closed sets then implies

$$\begin{aligned}
\limsup_{T\to\infty}\frac{1}{T}\log P\left(\eta_T\in D\right) &\leq \limsup_{T\to\infty}\frac{1}{T}\log P\left(\eta_T\in\cup_{j=1}^{J}\overline{N_{\delta_{v_j}}\left(\nu_j\right)}\right)\\
&\leq \max_{j=1,\ldots.J}\left\{\limsup_{T\to\infty}\frac{1}{T}\log P\left(\eta_T\in\overline{N_{\delta_{v_j}}\left(\nu_j\right)}\right)\right\}\\
&\leq \max_{j=1,\ldots.J}\left\{-I\left(\nu_j\right)+\varepsilon\right\}\\
&\leq -\inf_{j=1,\ldots,J} I\left(\nu_j\right)+\varepsilon\\
&\leq -\inf_{\eta\in D} I\left(\eta\right)+\varepsilon,
\end{aligned}$$

and since ε is arbitrary, the large deviation upper bound holds for D.

3. Laplace lower bound

We turn to the proof of the reverse inequality

$$\limsup_{T\to\infty}-\frac{1}{T}\log E\left[\exp\left\{-TF(\eta_T)\right\}\right]\leq\inf_{\eta\in\mathcal{P}(S)}\left[F\left(\eta\right)+I\left(\eta\right)\right]. \tag{69}$$

Let F be a nonnegative bounded and continuous function. Fix an arbitrary $\varepsilon>0$ and choose η such that

$$F\left(\eta\right)+I\left(\eta\right)\leq\inf_{\nu\in\mathcal{P}(S)}\left[F\left(\nu\right)+I\left(\nu\right)\right]+\varepsilon. \tag{70}$$

As pointed out in Remark 3, H defined in (21) is dense in $\mathcal{P}\left(S\right)$. Since I was extended from H to $\mathcal{P}\left(S\right)$ via lower semicontinuous regularization, we can assume without loss of generality that $\eta\ll\pi$. Define $\theta\doteq d\eta/d\pi$. We now argue we can further assume there exists $\delta>0$ such that

$$\delta\leq\theta\left(x\right)\leq\frac{1}{\delta} \tag{71}$$

for all $x\in S$. If $\eta^{\delta}\doteq\left(1-\delta\right)\eta+\delta\pi$ then $d\eta^{\delta}/d\pi\geq\delta$, and the continuity of F and the convexity of I imply that the difference between $F\left(\eta^{\delta}\right)+I\left(\eta^{\delta}\right)$ and $F\left(\eta\right)+I\left(\eta\right)$ can be made arbitrarily small.

Thus we can assume θ is uniformly bounded from below away from zero. Let $n \in \mathbb{N}$, and define

$$\eta^n(dx) \doteq \theta(x) 1_{\{\theta(x)\le n\}}\pi(dx) + \frac{\eta(\{x : \theta(x) > n\})}{\pi(\{x : \theta(x) > n\})} 1_{\{\theta(x)>n\}}\pi(dx).$$

Then $d\eta^n/d\pi \le \eta(\{x : \theta(x) > n\})/\pi(\{x : \theta(x) > n\})$, and since $\eta \ll \pi$ implies $\pi(\{x : \theta(x) > n\}) \to 0$, η^n converges weakly to η. It then follows from the continuity of F and lower semicontinuity of I that we can choose η satisfying (70) with 2ε replacing ε and also (71). Hence we assume η satisfies (70) and (71). Furthermore by Lusin's Theorem [**12**, Theorem 7.10], we can also assume that θ is continuous.

The proof of the lower bound will use the following representation. The infimum in the representation is taken over all control measures $\{\bar{\alpha}_i, \bar{\sigma}_i\}$, and the properties of such measures and how $\bar{\eta}_T$ and $\bar{R}_T$ are constructed from them were discussed immediately above the similar representation (28). The proof of the lemma is given in the appendix.

Lemma 7. *Let $F : \mathcal{P}(S) \to \mathbb{R}$ be bounded and continuous. Then*

$$-\frac{1}{T}\log E\left[\exp\{-TF(\eta_T)\}\right] = \inf E\left[F(\bar{\eta}_T) + \frac{1}{T}\sum_{i=1}^{\bar{R}_T}\left(R(\bar{\alpha}_{i-1}\,\|\alpha) + R(\bar{\sigma}_i\,\|\sigma)\right)\right]$$

where the infimum is taken over all control measures $\{\bar{\alpha}_i, \bar{\sigma}_i\}$.

Suppose that given any measure $\eta \in \mathcal{P}(S)$ satisfying (70) and (71), one can construct $\bar{\alpha}_i$ and $\bar{\sigma}_i$ such that given any subsequence of T, there is a further subsequence T_n such that

$$\lim_{T_n\to\infty} E\left[F(\bar{\eta}_{T_n}) + \frac{1}{T_n}\sum_{i=1}^{\bar{R}_{T_n}}\left(R(\bar{\alpha}_{i-1}\,\|\alpha) + R(\bar{\sigma}_i\,\|\sigma)\right)\right] = F(\eta) + I(\eta).$$

Then Lemma 7 implies the Laplace lower bound (115). The construction of suitable $\bar{\alpha}_i$ and $\bar{\sigma}_i$ turns on many of the same constructions as those used in the proof of the second part of Lemma 6. We first define $\mu \in \mathcal{P}(S \times S)$ as in (59). Then automatically $[\mu]_1 = [\mu]_2$, and hence if we define p as the regular conditional probability such that $\mu = [\mu]_1 \otimes p$, then $[\mu]_1$ is invariant under p [**7**, Lemma 8.5.1 (a)]. Define $\bar{\alpha}_i \doteq p$ for each

i, and let $\{\bar{X}_i\}$ be the corresponding Markov chain. Next define $\rho(dx) \doteq q(x)\eta(dx)$ and

$$\kappa(x) \doteq \frac{d[\mu]_1}{d\rho}(x) = \frac{1}{Q\theta(x)} \frac{d[\mu]_1}{d\widetilde{\pi}}(x). \tag{72}$$

By (71), there is $M < \infty$ such that $1/M \le \kappa \le M$, and due to the continuity of θ, κ is also continuous. Notice that

$$\eta(dx) = (q(x)\kappa(x))^{-1}[\mu]_1(dx). \tag{73}$$

Assumption (71) guarantees that

$$-\log \int_{S\times S} \theta^{1/2}(x)\theta^{1/2}(y)(\widetilde{\pi}\otimes\alpha)(dx,dy) < \infty \text{ and } -\int_S \log\theta(x)[\mu]_1(dx) < \infty,$$

(60) then implies that $R(\mu\|\widetilde{\pi}\otimes\alpha) < \infty$. Define A as in (67). Let $\bar{\sigma}_i$ be the exponential distribution with mean $[A\kappa(\bar{X}_{i-1})]^{-1}$ for each i. Thus we can construct a Markov jump process $\bar{X}(t)$ using $\bar{\alpha}_i$ and $\bar{\sigma}_i$ instead of α and σ, and the infinitesimal $\bar{\mathcal{L}}$ generator will be bounded and continuous and takes the form:

$$\bar{\mathcal{L}}f(x) = A\kappa(x)q(x)\int_S [f(y)-f(x)]p(x,dy).$$

(73) and the fact that $[\mu]_1$ is invariant under p imply $\int_S (\bar{\mathcal{L}}f(x))\eta(dx) = 0$, and by Echeverria's Theorem [**11**, Theorem 4.9.17], η is an invariant distribution of the continuous time process $\bar{X}$. We claim that η is the unique invariant distribution of $\bar{X}$. Indeed, by [**11**, Proposition 4.9.2] any invariant distribution ν for $\bar{X}$ satisfies $\int_S (\bar{\mathcal{L}}f(x))\nu(dx) = 0$. If we define

$$\widetilde{\nu}(dx) \doteq \frac{A\kappa(x)q(x)\nu(dx)}{\int_S A\kappa(x)q(x)\nu(dx)},$$

then $\widetilde{\nu}$ is invariant under p. However, the transitivity condition of α and [**7**, Lemma 8.6.3(c)] the invariant measure under p is unique, and hence the invariant measure of $\bar{X}$ is also unique. By the definition of $\bar{\eta}_T$ in (27),

$$\begin{aligned}\bar{\eta}_T(\cdot) &= \frac{1}{T}\int_0^T \delta_{\bar{X}(t)}(\cdot)\,dt \\ &= \frac{1}{T}\left[\sum_{i=1}^{\bar{R}_T-1} \delta_{\bar{X}_{i-1}}(dx)\frac{\bar{\tau}_i}{q(\bar{X}_{i-1})} + \delta_{\bar{X}_{\bar{R}_T-1}}(dx)\left(T - \sum_{i=1}^{\bar{R}_T-1}\frac{\bar{\tau}_i}{q(\bar{X}_{i-1})}\right)\right]. \end{aligned} \tag{74}$$

Since S is compact we can extract a subsequence of T such that $\bar{\eta}_T$ converges weakly, and by [**11**, Theorem 4.9.3] this weak limit is η. We claim the following along the same subsequence.

LEMMA 8. *$E\left[\bar{R}_T/T\right] \to A$, $E\left[\sum_{i=1}^{\bar{R}_T} \delta_{\bar{X}_{i-1}}(dx)/T\right] \to A[\mu]_1(dx)$ and $E\left[\sum_{i=1}^{\bar{R}_T} R(\bar{\sigma}_i \| \sigma)/T\right]$ $\to \int_S \ell(A\kappa(x))\, q(x)\, \eta(dx)$.*

PROOF. Since $\bar{\eta}_T \to \eta$ weakly, we have for any bounded and continuous function f on the space of subprobability measures on S, $\lim_{T\to\infty} E[f(\bar{\eta}_T)] = f(\eta)$. To prove the first half of the lemma, we define f by

$$f(\nu) \doteq \int_S \kappa(x)\, q(x)\, \nu(dx).$$

Since both κ and q are bounded and continuous, f is also bounded and continuous. Using (73)

$$f(\eta) = \int_S \kappa(x)\, q(x)\, \eta(dx) = \int_S [\mu]_1(dx) = 1. \tag{75}$$

Thus $\lim_{T\to\infty} E[f(\bar{\eta}_T)] = 1$. Now by (74) and the definition of $\bar{R}_T$

$$\begin{aligned}
&E\left[\left| f(\bar{\eta}_T) - f\left(\frac{1}{T}\sum_{i=1}^{\bar{R}_T} \delta_{\bar{X}_{i-1}}(dx)\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)}\right)\right|\right] \\
&= \frac{1}{T}E\left[\kappa\left(\bar{X}_{\bar{R}_T-1}\right) q\left(\bar{X}_{\bar{R}_T-1}\right)\left(\sum_{i=1}^{\bar{R}_T}\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)} - T\right)\right] \\
&\leq \frac{1}{T}E\left[\kappa\left(\bar{X}_{\bar{R}_T-1}\right) q\left(\bar{X}_{\bar{R}_T-1}\right)\frac{\bar{\tau}_{\bar{R}_T}}{q\left(\bar{X}_{\bar{R}_T-1}\right)}\right] \\
&\leq \frac{M}{T}E\left[\bar{\tau}_{\bar{R}_T}\right] \\
&\leq \frac{AM^2}{T} \to 0
\end{aligned}$$

as $T \to \infty$. Hence

$$\lim_{T\to\infty} E\left[f\left(\frac{1}{T}\sum_{i=1}^{\bar{R}_T}\delta_{\bar{X}_{i-1}}(dx)\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)}\right)\right] = 1.$$

Recall that $\mathcal{F}_i$ is the σ-algebra generated by $\left\{\left(\bar{X}_0,\ldots,\bar{X}_i\right),\left(\bar{\tau}_1,\ldots,\bar{\tau}_i\right)\right\}$. Then

$$
\begin{aligned}
&E\left[f\left(\frac{1}{T}\sum_{i=1}^{\bar{R}_T}\delta_{\bar{X}_{i-1}}(dx)\,\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)}\right)\right]\\
&=\frac{1}{T}E\left[\sum_{i=1}^{\bar{R}_T}\kappa\left(\bar{X}_{i-1}\right)q\left(\bar{X}_{i-1}\right)\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)}\right]\\
&=\frac{1}{T}E\left[\sum_{i=1}^{\infty}\kappa\left(\bar{X}_{i-1}\right)\bar{\tau}_i\mathbf{1}\left(\sum_{j=1}^{i-1}\frac{\bar{\tau}_j}{q\left(\bar{X}_{j-1}\right)}\leq T\right)\right]\\
&=\frac{1}{T}\sum_{i=1}^{\infty}E\left[E\left[\kappa\left(\bar{X}_{i-1}\right)\bar{\tau}_i\mathbf{1}\left(\sum_{j=1}^{i-1}\frac{\bar{\tau}_j}{q\left(\bar{X}_{j-1}\right)}\leq T\right)\middle|\mathcal{F}_{i-1}\right]\right]\\
&=\frac{1}{T}\sum_{i=1}^{\infty}E\left[\kappa\left(\bar{X}_{i-1}\right)\mathbf{1}\left(\sum_{j=1}^{i-1}\frac{\bar{\tau}_j}{q\left(\bar{X}_{j-1}\right)}\leq T\right)E\left[\bar{\tau}_i|\mathcal{F}_{i-1}\right]\right]\\
&=\frac{1}{T}\sum_{i=1}^{\infty}E\left[\kappa\left(\bar{X}_{i-1}\right)\mathbf{1}\left(\sum_{j=1}^{i-1}\frac{\bar{\tau}_j}{q\left(\bar{X}_{j-1}\right)}\leq T\right)\frac{1}{A\kappa\left(\bar{X}_{i-1}\right)}\right]\\
&=\frac{1}{AT}E\left[\sum_{i=1}^{\infty}\mathbf{1}\left(\sum_{j=1}^{i-1}\frac{\bar{\tau}_j}{q\left(\bar{X}_{j-1}\right)}\leq T\right)\right]\\
&=\frac{1}{A}E\left[\frac{\bar{R}_T}{T}\right].
\end{aligned}
$$

This completes the proof of the first statement in the lemma.

For the second statement in the lemma, using (73) and the fact that $\bar{\eta}_T$ converges weakly to η, we have $q(x)\,\kappa(x)\,\bar{\eta}_T(dx)$ converges weakly to $[\mu]_1(dx)$. Recall the definition of $\bar{\eta}_T$ in (74), we can bound the total variational difference between the following two measures:

$$
\begin{aligned}
&\left\| E\left[q(x)\kappa(x)\bar{\eta}_T(dx)\right] - E\left[q(x)\kappa(x)\left(\frac{1}{T}\sum_{i=1}^{\bar{R}_T}\delta_{\bar{X}_{i-1}}(dx)\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)}\right)\right]\right\|_{\mathrm{TV}} \\
&= \left\| E\left[q(x)\kappa(x)\bar{\eta}_T(dx) - q(x)\kappa(x)\left(\frac{1}{T}\sum_{i=1}^{\bar{R}_T}\delta_{\bar{X}_{i-1}}(dx)\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)}\right)\right]\right\|_{\mathrm{TV}} \\
&= \frac{1}{T}\left\| E\left[q\left(\bar{X}_{\bar{R}_T-1}\right)\kappa\left(\bar{X}_{\bar{R}_T-1}\right)\delta_{\bar{R}_T-1}(dx)\left(\sum_{i=1}^{\bar{R}_T}\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)} - T\right)\right]\right\|_{\mathrm{TV}} \\
&\leq \frac{1}{T}\left\| E\left[q\left(\bar{X}_{\bar{R}_T-1}\right)\kappa\left(\bar{X}_{\bar{R}_T-1}\right)\delta_{\bar{R}_T-1}(dx)\frac{\bar{\tau}_{\bar{R}_T}}{q\left(\bar{X}_{\bar{R}_T-1}\right)}\right]\right\|_{\mathrm{TV}} \\
&\leq \frac{M}{T}E\left[\bar{\tau}_{\bar{R}_T}\right] \\
&\leq \frac{AM^2}{T} \to 0.
\end{aligned}
$$

Hence

$$
E\left[q(x)\kappa(x)\left(\frac{1}{T}\sum_{i=1}^{\bar{R}_T}\delta_{\bar{X}_{i-1}}(dx)\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)}\right)\right] \text{ converges weakly to } [\mu]_1(dx).
$$

Next, we point out

$$
E\left[\frac{1}{T}\sum_{i=1}^{\bar{R}_T}\delta_{\bar{X}_{i-1}}(dx)\right] = A\cdot E\left[q(x)\kappa(x)\left(\frac{1}{T}\sum_{i=1}^{\bar{R}_T}\delta_{\bar{X}_{i-1}}(dx)\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)}\right)\right], \tag{76}
$$

which implies the second statement, i.e. $E\left[\sum_{i=1}^{\bar{R}_T}\delta_{\bar{X}_{i-1}}(dx)/T\right]$ converges weakly to $A[\mu]_1(dx)$. We start from the right hand side of (76),

$$
\begin{aligned}
& A \cdot E\left[q(x)\kappa(x)\left(\frac{1}{T}\sum_{i=1}^{\bar{R}_T}\delta_{\bar{X}_{i-1}}(dx)\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)}\right)\right] \\
= \; & \frac{A}{T}E\left[q(x)\kappa(x)\left(\sum_{i=1}^{\infty}\delta_{\bar{X}_{i-1}}(dx)\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)}\right)\mathbf{1}\left(\sum_{j=1}^{i-1}\frac{\bar{\tau}_j}{q\left(\bar{X}_{j-1}\right)}\leq T\right)\right] \\
= \; & \frac{A}{T}\sum_{i=1}^{\infty}E\left[q\left(\bar{X}_{i-1}\right)\kappa\left(\bar{X}_{i-1}\right)\delta_{\bar{X}_{i-1}}(dx)\mathbf{1}\left(\sum_{j=1}^{i-1}\frac{\bar{\tau}_j}{q\left(\bar{X}_{j-1}\right)}\leq T\right)\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)}\right] \\
= \; & \frac{A}{T}\sum_{i=1}^{\infty}E\left[\kappa\left(\bar{X}_{i-1}\right)\delta_{\bar{X}_{i-1}}(dx)\mathbf{1}\left(\sum_{j=1}^{i-1}\frac{\bar{\tau}_j}{q\left(\bar{X}_{j-1}\right)}\leq T\right)\bar{\tau}_i\right] \\
= \; & \frac{A}{T}\sum_{i=1}^{\infty}E\left[\kappa\left(\bar{X}_{i-1}\right)\delta_{\bar{X}_{i-1}}(dx)\mathbf{1}\left(\sum_{j=1}^{i-1}\frac{\bar{\tau}_j}{q\left(\bar{X}_{j-1}\right)}\leq T\right)E\left[\bar{\tau}_i|\mathcal{F}_{i-1}\right]\right] \\
= \; & \frac{A}{T}\sum_{i=1}^{\infty}E\left[\kappa\left(\bar{X}_{i-1}\right)\delta_{\bar{X}_{i-1}}(dx)\mathbf{1}\left(\sum_{j=1}^{i-1}\frac{\bar{\tau}_j}{q\left(\bar{X}_{j-1}\right)}\leq T\right)\frac{1}{A\kappa\left(\bar{X}_{i-1}\right)}\right] \\
= \; & \frac{1}{T}\sum_{i=1}^{\infty}E\left[\delta_{\bar{X}_{i-1}}(dx)\mathbf{1}\left(\sum_{j=1}^{i-1}\frac{\bar{\tau}_j}{q\left(\bar{X}_{j-1}\right)}\leq T\right)\right] \\
= \; & \frac{1}{T}E\left[\sum_{i=1}^{\bar{R}_T}\delta_{\bar{X}_{i-1}}(dx)\right].
\end{aligned}
$$

The proof of the third statement is similar. Define f by

$$
f(\nu) \doteq \int_S \ell(A\kappa(x))\,q(x)\,\nu(dx).
$$

Then as before,

$$
f(\eta) = \lim_{T\to\infty}E\left[f(\eta_T)\right] = \lim_{T\to\infty}E\left[f\left(\frac{1}{T}\sum_{i=1}^{\bar{R}_T}\delta_{\bar{X}_{i-1}}(dx)\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)}\right)\right].
$$

Using $g(x) = x\ell(1/x)$ and Lemma 4, we have

$$
\begin{aligned}
&E\left[f\left(\frac{1}{T}\sum_{i=1}^{\bar{R}_T}\delta_{\bar{X}_{i-1}}(dx)\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)}\right)\right] \\
&= E\left[\frac{1}{T}\sum_{i=1}^{\bar{R}_T}\ell\left(A\kappa\left(\bar{X}_{i-1}\right)\right)\bar{\tau}_i\right] \\
&= \frac{1}{T}\sum_{i=1}^{\infty}E\left[\ell\left(A\kappa\left(\bar{X}_{i-1}\right)\right)\mathbf{1}\left(\sum_{j=1}^{i-1}\frac{\bar{\tau}_j}{q\left(\bar{X}_{j-1}\right)}\leq T\right)E\left[\bar{\tau}_i|\mathcal{F}_{i-1}\right]\right] \\
&= E\left[\frac{1}{T}\sum_{i=1}^{\bar{R}_T}\frac{1}{A\kappa\left(\bar{X}_{i-1}\right)}\ell\left(A\kappa\left(\bar{X}_{i-1}\right)\right)\right] \\
&= E\left[\frac{1}{T}\sum_{i=1}^{\bar{R}_T}g\left(\frac{1}{A\kappa\left(\bar{X}_{i-1}\right)}\right)\right] \\
&= E\left[\frac{1}{T}\sum_{i=1}^{\bar{R}_T}R\left(\bar{\sigma}_i\left\|\sigma\right.\right)\right],
\end{aligned}
$$

and the third part of the lemma follows. $\square$

Now the Laplace lower bound is straightforward. The definition of μ in (59), the continuity of θ, and the bound (71) imply $x \to R\left(p(x,\cdot)\left\|\alpha(x,\cdot)\right.\right)$ is bounded and continuous. By Lemma 8 and the chain rule for relative entropy,

$$
\begin{aligned}
&\lim_{T\to\infty}E\left[F\left(\bar{\eta}_T\right)+\frac{1}{T}\sum_{i=1}^{\bar{R}_T}\left(R\left(\bar{\alpha}_{i-1}\left\|\alpha\right.\right)+R\left(\bar{\sigma}_i\left\|\sigma\right.\right)\right)\right] \\
&= \lim_{T\to\infty}E\left[F\left(\bar{\eta}_T\right)\right]+\lim_{T\to\infty}\int_S R\left(p(x,\cdot)\left\|\alpha(x,\cdot)\right.\right)E\left[\frac{1}{T}\sum_{i=1}^{\bar{R}_T}\delta_{\bar{X}_{i-1}}(dx)\right] \\
&+\lim_{T\to\infty}E\left[\frac{1}{T}\sum_{i=1}^{\bar{R}_T}R\left(\bar{\sigma}_i\left\|\sigma\right.\right)\right] \\
&= F(\eta)+AR\left(\mu\left\|[\mu]_1\otimes\alpha\right.\right)+\int_S\ell\left(A\kappa(x)\right)q(x)\,\eta(dx)\,.
\end{aligned}
$$

Returning to the proof of the second part of Lemma 6, we find that with this choice of A, μ and κ, the rate function $I(\eta)$ coincides with $AR\left(\mu\left\|[\mu]_1\otimes\alpha\right.\right)+$

$\int_S \ell\left(A\kappa\left(x\right)\right) q\left(x\right) \eta\left(dx\right)$ (note that this η corresponds to a special of Lemma 6 where $\Theta \doteq \{x \in S : \theta\left(x\right) = 0\}$ is empty). This completes the proof of the Laplace lower bound.

4. On the boundedness of the rate function

What we have done so far in this chapter and the last one is to point out the explicit formula for the large deviations rate function of the empirical measures for the Markov jump processes and provide a proof of the large deviations principle. The reason we do this in such detail is that this result is not covered by the Donsker-Varadhan large deviations theory. If we look at conditions in the Donsker-Varadhan large deviations theory for empirical measures carefully, we note that the underlying Markov process must possess a "diffusive" dynamics term in order to move the state in a continuous fashion, see (11). However, for Markov jump processes like the one we gave in Chapter 2 Section 2, this is not the case. In such models, the process only moves when a "jump" happens, there is no continuous accumulation of movements.

Let us consider a Markov jump process satisfying all the conditions in Chapter 3 Section 3 that has π as its invariant distribution. In order to hit a different probability measure $\eta \in \mathcal{P}\left(S\right)$, we need to perturb the original dynamics, which includes the distortion of the Markov chain transition probability α and the distortion of the exponential holding time σ. Each of these distortions must pay a relative entropy cost, and the minimum of the sum of these costs asymptotically approximates the rate function $I\left(\eta\right)$. As we have noticed, the rate function I is always bounded, which is quite different from Markov processes with a "diffusive" term. Especially when η is not absolutely continuous with respect to π, the rate function for a "diffusive" Markov process is infinite. In contrast, for a Markov jump process, when η is not absolutely continuous with respect to π, the relative entropy cost from distortion of α can be made arbitrarily small, and the rate function is almost the sole result of the distortion of σ. We will make this point explicit via the following example.

Recall the model we stated in Chapter 3 Section 2, where the state space S is $[0,1]$, the jump intensity is $q \equiv 1$ and for each $x \in [0,1]$, $\alpha(x,\cdot)$ is the uniform distribution on $[0,1]$. The invariant distribution π is clearly the uniform distribution on $[0,1]$. Now consider a Dirac measure $\eta \doteq \delta_{1/2}$ as our target measure. η is not absolutely continuous with respect to π. However, we can approximate η weakly via a sequence of probability measures that are absolutely continuous with respect to π. Define for each $n \in \mathbb{N}$, a probability measure η^n by its Radon-Nikodym derivative θ^n with respect to π as

$$\theta^n(x) \doteq \begin{cases} n-1 & \text{for } x \in \left(\frac{1}{2}-\frac{1}{2n}, \frac{1}{2}+\frac{1}{2n}\right) \\ \frac{1}{n-1} & \text{otherwise} \end{cases}.$$

Using the formula for rate function in (43), we have

$$I(\eta^n) = 1 - \left(\int_0^1 (\theta^n(x))^{1/2}\,dx\right)\left(\int_0^1 (\theta^n(y))^{1/2}\,dy\right) = 1 - \frac{4(n-1)}{n^2}.$$

According to the definition of rate function in Chapter 2 Section 4, the rate function is bounded above by 1. However $I(\eta^n) \to 1$ as $n \to \infty$, and one can check that this is true for any sequence corresponding to absolutely continuous measures converging weakly to η. Indeed, let η^n be a sequence of measures that are absolutely continuous with respect to π and weakly converge to η. We prove that $I(\eta^n) \to 1$. Using formula (22), it suffice to show

$$\int_0^1 (\theta^n(x))^{1/2}\,dx \to 0.$$

Define $B_\delta(1/2)$ as $(1/2-\delta, 1/2+\delta)$, then since η^n converges weakly to η, we know that

$$\int_0^1 \theta^n(x)\,\mathbf{1}(x \notin B_\delta(1/2))\,dx \to 0$$

as $n \to \infty$. One can decompose $\int_0^1 (\theta^n(x))^{1/2}\,dx$ into two parts:

$$\begin{aligned}
&\int_0^1 (\theta^n(x))^{1/2}\,dx \\
&= \int_0^1 (\theta^n(x))^{1/2}\,\mathbf{1}(x \in B_\delta(1/2))\,dx \\
&+ \int_0^1 (\theta^n(x))^{1/2}\,\mathbf{1}(x \notin B_\delta(1/2))\,dx.
\end{aligned}$$

We use Cauchy-Schwartz on both terms,

$$\begin{aligned}
&\int_0^1 (\theta^n(x))^{1/2}\,\mathbf{1}(x\in B_\delta(1/2))\,dx \\
&\le \sqrt{\int_0^1 \theta^n(x)\,dx \int \mathbf{1}(x\in B_\delta(1/2))\,dx} \\
&\le \sqrt{2\delta},
\end{aligned}$$

and

$$\begin{aligned}
&\int_0^1 (\theta^n(x))^{1/2}\,\mathbf{1}(x\notin B_\delta(1/2))\,dx \\
&\le \sqrt{\int_0^1 \theta^n(x)\,\mathbf{1}(x\notin B_\delta(1/2))\,dx} \\
&\cdot \sqrt{\int_0^1 \mathbf{1}(x\notin B_\delta(1/2))\,dx} \\
&\le \sqrt{\int_0^1 \theta^n(x)\,\mathbf{1}(x\notin B_\delta(1/2))\,dx}
\end{aligned}$$

which converges to 0 as $n\to\infty$. By nature of the lower semicontinuous extension, $I(\eta)$ equals to $\underline{\lim} I(\eta^n)$ for at least one such sequence η^n, we conclude that $I(\eta)=1$.

We now consider a fixed $n\in\mathbb{N}$ and examine the perturbed dynamics that can hit measure η^n. This is most easily seen by examining the minimizer in the variational formula for the rate function, which we have constructed during the proof of the Laplace principle lower bound in Chapter 3 Section 3. Recall that $\bar{\sigma}_i(\cdot)$ and $\bar{\alpha}_i(\cdot)$ are perturbed dynamics for the exponential holding time and the Markov chain, $\bar{\sigma}_i(\cdot)$ depends on $\left\{\bar{X}_0,\bar{\tau}_1,\bar{X}_1,\bar{\tau}_2,\ldots,\bar{X}_{i-1}\right\}$ and $\bar{\alpha}_i(\cdot)$ depends on $\left\{\bar{X}_0,\bar{\tau}_1,\bar{X}_1,\bar{\tau}_2\ldots,\bar{X}_{i-1},\bar{\tau}_i\right\}$. $\bar{\tau}_i$ and $\bar{X}_i$ are chosen recursively according to stochastic kernels $\bar{\sigma}_i(\cdot)$ and $\bar{\alpha}_i(\cdot)$. Specifically, $\bar{s}_i$ is defined by (15) using $\bar{X}_i$ and $\bar{\tau}_i$; $\bar{R}_T$ is defined by (26) using $\bar{s}_i$; and $\bar{\eta}_T$ is defined by (27) using $\bar{X}_i$, $\bar{\tau}_i$ and $\bar{R}_T$. Following the regime in Chapter 3 Section 3, we first define $\mu\in\mathcal{P}(S\times S)$ as in (59). Thus μ is a product measure. As before, we use $[\mu]_1$ to denote the first marginal of μ and p to denote the regular conditional probability such that $\mu=[\mu]_1\otimes p$. Since μ is a product measure defined by (59), $[\mu]_1$

and p are the same measure and the density with respect to π can be calculated as

$$\text{(77)} \qquad \frac{d\left[\mu\right]_1}{d\pi}(x) = \begin{cases} \frac{n}{2} & \text{for } x \in \left(\frac{1}{2} - \frac{1}{2n}, \frac{1}{2} + \frac{1}{2n}\right) \\ \frac{n}{2(n-1)} & \text{otherwise} \end{cases} .$$

As in Chapter 3 Section 3 let $\bar{\alpha}_i \doteq p$ for each i. A direct calculation of A using formula (67) shows that $A = 4\left(n-1\right)/n^2$. Also, κ defined in (63) reduces to

$$\kappa\left(x\right) = \begin{cases} \frac{n}{2(n-1)} & \text{for } x \in \left(\frac{1}{2} - \frac{1}{2n}, \frac{1}{2} + \frac{1}{2n}\right) \\ \frac{n}{2} & \text{otherwise} \end{cases} .$$

As in Chapter 3 Section 3, $\bar{\sigma}_i$ should be the exponential distribution with mean $\left[A\kappa\left(\bar{X}_{i-1}\right)\right]^{-1}$. Hence if $\bar{X}_{i-1}$ falls into $(1/2 - 1/(2n), 1/2 + 1/(2n))$, $\bar{\sigma}_i$ would be the exponential distribution with mean $n/2$, otherwise $\bar{\sigma}_i$ would be the exponential distribution with mean $n/\left[2\left(n-1\right)\right]$. Now our perturbed Markov jump process, denoted by $\bar{X}\left(t\right)$, is constructed using $\bar{\alpha}_i$ and $\bar{\sigma}_i$ defined as above. As proved in Lemma 8, the expected value of the relative entropy cost

$$\frac{1}{T}\sum_{i=1}^{\bar{R}_T}\left(R\left(\bar{\alpha}_{i-1}\left\|\alpha\right.\right) + R\left(\bar{\sigma}_i\left\|\sigma\right.\right)\right)$$

would convergence to

$$I\left(\eta^n\right) = AR\left(\mu\left\|\left[\mu\right]_1\otimes\alpha\right.\right) + \int_0^1 \ell\left(A\kappa\left(x\right)\right)\eta^n\left(dx\right)$$

as $T \to \infty$. We have noticed that $p\left(x,dy\right) = \left[\mu\right]_1\left(dy\right)$ and $\alpha\left(x,dy\right) = \pi\left(dy\right)$, by using (77)

$$\begin{aligned} & AR\left(\mu\left\|\left[\mu\right]_1\otimes\alpha\right.\right) \\ = \; & A\int_0^1 R\left(p\left(x,\cdot\right)\left\|\alpha\left(x,\cdot\right)\right.\right)\left[\mu\right]_1\left(dx\right) \\ = \; & \frac{4\left(n-1\right)}{n^2}\left(\log n - \log 2 - \frac{\log\left(n-1\right)}{2}\right) \end{aligned}$$

which converges to 0 as $n \to \infty$. Hence the relative entropy cost that comes from the distortion of the transition function of the Markov chain converges to 0. For the second term, by taking A, η^n and κ in place we have

$$\int_0^1 \ell\left(A\kappa\left(x\right)\right)\eta^n\left(dx\right) = \frac{2\left(n-1\right)}{n^2}\left(\log\left(n-1\right) + 2\log 2 - \log n\right) - \frac{4\left(n-1\right)}{n^2} + 1$$

which converges to 1 as $n \to \infty$. Thus as $n \to \infty$, in other words, as η^n approaches the target distribution η, the relative entropy cost that comes from the distortion of Markov chain vanishes, and the rate function becomes solely determined by the relative entropy cost that comes from the distortion of exponential waiting times. We can replace the target measure η with any discrete measure. Following the same logic as above, one can see that the perturbed Markov process hits a target measure by slowing down the exponential clock for transition states near the points with positive mass. By arranging the relative exponential holding times among different atoms, one can hit any such measure with relative entropy cost that comes solely from the distortion of these exponential clocks.

CHAPTER 4

An application to Markov chain Monte Carlo

In this chapter, we discuss an application of the explicit formula of the large deviations rate function to MCMC.

As we have pointed out in the Introduction, there are numerous ways to characterize the efficiency of MCMC algorithms. In this thesis, we propose to use the large deviations principle as a performance measure for the various Monte Carlo schemes. Specifically, under the situation where one can obtain an explicit formula for the large deviations rate function of the empirical measures, the algorithm with a larger rate function would asymptotically have a faster rate of convergence. As we have seen in Chapter 2 and Chapter 3, an explicit formula for the large deviations rate function can often be obtained for reversible Markov processes. As it turned out, a rich class of MCMC schemes do satisfy the reversibility condition. We will focus our interest on one such algorithm: the parallel tempering algorithm, and its distributional limit: the (full) infinite swapping algorithm as well as the more implementable version: partial infinite swapping algorithm.

Parallel tempering mechanism [**9, 13, 20, 21**] was introduced to tackle a common issue appeared in MCMC simulation for a Gibbs distribution. A Gibbs distribution takes the form of $e^{-V(x)/\tau}dx$, where V is some potential function defined on $\mathbb{R}^d$ and τ is a variable that represents the scaled temperature. The challenge in simulating the Gibbs distribution lies in the fact that we only know the density function up to a normalizing constant. A Metropolis-Hastings type MCMC algorithm is perfectly suitable in this situation, since no prior knowledge of the normalizing constant is needed in the simulation. However, when the target distribution π has multiple isolated parts, which is the case when the potential function V has multiple local minima that are separated by high energy barriers, the MCMC sampler often appears to be “trapped”in the neighborhood of a local minimum, which results in the slow convergence of the Monte Carlo. This is especially the case when the temperature τ is small, which exaggerates the isolation among distinguished local parts of the Gibbs distribution. Parallel tempering tries to overcome this difficulty by conducting a series of Markov chain simulations in parallel. Each of these MCMC simulations are performed at a different temperature. As we pointed out, the high energy barriers are easier to cross for high temperature simulations, and this ease of movement is transferred down through the temperature ladder by “particle exchanges”to improve the simulations for lower temperatures.

We will make the parallel tempering idea clear in the coming section. Then we will examine the explicit formula for the large deviations rate function of the empirical measures, and discover a natural limit in distribution of the parallel tempering algorithm. This limit results in a new class of MCMC algorithms: the infinite swapping algorithm. We will then devote the next chapter into discussing the practical implementation issues.

1. Parallel tempering: diffusion model with two temperatures

Although the implementation of parallel tempering uses a discrete time model, the motivation for the infinite swapping limit is best illustrated in the setting where the Markov process is a continuous time diffusion process. It is in this case that the large deviation rate function, as well as the construction of a process that is distributionally equivalent to the infinite swapping limit, is simplest. In order to minimize the notational overhead, we discuss in detail the two temperatures case. The extension to models with multiple temperatures is obvious and will be stated in later sections.

Recall that V is the potential function defined on $\mathbb{R}^d$ in the definition of Gibbs measure and we assume throughout this thesis the growth condition

$$\lim_{r\to\infty} \inf_{x:|x|\geq r} \langle \nabla V(x), x/|x| \rangle = \infty \tag{78}$$

to ensure ergodicity.

We use $\tau_1 < \tau_2$ to denote the two temperatures under which the Gibbs distributions we are simulating, and we denote the product Gibbs measure density by π, hence

$$\pi(x_1, x_2) \propto e^{-V(x_1)/\tau_1} e^{-V(x_2)/\tau_2}.$$

We use the notation "proportional to"here because we only know π up to a normalizing constant. If W_1 and W_2 are independent Wiener processes, then the following stochastic differential equation

$$\begin{aligned} dX_1 &= -\nabla V(X_1)dt + \sqrt{2\tau_1}dW_1 \\ dX_2 &= -\nabla V(X_2)dt + \sqrt{2\tau_2}dW_2 \end{aligned} \tag{79}$$

defines a unique (in distribution) Markov process taking values in $\mathbb{R}^d \times \mathbb{R}^d$ and one can prove that under the growth condition (78),

$$\mu(x_1, x_2) = \pi(x_1, x_2)\, dx_1 dx_2 = e^{-V(x)/\tau_1} e^{-V(x)/\tau_2} dx_1 dx_2 \tag{80}$$

defines the unique invariant distribution of the Markov process (X_1, X_2). As we have mentioned, the basic idea of parallel tempering is to bring "particle swap"into the picture. Hence at random times the X_1 component is moved to the current location of the X_2 component, and vice versa. Swapping is done according to a state dependent intensity, and the resulting process is actually a Markov jump diffusion. The form of the jump intensity can be selected so that the invariant distribution remains the same, and thus the empirical measure of X_1 can still be used to approximate μ. Specifically, the jump intensity or swapping intensity is of the Metropolis form $ag(X_1, X_2)$, where

$$g(x_1, x_2) = 1 \wedge \frac{\pi(x_2, x_1)}{\pi(x_1, x_2)} \tag{81}$$

and $a \in (0, \infty)$ is a constant. Note that the calculation of g does not require the knowledge of the normalizing constants. By swapping the particles at these state dependent random times, the process (X_1, X_2) which we shall now denote by (X_1^a, X_2^a) is still Markovian. We can calculate the infinitesimal generator (recall the definition in (5)) of this Markov process and denote it by $\mathcal{L}^a$. Hence for smooth functions $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ we have

$$\begin{aligned}\mathcal{L}^a f(x_1, x_2) = &- \langle \nabla_{x_1} f(x_1, x_2), \nabla V(x_1)\rangle - \langle \nabla_{x_2} f(x_1, x_2), \nabla V(x_2)\rangle \\ &+ \tau_1 \mathrm{tr}\left[\nabla^2_{x_1 x_1} f(x_1, x_2)\right] + \tau_2 \mathrm{tr}\left[\nabla^2_{x_2 x_2} f(x_1, x_2)\right] \\ &+ a g(x_1, x_2)\left[f(x_2, x_1) - f(x_1, x_2)\right],\end{aligned}$$

where $\nabla_{x_i} f$ and $\nabla^2_{x_i x_i} f$ denote the gradient and the Hessian matrix with respect to x_i, respectively, and tr denotes trace.

Straightforward calculations show that for any smooth function f which vanishes at infinity

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \mathcal{L}^a f(x_1, x_2) \mu(dx_1 dx_2) = 0. \tag{82}$$

Since the condition (78) implies that $V(x) \to \infty$ as $|x| \to \infty$, by the Echeverria's Theorem [**11**, Theorem 4.9.17], μ is the unique invariant probability distribution of the process (X_1^a, X_2^a).

2. Large deviations rate function for the empirical measures

As we have mentioned, we will use the large deviations principle as an MCMC performance measure. We want to study the convergence rate of the empirical measures defined by

$$\eta_T^a \doteq \frac{1}{T}\int_0^T \delta_{(X_1^a(t), X_2^a(t))} dt$$

to the invariant distribution μ defined by (80). This convergence is guaranteed by ergodic theory [**1**] under the growth condition (78) and the following transitivity condition: there are times T_1 and T_2 such that for any $x, y \in \mathbb{R}^d \times \mathbb{R}^d$,

$$\int_0^{T_1} e^{-t} p(x, dz; t) dt \ll \int_0^{T_2} e^{-t} p(y, dz; t) dt.$$

We will state a uniform large deviations principle for the analogous Markov jump processes in Chapter 4 Section 5.3 and prove it in the appendix. For now, we will state the following result and omit the largely straightforward calculation since its role here is motivational.

Let ν be a probability measure on $\mathbb{R}^d \times \mathbb{R}^d$ with smooth density. Define $\theta(x_1, x_2) \doteq [d\nu/d\mu](x_1, x_2)$. Then $I^a(\nu)$ can be expressed as

$$I^a(\nu) = J_0(\nu) + aJ_1(\nu), \tag{83}$$

where

$$J_0(\nu) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \frac{1}{8\theta(x_1, x_2)^2} \left[\tau_1 \left\|\nabla_{x_1}\theta(x_1, x_2)\right\|^2 + \tau_2 \left\|\nabla_{x_2}\theta(x_1, x_2)\right\|^2\right] \nu(dx_1dx_2)$$
$$J_1(\nu) = \int_{\mathbb{R}^d \times \mathbb{R}^d} g(x_1, x_2)\ell\left(\sqrt{\frac{\theta(x_2, x_1)}{\theta(x_1, x_2)}}\right) \nu(dx_1dx_2),$$

and where $\ell(z) = z\log z - z + 1$ for $z \geq 0$ is familiar from the large deviations theory for jump processes.

The key observation is that the rate function $I^a(\nu)$ is affine in the swapping rate a, with $J_0(\nu)$ the rate function in the case of no swapping. Furthermore, $J_1(\nu) \geq 0$ with equality if and only if $\theta(x_2, x_1) = \theta(x_1, x_2)$ for ν-a.e. (x_1, x_2). This form of the rate function, and in particular its monotonicity in a, motivates the study of the *infinite swapping limit* as $a \to \infty$.

Remark 7. The limit of the rate function I^a satisfies

$$I^\infty(\nu) \doteq \lim_{a\to\infty} I^a(\nu) = \begin{cases} J_0(\nu) & \theta(x_1, x_2) = \theta(x_2, x_1) \ \nu\text{-a.s.,} \\ \infty & \textit{otherwise.} \end{cases}$$

Hence for $I^\infty(\nu)$ to be finite it is necessary that ν put exactly the same relative weight as μ on the points (x_1, x_2) and (x_2, x_1). Note that if a process could be constructed with I^∞ as its rate function, then with the large deviations rate as our criteria such a process improves on parallel tempering with finite swapping rate in exactly those situations where parallel tempering improves on the process with no swapping at all.

3. Infinite swapping: distributional limit of parallel tempering

From a practical perspective, it may appear that there are limitations on how much benefit one obtains by letting $a \to \infty$. When implemented in discrete time, the overall jump intensity corresponds to the generation of roughly a independent random variables that are uniform on $[0,1]$ for each corresponding unit of continuous time, and based on each uniform variable a comparison is made to decide whether or not to make the swap. Hence even for fixed and finite T, the computations required to simulate a single trajectory scale like a as $a \to \infty$. Thus it is of interest if one can gain the benefit of the higher swapping rate without all the computational burden. This turns out to be possible, but requires that we view the prelimit processes in a different way.

It is clear that the processes (X_1^a, X_2^a) are not tight as $a \to \infty$, since the number of discontinuities of size $O(1)$ will grow without bound in any time interval of positive length. In order to obtain a limit, we consider alternative processes defined by

$$
\begin{aligned}
dY_1^a &= -\nabla V(Y_1^a)dt + \sqrt{2\tau_1 1_{\{Z^a=0\}} + 2\tau_2 1_{\{Z^a=1\}}}\, dW_1 \\
dY_2^a &= -\nabla V(Y_2^a)dt + \sqrt{2\tau_2 1_{\{Z^a=0\}} + 2\tau_1 1_{\{Z^a=1\}}}\, dW_2
\end{aligned}
\tag{84}
$$

where Z^a is a jump process that switches from state 0 to state 1 with intensity $ag(Y_1^a, Y_2^a)$ and from state 1 to state 0 with intensity $ag(Y_2^a, Y_1^a)$. Compared to conventional parallel tempering, the processes (Y_1^a, Y_2^a) swap the diffusion coefficients at the jump times rather than the physical locations of two particles with constant diffusion coefficients. For this reason, we refer to the solution to (84) as the *temperature swapped process,* in order to distinguish it from the *particle swapped process* (X_1^a, X_2^a). We illustrate these processes in Figure 1. Note that the solid line and the dotted line represent X_1^a and X_2^a, respectively. These processes have more and more frequent jumps of size $O(1)$ as $a \to \infty$. In contrast, the process (Y_1^a, Y_2^a) have varying diffusion coefficient. The figure attempts to also suggest features of the discrete time setting, with both successful and failed swap attempts.

Clearly the empirical measure of (Y_1^a, Y_2^a) does not provide an approximation to μ. Instead, we should shift attention between (Y_1^a, Y_2^a) and (Y_2^a, Y_1^a) depending on the value of Z^a. Indeed, the random probability measures

$$
\eta_T^a = \frac{1}{T}\int_0^T \left[1_{\{Z^a(t)=0\}}\delta_{(Y_1^a(t), Y_2^a(t))} + 1_{\{Z^a(t)=1\}}\delta_{(Y_2^a(t), Y_1^a(t))}\right] dt \tag{85}
$$

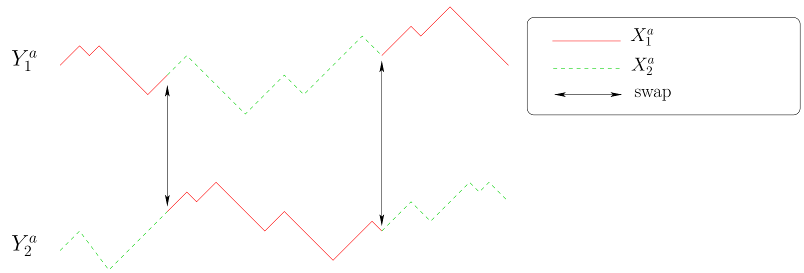

FIGURE 1. Temperature swapped and particle swapped processes

have the same distribution as

$$\frac{1}{T}\int_0^T \delta_{(X_1^a(s),X_2^a(s))}ds,$$

and hence converge to μ at the same rate. However, these processes and measures have well defined limits in distribution as $a \to \infty$. More precisely, we have the following result. For the proof see [**15**]. Related (but more complex) calculations are needed to prove the uniform large deviations result given in Theorem 7.

THEOREM 6. *Assume that ∇V is locally Lipschitz continuous. Then for each T the sequence (Y_1^a, Y_2^a, η_T^a) converges in distribution to $(Y_1^\infty, Y_2^\infty, \eta_T)$ as $a \to \infty$, where (Y_1^∞, Y_2^∞) is the unique strong solution to*

$$\begin{aligned} dY_1^\infty &= -\nabla V(Y_1^\infty)dt + \sqrt{2\tau_1\rho(Y_1^\infty, Y_2^\infty) + 2\tau_2\rho(Y_2^\infty, Y_1^\infty)}dW_1 \\ dY_2^\infty &= -\nabla V(Y_2^\infty)dt + \sqrt{2\tau_2\rho(Y_1^\infty, Y_2^\infty) + 2\tau_1\rho(Y_2^\infty, Y_1^\infty)}dW_2, \end{aligned} \tag{86}$$

$$\eta_T^\infty = \frac{1}{T}\int_0^T \left[\rho(Y_1^\infty(t), Y_2^\infty(t))\delta_{(Y_1^\infty(t),Y_2^\infty(t))} + \rho(Y_2^\infty(t), Y_1^\infty(t))\delta_{(Y_2^\infty(t),Y_1^\infty(t))}\right] dt, \tag{87}$$

and

$$\rho(x_1, x_2) \doteq \frac{\pi(x_1, x_2)}{\pi(x_2, x_1) + \pi(x_1, x_2)}.$$

The existence and form of the limit are due to the time scale separation between the fast Z^a process and the slow (Y_1^a, Y_2^a) process. To give an intuitive explanation of the limit dynamics, consider the prelimit processes (84). Suppose that on a small time interval, the value of the slow process (Y_1^a, Y_2^a) does not vary much, say $(Y_1^a, Y_2^a) \approx (x_1, x_2)$. Given the dynamics of the binary process Z^a, it is easy to verify that as

a tends to infinity the fractions of time that $Z^a = 0$ and $Z^a = 1$ are $\rho(x_1, x_2)$ and $\rho(x_2, x_1)$, respectively. This leads to the limit dynamics (86). When mapped back to the particle swapped process, $\rho(x_1, x_2)$ and $\rho(x_2, x_1)$ account for the fraction of time that $(X_1^a, X_2^a) = (x_1, x_2)$ and $(X_1^a, X_2^a) = (x_2, x_1)$, respectively, which naturally leads to the limit weighted empirical measure (87).

The weights ρ_1 and ρ_2 do not depend on the unknown normalization constant, and in fact

$$\rho(x_1, x_2) = \frac{e^{-\frac{V(x_1)}{\tau_1} - \frac{V(x_2)}{\tau_2}}}{e^{-\frac{V(x_1)}{\tau_1} - \frac{V(x_2)}{\tau_2}} + e^{-\frac{V(x_2)}{\tau_1} - \frac{V(x_1)}{\tau_2}}} \tag{88}$$

and

$$\rho(x_2, x_1) = 1 - \rho(x_1, x_2) = \frac{e^{-\frac{V(x_2)}{\tau_1} - \frac{V(x_1)}{\tau_2}}}{e^{-\frac{V(x_1)}{\tau_1} - \frac{V(x_2)}{\tau_2}} + e^{-\frac{V(x_2)}{\tau_1} - \frac{V(x_1)}{\tau_2}}}.$$

The following properties of the limit system are worth noting.

1. Instantaneous equilibration of multiple locations. Observe that the lower temperature component of this modified "empirical measure," i.e., the first marginal, uses contributions from both components at all times, corrected according to the weights. The form of the weights in (87) guarantees that the contributions to η_T^∞ from locations (Y_1^∞, Y_2^∞) and (Y_2^∞, Y_1^∞) are at any time perfectly balanced according to the invariant distribution on product space.

2. Symmetry and invariant distribution. While the marginals of η_T^∞ play very different roles, the dynamics of Y_1^∞ and Y_2^∞ are actually symmetric. Using the Echeverria's Theorem [**11**, Theorem 4.9.17], it can be shown that the unique invariant distribution of the process (Y_1^∞, Y_2^∞) has the density

$$\frac{1}{2}[\pi(x_1, x_2) + \pi(x_2, x_1)].$$

It then follows from the ergodic theorem that $\eta_T^\infty \Rightarrow \mu$ w.p.1 as $T \to \infty$. This is hardly surprising, since μ is the invariant distribution for the prelimit processes (X_1^a, X_2^a).

3. Escape from local minima. Finally it is worth commenting on the behavior of the diffusion coefficients as a function of the relative positions of Y_1^∞ and Y_2^∞ on an energy landscape. Recall that $\tau_1 < \tau_2$. Suppose that $Y_1^\infty(t)$ is near the bottom of

a local minimum (which for simplicity we set to be zero), while $Y_2^\infty(t)$ is at a higher energy level, perhaps within the same local minimum. Then

$$\rho(y_1,y_2) \approx \frac{e^{-\frac{V(y_2)}{\tau_2}}}{e^{-\frac{V(y_2)}{\tau_2}} + e^{-\frac{V(y_2)}{\tau_1}}} \approx 1,\ \rho(y_2,y_1) = 1-\rho(y_1,y_2) \approx 0.$$

Thus to some degree the dynamics look like

$$\begin{aligned} dY_1^\infty &= -\nabla V(Y_1^\infty)dt + \sqrt{2\tau_1}dW_1 \\ dY_2^\infty &= -\nabla V(Y_2^\infty)dt + \sqrt{2\tau_2}dW_2, \end{aligned}$$

i.e., the particle higher up on the energy landscape is given the greater diffusion coefficient, while the one near the bottom of the well is automatically given the lower coefficient. Hence the particle which is already closer to escaping from the well is automatically given the greater noise (within the confines of (τ_1,τ_2)). Recalling the role of the higher temperature particle is to more assiduously explore the landscape in parallel tempering, this is an interesting property.

One can apply results from [**5**] to show that the empirical measure of the infinite swapping limit $\{\eta_T^\infty : T>0\}$ satisfies a large deviations principle with rate function I^∞ as defined in Remark 7. However, to justify the claim that the infinite swapping model is truly superior to the finite swapping variant (note that $I^\infty \geq I^a$ for any finite a), one should establish a *uniform* large deviations principle, which would show that I^∞ is the correct rate function for any sequence $\{a_T : T>0\} \subset [0,\infty]$ such that $a_T \to \infty$ as $T \to \infty$. We omit the proof here, since in Theorem 7 the analogous result will be proved in the setting of continuous time jump Markov processes.

4. Infinite swapping: a generalization to multiple temperatures

In practice parallel tempering uses swaps between more than two temperatures. A key reason is that if the gap between the temperatures is too large then the probability of a successful swap under the [discrete time version of the] Metropolis rule (81) is far too low for the exchange of information to be effective. A natural generalization is to introduce, to the degree that computational feasibility is maintained, a ladder of higher temperatures, and then attempt pairwise swaps between particles. There are a variety of schemes used to select which pair to attempt the swap, including deterministic and randomized rules for selecting only adjacent temperatures or arbitrary pair of temperatures. However, if one were to replace any of these particle swapped processes with its equivalent temperature swapped analogue and consider the infinite swapping limit, one would get the same system of process dynamics and weighted empirical measures which we now describe.

Suppose that besides the lowest temperature τ_1 (in many cases the temperature of principal interest), we introduce the collection of higher temperatures

$$\tau_1 < \tau_2 < \cdots < \tau_K.$$

Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_K) \in (\mathbb{R}^d)^K$ be a generic point in the state space of the process and define a product Gibbs distribution with the density

$$\pi(\boldsymbol{y}) = \pi(y_1, y_2, \ldots, y_K) \propto e^{-V(y_1)/\tau_1} e^{-V(y_2)/\tau_2} \cdots e^{-V(y_K)/\tau_K}.$$

The limit of the temperature swapped processes with K temperatures takes the form

$$\begin{aligned}
dY_1^{\infty} &= -\nabla V\left(Y_1^{\infty}\right) dt + \sqrt{2\rho_{11}\tau_1 + 2\rho_{12}\tau_2 + \ldots + 2\rho_{1K}\tau_K}\, dW_1 \\
dY_2^{\infty} &= -\nabla V\left(Y_2^{\infty}\right) dt + \sqrt{2\rho_{21}\tau_1 + 2\rho_{22}\tau_2 + \ldots + 2\rho_{2K}\tau_K}\, dW_2 \\
&\vdots \\
dY_K^{\infty} &= -\nabla V\left(Y_K^{\infty}\right) dt + \sqrt{2\rho_{K1}\tau_1 + 2\rho_{K2}\tau_2 + \ldots + 2\rho_{KK}\tau_K}\, dW_K.
\end{aligned}$$

To define these weights ρ_{ij} it is convenient to introduce some new notation.

Let S_K be the collection of all bijective mappings from $\{1, 2, \ldots, K\}$ to itself. S_K has $K!$ elements, each of which corresponds to a unique permutation of the set $\{1, 2, \ldots, K\}$, and S_K forms a group with the group action defined by composition. Let σ^{-1} denote the inverse of σ. Furthermore, for each $\sigma \in S_K$ and every $\boldsymbol{y} = (y_1, y_2, \ldots, y_K) \in (\mathbb{R}^d)^K$, define $\boldsymbol{y}_\sigma \doteq \left(y_{\sigma(1)}, y_{\sigma(2)}, \ldots, y_{\sigma(K)}\right)$.

At the level of the prelimit particle swapped process, we interpret the permutation σ to correspond to event that the particles at location $\boldsymbol{y} = (y_1, y_2, \ldots, y_K)$

are swapped to the new location $\boldsymbol{y}_\sigma = \left(y_{\sigma(1)}, y_{\sigma(2)}, \ldots, y_{\sigma(K)}\right)$. Under the temperature swapped process, this corresponds to the event that particles initially assigned temperatures in the order $\tau_1, \tau_2, \ldots, \tau_K$ have now been assigned the temperatures $\tau_{\sigma^{-1}(1)}, \tau_{\sigma^{-1}(2)}, \ldots, \tau_{\sigma^{-1}(K)}$.

The identification of the infinite swapping limit of the temperature swapped processes is very similar to that of the two temperature model in the previous section. By exploiting the time-scale separation, one can assume that in a small time interval the only motion is due to temperature swapping and the motion due to diffusion is negligible. Hence the fraction of time that the permutation σ is in effect should again be proportional to the relative weight assigned by the invariant distribution to $\boldsymbol{y}_\sigma$, that is,

$$\pi(\boldsymbol{y}_\sigma) = \pi\left(y_{\sigma(1)}, y_{\sigma(2)}, \ldots, y_{\sigma(K)}\right).$$

Thus if

$$w(\boldsymbol{y}) \doteq \frac{\pi(\boldsymbol{y})}{\sum_{\theta \in S_K} \pi(\boldsymbol{y}_\theta)},$$

then the fraction of time that the permutation σ is in effect is $w(\boldsymbol{y}_\sigma)$. Note that for any $\boldsymbol{y}$,

$$\sum_{\sigma \in S_K} w(\boldsymbol{y}_\sigma) = 1.$$

Going back to the definition of the weights $\rho_{ij}(\boldsymbol{y})$, $i, j = 1, \ldots, K$, it is clear that they represent the limit proportion of time that the i-th particle is assigned the temperature j and hence will satisfy

$$\rho_{ij}(\boldsymbol{y}) = \sum_{\sigma:\ \sigma(j)=i} w(\boldsymbol{y}_\sigma).$$

Likewise the replacement for the empirical measure, accounting as it does for mapping the temperature swapped process back to the particle swapped process, is given by

$$\eta_T^\infty = \frac{1}{T}\int_0^T \sum_{\sigma \in S_K} w(\boldsymbol{Y}_\sigma^\infty(t))\delta_{\bar{\boldsymbol{Y}}_\sigma^\infty(t)} dt, \tag{89}$$

where $\boldsymbol{Y}_\sigma^\infty(t) \doteq [\boldsymbol{Y}^\infty(t)]_\sigma = (Y_{\sigma(1)}^\infty(t), Y_{\sigma(2)}^\infty(t), \ldots, Y_{\sigma(K)}^\infty(t))$.

The instantaneous equilibration property still holds for the infinite swapping system with multiple temperatures. That is, at any time $t \in [0, T]$ and given a current position $\boldsymbol{Y}^{\infty}(t) = \boldsymbol{y}$, the weighted empirical measure η_T^{∞} has contributions from all locations of the form $\boldsymbol{y}_{\sigma}, \sigma \in S_K$, balanced exactly according to their relative contributions from the invariant density $\pi\left(\boldsymbol{y}_{\sigma}\right)$. The dynamics of $\boldsymbol{Y}^{\infty}$ are again symmetric, and the density of the invariant distribution at point $\boldsymbol{y}$ is

$$\frac{1}{K!} \sum_{\sigma \in S_K} \pi(\boldsymbol{y}_{\sigma}).$$

REMARK 8. The infinite swapping process described above allows the most effective communication between all temperatures, and is the "best" in the sense that it leads to the largest large deviations rate function and hence the fastest rate of convergence. However, computation of the coefficients becomes very demanding for even moderate values of K, since one needs to evaluate $K!$ terms from all possible permutations. In Chapter 5 we discuss a more tractable and easily implementable family of schemes which are essentially approximations to the infinite swapping model presented in the current section and have very similar performance. We call the current model the full infinite swapping model since it uses the whole permutation group S_K, as opposed to the partial infinite swapping model which we will discuss in the next chapter where only subgroups of S_K are used.

5. Infinite swapping for Markov jump processes

The continuous time diffusion model is a convenient vehicle to convey the main idea of infinite swapping. In practice, however, algorithms are implemented in discrete time. In this section we discuss continuous time pure jump Markov processes and the associated infinite swapping limit. The purpose of this intermediate step is to serve as a bridge between the diffusion and discrete time Markov chain models. These two types of processes have some subtle differences regarding the infinite swapping limit which is best illustrated through the continuous time jump Markov model.

In this section we discuss the two-temperature model, and omit the completely analogous multiple-temperature counterpart. We will not refer to temperatures τ_1 and τ_2 to distinguish dynamics. Instead, let $\alpha_1(x, dy)$ and $\alpha_2(x, dy)$ be two probability transition kernels on $\mathbb{R}^d$ given $\mathbb{R}^d$. One can think of α_i as the dynamics under temperature τ_i for $i = 1, 2$. We assume that for each $i = 1, 2$ the stationary distribution μ_i associated with the transition kernel α_i admits the density π_i in order to be consistent with the diffusion models, and define

$$\mu = \mu_1 \times \mu_2, \ \pi(x_1, x_2) = \pi_1(x_1)\pi_2(x_2).$$

We assume that the kernels are Feller and have a density that is uniformly bounded with respect to Lebesgue measure. These conditions would always be satisfied in practice. Finally, we assume that the detailed balance or reversibility condition holds, that is,

$$\alpha_i(x, dz)\pi_i(x)dx = \alpha_i(z, dx)\pi_i(z)dz. \tag{90}$$

5.1. Model setup. In the absence of swapping [i.e., swapping rate $a = 0$], the dynamics of the system are as follows. Let $\boldsymbol{X}^0 = \{\boldsymbol{X}^0(t) = (X_1^0(t), X_2^0(t)) : t \geq 0\}$ denote a continuous time process taking values in $\mathbb{R}^d \times \mathbb{R}^d$. The probability transition kernel associated with the embedded Markov chain, denoted by $\bar{\boldsymbol{X}}^0 = \{(\bar{X}_1^0(j), \bar{X}_2^0(j)) : j = 0, 1, \ldots\}$, is

$$P\{\bar{\boldsymbol{X}}^0(j+1) \in (dy_1, dy_2) | \bar{\boldsymbol{X}}^0(j) = (x_1, x_2)\} = \alpha_1(x_1, dy_1)\alpha_2(x_2, dy_2).$$

Without loss of generality, we assume that the jump times occur according to a Poisson process with rate one. In other words, let $\{\tau_i\}$ be a sequence of independent and identically distributed (iid) exponential random variables with rate one that are independent of $\boldsymbol{X}^0$. Then

$$\boldsymbol{X}^0(t) = \bar{\boldsymbol{X}}^0(j), \text{ for } \sum_{i=1}^{j} \tau_i \leq t < \sum_{i=1}^{j+1} \tau_i.$$

The infinitesimal generator of $\boldsymbol{X}^0$ is such that for a given smooth function f,

$$\mathcal{L}^0 f(x_1, x_2) = \int_{\mathbb{R}^d \times \mathbb{R}^d} [f(y_1, y_2) - f(x_1, x_2)]\alpha_1(x_1, dy_1)\alpha_2(x_2, dy_2).$$

Owing to the detailed balance condition (90), the operator $\mathcal{L}^0$ is self-adjoint. Using arguments similar to but simpler than those used to prove the uniform LDP in Theorem 7, the large deviations rate function I^0 associated with the occupation measure

$$\eta_T^0 = \frac{1}{T}\int_0^T \delta_{\boldsymbol{X}^0(t)}\, dt$$

can be explicitly identified: for any probability measure ν on $\mathbb{R}^d \times \mathbb{R}^d$ with $\nu \ll \mu$ and $\theta = d\nu/d\mu$,

$$I^0(\nu) = 1 - \int_{(\mathbb{R}^d\times\mathbb{R}^d)^2} \sqrt{\theta(x_1,x_2)\theta(y_1,y_2)}\pi(x_1,x_2)\alpha_1(x_1,dy_1)\alpha_2(x_2,dy_2)\, dx_1dx_2,$$

and I^0 is extended to all of $\mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ by lower semicontinuous regularization.

5.2. Finite swapping model. Denote by $\boldsymbol{X}^a = \{(X_1^a(t), X_2^a(t)) : t \geq 0\}$ the state process of the finite swapping model with swapping rate a, and let $\bar{\boldsymbol{X}}^a = \{(\bar{X}_1^a(j), \bar{X}_2^a(j)) : j = 0, 1, \ldots\}$ be the embedded Markov chain. The probability transition kernel for $\bar{\boldsymbol{X}}^a$ is

$$\begin{aligned} P\{\bar{\boldsymbol{X}}^a(j+1) \in (dy_1, dy_2) | \bar{\boldsymbol{X}}^a(j) = (x_1,x_2)\} \;&=\; \frac{1}{a+1}\alpha_1(x_1,dy_1)\alpha_2(x_2,dy_2) \\ + \frac{a}{a+1}\left[g(x_1,x_2)\delta_{(x_2,x_1)}(dy_1,dy_2) + (1-g(x_1,x_2))\delta_{(x_1,x_2)}(dy_1,dy_2)\right], \end{aligned}$$

where g is defined as in (81). Furthermore, let $\{\tau_i^a\}$ be a sequence of iid exponential random variables with rate $(a+1)$ and define

$$\boldsymbol{X}^a(t) = \bar{\boldsymbol{X}}^a(j), \text{ for } \sum_{i=1}^{j}\tau_i^a \leq t < \sum_{i=1}^{j+1}\tau_i^a.$$

In other words, the jumps occur according to a Poisson process with rate $a+1$. Note that there are two types of jumps. At any jump time, with probability $1/(a+1)$ it will be a jump according to the underlying probability transition kernels α_1 and α_2. With probability $a/(a+1)$ it will be an attempted swap which will succeed with the probability determined by g. As a grows, the swap attempts become more and more frequent. However, the time between two consecutive jumps of the first type will have the same distribution as

$$S^a = \sum_{i=1}^{N^a}\tau_i^a$$

where N^a is a geometric random variable with parameter $1/(a+1)$. It is easy to argue that for any a the distribution of S^a is exponential with rate one. This observation will be useful when we derive the infinite swapping limit.

The infinitesimal generator $\mathcal{L}^a$ of $\bar{\boldsymbol{X}}^a$ is such that for any smooth function f on $\mathbb{R}^d \times \mathbb{R}^d$

$$\mathcal{L}^a f(x_1, x_2) = \int_{\mathbb{R}^d \times \mathbb{R}^d} [f(y_1, y_2) - f(x_1, x_2)] \alpha_1(x_1, dy_1) \alpha_2(x_2, dy_2)$$
$$+ ag(x_1, x_2)[f(x_2, x_1) - f(x_1, x_2)].$$

It is not difficult to check that the stationary distribution of $\bar{\boldsymbol{X}}^a$ remains μ and that $\mathcal{L}^a$ is self-adjoint. As before, the large deviations rate function I^a for the occupation measure

$$\eta_T^a = \frac{1}{T} \int_0^T \delta_{\boldsymbol{X}^a(t)} \, dt \tag{91}$$

can be explicitly identified. Indeed, for any probability measure ν on $\mathbb{R}^d \times \mathbb{R}^d$ with $\nu \ll \mu$ and $\theta = d\nu/d\mu$

$$I^a(\nu) = I^0(\nu) + aJ(\nu),$$

where

$$J(\nu) = \int_{\mathbb{R}^d \times \mathbb{R}^d} g(x_1, x_2) \ell \left(\sqrt{\frac{\theta(x_2, x_1)}{\theta(x_1, x_2)}} \right) \nu(dx_1, dx_2).$$

Note that as before, I^a is monotonically increasing with respect to a. Since $J(\nu) \geq 0$ with equality if and only if $\theta(x_1, x_2) = \theta(x_2, x_1)$ ν-a.e., we have

$$I^\infty(\nu) \doteq \lim_{a \to \infty} I^a(\nu) = \begin{cases} I^0(\nu) & \text{if } \theta(x_1, x_2) = \theta(x_2, x_1) \ \nu\text{-a.s.}, \\ \infty & \text{otherwise.} \end{cases} \tag{92}$$

5.3. Infinite swapping limit. The infinite swapping limit for $\boldsymbol{X}^a$ as $a \to \infty$ can be similarly obtained by considering the corresponding temperature swapped processes. Since the times between jumps determined by α_1 and α_2 are always exponential with rate one, the infinite swapping limit $\boldsymbol{Y}^\infty = (Y_1^\infty, Y_2^\infty)$ is a pure jump

Markov process where jumps occur according to a Poisson process with rate one. In other words,

$$\boldsymbol{Y}^{\infty}(t) = \bar{\boldsymbol{Y}}^{\infty}(j), \text{ for } \sum_{i=1}^{j} \tau_i \leq t < \sum_{i=1}^{j+1} \tau_i,$$

where $\bar{\boldsymbol{Y}}^{\infty}$ is the embedded Markov chain and $\{\tau_i\}$ a sequence of iid exponential random variables with rate one. Furthermore, the probability transition kernel for $\bar{\boldsymbol{Y}}^{\infty}$ is

$$\begin{aligned}(93) \qquad & P\{\bar{\boldsymbol{Y}}^{\infty}(j+1) \in (dz_1, dz_2) | \bar{\boldsymbol{Y}}^{\infty}(j) = (y_1, y_2)\} \\ & = \rho(y_1, y_2)\alpha_1(y_1, dz_1)\alpha_2(y_2, dz_2) + \rho(y_2, y_1)\alpha_2(y_1, dz_1)\alpha_1(y_2, dz_2),\end{aligned}$$

where the weight function ρ is defined as in Theorem 6. It is not difficult to argue that the stationary distribution for $\boldsymbol{Y}^{\infty}$ is

$$\bar{\mu}(dy_1, dy_2) = \frac{1}{2}[\pi(y_1, y_2) + \pi(y_2, y_1)]\, dy_1 dy_2,$$

and the weighted occupation measure

$$(94) \quad \eta_T^{\infty} = \frac{1}{T}\int_0^T \left[\rho(Y_1^{\infty}(t), Y_2^{\infty}(t))\delta_{(Y_1^{\infty}(t), Y_2^{\infty}(t))} + \rho(Y_2^{\infty}(t), Y_1^{\infty}(t))\delta_{(Y_2^{\infty}(t), Y_1^{\infty}(t))}\right] dt$$

converges to $\mu(dx_1, dx_2) = \pi(x_1, x_2)dx_1dx_2$ as $T \to \infty$. It is obvious that the dynamics of the infinite swapping limit are symmetric and instantaneously equilibrate the contribution from (Y_1, Y_2) and (Y_2, Y_1) according to the invariant measure, owing to the weight function ρ.

We have the following uniform large deviations principle result, which justifies the superiority of infinite swapping model. Its proof is deferred to the appendix. It should be noted that rate identification is not covered by the existing literature, even in the case of a fixed swapping rate, due to the pure jump nature of the process.

THEOREM 7. *The occupation measure $\{\eta_T^{\infty} : T > 0\}$ satisfies a large deviations principle with rate function I^{∞}. More generally, define the finite swapping model as in Chapter 4 Section 5.2. Consider any sequence $\{a_T : T > 0\} \subset [0, \infty]$ such that $a_T \to \infty$ as $T \to \infty$, and interpret $a_T < \infty$ to mean that $\eta_T^{a_T}$ is defined by (91) with*

$a = a_T$, and $a_T = \infty$ to mean that $\eta_T^{a_T}$ is defined by (94). Then $\{\eta_T^{a_T} : T > 0\}$ satisfies an LDP with the rate function I^∞ defined in equation (92).

CHAPTER 5

Discrete time infinite swapping model and implementation issues

1. Discrete time algorithms

1.1. Conventional parallel tempering algorithms. In the discrete time, multi-temperature algorithms that are actually implemented, a swap is attempted after a deterministic or random number of time steps, with a success probability of the form (81). The two temperatures corresponding to particles for which a swap is attempted can be chosen according to a deterministic or random schedule, and as noted previously are usually adjacent since otherwise the success probability (81) will be too small to allow efficient exchange of information.

As before it suffices to describe the algorithm in the setting of two temperatures. As in Section 5, let $\alpha_i(x, dy)$ denote the probability transition kernel for temperature τ_i whose stationary distribution has a density π_i for $i = 1, 2$. For now let $N = 1/a$ be a fixed positive integer that determines the frequency of swap attempts. Let $\boldsymbol{X} = \{(X_1(j), X_2(j)) : j = 0, 1, \ldots\}$ denote the state process. Then the evolution of the dynamics is as follows. For any integer $k \geq 1$ and $(k-1)(N+1) \leq j \leq k(N+1)-2$,

$$P\{\boldsymbol{X}(j+1) \in (dy_1, dy_2) | \boldsymbol{X}(j) = (x_1, x_2)\} = \alpha_1(x_1, dy_1)\alpha_2(x_2, dy_2)$$

and for $j = k(N+1) - 1$,

$$P\{\boldsymbol{X}(j+1) = (x_2, x_1) | \boldsymbol{X}(j) = (x_1, x_2)\} = g(x_1, x_2),$$

$$P\{\boldsymbol{X}(j+1) = (x_1, x_2) | \boldsymbol{X}(j) = (x_1, x_2)\} = 1 - g(x_1, x_2).$$

Thus a swap is attempted after every N ordinary time steps based on the underlying transition kernels α_1 and α_2. The case $N = 1/a$ with a an integer greater than one corresponds to the case where multiple swaps are attempted between two ordinary time steps. The unique invariant distribution of X is $\mu(dx_1dx_2) = \pi(x_1,x_2)\,dx_1dx_2$, regardless of the value of N, and the occupation measure

$$\frac{1}{J}\sum_{j=0}^{J-1}\delta_{\boldsymbol{X}(j)}$$

converges to μ as $J\to\infty$ almost surely.

Remark 9. Note that N could be random. For example, if N is chosen to be a geometric random variable with mean λ, then X is exactly the embedded Markov chain of the pure jump Markov process X^a with $a = 1/\lambda$ in Subsection 5.2.

1.2. Infinite swapping model. As with the continuous time case, to produce a well-defined limit one must consider the temperature swapped process and then consider the limit as swapping frequency tends to infinity. It turns out that the limit is exactly the embedded Markov chain for the pure jump Markov process in Subsection 5.3. That is, the infinite swapping limit in discrete time is a Markov chain $\boldsymbol{Y}^\infty = \{(Y_1^\infty(j), Y_2^\infty(j)) : j = 0,1,\ldots\}$ with the transition kernel

$$\rho(y_1,y_2)\alpha_1(y_1,dz_1)\alpha_2(y_2,dz_2) + \rho(y_2,y_1)\alpha_2(y_1,dz_1)\alpha_1(y_2,dz_2). \tag{95}$$

The corresponding weighted empirical measure is

$$\eta_J^\infty \doteq \frac{1}{J}\sum_{j=0}^{J-1}\left[\rho(Y_1^\infty(j),Y_2^\infty(j))\delta_{(Y_1^\infty(j),Y_2^\infty(j))} + \rho(Y_2^\infty(j),Y_1^\infty(j))\delta_{(Y_2^\infty(j),Y_1^\infty(j))}\right]. \tag{96}$$

The generalization to multiple temperatures is also straightforward. Suppose that there are K temperatures. Denote the infinite swapping limit process by $\boldsymbol{Y} = \{\boldsymbol{Y}(j) : j = 0,1,\ldots\}$, which is a Markov chain taking values in the space $(\mathbb{R}^d)^K$. Given that the current state of the chain is $\boldsymbol{Y}^\infty(j) = \boldsymbol{y} = (y_1,\ldots,y_K)$, define as before the weights

$$w(\boldsymbol{y}) \doteq \frac{\pi(\boldsymbol{y})}{\sum_{\theta\in S_K}\pi(\boldsymbol{y}_{\boldsymbol{\theta}})}.$$

Then the transition kernel of $\boldsymbol{Y}^\infty$ is

$$P(\boldsymbol{Y}^\infty(j+1)\in d\boldsymbol{z}|\boldsymbol{Y}^\infty(j)=\boldsymbol{y}) = \sum_{\sigma\in S_K} w(\boldsymbol{y}_\sigma)\alpha_1(y_{\sigma(1)},dz_{\sigma(1)})\cdots\alpha_K(y_{\sigma(K)},dz_{\sigma(K)}).$$

The discrete time numerical approximation to the invariant distribution is

$$\eta_J^{\infty} \doteq \frac{1}{J} \sum_{j=0}^{J-1} \sum_{\sigma \in S_K} w(\boldsymbol{Y}_\sigma^{\infty}(j)) \delta_{\boldsymbol{Y}_\sigma^{\infty}(j)}.$$

Remark 10. It is not difficult to derive large deviations principles for the discrete time finite swapping or infinite swapping models. However, it remains an open question whether the rate function is monotonic with respect to the swap rate (frequency). However, the discrete time large deviations rate function can be obtained from that of the continuous time pure jump Markov process models through the contraction principle, and the two coincide in the limit as the transition kernels α_i correspond to an infinitesimal time step for the diffusion process (79). Hence the discrete time rate function will be at least approximately monotone, and in this sense the infinite swapping limit should (at least approximately) dominate all finite swapping algorithms. This is supported by the data presented in Section 3 and the much more extensive empirical study presented in [**16**].

2. Partial infinite swapping

As noted in Section 4, the number of weights and their calculation can become unwieldy for infinite swapping even when the number of temperatures is moderate. In this section we construct algorithms that maintain most of the benefit of the infinite swapping algorithm but at a much lower computational cost. In the first subsection we describe the infinite swapping limit models when only a subgroup of the permutations of the particles (respectively, temperatures) are allowed by the prelimit particle (respectively, temperature) swapped process. The computational complexity of these limit models will be controlled by limiting the number of permutations that communicate with each other through the swapping mechanism. The infinite swapping models in this subsection will be called *partial infinite swapping models,* as opposed to the full infinite swapping models in the previous section. The second subsection shows how such partial infinite swapping schemes can be interwoven to approximate the full infinite swapping model.

2.1. Partial infinite swapping models. We consider subsets A of S_K with the property that A is an algebraic subgroup of S_K. That is,

1. the identity belongs to A;
2. if $\sigma_1, \sigma_2 \in A$ then $\sigma_1 \circ \sigma_2 \in A$, where $\circ$ denotes composition;
3. if $\sigma \in A$ then $\sigma^{-1} \in A$.

Although one can write down a partial infinite swapping model that corresponds to instantaneous equilibration for an arbitrary subset A, it is only when A is a subgroup that the corresponding partial infinite swapping process has an interpretation as the limit of parallel tempering type processes. When alternating between partial infinite swapping processes, a "handoff" rule will be needed, and it is only for those which correspond to subgroups that such a handoff rule is well defined. This point is discussed in some detail in the next section.

The definition of the partial infinite swapping process based on A is completely analogous to that of the full infinite swapping process. The state process $\{\boldsymbol{Y}(j) : j = 0, 1, \ldots\}$ is a Markov chain with the transition kernel

$$\alpha^A(\boldsymbol{y}, d\boldsymbol{z}) \doteq \sum_{\sigma \in A} \tilde{w}^A\left(\boldsymbol{y}_\sigma\right) \alpha_1(y_{\sigma(1)}, dz_{\sigma(1)}) \cdots \alpha_K(y_{\sigma(K)}, dz_{\sigma(K)}) \tag{97}$$

and the weighted empirical measure is

$$\tilde{\eta}_J \doteq \frac{1}{J} \sum_{j=0}^{J-1} \sum_{\sigma \in A} \tilde{w}^A(\boldsymbol{Y}_\sigma(j)) \delta_{\boldsymbol{Y}_\sigma(j)},$$

where the weight function $\tilde{w}^A$ is defined by

$$\tilde{w}^A\left(\boldsymbol{y}\right) \doteq \frac{\pi\left(\boldsymbol{y}\right)}{\sum_{\theta \in A} \pi\left(\boldsymbol{y}_\theta\right)}, \tag{98}$$

and satisfies for any $\boldsymbol{y}$

$$\sum_{\sigma \in A} \tilde{w}^A(\boldsymbol{y}_\sigma) = 1.$$

We omit the dependence on both $a = \infty$ and A from the notation. Note that in contrast with the full swapping system, it is only those permutations of $\boldsymbol{y}$ corresponding to $\sigma \in A$ that are balanced according to the invariant distribution in their contributions to $\tilde{\eta}_J$.

To illustrate the construction we present a few examples. With a standard abuse of notation denote the permutation σ such that $\sigma(i) = a_i$ by the form $(a_1, a_2, \ldots, a_K)$. In particular, $(1, 2, \ldots, K)$ is the identity of the group S_K.

Example 1. *Let $K = 4$ and $A = \{(1,2,3,4), (2,1,3,4)\}$. This corresponds to only allowing swaps between temperatures τ_1 and τ_2 at the prelimit. Define*

$$\tilde{w}(\boldsymbol{y}) = \frac{\pi(y_1, y_2, y_3, y_4)}{\pi(y_1, y_2, y_3, y_4) + \pi(y_2, y_1, y_3, y_4)}$$

The probability transition kernel of the corresponding partial infinite swapping process is given by

$$\tilde{w}\,(y_1, y_2, y_3, y_4)\, \alpha_1(y_1, dz_1)\alpha_2(y_2, dz_2)\alpha_3(y_3, dz_3)\alpha_4(y_4, dz_4)$$
$$+ \tilde{w}\,(y_2, y_1, y_3, y_4)\, \alpha_1(y_2, dz_2)\alpha_2(y_1, dz_1)\alpha_3(y_3, dz_3)\alpha_4(y_4, dz_4)$$

and the contribution to the weighted empirical measure is

$$\tilde{w}\,(Y_1, Y_2, Y_3, Y_4)\, \delta_{(Y_1,Y_2,Y_3,Y_4)} + \tilde{w}\,(Y_2, Y_1, Y_3, Y_4)\, \delta_{(Y_2,Y_1,Y_3,Y_4)}.$$

Note that with π_{ij} denoting the marginal invariant distribution on the i-th and j-th components, the weight function can be written as

$$\tilde{w}(\boldsymbol{y}) = \frac{\pi_{12}(y_1, y_2)}{\pi_{12}(y_1, y_2) + \pi_{12}(y_2, y_1)},$$

which is consistent with the weights of the two-temperature model in Theorem 6.

Example 2. *Again take $K = 4$, but this time use the subgroup generated by $(2,1,3,4)$ and $(1,2,4,3)$, i.e., $A = \{(1,2,3,4), (2,1,3,4), (1,2,4,3), (2,1,4,3)\}$. Then the dynamics are given by*

$$\tilde{w}\,(y_1, y_2, y_3, y_4)\, \alpha_1(y_1, dz_1)\alpha_2(y_2, dz_2)\alpha_3(y_3, dz_3)\alpha_4(y_4, dz_4)$$
$$+ \tilde{w}\,(y_2, y_1, y_3, y_4)\, \alpha_1(y_2, dz_2)\alpha_2(y_1, dz_1)\alpha_3(y_3, dz_3)\alpha_4(y_4, dz_4)$$
$$+ \tilde{w}\,(y_1, y_2, y_4, y_3)\, \alpha_1(y_1, dz_1)\alpha_2(y_2, dz_2)\alpha_3(y_4, dz_4)\alpha_4(y_3, dz_3)$$
$$+ \tilde{w}\,(y_2, y_1, y_4, y_3)\, \alpha_1(y_2, dz_2)\alpha_2(y_1, dz_1)\alpha_3(y_4, dz_4)\alpha_4(y_3, dz_3)$$

where the weight function $\tilde{w}$ is defined by

$$\widetilde{w}\left(\boldsymbol{y}\right)=\frac{\pi\left(y_1,y_2,y_3,y_4\right)}{\pi\left(y_1,y_2,y_3,y_4\right)+\pi\left(y_2,y_1,y_3,y_4\right)+\pi\left(y_1,y_2,y_4,y_3\right)+\pi\left(y_2,y_1,y_4,y_3\right)}.$$

The contribution to the weighted empirical measure is

$$\begin{aligned}&\tilde{w}\left(Y_1,Y_2,Y_3,Y_4\right)\delta_{\left(\bar{Y}Y_1,Y_2,Y_3,Y_4\right)}+\tilde{w}\left(Y_2,Y_1,Y_3,Y_4\right)\delta_{(Y_2,Y_1,Y_3,Y_4)}\\&\quad+\tilde{w}\left(Y_1,Y_2,Y_4,Y_3\right)\delta_{(Y_1,Y_2,Y_4,Y_3)}+\tilde{w}\left(Y_2,Y_1,Y_4,Y_3\right)\delta_{(Y_2,Y_1,Y_4,Y_3)}.\end{aligned}$$

EXAMPLE 3. *We let $K=3$ and take A to be the subgroup of S_K generated by the rotation $(2,3,1)$, i.e., $A=\{(1,2,3),(2,3,1),(3,1,2)\}$. Then the dynamics are given by*

$$\begin{aligned}&\tilde{w}\left(y_1,y_2,y_3\right)\alpha_1(y_1,dz_1)\alpha_2(y_2,dz_2)\alpha_3(y_3,dz_3)\\&\quad+\tilde{w}\left(y_2,y_3,y_1\right)\alpha_1(y_2,dz_2)\alpha_2(y_3,dz_3)\alpha_3(y_1,dz_1)\\&\quad+\tilde{w}\left(y_3,y_1,y_2\right)\alpha_1(y_3,dz_3)\alpha_2(y_1,dz_1)\alpha_3(y_2,dz_2)\end{aligned}$$

where

$$\tilde{w}\left(\boldsymbol{y}\right)=\frac{\pi(y_1,y_2,y_3)}{\pi(y_1,y_2,y_3)+\pi\left(y_2,y_3,y_1\right)+\pi\left(y_3,y_1,y_2\right)}$$

and the contribution to the weighted empirical measure is

$$\tilde{w}_1\left(Y_1,Y_2,Y_3\right)\delta_{(Y_1,Y_2,Y_3)}+\tilde{w}_2(Y_2,Y_3,Y_1)\delta_{(Y_2,Y_3,Y_1)}+\tilde{w}_3\left(Y_3,Y_1,Y_3\right)\delta_{(Y_3,Y_1,Y_3)}.$$

The first two examples would correspond to the infinite swapping limit of a standard parallel tempering process, where swaps between only 1 and 2 are allowed in the first example, and swaps between 1 and 2 and swaps between 3 and 4 are allowed in the second. Note that the computational complexity does not increase significantly between the first and second example. The third example corresponds to a very different sort of prelimit process, in which "rotations" of the coordinates $(y_1,y_2,y_3)\rightarrow(y_2,y_3,y_1)\rightarrow(y_3,y_1,y_2)\rightarrow(y_1,y_2,y_3)$ are allowed. One can devise a Metropolis type rule that allows such "swaps" and yields the indicated infinite swapping system.

2.2. Approximating full infinite swapping by partial swapping. In this section we consider the issue of alternating between such partial infinite swapping systems to approximate the full infinite swapping limit. Let A and B be subgroups of S_K. A and B are said to *generate* S_K if the smallest subgroup that contains A and B is S_K itself. Note that the total number of permutations in $A \cup B$ can be significantly smaller than $K!$, the size of S_K. In fact, it is possible to construct subgroups A and B that generate S_K and that the total number of permutations in $A \cup B$ is of order K. There is an obvious extension to more than two subgroups.

Example 4. *Let $K = 4$ and let A be generated by $\{(2, 1, 3, 4), (1, 3, 2, 4)\}$ and B be generated by $\{(1, 3, 2, 4), (1, 2, 4, 3)\}$, respectively. Thus A is the collection of 6 permutations that fix the last component and allow all rearrangements of the first three, while B fixes the first component and allows all rearrangements of the last three. Then A and B generate S_K.*

Example 5. *Let $K = 4$ and let A and B be subgroups generated by $\{(2, 1, 3, 4)\}$ and $\{(2, 3, 4, 1)\}$, respectively. In other words, $A = \{(1, 2, 3, 4), (2, 1, 3, 4)\}$ corresponds to only allowing permutations between the first two components, while $B = \{(1, 2, 3, 4), (2, 3, 4, 1), (3, 4, 1, 2), (4, 1, 2, 3)\}$ corresponds to cycling of the four temperatures. Then A and B generate S_K.*

To keep the computational cost controlled, one can approximate the full infinite swapping model by alternating between partial infinite swapping processes whose associated subgroups generate the whole group. However, one must be careful in how the "handoff" is made when switching between different partial swapping models. It turns out that one cannot simply switch between different partial infinite swapping dynamics (i.e., transition kernels). Recall that in order to get a consistent approximation to the desired target invariant distribution we do not use the empirical measure generated by $\boldsymbol{Y}$, but rather a carefully constructed weighted empirical measure that works with several permutations of $\boldsymbol{Y}$. Simply switching the dynamics and weights will in fact produce an algorithm that may not converge to the target distribution.

To see how one should design a handoff rule, note that if one considers a collection of transition kernels each having the same invariant distribution and alternates between them in a way that does not depend on the outcomes prior to a switch, then the resulting empirical measure will in fact converge to the common invariant distribution. This fact is used (at least implicitly) in the parallel tempering algorithm itself, where one alternates the pair of particles being considered for swapping according to deterministic or random rules so long as the random rules do not rely on previously observed outcomes.

Now we use the fact that each partial infinite swapping model is a limit of either a parallel tempering algorithm where only some pairs of particles are considered for swapping, or some more general form of parallel tempering which would allow groups of particles to simultaneously swap (according to an appropriate Metropolis-type acceptance rule). An example in the earlier category would be A in Example 4, which arises if only the pairs corresponding to temperatures τ_1, τ_2 and τ_2, τ_3 are allowed to swap, whereas an example in the latter category would be B in Example 5, which corresponds to allowing the particle at temperature τ_i to move to the location of the particle at temperature τ_{i-1} (with $\tau_0 = \tau_4$), and the reverse. Furthermore, each of these partial infinite swapping models arises as a limit of transition kernels of the corresponding temperature swapped processes which preserve the same common invariant distribution. In taking the limit as the swap rate tends to infinity, the correspondence between particle locations for a particle swapped process and the "instantaneously equilibrated" temperature swapped process Y is lost. However, one can construct a consistent algorithm by *reconstructing this correspondence*. In fact one should choose the particle location according to the probabilities (under the invariant distribution) associated with the various permutations in the subgroup. See Subsection 2.3 for more detailed discussion on the intuition behind this approximation algorithm.

We next present an algorithm for alternating between two partial infinite swapping dynamics. The restriction to two is for notational convenience only. Suppose that the dynamics are indexed by the corresponding subgroups A and B, and that n_A steps of subgroup A are to be alternated with n_B steps of subgroup B. For simplicity we do not describe a "burn-in" period. As in (97) and (98) we let $\alpha^A(\boldsymbol{y}, d\boldsymbol{z})$ and $\tilde{w}^A(\boldsymbol{y})$ denote the transition kernel for A and the weights allocated to the permutation $\sigma \in A$, respectively, and similarly for B.

ALGORITHM 1. *(Approximation to full infinite swapping)*

(1) *Initialization:* $\boldsymbol{X}^A(0) = \boldsymbol{Y}(0;0) \in (\mathbb{R}^d)^K, \ell = 1$.

(2) *Loop* ℓ*:*

 (a) *Initialization for* A *dynamics: set* $\boldsymbol{Y}(\ell;0) = \boldsymbol{X}^A(\ell - 1)$.

 (b) *Subgroup* A *dynamics: update* $\boldsymbol{Y}(\ell;k), k = 1, \ldots, n_A$ *according to the transition kernel* α^A*, and add*

$$\sum_{\sigma \in A} \tilde{w}^A(\boldsymbol{Y}_\sigma(\ell;k))\delta_{Y_{\boldsymbol{\sigma}}(\ell;k)}$$

 to the un-normalized empirical measure.

 (c) *Reconstructing particle locations at the end of* A *dynamics: Let* $\boldsymbol{X}^B(\ell)$ *be a random sample from the set* $\{\boldsymbol{Y}_{\boldsymbol{\sigma}}(\ell;n_A) : \sigma \in A\}$ *according to the weights* $\{\tilde{w}^A(\boldsymbol{Y}_\sigma(\ell;n_A)) : \sigma \in A\}$.

 (d) *Initialization for* B *dynamics: set* $\boldsymbol{Y}(\ell;n_A) = \boldsymbol{X}^B(\ell)$.

 (e) *Subgroup* B *dynamics: update* $\boldsymbol{Y}(\ell;k), k = n_A+1, \ldots, n_A+n_B$ *according to the transition kernel* α^B*, and add*

$$\sum_{\sigma \in B} \tilde{w}^B(\boldsymbol{Y}_\sigma(\ell;k))\delta_{Y_{\boldsymbol{\sigma}}(\ell;k)}$$

 to the un-normalized empirical measure.

 (f) *Reconstructing particle locations at the end of* B *dynamics: Let* $\boldsymbol{X}^A(\ell)$ *be a random sample from the set* $\{\boldsymbol{Y}_{\boldsymbol{\sigma}}(\ell;n_A+n_B) : \sigma \in B\}$ *according to the weights* $\{\tilde{w}^B(\boldsymbol{Y}_\sigma(\ell;n_A+n_B)) : \sigma \in B\}$.

 (g) *Set* $\ell = \ell + 1$ *and loop back to (a).*

(3) *Normalize the empirical measure.*

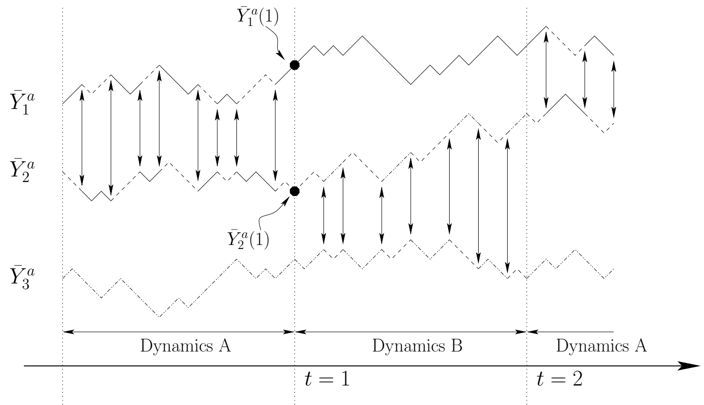

FIGURE 2. Approximation via partial infinite swapping

2.3. Discussions on the approximation. In this section we further discuss the intuition underlying the handoff rule between different partial infinite swapping dynamics and the approximation algorithm of the previous section. We temporarily assume that the model is in continuous time since the intuition is most transparent in this case.

For simplicity let us assume that there are three temperatures and two groups $A = \{(1,2,3),(2,1,3)\}$ and $B = \{(1,2,3),(1,3,2)\}$. That is, under group A dynamics only pairwise swaps between the temperatures τ_1 and τ_2 are allowed, while under the group B dynamics, only the swaps between τ_2 and τ_3 are allowed. See Figure 2.

Consider the following prelimit swapping model. Let the swap rate be a. The dynamics corresponding to group A and group B will be alternated on time intervals of length h. Hence on the interval $[2kh,(2k+1)h)$ the particle swapped process (X_1^a, X_2^a, X_3^a) only involves swaps between temperatures τ_1 and τ_2. One can easily construct the corresponding temperature swapped process (Y_1^a, Y_2^a, Y_3^a) as before. Note that $X_3^a = Y_3^a$ on this time interval. Similarly, on the interval $[(2k+1)h,(2k+2)h)$, only swaps between τ_2 and τ_3 are allowed and on this interval $X_1^a = Y_1^a$. Note that there is no ambiguity for the prelimit processes at the switch times $t = h, 2h, \ldots,$ since the locations of the particles (X_1^a, X_2^a, X_3^a) are known.

Now consider the limit as $a \to \infty$ with h being fixed. Without loss of generality, we will only discuss how to deal with the switch of the dynamics at time $t = h$.

On the time interval $[0, h)$ we have the partial infinite swapping limit process $\boldsymbol{Y}^A = (Y_1, Y_2, Y_3)$ that corresponds to the group A. Similarly it is clear that on the time interval $[h, 2h)$ we should have the partial infinite swapping process $\boldsymbol{Y}^B$ corresponding to the group B. The problem is, however, by taking the limit, we lose the information on the locations of the particles (X_1^a, X_2^a, X_3^a). Unless we can somehow recover this information at the switch time $t = h$ to assign $\boldsymbol{Y}^B(h)$, we cannot determine the dynamics of $\boldsymbol{Y}^B$ on $[h, 2h)$. The key is to recall that the infinite swapping limit instantaneously equilibrates multiple locations according to the invariant distribution. In other words, given $\boldsymbol{Y}^A(h-) = \boldsymbol{y} = (y_1, y_2, y_3)$, the locations of the particles are distributed according to

$$\sum_{\sigma \in A} \tilde{w}^A(\boldsymbol{y}_\sigma)\delta_{\boldsymbol{y}_\sigma} = \frac{\pi(y_1, y_2, y_3)\delta_{(y_1,y_2,y_3)} + \pi(y_2, y_1, y_3)\delta_{(y_2,y_1,y_3)}}{\pi(y_1, y_2, y_3) + \pi(y_2, y_1, y_3)}.$$

Therefore, in order to identify the locations of the particles at time h, we will take a random sample from this distribution once $\boldsymbol{Y}^A(h-)$ is known. This explains the handoff rule used at the switch times of partial infinite swapping processes.

Now we let $h \to 0$. Since A and B generate the whole permutation group S_K, it is easy to check that at each time instant, the locations of $\{\boldsymbol{y}_\sigma : \sigma \in S_K\}$ are equilibrated according to their invariant distribution, and therefore in the limit we will attain the full infinite swapping model. This can be made rigorous by exploiting the time scale separation between the slow diffusion processes $(\boldsymbol{Y}^A, \boldsymbol{Y}^B)$ and the fast switching process. We omit the proof because the discussion is largely motivational.

Coming back to the discrete time partial infinite swapping model, it is clear that Algorithm 1 is nothing but a straightforward adaption of the preceding discussion to discrete time. The only difference is that one cannot establish an analogous result regarding approximation to the full infinite swapping model as in continuous time. The subtlety here is that in continuous time, as $h \to 0$, one can basically ignore any effect from the diffusion on any small time interval and assume that the process is only making jumps between different permutations of a fixed triple (y_1, y_2, y_3). This time scale separation is no longer valid in discrete time. In this setting, the performance of a scheme based on interweaving partial infinite swapping schemes lies between parallel tempering and full infinite swapping, and computational results suggest that it is closer to the latter than the former.

The issue of which interwoven partial schemes will perform best is an open question. In practice we have used schemes of the following form. Suppose that a set of say 45 temperatures is given. We then partition 45 into blocks of sizes $3, 6, \ldots, 6$, with the first block containing the lowest three temperatures, the second block the next six, and so on. Dynamic A then is given by allowing all permutations within each block. Note that the complexity of the coefficients is then no worse than 6!. In Dynamic B we use the partition $6, 6, \ldots, 6, 3$. The form of the partial scheme is heuristically motivated by allowing the largest possible overlap between the different blocks when switching between dynamics, subject to the constraint that blocks be of size no greater than 6.

3. Numerical examples

In this section we present data comparing parallel tempering at various swap rates and both full and partial infinite swapping. We present what we call "relaxation studies." The quantity of interest is the average potential energy of the lowest temperature component under the invariant distribution. In these studies, the system is run a long time to reach equilibrium, after which it is repeatedly pushed out of equilibrium and we measure the time needed to "relax" back to equilibrium. Each cycle consists of temporarily raising the temperatures (heating) of some of the lowest temperature components for a number of steps sufficient to push the average potential energy away from the "true" value (as measured by either sample or time averages). The temperatures are then returned to their "true" values (cooling) for a fixed number of steps, and the process is then repeated 2,000 times. We plot the average of the 2,000 samples as a function of the number of moves, and the performance of the algorithm is captured by the rate at which these averages approach the correct value.

A simple relaxation study for a four-temperature 13-atom LJ cluster is displayed in the following. One cycle of the four-temperatures variations is shown in Figure 3. The Metropolis Monte Carlo dynamics under each temperature is simulated using the "smart Monte Carlo" (SMC) scheme of [**18**]. In this example one period consists of 600 SMC moves, with moves 50-450 being the "cooling"portion and the remainder the "heating"portion of the cycle. As this cycle is repeated, potential energies corresponding to various numbers of SMC moves are accumulated and "relaxation"curves of the type shown in Figure 4 are generated. The results in Figure 4 are generated using 2000 heating/cooling cycles of the type displayed in Figure 3 using complete (full) infinite swapping (INS) sampling. For comparison purposes, the values of the average potential energies for the high and low-temperatures shown in Figure 4 are computed in separate INS simulations using 6×10^5 SMC moves. The rates at which the heating/cooling curves approach their limiting equilibrium values can be seen in Figure 4.

Relaxation studies allow us to compare the rates of equilibration for different sampling methods. In Figure 5, for example, we compare the cooling portions of the lowest of the four temperatures shown in Figure 4 obtained using various sampling methods, including INS sampling, parallel tempering, and conventional single-temperature Metropolis methods. The parallel tempering results shown in Figure 5 are labeled with the percentages of trial swap moves involved. Specifically, the percentage stated corresponds to the probability that during each update of the coordinates there is an attempted parallel tempering exchange between one, randomly chosen, nearest-neighbor pair of temperatures. In this notation, conventional, single-temperature Metropolis results correspond to the zero swapping limit of parallel tempering. The other computational parameters used for the studies of Figure 5 are the same as those in Figure 4.

The Lennard-Jones cluster of 13 atoms is not a particularly demanding problem, we present it here so a comparison can be made between INS and partial infinite swapping (PINS). Figure 6 compares relaxation performance of the INS scheme with

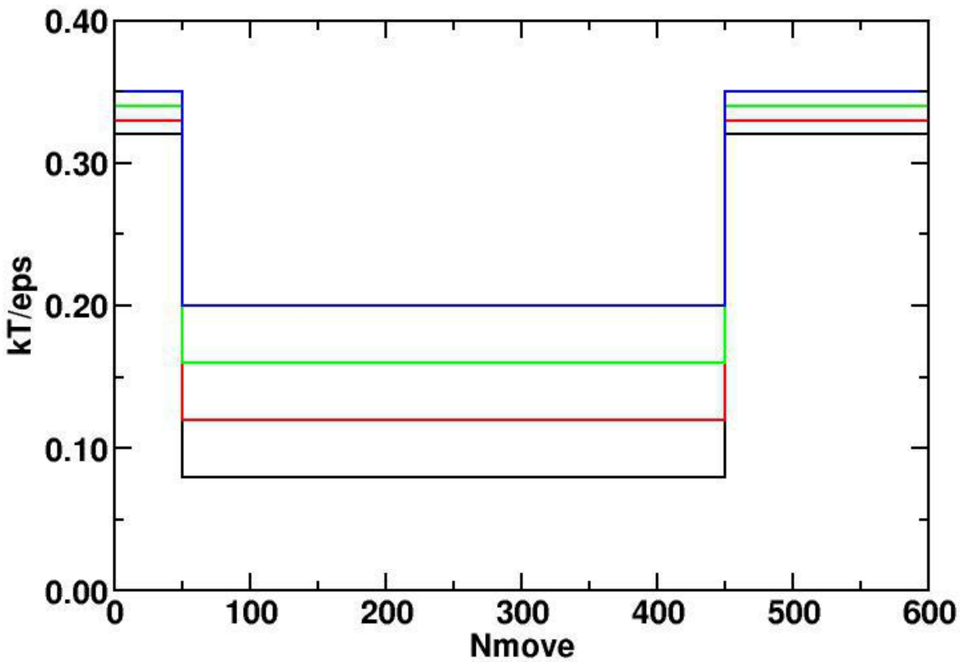

Figure 3. Temperature profiles for the four-temperature, LJ-13 relaxation simulations

PT with different swapping frequencies. We see that the most efficient of the PT schemes appears to use an attempted swap rate of around 64%. (The rates that would typically be used in such calculations are in the range of 5-10%.) Figure 7 magnifies a portion of the graph, but plots only the best parallel tempering result and adds a PINS result based on blocks of the form 1-3 and 3-1 (a lowest temperature and a group of three higher temperatures for one chain and group of three lower temperatures and a single highest temperature for the other, and with a handoff at each Metropolis step). Little difference is observed between the INS and PINS, though both appear to have an obvious advantage over the best PT scheme.

The 38-atom Lennard-Jones cluster is an appreciably more complex system. Data for this example is obtained using a 45 temperature ensemble is given in Figure 8. Heating and cooling profiles used are of the generic type illustrated in Figure 3. In the present study, however, the cycles consist of 1200 SMC moves, each of one unit

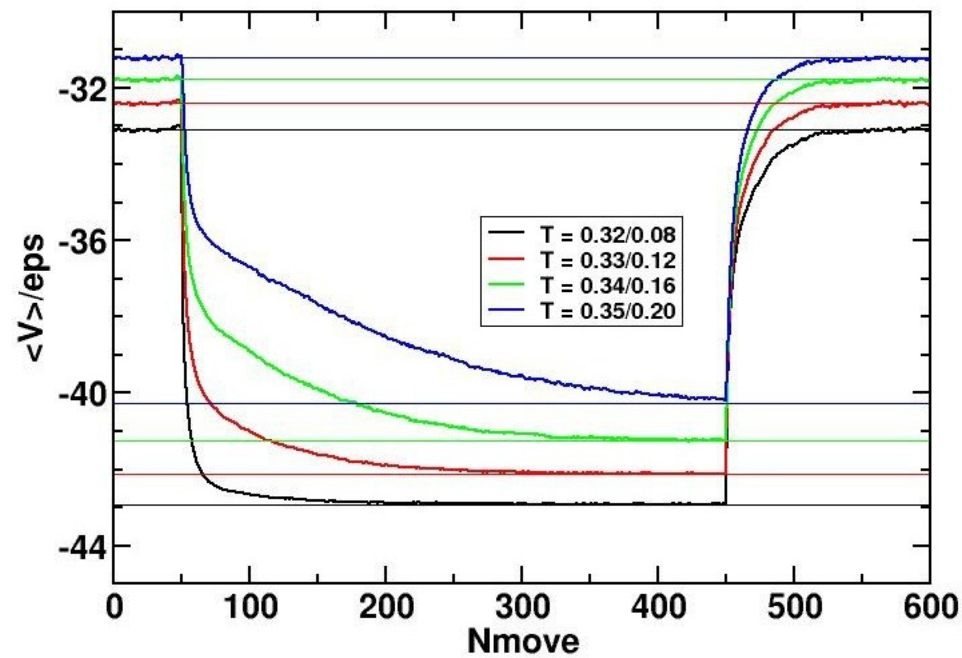

FIGURE 4. Average potential energies for the LJ-13 infinite swapping relaxation simulations performed under INS scheme using the temperature profiles in Figure 3

Lennard-Jones time duration. The cooling segment is taken as the portion of the cycle from moves 200 to 800 with the remainder being the heating portion. During the cooling portion of the cycle the 45 temperatures in the ensemble cover the range from (0.050-0.210) in temperature steps of 0.005, and from (0.210 - 0.330) in steps of 0.010 while during the heating portion of the cycle temperatures less than or equal to 0.150 are set equal to 0.150. Because full infinite swapping is impossible for this larger computational ensemble, we use the partial infinite swapping. Specifically, each of the chains is composed of seven, six temperature blocks plus one smaller block of three temperatures for a total of 45 temperatures. In one chain, the three-temperature block is made up of the lowest three temperatures, in the other chain the highest three. For comparison, results are also presented for parallel tempering and "floating"partial infinite swapping, which utilizes a symmetrized band of temperatures that moves

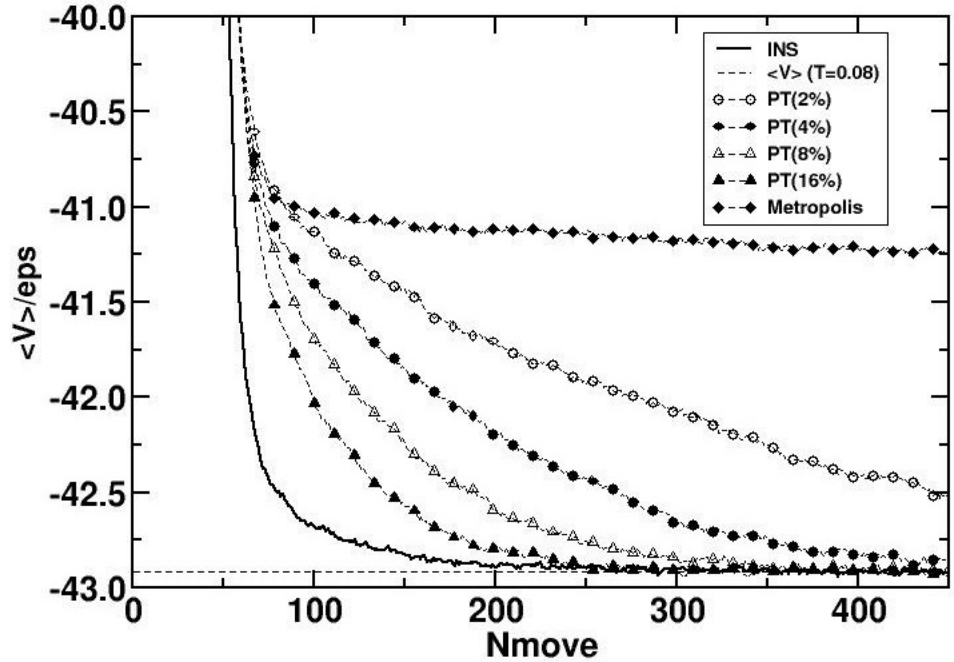

FIGURE 5. The cooling portion of the lowest of the four-temperature studies of Figure 4. Results compared INS, PT with different frequencies and ordinary Metropolis sampling with the numerically exact value represented by the dashed line

randomly within the larger computational ensemble, for details see [**16**]. The results shown in Figure 8 are obtained using 500 thermal cycles.

The 38-atom Lennard-Jones cluster has an interesting landscape. In particular, while the global and lowest-lying local minima are similar in energy, the minimum energy pathway that separates them involves appreciably higher energies and contains 13 separate barriers [**22**, Chapter 8.3]. As discussed in [**16**], the partial infinite swapping approach is appreciably more effective than conventional tempering approaches in providing a proper sampling of this complex potential energy landscape.

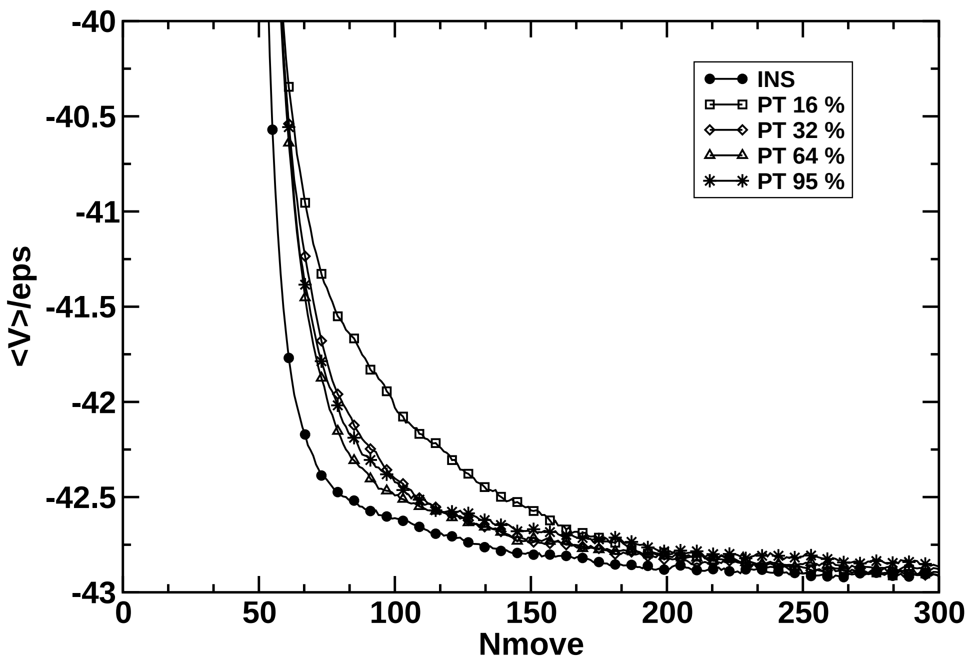

FIGURE 6. Cooling portion of the LJ-13 relaxation study: INS and PT with different swapping frequencies

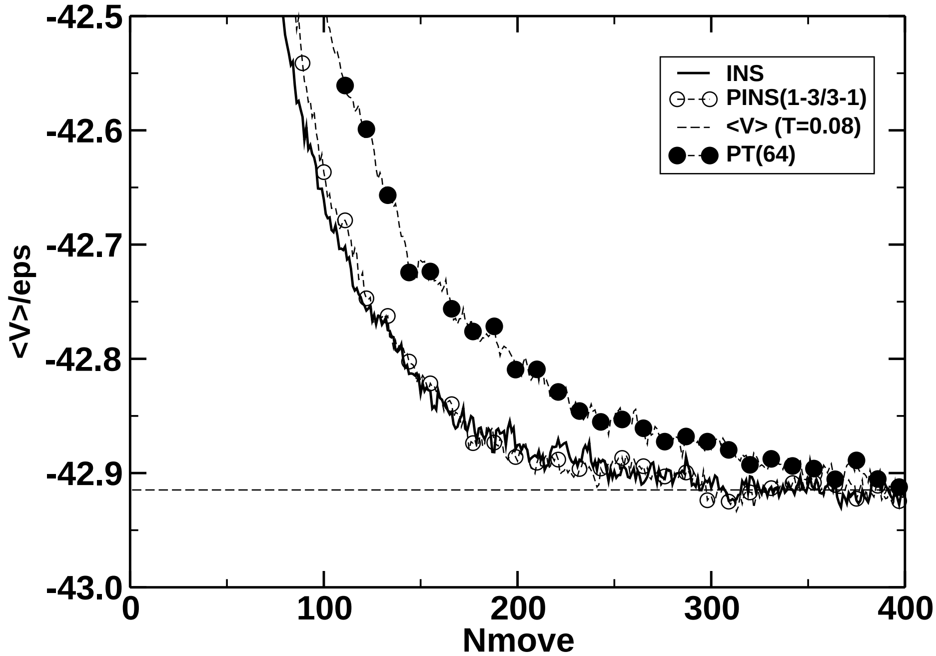

FIGURE 7. Cooling portion of LJ-13 relaxation study: INS, PT with optimal swapping frequency and PINS with temperature grouping strategies 1-3 and 3-1

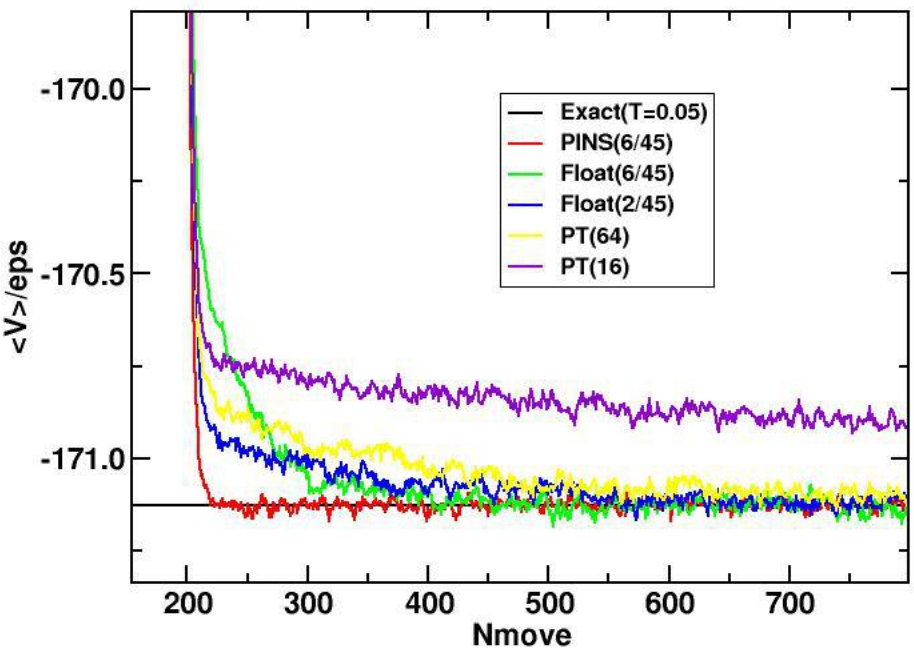

Figure 8. Cooling portion of LJ-38 relaxation study: PINS with temperature grouping strategies $3-6-\cdots 6$ and $6-\cdots-6-3$, floating PINS (Float) using two and six-temperatures band, and PT with swapping frequencies 16% and 64%

CHAPTER 6

Conclusion

The main result of this thesis is to explore the applications of large deviations theory to characterize the rate of convergence of Markov processes. We have reviewed several frequently used criteria as the performance measures of a Markov chain Monte Carlo. Unfortunately, none of the conventional method is conveniently applicable to the convergence of the empirical measures directly. On the other hand, a large deviations principle for the empirical measures can be established for an extensive class of Markov processes. We suggest in this thesis to use the large deviations rate function as a criterion to examine the effectiveness of Markov chain Monte Carlo.

It is well known that an explicit Donsker-Varadhan formula for the rate function of the empirical measures is available when a continuous time Markov process is self-adjoint and the transition function for all choices of time and spacial variables have densities with respect to a single σ-finite reference measure. However, for a Markov jump process, where the latter is not satisfied, an explicit formula is not provided in any existing literature. In fact, a large deviations theory for the empirical measures of Markov jump processes is not even available, since the Donsker-Varadhan result of empirical measures does not cover the jump case. In the first part of this thesis, we establish such an LDP result as well as an explicit formula.

In the second part of the thesis, we focus our attention to a particular MCMC scheme: the parallel tempering algorithm. The Markov chain constructed for parallel tempering implementation is one example of a reversible Markov process with jump aspect. By studying the rate function, we discover that the rate function is monotone in the swapping rate and hence arrive at the distributional limit of parallel tempering algorithm: the infinite swapping model. Intuitively, infinite swapping algorithm improves the performance of parallel tempering by allocating more effectively different parts of the simulation to do the job they are best at and facilitating the communication among them. Although intuitively and theoretically dominant to parallel tempering algorithm, infinite swapping scheme is not tractable due to its factorial scaling computational overhead. We thus propose a partial infinite swapping algorithm, which somewhat captures the infinite swapping improvement without incurring too much computation burden. Some numerical results are displayed to compare the performance between parallel tempering and full/partial infinite swapping.

APPENDIX A

Proof of inequality (54)

Proof. Recall that $R(\mu \,\|[\mu]_1 \otimes \alpha) < \infty$, where $[\mu]_1 = [\mu]_2$ and $\widetilde{\pi}$ is invariant under α. We want to prove

$$R([\mu]_1 \,\|\widetilde{\pi}) < \infty. \tag{99}$$

As stated in Chapter 2 Section 3, there exist an integer N and a real number $c > 0$ such that

$$\alpha^N(x, \cdot) \leq c\widetilde{\pi}(\cdot)$$

for all $x \in S$. Now let p be the regular conditional probability such that $\mu = [\mu]_1 \otimes p$. Hence we have

$$R(\mu \,\|[\mu]_1 \otimes \alpha) = R([\mu]_1 \otimes p \,\|[\mu]_1 \otimes \alpha) < \infty.$$

The chain rule of relative entropy implies that

$$R\left([\mu]_1 \otimes p \otimes \underset{N}{\cdots} \otimes p \,\middle\|\, [\mu]_1 \otimes \alpha \otimes \underset{N}{\cdots} \otimes \alpha\right) = N \cdot R([\mu]_1 \otimes p \,\|[\mu]_1 \otimes \alpha) < \infty. \tag{100}$$

Indeed, since $[\mu]_1$ is invariant under p, we have for any integer n, the n-th marginal of $\left[[\mu]_1 \otimes p \otimes \underset{n-1}{\cdots} \otimes p\right]$ is:

$$\left[[\mu]_1 \otimes p \otimes \underset{n-1}{\cdots} \otimes p\right]_n = [\mu]_1 .$$

Hence one can prove (100) by induction:

$$\begin{aligned}
&R\left([\mu]_1 \otimes p \otimes \underset{n}{\cdots} \otimes p \,\middle\|\, [\mu]_1 \otimes \alpha \otimes \underset{n}{\cdots} \otimes \alpha\right) \\
&= R\left([\mu]_1 \otimes p \otimes \underset{n-1}{\cdots} \otimes p \,\middle\|\, [\mu]_1 \otimes \alpha \otimes \underset{n-1}{\cdots} \otimes \alpha\right) + \int_S R(p \,\|\alpha) \left[[\mu]_1 \otimes p \otimes \underset{n-1}{\cdots} \otimes p\right]_n \\
&= (n-1) \cdot R([\mu]_1 \otimes p \,\|[\mu]_1 \otimes \alpha) + \int_S R(p \,\|\alpha) [\mu]_1 \\
&= n \cdot R([\mu]_1 \otimes p \,\|[\mu]_1 \otimes \alpha) .
\end{aligned}$$

We use notation $[\nu]_{k|j}$ to denote the regular conditional probability measure of the k-th marginal of ν given the j-th marginal of ν. Note that one can define a mapping from $\mathcal{P}\left(S^{N+1}\right)$ to $\mathcal{P}\left(S^2\right)$ such that each $\nu \in \mathcal{P}\left(S^{N+1}\right)$ is mapped to $[\nu]_1 \otimes [\nu]_{N+1|1}$, since the relative entropy for induced measures is always smaller, we have

$$R\left([\mu]_1 \otimes p^N \left\| [\mu]_1 \otimes \alpha^N\right.\right) < \infty,$$

where p^N and α^N are used to denote the N-step transition probability function. Now since $[\mu]_1$ is invariant under p, it is also invariant under p^N, hence $\left[[\mu]_1 \otimes p^N\right]_2 = [\mu]_1$. Using the chain rule of relative entropy again we have

$$R\left([\mu]_1 \left\| \left[[\mu]_1 \otimes \alpha^N\right]_2\right.\right) < \infty.$$

Now this would imply (99), indeed we have

$$\begin{aligned}
\infty &> R\left([\mu]_1 \left\| \left[[\mu]_1 \otimes \alpha^N\right]_2\right.\right) \\
&= R\left([\mu]_1 \| \widetilde{\pi}\right) - \log \int_S \frac{d\left([\mu]_1 \otimes \alpha^N\right)_2}{d\widetilde{\pi}} [\mu]_1 \\
&\geq R\left([\mu]_1 \| \widetilde{\pi}\right) - \log c.
\end{aligned}$$

$\square$

APPENDIX B

Proof of Lemma 7

Proof. Define for each $k \in \mathbb{N}_+$

$$\eta_T^k(\cdot) \doteq \frac{1}{T}\left[\sum_{i=1}^{R_T\wedge k-1} \delta_{X_{i-1}}(\cdot)\frac{\tau_i}{q(X_{i-1})} + \delta_{X_{R_T\wedge k-1}}(\cdot)\left(T - \sum_{i=1}^{R_T\wedge k-1}\frac{\tau_i}{q(X_{i-1})}\right)\right].$$

Now for any measure $\nu^k \in \mathcal{P}\left((S\times\mathbb{R}_+)^k\right)$, we can decompose ν^k as

$$\nu^k = \bar{\alpha}_0 \otimes \bar{\sigma}_1 \otimes \bar{\alpha}_1 \otimes \bar{\sigma}_2 \otimes \cdots \otimes \bar{\alpha}_{k-1} \otimes \bar{\sigma}_k. \tag{101}$$

Choose the barred random variables $\bar{X}_i$ and $\bar{\tau}_i$ according to $\bar{\alpha}_i$ and $\bar{\sigma}_i$ as before and define the corresponding $\bar{R}_T \wedge k$ the following way: if $\sum_{i=1}^{k} \bar{\tau}_i / q\left(\bar{X}_{i-1}\right) > T$, then $\bar{R}_T \wedge k \doteq \bar{R}_T$ where $\bar{R}_T$ is the integer that validates

$$\sum_{i=1}^{\bar{R}_T-1}\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)} \le T < \sum_{i=1}^{\bar{R}_T}\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)},$$

otherwise define $\bar{R}_T \wedge k \doteq k$. We also define

$$\bar{\eta}_T^k(\cdot) \doteq \frac{1}{T}\left[\sum_{i=1}^{\bar{R}_T\wedge k-1} \delta_{\bar{X}_{i-1}}(\cdot)\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)} + \delta_{\bar{X}_{\bar{R}_T\wedge k-1}}(\cdot)\left(T - \sum_{i=1}^{\bar{R}_T\wedge k-1}\frac{\bar{\tau}_i}{q\left(\bar{X}_{i-1}\right)}\right)\right]. \tag{102}$$

If we denote the multi-dimensional probability measure from the original dynamics by $\mu^k \in \mathcal{P}\left((S\times\mathbb{R}_+)^k\right)$, i.e.

$$\mu^k \doteq (\otimes_k \alpha) \times \left(\prod_k \sigma\right),$$

then applying Lemma 1, we have

$$-\frac{1}{T}\log E\left[\exp\left\{-TF\left(\eta_T^k\right)\right\}\right] = \inf_{\nu^k\in\mathcal{P}\left((S\times\mathbb{R}_+)^k\right)}\left[\int_{(S\times\mathbb{R}_+)^k} F\left(\bar{\eta}_T^k\right)\nu^k + \frac{1}{T}R\left(\nu^k \,\middle\|\, \mu^k\right)\right]. \tag{103}$$

If we apply Theorem 5 repeatedly to $R\left(\nu^k \left\| \mu^k\right.\right)$, we would obtain

$$R\left(\nu^k \left\| \mu^k\right.\right) = E\left[\sum_{i=1}^{k}\left(R\left(\bar{\alpha}_{i-1}\left\|\alpha\right.\right) + R\left(\bar{\sigma}_i\left\|\sigma\right.\right)\right)\right].$$

We can hence rewrite (103) as

(104)

$$-\frac{1}{T}\log E\left[\exp\left\{-TF\left(\eta_T^k\right)\right\}\right] = \inf_{\nu^k\in\mathcal{P}\left((S\times\mathbb{R}_+)^k\right)} E\left[F\left(\bar{\eta}_T^k\right) + \frac{1}{T}\sum_{i=1}^{k}\left(R\left(\bar{\alpha}_{i-1}\left\|\alpha\right.\right) + R\left(\bar{\sigma}_i\left\|\sigma\right.\right)\right)\right]$$

Now for each $\nu^k \in \mathcal{P}\left((S\times\mathbb{R}_+)^k\right)$, we construct another measure $\widehat{\nu}^k \in \mathcal{P}\left((S\times\mathbb{R}_+)^k\right)$ recursively as follows: define $\widehat{\alpha}_0 \doteq \bar{\alpha}_0$ and $\widehat{\sigma}_1 \doteq \bar{\sigma}_1$. For all $2 \le i \le k$, define $\widehat{\alpha}_{i-1}$ and $\widehat{\sigma}_i$ as

$$\left(\widehat{\alpha}_{i-1}, \widehat{\sigma}_i\right) = \begin{cases} \left(\bar{\alpha}_{i-1}, \bar{\sigma}_i\right) & \text{if } \sum_{j=1}^{i-1}\frac{\bar{\tau}_j}{q\left(\bar{X}_{j-1}\right)} \le T \\ (\alpha, \sigma) & \text{otherwise} \end{cases}.$$

We then define $\widehat{\nu}^k$ using $\widehat{\alpha}_i$ and $\widehat{\sigma}_i$ by (101). Following from the definition (102) we have $E\left[F\left(\widehat{\eta}_T^k\right)\right] = E\left[F\left(\bar{\eta}_T^k\right)\right]$. And

$$\begin{aligned} E\left[\sum_{i=1}^{k}\left(R\left(\widehat{\alpha}_{i-1}\left\|\alpha\right.\right) + R\left(\widehat{\sigma}_i\left\|\sigma\right.\right)\right)\right] &= E\left[\sum_{i=1}^{\widehat{R}_T\wedge k}\left(R\left(\widehat{\alpha}_{i-1}\left\|\alpha\right.\right) + R\left(\widehat{\sigma}_i\left\|\sigma\right.\right)\right)\right] \\ &= E\left[\sum_{i=1}^{\bar{R}_T\wedge k}\left(R\left(\bar{\alpha}_{i-1}\left\|\alpha\right.\right) + R\left(\bar{\sigma}_i\left\|\sigma\right.\right)\right)\right] \\ &\le E\left[\sum_{i=1}^{k}\left(R\left(\bar{\alpha}_{i-1}\left\|\alpha\right.\right) + R\left(\bar{\sigma}_i\left\|\sigma\right.\right)\right)\right]. \end{aligned}$$

Hence we can refine (104) as

(105)

$$-\frac{1}{T}\log E\left[\exp\left\{-TF\left(\eta_T^k\right)\right\}\right] = \inf_{\nu^k\in\mathcal{P}\left((S\times\mathbb{R}_+)^k\right)} E\left[F\left(\bar{\eta}_T^k\right) + \frac{1}{T}\sum_{i=1}^{\bar{R}_T\wedge k}\left(R\left(\bar{\alpha}_{i-1}\left\|\alpha\right.\right) + R\left(\bar{\sigma}_i\left\|\sigma\right.\right)\right)\right].$$

Now, since we have the pointwise convergence of both $R_T\wedge k \to R_T$ and $\bar{R}_T\wedge k \to \bar{R}_T$ as $k\to\infty$, by the dominated convergence theorem,

$$\lim_{k\to\infty} -\frac{1}{T}\log E\left[\exp\left\{-TF\left(\eta_T^k\right)\right\}\right] = -\frac{1}{T}\log E\left[\exp\left\{-TF\left(\eta_T\right)\right\}\right],$$

$$\text{and } \lim_{k\to\infty} E\left[F\left(\bar{\eta}_T^k\right)\right] = E\left[F\left(\bar{\eta}_T\right)\right].$$

Also by the monotone convergence theorem, we have

$$\lim_{k\to\infty} E\left[\sum_{i=1}^{\bar{R}_T\wedge k}\left(\left(R\left(\bar{\alpha}_{i-1}\,\|\alpha\right)+R\left(\bar{\sigma}_i\,\|\sigma\right)\right)\right)\right]=E\left[\sum_{i=1}^{\bar{R}_T}\left(\left(R\left(\bar{\alpha}_{i-1}\,\|\alpha\right)+R\left(\bar{\sigma}_i\,\|\sigma\right)\right)\right)\right].$$

Hence by taking limits to both sides of (105), we arrive at

$$-\frac{1}{T}\log E\left[\exp\left\{-TF\left(\eta_T\right)\right\}\right]=\inf\overline{E}\left[F\left(\bar{\eta}_T\right)+\frac{1}{T}\sum_{i=1}^{\bar{R}_T}\left(\left(R\left(\bar{\alpha}_{i-1}\,\|\alpha\right)+R\left(\bar{\sigma}_i\,\|\sigma\right)\right)\right)\right]$$

where the infimum is taken over all controlled measures $\{\bar{\alpha}_i, \bar{\sigma}_i\}$. We have proved the lemma. $\square$

APPENDIX C

Proof of Theorem 7

Throughout the proof, we let $S = S_1 \times S_1$, where $S_1 \subset \mathbb{R}^d$ is convex and compact. Let $\mathcal{P}(S)$ denote the Polish space of all probability measures on S equipped with the topology of weak convergence. For any probability measure $\nu \in \mathcal{P}(S)$, define its mirror image $\nu^R \in \mathcal{P}(S)$ by requiring

$$\nu^R(A \times B) = \nu(B \times A)$$

for all Borel sets $A, B \subset \mathbb{R}^d$. Furthermore, as in the rest of the paper, a bold symbol $\boldsymbol{x} \in S$ means $\boldsymbol{x} = (x_1, x_2)$, where $x_1, x_2 \in \mathbb{R}^d$, and $\boldsymbol{x}^R = (x_2, x_1)$. We also use the notation

$$\alpha(\boldsymbol{x}, d\boldsymbol{y}) \doteq \alpha_1(x_1, dy_1)\alpha_2(x_2, dy_2),$$

which is a probability transition kernel defined on S given S.

To prove the uniform large deviation principle, it suffices to prove the equivalent uniform Laplace principle [**7**, Chapter 1]. To simplify the proof we have assumed that S is compact. This would be the case if, e.g., V is defined with periodic boundary conditions. The general case can be handled under (78) by using V as a Lyapunov function [**7**, Section 8.2]. It will be convenient to split this into upper and lower bounds. We also consider just the (more complicated) case where $a_T \to \infty$ but $a_T < \infty$ for each T. Allowing $a_T = \infty$ requires a different notation to handle this special case, but does not change the structure of the proof otherwise.

We will show for any bounded continuous function $F : \mathcal{P}(S) \to \mathbb{R}$ that

$$\liminf_{T\to\infty} -\frac{1}{T} \log E\left[\exp\{-TF(\eta_T^{a_T})\}\right] = \inf_{\nu\in\mathcal{P}(S)} \left[F(\nu) + I^\infty(\nu)\right]. \tag{106}$$

By adding a constant to both sides we can assume $F \geq 0$, and do so for the rest of this section.

The proof of the uniform large deviation principle is based on the weak convergence approach. The proof is complicated by the multiscale aspect of the fast swapping process, as well as the fact that $\eta_T^{a_T}$ is a weighted empirical measure that involves this fast process.

0.1. Preliminary results.

0.1.1. *A representation.* We first state a stochastic control representation for the left hand side of (106). As with the derivation of the infinite swapping process via weak convergence, it will be necessary to work with the (distributionally equivalent) temperature swapped processes for tightness to hold. In the representation, all random variables used to construct $\eta_T^{a_T}$ are replaced by random variables whose distribution is selected, and both the distributions and the random variables will be distinguished from their uncontrolled, original counterparts by an overbar. For this reason, while the continuous time process is denoted by $\boldsymbol{Y}^a(t)$, we change notation and use $\boldsymbol{U}^a(j)$ rather than $\bar{\boldsymbol{Y}}^a(j)$ to denote the discrete time process.

We first construct the temperature swapped process. Let $\alpha(\boldsymbol{x}, d\boldsymbol{y}|0) = \alpha(\boldsymbol{x}, d\boldsymbol{y})$ and $\alpha(\boldsymbol{x}, d\boldsymbol{y}|1) = \alpha(\boldsymbol{x}^R, d\boldsymbol{y}^R)$, and let $\{N_j^a, j = 0, 1, \ldots\}$ be iid geometric random variables with parameter $1/(1+a)$, i.e. geometric random variables with mean a. Then the random variables $\{\boldsymbol{U}^a(j), j = 0, 1, \ldots\}$, $\{M_\ell^a(j), j = 0, 1, \ldots, \ell = 0, 1, \ldots, N_j^a\}$ are constructed recursively as follows. Given $M_0^a(j) = z$ and $\boldsymbol{U}^a(j) = \boldsymbol{x}$, $\boldsymbol{U}^a(j+1)$ is distributed according to $\alpha(\boldsymbol{x}, d\boldsymbol{y}|z)$. The process $M_\ell^a(j), \ell = 0, 1, \ldots, N_j^a$ is a Markov chain with states $\{0, 1\}$ and transition probabilities

$$
\begin{array}{cc}
p(0,0|\boldsymbol{x}) = g(\boldsymbol{x}) & p(0,1|\boldsymbol{x}) = 1 - g(\boldsymbol{x}) \\
p(1,0|\boldsymbol{x}) = 1 - g(\boldsymbol{x}^R) & p(1,0|\boldsymbol{x}) = g(\boldsymbol{x}^R)
\end{array}. \tag{107}
$$

The initial value for the subsequent interval is given by $M_0^a(j+1) = M_{N_j^a}^a(j)$. Letting $\{\tau_{j,\ell}^a, i = 0, 1, \ldots, \ell = 0, 1, \ldots, N_j^a - 1\}$ be iid exponential random variables with mean $1/a$, the temperature swapped process in continuous time is then given by

$$
\boldsymbol{Y}^a(t) = \boldsymbol{U}^a(j) \text{ for } \sum_{i=0}^{j-1} \sum_{\ell=0}^{N_j^a - 1} \tau_{i,\ell}^a \leq t < \sum_{i=0}^{j} \sum_{\ell=0}^{N_j^a - 1} \tau_{i,\ell}^a,
$$

(with the convention that the sum from 0 to -1 is 0),

$$Z^a(t) = M_\ell^a(j) \text{ for } \sum_{i=0}^{j-1}\sum_{k=0}^{\ell-1}\tau_{i,k}^a \leq t < \sum_{i=0}^{j-1}\sum_{k=0}^{\ell}\tau_{i,k}^a,$$

and lastly the ordinary and weighted empirical measures are given by

$$\psi_T^a = \frac{1}{T}\int_0^T \delta_{\mathbf{Y}^a(t)}dt \text{ and } \eta_T^a = \frac{1}{T}\int_0^T \left[1_{\{Z^a(t)=0\}}\delta_{\mathbf{Y}^a(t)} + 1_{\{Z^a(t)=1\}}\delta_{\mathbf{Y}^a(t)^R}\right] dt.$$

Let σ^a denote the exponential distribution with mean $1/a$ and let β^a denote the geometric distribution with mean a. For the representation, all distributions [e.g., $\alpha(\boldsymbol{x}, d\boldsymbol{y}|z)$], can be perturbed from their original form, but such a perturbation pays a relative entropy cost. We distinguish the new distributions and random variables by using an overbar. Given $T \in (0,\infty)$, let R^a and K^a be the discrete time indices when the continuous time parameter reaches T, i.e.,

$$\sum_{i=0}^{R^a-2}\sum_{k=0}^{N_i^a-1}\tau_{i,k}^a + \sum_{k=0}^{K^a-1}\tau_{R^a-1,k}^a \leq T < \sum_{i=0}^{R^a-2}\sum_{k=0}^{N_i^a-1}\tau_{i,k}^a + \sum_{k=0}^{K^a}\tau_{R^a-1,k}^a. \tag{108}$$

In this representation the barred quantities are constructed analogously to their unbarred counterparts. Thus, e.g., $\bar{R}^a$ and $\bar{N}_i^a$ are defined by (108) but with $\tau_{i,k}^a$ replaced by $\bar{\tau}_{i,k}^a$. Random variables corresponding to any given value of j are constructed in the order $\bar{\boldsymbol{U}}^a(j+1), \bar{N}_j^a, \bar{M}_\ell^a(j), \bar{\tau}_{i,\ell}^a, \ell = 0, 1, \ldots, \bar{N}_j^a$, and then j is updated to $j+1$. Barred measures, which are also allowed to depend on discrete time, are used to construct the corresponding barred random variables, e.g., $\bar{\boldsymbol{U}}^a(j+1)$ is (conditionally) distributed according to $\bar{\alpha}_j(\bar{\boldsymbol{U}}^a(j), \cdot|\bar{M}_0^a(j))$. The infimum is over all collections of measures $\{\bar{\alpha}_j, \bar{\beta}_j^a, \bar{p}_{j,\ell}, \bar{\sigma}_{j,\ell}^a\}$ and, although this is not denoted explicitly, any particular measure can depend on all previously constructed random variables. To simplify notation we let $\bar{N}_{\bar{R}^a}^a$ denote $\bar{K}^a$. We state the representation for $\{\eta_T^a\}$, and note that an analogous representation holds for $\{\psi_T^a\}$.

LEMMA 9. *Let $G : \mathcal{P}(S) \times \mathbb{N} \to \mathbb{R}$ be bounded from below and measurable. Then the representation*

(109)

$$
\begin{aligned}
&-\frac{1}{T}\log E\left[\exp\{-TG(\eta_T^a, R^a)\}\right] \\
&\quad = \inf E\left[G(\bar{\eta}_T^a, \bar{R}^a) + \frac{1}{T}\sum_{i=0}^{\bar{R}^a-1}\left[R\left(\bar{\alpha}_i(\bar{\boldsymbol{U}}^a(i), \cdot|\bar{M}_0^a(i))\,\middle\|\,\alpha(\bar{\boldsymbol{U}}^a(i), \cdot|\bar{M}_0^a(i))\right) + R\left(\bar{\beta}_i^a\,\|\beta^a\right)\right]\right. \\
&\qquad \left. + \frac{1}{T}\sum_{i=0}^{\bar{R}^a-1}\sum_{k=0}^{\bar{N}_i^a-1}\left[R\left(\bar{p}_{i,k}(\bar{M}_k^a(i), \boldsymbol{\cdot}|\bar{\boldsymbol{U}}^a(i))\,\middle\|\,p(\bar{M}_k^a(i), \boldsymbol{\cdot}|\bar{\boldsymbol{U}}^a(i))\right) + R\left(\bar{\sigma}_{i,k}^a\,\|\sigma^a\right)\right]\right]
\end{aligned}
$$

is valid.

The proof of such representations follow from the chain rule for relative entropy (see, e.g., [**7**, Section B.2]). A novel feature of the representation here is that the total number of discrete time steps is random. However, this case can easily be reduced to the case with a fixed deterministic number of steps. In fact, the key steps of the proof of this lemma follows the same routine as in the proof of Lemma 7, we omit the proof here to avoid redundancy.

0.1.2. *Rate for the ordinary empirical measure.* **Notation for marginals.** We will frequently factor measures on product spaces in the proof. For a (deterministic) probability measure ν on a product space such as $S^1 \times S^2 \times S^3$, with each S^i a Polish space, we use notation such as $\nu_{1,2}$ to denote the marginal distribution on the first 2 components, and notation such as $\nu_{1|3}$ to denote the conditional distribution on the first component given the third. When ν is a measurable random measure these can all be chosen so that they are also measurable.

We will make use of the rate function for the ordinary empirical measure. Let $\varphi(\boldsymbol{x}, d\boldsymbol{y}) = \alpha(\boldsymbol{x}, d\boldsymbol{y}|0)\rho_0(\boldsymbol{x}) + \alpha(\boldsymbol{x}, d\boldsymbol{y}|1)\rho_1(\boldsymbol{x})$, where $\rho_0(\boldsymbol{x}) = \rho(\boldsymbol{x})$ and $\rho_1(\boldsymbol{x}) = \rho(\boldsymbol{x}^R)$, and let $\bar{\mu}(d\boldsymbol{x}) = [\pi(\boldsymbol{x}) + \pi(\boldsymbol{x}^R)]d\boldsymbol{x}/2$ be its unique invariant probability distribution. If γ is absolutely continuous with respect to $\bar{\mu}$ with $\kappa(\boldsymbol{x}) = [d\gamma/d\bar{\mu}](\boldsymbol{x})$, then set

$$K(\gamma) = 1 - \int \sqrt{\kappa(\boldsymbol{x})\kappa(\boldsymbol{y})}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y}).$$

Note that K is convex. We then extend the definition to all of $\mathcal{P}(S)$ via lower semicontinuous regularization with respect to the weak topology. Thus if $\gamma_i \to \gamma$ in the weak topology and if each γ_i is absolutely continuous with respect to $\bar{\mu}$, then $\liminf_i K(\gamma_i) \geq K(\gamma)$, and we have equality for at least one such sequence. Note that since $\bar{\mu}$ is mutually absolutely continuous with respect to Lebesgue measure, this means that $K(\gamma) \leq 1$ for all $\gamma \in \mathcal{P}(S)$.

The following lemma will help relate weak limits of quantities in the representations to the rate function of the ordinary empirical measure.

LEMMA 10. *Let $\gamma \in \mathcal{P}(S)$ be absolutely continuous with respect to $\bar{\mu}$ and let $\kappa = [d\gamma/d\bar{\mu}]$. Assume $A \in (0,\infty)$, $\nu \in \mathcal{P}(S\times S)$ is such that $[\nu]_1 = [\nu]_2$, $R(\nu \| [\nu]_1 \otimes \varphi) < \infty$, r is such that $r[\nu]_1 = \kappa\bar{\mu}$, and $0 \leq -\int \log r(\boldsymbol{y})[\nu]_1(d\boldsymbol{y}) < \infty$. Then*

$$\begin{aligned} K(\gamma) &= 1 - \int \sqrt{\kappa(\boldsymbol{x})\kappa(\boldsymbol{y})}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y}) \\ &\leq AR(\nu\|[\nu]_1 \otimes \varphi) - A\int \log r(\boldsymbol{y})[\nu]_1(d\boldsymbol{y}) + A\log A - A + 1. \end{aligned} \tag{110}$$

PROOF. Let C be the set where $\kappa(\boldsymbol{x}) = 0$. Then $-\int \log r(\boldsymbol{y})[\nu]_1(d\boldsymbol{y}) < \infty$ implies $r(\boldsymbol{y}) > 0$ a.s. with respect to $[\nu]_1(d\boldsymbol{y})$, and we also have $\int r(\boldsymbol{y})[\nu]_1(d\boldsymbol{y}) = 1$, so that $r(\boldsymbol{y}) < \infty$ a.s. with respect to $[\nu]_1(d\boldsymbol{y})$. It follows that $[\nu]_1(C) = 0$. Now suppose that

$$\int \sqrt{\kappa(\boldsymbol{x})\kappa(\boldsymbol{y})}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y}) = 0.$$

Then $\kappa(\boldsymbol{x})\kappa(\boldsymbol{y}) = 0$ a.s. with respect to Lebesgue measure, and so if $\kappa(\boldsymbol{x}) \neq 0$ then $\varphi(\boldsymbol{x}, \{\boldsymbol{y} : \kappa(\boldsymbol{y}) > 0\}) = 0$, or $\varphi(\boldsymbol{x}, C) = 1$. Thus $\nu((S\backslash C) \times C) = 0$, while $[\nu]_1 \otimes \varphi((S\backslash C) \times C) = 1$, which implies $R(\nu\|[\nu]_1 \otimes \varphi) = \infty$, which is a contradiction. We conclude that

$$\int \sqrt{\kappa(\boldsymbol{x})\kappa(\boldsymbol{y})}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y}) > 0.$$

Since also

$$\int \sqrt{\kappa(\boldsymbol{x})\kappa(\boldsymbol{y})}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y}) \leq 1,$$

it follows that

$$-\log \int \sqrt{\kappa(\boldsymbol{x})\kappa(\boldsymbol{y})}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y}) \in [0,\infty).$$

Since $R(\nu\|[\nu]_1 \otimes \varphi) < \infty$ and since $\bar{\mu}$ is the invariant distribution of φ it follows that $R([\nu]_1\|\bar{\mu}) < \infty$ [**7**, Lemma 8.6.2]. This means that

$$\int \log \frac{\kappa(\boldsymbol{x})}{r(\boldsymbol{x})}[\nu]_1(d\boldsymbol{x}) < \infty,$$

and since $-\int \log r(\boldsymbol{y})[\nu]_1(d\boldsymbol{y}) < \infty$, it follows that $-\int \log \kappa(\boldsymbol{x})[\nu]_1(d\boldsymbol{x}) > -\infty$. From $[\nu]_1 = [\nu]_2$ we conclude that

$$-\frac{1}{2}\int \left[\log \kappa(\boldsymbol{x}) + \log \kappa(\boldsymbol{y})\right] \nu(d\boldsymbol{x}, d\boldsymbol{y}) > -\infty. \tag{111}$$

By relative entropy duality ([**7**, Proposition 1.4.2])

$$\begin{aligned}
-\log \int \sqrt{\kappa(\boldsymbol{x})\kappa(\boldsymbol{y})}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y}) &= -\log \int e^{\frac{1}{2}[\log \kappa(\boldsymbol{x}) + \log \kappa(\boldsymbol{y})]}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y}) \\
&\leq R(\nu\|\bar{\mu} \otimes \varphi) - \frac{1}{2}\int \left[\log \kappa(\boldsymbol{x}) + \log \kappa(\boldsymbol{y})\right] \nu(d\boldsymbol{x}, d\boldsymbol{y})
\end{aligned}$$

is valid as long as the right hand side is not of the form $\infty - \infty$, which is true by (111). The chain rule then gives

$$\begin{aligned}
&-\log \int e^{\frac{1}{2}[\log \kappa(\boldsymbol{x}) + \log \kappa(\boldsymbol{y})]}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y}) \\
&\quad\leq R(\nu\|\bar{\mu} \otimes \varphi) - \int \log \kappa(\boldsymbol{x})[\nu]_1(d\boldsymbol{x}) \\
&\quad= R(\nu\|[\nu]_1 \otimes \varphi) + R([\nu]_1\|\bar{\mu}) - \int \log \kappa(\boldsymbol{x})[\nu]_1(d\boldsymbol{x}) \\
&\quad= R(\nu\|[\nu]_1 \otimes \varphi) - \int \log r(\boldsymbol{x})[\nu]_1(d\boldsymbol{x}),
\end{aligned}$$

and thus

$$-\int \sqrt{\kappa(\boldsymbol{x})\kappa(\boldsymbol{y})}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y}) \leq -e^{-R(\nu\|[\nu]_1\otimes\varphi) + \int \log r(\boldsymbol{x})[\nu]_1(d\boldsymbol{x})}.$$

Then (110) follows from the fact that if $a \in \mathbb{R}$ and $b \in (0, \infty)$ then $-e^{-a} \leq ab + b\log b - b$ by taking $a = R(\nu\|[\nu]_1 \otimes \varphi) - \int \log r(\boldsymbol{x})[\nu]_1(d\boldsymbol{x})$ and $b = A$. $\square$

0.1.3. *Decomposition of the exponential distribution.* In the construction of η_T^a we used independent exponential random variables $\tau_{i,k}^a$ and geometric random variables N_i^a, and the fact that for each i

$$\sum_{\ell=0}^{N_i^a-1} \tau_{i,\ell}^a$$

is exponential with mean one. This decomposition corresponds to a relationship in relative entropies, which we now state.

LEMMA 11. *For $a \in (0,\infty)$, let $\bar{N}^a$ be distributed according to a random probability measure $\bar{\beta}^a$ on $\{0,1,\ldots\}$. Given $\bar{N}^a = \ell$, for $k \in \{0,1,\ldots l\}$ let $\bar{\tau}_k^a(\ell)$ be distributed according to a random probability measure $\bar{\sigma}_k^a(\ell)$ on $[0,\infty)$. Let $\bar{\sigma}$ be the distribution of the random variable*

$$\bar{\tau} = \sum_{k=0}^{\bar{N}^a-1} \bar{\tau}_k^a(\bar{N}^a).$$

Then

$$E\left[R\left(\bar{\beta}^a \left\| \beta^a\right.\right) + \sum_{k=0}^{\bar{N}^a-1} R\left(\bar{\sigma}_k^a(\bar{N}^a) \left\| \sigma^a\right.\right)\right] \geq E\left[R\left(\bar{\sigma} \left\| \sigma^1\right.\right)\right].$$

PROOF. Define a measure μ on the space $\mathbb{N}_+ \times \prod_{i=0}^{\infty}[0,\infty)$ as follows. For any $\ell \in \mathbb{N}_+$ and any sequence $A_0, A_1, \ldots$ of Borel measurable subsets of $[0,\infty)$,

$$\mu\left(\{l\} \times A_0 \times A_1 \times \ldots\right) = \beta^a\left(l\right) \prod_{k=0}^{\infty} \sigma^a\left(A_k\right),$$

and similarly define the measure $\bar{\mu}$ by

$$\bar{\mu}\left(\{l\} \times A_0 \times A_1 \times \ldots\right) = \bar{\beta}^a\left(l\right) \left(\prod_{k=0}^{\ell-1} \bar{\sigma}_k^a(\ell)\left(A_k\right)\right) \left(\prod_{k=\ell}^{\infty} \sigma^a\left(A_k\right)\right).$$

Then by the chain rule of relative entropy

$$E\left[R\left(\bar{\mu} \left\| \mu\right.\right)\right] = E\left[R\left(\bar{\beta}^a \left\| \beta^a\right.\right) + \sum_{\ell=0}^{\infty} \sum_{k=0}^{\ell-1} R\left(\bar{\sigma}_k^a(\ell) \left\| \sigma^a\right.\right) \bar{\beta}^a\left(\ell\right)\right].$$

Since

$$\begin{aligned} E\left[\sum_{k=0}^{\bar{N}^a-1} R\left(\bar{\sigma}_k^a(\bar{N}^a) \left\| \sigma^a\right.\right)\right] &= E\left[E\left[\left.\sum_{k=0}^{\bar{N}^a-1} R\left(\bar{\sigma}_k^a(\bar{N}^a) \left\| \sigma^a\right.\right)\right| \bar{N}^a\right]\right] \\ &= E\left[\sum_{\ell=0}^{\infty} \sum_{k=0}^{\ell-1} R\left(\bar{\sigma}_k^a(\ell) \left\| \sigma^a\right.\right) \bar{\beta}^a\left(\ell\right)\right], \end{aligned}$$

it follows that

$$E\left[R\left(\bar{\mu}\,\|\mu\right)\right]=E\left[R\left(\bar{\beta}^{a}\,\|\beta^{a}\right)+\sum_{k=0}^{\bar{N}^{a}-1}R\left(\bar{\sigma}_{k}^{a}(\bar{N}^{a})\,\|\sigma^{a}\right)\right].$$

Observe that $\bar{\tau}$ can be written as a measurable mapping on $\mathbb{N}_{+}\times\prod_{i=0}^{\infty}[0,\infty)$ and that $\bar{\sigma}$ and σ^{1} are the distributions induced on $[0,\infty)$ under that map by $\bar{\mu}$ and μ, respectively. Since relative entropy can only decrease under such a mapping, it follows that $ER\left(\bar{\mu}\,\|\mu\right)\geq ER\left(\bar{\sigma}\,\|\sigma^{1}\right)$. □

0.2. Lower bound. The proof of the lower bound will be partitioned into three cases according to a parameter $C\in(1,\infty)$. After the three cases have been argued, the proof of the lower bound will be completed. The first two cases are very simple, and give estimates when R^{a}/T is small (i.e., unusually few exponential clocks with mean 1 are needed before time T is reached) or when R^{a}/T is large (i.e., an unusually large number of such clocks are needed to reach tome T). The processes $\{\boldsymbol{U}^{a}(j)\}$ and $\{M_{\ell}^{a}(j)\}$ play no role in these estimates, and the required estimates follow from Chebyshev's inequality as it is used in the proof of Crámer's Theorem. We will need the function $h\left(b\right)=-\log b+b-1,\ b\in[0,\infty)$, which satisfies $\inf\left\{R\left(\gamma\,\|\sigma^{1}\right):\int u\gamma\left(du\right)=b\right\}=h\left(b\right)$.

0.2.1. *The case $R^{a}/T\leq 1/C$.* Let $F:\mathcal{P}(S)\to\mathbb{R}$ be non-negative and continuous. Then

$$\begin{aligned}
&-\frac{1}{T}\log E\left[\exp\{-TF(\eta_{T}^{a})-\infty 1_{\{[0,1/C]\}}(R^{a}/T)\}\right]\\
&\qquad=-\frac{1}{T}\log E\left[\exp\left\{-TF(\eta_{T}^{a})\right\}1_{\{[0,1/C]\}}(R^{a}/T)\right]\\
&\qquad\geq-\frac{1}{T}\log P\left\{\sum_{i=0}^{\lfloor T/C\rfloor+1}\tau_{i}^{1}\geq T\right\}.
\end{aligned}$$

By Chebyshev's inequality, for $\alpha\in(0,1)$

$$\begin{aligned}
P\left\{\sum_{i=0}^{\lfloor T/C\rfloor+1}\tau_{i}^{1}\geq T\right\}&=P\left\{e^{\alpha\sum_{i=0}^{\lfloor T/C\rfloor+1}\tau_{i}^{1}}\geq e^{\alpha T}\right\}\\
&\leq\exp\left\{\left(\lfloor T/C\rfloor+2\right)\left(\log\frac{1}{1-\alpha}-\alpha C\right)\right\}.
\end{aligned}$$

Optimizing this inequality over $\alpha \in (0,1)$ gives

$$\begin{aligned}
&\liminf_{T\to\infty} -\frac{1}{T}\log E\left[\exp\{-TF(\eta_T^a) - \infty 1_{\{[0,1/C]\}}(R^a/T)\}\right] \\
&\quad \geq \liminf_{T\to\infty} \frac{1}{T}\left(\lfloor T/C \rfloor + 2\right) h\left(C\right) \\
&\quad = \frac{h\left(C\right)}{C}.
\end{aligned}$$

Note that $h\left(C\right)/C \to 1$ as $C \to \infty$.

0.2.2. *The case* $R^a/T \geq C$.. With F as in the last section, an analogous argument gives

$$\begin{aligned}
&\liminf_{T\to\infty} -\frac{1}{T}\log E\left[\exp\{-TF(\eta_T^a) - T\infty 1_{\{[C,\infty)\}}(R^a/T)\}\right] \\
&\quad \geq \liminf_{T\to\infty} \frac{1}{T}\left(\lfloor T\left(C-1\right) \rfloor + 2\right) h\left(\frac{1}{C-1}\right) \\
&\quad = \left(C-1\right) h\left(\frac{1}{C-1}\right).
\end{aligned}$$

Note that $\left(C-1\right) h\left(1/\left(C-1\right)\right) \to \infty$ as $C \to \infty$. .

0.2.3. *The case* $1/C \leq R^a/T \leq C$.. To analyze this case it will be sufficient to consider any deterministic sequence $\{r^a\}$ such that $1/C \leq r^a/T \leq C$, and such that $r^a/T \to A$ as $T \to \infty$, and obtain lower bounds on

$$\liminf_{T\to\infty} -\frac{1}{T}\log E\left[\exp\{-TF(\eta_T^a) - \infty 1_{\{r^a/T\}}(R^a/T)\}\right].$$

The representation from Lemma 9 will be applied. We first note that the representation will include a term of the form $\infty 1_{\{r^a/T\}}(\bar{R}^a/T)$ on the right hand side. We can remove this term if we restrict the infimum to controls for which $\bar{R}^a = r^a$ w.p.1, and do so for notational convenience. The representation thus becomes

(112)

$$-\frac{1}{T}\log E\left[\exp\{-TF(\eta^a_T)-\infty 1_{\{r^a\}}(R^a)\}\right]$$
$$=\inf E\left[F(\bar\eta^a_T)+\frac{1}{T}\sum_{i=0}^{r^a-1}\left[R\left(\bar\alpha_i(\bar{\boldsymbol{U}}^a(i),\cdot|\bar M^a_0(i))\,\|\alpha(\bar{\boldsymbol{U}}^a(i),\cdot|\bar M^a_0(i))\right)+R\left(\bar\beta^a_i\,\|\beta^a\right)\right]\right.$$
$$\left.+\frac{1}{T}\sum_{i=0}^{r^a-1}\sum_{k=0}^{\bar N^a_i-1}\left[R\left(\bar p_{i,k}(\bar M^a_k(i),\boldsymbol{\cdot}|\bar{\boldsymbol{U}}^a(i))\,\|p(\bar M^a_k(i),\boldsymbol{\cdot}|\bar{\boldsymbol{U}}^a(i))\right)+R\left(\bar\sigma^a_{i,k}\,\|\sigma^a\right)\right]\right].$$

There are four relative entropy sums in (112). Since F is bounded from below, the lower bound holds vacuously unless each such term is uniformly bounded.

We first show that the empirical distribution on $\bar N^a_i$ converges to δ_∞ by using a martingale argument. For $C_1, C_2\subset[0,\infty)$, let

$$\xi^T(C_1\times C_2)\doteq\frac{1}{r^a}\sum_{i=0}^{r^a-1}\delta_{\bar N^a_i}(C_1)\bar\beta^a_i(C_2).$$

Consider the one-point compactification of $[0,\infty)$, which we identify with $[0,\infty]$, and left $f:[0,\infty]\to\mathbb{R}$ be bounded and continuous. Then for any $\varepsilon>0$

(113)
$$P\left\{\left|\int[f(z_1)-f(z_2)]\xi^T(dz_1\times dz_2)\right|\geq\varepsilon\right\}$$
$$=P\left\{\left|\frac{1}{r^a}\sum_{i=0}^{r^a-1}\left(f(\bar N^a_i)-\int f(n)\bar\beta^a_i(dn)\right)\right|\geq\varepsilon\right\}$$
$$\leq\frac{1}{\varepsilon^2}E\left[\left|\frac{1}{r^a}\sum_{i=0}^{r^a-1}\left(f(\bar N^a_i)-\int f(n)\bar\beta^a_i(dn)\right)\right|^2\right]$$
$$\leq\frac{\|f\|^2_\infty}{\varepsilon^2 r^a}$$
$$\to 0,$$

and thus any weak limit of ξ^T has identical marginals. Next note that by Jensen's inequality and since $r^a/T\to A\in(0,\infty)$, the uniform bound on the second relative entropy sum implies that $ER\left([\xi^T]_2\,\|\beta^a\right)$ is uniformly bounded. Using that

$\inf\{R(\gamma \| \beta_a) : \int u\gamma(du) = b\} = b\log\frac{b}{a} + (1+b)\log\frac{1+a}{1+b}$, is follows that the weak limit of $[\xi^T]_2 = \delta_\infty$ w.p.1.

Next we consider the asymptotic properties of the collection $\{\bar{M}_\ell^a(i)\}$, under boundedness of the third relative entropy sum. We use that the empirical measure of the $\{\bar{N}_i^a\}$ tends to δ_∞. Since $p(\cdot,\cdot|\bar{\boldsymbol{U}}^a(i))$ is the transition function of a finite state Markov chain this means that asymptotically the $\bar{M}_k^a(i), k = 0,\ldots,\bar{N}_i^a$ are samples from the transition probability (107) with $\boldsymbol{x} = \bar{\boldsymbol{U}}^a(i)$, and in particular that $\bar{M}_0^a(i+1)$ is asymptotically conditionally independent with distribution $(\rho(\boldsymbol{x}), 1-\rho(\boldsymbol{x}))$. Indeed, a martingale argument similar to (113) shows that if

$$\mu^T(C_1 \times C_2) \doteq \frac{1}{r^a}\sum_{i=0}^{r^a-1}\delta_{\bar{\boldsymbol{U}}^a(i)}(C_1)\delta_{\bar{M}_0^a(i)}(C_2),$$

and if μ^∞ is the weak limit of any convergent subsequence (which must exist by compactness), then

$$\left([\mu^\infty]_{2|1}(\{0\}\,|\boldsymbol{y}), [\mu^\infty]_{2|1}(\{1\}\,|\boldsymbol{y})\right) = (\rho(\boldsymbol{y}), 1-\rho(\boldsymbol{y}))\ \ [\mu^\infty]_1\text{-a.s.,} \tag{114}$$

and the same is true for the empirical measure of $\{\bar{M}_k^a(i), k = 0,\ldots,\bar{N}_i^a\}$.

We next remove the third relative entropy sum from the representation (112), and obtain a lower bound for the right hand side using Lemma 11. Let $\hat{\sigma}_i^a$ denote the distribution of the random variable

$$\hat{\tau}_i^a = \sum_{\ell=0}^{\bar{N}_i^a-1}\bar{\tau}_{i,\ell}^a$$

Then we have the lower bound

(115)

$$-\frac{1}{T}\log E\left[\exp\{-TF(\eta_T^a) - \infty 1_{\{r^a\}}(R^a)\}\right]$$
$$\geq \inf E\left[F(\bar{\eta}_T^a) + \frac{1}{T}\sum_{i=0}^{r^a-1}\left[R\left(\bar{\alpha}_i(\bar{\boldsymbol{U}}^a(i),\cdot|\bar{M}_0^a(i))\,\|\alpha(\bar{\boldsymbol{U}}^a(i),\cdot|\bar{M}_0^a(i))\right) + R\left(\hat{\sigma}_i^a\,\|\sigma^1\right)\right]\right].$$

To study the lower bound of (115) we introduce the measures

$$\kappa^T(C_1 \times C_2 \times C_3) \doteq \frac{1}{r^a} \sum_{i=0}^{r^a-1} \delta_{\bar{\boldsymbol{U}}^a(i)}(C_1)\bar{\alpha}_i(\bar{\boldsymbol{U}}^a(i), C_2|\bar{M}_0^a(i))\delta_{\bar{M}_0^a(i)}(C_3)$$

$$\xi^T(C_1 \times C_2) \doteq \frac{1}{T} \sum_{i=0}^{r^a-1} \delta_{\bar{\boldsymbol{U}}^a(i)}(C_1)\delta_{(\hat{m}_i^a)^{-1}}(C_2)\hat{m}_i^a,$$

where $\hat{m}_i^a$ is the conditional mean of $\hat{\tau}_i^a$ for $i = 0, \ldots, r^a - 2$, and $\hat{m}_{r^a-1}^a$ is the median of the conditional mean of $\hat{\tau}_{r^a-1}^a$, 1 and $(T - \sum_{i=0}^{r^a-2} \hat{\tau}_i^a)$. Notice that the restriction on the control measures implies

$$\sum_{i=0}^{r^a-2} \hat{\tau}_i^a \leq T < \sum_{i=i}^{r^a-1} \hat{\tau}_i^a.$$

We introduce the function $\ell : \mathbb{R}_+ \to \mathbb{R}_+$ given by $\ell(b) = b \log b - b + 1$. Note that $h(b) = b\ell(1/b)$. The sequence $\{\kappa^T\}$ is tight because S and $\{0,1\}$ are compact. Using the fact that $\hat{\sigma}_i^a$ selects the conditional distribution of $\hat{\tau}_i^a$ and $\inf\{R(\gamma \| \sigma^1) : \int u\gamma(du) = b\} = h(b)$, we have

$$\begin{aligned} E\left[\frac{1}{T}\sum_{i=0}^{r^a-1} R\left(\hat{\sigma}_i^a \| \sigma^1\right)\right] &\geq E\left[\frac{1}{T}\sum_{i=0}^{r^a-1} h(\hat{m}_i^a)\right] \\ &= E\left[\frac{1}{T}\sum_{i=0}^{r^a-1} \hat{m}_i^a \ell([\hat{m}_i^a]^{-1})\right] \\ &= E\left[\int \ell(z)\xi^T(du \times dz)\right]. \end{aligned} \tag{116}$$

Notice that for the inequality in (116) to hold, we used the fact that h is monotonously increasing in $(1, \infty)$ and $\hat{m}_{r^a-1}^a$ is the median of the conditional mean of $\hat{\tau}_{r^a-1}^a$, 1 and $(T - \sum_{i=0}^{r^a-2} \hat{\tau}_i^a)$. Hence the uniform bound on the relative entropy sum gives that $\{\xi^T\}$ is tight. Let $f : S \to \mathbb{R}$ be bounded and continuous. Since $\bar{\alpha}_i(\bar{\boldsymbol{U}}^a(i), \cdot|\bar{M}_0^a(i))$

selects the conditional distribution of $\bar{\boldsymbol{U}}^a(i+1)$, for $\varepsilon > 0$

$$\begin{aligned}
&P\left\{\left|\sum_{z=1}^{2}\int [f(\boldsymbol{y}_1) - f(\boldsymbol{y}_2)]\kappa^T(d\boldsymbol{y}_1 \times d\boldsymbol{y}_2 \times dz)\right| \geq \varepsilon\right\} \\
&\quad = P\left\{\left|\frac{1}{r^a}\sum_{i=0}^{r^a-1}\left(f(\bar{\boldsymbol{U}}^a(i+1)) - \int f(\boldsymbol{y})\bar{\alpha}_i(\bar{\boldsymbol{U}}^a(i), d\boldsymbol{y}|\bar{M}_0^a(i))\right)\right| \geq \varepsilon\right\} \\
&\quad \leq \frac{1}{\varepsilon^2}E\left[\left|\frac{1}{r^a}\sum_{i=0}^{r^a-1}\left(f(\bar{\boldsymbol{U}}^a(i+1)) - \int f(\boldsymbol{y})\bar{\alpha}_i(\bar{\boldsymbol{U}}^a(i), d\boldsymbol{y}|\bar{M}_0^a(i))\right)\right|^2\right] \\
&\quad \leq \frac{\|f\|_\infty^2}{\varepsilon^2 r^a} \\
&\quad \to 0.
\end{aligned}$$

We find that

$$[\kappa^\infty]_1 = [\kappa^\infty]_2 \text{ w.p.1.} \tag{117}$$

Using Fatou's lemma and the definition of φ, we get the lower bound $AR([\kappa^\infty]_{1,2}\|[\kappa^\infty]_1 \otimes \varphi)$ on the weak limit of the corresponding relative entropies (the second sum in (115)).

With regard to ξ^∞, comparing the form of $\left[\xi^T\right]_1$ and $\bar{\psi}_T^a$ and using the martingale argument used for (113) and (117) gives

$$[\xi^\infty]_1 = \bar{\psi}_\infty.$$

Observe that

$$\int z\xi^T(d\boldsymbol{x} \times dz) = \frac{r^a}{T}\left[\kappa^T\right]_1(d\boldsymbol{x}).$$

Because of the superlinearity of ℓ we have uniform integrability, and thus passing to the limit gives

$$\int z[\xi^\infty]_1(d\boldsymbol{x})[\xi^\infty]_{2|1}(dz|\boldsymbol{x}) = A[\kappa^\infty]_1(d\boldsymbol{x}).$$

Hence with the definition $b(\boldsymbol{x}) = \int z[\xi^\infty]_{2|1}(dz|\boldsymbol{x})$, $[d[\kappa^\infty]_1/d[\xi^\infty]_1](\boldsymbol{x}) = b(\boldsymbol{x})/A$. Using Fatou's lemma, Jensen's inequality and the weak convergence, we get the following lower bound on the corresponding relative entropies (the first sum in (115)):

$$\begin{aligned}
\liminf_{T\to\infty} &\int \ell(z)\xi^T(du\times dz) \\
&\geq \int \ell(z)\xi^\infty(du\times dz) \\
&\geq \int \ell\left(\int z[\xi^\infty]_{2|1}(dz|\boldsymbol{x})\right)[\xi^\infty]_1(d\boldsymbol{x}) \\
&= \int \ell\left(b(\boldsymbol{x})\right)[\xi^\infty]_1(d\boldsymbol{x}) \\
&= A\int h(1/b(\boldsymbol{x}))[\kappa^\infty]_1(d\boldsymbol{x}).
\end{aligned}$$

Let $r(\boldsymbol{y}) = A/b\left(\boldsymbol{y}\right)$. Then we can write the combined lower bound on the relative entropies as

$$\begin{aligned}
&AR([\kappa^\infty]_{1,2}\|[\kappa^\infty]_1\otimes\varphi) + A\int h(1/b(\boldsymbol{x}))[\kappa^\infty]_1(d\boldsymbol{x}) \\
&= AR([\kappa^\infty]_{1,2}\|[\kappa^\infty]_1\otimes\varphi) - A\int \log r(\boldsymbol{y})[\kappa^\infty]_1(d\boldsymbol{y}) + A\log A - A + 1. \qquad (118)
\end{aligned}$$

We next consider $\bar{\psi}_T^a$ and $\bar{\eta}_T^a$. Using the limiting properties of the $\bar{M}_k^a(i)$, we have

$$\begin{aligned}
\bar{\eta}_T^a(C) &= \frac{1}{T}\int_0^T \left[1_{\left\{\bar{Z}^a(t)=0\right\}}\delta_{\bar{\boldsymbol{Y}}^a(t)}(C) + 1_{\left\{\bar{Z}^a(t)=1\right\}}\delta_{\bar{\boldsymbol{Y}}^a(t)^R}(C)\right]dt \\
&\to \int \left[\rho(\boldsymbol{y})\delta_{\boldsymbol{y}}(C) + (1-\rho(\boldsymbol{y}))\delta_{\boldsymbol{y}^R}(C)\right][\xi^\infty]_1 \\
&= \bar{\eta}_\infty(C)
\end{aligned}$$

and

$$\bar{\psi}_T^a(C) = \frac{1}{T}\int_0^T \delta_{\bar{\boldsymbol{Y}}^a(t)}(C)dt \to \int_C [\xi^\infty]_1 = \bar{\psi}_\infty(C).$$

Note that this implies the relation

$$\bar{\eta}_\infty(C) = \int_C \left[\rho(\boldsymbol{y})\bar{\psi}_\infty(d\boldsymbol{y}) + \rho(\boldsymbol{y}^R)\bar{\psi}_\infty(d\boldsymbol{y}^R)\right]. \qquad (119)$$

Finally we consider the weighted empirical measure. By (118) we have the lower bound $E\left[F(\bar{\eta}_\infty) + K(\bar{\psi}_\infty)\right]$ for the limit inferior of the right hand side of (112), where $\bar{\eta}_\infty$ and $\bar{\psi}_\infty$ are related by (119). Thus we need only show that

$$K(\bar{\psi}_\infty) \geq I^0(\bar{\eta}_\infty) = I^\infty(\bar{\eta}_\infty).$$

The equality follows from the definition of I^{∞} and (119). Let $a(\boldsymbol{y}) = [d\bar{\psi}_{\infty}/d\bar{\mu}](\boldsymbol{y})$. Then

$$K(\bar{\psi}_{\infty}) = 1 - \int \sqrt{a(\boldsymbol{x})a(\boldsymbol{y})}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y})$$

and

$$I^{0}(\bar{\eta}_{\infty}) = 1 - \int \frac{1}{2}\sqrt{\left(a(\boldsymbol{x}) + a(\boldsymbol{x}^{R})\right)\left(a(\boldsymbol{y}) + a(\boldsymbol{y}^{R})\right)}\mu(d\boldsymbol{x})\alpha(\boldsymbol{x}, d\boldsymbol{y}).$$

Using that $\rho(\boldsymbol{x}) = 1 - \rho(\boldsymbol{x}^{R})$ and symmetry

$$\begin{aligned}
&\int \sqrt{a(\boldsymbol{x})a(\boldsymbol{y})}\frac{1}{2}\left[\mu(d\boldsymbol{x}) + \mu(d\boldsymbol{x}^{R})\right]\left(\rho(\boldsymbol{x})\alpha(\boldsymbol{x}, d\boldsymbol{y}) + \rho(\boldsymbol{x}^{R})\alpha(\boldsymbol{x}^{R}, d\boldsymbol{y}^{R})\right) \\
&= \frac{1}{2}\int \left(\sqrt{a(\boldsymbol{x})a(\boldsymbol{y})} + \sqrt{a(\boldsymbol{x}^{R})a(\boldsymbol{y}^{R})}\right)\mu(d\boldsymbol{x})\left(\rho(\boldsymbol{x})\alpha(\boldsymbol{x}, d\boldsymbol{y}) + \rho(\boldsymbol{x}^{R})\alpha(\boldsymbol{x}^{R}, d\boldsymbol{y}^{R})\right) \\
&= \frac{1}{2}\int \left(\sqrt{a(\boldsymbol{x})a(\boldsymbol{y})} + \sqrt{a(\boldsymbol{x}^{R})a(\boldsymbol{y}^{R})}\right)\mu(d\boldsymbol{x})\alpha(\boldsymbol{x}, d\boldsymbol{y}).
\end{aligned}$$

We now use

$$\sqrt{a(\boldsymbol{x})a(\boldsymbol{y})} + \sqrt{a(\boldsymbol{x}^{R})a(\boldsymbol{y}^{R})} \leq \sqrt{\left(a(\boldsymbol{x}) + a(\boldsymbol{x}^{R})\right)\left(a(\boldsymbol{y}) + a(\boldsymbol{y}^{R})\right)}$$

to obtain $K(\bar{\psi}_{\infty}) \geq I^{0}(\bar{\eta}_{\infty})$. [We note for later use that given $\bar{\eta}$ that is absolutely continuous with respect to Lebesgue measure and such that $I^{\infty}(\bar{\eta}) < \infty$, $K(\bar{\psi}) = I^{0}(\bar{\eta})$ can be shown for a $\bar{\psi}$ that maps to $\bar{\eta}$ by taking $\bar{\psi} = [\bar{\eta} + \bar{\eta}^{R}]/2$.]

0.2.4. *Combining the cases.* In the last subsection we showed that for any sequence r^{a} such that $r^{a}/T \to A \in [1/C, C]$,

$$\begin{aligned}
\liminf_{T\to\infty} -\frac{1}{T}\log E\left[\exp\{-TF(\eta_{T}^{a}) - \infty 1_{\{r^{a}/T\}}(R^{a}/T)\}\right] &\geq E\left[F(\bar{\eta}_{\infty}) + I^{\infty}(\bar{\eta}_{\infty})\right] \\
&\geq \inf_{\nu\in\mathcal{P}(S)}\left[F(\nu) + I^{\infty}(\nu)\right],
\end{aligned}$$

and an argument by contradiction shows that the bound is uniform in A. Thus

$$\begin{aligned}
&\liminf_{T\to\infty} -\frac{1}{T}\log\left\{\sum_{r^{a}=\lceil\frac{T}{C}\rceil}^{\lfloor TC\rfloor} E\left[\exp\{-TF(\eta_{T}^{a}) - \infty 1_{\{r^{a}/T\}}(R^{a}/T)\}\right]\right\} \\
&\qquad \geq \liminf_{T\to\infty} -\frac{1}{T}\log\left\{TC\cdot\bigvee_{r^{a}=\lceil\frac{T}{C}\rceil}^{\lfloor TC\rfloor} E\left[\exp\{-TF(\eta_{T}^{a}) - \infty 1_{\{r^{a}/T\}}(R^{a}/T)\}\right]\right\} \\
&\qquad \geq \inf_{\nu\in\mathcal{P}(S)}\left[F(\nu) + I^{\infty}(\nu)\right].
\end{aligned}$$

We now partition $E\left[\exp\{-TF(\eta_T^a)\right]$ according to the various cases to obtain the overall lower bound

$$\liminf_{T\to\infty} -\frac{1}{T}\log E\left[\exp\{-TF(\eta_T^{a_T})\}\right]$$
$$\geq \min\left\{\inf_{\nu\in\mathcal{P}(S)}\left[F(\nu)+I^\infty(\nu)\right], \frac{h\left(C\right)}{C}, \left(C-1\right)h\left(\frac{1}{C-1}\right)\right\}.$$

Letting $C\to\infty$ and using the fact that $I^\infty\leq 1$, we have the desired lower bound

$$\liminf_{T\to\infty} -\frac{1}{T}\log E\left[\exp\{-TF(\eta_T^{a_T})\}\right] \geq \inf_{\nu\in\mathcal{P}(S)}\left[F(\nu)+I^\infty(\nu)\right].$$

0.3. Upper bound. The proof of the reverse inequality

$$\limsup_{T\to\infty} -\frac{1}{T}\log E\left[\exp\{-TF(\eta_T^{a_T})\}\right] \leq \inf_{\nu\in\mathcal{P}(S)}\left[F(\nu)+I^\infty(\nu)\right] \tag{120}$$

is simpler. Let bounded and continuous F be given. Given $\varepsilon>0$, we can find ν that is absolutely continuous with respect to Lebesgue measure and for which

$$\left[F(\nu)+I^\infty(\nu)\right] \leq \inf_{\nu\in\mathcal{P}(S)}\left[F(\nu)+I^\infty(\nu)\right]+\varepsilon.$$

Next we use that I^∞ is convex and $I^\infty(\mu)=0$ to find $\tau>0$ such that

$$\left[F(\nu^\tau)+I^\infty(\nu^\tau)\right] \leq \left[F(\nu)+I^\infty(\nu)\right]+\varepsilon$$

when $\nu^\tau=\tau\mu+(1-\tau)\nu$. Note that if $\theta(\boldsymbol{x})=[d\nu/d\mu](\boldsymbol{x})$ and $\theta^\tau(\boldsymbol{x})=[d\nu^\tau/d\mu](\boldsymbol{x})$, then $\theta^\tau(\boldsymbol{x})\geq\tau>0$ and so $I^\infty(\nu^\tau)<1$.

We will construct a control to use in the representation such $\bar{\eta}_T^{a_T}$ will converge w.p.1 to ν^τ and RE^T will converge w.p.1 to $I^\infty(\nu^\tau)$. These convergences will follow from the ergodic theorem, and the bound $\theta^\tau(\boldsymbol{x})\geq\tau$ will be used to argue that the ergodicity of the original process is inherited by the controlled process.

We now proceed to the construction. Let $\varsigma(d\boldsymbol{x})=[\nu^\tau(d\boldsymbol{x})+\nu^\tau(d\boldsymbol{x}^R)]/2$, so that $I^\infty(\nu^\tau)=K(\varsigma)$. Let $\kappa(\boldsymbol{x})=[d\varsigma/d\bar{\mu}](\boldsymbol{x})\geq\tau/2$. Then

$$K(\varsigma)=1-\int\sqrt{\kappa(\boldsymbol{x})\kappa(\boldsymbol{y})}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x},d\boldsymbol{y}).$$

Since κ is uniformly bounded from below we have the dual relationship

$$
\begin{aligned}
-\log \int \sqrt{\kappa(\boldsymbol{x})\kappa(\boldsymbol{y})}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y}) &= -\log \int e^{\frac{1}{2}[\log\kappa(\boldsymbol{x})+\log\kappa(\boldsymbol{y})]}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y}) \\
&= R(\eta\|\bar{\mu}\otimes\varphi) - \frac{1}{2}\int [\log\kappa(\boldsymbol{x}) + \log\kappa(\boldsymbol{y})]\,\nu(d\boldsymbol{x}, d\boldsymbol{y}),
\end{aligned}
$$

where

$$
\eta(C) = \frac{\int_C e^{\frac{1}{2}[\log\kappa(\boldsymbol{x})+\log\kappa(\boldsymbol{y})]}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y})}{\int_{S^2} e^{\frac{1}{2}[\log\kappa(\boldsymbol{x})+\log\kappa(\boldsymbol{y})]}\bar{\mu}(d\boldsymbol{x})\varphi(\boldsymbol{x}, d\boldsymbol{y})}.
$$

Since $\sqrt{\kappa(\boldsymbol{x})\kappa(\boldsymbol{y})}$ is symmetric in $\boldsymbol{x}$ and $\boldsymbol{y}$ automatically $[\eta]_1 = [\eta]_2$, and using the bound $\kappa(\boldsymbol{x}) \geq \tau/2$ we can factor $\eta(d\boldsymbol{x}\times d\boldsymbol{y}) = [\eta]_1(d\boldsymbol{x})\bar{\varphi}(\boldsymbol{x}, d\boldsymbol{y})$, where $\bar{\varphi}$ is the transition kernel of a uniformly ergodic Markov chain. Let

$$
\bar{\alpha}(\boldsymbol{x}, d\boldsymbol{y}|0) = \bar{\alpha}(\boldsymbol{x}, d\boldsymbol{y}|1) = \bar{\varphi}(\boldsymbol{x}, d\boldsymbol{y}).
$$

Notice that from the definition of $\bar{\varphi}$, $\bar{\varphi}(\boldsymbol{x}, d\boldsymbol{y}) = \bar{\varphi}(\boldsymbol{x}^R, d\boldsymbol{y}^R)$, hence $\bar{\alpha}$ has the property that

$$
\bar{\alpha}(\boldsymbol{x}^R, d\boldsymbol{y}^R|0) = \bar{\alpha}(\boldsymbol{x}, d\boldsymbol{y}|1)
$$

These will be the transition kernels used to construct the $\bar{\boldsymbol{U}}^a(i)$.

Now of course the invariant distribution of $\bar{\varphi}$ is $[\eta]_1$, and not the desired distribution ς. Let $r(\boldsymbol{x}) = [d\varsigma/d[\eta]_1](\boldsymbol{x})$. Then $r(\boldsymbol{x})$ identifies the way in which the distribution of the random variables $\bar{N}_j^a$ should be modified so that in the continuous time the empirical measure $[\eta]_1$ is reshaped into ς. Choose A so that equality holds in (110), and set $b(\boldsymbol{x}) = Ar(\boldsymbol{x})$. Then $\bar{\beta}_i^a$ is chosen to be $\beta^{ab(\boldsymbol{x})}$. Consistent with the analysis of the upper bound, we do not perturb the distribution of the other variables, so that $\bar{p}_{i,k}(\cdot, \bullet|\bullet) = p(\cdot, \bullet|\bullet)$ and $\bar{\sigma}_{j,\ell}^a = \sigma^a$. We then construct the controlled processes using these measures in exact analogy with the construction of the original process.

With this choice and Lemma 9 we obtain the bound

$$
\begin{aligned}
&-\frac{1}{T}\log E\left[\exp\{-TF(\eta_T^{a_T})\}\right] \\
&\leq E\left[F(\bar{\eta}_T^a) + \frac{1}{T}\sum_{i=0}^{\bar{R}^a-1}\left[R\left(\bar{\alpha}(\bar{\boldsymbol{U}}^a(i), \cdot|\bar{M}_0^a(i))\,\|\alpha(\bar{\boldsymbol{U}}^a(i), \cdot|\bar{M}_0^a(i))\right) + R\left(\beta^{ab(\bar{\boldsymbol{U}}^a(i))}\,\|\beta^a\right)\right]\right].
\end{aligned}
$$

Now apply the ergodic theorem, and use that w.p.1 $A^T = \bar{R}^a/T \to 1/\int b(\boldsymbol{x})[\eta]_1(d\boldsymbol{x}) = A$, and also that asymptotically the conditional distribution of $\bar{M}_0^a(i)$ is given by $(\rho_0(\bar{\boldsymbol{U}}^a(i)), \rho_1(\bar{\boldsymbol{U}}^a(i)))$. Then right hand side of the last display converges to we have the limit

$$
\begin{aligned}
&F(\nu^\tau) + \frac{1}{\int b(\boldsymbol{x})[\eta]_1(d\boldsymbol{x})} \int \left[\sum_{z=1}^{2} R\left(\bar{\alpha}(\boldsymbol{x}, d\boldsymbol{y}|z) \,\|\alpha(\boldsymbol{x}, d\boldsymbol{y}|z)\right) \rho_z(\boldsymbol{x}) \right] r(\boldsymbol{x})[\eta]_1(d\boldsymbol{x}) \\
&\quad + \frac{1}{\int b(\boldsymbol{x})[\eta]_1(d\boldsymbol{x})} \int \left(-\log b(\boldsymbol{x}) + b(\boldsymbol{x}) - 1\right) b(\boldsymbol{x})[\eta]_1(d\boldsymbol{x}) \\
&= F(\nu^\tau) + A \int R\left(\bar{\varphi}(\boldsymbol{x}, d\boldsymbol{y}) \,\|\varphi(\boldsymbol{x}, d\boldsymbol{y})\right) \varsigma(d\boldsymbol{x}) - A \int \log r(\boldsymbol{x}) \varsigma(d\boldsymbol{x}) + A \log A - A + 1 \\
&= F(\nu^\tau) + K(\varsigma) \\
&= F(\nu^\tau) + I^\infty(\nu^\tau).
\end{aligned}
$$

This completes the proof of (120) and also the proof of Theorem 7.

Bibliography

[1] L. Breiman. *Probability Theory.* Addison-Wesley, Reading, Mass., 1968.

[2] M.K. Cowles and B.P. Carlin. Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91:883–904, 1996.

[3] P. Diaconis and D. Stroock. Geometric bounds for eigenvalues of Markov chains. *Ann. Appl. Prob.*, 1:36–61, 1991.

[4] J.D. Doll, N. Plattner, D.L. Freeman, Y. Liu, and P. Dupuis. Rare-event sampling: occupation-based performance measures for parallel tempering and infinite swapping Monte Carlo methods. *J. Chem. Phys.*, 137:204112, 2012.

[5] M.D. Donsker and S.R.S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, I. *Comm. Pure Appl. Math.*, 28:1–47, 1975.

[6] M.D. Donsker and S.R.S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, III. *Comm. Pure Appl. Math.*, 29:389–461, 1976.

[7] P. Dupuis and R. S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations.* John Wiley & Sons, New York, 1997.

[8] P. Dupuis, Y. Liu, N. Plattner, and J.D. Doll. On the infinite swapping limit for parallel tempering. *Multiscale Model. Simul.*, 10:986–1022, 2012.

[9] D.J. Earl and M.W. Deem. Parallel tempering: theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.*, 7:3910–3916, 2005.

[10] R.S. Ellis and A.D. Wyner. Uniform large deviation property of the empirical process of a Markov chain. *Ann. Probab.*, 17:1147–1151, 1989.

[11] S.N. Ethier and T.G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley, New York, 1986.

[12] G.B. Folland. *Real Analysis: Modern Techniques and Their Applications.* John Wiley & Sons, New York, 1999.

[13] 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. American Statistical Association, New York, 1991.

[14] C.J. Geyer. Practical Markov chain Monte Carlo. *Statistical Science*, 7:473–511, 1992.

[15] H. J. Kushner. *Approximation and Weak Convergence Methods for Random Processes with Applications to Stochastic System Theory.* MIT Press, Cambridge, MA, 1984.

[16] N. Plattner, J.D. Doll, P. Dupuis, H. Wang, Y. Liu, and J.E. Gubernatis. An infinite swapping approach to the rare-event sampling problem. *J. Chem. Phys.*, 135:134111, 2011.

[17] J. Rosenthal. Convergence rates for Markov chains. *SIAM Review*, 37:387–405, 1995.

[18] P.J. Rossky, J.D. Doll, and H.L. Friedman. Brownian dynamics as smart Monte Carlo. *J. Chem. Phys.*, 69:4628–4633, 1978.

[19] Walter Rudin. *Functional Analysis.* McGraw-Hill, New York, 1991.

[20] Y. Sugita and Y. Okamoto. The incomplete beta function law for parallel tempering sampling of classical canonical systems. *Chem. Phys. Lett.*, 314:141–151, 1999.

[21] R.H. Swendsen and J.S. Wang. Replica Monte Carlo simulation of spin glasses. *Phys. Rev. Lett.*, 57:2607–2609, 1986.

[22] D.J. Wales. *Energy Landscapes with Applications of Clusters, Biomolecules and Glasses.* Cambridge University Press, Cambridge, 2003.

[23] Wai Kong Yuen. Applications of geometric bounds to the convergence rate of Marko chains on $\mathbb{R}^n$. *Stochastic Processes and their Applications*, 87:1–23, 2000.