

Analysis of an Interacting Particle Method for Rare Event Estimation

by

Yi Cai

M.S., Brown University, Providence, RI, 2007

B.S., Peking University, Beijing, China 2006

B.A., Peking University, Beijing, China 2006

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 2012

Abstract of “Analysis of An Interacting Particle Method for Rare Event Estimation”
by Yi Cai, Ph.D., Brown University, May 2012

This thesis is a large deviations study for the performance of an interacting particle method for rare event estimation. The analysis is restricted to a one-dimensional setting, though even in this restricted setting a number of new techniques must be developed. In contrast to the large deviations analysis of related algorithms, for interacting particle schemes it is an occupation measure analysis that is relevant, and within this framework many standard assumptions (stationarity, Feller property) can no longer be assumed. The methods developed are not limited to the question of performance analysis, and in fact give the full large deviations principle for such systems.

© Copyright 2012 by Yi Cai

This dissertation by Yi Cai 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, Ph.D., Advisor

Recommended to the Graduate Council

Date_____

Hongjie Dong, Ph.D., Reader

Date_____

Kavita Ramanan, Ph.D., Reader

Approved by the Graduate Council

Date_____

Peter Weber, Dean of the Graduate School

To My Parents

The Vita of Yi Cai

Yi Cai was born in Nanchang, Jiangxi province, China on January 8, 1985.

In 2002, she started her undergraduate study at Peking University, and received both a Bachelor of Science degree in Theoretical Mechanics and a Bachelor of Arts degree in Economics four years later.

She came to the United States in 2006 and has been attending the Ph.D. program in the Division of Applied Mathematics at Brown University. During the Ph.D. program, she received a Master of Science degree in Applied Mathematics in 2007. This dissertation was defended in April 27, 2012.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor Prof. Paul Dupuis. This dissertation would have not been possible without his support, understanding and encouragement. Coming from a non-mathematical background, I started out with not much knowledge in large deviations and few experience in scientific research. Prof Dupuis, with unsurpassed patience, helped me overcome those challenges and pointed me to the right direction. His rigorous attitude towards scientific research have been a tremendous influence on me and shaped me profoundly. I am very fortunate to have him as a mentor.

I would also like to thank my other committee members Prof. Hongjie Dong and Prof. Kavita Ramanan for their helpful suggestion and comments during the completion of this dissertation.

I am grateful to the Division of Mathematics and graduate school of Brown University for generous financial support in the past six years. I also want to thank all of my friends in Brown for helping me in every aspect. My special appreciation goes to Jingjing Wang for her friendship and support, which helped me overcome many crisis situations.

Last but not least, I would like to thank my parents for their unconditional love. They have always been supportive of everything of me even though they are at the other side of the world.

Contents

Acknowledgments	vi
0 Introduction and Preliminaries	1
0.1 Monte Carlo Simulation and Asymptotic Optimality	1
0.2 Beyond Monte Carlo Simulation	4
0.3 Preliminaries	10
I Problem Formulation and Main Results	16
1 Notation and Problem Formulation	17
1.1 Process Model and the Event of Interest	17
1.2 Interacting Particle System (IPS) Estimator	19
1.3 Reformulation as an Occupation Measure Problem	22
1.4 First Entrance Distributions and the Feller Property	23
2 Statement of the Main Results	28
II Large Deviation Analysis	30
3 Relative Entropy Representation	31
3.1 The Representation Formula	31
3.2 The Limiting Behavior of the Underlying Transportation Kernel . . .	37

4	The Large Deviation Upper Bound	42
4.1	Notation and Preliminaries	42
4.2	Tightness and Characterization of Limits	46
4.3	Proof of the Lower Bound (2.2)	57
5	The Large Deviation Lower Bound	60
5.1	Some Important Assumptions and Ergodicity of the Limit Underlying Transition Kernel	60
5.2	The Construction of Ergodic IPS Transition Kernels	63
5.3	The Construction of Ergodic Controls	70
5.4	Proof of the Upper Bound (2.3)	77
5.5	Appendix	85
6	Asymptotics of the Decay Rate as $N \rightarrow \infty$	95
6.1	Proof of Theorem 6.1	98
III	Numerical Results	101
7	Computational Results	102
	Bibliography	113
	Index	116

List of Figures

7.1	large deviation rate of the second moment of the interacting particle system estimator for different m and N	104
7.2	trade-off between N and K with fixed effort, $N \times K = 1,000,000$, m $= 10$	106
7.3	trade-off between N and K with fixed effort, $N \times K = 1,000,000$, m $= 50$	107
7.4	trade-off between N and K with fixed effort, $N \times K = 1,000,000$, m $= 10$	108
7.5	sample error, $m = 10$, fixed K	109
7.6	sample error, $m = 50$, fixed K	109
7.7	standard error, $m = 10$, fixed K	110
7.8	standard error, $m = 50$, fixed K	110
7.9	standard error, $m = 10$, fixed N	111
7.10	decay rate in N , $m = 10$, fixed K	111
7.11	decay rate in N , $m = 50$, fixed K	112

Chapter 0

Introduction and Preliminaries

0.1 Monte Carlo Simulation and Asymptotic Optimality

A basic and common technique for the estimation of probabilities and functionals of probability measures is the standard Monte Carlo simulation. Let X be a random variable taking values in $\mathbb{R}$ distributed according to $\mu \in \mathcal{P}(\mathbb{R})$. To estimate the expectation of X , denoted by $\mathbb{E}X$, one can generate K number of *independently and identically distributed (iid)* samples from μ , say $X_0, X_1, \dots, X_{K-1}$, and the estimate for $\mathbb{E}X$ is just the sample mean:

$$Q_K \doteq \frac{1}{K} \sum_{i=0}^{K-1} X_i.$$

It is easy to check that this estimator is unbiased. Furthermore, Law of Large Numbers (LLN) ensures that this estimate converges to the correct value as the number of samples K goes to ∞ . The sample size K can be chosen by some predetermined precision requirement.

However, when the events of interest are of small probabilities (termed *rare*

events), the standard Monte Carlo simulation will require an enormous number of samples for the variance of the resulting estimate to be comparable to the unknown probability. To be precise, we introduce the large deviation methodology.

We begin by giving the definition of rate function and large deviation principle. Let $\mathcal{X}$ be a complete separable metric space, i.e. a *Polish space*.

Definition 0.1 (Rate Function). *A function I on $\mathcal{X}$ is called a rate function on $\mathcal{X}$, or simply a rate function, if I maps $\mathcal{X}$ into $[0, \infty]$ and if for each $M < \infty$ the level set $\{x \in \mathcal{X} : I(x) \leq M\}$ is a compact subset of $\mathcal{X}$.*

Let $\{X^m, m \in \mathbb{N}\}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in a Polish space $\mathcal{X}$.

Definition 0.2 (Large deviation Principle). *Let I be a rate function on $\mathcal{X}$. The sequence $\{X^m\}$ is said to satisfy the large deviation principle on $\mathcal{X}$ with rate function I if the following two conditions hold.*

(a) *Large deviation upper bound. For each closed subset F of $\mathcal{X}$*

$$\limsup_{m \rightarrow \infty} \frac{1}{m} \log \mathbb{P}\{X^m \in F\} \leq - \inf_{x \in F} I(x).$$

(b) *Large deviation lower bound. For each open subset G of $\mathcal{X}$*

$$\liminf_{m \rightarrow \infty} \frac{1}{m} \log \mathbb{P}\{X^m \in G\} \geq - \inf_{x \in G} I(x).$$

With the definition of large deviation principle in hand, we now give a concrete example of the ill-performance of standard Monte Carlo for rare events simulation, but also keep in mind that the issues hold in general.

Let $Y^0, Y^1, \dots$ be a sequence of iid $\mathbb{R}^d$ -valued random variables with distribution μ , and assume the moment generating function $\int_{\mathbb{R}^d} \exp\langle \alpha, y \rangle \mu(dy)$ is finite for all $\alpha \in \mathbb{R}^d$. Let $S^m = Y^0 + Y^1 + \dots + Y^{m-1}$. For Borel sets $A \subset \mathbb{R}^d$, Cramér's

Theorem [6, Theorem 2.2.3] guarantees a large deviation principle for S^m/m . Let $p^m \doteq \mathbb{P}(S^m/m \in A)$. We also define the *indicator function* by

$$\mathbf{1}_A(x) = \begin{cases} 1 & x \in A \\ 0 & \text{otherwise} \end{cases}.$$

Now let's consider the Monte Carlo approximation for p^m for some fixed value m . In terms of the notation introduced previously, we have $X = \mathbf{1}_A(S^m/m)$, the straightforward Monte Carlo would require the generation of many iid copies of X (i.e., of S^m). The variance of Q_K is given by

$$\text{var}[Q_K] = \text{var}\left[\frac{1}{K} \sum_{i=0}^{K-1} \mathbf{1}_A\left(\frac{S_i^m}{m}\right)\right] = \frac{p^m(1-p^m)}{K}.$$

If $p^m \rightarrow 0$ as $m \rightarrow \infty$, this quantity approaches zero. However, the relative error is

$$\text{relative error} \doteq \frac{\text{standard deviation of } Q_K}{\text{mean of } Q_K} = \frac{\sqrt{p^m(1-p^m)}}{\sqrt{K}p^m}.$$

Since $\sqrt{p^m(1-p^m)}/p^m \rightarrow \infty$, a large sample size (i.e., K) is required for the estimator to achieve a reasonable relative error bound (at least when m is large). In fact, since p^m scales according to a large deviation principle, K must grow exponentially in m if a bounded relative error is to be maintained. As a result, standard Monte Carlo simulation is rarely used to estimate rare event probabilities and more efficient alternatives are explored.

Before introducing other more “efficient” schemes, a natural question to ask is: What is a mathematical criteria for “efficient”? The above discussion suggests a solution: for fixed m , the faster the variance decays, the smaller sample size K is required. We define those schemes that achieve the “fastest” rate of convergence as asymptotically optimal. The formal mathematical definition is formulated as following.

Asymptotic Optimality

Assume random variables $\{X^m, m = 0, 1, \dots\}$ satisfies a large deviation limit:

$$\lim_{m \rightarrow \infty} \frac{1}{m} \log \mathbb{E} X^m = -\rho,$$

for some $\rho \geq 0$. In order to estimate $\mathbb{E} X^m$, an alternative random variable $\bar{X}^m$ is constructed such that $\mathbb{E} X^m = \mathbb{E} \bar{X}^m$. Similarly to standard Monte Carlo, one generates iid replications $(\bar{X}_0^m, \dots, \bar{X}_{K-1}^m)$ of $\bar{X}^m$ and obtain an unbiased estimator by averaging:

$$\bar{Q}_K^m \doteq \frac{1}{K} \sum_{i=0}^{K-1} \bar{X}_i^m.$$

The rate of convergence associated with this estimator is determined by the variance of the summands. Since unbiasedness implies that the first moment is fixed, it suffices to consider the second moment $\mathbb{E}(\bar{X}_0^m)^2$. However, it follows from Jensen's inequality that

$$\limsup_{m \rightarrow \infty} -\frac{1}{m} \log \mathbb{E} (\bar{X}_0^m)^2 \leq \lim_{m \rightarrow \infty} -\frac{1}{m} \log (\mathbb{E} \bar{X}_0^m)^2 = 2\rho. \quad (0.1)$$

The estimator $\bar{Q}_K^m$ is said to be *asymptotically optimal* if this upper bound is achieved, i.e. if

$$\lim_{m \rightarrow \infty} -\frac{1}{m} \log \mathbb{E} (\bar{X}_0^m)^2 = 2\rho.$$

This thesis studies the performance of one of those more efficient schemes for rare event simulation, namely, interacting particle scheme. The above concept will be the focus of the analysis. We will come back to the definition of asymptotic optimality in Chapter 1.

0.2 Beyond Monte Carlo Simulation

The two most widely considered alternatives to the standard Monte Carlo simulation are those based on change-of-measure techniques and those based on branching

processes. The schemes that this dissertation is interested in belongs to the second category.

Importance Sampling Schemes

The first alternative is usually referred as *importance sampling*. Suppose that the probability of interest takes the form $p = \mathbb{P}(X \in A) = \mu(A)$, where A is a subset of some reasonably regular space and μ a probability measure. Contrast to ordinary Monte Carlo which samples directly from μ , importance sampling uses an alternative sampling distribution ν , generates iid samples $\bar{X}_i$ from ν , and then estimates via the sample mean of $[d\mu/d\nu](\bar{X}_i)\mathbf{1}_{\{\bar{X}_i \in A\}}$. The Radon-Nikodym derivative $[d\mu/d\nu](\bar{X}_i)$ guarantees that the estimate is unbiased. The goal is to choose ν so that the individual samples $[d\mu/d\nu](\bar{X}_i)\mathbf{1}_{\{\bar{X}_i \in A\}}$ cluster tightly around p , thereby reducing the variance. The characterization of optimal achievable performance of importance sampling schemes in the large deviation is discussed in [10, 11] in terms of a deterministic differential game. And [12] suggests a way to construct importance sampling schemes based on subsolutions of the Isaacs equation.

Branching Type Schemes

For branching type schemes, the most discussed application is the first entrance probabilities, and to continue the discussion we specialized to that case. Consider a Markov process $X = \{X_i, i = 0, 1, \dots\}$, and the probability of interest is

$$p = \mathbb{P}\{\tau_A \leq \tau_{A \cup B} | X(0) = x\}$$

for sets A and B , where $\tau_C = \min\{i \geq 0 : X_i \in C\}$. We consider just the discrete time case, but note that the continuous time case is also of interest. Suppose that B is a “typical set” and that A is rare, in that with high probability (perhaps *very* high probability) B is entered prior to A .

When using branching processes, the idea is to branch the process in such a way that the a priori rare outcome becomes more likely for at least some of the descendants (or particles), and then normalize to produce a single unbiased estimate of p . As with ordinary Monte Carlo one can then generate a number of independent copies of the branched process, and use the sample average to estimate p . Since the copies are independent the performance of the scheme is determined by the variance of a single copy. Key questions in the construction of the branching process are then *when to branch* (the triggers) and *how much to branch* (the rates). A natural and standard approach to defining triggers is the use of thresholds, i.e., nested sets $\{C_j, j = 0, 1, \dots, m\}$ with

$$C_0 \supset C_1 \supset \dots \supset C_m \supset A,$$

and a rule for branching when these sets are entered. For example, branching of a particular particle could occur the first time either it or perhaps some selected collection of particles containing it first reaches a threshold set.

Within a framework where branching is triggered by thresholds there are many possible variants. In what appears to be the oldest class of such methods, particles do not interact, and branching of any particular particle occurs the first time it reaches a threshold that has not been reached before. The number of descendants could be random and it could depend on other factors such as the position of the particle when entering the threshold, but it is independent of the state of all other particles at the time of the branching. (If more than one new threshold is entered at the same time then the distribution of the number of descendants is an appropriate convolution.) The mean number of descendants is often a constant (an assumption which for purposes of discussion we will assume true), and hence these types of schemes are sometimes called *Fixed Rate Branching* schemes. Besides the basic fixed branching rate scheme [16, 17, 19], more efficient variants which kill off particles that are of little use while remaining unbiased have been developed [3, 18, 22, 23, 24].

There are many challenges in the design of such schemes. For a large scale problem care must be taken to ensure that the total number of particles, whose distribution obviously depends on the precise placement of the thresholds, is not too large. At the same time, significant reduction of the variance of the estimator requires that enough particles make it all the way to the rare set. The challenge of the design problem is to find the proper balance between these opposing requirements.

In a second class of methods, a particle can be branched only when all particles that started from the previous threshold (say C_{j-1}) have either reached C_j or B . The branching here will not be conditionally independent of the state of all other particles, and in fact the branching is done so as to maintain a fixed and deterministic number (say N) of particles at each threshold. To be precise, suppose that $\{X_{j-1,n}, n = 1, \dots, N\}$ denote the locations of N particles in threshold C_{j-1} . Starting with these locations, particles are independently evolved under the dynamics of X until either C_j or B is reached. Suppose that $N_j \leq N$ particles reach C_j before B , and let $\mathcal{S}_j$ denote the collection of corresponding locations. In order to maintain N particles, we let $X_{j,n}$ for $n = 1, \dots, N_j$ be chosen to coincide with the N_j positions in $\mathcal{S}_j$ (the order is unimportant), and then construct $X_{j,n}$ for $n = N_j + 1, \dots, N$ by independently sampling from $\mathcal{S}_j$ according to the uniform distribution. [If $N_j = 0$ the simulation terminates and zero is returned as the estimate.] Repeating this process over say m thresholds which separate the initial point from A , we get an unbiased estimate

$$\hat{p} = \prod_{j=1}^m \frac{N_j}{N} \quad (0.2)$$

of $\mathbb{P}\{\tau_A \leq \tau_{A \cup B}\}$.

Methods of this sort are studied in [4, 5, 14, 15], and since the total computational effort (or at least the total number of particles) is a priori known, they are referred to as *Fixed Effort Branching* methods. Other names used for these sorts of schemes are *interacting particle systems* (IPS) and *genealogical particle methods*.

Large deviation analysis of both standard splitting and the variation known as

RESTART have been carried out in [2] and [3]. Under the large deviation scaling (and for light-tailed processes), it is natural to let the thresholds be defined as level sets of a function V that is independent of m , and then let the “gaps” between levels scale proportionally to $1/m$. Associated to the estimation problem is a partial differential equation [PDE] (specifically, a first order Hamilton-Jacobi-Bellman equation) defined in terms of the sets A and B and the large deviation rate function of X^m . The first conclusion of [2] and [3] is that the total computational effort associated with generating a single sample scales subexponentially with m if and only if $V/\log R$ is a *subsolution* to this PDE, where R is the mean number of particles produced at each branching. It is well known that p^m decays like $e^{-mW(x)}$, where W is a weak-sense *solution* to the same PDE. The second conclusion is that the second moment of the estimator (defined via the thresholds of V and with mean branching rate R) decays like $e^{-m[W(x)+V(x)/\log R]}$, where x is the starting point $X^m(0) = x$. Hence, in this rough exponential sense, the standard deviation of the estimator is of the same size as the probability of interest if and only if $V(x)/\log R$ achieves the maximum possible value at x [i.e., $W(x)$]. When this is not the case the ratio of the standard deviation to the probability grows exponentially as $e^{m[W(x)-V(x)/\log R]/2}$. Hence significant improvement over ordinary Monte Carlo [i.e., $V(x) = 0$] is still possible, and the improvement is quantified by $V(x)$.

In contrast, very little is known regarding the performance of fixed effort schemes in the small probability limit. Analysis to date have studied the variance when the number of particles is first sent to ∞ (so that exchangeability and asymptotically negligible interaction between particles can be exploited), and then the probability is made small. Since fixed effort schemes are at least in part motivated by a desire to control computational effort in a predictable way, this is not an ideal measure of performance. One would prefer to keep the number of particles fixed while letting the probability tend to zero and, as with the fixed rate schemes, compare the standard deviation of the estimator with the probability. It is not surprising that such an

analysis is difficult. With standard splitting, any two particles that reach A have a fairly simple dependence. At some time in the past these particles have a last common ancestor. Prior to that the particles have the same history and after that the evolutions of the two are independent. One can partition particles according to the time and location of the last ancestor, and in this way exploit this relatively simple dependence. More complex implementations are more difficult to analyze, but the dependence between particles is still limited.

The situation with fixed effort schemes is very different. Here dependence between particles is reintroduced with each resampling, which makes the analysis much more difficult. In the non-interacting case, one can analyze the variance of the estimator using large deviation estimates and techniques related to Freidlin-Wentzell type “small noise” asymptotics. In the interacting case the relevant techniques are related to those used to analyze the empirical or occupation measure of a Markov process, though there are numerous difficulties that make the analysis significantly more technical and complex. Among the difficult features besides the recoupling mentioned previously are: (i) the natural time scale for the process is not the natural time scale for the algorithm, which is in fact indexed by thresholds, (ii) the transition kernels associated with threshold crossing typically do not satisfy the Feller property, (iii) the transition kernels also depend on the large deviation parameter m , and (iv) the derived estimates should be valid all the way from a small number of particles (where large deviation effects dominate) to many particles (where law of large numbers effects dominate).

In this thesis we consider only the one dimensional problem, which is already somewhat demanding, and leave the multidimensional problem for further study. An outline of the thesis is as follows. In the next chapter we formulate the problem and the algorithm in precise terms, and state most of the assumptions that will be used. We also discuss the failure of the Feller property mentioned previously, and indicate how this issue will be addressed in the analysis. Chapter 2 gives the statement of

the main result, which is the large deviation decay rate for the second moment of the IPS estimator for an arbitrary number of particles. Chapter 3 develops a convenient stochastic control representation for this second moment. Chapter 4 proves tightness of measures and processes appearing in the representation, characterizes weak limits, and proves the large deviation upper bound. Ergodicity properties for controlled processes are discussed in Chapter 5, after which the large deviation lower bound is proved. Chapter 6 studies the decay rate as a function of the number of particles. Among the conclusions of this section are that for any fixed N the IPS method has a suboptimal rate of convergence of the second moment to zero, and with a well defined “gap.” However, another conclusion is that the IPS method is also less sensitive to design issues than competing schemes, at least in the one dimensional setting. The issue is less clear cut for multidimensional problems.

0.3 Preliminaries

In this section, we state some useful concepts and theorems which will be needed throughout this thesis.

We begin by giving the definition of Laplace principle which can be seen as an alternative definition for large deviation principle (Definition 0.2). See [9, Section 1.2] for detail discussions. In this dissertation, we will use this definition to prove the large deviation property of the second moment of IPS estimator.

Definition 0.3 (Laplace Principle). *Let I be a rate function on $\mathcal{X}$. The sequence $\{X^m\}$ is said to satisfy the Laplace principle on $\mathcal{X}$ with rate function I if for all bounded continuous functions h*

$$\lim_{m \rightarrow \infty} \frac{1}{m} \log \mathbb{E} \{ \exp [-mh(X^m)] \} = - \inf_{x \in \mathcal{X}} \{ h(x) + I(x) \}.$$

The term Laplace principle upper bound refers to the validity of

$$\limsup_{m \rightarrow \infty} \frac{1}{m} \log \mathbb{E} \{ \exp [-mh(X^m)] \} \leq - \inf_{x \in \mathcal{X}} \{ h(x) + I(x) \}$$

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

$$\liminf_{m \rightarrow \infty} \frac{1}{m} \log \mathbb{E} \{ \exp [-mh(X^m)] \} \geq - \inf_{x \in \mathcal{X}} \{ h(x) + I(x) \}.$$

for all bounded continuous functions h .

Another important concept is relative entropy, which will play a key role in the weak convergence approach to large deviations via the variational formula given in part (a) of Proposition 0.5. The derivation of this formula requires a measurable space $(\mathcal{S}, \mathcal{A})$. We denote by $\mathcal{P}(\mathcal{S})$ the set of probability measures on $(\mathcal{S}, \mathcal{A})$.

Definition 0.4 (Relative Entropy). *Given $\gamma, \theta \in \mathcal{P}(\mathcal{S})$, the relative entropy of γ with respect to θ is defined by*

$$R(\gamma \parallel \theta) = \int_{\mathcal{S}} \log \left(\frac{d\gamma}{d\theta}(x) \right) \gamma(dx)$$

if γ is absolutely continuous with respect to θ , i.e. $\gamma \ll \theta$, and $\log([d\gamma/d\theta](x))$ is integrable with respect to γ , and by ∞ otherwise.

The next proposition [9, Proposition 1.4.2], which states the variational formula, is a cornerstone for the weak convergence approach.

Proposition 0.5. *Let $(\mathcal{S}, \mathcal{A})$ be a measurable space, k a bounded measurable function mapping $\mathcal{S}$ into $\mathbb{R}$, and θ a probability measure on $\mathcal{S}$. The following conclusions hold.*

(a) *We have the variational formula*

$$- \log \int_{\mathcal{S}} e^{-k} d\theta = \inf_{\gamma \in \mathcal{P}(\mathcal{S})} \left\{ R(\gamma \parallel \theta) + \int_{\mathcal{S}} k d\gamma \right\}. \quad (0.3)$$

(b) Let γ_0 denote the probability measure on $\mathcal{S}$ which is absolutely continuous with respect to θ and satisfies

$$\frac{d\gamma_0}{d\theta}(x) \doteq e^{-k(x)} \cdot \frac{1}{\int_{\mathcal{S}} e^{-k} d\theta}.$$

Then the infimum in the variational formula (0.3) is attained uniquely at γ_0 .

An extended version of Proposition 0.5 will also be needed, given in [9, Proposition 4.5.1].

Proposition 0.6. *Let $\mathcal{S}$ be a Polish space, k a measurable function mapping $\mathcal{S}$ into $\mathbb{R}$ which is either bounded from below or bounded from above, and θ a probability measure on $\mathcal{S}$. Then we have the variational formula*

$$-\log \int_{\mathcal{S}} e^{-k} d\theta = \inf_{\gamma \in \Delta(\mathcal{S})} \left\{ R(\gamma \parallel \theta) + \int_{\mathcal{S}} k d\gamma \right\},$$

where $\Delta(\mathcal{S}) \doteq \{\gamma \in \mathcal{P}(\mathcal{S}) : R(\gamma \parallel \theta) < \infty\}$.

The next lemma and theorem [9, Lemma 1.4.3, Theorem B.2.1] investigate some properties of relative entropy: convexity, lower semicontinuity, compactness of level sets, uniform integrability of sequences of measures satisfying a suitable relative entropy bound, factorization, approximation by sums and chain rule, etc.

Let $\mathcal{X}$ be a Polish space. We denote by $\mathcal{C}_b(\mathcal{X})$ the space of bounded continuous functions mapping $\mathcal{X}$ into $\mathbb{R}$ and by $\Psi_b(\mathcal{X})$ the space of bounded measurable functions mapping $\mathcal{X}$ into $\mathbb{R}$.

Lemma 0.7. *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces. The relative entropy $R(\cdot \parallel \cdot)$ has the following properties.*

(a) (**Donsker-Varadhan variational formula.**) For each γ and θ in $\mathcal{P}(\mathcal{X})$

$$\begin{aligned} R(\gamma \parallel \theta) &= \sup_{g \in \mathcal{C}_b(\mathcal{X})} \left\{ \int_{\mathcal{X}} g d\gamma - \log \int_{\mathcal{X}} e^g d\theta \right\} \\ &= \sup_{\psi \in \Psi_b(\mathcal{X})} \left\{ \int_{\mathcal{X}} \psi d\gamma - \log \int_{\mathcal{X}} e^\psi d\theta \right\}. \end{aligned} \quad (0.4)$$

(b) $R(\gamma \parallel \theta)$ is a convex, lower semicontinuous function of $(\gamma, \theta) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$. In particular, $R(\gamma \parallel \theta)$ is a convex, lower semicontinuous function of each variable γ or θ separately. In addition, for fixed $\theta \in \mathcal{P}(\mathcal{X})$ $R(\cdot \parallel \theta)$ is strictly convex on the set $\{\gamma \in \mathcal{P}(\mathcal{X}) : R(\gamma \parallel \theta) < \infty\}$.

(c) For each $\theta \in \mathcal{P}(\mathcal{X})$ $R(\cdot \parallel \theta)$ has compact level sets. That is, for each $M < \infty$ the set $\{\gamma \in \mathcal{P}(\mathcal{X}) : R(\gamma \parallel \theta) \leq M\}$ is a compact subset of $\mathcal{P}(\mathcal{X})$.

(d) Let $\mathcal{X} = \mathbb{R}^d$ and let $\{\gamma_n, n \in \mathbb{N}\}$ and $\{\theta_n, n \in \mathbb{N}\}$ be sequences in $\mathcal{P}(\mathbb{R}^d)$. Assume that for each $\alpha \in \mathbb{R}^d$

$$\sup_{n \in \mathbb{N}} \int_{\mathbb{R}^d} \exp \langle \alpha, y \rangle \theta_n(dy) < \infty \text{ and } \sup_{n \in \mathbb{N}} R(\gamma_n \parallel \theta_n) < \infty.$$

Then $\{\gamma_n\}$ is both tight and uniformly integrable in the sense that

$$\lim_{C \rightarrow \infty} \sup_{n \in \mathbb{N}} \int_{\{y \in \mathbb{R}^d : \|y\| > C\}} \|y\| \gamma_n(dy) = 0.$$

(e) Let $\mathcal{X} = \mathbb{R}^d$ and let γ and θ be probability measures on $\mathbb{R}^d$. Assume that for each $\alpha \in \mathbb{R}^d$

$$\int_{\mathbb{R}^d} \exp \langle \alpha, y \rangle \theta(dy) < \infty$$

and that $R(\gamma \parallel \theta) < \infty$. Then $\int_{\mathbb{R}^d} \|y\| \gamma(dy) < \infty$.

(f) Let $\sigma(dy|x)$ and $\tau(dy|x)$ be stochastic kernels on $\mathcal{Y}$ given $\mathcal{X}$ and θ a probability measure on $\mathcal{X}$. Then the function mapping $x \in \mathcal{X} \mapsto R(\sigma(\cdot \parallel x) \parallel \tau(\cdot \parallel x))$ is

measurable and

$$\int_{\mathcal{X}} R(\sigma(\cdot|x) \parallel \tau(\cdot|x)) \theta(dx) = R(\theta \otimes \sigma \parallel \theta \otimes \tau).$$

(g) We denote by Π the class of all finite measurable partitions of $\mathcal{X}$. Then for each γ and θ in $\mathcal{P}(\mathcal{X})$

$$R(\gamma \parallel \theta) = \sup_{\pi \in \Pi} \sum_{A \in \pi} \gamma(A) \log \frac{\gamma(A)}{\theta(A)},$$

where the summand equals 0 if $\gamma(A) = 0$ and equals ∞ if $\gamma(A) > 0$ and $\theta(A) = 0$. In addition, if A is any Borel subset of $\mathcal{X}$, then

$$R(\gamma \parallel \theta) \geq \gamma(A) \log \frac{\gamma(A)}{\theta(A)} - 1.$$

Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces and let $\tau(dy|x)$ be a family of probability measures on $\mathcal{Y}$ parameterized by $x \in \mathcal{X}$. We call $\tau(dy|x)$ a *stochastic kernel* on $\mathcal{Y}$ given $\mathcal{X}$ if for every Borel subset E of $\mathcal{Y}$ the function mapping $x \in \mathcal{X} \mapsto \tau(E|x) \in [0, 1]$ is measurable.

Theorem 0.8 (Chain Rule). *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces and α and β probability measures on $\mathcal{X} \times \mathcal{Y}$. We denote by α_1 and β_1 the first marginals of α and β and by $\alpha(dy|x)$ and $\beta(dy|x)$ the stochastic kernels on $\mathcal{Y}$ given $\mathcal{X}$ for which we have the decompositions*

$$\alpha(dx \times dy) = \alpha_1(dx) \otimes \alpha(dy|x) \text{ and } \beta(dx \times dy) = \beta_1(dx) \otimes \beta(dy|x).$$

Then the function mapping $x \in \mathcal{X} \mapsto R(\alpha(\cdot|x) \parallel \beta(\cdot|x))$ is measurable and

$$R(\alpha \parallel \beta) = R(\alpha_1 \parallel \beta_1) + \int_{\mathcal{X}} R(\alpha(\cdot|x) \parallel \beta(\cdot|x)) \alpha_1(dx).$$

The next theorem [9, Theorem A.5.6] states the decomposition of stochastic kernels.

Theorem 0.9 (Decomposition Theorem). *Let $\sigma(dx \times dy|z)$ be a stochastic kernel on $\mathcal{X} \times \mathcal{Y}$ given $\mathcal{S}$ and θ a probability measure on $(\mathcal{S}, \mathcal{A})$. We denote by $\sigma_1(dx|z)$ the first marginal of $\sigma(dx \times dy|z)$, which is the stochastic kernel on $\mathcal{X}$ given $\mathcal{S}$ defined by $\sigma_1(A|z) \doteq \sigma(A \times \mathcal{Y}|z)$ for Borel sets A and $z \in \mathcal{Y}$. Then there exists a stochastic kernel $\sigma(dy|x, z)$ on $\mathcal{Y}$ given $\mathcal{X} \times \mathcal{S}$ such that θ -a.s. for $z \in \mathcal{S}$*

$$\sigma(A \times B|z) = \int_{A \times B} \sigma_1(dx|z) \sigma(dy|x, z) = \int_A \sigma(B|x, z) \sigma_1(dx|z)$$

for all Borel subesets A of $\mathcal{X}$ and B of $\mathcal{Y}$. This decomposition is summarized as $\sigma(dx \times dy|z) = \sigma_1(dx|z) \otimes \sigma(dy|x, z)$.

Assume $p(x, dy)$ a transition probability function defined on polish space $\mathcal{S}$. The Feller property ([9, Condition 8.3.1]) is defined as following:

Definition 0.10 (Feller Property). *The function mapping $x \in \mathcal{S} \mapsto p(x, \cdot) \in \mathcal{P}(\mathcal{S})$ is continuous in the topology of weak convergence on $\mathcal{P}(\mathcal{S})$; i.e., if $\{x_n, n \in \mathbb{N}\}$ is any sequence in $\mathcal{S}$ such that $x_n \rightarrow x \in \mathcal{S}$, then $p(x_n, \cdot) \Rightarrow p(x, \cdot)$.*

Part I

Problem Formulation and Main Results

Chapter 1

Notation and Problem Formulation

In this chapter we formulate the problem of interest, which is the analysis via large deviations of the performance of an *interacting particle system* (IPS) for rare event estimation. We begin by identifying the event of interest and the related large deviation asymptotics in the next section. In the following section the estimator is constructed and standard performance criteria for rare event estimation are reviewed. In the last section we show how these criteria are reformulated in terms of the occupation measure of a related stochastic process.

1.1 Process Model and the Event of Interest

In this thesis, we restrict discussion to a one dimensional process, but within that framework allow a state dependent discrete time Markov model. The process model is constructed as follows. Let $\{\theta_i(x), i \in \mathbb{N}_0, x \in \mathbb{R}\}$ be a sequence of independent and identically distributed (iid) random fields with the common distribution

$$\mathbb{P}\{\theta_i(x) \in dy\} = \mu(dy|x). \quad (1.1)$$

For $m \in \mathbb{N}$, define a Markov chain by setting $Z_0^m = 0$ and recursively defining

$$Z_{i+1}^m = Z_i^m + \frac{1}{m} \theta_i(Z_i^m). \quad (1.2)$$

Here m is the large deviation parameter, and we are interested in events that become rare as $m \rightarrow \infty$. We also consider a continuous time interpolant, which is defined by $Z^m(t) = Z_i^m$ if $t \in [i/m, i/m + 1/m)$. We consider $Z^m(\cdot)$ to take values in the standard Skorokhod space.

Conditions 1.1 and 1.3 introduced below will be assumed throughout the paper. Condition 1.2 will be replaced by the strictly stronger Condition 1.4 when discussing the performance of the IPS scheme.

Condition 1.1 (Stability). $\sup_{x \in \mathbb{R}} \mathbb{E} \theta_i(x) < 0$.

The event of interest is then

$$p^m \doteq \mathbb{P} \{ Z_{\sigma^m}^m \geq 1 \mid Z_0^m = 0 \}, \quad (1.3)$$

where $\sigma^m = \min \{ i \geq 0 : Z_i^m \in [1, \infty) \cup [-\infty, -1/m) \}$. Although the methods we develop below can be used to analyze a variety of problem formulations, the event studied here is common in the rare event literature for the analysis of algorithms. To identify the rate of decay via large deviations, we need some standard notation and conditions. For the event of interest it is only the definition of $\mu(dy|x)$ on $[-1/m, 1]$ that matters, and so we pose conditions henceforth only on $x \in [-1, 1]$. For each such x and $\alpha \in \mathbb{R}$, let

$$H(x, \alpha) = \log \int_{\mathbb{R}} e^{\alpha y} \mu(dy|x). \quad (1.4)$$

For a Polish space S , let $\mathcal{P}(S)$ denote the set of probability measures on S , equipped with the topology of weak convergence.

Condition 1.2. For each $\alpha \in \mathbb{R}$, $\sup_{x \in [-1, 1]} H(x, \alpha) < \infty$, and the mapping $x \in \mathbb{R} \mapsto \mu(\cdot|x) \in \mathcal{P}(\mathbb{R})$ is continuous in the topology of weak convergence on $\mathcal{P}(\mathbb{R})$.

The following properties are well known under the last condition [9]. Let $L(x, \cdot)$ be the Legendre transform of $H(x, \cdot)$, i.e., for $\beta \in \mathbb{R}$

$$L(x, \beta) = \sup_{\alpha \in \mathbb{R}} [\alpha\beta - H(x, \alpha)].$$

Then $L(x, \beta) \geq 0$ and $L(x, \beta) = 0$ if and only if $\beta = \mathbb{E}\theta_i(x)$. Condition 1.2 is sufficient for Z^m to satisfy a large deviation upper bound on path space. Various sets of conditions suffice for the corresponding lower bound, and the particular condition that is assumed below is sufficient for the lower bound and the existence of a large deviation limit for p^m . See [9] for alternatives.

Condition 1.3. *The interior of the convex hull of the support of $\mu(\cdot|x)$ is independent of x and contains 0.*

Under all three conditions we can assert that

$$\lim_{m \rightarrow \infty} \frac{1}{m} \log p^m = \lim_{m \rightarrow \infty} \frac{1}{m} \log \mathbb{P} \{Z_{\sigma^m}^m \geq 1 \mid Z_0^m = 0\} = -\rho, \quad (1.5)$$

where

$$\begin{aligned} \rho &= \inf \left\{ \int_0^\tau L(\phi(s), \dot{\phi}(s)) ds : \phi(0) = 0, \phi(\tau) = 1, \phi(s) \in (0, 1) \text{ for } s \in (0, \tau) \right\} \\ &\in (0, \infty), \end{aligned}$$

and where the integral is taken to be ∞ if ϕ is not absolutely continuous.

1.2 Interacting Particle System (IPS) Estimator

While the large deviation asymptotics described in the last section give an estimate of the probability for any m , this estimate can be off significantly for values of m that are of interest. For this reason there is much interest in improved approximations, and in particular in higher accuracy Monte Carlo approximations.

The construction of an IPS estimator for p^m can be described as follows. We introduce a sequence of sets (often defined in higher dimensions via *level sets*) of the form

$$C_m^m \subset C_{m-1}^m \subset \cdots \subset C_0^m.$$

In the present one dimensional setting we will use $C_j^m = \{x : x \geq j/m\}$. (An interesting feature of the IPS scheme that will be proved is that, in the one dimensional setting, it is relatively insensitive to this choice). The construction of the estimate is recursive and indexed by stages, which are identified with the entrance into a corresponding set (we will use the terms *stage* and *threshold* interchangeably). The construction at each stage is partitioned into *transportation* and *resampling* aspects. For particles $n = 1, \dots, N$ and thresholds $j = 0, \dots, m-1$, define stopping times $\sigma_{j,n}^m$ and stopped processes recursively as follows. Let $\{\theta_{j,n,i}(x), x \in [-1, 1]\}$ be random fields, which are independent and identically distributed for distinct values of $j \in \{1, \dots, m-1\}$, $n \in \{1, \dots, N\}$ and $i \in \mathbb{N}_0$.

Stage $j = 0$. Simulate N copies of (1.2), i.e., define

$$Z_{0,n,i+1}^m = Z_{0,n,i}^m + \frac{1}{m} \theta_{0,n,i}(Z_{0,n,i}^m), \quad Z_{0,n,0}^m = 0,$$

for $n = 1, \dots, N$ and let

$$\sigma_{0,n}^m = \min \{i \geq 0 : Z_{0,n,i}^m \in C_1^m \cup (-\infty, -1/m]\},$$

be the stopping time of the n th particle hitting the next stage C_1^m , or being *absorbed* by $(-\infty, -1/m]$.

Stages $j = 1, \dots, m-1$. Let $Y_{j,n}^m = Z_{j-1,n,\sigma_{j-1,n}^m}^m$, i.e., the location of the n th particle when it first enters C_j^m or $(-\infty, -1/m]$. Let $\mathcal{S}_j^m = \{Y_{j,n}^m : Y_{j,n}^m \in C_j^m\}$ and let $N_j^m \leq N$ be the number of elements of $\mathcal{S}_j^m$. Set $X_{j,n}^m = Y_{j,n}^m$ if $Y_{j,n}^m \in C_j^m$, and choose $X_{j,n}^m$ for indices n not corresponding to an element of $\mathcal{S}_j^m$ uniformly from $\mathcal{S}_j^m$.

Define processes $Z_{j,n,i}^m$ via $Z_{j,n,0}^m = X_{j,n}^m$ and

$$Z_{j,n,i+1}^m = Z_{j,n,i}^m + \frac{1}{m} \theta_{j,n,i} (Z_{j,n,i}^m),$$

for $n = 1, \dots, N$, and again, let

$$\sigma_{j,n}^m = \min \{i \geq 0 : Z_{j,n,i}^m \in C_{j+1}^m \cup (-\infty, -1/m]\}.$$

In general, the particles are thought of as evolving according to a two-step, genetic type mechanism, though we eschew the associated terminology here:

$$(X_{j,1}^m, \dots, X_{j,N}^m) \xrightarrow{\text{transportation}} (Y_{j+1,1}^m, \dots, Y_{j+1,N}^m) \xrightarrow{\text{resampling}} (X_{j+1,1}^m, \dots, X_{j+1,N}^m).$$

Since each particle is distributed according to the first entrance distribution into threshold j , N_j^m/N is an unbiased estimate of

$$\mathbb{P} \{Z^m \text{ reaches threshold } j \mid Z^m \text{ reached threshold } j-1\}.$$

Thus,

$$\hat{p}^m = \prod_{j=1}^m \frac{N_j^m}{N} \tag{1.6}$$

is in fact an unbiased estimate of

$$p^m = \prod_{j=1}^m \mathbb{P} \{Z^m \text{ reaches threshold } j \mid Z^m \text{ reached threshold } j-1\}.$$

The standard measure of performance in these circumstances is the variance of $\hat{p}^m$, which is to be minimized. Since $\mathbb{E} \hat{p}^m = p^m$, minimization of the variance is equivalent to minimization of the second moment $\mathbb{E}(\hat{p}^m)^2$. Recall the discussion in Chapter 0, we say the estimator $\hat{p}^m$ is *asymptotically optimal* if the upper bound is

achieved, i.e., if

$$\lim_{m \rightarrow \infty} -\frac{1}{m} \log \mathbb{E} (\hat{p}^m)^2 = 2\rho.$$

Of course, besides performance one must consider the computational cost to produce each sample $\hat{p}^m$. However, (0.1) provides a bound on the best possible performance.

1.3 Reformulation as an Occupation Measure

Problem

The IPS approach to rare event estimation requires a very different set of techniques for its analysis than approaches based on branching processes that use a pre-determined branching rate. In particular, the statistical coupling of particles that occurs at each resampling stage prevents direct use of the “small noise” large deviation properties of the underlying distribution of Z^m . Instead, large deviation asymptotics for the occupation measures associated with the particles that make it to the next threshold are relevant. Indeed, if

$$L^m = \frac{1}{m} \sum_{j=1}^m \delta_{N_j^m}, \quad (1.7)$$

then the performance measure

$$-\frac{1}{m} \log \mathbb{E} (\hat{p}^m)^2 = -\frac{1}{m} \log \mathbb{E} \prod_{j=1}^m \left(\frac{N_j^m}{N} \right)^2$$

can be rewritten in the form

$$\hat{W}_N^m = -\frac{1}{m} \log \mathbb{E} \left[e^{2 \sum_{j=1}^m \log \left(\frac{N_j^m}{N} \right)} \right] = -\frac{1}{m} \log \mathbb{E} \left[e^{2m \int \log \left(\frac{n}{N} \right) L^m(dn)} \right]. \quad (1.8)$$

This exhibits the performance measure as a Laplace functional of the occupation measure L^m . However, for many reasons the analysis of L^m is nonstandard, including

the following.

1. The “transition kernel” for N_j^m is only implicitly defined in terms of the basic problem data [i.e., the process dynamics $\mu(dy|x)$], and involves a two-stage (transportation and resampling) construction.
2. The transition kernel will depend on x through $\mu(dy|x)$. However, x here is identified with the stage of the construction, and so plays the role of a “time” variable and not a state variable.

In the following chapters we carry out a full large deviations analysis of $\hat{W}^m$ in the limit $m \rightarrow \infty$. However, the resulting rate will depend in a nontrivial way on the fixed number of particles N . As a consequence, the dependence of the rate constant on the number of particles should also be analyzed.

1.4 First Entrance Distributions and the Feller Property

As discussed in the last section, the large deviation properties of the IPS estimator are intimately connected with the large deviation properties of a related empirical measure on the number of particles making it to the next threshold (*success particles*). The number of success particles is not by itself Markovian, and a transition kernel on first entrance distributions plays an essential role in its analysis. In this section the transition kernel that arises in the limit will be defined, and properties that are needed for a large deviation analysis will be discussed.

Anticipating a change in notation later on, we replace x by t , and recall that for $t \in [0, 1]$, $\theta_i(t)$ denotes a sequence of iid random variables with distribution $\mu(dy|t)$. Define $Z_i(t)$ by $Z_{i+1}(t) = Z_i(t) + \theta_i(t)$ and set $\sigma(t) = \min \{i \geq 0 : Z_i(t) \geq 1\}$. Note that t here plays the role of a parameter. We next define a family of stochastic kernels indexed by t . The kernels will be measures on $[0, \infty) \cup \{\Delta\}$, where Δ corresponds to

the event that $Z_i(t)$ reaches $-\infty$ before exceeding 1, and thereby enters a “cemetery” state. For $v \in \mathbb{R}$ and $A \subset [0, \infty) \cup \{\Delta\}$, set

$$\gamma(A|v, t) = \begin{cases} \mathbb{P}\{Z_{\sigma(t)}(t) - 1 \in A, \sigma(t) < \infty \mid Z_0(t) = v\} & v \in (-\infty, 1), A \subset [0, \infty) \\ \mathbb{P}\{\sigma(t) = \infty \mid Z_0(t) = v\} & v \in (-\infty, 1), A = \{\Delta\} \\ \delta_{v-1}(A) & v \in [1, \infty), A \subset [0, \infty) \end{cases}, \quad (1.9)$$

where δ_a denotes the Dirac function at a .

For each fixed t , $v \rightarrow \gamma(\cdot|v, t)$ is a transition kernel on $[0, \infty) \cup \{\Delta\}$ that reflects the distributional properties of the “overshoot” i.e., the location of the process when it first enters the next threshold given that the current overshoot is $v \neq \Delta$, when the “local” characteristics of the increments in the underlying process are determined by $\mu(dy|t)$. It is the local transition kernel that will be relevant in the IPS scheme in the limit $m \rightarrow \infty$, in a sense that will be made precise in Lemma 3.2.

Before proceeding we review some issues concerning large deviations for the occupation measures of a Markov chain. The first papers on this topic were [7, 8], which were subsequently generalized in many directions. One of the key assumptions in [7] and most other work on the problem is that the transition kernel satisfy the Feller property as in Condition 0.10. In the present context this would be the requirement that for each $t \in [0, 1]$, the map

$$v \rightarrow \gamma(\cdot|v, t)$$

is continuous in the topology of weak convergence for $v \in [0, \infty) \cup \{\Delta\}$. While such an assumption is reasonable in many circumstances, it is restrictive in the current setting, since γ is defined by a first entrance distribution. The Feller property can be weakened, as in [9, Section 9.2], which requires that the values of v where it fails are in some sense “negligible” and thus one is still in an essentially Feller type framework. In fact, one obtains a rate function that is of exactly the same form as

in the Feller case, and we will follow a similar approach here. However, for other, less structured weakenings of the Feller property, the resulting processes can display unusual large deviation properties with very non-standard forms for the rate function (when such processes can be analyzed at all). Indeed, many strange behaviors can occur when the dynamics of the chain allow the process to spend significant time near the discontinuities [13].

A second property that will be needed to establish large deviation bounds is regularity of $\gamma(\cdot|v, t)$ in t . Using the Markov property and a standard conditioning argument, one can uniquely characterize $\gamma(\cdot|v, t)$ for $v \in (-\infty, 1)$ and Borel sets $A \subset [0, \infty)$ via the equation [see the derivation of (3.11)]

$$\gamma(A|v, t) = \int_{\{y: v+y < 1\}} \gamma(A|v+y, t) \mu(dy|t) + \mu(A+1-v|t). \quad (1.10)$$

It is apparent from this equation that the atoms of $\mu(\cdot|t)$ play an essential role in determining whether a Feller property or a “near” Feller property holds, and that regularity in t will impose further conditions. It should also be noted that even if $\mu(\cdot|t)$ has a density that is as smooth as desired with respect to Lebesgue measure, $\gamma(\cdot|v, t)$ will not satisfy the ordinary Feller property because of the discontinuity at $v = 1$. We will use the following assumption, which is general enough to cover many standard models and simple to verify. Note that this condition implies Condition 1.2.

Condition 1.4. *Consider reference probability measures ν_d and ν_c , each with a finite moment generating function, such that ν_c is absolutely continuous with respect to Lebesgue measure and there is an integer q such that the support of ν_d is contained in $\mathbb{Z}/q$. We then assume $\mu(dx|t) = f_d(x, t)\nu_d(dx) + f_c(x, t)\nu_c(dx)$, where $f_d(x, t)$ and $f_c(x, t)$ are continuous and uniformly positive and bounded in x and t .*

Under this condition either the Feller property will hold, or a near Feller property will hold so that the large deviation properties are as if the Feller property were true.

The next lemma identifies the places where the Feller property fails, and also shows that continuity from the right always holds.

Lemma 1.5. *Assume Conditions 1.1 and 1.4. Then for each $t \in [0, 1]$*

$$\{v \in [0, 1] : v \rightarrow \gamma(\cdot|v, t) \text{ is discontinuous}\} \subset (0, 1] \cap [\mathbb{Z}/q],$$

and $\gamma(\cdot|v, t)$ is continuous from the right for all v .

Proof. Let $t \in [0, 1]$ be given, and let $\theta_i = \theta_i(t)$ be iid with distribution $\mu(\cdot|t)$ (for simplicity we omit the t -dependence of θ_i). For $v \in [0, 1]$, define

$$R_v \doteq \sum_{i=1}^{\infty} \left[v + \sum_{k=1}^i \theta_k - 1 \right] 1_{\{v + \sum_{k=1}^{\ell} \theta_k < 1 \text{ for } \ell=1, \dots, i-1\}} 1_{\{v + \sum_{k=1}^i \theta_k \geq 1\}}.$$

Then with the understanding that $v + \sum_{k=1}^{\ell} \theta_k < 1$ for all $\ell < \infty$ means $R_v = \Delta$, the distribution of R_v is $\gamma(\cdot|v, t)$. Thus for a bounded and continuous function $f : \mathbb{R} \rightarrow \mathbb{R}$

$$\begin{aligned} & \int_{\mathbb{R}} f(x) \gamma(dx|v, t) \\ &= \mathbb{E} \left[\sum_{i=1}^{\infty} f \left(v + \sum_{k=1}^i \theta_k - 1 \right) 1_{\{v + \sum_{k=1}^{\ell} \theta_k < 1 \text{ for } \ell=1, \dots, i-1\}} 1_{\{v + \sum_{k=1}^i \theta_k \geq 1\}} \right]. \end{aligned}$$

Suppose that v is not an element of $(0, 1] \cap [\mathbb{Z}/q]$. Then for each i the indicator functions in the last display are continuous at v w.p.1, and so by the Lebesgue Dominated Convergence Theorem, if $v_j \rightarrow v$ as $j \rightarrow \infty$ then

$$\int_{\mathbb{R}} f(x) \gamma(dx|v_j, t) \rightarrow \int_{\mathbb{R}} f(x) \gamma(dx|v, t).$$

The analogous argument applies when $v \in (0, 1] \cap [\mathbb{Z}/q]$ if we restrict to $v_j \downarrow v$, which proves the right continuity. $\square$

We will also need the following result, which indicates how the places where the

Feller property fails can be avoided. Let $\mu_n \otimes \gamma(A \times B|t) = \int_A \gamma(B|v, t) \mu_n(dv)$, and similarly for $\mu \otimes \gamma$.

Lemma 1.6. *Fix $t \in [0, 1]$, and assume that*

$$\{v \in [0, 1] : v \rightarrow \gamma(\cdot|v, t) \text{ is discontinuous}\} \subset (0, 1] \cap [\mathbb{Z}/q],$$

and $\gamma(\cdot|v, t)$ is continuous from the right for all v . Let $\mu_n \Rightarrow \mu$, where $\mu_n \in \mathcal{P}([0, 1])$. Assume that for all $\varepsilon > 0$ there is $\delta > 0$ such that $\mu_n(\cup_{k=1}^q([k/q] - \delta, [k/q])) \leq \varepsilon$. Then $\mu_n \otimes \gamma(dv \times dy|t) \Rightarrow \mu \otimes \gamma(dv \times dy|t)$.

Proof. To simplify notation we omit the t dependence. We need to show for bounded and continuous g and h that

$$\int g(v)h(y)(\mu_n \otimes \gamma)(dv \times dy) \rightarrow \int g(v)h(y)(\mu \otimes \gamma)(dv \times dy).$$

Now $f(v) = \int h(y)\gamma(dy|v)$ is continuous from the right for all v and from the left save (possibly) those v of the form k/q . Let $f^\delta(v)$ be continuous, have the same uniform bound as f , and equal to f outside $\cup_{k=1}^q([k/q] - \delta, [k/q])$. Then

$$\left| \int g(v)\mu_n(dv) \int h(y)\gamma(dy|v) - \int g(v)\mu_n(dv)f^\delta(v) \right| \leq 2\varepsilon \|h\|_\infty \|g\|_\infty.$$

Since $\mu_n \Rightarrow \mu$ implies $\mu(\cup_{k=1}^q([k/q] - \delta, [k/q])) \leq \varepsilon$,

$$\left| \int g(v)\mu(dv) \int h(y)\gamma(dy|v) - \int g(v)\mu(dv)f^\delta(v) \right| \leq 2\varepsilon \|h\|_\infty \|g\|_\infty,$$

and since $f^\delta(x)$ is continuous, $\mu_n \Rightarrow \mu$ implies

$$\limsup_{n \rightarrow \infty} \left| \int g(v)h(y)\mu_n(dv)\gamma(dy|v) - \int g(v)h(y)\mu(dv)\gamma(dy|v) \right| \leq 4\varepsilon \|h\|_\infty \|g\|_\infty.$$

The result now follows since $\varepsilon > 0$ is arbitrary. □

Chapter 2

Statement of the Main Results

In this chapter we present the main theoretical result, which is the limit of the second moment for the IPS estimator as $m \rightarrow \infty$.

For $n = 0, 1, \dots, N$, let $F^N(n) = -2 \log(n/N)$ (with the understanding that $F^N(0) = \infty$), and for $\ell \in \mathcal{P}(\{0, \dots, N\})$ let

$$\langle F^N, \ell \rangle = \sum_{n=0}^N F^N(n) \ell(n). \quad (2.1)$$

Recall that $\hat{W}_N^m$ is defined in (1.8) to be $-(1/m) \log \mathbb{E}(\hat{p}^m)^2$. It will be evident from the theorem statement that the decay rate for the second moment depends on the number of particles N . The dependence of the rate on N is discussed in Chapter 6.

As discussed in section 1.3, study of the IPS requires the large deviation analysis of an occupation measure problem with a “time” dependent transition kernel. To distinguish a point in the underlying state space of the process from variables that identify stages and hence this time variable, we use t rather than x to correspond to the normalized limits of the stage under the scaling j/m . (Its use in this manner motivated the switch from $\mu(dy|x)$ to $\mu(dy|t)$ previously.)

The local rate function $G_N(t, \phi)$ introduced below will be explicitly defined in Section 4.3 [equation (4.22)], after we have discussed the weak limits of certain con-

trolled IPS models that appear in a representation for $\hat{W}_N^m$. The second argument ϕ will be a probability measure on $\{1, \dots, N\}$. The local rate will depend on t since the number of particles which reach the next level and their locations depend on the stage through $\mu(dy|t)$. Thus the large deviation problem involves a probability measure-valued process with non-stationary increments, whose increments are essentially occupation measures that reflect the local properties (i.e., $\mu(dy|x)$ for x near t). The large deviation lower bound require a communication condition and a condition of locally uniform ergodicity on $\gamma(\cdot|v, t)$ when thought of as a transition kernel for each fixed t . These conditions are stated in Chapter 4.

Remark 2.1. This thesis is concerned with the efficiency of the IPS system, and for this problem the function of interest is $\langle F^N, L^m \rangle$. However, the proof of Theorem 2.2 goes through without change if $\langle F^N, L^m \rangle$ is replaced by any bounded and continuous function of L^m . One thereby obtains the full large deviations principle for $\{L^m\}$ with the rate function

$$I_N(\nu) = \inf \left\{ \int_0^1 G_N(t, \phi(t)) dt : \nu = \int_0^1 \phi(s) ds \right\}.$$

Theorem 2.2. *Assume Conditions 1.1, 1.2, and 1.4 and define $G_N : [0, 1] \times P(\{0, \dots, N\}) \rightarrow [0, \infty)$ by (4.22). Then*

$$\liminf_{m \rightarrow \infty} \hat{W}_N^m \geq \inf_{\nu \in \mathcal{P}(\{1, \dots, N\})} \inf_{\phi: \nu = \int_{[0,1]} \phi(t) dt} \left\{ \int_{[0,1]} G_N(t, \phi(t)) dt + \langle F^N, \nu \rangle \right\}. \quad (2.2)$$

If in addition Conditions 5.3 and 5.1 hold, then

$$\limsup_{m \rightarrow \infty} \hat{W}_N^m \leq \inf_{\nu \in \mathcal{P}(\{1, \dots, N\})} \inf_{\phi: \nu = \int_{[0,1]} \phi(t) dt} \left\{ \int_{[0,1]} G_N(t, \phi(t)) dt + \langle F^N, \nu \rangle \right\}. \quad (2.3)$$

Part II

Large Deviation Analysis

Chapter 3

Relative Entropy Representation

The key to the proof of Theorem 2.2 is a convenient relative entropy/control representation. The first step is to introduce notation which focuses attention on the key events of interest, i.e., the particles make it to the next level and their distribution (via the “transportation” measure) and the resampling from among these particles. The next step is to identify a “local” ergodicity which is central to the form of the local rate function $G_N(t, \phi)$. This will require several steps of rewriting the representation to account for changes of variable, which are needed to rescale and properly center the “transportation” measure. Once this is done and a suitable representation is obtained, we can analyze the asymptotics of $\hat{W}_N^m$ using this representation and weak convergence methods. The derivation of the representation is the purpose of the present chapter.

3.1 The Representation Formula

We first define the single particle (*underlying*) *transportation kernels* at the prelimit level by

$$\alpha_j^m(dy|x) = \mathbb{P} \left\{ Z_{\sigma_j^m}^m \in dy \mid Z_0^m = x \right\}, \quad (3.1)$$

where $x \in C_j^m \setminus C_{j+1}^m$, $\sigma_j^m = \inf\{i \geq 0 : Z_i^m \in C_{j+1}^m \cup (-\infty, -1/m]\}$, and where $\{Z_i^m\}$ is the underlying process model defined in (1.2). Note that α_j^m is a stochastic kernel on $(-\infty, -1/m] \cup C_{j+1}^m$ given $C_j^m \setminus C_{j+1}^m$.

Next we define a rescaled version of the original process according to

$$\check{Z}_{i+1}^m = \check{Z}_i^m + \theta_i \left(\check{Z}_i^m / m \right). \quad (3.2)$$

It is easy to check that $\check{Z}_i^m = mZ_i^m$ if $\check{Z}_0^m = mZ_0^m$. The thresholds for the rescaled process are stretched versions of those used for the original process: $\check{C}_j = mC_j^m = [j, \infty)$. With notation that follows that of the original IPS model, we denote the location of particles before and after resampling at stage j by $(\check{Y}_{j,1}^m, \dots, \check{Y}_{j,N}^m)$ and $(\check{X}_{j,1}^m, \dots, \check{X}_{j,N}^m)$. Then the related transportation kernel is defined by

$$\begin{aligned} \check{\alpha}_j^m(dy|x) &= \mathbb{P} \left\{ \check{Z}_{\check{\sigma}_j^m}^m \in dy \mid \check{Z}_0^m = x \right\} \\ &= \mathbb{P} \left\{ mZ_{\sigma_j^m}^m \in dy \mid mZ_0^m = x \right\} \\ &= \alpha_j^m(dy/m \mid x/m), \end{aligned}$$

where $x \in \check{C}_j \setminus \check{C}_{j+1}$ and $\check{\sigma}_j^m = \inf\{i \geq 0 : \check{Z}_i^m \in \check{C}_{j+1} \cup (-\infty, -1]\}$. In comparison with the original transportation kernel defined in (3.1), $\check{\alpha}_j^m$ is now a stochastic kernel on $(-\infty, -1] \cup \check{C}_{j+1}$ given $\check{C}_j \setminus \check{C}_{j+1}$. Note that both sets no longer depend on m . Later in this section a *centering* convention will be introduced, so that all the transportation kernels are on the same fixed space. *For the rest of this chapter we will focus on the rescaled processes and related measures. This is without loss, in that all quantities of interest are the same for both systems. With an abuse of notation, we omit the "m" from the notation.*

To obtain a control representation, we exploit an underlying Markovian structure. Slightly complicating the description and notation is the fact that the evolution interweaves the transportation and resampling steps, and it will be useful in the representation to keep them distinct. At any given stage j , $(X_{j,1}^m, \dots, X_{j,N}^m)$ takes

values in $C_j^N \subset \mathbb{R}^N$. Given $(X_{j,1}^m, \dots, X_{j,N}^m) = (x_1, \dots, x_N)$, the transition kernel for the transportation stage is

$$\Lambda_j^m(dy_1, \dots, dy_N | x_1, \dots, x_N) = \prod_{n=1}^N [1_{C_j \setminus C_{j+1}}(x_n) \alpha_j^m(dy_n | x_n) + 1_{C_{j+1}}(x_n) \delta_{x_n}(dy_n)] . \quad (3.3)$$

One can construct a measurable mapping $g_{j+1}^m = (g_{\text{num},j+1}^m, g_{\text{loc},j+1}^m)$ such that

$$\begin{aligned} g_{\text{num},j+1}^m(Y_{j+1,1}^m, \dots, Y_{j+1,N}^m) &= N_{j+1}^m \\ g_{\text{loc},j+1}^m(Y_{j+1,1}^m, \dots, Y_{j+1,N}^m) &= (\tilde{Y}_{j+1,1}^m, \dots, \tilde{Y}_{j+1,N_{j+1}^m}^m, 0, \dots, 0), \end{aligned}$$

where $\{\tilde{Y}_{j+1,i}^m\}$ are the locations of particles that made it up to the next level, and the trailing zeros are place holders which keep the dimension of the vector equal to N . In fact, there are many such mappings, in that they may produce different orders for the $\tilde{Y}_{j+1,k}^m$, and the particular one used is not important. Now suppose that given $N_{j+1}^m = n$ and $\tilde{Y}_{j+1,k}^m = \tilde{y}_k$, $k \leq n$, a random vector $(\sigma_1, \dots, \sigma_N)$ is generated according to the product measure

$$\Theta(\cdot | n) = \otimes_{i=1}^n \delta_i \otimes_{i=n+1}^N U_n, \quad (3.4)$$

where U_n is the uniform measure on $\{1, \dots, n\}$. The role of the $\otimes_{i=1}^n \delta_i$ is also as a placeholder, and simply says that the first n particles have already been determined, and thus resampling is only used to determine the location of the remaining $N - n$ particles. Then we can set

$$X_{j+1,k}^m = \begin{cases} \tilde{y}_k & k = 1, \dots, n \\ \tilde{y}_{\sigma_k} & k = n+1, \dots, N \end{cases} = \tilde{y}_{\sigma_k}.$$

There is a measurable map h_j^m that takes $(g_{j+1}^m(Y_{j+1,1}^m, \dots, Y_{j+1,N}^m), \sigma_1, \dots, \sigma_N)$ into $(X_{j+1,1}^m, \dots, X_{j+1,N}^m)$.

With these definitions, the IPS process can be built on the following probability space. Let $\mathcal{X} = \{(-\infty, -1] \cup [0, \infty)\}^N$ and $\mathcal{Y} = \{1, \dots, N\}^N$, and set

$$\mathcal{Z} = [\mathcal{X} \times \mathcal{Y}]^m.$$

To simplify the notation, we denote corresponding vectors by bold symbols, e.g., $\mathbf{x}_0 = (x_{0,1}, \dots, x_{0,N})$. We define a probability measure on $\mathcal{Z}$ by

$$\begin{aligned} \mathbb{P}^m(d\mathbf{y}_1, d\boldsymbol{\sigma}_1, \dots, d\mathbf{y}_m, d\boldsymbol{\sigma}_m) \\ = \Lambda_0^m(d\mathbf{y}_1 | \mathbf{x}_0) \Theta(d\boldsymbol{\sigma}_1 | g_{\text{num},1}^m(\mathbf{y}_1)) \Lambda_1^m(d\mathbf{y}_2 | h_1^m(g_1^m(\mathbf{y}_1), \boldsymbol{\sigma}_1)) \Theta(d\boldsymbol{\sigma}_2 | g_{\text{num},2}^m(\mathbf{y}_2)) \cdots \end{aligned}$$

Note that the distribution induced by $h_1^m(g_1^m(\mathbf{y}_1), \boldsymbol{\sigma}_1)$ when $\mathbf{y}_1$ and $\boldsymbol{\sigma}_1$ have distribution $\Lambda_0^m(d\mathbf{y}_1 | \mathbf{x}_0) \Theta(d\boldsymbol{\sigma}_1 | g_{\text{num},1}^m(\mathbf{y}_1))$ is the same as that of $(X_{1,1}^m, \dots, X_{1,N}^m)$ given $(X_{0,1}^m, \dots, X_{0,N}^m) = (x_{0,1}, \dots, x_{0,N})$. We can then expand the expectation of (1.8) as

$$\hat{W}_N^m = -\frac{1}{m} \log \int_{\mathcal{Z}} e^{2 \sum_{j=1}^m \log \left(\frac{g_{\text{num},j}^m(\mathbf{y}_j)}{N} \right)} \mathbb{P}^m(d\mathbf{y}_1, d\boldsymbol{\sigma}_1, \dots, d\mathbf{y}_m, d\boldsymbol{\sigma}_m).$$

Applying Proposition 0.5, the relative entropy representation for exponential integrals, we have

$$\begin{aligned} & -\frac{1}{m} \log \int_{\mathcal{Z}} e^{2 \sum_{j=1}^m \log \left(\frac{g_{\text{num},j}^m(\mathbf{y}_j)}{N} \right)} \mathbb{P}^m(d\mathbf{y}_1, d\boldsymbol{\sigma}_1, \dots, d\mathbf{y}_m, d\boldsymbol{\sigma}_m) \\ & = \inf_{\mathbb{Q}^m} \left[\frac{1}{m} R(\mathbb{Q}^m \| \mathbb{P}^m) - \frac{1}{m} \int_{\mathcal{Z}} 2 \sum_{j=1}^m \log \left(\frac{g_{\text{num},j}^m(\mathbf{y}_j)}{N} \right) \mathbb{Q}^m(d\mathbf{y}_1, d\boldsymbol{\sigma}_1, \dots, d\mathbf{y}_m, d\boldsymbol{\sigma}_m) \right]. \end{aligned}$$

We now factor $\mathbb{Q}^m(d\mathbf{y}_1, d\boldsymbol{\sigma}_1, \dots, d\mathbf{y}_m, d\boldsymbol{\sigma}_m)$ to reflect the way that $\mathbb{P}^m$ is factored. While $\mathbb{P}^m$ exhibits a type of Markovian structure, the same cannot be assumed of $\mathbb{Q}^m$. However, we can always factor it so that the resulting conditional distributions

do not anticipate the future. Thus we write

$$\mathbb{Q}^m(dy_1, d\sigma_1, \dots, dy_m, d\sigma_m) = \bar{\Lambda}_0^m(dy_1) \bar{\Theta}_1^m(d\sigma_1) \bar{\Lambda}_1^m(dy_2) \cdots \bar{\Lambda}_{m-1}^m(dy_m) \bar{\Theta}_m^m(d\sigma_m),$$

where $\bar{\Theta}_j^m(d\sigma_j) \in \mathcal{P}(\mathcal{Y})$ is implicitly a function of $(\mathbf{y}_1, \sigma_1, \dots, \mathbf{y}_j)$ and $\bar{\Lambda}_{j-1}^m(dy_j) \in \mathcal{P}((-\infty, -1] \cup C_j)$ is implicitly a function of $(\mathbf{y}_1, \sigma_1, \dots, \mathbf{y}_{j-1}, \sigma_{j-1})$. Denote the distribution of $(\bar{\mathbf{Y}}_1^m, \bar{\sigma}_1^m, \dots, \bar{\mathbf{Y}}_m^m, \bar{\sigma}_m^m)$ by $\mathbb{Q}^m$. Then using the chain rule for relative entropy Theorem 0.8 to decompose $R(\mathbb{Q}^m \parallel \mathbb{P}^m)$, we can write the variational representation as

$$\begin{aligned} \inf_{\{\bar{\Lambda}_j^m, \bar{\Theta}_{j+1}^m\}} \bar{\mathbb{E}} & \left[\frac{1}{m} \sum_{j=0}^{m-1} R(\bar{\Lambda}_j^m(\cdot) \parallel \Lambda_j^m(\cdot | h_j^m(g_j^m(\bar{\mathbf{Y}}_j^m), \bar{\sigma}_j^m))) \right. \\ & \left. + \frac{1}{m} \sum_{j=1}^m R(\bar{\Theta}_j^m(\cdot) \parallel \Theta(\cdot | g_{\text{num},j}^m(\bar{\mathbf{Y}}_j^m))) - \frac{1}{m} \sum_{j=1}^m 2 \log \left(\frac{g_{\text{num},j}^m(\bar{\mathbf{Y}}_j^m)}{N} \right) \right], \end{aligned} \quad (3.5)$$

where $\bar{\mathbb{E}}$ denotes expected value with respect to $\mathbb{Q}^m$.

To proceed with the analysis, we need to introduce a centering in the previous definitions. The usefulness of the centering is due to the fact that the environment presented to each of the transportation and resampling measures is in some sense stationary relative to the starting location C_j of the particles at stage j . Our goal is analyze what amounts to an empirical measure on the number and locations of particles that make it to successive levels. For this to work out, the various measures should be defined on a fixed space that does not depend on j . In fact, it is only the *difference* of particle positions before and after transportation that we need to record. By a change of variable we can redefine the transportation kernel α_j^m so that rather than being on $(-\infty, -1] \cup C_{j+1}$ given a point in $C_j \setminus C_{j+1}$, we consider them as measures on the fixed space $\Delta \cup [0, \infty)$ given $[0, 1)$, where Δ will be called the “cemetery” state and plays the role of $(-\infty, -1]$, while $[0, \infty)$ plays the role of C_{j+1} . We make the analogous change with respect to the N-particle transportation kernels Λ_j^m defined in (3.3). Under this convention, the mappings $g_j^m = (g_{\text{num},j}^m, g_{\text{loc},j}^m)$ and

h_j^m now have domains and ranges

$$\begin{aligned} g &: \{\Delta \cup [0, \infty)\}^N \rightarrow \{1, \dots, N\} \times [0, \infty)^N \\ h &: \{1, \dots, N\} \times [0, \infty)^N \times \mathcal{Y} \rightarrow [0, \infty)^N. \end{aligned}$$

Thus $g = (g_{\text{num}}, g_{\text{loc}})$, where g_{num} counts the number of particles that make it up to $[0, \infty)$ (termed success particles), and g_{loc} returns the locations of these particles.

Again we will abuse notation and use the same labels as before for these new centered measures. In terms of the *original random variables*, we can exhibit the centered and rescaled transportation kernel, for $x \in [0, 1)$, as

$$\alpha_j^m(A|x) = \begin{cases} \mathbb{P} \left\{ \left(mZ_{\sigma_j^m}^m - (j+1) \right) \in A \mid (mZ_0^m - j) = x \right\} & \text{if } A \subset [0, \infty) \\ \mathbb{P} \left\{ \left(mZ_{\sigma_j^m}^m - (j+1) \right) \in (-\infty, -1] \mid (mZ_0^m - j) = x \right\} & \text{if } A = \{\Delta\} \end{cases}. \quad (3.6)$$

The transportation kernel is then

$$\Lambda_j^m(dy_1, \dots, dy_N | x_1, \dots, x_N) = \prod_{n=1}^N \left[\mathbf{1}_{[0,1)}(x_n) \alpha_j^m(dy_n | x_n) + \mathbf{1}_{[1,\infty)}(x_n) \delta_{x_n-1}(dy_n) \right]. \quad (3.7)$$

Note that all the transportation kernels are measures on the same space, i.e., $\{\Delta \cup [0, \infty)\}^N$. We thereby obtain the following variational representation, which is now amenable to weak convergence analysis:

$$\begin{aligned} V_N^m &= \inf_{\{(\bar{\Lambda}_j^m, \bar{\Theta}_{j+1}^m)\}} \bar{\mathbb{E}} \left[\frac{1}{m} \sum_{j=0}^{m-1} R(\bar{\Lambda}_j^m(\cdot) \parallel \Lambda_j^m(\cdot | h(g(\bar{\mathbf{Y}}_j^m), \bar{\boldsymbol{\sigma}}_j^m))) \right. \\ &\quad \left. + \frac{1}{m} \sum_{j=1}^m R(\bar{\Theta}_j^m(\cdot) \parallel \Theta(\cdot | g_{\text{num}}(\bar{\mathbf{Y}}_j^m))) - \frac{1}{m} \sum_{j=1}^m 2 \log \left(\frac{g_{\text{num}}(\bar{\mathbf{Y}}_j^m)}{N} \right) \right], \end{aligned} \quad (3.8)$$

where V_N^m is referred as the minimal cost function.

Without loss we can restrict to controls for which the cost is finite. This implies the finiteness of the relative entropies, which in turn implies that the resampling

controls $\bar{\Theta}_j^m$ should have the same support as $\Theta(\cdot | g_{\text{num}}(\bar{\mathbf{Y}}_j^m))$, w.p.1. Recalling that $\Theta(\cdot | n) = \otimes_{i=1}^n \delta_i \otimes_{i=n+1}^N U_n$, this means that $\bar{\Theta}_j^m(\cdot | \omega)$ will inherit the property that the resampling control leaves the $\bar{N}_j^m = g_{\text{num}}(\bar{\mathbf{Y}}_j^m)$ particles that made it up alone, and sample from among these particles (though perhaps with a distribution not equal to the uniform distribution on $\{1, \dots, \bar{N}_j^m\}$) to determine the locations of the remaining $N - g_{\text{num}}(\bar{\mathbf{Y}}_j^m)$ particles.

We formally state the representation so derived.

Theorem 3.1 (Representation Theorem). *Let Θ, Λ_j^m and V_N^m be defined by (3.4), (3.7) and (3.8), respectively. Then for all $m \in \mathbb{N}$*

$$\hat{W}_N^m = -\frac{1}{m} \log \mathbb{E} \prod_{j=1}^m \left(\frac{N_j^m}{N} \right)^2$$

is equal to the minimal cost function V_N^m .

3.2 The Limiting Behavior of the Underlying Transportation Kernels

To evaluate the asymptotics of $\hat{W}_N^m$ and identify the rate function, we will need to identify the limit of Λ_j^m . The rescaled and centered transportation kernels $\alpha_j^m(dy | x)$ are related to the asymptotic first entrance distributions defined in (1.9). The following lemma makes the relation precise.

Lemma 3.2. *Assume Conditions 1.1 and 1.4. Define $\alpha_j^m(dy | x)$ by (3.6), and $\gamma(dy | x, t)$ by (1.9), and assume that $j/m \rightarrow t \in (0, 1)$ as $m \rightarrow \infty$. Then for $x \in [0, 1)$*

$$\alpha_j^m(\cdot | x) \Rightarrow \gamma(\cdot | x, t),$$

and thus for $x \in [0, \infty)$

$$1_{[0,1)}(x)\alpha_j^m(\cdot|x) + 1_{[1,\infty)}(x)\delta_{x-1}(\cdot) \Rightarrow \gamma(\cdot|x, t). \quad (3.9)$$

Moreover, for each $x \in [0, 1]$ $t \rightarrow \gamma(\cdot|x, t)$ is continuous for $t \in [0, 1]$.

Proof. We recall that $\alpha_j^m(\cdot|x)$ is defined in (3.6) by setting

$$Z_{i+1} = Z_i + \theta_i(Z_i/m), \quad Z_0 = j + x, \quad (3.10)$$

and then stopping this process the first time σ_j^m when either $Z_i \geq j + 1$ or $Z_i \leq -1$, with the latter event meaning that mass is assigned to Δ , and $\alpha_j^m(A|x) = \mathbb{P}\{Z_{\sigma_j^m} \in j + 1 + A\}$ when $A \subset [0, \infty)$. Let $b \in (0, 1)$, and define

$$\sigma_j^m(b) = \inf\{i \geq 0 : Z_i \geq j + 1 \text{ or } Z_i \leq jb\}.$$

The parameter b is used to localize the problem, so that we only have to pay attention to the process when Z_i/m is very close to t . Under the stability condition (Condition 1.1), given any $a > 0$, for all sufficiently large m we have $\mathbb{P}\{Z_{\sigma_j^m} \geq j + 1 \text{ and } Z_{\sigma_j^m(b)} \leq jb\} \leq a$. It follows that if we define “barred” measures by $\bar{\alpha}_j^m(A|x) = \mathbb{P}\{Z_{\sigma_j^m(b)} \in j + 1 + A\}$ and $\bar{\alpha}_j^m(\Delta|x) = \mathbb{P}\{Z_{\sigma_j^m(b)} < j + 1\}$, then modulo an error in the total variation norm of size $2a$, $\alpha_j^m(\cdot|x)$ and $\bar{\alpha}_j^m(\cdot|x)$ have the same weak limit. We can also define $\bar{\gamma}(\cdot|x, t)$ in the analogous way, and obtain the same total variation norm error bound. Therefore, it suffices to show that $\bar{\alpha}_j^m(\cdot|x) \Rightarrow \bar{\gamma}(\cdot|x, t)$.

Again using Condition 1.1, $\bar{\alpha}_j^m(A|x)$ can be characterized as the unique solution

to the integral equation

$$\begin{aligned}
\bar{\alpha}_j^m(A|x) &= \mathbb{P}\{Z_{\sigma_j^m(b)} \in j+1+A\} \\
&= \mathbb{P}\left\{\left\{Z_{\sigma_j^m(b)} \in j+1+A\right\} \left(1_{\{\sigma_j^m(b)>1\}} + 1_{\{\sigma_j^m(b)=1\}} \cap \{Z_{\sigma_j^m(b)} \geq j+1\}\right)\right\} \\
&= \int_{\{y:jb/m < x+y < 1\}} \bar{\alpha}_j^m(A|x+y) \mu\left(dy \left| \frac{j+x}{m} \right.\right) + \mu\left(A+1-x \left| \frac{j+x}{m} \right.\right).
\end{aligned} \tag{3.11}$$

Under Condition 1.4, given $b < 1$ there exists a $c_b < 1$ such that for $b^2t \leq s \leq t/b^2$

$$c_b \mu(dy|t) \leq \mu(dy|s) \leq \frac{1}{c_b} \mu(dy|t), \tag{3.12}$$

where $c_b \uparrow 1$ as $b \uparrow 1$.

For fixed b , choose m big enough that $[bj/m, (j+1)/m] \subset [b^2t, t/b^2]$. Define $D_k^m \doteq \{x_i : jb/m < \sum_{i=0}^\ell x_i < 1 \text{ for } \ell = 0, \dots, k\}$. Then for $n \in \mathbb{N}$ and all $x \in [bj/m, (j+1)/m]$, applying (3.12) and (3.11) recursively gives

$$\begin{aligned}
\bar{\alpha}_j^m(A|x) &\leq \left(\frac{1}{c_b}\right)^n \int_{D_n^m} \bar{\alpha}_j^m\left(A \left| \sum_{i=0}^n x_i \right.\right) \mu(dx_n|t) \cdots \mu(dx_1|t) \\
&\quad + \sum_{k=0}^{n-1} \left(\frac{1}{c_b}\right)^{k+1} \int_{D_k^m} \mu\left(A+1 - \sum_{i=0}^k x_i \left| t \right.\right) \mu(dx_k|t) \cdots \mu(dx_1|t),
\end{aligned} \tag{3.13}$$

where $x = x_0$. Note also that $\bar{\gamma}(\cdot|x, t)$ solves the equation analogous to (3.11), save that $\mu(dy|(j+x)/m)$ is replaced everywhere by $\mu(dy|t)$. Hence a calculation analogous to the one used in (3.13) gives

$$\begin{aligned}
\bar{\gamma}(A|x, t) &= \int_{D_n^m} \bar{\gamma}\left(A \left| \sum_{i=0}^n x_i, t \right.\right) \mu(dx_n|t) \cdots \mu(dx_1|t) \\
&\quad + \sum_{k=0}^{n-1} \int_{D_k^m} \mu\left(A+1 - \sum_{i=0}^k x_i \left| t \right.\right) \mu(dx_k|t) \cdots \mu(dx_1|t).
\end{aligned} \tag{3.14}$$

Condition 1.1 implies that for each $\varepsilon > 0$ there exists $M < \infty$, such that the probability that the random walk escapes from $[-1, j+1]$ in M steps is greater than ε uniformly in x , and thus

$$\mathbb{P} \{ \sigma_j^m(b) \leq M \mid Z_0 = x \} \geq \mathbb{P} \{ \sigma_j^m \leq M \mid Z_0 = x \} > \varepsilon.$$

On the other hand, the first term in (3.13) can be bounded above in terms of the probability of not escaping in n steps. In fact, for $n = kM + l$, $k \geq 1$, $l < M$

$$\begin{aligned} & \int_{D_n^m} \bar{\alpha}_j^m \left(A \left| \sum_{i=0}^n x_i \right. \right) \mu(dx_n|t) \cdots \mu(dx_1|t) \\ &= \mathbb{P} \{ \sigma_j^m(b) > n, Z_n \in j+1+A \mid x \} \\ &\leq (1-\varepsilon)^k \\ &= \left\{ [1-\varepsilon]^{1/(M+1)} \right\}^{(M+1)k} \\ &\leq \rho^n, \end{aligned}$$

where $\rho = \rho(\varepsilon, M) = [1-\varepsilon]^{1/(M+1)} < 1$. Repeating the analysis for $\bar{\gamma}$, we get

$$\int_{D_n^m} \gamma \left(A \left| \sum_{i=0}^n x_i \right. \right) \mu(dx_n|t) \cdots \mu(dx_1|t) \leq \rho^n.$$

Pick $b < 1$ such that c_b satisfies $\rho/c_b < 1$. Then the series in (3.13) is convergent, and for n large enough

$$\begin{aligned} & \bar{\alpha}_j^m(A|x) \\ &\leq \left(\frac{\rho}{c_b} \right)^n + \sum_{k=0}^{n-1} \left(\frac{1}{c_b} \right)^{k+1} \int_{D_k^m} \mu \left(A+1 - \sum_{i=0}^k x_i \mid t \right) \mu(dx_k|t) \cdots \mu(dx_1|t) \\ &\leq \left(\frac{\rho}{c_b} \right)^n + \left(\frac{1}{c_b} \right)^n \sum_{k=0}^{n-1} \int_{D_k^m} \mu \left(A+1 - \sum_{i=0}^k x_i \mid t \right) \mu(dx_k|t) \cdots \mu(dx_1|t) \\ &\leq \left(\frac{\rho}{c_b} \right)^n + \left(\frac{1}{c_b} \right)^n \bar{\gamma}(A|x, t). \end{aligned}$$

With a similar computation we obtain the lower bound

$$\begin{aligned}
\bar{\alpha}_j^m(A|x) &\geq c_b^n \int_{D_n^m} \bar{\alpha}_j^m \left(A \left| \sum_{i=0}^n x_i \right. \right) \mu(dx_n|t) \cdots \mu(dx_1|t) \\
&\quad + \sum_{k=0}^{n-1} c_b^{k+1} \int_{D_k^m} \mu \left(A + 1 - \sum_{i=0}^k x_i \left| t \right. \right) \mu(dx_k|t) \cdots \mu(dx_1|t) \\
&\geq c_b^n (\bar{\gamma}(A|x, t) - \rho^n).
\end{aligned}$$

Combining these two inequalities gives

$$c_b^n (\bar{\gamma}(A|x, t) - \rho^n) \leq \bar{\alpha}_j^m(\cdot|x) \leq (1/c_b)^n (\bar{\gamma}(A|x, t) + \rho^n).$$

The claim now follows by sending $m \rightarrow \infty$ (and using $j/m \rightarrow t$), then $b \uparrow 1$, and finally $n \rightarrow \infty$.

The last sentence of the lemma follows from a very similar construction, and is hence omitted. $\square$

Chapter 4

The Large Deviation Upper Bound

In this chapter we prove (2.2), which corresponds to a large deviation upper bound. We first rewrite that control representation (3.8) so that the controls and controlled process are in a space that is independent of m , which facilitates the use of weak convergence arguments. Tightness is addressed at the beginning of Section 4.2, and then weak convergence methods are used to relate the limits of the controls to the limits of the processes. It is this asymptotic relationship which identifies the local rate function, and in Section 4.3 these results are used to prove a large deviation upper bound with the given rate.

4.1 Notation and Preliminaries

In this section we make a few final adjustments to the relative entropy representation. A number of stochastic kernels indexed by $t \in [0, 1]$ will be needed. To simplify the notation, the definition at $t = 1$ is usually omitted. This is without loss, since these kernels are always integrated with respect to Lebesgue measure.

Recall the control representation (3.8). Let $\bar{\mathbf{X}}_j^m = h(g(\bar{\mathbf{Y}}_j^m), \bar{\boldsymbol{\sigma}}_j^m)$ and $\bar{N}_j^m = g_{\text{num}}(\bar{\mathbf{Y}}_j^m)$. These quantities have the exactly same interpretation as for the original process, but with respect to the controlled distributions. Thus $\bar{N}_j^m$ is the number

of particles making it to threshold j , and $\bar{\mathbf{X}}_j^m$ are the locations after resampling. We will also want to keep track of the dependence of the transportation control as a function of the state of the process $\bar{\mathbf{X}}_j^m$. For $m \in \mathbb{N}$ and $t \in [0, 1]$ define the stochastic kernel

$$\bar{\xi}^m(d\mathbf{x}, d\mathbf{y} | t) \doteq \delta_{\bar{\mathbf{X}}_j^m}(d\mathbf{x}) \times \bar{\Lambda}_j^m(d\mathbf{y}) \text{ if } t \in \left[\frac{j}{m}, \frac{j+1}{m} \right), \quad j = 0, \dots, m-1,$$

where $\bar{\Lambda}_j^m(d\mathbf{y}) = \bar{\Lambda}_j^m(d\mathbf{y} | \bar{\mathbf{Y}}_1, \bar{\boldsymbol{\sigma}}_1, \dots, \bar{\mathbf{Y}}_j, \bar{\boldsymbol{\sigma}}_j)$. Also define

$$\bar{\eta}^m(dn, d\boldsymbol{\sigma} | t) \doteq \delta_{\bar{N}_{j+1}^m}(dn) \times \bar{\Theta}_{j+1}^m(d\boldsymbol{\sigma}) \text{ if } t \in \left[\frac{j}{m}, \frac{j+1}{m} \right), \quad j = 0, \dots, m-1,$$

where $\bar{\Theta}_j^m(d\boldsymbol{\sigma}) = \bar{\Theta}_j^m(d\boldsymbol{\sigma} | \bar{\mathbf{Y}}_1, \bar{\boldsymbol{\sigma}}_1, \dots, \bar{\mathbf{Y}}_j)$. Random probability measures $\bar{\xi}^m \in \mathcal{P}([0, \infty)^N \times \{\Delta \cup [0, \infty)\}^N \times [0, 1])$ and $\bar{\eta}^m \in \mathcal{P}(\{1, \dots, N\} \times \mathcal{Y} \times [0, 1])$ are then defined by

$$\bar{\xi}^m(\mathbf{A}_1 \times \mathbf{A}_2 \times B) \doteq \int_B \bar{\xi}^m(\mathbf{A}_1 \times \mathbf{A}_2 | t) dt, \quad (4.1)$$

$$\bar{\eta}^m(C_1 \times C_2 \times B) \doteq \int_B \bar{\eta}^m(C_1 \times C_2 | t) dt \quad (4.2)$$

for Borel subsets $\mathbf{A}_1$ of $[0, \infty)^N$, $\mathbf{A}_2$ of $\{\Delta \cup [0, \infty)\}^N$, C_1 of $\{1, \dots, N\}$, C_2 of $\mathcal{Y}$, and B of $[0, 1]$. Again, we use the bold uppercase letter to denote the subsets of higher dimensions. Similarly, measures related to the distributions used to define the original process are defined by

$$r^m(d\mathbf{x}, d\mathbf{y} | t) \doteq \delta_{\bar{\mathbf{X}}_j^m}(d\mathbf{x}) \times \Lambda_j^m(d\mathbf{y} | \bar{\mathbf{X}}_j^m) \text{ if } t \in \left[\frac{j}{m}, \frac{j+1}{m} \right), \quad j = 0, \dots, m-1,$$

$$q^m(dn, d\boldsymbol{\sigma} | t) \doteq \delta_{\bar{N}_{j+1}^m}(dn) \times \Theta(d\boldsymbol{\sigma} | \bar{N}_{j+1}^m) \text{ if } t \in \left[\frac{j}{m}, \frac{j+1}{m} \right), \quad j = 0, \dots, m-1,$$

and

$$r^m(\mathbf{A}_1 \times \mathbf{A}_2 \times B) \doteq \int_B r^m(\mathbf{A}_1 \times \mathbf{A}_2 | t) dt, \quad (4.3)$$

$$q^m(C_1 \times C_2 \times B) \doteq \int_B q^m(C_1 \times C_2 | t) dt, \quad (4.4)$$

where $\mathbf{A}_i$, C_i , B are as above.

The chain rule for relative entropy Theorem 0.8 is then used to rewrite (3.8) in terms of these measures. To be precise, we repeatedly use the fact that for probability measures and stochastic kernels defined on the appropriate spaces

$$R(\mu(dz|u)\gamma(du) \parallel \nu(dz|u)\gamma(du)) = \int_{\mathcal{U}} R(\mu(\cdot|u) \parallel \nu(\cdot|u)) \gamma(du)$$

to insert the Dirac functions appearing in the definitions of $\bar{\xi}^m$, $\bar{\eta}^m$, r^m and q^m without any change in the cost. For $\ell \in \mathcal{P}(\{1, \dots, N\})$ recall the definition

$$\langle F^N, \ell \rangle = \sum_{n=0}^N -2 \log \left(\frac{n}{N} \right) \ell(n).$$

With the $1/m$ factors accounted for by the piecewise constant nature of the measures and the Lebesgue integral, (3.8) can be restated as

$$\hat{W}_N^m = \inf_{\{\bar{\xi}^m, \bar{\eta}^m\}} \bar{\mathbb{E}} \left\{ \int_0^1 [R(\bar{\xi}^m(\cdot|t) \parallel r^m(\cdot|t)) + R(\bar{\eta}^m(\cdot|t) \parallel q^m(\cdot|t))] dt + \langle F^N, \bar{L}^m \rangle \right\}, \quad (4.5)$$

where

$$\bar{L}^m = \frac{1}{m} \sum_{j=1}^m \delta_{\bar{N}_j^m} \quad (4.6)$$

is the controlled empirical measure.

The proof of Theorem 2.2 uses weak convergence arguments to analyze the limit of (4.5). Before proving the tightness results for our problem, it is useful to recall some criteria for the tightness of probability measures. In the lemmas to follow, $\mathcal{V}$

is a Polish space and A is an arbitrary index set. A *tightness function* on $\mathcal{V}$ is a mapping $g : \mathcal{V} \rightarrow [0, \infty]$ such that for each $M \in [0, \infty)$, $\{v : g(v) \leq M\}$ is compact. The first result is given in [9, Theorem A.3.17].

Lemma 4.1. *Let g be a tightness function on $\mathcal{V}$. Define a function $G(\theta)$ mapping $\mathcal{P}(\mathcal{V})$ into $\mathbb{R} \cup \{\infty\}$ by*

$$G(\theta) = \int_{\mathcal{V}} g d\theta.$$

Then the following conclusions hold.

- (a) *For each $M < \infty$, the set $Z_G(M) = \{\theta \in \mathcal{P}(\mathcal{V}) : G(\theta) \leq M\}$ is tight.*
- (b) *G is a tightness function on $\mathcal{P}(\mathcal{V})$.*

The next result can be found in [20, Theorem 6.1].

Lemma 4.2. *A family $\{\nu_a : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow \mathcal{P}(\mathcal{V}), a \in A\}$ of random probability measures on a Polish space $\mathcal{V}$ is tight if and only if the family of (deterministic) probability measures $\{\mathbb{E}\nu_a, a \in A\} \subset \mathcal{P}(\mathcal{V})$ is tight.*

The next lemma gives an elementary sufficient condition for the tightness of empirical measures.

Lemma 4.3. *Let $\{\mathbf{X}_j \in \mathbb{R}^N, j = 0, \dots, m-1\}$ be a collection of random vectors. Define the empirical measure $J^m = \frac{1}{m} \sum_{j=0}^{m-1} \delta_{\mathbf{X}_j}$. Suppose that given $\varepsilon > 0$ there exists $b < \infty$ such that for all $n \in \{1, \dots, N\}$*

$$\frac{1}{m} \sum_{j=0}^{m-1} \mathbb{P}(|X_{j,n}| > b) < \varepsilon.$$

Then $\{J^m\}$ is tight.

Proof. By Lemma 4.2, it suffices to show that given $\delta > 0$, there is a compact set $\mathbf{K}$ such that $\mathbb{E}J^m(\mathbf{K}^c) < \delta$ for all m . Given δ , let $\varepsilon = \delta/N$ and then choose b so that

the previous display holds. Letting $\mathbf{K} = [-b, b]^N$, we have

$$\mathbb{E}J^m(\mathbf{K}^c) = \frac{1}{m} \sum_{j=0}^{m-1} \mathbb{P}(\mathbf{X}_j \in \mathbf{K}^c) \leq \frac{1}{m} \sum_{j=0}^{m-1} \sum_{n=1}^N \mathbb{P}(|X_{j,n}| > b) < N\varepsilon = \delta.$$

□

4.2 Tightness and Characterization of Limits

We now proceed to prove the tightness of both the underlying measure and the *admissible control measures*. Note that, owing to the rescaling and centering, measures such as $\alpha_j^m(dy|x)$ are stochastic kernels on $\Delta \cup [0, \infty)$.

Proposition 4.4. *Assume Conditions 1.1 and 1.2. For each $m \in \mathbb{N}$, consider any admissible control sequence $\{\bar{\Lambda}_j^m, j = 0, \dots, m-1\}$ and $\{\bar{\Theta}_j^m, j = 1, \dots, m\}$ such that*

$$\begin{aligned} \sup_{m \in \mathbb{N}} \bar{\mathbb{E}} \left\{ \frac{1}{m} \sum_{j=0}^{m-1} R(\bar{\Lambda}_j^m(\cdot) \parallel \Lambda_j^m(\cdot | h(g(\bar{\mathbf{Y}}_j^m), \bar{\boldsymbol{\sigma}}_j^m))) \right\} &< \infty \\ \sup_{m \in \mathbb{N}} \bar{\mathbb{E}} \left\{ \frac{1}{m} \sum_{j=1}^m R(\bar{\Theta}_j^m(\cdot) \parallel \Theta(\cdot | g_{num}(\bar{\mathbf{Y}}_j^m))) \right\} &< \infty, \end{aligned}$$

where $\bar{\Lambda}_j^m(\cdot) = \bar{\Lambda}_j^m(\cdot | \bar{\mathbf{Y}}_1, \bar{\boldsymbol{\sigma}}_1, \dots, \bar{\mathbf{Y}}_j, \bar{\boldsymbol{\sigma}}_j)$ and $\bar{\Theta}_j^m(\cdot) = \bar{\Theta}_j^m(\cdot | \bar{\mathbf{Y}}_1, \bar{\boldsymbol{\sigma}}_1, \dots, \bar{\mathbf{Y}}_j)$. Then the following results hold.

1. $\{r^m, m \in \mathbb{N}\}$ with r^m defined in (4.3) is tight.
2. Let the admissible control measures $(\bar{\xi}^m, \bar{\eta}^m)$ be defined by equations (4.1) and (4.2). Then $\{(\bar{\xi}^m, \bar{\eta}^m), m \in \mathbb{N}\}$ is tight.

Proof. Part 1 (tightness of r_j^m). We divide the proof for the tightness of r^m into three steps.

Step 1: the tightness of α^m . Let $H(x, \alpha)$ be the moment generating function defined in (1.4). By Condition 1.2, $H(\alpha) \doteq \sup_{x \in [-1, 1]} H(x, \alpha) < \infty$. Given $j = 0, \dots, m-1$,

let $\{\hat{\theta}_{j,i}^m(x), i \in \mathbb{N}, x \in [-1, 1]\}$ be iid random fields with $\hat{\theta}_{j,i}^m(x) \sim \mu(\cdot | x + j/m)$ and $\{Z_{j,i}^m\}$ be the Markov chain defined by

$$Z_{j,i+1}^m = Z_{j,i}^m + \hat{\theta}_{j,i}^m(Z_{j,i}^m/m), \quad Z_{j,0}^m = z < 0.$$

Then for each fixed α , the sequence $T_{j,i}^m = \exp\{\alpha Z_{j,i}^m - iH(\alpha)\}$ is a supermartingale with respect to the filtration $\mathcal{F}_{j,i}^m = \sigma\{Z_{j,s}^m : s \leq i\}$. Indeed,

$$\begin{aligned} \mathbb{E}[T_{j,i+1}^m | \mathcal{F}_{j,i}^m] &= T_{j,i}^m \mathbb{E}\left[\exp\left\{\alpha \hat{\theta}_{j,i}^m(Z_{j,i}^m/m) - H(\alpha)\right\} | \mathcal{F}_{j,i}^m\right] \\ &= T_{j,i}^m \exp\left\{H(Z_{j,i}^m/m + j/m, \alpha) - H(\alpha)\right\} \\ &\leq T_{j,i}^m. \end{aligned}$$

Define the stopping time $\tau_j^m = \inf\{k > 0 : Z_{j,k}^m \notin (-j-1, 0)\}$. Then by the Optional Sampling Theorem, for $\alpha > 0$

$$\mathbb{E} \exp\left\{\alpha Z_{j,\tau_j^m}^m - \tau_j^m H(\alpha)\right\} \leq e^{\alpha z} < 1.$$

If we can prove that $H(\alpha^*) < 0$ for some $\alpha^* > 0$, then Chebyshev's inequality yields

$$\begin{aligned} \mathbb{P}\left\{Z_{j,\tau_j^m}^m \geq b\right\} &\leq e^{-\alpha^* b} \mathbb{E} \exp\left\{\alpha^* Z_{j,\tau_j^m}^m\right\} \\ &\leq e^{-\alpha^* b} \mathbb{E} \exp\left\{\alpha^* Z_{j,\tau_j^m}^m - \tau_j^m H(\alpha^*)\right\} \\ &\leq e^{-\alpha^* b}, \end{aligned}$$

which implies the tightness of $\{(Z_{j,\tau_j^m}^m) \vee 0, j = 1, \dots, m, m \in \mathbb{N}\}$. For later use, we note that $Z_{j,i}^m - z$ has the same distribution as the process $Z_i - (j+x)$ (equation (3.10)) that appears in the definition of $\alpha_j^m(\cdot|x)$ when $z = x - 1$, and so we have

$$\alpha_j^m([b, \infty)|x) \leq e^{-\alpha^* b}. \quad (4.7)$$

We are now left to show the existence of such α^* . According to Condition 1.1, $D_\alpha H(x, 0) = \mathbb{E}\hat{\theta}_{0,i}(x) < 0$ for all $x \in [-1, 1]$. Hence for each x there is $\alpha_x > 0$ such that $H(x, \alpha) < 0$ for all $\alpha \in (0, \alpha_x)$, and by the continuity of H we can also take α_x to be continuous in x . Since $[-1, 1]$ is compact, $\inf \alpha_x > 0$, hence such α^* exists.

Step 2: tightness of $\bar{\xi}_1^m$. By assumption, there exists $M < \infty$ such that

$$\sup_{m \in \mathbb{N}} \bar{\mathbb{E}} R(\bar{\xi}^m \| r^m) = \sup_{m \in \mathbb{N}} \bar{\mathbb{E}} \left\{ \frac{1}{m} \sum_{j=0}^{m-1} R(\bar{\Lambda}_j^m(\cdot) \| \Lambda_j^m(\cdot | h(g(\bar{\mathbf{Y}}_j^m), \bar{\boldsymbol{\sigma}}_j^m))) \right\} \leq M. \quad (4.8)$$

According to the variational formula in Proposition 0.6 or Lemma 0.7(a), for any measurable function ψ bounded from below and for each $m \in \mathbb{N}$, $j \leq m$

$$\bar{\mathbb{E}} \left\{ \int \psi d\bar{\Lambda}_j^m - \log \int e^\psi d\Lambda_j^m \right\} \leq \bar{\mathbb{E}} R(\bar{\Lambda}_j^m \| \Lambda_j^m). \quad (4.9)$$

Define the measurable function

$$\psi(\mathbf{y}) = \begin{cases} c & y_n > b \\ 0 & \text{otherwise} \end{cases}.$$

Then applying the inequality (4.7) to each component of $\bar{\mathbf{Y}}_j^m$ and (4.9) yield

$$\begin{aligned} c\bar{\mathbb{P}}\{\bar{Y}_{j,n}^m \geq b\} &= \bar{\mathbb{E}} \left\{ \int \psi d\bar{\Lambda}_j^m \right\} \\ &\leq \bar{\mathbb{E}} R(\bar{\Lambda}_j^m \| \Lambda_j^m) + \log [1 + e^c e^{-\alpha^* b}]. \end{aligned}$$

For $n \in \{1, \dots, N\}$ and any Borel set K of $[0, \infty)$, $\bar{X}_{j,n}^m \in K$ is possible only if $\bar{Y}_{j,n'}^m \in K$ for some $n' \in \{1, \dots, N\}$. Therefore, given any $\varepsilon > 0$ there are $b < \infty$ and $c < \infty$ such that

$$\frac{1}{m} \sum_{j=0}^{m-1} \bar{\mathbb{P}}\{\bar{X}_{j,n}^m \geq b\} \leq \frac{1}{m} \sum_{j=0}^{m-1} \sum_{n=1}^N \bar{\mathbb{P}}\{\bar{Y}_{j,n}^m \geq b\} \leq \frac{NM}{c} + \frac{N}{c} \log [1 + e^{c-\alpha^* b}] \leq \varepsilon. \quad (4.10)$$

With $\bar{\xi}_1^m$ playing the role of J^m , Lemma 4.3 then yields the tightness of $\{\bar{\xi}_1^m\}$.

Step 3: tightness of r^m . If we can show that the measures $\frac{1}{m} \sum_{j=0}^{m-1} \Lambda_j^m$ also tight, then tightness of $\{r^m\}$ will follow. From the definition of $\Lambda_j^m(\cdot|\mathbf{x})$ in (3.7) we have the upper bound

$$\Lambda_j^m(A_1 \times \cdots \times A_N | \mathbf{x}) \leq \sum_{n=1}^N [\mathbf{1}_{[0,1)}(x_n) \alpha_j^m(A_n | x_n) + \mathbf{1}_{[1,\infty)}(x_n) \delta_{x_n-1}(A_n)].$$

Recall that we have shown $\alpha_j^m([b, \infty)|x) \leq e^{-\alpha^* b}$. Given $\varepsilon > 0$, choose $b, c < \infty$ satisfying both $e^{-\alpha^* b} < \varepsilon$ and (4.10) and let $\mathbf{K} = [b, \infty)^N$. Then

$$\begin{aligned} & \bar{\mathbb{E}} \left\{ \frac{1}{m} \sum_{j=0}^{m-1} \Lambda_j^m(\mathbf{K}^c | \bar{\mathbf{X}}_j^m) \right\} \\ & \leq \bar{\mathbb{E}} \left\{ \frac{1}{m} \sum_{j=0}^{m-1} \sum_{n=1}^N \left[\mathbf{1}_{[0,1)}(\bar{X}_{j,n}^m) \alpha_j^m([b, \infty) | \bar{X}_{j,n}^m) + \mathbf{1}_{[1,\infty)}(\bar{X}_{j,n}^m) \delta_{\bar{X}_{j,n}^m-1}([b, \infty)) \right] \right\} \\ & < \varepsilon + \frac{1}{m} \sum_{j=0}^{m-1} \sum_{n=1}^N \bar{\mathbb{P}} \{ \bar{X}_{j,n}^m \geq b+1 \} \\ & < (N+1)\varepsilon. \end{aligned}$$

Applying Lemma 4.2, we complete the proof of the tightness of $\{r^m\}$.

Part 2 (tightness of $\{(\bar{\xi}^m, \eta_j^m)\}$). Tightness of r^m implies that for each $i \in \mathbb{N}$ there exists a compact set $\mathbf{K}_i$ of $\mathcal{X} \times \mathcal{Y}$ such that $\bar{\mathbb{E}} r^m(\mathbf{K}_i) \geq 1 - 2^{-i}$ for all m . Suppose we define a by

$$a(\cdot) = 1 + \sum_{i=1}^{\infty} i \mathbf{1}_{\mathbf{K}_i^c}(\cdot).$$

Then

$$\sup_{m \in \mathbb{N}} \bar{\mathbb{E}} \int a dr^m \leq 1 + \sum_{i=1}^{\infty} i 2^{-i} = M_1 < \infty.$$

Letting $U(\cdot) = \log a(\cdot)$, it follows that $U \geq 0$ and U has compact level sets. Thus

U is a tightness function such that

$$\sup_{m \in \mathbb{N}} \bar{\mathbb{E}} \int e^U dr^m \leq M_1.$$

Then, again applying the variational formula as in (4.9), Jensen's inequality, and the uniform bound (4.8),

$$\begin{aligned} \sup_{m \in \mathbb{N}} \bar{\mathbb{E}} \left\{ \int U d\bar{\xi}^m \right\} &\leq M + \sup_{m \in \mathbb{N}} \bar{\mathbb{E}} \left\{ \log \int e^U dr^m \right\} \\ &\leq M + \log \sup_{m \in \mathbb{N}} \bar{\mathbb{E}} \left\{ \int e^U dr^m \right\} \\ &< \infty. \end{aligned}$$

Lemma 4.1 then implies the tightness of $\{\bar{\xi}^m\}$.

Recall that $\mathcal{Y} = \{1, \dots, N\}^N$. Since the measures $\bar{\eta}^m$ take values in the compact set $\mathcal{P}(\{1, \dots, N\} \times \mathcal{Y} \times [0, 1])$, tightness of $\{\bar{\eta}^m\}$ is automatic. $\square$

Next we state the main weak convergence result. Recall that for a probability measure such as $\bar{\xi}$ on a product space, $\bar{\xi}_i$ will indicate the i th marginal. The first two statements are straightforward, and serve to establish notation and decompositions of the limit measures. The last two statements identify key equilibrium properties of the joint distribution of the location and number of the particles that make it to the next threshold. Specifically, equation (4.16) identifies a kind of local (in time or stage) stationary distribution on the particle locations after resampling and in the limit for the controlled IPS system, while (4.17) gives the corresponding distribution on the number of particles that make it to the next level. These relations, which link the equilibrium distributions to the limits of the controls, are the main ingredient in the definition of the local rate function for the IPS process.

Lemma 4.5. *Assume Conditions 1.1 and 1.4. Let $(\bar{\xi}^m, \bar{\eta}^m)$ be the admissible control measures defined by equations (4.1) and (4.2) and let r^m and q^m be defined by (4.3)*

and (4.4). Also, let $\bar{L}^m$ be the controlled empirical measure defined in (4.6), and assume that

$$M \doteq \sup_{m \in \mathbb{N}} \mathbb{E} \left\{ \int_0^1 [R(\bar{\xi}^m(\cdot|t) \| r^m(\cdot|t)) + R(\bar{\eta}^m(\cdot|t) \| q^m(\cdot|t))] dt \right\} < \infty. \quad (4.11)$$

Then every subsequence of $(\bar{\xi}^m, \bar{\eta}^m, r^m, q^m, \bar{L}^m)$ has a subsubsequence that converges in distribution to a limit $(\bar{\xi}, \bar{\eta}, r, q, \bar{L})$ with the following properties.

1. There exists a probability space $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{\mathbb{P}})$ such that the limiting quantities $(\bar{\xi}, \bar{\eta})$ can be realized, respectively, as stochastic kernels on $[0, \infty)^N \times \{\Delta \cup [0, \infty)\}^N \times [0, 1]$ and $\{1, \dots, N\} \times \mathcal{Y} \times [0, 1]$ given $\bar{\Omega}$. Furthermore there are stochastic kernels $\bar{\xi}(d\mathbf{x} \times d\mathbf{y} | t, \omega)$ and $\bar{\eta}(dn \times d\boldsymbol{\sigma} | t, \omega)$ on $[0, \infty)^N \times \{\Delta \cup [0, \infty)\}^N$ and $\{1, \dots, N\} \times \mathcal{Y} \times [0, 1]$ given $[0, 1] \times \bar{\Omega}$, such that

$$\begin{aligned} \bar{\xi}(d\mathbf{x} \times d\mathbf{y} \times dt | \omega) &= \bar{\xi}(d\mathbf{x} \times d\mathbf{y} | t, \omega) dt, \\ \bar{\eta}(dn \times d\boldsymbol{\sigma} \times dt | \omega) &= \bar{\eta}(dn \times d\boldsymbol{\sigma} | t, \omega) dt. \end{aligned}$$

2. Define marginal distributions by

$$\bar{\xi}_1(\mathbf{A}_1 | t, \omega) = \bar{\xi}(\mathbf{A}_1 \times \{\Delta \cup [0, \infty)\}^N | t, \omega), \quad \bar{\eta}_1(C_1 | t, \omega) = \bar{\eta}(C_1 \times \mathcal{Y} | t, \omega),$$

$$\bar{\xi}_2(\mathbf{A}_2 | t, \omega) = \bar{\xi}([0, \infty)^N \times \mathbf{A}_2 | t, \omega), \quad \bar{\eta}_2(C_2 | t, \omega) = \bar{\eta}(\{1, \dots, N\} \times C_2 | t, \omega).$$

Then there exist stochastic kernels $\bar{\Lambda}(d\mathbf{y} | \mathbf{x}, t, \omega)$ on $\{\Delta \cup [0, \infty)\}^N$ given $[0, \infty)^N \times [0, 1] \times \bar{\Omega}$ and $\bar{\Theta}(d\boldsymbol{\sigma} | n, t, \omega)$ on $\mathcal{Y}$ given $\{1, \dots, N\} \times [0, 1] \times \bar{\Omega}$, such that $\bar{\mathbb{P}}$ -a.s. for $\omega \in \bar{\Omega}$

$$\int_B \bar{\xi}(\mathbf{A}_1, \mathbf{A}_2 | t, \omega) dt = \int_B \int_{\mathbf{A}_1} \bar{\Lambda}(\mathbf{A}_2 | \mathbf{x}, t, \omega) \bar{\xi}_1(d\mathbf{x} | t, \omega) dt, \quad (4.12)$$

$$\int_B \bar{\eta}(C_1, C_2 | t, \omega) dt = \int_B \int_{C_1} \bar{\Theta}(C_2 | n, t, \omega) \bar{\eta}_1(dn | t, \omega) dt \quad (4.13)$$

for Borel subsets $\mathbf{A}_1$ of $[0, \infty)^N$, $\mathbf{A}_2$ of $\{\Delta \cup [0, \infty)\}^N$, C_1 of $\{1, \dots, N\}$, C_2 of $\mathcal{Y}$, and B of $[0, 1]$.

3. Statements analogous to Parts 1 and 2 above hold for r and q , and for $\bar{\mathbb{P}}$ -a.s. for $\omega \in \bar{\Omega}$

$$r(\mathbf{A}_1, \mathbf{A}_2, B | \omega) = \int_B \int_{\mathbf{A}_1} \Lambda(\mathbf{A}_2 | \mathbf{x}, t) \bar{\xi}_1(d\mathbf{x} | t, \omega) dt, \quad (4.14)$$

where $\Lambda(\cdot | \mathbf{x}, t) = \otimes_{i=1}^n \gamma(\cdot | x_i, t)$, and with γ defined by (1.9). Also, $q(dn, d\sigma, dt)$ has the form

$$q(C_1, C_2, B | \omega) = \int_B \int_{C_1} \Theta(C_2 | n) \bar{\eta}_1(dn | t, \omega) dt, \quad (4.15)$$

where $\Theta(\cdot | n) = \otimes_{i=1}^n \delta_i \otimes_{j=n+1}^N U_n$ is the product measure defined in (3.4), and $\mathbf{A}_i$, C_i and B take the same form as in Part 2.

4. $\bar{\mathbb{P}}$ -a.s. for $\omega \in \bar{\Omega}$, we have

$$\begin{aligned} \int_0^t \bar{\xi}_1(\mathbf{A} | s, \omega) ds &= \int_0^t \int \delta_{\{h(g(\mathbf{y}), \sigma)\}}(\mathbf{A}) \bar{\Theta}(d\sigma | g_{num}(\mathbf{y}), s, \omega) \bar{\xi}_2(d\mathbf{y} | s, \omega) ds \\ \int_0^t \bar{\eta}_1(\{n\} | s, \omega) ds &= \int_0^t \int \mathbf{1}_{\{g_{num}(\mathbf{y})=n\}} \bar{\xi}_2(d\mathbf{y} | s, \omega) ds. \end{aligned}$$

Therefore, a.s. with respect to Lebesgue measure

$$\bar{\xi}_1(\mathbf{A} | s, \omega) = \int \delta_{\{h(g(\mathbf{y}), \sigma)\}}(\mathbf{A}) \bar{\Theta}(d\sigma | g_{num}(\mathbf{y}), s, \omega) \bar{\xi}_2(d\mathbf{y} | s, \omega) \quad (4.16)$$

$$\bar{\eta}_1(\{n\} | s, \omega) = \int \mathbf{1}_{\{g_{num}(\mathbf{y})=n\}} \bar{\xi}_2(d\mathbf{y} | s, \omega). \quad (4.17)$$

Furthermore, for any subset C of $\{1, \dots, N\}$

$$\bar{L}(C) = \int_0^1 \bar{\eta}_1(C | t, \omega) dt. \quad (4.18)$$

Remark 4.6. We suppress ω variable in later discussion from time to time, but keep in mind that the measures are random.

Proof. Part 1. Since $(\bar{\xi}^m, \bar{\eta}^m) \rightarrow (\bar{\xi}, \bar{\eta})$ in distribution, by the Skorokhod Representation Theorem, we can assume that w.p.1 $(\bar{\xi}^m, \bar{\eta}^m) \Rightarrow (\bar{\xi}, \bar{\eta})$ on some probability space $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{\mathbb{P}})$. By definition, $\bar{\xi}^m([0, \infty)^N \times \{\Delta \cup [0, \infty)\}^N \times dt) = [\bar{\xi}^m]_3(dt) = \lambda(dt)$, where λ is Lebesgue measure on $[0, 1]$. It follows easily from the convergence that w.p.1 $[\bar{\xi}]_3(dt) = \lambda(dt)$. Hence by Theorem 0.9, there exists a stochastic kernel $\bar{\xi}(d\mathbf{x} \times d\mathbf{y} | t, \omega)$ on $[0, \infty)^N \times \{\Delta \cup [0, \infty)\}^N$ given $[0, 1] \times \bar{\Omega}$ such that w.p.1

$$\bar{\xi}(d\mathbf{x} \times d\mathbf{y} \times dt | \omega) = \bar{\xi}(d\mathbf{x} \times d\mathbf{y} | t, \omega) dt.$$

The decomposition of $\bar{\eta}$ is proved in the same fashion.

Part 2. The stated decompositions follow from another application of Theorem 0.9.

Part 3. If $(x, t) \rightarrow \gamma(\cdot | x, t)$ were continuous then $(\mathbf{x}, t) \rightarrow \Lambda(\cdot | \mathbf{x}, t)$ would also be continuous, and the form of r would follow from the fact that by Lemma 3.2 $\Lambda_j^m(\cdot | \mathbf{x}) \rightarrow \Lambda(\cdot | \mathbf{x}, t)$ uniformly, and the weak convergence $\bar{\xi}^m \Rightarrow \bar{\xi}$. Now in fact $(x, t) \rightarrow \gamma(\cdot | x, t)$ is not continuous, though under the given conditions it is continuous in t and $x \rightarrow \gamma(\cdot | x, t)$ is always right continuous and left continuous when x is not of the form $(0, 1] \cap [\mathbb{Z}/q]$ (see Lemmas 1.5 and 3.2). Thus to obtain the decomposition via Lemma 1.6, it is sufficient to show that under (4.11), $\bar{\mathbb{E}} \bar{\xi}_1^m(\{[K_\delta^c]^N\}^c)$ can be made as small as desired by choosing $\delta > 0$ sufficiently small, where $K_\delta = \cup_{k=1}^q ([k/q] - \delta, [k/q])$ and $K_\delta^c = [0, \infty) \setminus K_\delta$.

We first claim that it is sufficient to show that given $\varepsilon > 0$,

$$\bar{\mathbb{E}} \frac{1}{m} \sum_{n=1}^N \sum_{j=0}^{m-1} \bar{\Lambda}_{j,n}^m(K_\delta) < \varepsilon \quad (4.19)$$

for some $\delta > 0$, where $\bar{\Lambda}_{j,n}^m$ is the n th marginal of $\bar{\Lambda}_j^m$. Using the reasoning that lead

to (4.10), we have

$$\begin{aligned}
\bar{\mathbb{E}} \bar{\xi}_1^m(\{[K_\delta^c]^N\}^c) &\leq \frac{1}{m} \sum_{j=0}^{m-1} \sum_{n=1}^N \bar{\mathbb{P}}(\bar{X}_{j,n}^m \in K_\delta) \\
&\leq \frac{1}{m} \sum_{j=0}^{m-1} N \sum_{n=1}^N \bar{\mathbb{P}}(\bar{Y}_{j,n}^m \in K_\delta) \\
&= N \bar{\mathbb{E}} \frac{1}{m} \sum_{n=1}^N \sum_{j=0}^{m-1} \bar{\Lambda}_{j,n}^m(K_\delta).
\end{aligned}$$

We next show that for each $n = 1, \dots, N$, $\bar{\mathbb{E}} \frac{1}{m} \sum_{j=0}^{m-1} \bar{\Lambda}_{j,n}^m(K_\delta)$ can be made small, thus (4.19). Let $\Lambda_{j,n}^m(\cdot|x) = \mathbf{1}_{[0,1)}(x) \alpha_j^m(\cdot|x) + \mathbf{1}_{[1,\infty)}(x) \delta_{x-1}(\cdot)$. Applying once again Lemma 0.7(a) as in (4.9) and Jensen's inequality, for any bounded measurable function f

$$\begin{aligned}
&\bar{\mathbb{E}} \left\{ \int f d \left(\frac{1}{m} \sum_{j=0}^{m-1} \bar{\Lambda}_{j,n}^m(\cdot) \right) - \log \int e^f d \left(\frac{1}{m} \sum_{j=0}^{m-1} \Lambda_{j,n}^m(\cdot | h(g(\bar{\mathbf{Y}}_j^m), \bar{\boldsymbol{\sigma}}_j^m)) \right) \right\} \\
&\leq \bar{\mathbb{E}} \left\{ \frac{1}{m} \sum_{j=0}^{m-1} R(\bar{\Lambda}_j^m(\cdot) \| \Lambda_j^m(\cdot | h(g(\bar{\mathbf{Y}}_j^m), \bar{\boldsymbol{\sigma}}_j^m))) \right\} \\
&\leq M.
\end{aligned}$$

Assume that given ε , there exist $\delta > 0$ so that

$$\bar{\mathbb{E}} \frac{1}{m} \sum_{j=0}^{m-1} \Lambda_{j,n}^m(K_\delta | \bar{X}_{j,n}^m) < \varepsilon. \tag{4.20}$$

Substituting into the last display the function f that equals 0 on K_δ^c and $\log(1 + 1/\varepsilon)$

on K_δ , we have

$$\begin{aligned}
& \bar{\mathbb{E}} \left(\frac{1}{m} \sum_{j=0}^{m-1} \bar{\Lambda}_{j,n}^m(K_\delta) \right) \\
& \leq \frac{1}{\log(1 + 1/\varepsilon)} \left(M + \bar{\mathbb{E}} \log \left[1 + \frac{1}{\varepsilon} \left(\frac{1}{m} \sum_{j=0}^{m-1} \Lambda_{j,n}^m(K_\delta | \bar{X}_{j,n}^m) \right) \right] \right) \\
& \leq \frac{1}{\log(1 + 1/\varepsilon)} (M + \log 2).
\end{aligned}$$

This is made as small as desired by choosing $\varepsilon > 0$ small.

We complete by proving the claim (4.20). Recall that by definition $\Lambda_{j,n}^m(\cdot|x) = \mathbf{1}_{[0,1)}(x)\alpha_j^m(\cdot|x) + \mathbf{1}_{[1,\infty)}(x)\delta_{x-1}(\cdot)$, and that α_j^m has a uniformly bounded density with respect to Lebesgue measure everywhere except $(0, 1] \cap [\mathbb{Z}/q]$ (this follows the same proof as for $\gamma(\cdot|x, t)$). Thus by the right continuity of $x \mapsto \alpha_j^m(\cdot|x)$, $\alpha_j^m(K_\delta|x)$ is proportional to δ for all x save those in intervals of the form $([k/q], [k/q] + \delta)$. However, the only way a particular component of $\bar{\mathbf{X}}_j^m$ can be in this set (at least if $\delta < 1/q$) is if one of its summands in the random walk that ended at that component was not in the support of the discrete measure. When this happens, the distribution of this particular component has a bounded density with respect to Lebesgue measure, and hence the probability that it falls in K_δ is proportional to δ . Therefore, $\bar{\mathbb{E}} \frac{1}{m} \sum_{j=0}^{m-1} \Lambda_{j,n}^m(K_\delta | \bar{X}_{j,n}^m)$ is itself proportional to δ , which completes the argument.

The argument for (4.15) is similar but much simpler since the measure is on a discrete space and is omitted. Here we use that at the prelimit

$$q^m(dn, d\sigma, dt) = \delta_{\bar{N}_{\lceil mt \rceil}^m}(dn) \Theta(d\sigma | n) dt = [\bar{\eta}_1^m(dn | t) dt] \Theta(d\sigma | n).$$

Part 4. For $t \in [0, 1]$ define

$$\bar{J}_t^m = \frac{1}{m} \sum_{j=1}^{\lfloor mt \rfloor} \delta_{h(g(\bar{\mathbf{Y}}_j^m), \bar{\sigma}_j^m)}, \quad \bar{L}_t^m = \frac{1}{m} \sum_{j=1}^{\lfloor mt \rfloor} \delta_{g_{\text{num}}(\bar{\mathbf{Y}}_j^m)}.$$

Then for bounded and continuous functions u_1 mapping $[0, \infty)^N$ into $\mathbb{R}$ and u_2 mapping $\{1, \dots, N\}$ into $\mathbb{R}$, by definition

$$\begin{aligned} \int u_1(\mathbf{x}) \bar{J}_t^m(d\mathbf{x}) &= \frac{1}{m} \sum_{j=1}^{\lfloor mt \rfloor} u_1(\bar{\mathbf{X}}_j^m) = \int_0^{\lfloor mt \rfloor/m} \int u_1(\mathbf{x}) \bar{\xi}_1^m(d\mathbf{x}|s) ds, \\ \int u_2(n) \bar{L}_t^m(dn) &= \frac{1}{m} \sum_{j=1}^{\lfloor mt \rfloor} u_2(\bar{N}_j^m) = \int_0^{\lfloor mt \rfloor/m} \int u_2(n) \bar{\eta}_1^m(dn|s) ds. \end{aligned}$$

This implies

$$(\bar{J}_t^m(d\mathbf{x}), \bar{L}_t^m(dn)) \rightarrow \left(\int_0^t \bar{\xi}_1(d\mathbf{x}|s) ds, \int_0^t \bar{\eta}_1(dn|s) ds \right).$$

For $n \in \mathbb{N}$ and $j \in \{0, 1, \dots, m-1\}$, we define $\bar{\mathcal{F}}_j^m$ to be the σ -field generated by $\{(\bar{\mathbf{Y}}_k^m, \bar{\boldsymbol{\sigma}}_k^m), k \leq j\}$ and $\bar{\mathcal{G}}_j^m$ to be that generated by $\bar{\mathcal{F}}_{j-1}^m$ and $\bar{\mathbf{Y}}_j^m$. Then

$$\begin{aligned} u_1(h(g(\bar{\mathbf{Y}}_{j+1}^m), \bar{\boldsymbol{\sigma}}_{j+1}^m)) - \int u_1(h(g(\mathbf{y}), \boldsymbol{\sigma})) \bar{\Theta}_{j+1}^m(d\boldsymbol{\sigma}) \bar{\Lambda}_j^m(d\mathbf{y}), \\ u_2(g_{\text{num}}(\bar{\mathbf{Y}}_{j+1}^m)) - \int u_2(g_{\text{num}}(\mathbf{y})) \bar{\Lambda}_j^m(d\mathbf{y}) \end{aligned}$$

are martingale differences with respect to $\bar{\mathcal{F}}_j^m$ and $\bar{\mathcal{G}}_j^m$, respectively. Since

$$\begin{aligned} &\int u_1(\mathbf{x}) \bar{J}_t^m(d\mathbf{x}) - \int_0^{\lfloor mt \rfloor/m} \int u_1(h(g(\mathbf{y}), \boldsymbol{\sigma})) \bar{\eta}^m(d\boldsymbol{\sigma}|g_{\text{num}}(\mathbf{y}), s) \bar{\xi}_2^m(d\mathbf{y}|s) ds \\ &= \frac{1}{m} \sum_{j=1}^{\lfloor mt \rfloor} u_1(\bar{\mathbf{X}}_j^m) - \frac{1}{m} \sum_{j=0}^{\lfloor mt \rfloor-1} \int u_1(h(g(\mathbf{y}), \boldsymbol{\sigma})) \bar{\Theta}_{j+1}^m(d\boldsymbol{\sigma}) \bar{\Lambda}_j^m(d\mathbf{y}) \\ &= \frac{1}{m} \sum_{j=0}^{\lfloor mt \rfloor-1} \left[u_1(\bar{\mathbf{X}}_{j+1}^m) - \int u_1(h(g(\mathbf{y}), \boldsymbol{\sigma})) \bar{\Theta}_{j+1}^m(d\boldsymbol{\sigma}) \bar{\Lambda}_j^m(d\mathbf{y}) \right]. \end{aligned}$$

Denote the left hand side of the above equalities by R^m . Then Chebyshev's inequality

and a standard conditioning argument yield that for any $\varepsilon > 0$

$$\begin{aligned}
& \bar{\mathbb{P}} \{ |R^m| \geq \varepsilon \} \\
&= \bar{\mathbb{P}} \left\{ \left| \frac{1}{m} \sum_{j=0}^{\lfloor mt \rfloor - 1} \left[u_1(\bar{\mathbf{X}}_{j+1}^m) - \int u_1(h(g(\mathbf{y}), \boldsymbol{\sigma})) \bar{\Theta}_{j+1}^m(d\boldsymbol{\sigma}) \bar{\Lambda}_j^m(d\mathbf{y}) \right] \right| \geq \varepsilon \right\} \\
&\leq \frac{1}{\varepsilon^2} \bar{\mathbb{E}} \left\{ \frac{1}{m^2} \left(\sum_{j=0}^{\lfloor mt \rfloor - 1} \left[u_1(\bar{\mathbf{X}}_{j+1}^m) - \int u_1(h(g(\mathbf{y}), \boldsymbol{\sigma})) \bar{\Theta}_{j+1}^m(d\boldsymbol{\sigma}) \bar{\Lambda}_j^m(d\mathbf{y}) \right] \right)^2 \right\} \\
&= \frac{1}{\varepsilon^2} \bar{\mathbb{E}} \left\{ \frac{1}{m^2} \sum_{j=0}^{\lfloor mt \rfloor - 1} \left[u_1(\bar{\mathbf{X}}_{j+1}^m) - \int u_1(h(g(\mathbf{y}), \boldsymbol{\sigma})) \bar{\Theta}_{j+1}^m(d\boldsymbol{\sigma}) \bar{\Lambda}_j^m(d\mathbf{y}) \right]^2 \right\} \\
&\leq \frac{4 \|u\|_\infty^2}{m \varepsilon^2}.
\end{aligned}$$

A similar calculation can be done for $\bar{L}_t^m$, and so (4.16) and (4.17) follow. Lastly (4.18) follows from the fact that $\bar{L}_1^m = \bar{L}^m$. $\square$

4.3 Proof of the Lower Bound (2.2)

In this section, we introduce some related concepts, identify the rate function and give the prove for the first part to Theorem 2.2, i.e. (2.2).

Definition 4.7 (Feasible). *Probability measures $\xi \in \mathcal{P}([0, \infty)^N \times \{\Delta \cup [0, \infty)\}^N)$ and $\eta \in \mathcal{P}(\{1, \dots, N\} \times \mathcal{Y})$ are called feasible if the analogues of (4.16) and (4.17) hold, i.e.,*

$$\begin{aligned}
\xi_1(A) &= \int \delta_{h(g(\mathbf{y}), \boldsymbol{\sigma})}(A) \eta(d\boldsymbol{\sigma} | g_{num}(\mathbf{y})) \xi_2(d\mathbf{y}) \\
\eta_1(\{n\}) &= \int \mathbf{1}_{\{g_{num}(\mathbf{y})=n\}} \xi_2(d\mathbf{y}),
\end{aligned} \tag{4.21}$$

where ξ_i and η_i are the marginals and $\eta(d\boldsymbol{\sigma}|n)$ is defined by $\eta(dn \times d\boldsymbol{\sigma}) = \eta_1(dn) \eta(d\boldsymbol{\sigma}|n)$.

For $\phi \in \mathcal{P}(\{1, \dots, N\})$, define $A(\phi) \doteq \{\text{feasible } (\xi, \eta) \text{ such that } \eta_1 = \phi\}$. We

now give the form of *local rate function* G_N in Theorem 2.2:

$$G_N(t, \phi) = \inf_{(\xi, \eta) \in A(\phi)} \left\{ \int R(\xi(\cdot | \mathbf{x}) \| \Lambda(\cdot | \mathbf{x}, t)) \xi_1(d\mathbf{x}) + \int R(\eta(\cdot | n) \| \Theta(\cdot | n)) \eta_1(dn) \right\}, \quad (4.22)$$

where $\Lambda(A_1 \times \cdots \times A_n | \mathbf{x}, t) = \otimes_{i=1}^n \gamma(A_i | x_i, t)$.

Let $(\bar{\xi}^m, \bar{\eta}^m)$ give a value that is within $\varepsilon > 0$ of the infimum in (4.5). Tightness of $\{(\bar{\xi}^m, \bar{\eta}^m, r^m, q^m)\}$ has been proved along any sequence for which the relative entropy costs are bounded. For the purposes of proving a lower bound on $\hat{W}_N^m$, this can be assumed without loss. Tightness of $\{\bar{L}^m\}$ follows from the fact that $\bar{L}^m$ takes values in a compact set. We now evaluate the limit inferior of $\hat{W}_N^m$ along the subsequence of $n \in \mathbb{N}$ for which $(\bar{\xi}^m, \bar{\eta}^m, r^m, q^m, \bar{L}^m) \rightarrow (\bar{\xi}, \bar{\eta}, r, q, \bar{L})$. Then

$$\begin{aligned} & \liminf_{m \rightarrow \infty} \hat{W}_N^m + \varepsilon \\ & \geq \liminf_{m \rightarrow \infty} \mathbb{E} \left\{ \int_0^1 [R(\bar{\xi}^m(\cdot | t) \| r^m(\cdot | t)) + R(\bar{\eta}^m(\cdot | t) \| q^m(\cdot | t))] dt + \langle F^N, \bar{L}^m \rangle \right\} \\ & = \liminf_{m \rightarrow \infty} \mathbb{E} \{ R(\bar{\xi}^m \| r^m) + R(\bar{\eta}^m \| q^m) + \langle F^N, \bar{L}^m \rangle \} \\ & \geq \mathbb{E} \{ R(\bar{\xi} \| r) + R(\bar{\eta} \| q) + \langle F^N, \bar{L} \rangle \} \\ & = \mathbb{E} \left\{ \int_0^1 [R(\bar{\xi}(\cdot | t) \| r(\cdot | t)) + R(\bar{\eta}(\cdot | t) \| q(\cdot | t))] dt + \langle F^N, \bar{L} \rangle \right\} \\ & \geq \mathbb{E} \left\{ \int_0^1 G_N(t, \bar{\eta}_1(\cdot | t)) dt + \langle F^N, \bar{L} \rangle \right\} \\ & \geq \inf \left\{ \int_0^1 G_N(t, \phi(t)) dt + \langle F^N, \nu \rangle : \phi(\cdot) \in \mathcal{P}(\{1, \dots, N\}), \nu = \int_0^1 \phi(s) ds \right\}. \end{aligned}$$

The first line of this display is formula (4.5). The second line follows from the chain rule Lemma 0.7(f). The third line is a consequence of the convergence in distribution of $(\bar{\xi}^m, \bar{\eta}^m, r^m, q^m, \bar{L}^m) \rightarrow (\bar{\xi}, \bar{\eta}, r, q, \bar{L})$, the nonnegativity and lower semicontinuity of relative entropy, Fatou's Lemma, and the non-negativity of F^N . The forth line is again the chain rule, and the fifth is due to the almost sure feasibility implied by (4.16) and (4.17) and the definition of G_N . The final line follows from (4.18).

Note that the proof of the Laplace upper bound naturally provides a candidate for the rate function, which is defined for $\nu \in \mathcal{P}(\{1, \dots, N\})$ by

$$I_N(\nu) = \inf \left\{ \int_0^1 G_N(t, \phi(t)) dt : \nu = \int_0^1 \phi(s) ds \right\},$$

and with $G_N(t, \phi)$ defined as in (4.22).

Since $\varepsilon > 0$ is arbitrary, we have proved that every subsequence of the original sequence $\{\hat{W}_N^m, m \in \mathbb{N}\}$ has a subsubsequence satisfying

$$\liminf_{m \rightarrow \infty} \hat{W}_N^m \geq \inf_{\nu \in \mathcal{P}(\{1, \dots, N\})} \{I_N(\nu) + \langle F^N, \nu \rangle\}.$$

An argument by contradiction applied to an arbitrary subsequence shows that the entire sequence $\{\hat{W}_N^m, m \in \mathbb{N}\}$ satisfies this same lower limit.

Chapter 5

The Large Deviation Lower Bound

In this chapter we prove (2.3), which corresponds to a large deviation lower bound. To prove this bound, one starts with a near minimizer in the right hand side, and based on this constructs a control that can be applied in the prelimit stochastic control representation. A key qualitative property that is used to prove the convergence of the costs is some kind of ergodicity. As we will see, the nonstandard features of the problem that made the proof of the large deviation upper bound difficult complicate the analysis here as well. In particular, each of the following properties requires the introduction of new techniques and methods of analysis: the transition kernel on thresholds is in general not Feller; the transition kernels depend both on the large deviation parameter m and on the threshold index; and the underlying process includes the absorbing state Δ .

We begin by discussing the ergodicity properties of the limit transition kernels.

5.1 Some Important Assumptions and Ergodicity of the Limit Underlying Transition Kernel

To establish the large deviation lower bound, we need to introduce additional conditions on the underlying measures $\mu(dy|t)$ and limit transition kernel γ defined in

(1.9). Note that since γ contains a cemetery state Δ , the relevant ergodicity properties are with respect to the kernel that excludes the absorbing set. Note also that under Conditions 1.1, 1.3 and 1.4, $\gamma(\Delta|x, t)$ is bounded away from 0 and 1, uniformly in $t \in [0, 1]$ and $x \in [0, 1]$. Define a stochastic kernel $\tilde{\gamma}$ on $\mathcal{P}([0, \infty))$ given $[0, \infty) \times [0, 1]$ by

$$\tilde{\gamma}(A|x, t) = \frac{\gamma(A|x, t)}{1 - \gamma(\Delta|x, t)}$$

for Borel sets $A \subset [0, \infty)$.

Some type of transitivity or communication condition of the following sort is standard in the large deviation theory for the empirical measures of a Markov chain. Note that the assumption is made for each t due to the nonstationarity of the process, but under Condition 1.4 the condition holds for one t if and only if it holds for all $t \in [0, 1]$. Also, we can restrict to initial conditions $x \in [0, 1)$ since for initial conditions $x \geq 1$ the dynamics of the chain simply move the process to $x - 1$ with probability one.

Condition 5.1 (Transitivity). *For each $t \in [0, 1]$, there exist positive integers l_0 and n_0 such that for all $x, \zeta \in [0, 1)$*

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

Here, $\tilde{\gamma}^{(k)}$ is the k -step transition probability defined recursively by

$$\tilde{\gamma}^{(k+1)}(A|x, t) = \int \tilde{\gamma}(A|y, t) \tilde{\gamma}^{(k)}(dy|x, t)$$

for Borel sets A , with $\tilde{\gamma}^{(1)} = \tilde{\gamma}$.

An immediate consequence of Condition 5.1 is the *indecomposability* [1, Section 7.3] of $\tilde{\gamma}$ which is shown in the next proposition. The idea of the proof will be used again in the next section when we show that the IPS transition kernel, involving N particles, also satisfies a “transitivity” condition.

Proposition 5.2. *Assume Condition 5.1. Then for each $t \in [0, 1]$ $\tilde{\gamma}$ is indecomposable, i.e., there do not exist disjoint Borel subsets A_1 and A_2 of $[0, \infty)$ such that*

$$\tilde{\gamma}(A_1 | x, t) = 1 \text{ for all } x \in A_1 \text{ and } \tilde{\gamma}(A_2 | \zeta, t) = 1 \text{ for all } \zeta \in A_2. \quad (5.1)$$

Proof. We suppress t in the notation since it plays no role in the proof. Assume (5.1) holds. Then because the dynamics shift x to $x - 1$ when $x \geq 1$, there exist $x \in [0, 1) \cap A_1$ and $\zeta \in [0, 1) \cap A_2$ such that

$$\tilde{\gamma}(A_1 | x) = 1 \text{ and } \tilde{\gamma}(A_2 | \zeta) = 1.$$

It then follows from (5.1) that for any l and n in $\mathbb{N}$ we have $\tilde{\gamma}^{(l)}(A_1 | x) = 1$ and $\tilde{\gamma}^{(n)}(A_2 | \zeta) = 1$. According to Condition 5.1, the first of these equalities with $l = l_0$ guarantees that there exists $j \geq n_0$ such that $\tilde{\gamma}^{(j)}(A_1 | \zeta) > 0$. Since A_1 and A_2 are disjoint, for $\zeta \in A_2$ this inequality is incompatible with the equality $\tilde{\gamma}^{(n)}(A_2 | \zeta) = 1$ for any $n \in \mathbb{N}$. Hence the indecomposability of $\tilde{\gamma}(\cdot | x)$ follows. $\square$

In the proof of the lower bound, ergodicity will be used to derive various limits. However, owing to the nonstationarity there will be contributions from several different Markov processes, with the final state for one serving as the initial state for the next. For this reason, we need the following type of condition which ensures that the bounds coming from each process are independent of the starting state. See the discussion in [9, Section 8.4].

Condition 5.3 (Uniformity). *For any $t \in [0, 1]$ and $x \in [0, 1)$ there exists a probability measure β on $[0, \infty)$, $k \in \mathbb{N}$, and $c > 0$ such that for all $\zeta \in [0, 1)$ satisfying $|x - \zeta| < c$ and for all Borel sets A*

$$\sum_{i=1}^k \tilde{\gamma}^{(i)}(A | \zeta, t) \geq c\beta(A).$$

5.2 The Construction of Ergodic IPS Transition Kernels

As mentioned in the previous section, the underlying transition kernel γ contains a cemetery state. In order to introduce additional ergodicity property, we focused our discussion on $\tilde{\gamma}$, a conditioned version of γ by eliminating the cemetery state Δ . In this section, we move on to a group of IPS-type processes, which share the same transportation measure related to γ but have different resampling mechanisms. Again, such processes contain a cemetery states $\{\Delta\}^N$. However, based on the assumptions of $\tilde{\gamma}$ in the previous section, the renormalized version of the IPS-type processes also satisfy transitivity and uniformity conditions. Furthermore, such processes are geometrically ergodic.

Let $\mathcal{T}_1 = [0, \infty)^N$, $\mathcal{T}_2 = \{[0, \infty), \Delta\}^N$. Recall that for any admissible control with finite cost, the transportation control must be supported on $\mathcal{T}_2 \setminus \{\Delta\}^N$, i.e., with probability one at least one particle makes it to the next threshold. Hence it is relevant to study the renormalized transportation measure on $\mathcal{T}_2 \setminus \{\Delta\}^N$ defined by

$$\tilde{\Lambda}(\cdot | \mathbf{x}, t) = \frac{\Lambda(\cdot | \mathbf{x}, t)}{1 - \Lambda(\{\Delta\}^N | \mathbf{x}, t)} = \frac{\otimes_{i=1}^N \gamma(\cdot | x_i, t)}{1 - \otimes_{i=1}^N \gamma(\Delta | x_i, t)}. \quad (5.2)$$

Denote $c(\mathbf{x}, t) = 1 - \otimes_{i=1}^N \gamma(\Delta | x_i, t)$. Since $\gamma(\Delta | x, t)$ is bounded away from 0 and 1 uniformly in $t \in [0, 1]$ and $x \in [0, 1)$, $c(\mathbf{x}, t)$ is bounded away from 0 and 1 uniformly in $t \in [0, 1]$ and $\mathbf{x} \in \mathcal{T}_1 \setminus [1, \infty)^N$. We now consider $\tilde{\Lambda}$ together with any resampling measure that satisfies $\tilde{\Theta} \ll \Theta$, and denote the transition kernel by $\tilde{p}(\mathbf{x}, \cdot | t)$ on $\mathcal{T}_1$ given $\mathcal{T}_1$, i.e.

$$\tilde{p}(\mathbf{x}, \cdot | t) = \int_{\mathcal{T}_2 \times \mathcal{Y}} \delta_{h(g(\mathbf{y}), \boldsymbol{\sigma})}(\cdot) \tilde{\Theta}(d\boldsymbol{\sigma} | g_{\text{num}}(\mathbf{y})) \tilde{\Lambda}(d\mathbf{y} | \mathbf{x}, t). \quad (5.3)$$

Note the ambiguity of the notation: $\tilde{p}$ can be associated with a number of resampling

measures but always uses the transportation measure $\tilde{\Lambda}$.

Proposition 5.4. *Consider any resampling measure $\tilde{\Theta}$ such that $\tilde{\Theta} \ll \Theta$, and assume Conditions 1.1, 1.2 and 1.4. Then for each $t \in [0, 1]$, $\tilde{p}(\mathbf{x}, \cdot | t)$ has an invariant measure.*

Proof. The main difficulty is due to the fact that the process does not satisfy the Feller property. For $m \in \mathbb{N}$, $j = 0, \dots, m-1$, and $\tilde{\mathbf{X}}_0^m = 0$, define random variables $\tilde{\mathbf{X}}_0^m, \tilde{\mathbf{Y}}_1^m, \tilde{\mathbf{X}}_1^m, \tilde{\mathbf{Y}}_2^m, \dots$ recursively via the requirements

$$\begin{aligned} \mathbb{P} \left\{ \tilde{\mathbf{Y}}_{j+1}^m \in \cdot \mid \tilde{\mathbf{X}}_j^m = \mathbf{x} \right\} &= \tilde{\Lambda}(\cdot | \mathbf{x}, t), \quad \tilde{N}_{j+1}^m = g_{\text{num}}(\tilde{\mathbf{Y}}_{j+1}^m) \\ \mathbb{P} \left\{ \boldsymbol{\sigma}_{j+1}^m \in \cdot \mid \tilde{N}_{j+1}^m = n \right\} &= \tilde{\Theta}(\cdot | n, t), \quad \tilde{\mathbf{X}}_{j+1}^m = h(g(\tilde{\mathbf{Y}}_{j+1}^m), \boldsymbol{\sigma}_{j+1}^m). \end{aligned}$$

Let

$$\tilde{J}^m = \frac{1}{m} \sum_{j=1}^m \delta_{\tilde{\mathbf{X}}_j^m},$$

and denote

$$\tilde{\xi}^m(d\mathbf{x}, d\mathbf{y}) = \frac{1}{m} \sum_{j=1}^m \delta_{\tilde{\mathbf{X}}_j^m}(d\mathbf{x}) \times \tilde{\Lambda}(d\mathbf{y} | \mathbf{x}, t).$$

Then for any bounded and continuous function $u : \mathcal{T}_1 \rightarrow \mathbb{R}$,

$$\int u(\mathbf{x}) \tilde{J}^m(d\mathbf{x}) = \frac{1}{m} \sum_{j=1}^m u(\tilde{\mathbf{X}}_j^m) = \int u_1(\mathbf{x}) \tilde{\xi}_1^m(d\mathbf{x}).$$

Note that by construction $R(\tilde{\Lambda}(\cdot | \mathbf{x}, t) \| \Lambda(\cdot | \mathbf{x}, t)) = -\log c(\mathbf{x}, t) < \infty$ for all $\mathbf{x}$ and t , and $\tilde{\Theta} \ll \Theta$ implies $R(\tilde{\Theta}(\cdot | n, t) \| \Theta(\cdot | n)) < \infty$ since $\tilde{\Theta}$ and Θ are probability measures on a finite set. Thus by Proposition 4.4 $\{\tilde{\xi}^m\}$ is tight. Let $\tilde{\xi}^m$ converge in distribution to a limit $\tilde{\xi}$ along a subsequence. Along that same subsequence obviously

$$\tilde{J}^m(d\mathbf{x}) \rightarrow \tilde{\xi}_1(d\mathbf{x}).$$

We can then use the martingale difference argument and Condition 1.4 to deal with

the places where the Feller property fails just as in Part 4 of Lemma 4.5, and obtain

$$\int u(\mathbf{x}) \tilde{\xi}_1(d\mathbf{x}) = \int \int u(h(g(\mathbf{y}), \boldsymbol{\sigma})) \tilde{\Theta}(d\boldsymbol{\sigma} | g_{\text{num}}(\mathbf{y})) \tilde{\Lambda}(d\mathbf{y} | \mathbf{x}, t) \tilde{\xi}_1(d\mathbf{x})$$

w.p.1. We now use the facts that there is a countable set of measure determining functions u to conclude that w.p.1

$$\tilde{\xi}_1(\mathbf{A}) = \int \tilde{p}(\mathbf{x}, \mathbf{A} | t) \tilde{\xi}_1(d\mathbf{x})$$

for all Borel sets $\mathbf{A}$, i.e., $\tilde{\xi}_1$ is an invariant measure of $\tilde{p}(\mathbf{x}, \cdot | t)$. $\square$

For $t \in [0, 1]$, let $\tilde{\pi}(d\mathbf{x} | t)$ denote any of the invariant distributions of $\tilde{p}(\mathbf{x}, \cdot | t)$ just constructed. The next lemma shows that there is in fact only one such invariant probability and that the associated Markov process is geometrically ergodic, as well as other properties that will be needed later.

Lemma 5.5. *Consider any resampling measure $\tilde{\Theta}$ such that $\tilde{\Theta} \ll \Theta$, and assume Conditions 1.1, 1.2, 1.4, 5.1 and 5.3. The following conclusion hold for each $t \in [0, 1]$.*

1. *The Markov process associated with the transition kernel $\tilde{p}(\mathbf{x}, \cdot | t)$ and initial distribution $\tilde{\pi}(\cdot | t)$ is ergodic.*
2. *The transition kernel $\tilde{p}$ satisfies a transitivity condition of the same form as $\tilde{\gamma}$: there exist positive integers n_0 and l_0*

$$\sum_{i=l_0}^{\infty} \frac{1}{2^i} \tilde{p}^{(i)}(\mathbf{x}, d\mathbf{x}' | t) \ll \sum_{j=n_0}^{\infty} \frac{1}{2^j} \tilde{p}^{(j)}(\boldsymbol{\zeta}, d\boldsymbol{\zeta}' | t),$$

for $\mathbf{x}, \boldsymbol{\zeta} \in [0, 1)^N$.

3. *The transition kernel $\tilde{p}(\mathbf{x}, \cdot | t)$ satisfies the following uniformity condition. Given $\mathbf{x} \in [0, 1)^N$ there exists a probability measure $\tilde{\beta}$ on $[0, \infty)^N$, $\tilde{k} \in \mathbb{N}$ and $\tilde{c} > 0$*

such that for all $\zeta \in [0, 1)^N$ satisfying $\|\zeta - \mathbf{x}\| \leq \tilde{c}$ and for all Borel sets $\mathbf{A}$,

$$\sum_{i=1}^{\tilde{k}} \tilde{p}^{(i)}(\mathbf{x}, \mathbf{A} | t) \geq \tilde{c} \tilde{\beta}(\mathbf{A}).$$

In particular, the precompact set $[0, 1)^N$ is a small set [21, (5.14)] of the Markov process associated with transition kernel $\tilde{p}$.

4. The Markov process associated with transition kernel $\tilde{p}$ is geometrically ergodic.
5. Let $\mathbf{A}$ be a Borel set of $\mathcal{T}_1$ such that $\tilde{p}^{(l_0)}(\mathbf{x}_0, \mathbf{A} | t) > 0$ for some $\mathbf{x}_0 \in \mathcal{T}_1$ and $l_0 \in \mathbb{N}$. Then $\tilde{\pi}(\mathbf{A} | t) > 0$.

Proof. To simplify the notation we omit t .

Part 1. We first show that $\tilde{p}(\mathbf{x}, \cdot)$ is indecomposable. Ergodicity then follows from [9, Theorem A.4.5]. The proof is by contradiction. Suppose that there exist disjoint sets $\mathbf{A}_i \subset \mathcal{T}_1$, such that $\tilde{p}(\mathbf{x}_i, \mathbf{A}_i) = 1$ for all $\mathbf{x}_i \in \mathbf{A}_i$, $i = 1, 2$. Define

$$B_i = \bigcup_{\mathbf{x}_i \in \mathbf{A}_i} \bigcup_{n=1}^N \{x_n : \mathbf{x}_i = (x_1, \dots, x_N)\},$$

i.e., B_i is the collection of all possible individual particle locations for joint N -particle locations $\mathbf{x}_i$ in $\mathbf{A}_i$. Note that $\mathbf{A}_i \subset \{B_i\}^N$, and therefore $\tilde{p}(\mathbf{x}_i, \{B_i\}^N) = 1$ for all $\mathbf{x}_i \in \mathbf{A}_i$. Since $\gamma(\Delta | x_i) < 1$ for all $x_i \in B_i$, we must have $\gamma(B_i | x_i) > 0$ for all $x_i \in B_i$.

First, we claim that $\tilde{\gamma}(B_i | x_i) = 1$ for all $x_i \in B_i$, $i = 1, 2$. If not, then without loss of generality we may assume $i = 1$, and that there exists $z \in B_1$ and $C_1 \subset [0, \infty)$ with $C_1 \cap B_1 = \emptyset$, such that $\gamma(C_1 | z) > 0$. Let $\bar{\mathbf{x}} \in \mathbf{A}_1$ have z as its n th component. Consider the set

$$\mathbf{C}_1 = \{B_1\}^{n-1} \otimes C_1 \otimes \{B_1\}^{N-n} \subset \mathcal{T}_1.$$

By construction $\mathbf{C}_1 \cap \mathbf{A}_1 = \emptyset$, but

$$\begin{aligned} \tilde{p}(\bar{\mathbf{x}}, \mathbf{C}_1) &\geq \int_{\mathbf{C}_1 \times \mathcal{Y}} \delta_{h(g(\mathbf{y}), \boldsymbol{\sigma})}(\mathbf{C}_1) \tilde{\Theta}(d\boldsymbol{\sigma} | g_{\text{num}}(\mathbf{y})) \tilde{\Lambda}(d\mathbf{y} | \bar{\mathbf{x}}) \\ &\geq \gamma(C_1 | z) \prod_{j=1, j \neq n}^N \gamma(B_i | \bar{x}_j) \\ &> 0, \end{aligned}$$

which is a contradiction. Hence, $\tilde{\gamma}(B_i | x) = 1$ for all $x \in B_i$.

With this property established, if B_1 and B_2 were disjoint then we would have a contradiction to Proposition 5.2. We claim even more, which is that $[0, 1) \cap B_i, i = 1, 2$ are not disjoint. Indeed, if $x \in B_1 \cap B_2$, then it is automatic that $x - \lfloor x \rfloor \in [0, 1) \cap B_1 \cap B_2$.

Thus we may suppose that there exists $x_0 \in B_1 \cap B_2 \cap [0, 1)$, which, without loss of generality, is the first component of some $\mathbf{x}_i \in \mathbf{A}_i$, for $i = 1, 2$. Let

$$k = \max_{i,n} \{ \lceil x_{i,n} \rceil : \mathbf{x}_i = (x_{i,1}, \dots, x_{i,N}) \}. \quad (5.4)$$

We now construct an event $\mathbf{E}$ which satisfies $\tilde{p}^{(k)}(\mathbf{x}_i, \mathbf{E}) > 0$ for $i = 1, 2$. An event on path space will be constructed, which will imply the existence of an event of the desired form. We start at time 0 with particles in locations $\mathbf{x}_i$ with $i = 1, 2$. At each time $\ell = 0, \dots, k-1$, we require that any particle that is a descendant of the particle originally in the first location be returned to $[0, \infty)$, i.e., not to Δ . This can always be done with uniformly positive probability. All other particles are dealt with as follows:

- If a particle is in $[0, 1)$ but not a descendant then it moves to the state Δ .
- If a particle is at $x \in [1, \infty)$ then it is moved to $x - 1$.

Again both events can be realized with positive probability. The definition of k in (5.4) guarantees that at the end of k steps, all the descendants other than those

originate from the x_0 will be sent to the cemetery state Δ . In other words, we have identified an event with strictly positive probability defined in terms of the descendants of the particle originally at x_0 , all of which have exactly the same distribution for the two different initial conditions, and hence such a set $\mathbf{E}$ exists.

Part 2. We now prove that $\tilde{p}(\mathbf{x}, d\mathbf{y})$ also satisfies a transitivity condition with the same l_0 as in Condition 5.1. That is, we need to show:

there exists a fixed $\bar{n}_0 < \infty$ such that, given any $\mathbf{x} \in [0, 1)^N$, $\boldsymbol{\zeta} \in [0, 1)^N$ and $\mathbf{A} \subset \mathcal{T}_1$ with $\tilde{p}^{(i)}(\mathbf{x}, \mathbf{A}) > 0$ for some $i \geq l_0$, there is $j \geq \bar{n}_0$ such that $\tilde{p}^{(j)}(\boldsymbol{\zeta}, \mathbf{A}) > 0$.

Without loss of generality, we can assume $\mathbf{A} \subset \mathcal{T}_1$ is a rectangle, i.e., $\mathbf{A} = A_1 \times \cdots \times A_N$. In this case, for each A_n there exists $y_n \in \{x_1, \dots, x_N\}$ such that $\tilde{\gamma}^{(i)}(A_n | y_n) > 0$ for the same $i \geq l_0$. Condition 5.1 implies that for $\zeta_n \in [0, 1)$ there exists $j_n \geq n_0$ such that $\tilde{\gamma}^{(j_n)}(A_n | \zeta_n) > 0$. The problem now becomes finding a common j for all ζ_n .

Using the same occupation measure argument as in Proposition 5.4, one can construct an invariant distribution for $\tilde{\gamma}$. Let $\tilde{\kappa}$ denote this measure. We claim that if $\tilde{\gamma}^{(i)}(A | x) > 0$ for some $i \geq l_0$, $x \in [0, \infty)$ and Borel set A , then $\tilde{\kappa}(A) > 0$. This is due to the following standard argument. Define the function ι mapping $[0, 1)$ into $\mathbb{N}$ by

$$\iota(\zeta) = \min \{j \in \mathbb{N} : j \geq n_0 \text{ and } \tilde{\gamma}^{(j)}(A | \zeta) > 0\}.$$

Condition 5.1 guarantees that $\iota(\zeta)$ is finite for all ζ . Furthermore, $\iota(\zeta)$ is measurable. $[0, 1)$ can be written as disjoint union of Borel sets $\Sigma^{(j)}$, where

$$\Sigma^{(j)} = \{\zeta : \iota(\zeta) = j\}.$$

It must be true that $\tilde{\kappa}(\Sigma^{(j)}) > 0$ for some $j \in \mathbb{N}$ satisfying $j \geq n_0$. The definition of

$\Sigma^{(j)}$ implies that $\tilde{\gamma}^{(j)}(A|\zeta) > 0$ for all $\zeta \in \Sigma^{(j)}$. Hence,

$$\tilde{\kappa}(A) = \int \tilde{\gamma}^{(j)}(A|\zeta) \tilde{\kappa}(d\zeta) \geq \int_{\Sigma^{(j)}} \tilde{\gamma}^{(j)}(A|\zeta) \tilde{\kappa}(d\zeta) > 0.$$

This proves the claim.

Continuing with the same set A , we note that since $\tilde{\gamma}$ possesses an invariant probability measure and satisfies the transitivity condition, the corresponding Markov chain is ergodic. It follows from the Ergodic Theorem that

$$\int_A \frac{1}{n} \sum_{i=1}^n \tilde{\gamma}^{(i)}(A|\zeta) \tilde{\kappa}(d\zeta) \rightarrow \tilde{\kappa}(A) > 0,$$

and hence there is $k \in \mathbb{N}$ such that reasoning then gives

$$\int_A \tilde{\gamma}^{(k)}(A|\zeta) \tilde{\kappa}(d\zeta) > 0. \quad (5.5)$$

We now return to the proof of the transitivity. According to the previous paragraph, for each n there is $k_n \in \mathbb{N}$ such that (5.5) holds with k and A replaced by k_n and A_n . Also, we can find $\{c_1, \dots, c_N\} \in \mathbb{N}^N$ satisfying

$$j_1 + c_1 k_1 = \dots = j_N + c_N k_N \geq n_0.$$

It then follows that with $\bar{n}_0$ equal to the common value $j_n + c_n k_n$,

$$\begin{aligned} & \tilde{p}^{(\bar{n}_0)}(\zeta, \mathbf{A}) \\ & \geq \prod_{n=1}^N \left[\int_{A_n} \int_{A_n} \dots \int_{A_n} \tilde{\gamma}^{(j_n)}(dz_0|\zeta_n) \tilde{\gamma}^{(k_n)}(dz_1|z_0) \dots \tilde{\gamma}^{(k_n)}(dz_{c_n}|z_{c_n-1}) \right] \\ & > 0. \end{aligned}$$

Part 3. Let $d = \sup_{x \in [0,1], t \in [0,1]} \gamma(\Delta|x, t) < 1$. Without loss of generality, assume

that $\mathbf{A}$ is a rectangle, i.e., $\mathbf{A} = A_1 \times \cdots \times A_N$. Then

$$\tilde{p}^{(i)}(\boldsymbol{\zeta}, \mathbf{A}) \geq (1-d)^{iN} \otimes_{n=1}^N \tilde{\gamma}^{(i)}(A_n | \zeta_n).$$

Let $\tilde{k} = k$, $\tilde{c} \in (0, c(1-d)^{kN})$ and $\tilde{\beta}$ be the product of β . The uniformity property then follows for small enough but positive $\tilde{c}$ by Condition 5.3.

Part 4. It is easy to check that $\tilde{p}$ is ψ -irreducible [21, Proposition 4.2.2] by the uniformity property proved in Part 3. Furthermore, $[0, 1)^N$ is a *small set* of the process. Geometric ergodicity then follows from [21, Theorem 15.0.1].

Part 5. The proof is essentially the same as the argument used in Part 2 to show that $\tilde{\gamma}^{(i)}(A|x) > 0$ implies $\tilde{\kappa}(A) > 0$. See also [9, Lemma 8.6.2]. $\square$

5.3 The Construction of Ergodic Controls

In this section, we will show how to construct controls from the renormalized underlying transition kernel $\tilde{p}$ which inherits the ergodicity property, as in Lemma 5.6. Moreover, how to reconstruct a (state-dependent) random walk from a given transition kernel of a single particle, as in Lemma 5.7.

The following lemma tells us that given any target distribution on the number of successful particles, we can find a control with cost close to the optimum and for which the associated Markov process is geometrically ergodic and the transportation cost is bounded as a function of $\mathbf{x}$. Let

$$\tilde{\phi}(\{n\} | t) = \int \mathbf{1}_{\{g_{\text{num}}(\mathbf{y})=n\}} \tilde{\Lambda}(d\mathbf{y} | \mathbf{x}, t) \tilde{\pi}(d\mathbf{x} | t), \quad (5.6)$$

the distribution of the number of success particles associated with the renormalized measures of the underlying process model and a resampling distribution $\tilde{\Theta} \ll \Theta$, i.e., $(\tilde{\Lambda}, \tilde{\Theta}, \tilde{\pi})$. In the lemma $\|\cdot\|_v$ denotes the total variation norm. Again we suppress the t variable in the argument.

Lemma 5.6. *Assume Conditions 1.1, 1.3, 1.4, 5.1 and 5.3 and consider any $\varphi \in \mathcal{P}\{1, \dots, N\}$ satisfying $G_N(t, \varphi) < \infty$, where G_N is the local rate function defined in (4.22).*

1. *For $(\xi, \eta) \in A(\varphi)$ suppose that*

$$R(\xi \parallel \xi_1 \otimes \Lambda) + R(\eta \parallel \varphi \otimes \Theta) < \infty,$$

and recall that by the definition of feasibility $\xi_1 = [\xi]_1$ is an invariant distribution of the associated transition kernel. Let $\eta = \varphi \otimes \tilde{\Theta}$ define $\tilde{\Theta}$ and let $\tilde{\pi}$ be invariant for $(\tilde{\Lambda}, \tilde{\Theta})$. Then $\xi_1 \ll \tilde{\pi}$.

2. *Given $\delta > 0$ there exists $\bar{\phi} \in \mathcal{P}\{1, \dots, N\}$ such that $\|\bar{\phi} - \varphi\|_v \leq \delta$, and furthermore there exists $(\bar{\xi}, \bar{\eta}) \in A(\bar{\phi})$ such that the associated controlled Markov sequence $\bar{\mathbf{X}}_j$ is geometrically ergodic, and*

$$G_N(t, \bar{\phi}) \leq R(\bar{\xi} \parallel \bar{\xi}_1 \otimes \Lambda) + R(\bar{\eta} \parallel \bar{\phi} \otimes \Theta) \leq G_N(t, \varphi) + \delta. \quad (5.7)$$

3. *There exists a sequence of admissible controls $(\bar{\xi}^k, \bar{\eta})$ such that the associated transition kernels are geometrically ergodic and the transportation cost $R(\bar{\Lambda}^k(\cdot | \mathbf{x}) \parallel \Lambda(\cdot | \mathbf{x}))$ is bounded as a function of $\mathbf{x}$. Furthermore,*

$$R(\bar{\Lambda}^k(\cdot | \mathbf{x}) \parallel \Lambda(\cdot | \mathbf{x})) \rightarrow R(\bar{\Lambda}(\cdot | \mathbf{x}) \parallel \Lambda(\cdot | \mathbf{x})) \quad (5.8)$$

and $\|\bar{\xi}_1^k - \bar{\xi}_1\|_v \rightarrow 0$, where $\bar{\xi}^k = \bar{\xi}_1^k \otimes \bar{\Lambda}^k$, $\bar{\xi} = \bar{\xi}_1 \otimes \bar{\Lambda}$

We place the proof for Part 3 of the Lemma in the appendix at the end of this chapter, since it requires a proposition that will be stated later in the next section.

Proof. Part 1. Define $\bar{\Lambda}$ by the decompositions $\xi = \xi_1 \otimes \bar{\Lambda}$ and $\tilde{\Theta}$ defined as in the statement. Let $\bar{p}(\mathbf{x}, \cdot)$ and $\tilde{p}(\mathbf{x}, \cdot)$ be the transition kernels associated with $(\bar{\Lambda}, \tilde{\Theta})$

and $(\tilde{\Lambda}, \tilde{\Theta})$ according to the analogues of (5.3). Then applying Part 4 of Lemma 4.5 and Lemma 0.7(g) to get the first inequality

$$\begin{aligned}
& \int_{\mathcal{T}_1} R(\bar{p}(\mathbf{x}, \cdot) \parallel \tilde{p}(\mathbf{x}, \cdot)) \xi_1(d\mathbf{x}) \\
&= \int_{\mathcal{T}_1} \left[\int_{\mathcal{T}_1 \times \mathcal{Y} \times \{\mathcal{T}_2 \setminus \{\Delta^N\}\}} \delta_{h(g(\mathbf{y}), \boldsymbol{\sigma})}(d\mathbf{x}') \tilde{\Theta}(d\boldsymbol{\sigma} | g_{\text{num}}(\mathbf{y})) \bar{\Lambda}(d\mathbf{y} | \mathbf{x}) \right. \\
&\quad \times \log \frac{d \int_{\mathcal{Y} \times \{\mathcal{T}_2 \setminus \{\Delta^N\}\}} \delta_{h(g(\mathbf{y}), \boldsymbol{\sigma})}(\cdot) \tilde{\Theta}(d\boldsymbol{\sigma} | g_{\text{num}}(\mathbf{y})) \bar{\Lambda}(d\mathbf{y} | \mathbf{x})}{d \int_{\mathcal{Y} \times \{\mathcal{T}_2 \setminus \{\Delta^N\}\}} \delta_{h(g(\mathbf{y}), \boldsymbol{\sigma})}(\cdot) \tilde{\Theta}(d\boldsymbol{\sigma} | g_{\text{num}}(\mathbf{y})) \tilde{\Lambda}(d\mathbf{y} | \mathbf{x})}(\mathbf{x}') \Big] \xi_1(d\mathbf{x}) \\
&\leq \int_{\mathcal{T}_1} \left[\int_{\mathcal{T}_2 \setminus \{\Delta^N\}} \bar{\Lambda}(d\mathbf{y} | \mathbf{x}) \log \frac{d\bar{\Lambda}(\cdot | \mathbf{x})}{d\tilde{\Lambda}(\cdot | \mathbf{x})}(\mathbf{y}) \right] \xi_1(d\mathbf{x}) \\
&= \int_{\mathcal{T}_1} \left[\int_{\mathcal{T}_2 \setminus \{\Delta^N\}} \bar{\Lambda}(d\mathbf{y} | \mathbf{x}) \log \frac{d\bar{\Lambda}(\cdot | \mathbf{x})}{d\Lambda(\cdot | \mathbf{x})}(\mathbf{y}) \right] \xi_1(d\mathbf{x}) \\
&\quad + \int_{\mathcal{T}_1} \left[\int_{\mathcal{T}_2 \setminus \{\Delta^N\}} \bar{\Lambda}(d\mathbf{y} | \mathbf{x}) \log \frac{d\Lambda(\cdot | \mathbf{x})}{d\tilde{\Lambda}(\cdot | \mathbf{x})}(\mathbf{y}) \right] \xi_1(d\mathbf{x}) \\
&= \int_{\mathcal{T}_1} R(\bar{\Lambda}(\cdot | \mathbf{x}) \parallel \Lambda(\cdot | \mathbf{x})) \xi_1(d\mathbf{x}) + \int_{\mathcal{T}_1} \log c(\mathbf{x}) \xi_1(d\mathbf{x}) \\
&< \infty.
\end{aligned}$$

Hence the set $\mathbf{C} = \{\mathbf{x} \in \mathcal{T}_1 : \bar{p}(\mathbf{x}, \cdot) \ll \tilde{p}(\mathbf{x}, \cdot)\}$ satisfies $\xi_1(\mathbf{C}) = 1$. If we redefine $\bar{p}(\mathbf{x}, \cdot)$ to equal $\tilde{p}(\mathbf{x}, \cdot)$ for $\mathbf{x} \in \mathbf{C}^c$, then ξ_1 is also an invariant measure of the new transition kernel. Let $\bar{p}$ also denote this redefined kernel.

One can show by induction that

$$\bar{p}^{(l_0)}(\mathbf{x}, \cdot) \ll \tilde{p}^{(l_0)}(\mathbf{x}, \cdot) \quad \text{for all } \mathbf{x} \in \mathbf{C},$$

where l_0 is the number appearing in Lemma 5.5. Therefore for any Borel set $\mathbf{A}$ with $\xi_1(\mathbf{A}) > 0$, iterating $\xi_1 \bar{p} = \xi_1$ yields

$$\int \bar{p}^{(l_0)}(\mathbf{x}, \mathbf{A}) \xi_1(d\mathbf{x}) = \xi_1(\mathbf{A}) > 0.$$

Hence there exists a Borel set $\mathbf{B}$ such that $\xi_1(\mathbf{B}) > 0$ and $\bar{p}^{(l_0)}(\mathbf{x}, \mathbf{A}) > 0$ for all

$\mathbf{x} \in B$. It follows that $\tilde{p}^{(l_0)}(\mathbf{x}_0, \mathbf{A}) > 0$ for some $\mathbf{x}_0 \in \mathbf{B} \cap \mathbf{C}$. By part 5 of the previous lemma $\tilde{\pi}(\mathbf{A}) > 0$, and therefore $\xi_1 \ll \tilde{\pi}$.

Part 2. For the given $\tilde{\Theta}$ define $\tilde{\phi}$ by (5.6). For $a \in (0, 1)$, the probability measure $\varphi^a = (1 - a)\varphi + a\tilde{\phi}$ satisfies $\|\varphi^a - \varphi\|_v = a\|\varphi - \tilde{\phi}\|_v \leq 2a$. Hence to satisfy the first claim we can take $\bar{\phi} = \varphi^a$ for any $a \in (0, \delta/2]$. Next define probability measures for $a \in (0, \delta/2]$ by

$$\bar{\xi} = (1 - a)\xi + a\tilde{\pi} \otimes \tilde{\Lambda}, \text{ and } \bar{\eta} = (1 - a)\eta + a\tilde{\phi} \otimes \tilde{\Theta}. \quad (5.9)$$

Then $\bar{\xi}_1 = (1 - a)\xi_1 + a\tilde{\pi}$ and $\bar{\eta} = \bar{\phi} \otimes \tilde{\Theta}$. We first claim that $(\bar{\xi}, \bar{\eta}) \in A(\bar{\phi})$, i.e., $(\bar{\xi}, \bar{\eta})$ is feasible. Indeed,

$$\begin{aligned} & \int \delta_{h(g(\mathbf{y}), \boldsymbol{\sigma})}(\mathbf{A}) \tilde{\Theta}(d\boldsymbol{\sigma} | g_{\text{num}}(\mathbf{y})) \bar{\xi}_2(d\mathbf{y}) \\ &= \int \delta_{h(g(\mathbf{y}), \boldsymbol{\sigma})}(\mathbf{A}) \tilde{\Theta}(d\boldsymbol{\sigma} | g_{\text{num}}(\mathbf{y})) \left\{ (1 - a)\xi_2 + a \left[\tilde{\pi} \otimes \tilde{\Lambda} \right]_2 \right\} \\ &= [(1 - a)\xi_1 + a\tilde{\pi}](\mathbf{A}) \\ &= \bar{\xi}_1(\mathbf{A}) \end{aligned}$$

and

$$\begin{aligned} \int \mathbf{1}_{\{g_{\text{num}}(\mathbf{y})=n\}} \bar{\xi}_2(d\mathbf{y}) &= \int \mathbf{1}_{\{g_{\text{num}}(\mathbf{y})=n\}} \left\{ (1 - a)\xi_2 + a \left[\tilde{\pi} \otimes \tilde{\Lambda} \right]_2 \right\} \\ &= \left[(1 - a)\varphi + a\tilde{\phi} \right](\{n\}) \\ &= \bar{\phi}(\{n\}). \end{aligned}$$

Let q and $\bar{p}$ be the transition kernels defined as in (5.3) associated with the transportation and the resampling controls induced by (ξ, η) and $(\bar{\xi}, \bar{\eta})$ (note that $\bar{p}$ is based on $\bar{\xi}$ defined in (5.9) and is not the same as in Part 1). Then ξ_1 and $\bar{\xi}_1$ are

invariant measures of q and $\bar{p}$, respectively. Furthermore,

$$\bar{\xi}_1 \otimes \bar{p} = (1 - a) \xi_1 \otimes q + a \tilde{\pi} \otimes \tilde{p}. \quad (5.10)$$

We next verify (5.7). By the convexity of relative entropy

$$\begin{aligned} G_N(t, \bar{\phi}) &\leq R(\bar{\xi} \parallel \bar{\xi}_1 \otimes \Lambda) + R(\bar{\eta} \parallel \bar{\phi} \otimes \Theta) \\ &= R\left((1 - a) \xi + a \tilde{\pi} \otimes \tilde{\Lambda} \parallel (1 - a) \xi_1 \otimes \Lambda + a \tilde{\pi} \otimes \Lambda\right) \\ &\quad + R\left((1 - a) \eta + a \tilde{\phi} \otimes \tilde{\Theta} \parallel (1 - a) \varphi \otimes \Theta + a \tilde{\phi} \otimes \Theta\right) \\ &\leq (1 - a) [R(\xi \parallel \xi_1 \otimes \Lambda) + R(\eta \parallel \varphi \otimes \Theta)] \\ &\quad + a \left[R\left(\tilde{\pi} \otimes \tilde{\Lambda} \parallel \tilde{\pi} \otimes \Lambda\right) + R\left(\tilde{\phi} \otimes \tilde{\Theta} \parallel \tilde{\phi} \otimes \Theta\right) \right]. \end{aligned}$$

Using the facts that Θ is a measure on a finite space, that $\tilde{\Theta} \ll \Theta$, and finally the explicit relation between $\tilde{\Lambda}$ and Λ , we have that

$$R\left(\tilde{\pi} \otimes \tilde{\Lambda} \parallel \tilde{\pi} \otimes \Lambda\right) + R\left(\tilde{\phi} \otimes \tilde{\Theta} \parallel \tilde{\phi} \otimes \Theta\right) < \infty.$$

Therefore for sufficiently small $a > 0$,

$$G_N(t, \bar{\phi}) \leq \inf_{(\xi, \eta) \in A(\varphi)} [R(\xi \parallel \xi_1 \otimes \Lambda) + R(\eta \parallel \varphi \otimes \Theta)] + \delta = G_N(t, \varphi) + \delta.$$

We are now left to show the geometric ergodicity of the associated controlled sequence $\bar{\mathbf{X}}_j$. In fact, if we can prove that $\bar{p}$ and $\tilde{p}$ are mutually absolute continuous $\tilde{\pi}$ -a.s., then repeating the proof as in Lemma 5.5 yields the ergodicity of $\bar{\mathbf{X}}_j$. Since $\bar{\xi}_1 \ll \tilde{\pi}$ and $\bar{\xi}_1(\mathbf{A}) \geq a \tilde{\pi}(\mathbf{A})$ for all Borel sets $\mathbf{A}$ of $\mathcal{T}_1$, the Radon-Nikodym derivative $f(\mathbf{x}) = (d\bar{\xi}_1/d\tilde{\pi})(\mathbf{x})$ satisfies $f(\mathbf{x}) \in [a, \infty)$ $\tilde{\pi}$ -a.s. for $\mathbf{x} \in \mathcal{T}_1$. For any Borel sets $\mathbf{A}$ and $\mathbf{B}$, by (5.10)

$$\int_{\mathbf{A}} \bar{p}(\mathbf{x}, \mathbf{B}) f(\mathbf{x}) \tilde{\pi}(d\mathbf{x}) \geq a \tilde{\pi} \otimes \tilde{p}(\mathbf{A} \times \mathbf{B}) = a \int_{\mathbf{A}} \tilde{p}(\mathbf{x}, \mathbf{B}) \tilde{\pi}(d\mathbf{x}).$$

This implies that $\tilde{\pi}$ -a.s. for all $\mathbf{x} \in \mathcal{T}_1$

$$\bar{p}(\mathbf{x}, \mathbf{B}) \geq \frac{a}{f(\mathbf{x})} \tilde{p}(\mathbf{x}, \mathbf{B}). \quad (5.11)$$

On the other hand we know that $\int R(\bar{p}(\mathbf{x}, \cdot) \| \tilde{p}(\mathbf{x}, \cdot)) \bar{\xi}_1(d\mathbf{x}) < \infty$, which yields

$$\bar{p}(\mathbf{x}, \cdot) \ll \tilde{p}(\mathbf{x}, \cdot), \bar{\xi}_1\text{-a.s.} \quad (5.12)$$

Since $\bar{\xi}_1$ and $\tilde{\pi}$ are mutually absolute continuous, there exists a Borel set $\mathbf{C}$ such that $\bar{\xi}_1(\mathbf{C}) = \tilde{\pi}(\mathbf{C}) = 0$ and both (5.11) and (5.12) hold on the complement of $\mathbf{C}$. Then without loss of generality, we can redefine $\bar{p}(\mathbf{x}, \cdot)$ to equal $\tilde{p}(\mathbf{x}, \cdot)$ for $\mathbf{x} \in \mathbf{C}$. Therefore, $\bar{p}$ and $\tilde{p}$ are mutually absolutely continuous $\tilde{\pi}$ -a.s.

Part 3. See Section 5.5. □

As noted previously, among the features that distinguish this problem from the standard theory for occupation measures is the fact that the underlying transition kernels depend on m . This is a substantial difference, since in the representation at the prelimit it is these kernels that appear in the relative entropy cost, and so a control that is designed on the limit transition kernel may not have even a finite cost for the prelimit kernel. In other words, one cannot in any straightforward way apply a control that is optimal or nearly optimal for the limit to the prelimit, and instead it must be carefully adapted so that the costs at the prelimit approximate those at the limit. Such an adaptation is the concern of the next lemma.

In this lemma, we will relate perturbations on the transition kernel of a single particle to perturbations of the increments in the underlying random walk, and vice versa. In this context, a *controlled increment* will mean a distribution on increments that can depend in a Borel measurable way on the entire past of the random walk.

Lemma 5.7. *Let $\alpha_j^m(\cdot|x)$ be the rescaled single particle transition kernel defined in*

(3.6), and let $x \in [0, 1)$.

1. Given a sequence of controlled increments $\{\nu_{j,i}^m, i = 0, 1, \dots\}$, define the process $\bar{Z}_{j,i}^m$ by $\bar{Z}_{j,0}^m = j + x$ and $\bar{Z}_{j,i+1}^m = \bar{Z}_{j,i}^m + \bar{\theta}_{j,i}^m$, where $\bar{\theta}_{j,i}^m$ has the conditional distribution $\nu_{j,i}^m$. Let $\bar{\sigma}_j^m = \min\{i \geq 0 : \bar{Z}_{j,i}^m \notin (-1, j+1)\}$ and τ_j^m be the distribution of $\bar{Z}_{j,\bar{\sigma}_j^m}^m$. Then

$$\bar{\mathbb{E}} \left[\sum_{i=0}^{\bar{\sigma}_j^m-1} R(\nu_{j,i}^m \parallel \mu) \right] \geq R(\tau_j^m \parallel \alpha_j^m).$$

2. Conversely, given a controlled transition kernel $\tau(\cdot|x)$ that satisfies $R(\tau \parallel \alpha_j^m) < \infty$ and $d\tau(\cdot|x)/d\alpha_j^m(\cdot|x) \geq c$ for some $c > 0$, one can construct controlled increments $\{\nu_{j,i}^m\}$ and a process $\{\bar{Z}_{j,i}^m\}$, such that $\bar{Z}_{j,\bar{\sigma}_j^m}^m$ has distribution τ , and

$$R(\tau \parallel \alpha_j^m) = \bar{\mathbb{E}} \left[\sum_{i=0}^{\bar{\sigma}_j^m-1} R(\nu_{j,i}^m \parallel \mu) \right].$$

Remark 5.8. The conclusion of the previous lemma is also valid for the limit transition kernel $\gamma(\cdot|x, t)$. For instance, for each $t \in [0, 1]$ and given a controlled transition kernel τ , there exists a controlled process $\bar{Z}_{i+1}(t) = \bar{Z}_i(t) + \bar{\theta}_i(t)$ where the $\bar{\theta}_i(t)$ are iid with distribution $\nu_i(\cdot|t)$, and a stopping time $\bar{\sigma}(t) = \min\{i \geq 0 : \bar{Z}_i(t) \geq 1\}$, such that $\bar{Z}_{\bar{\sigma}(t)}(t)$ has distribution τ and

$$R(\tau \parallel \gamma(\cdot|x, t)) = \bar{\mathbb{E}} \left[\sum_{i=0}^{\bar{\sigma}(t)-1} R(\nu_i(\cdot|t) \parallel \mu(\cdot|t)) \right].$$

The next result will also be used in the construction of a nearly optimal control.

Proposition 5.9. The local rate function $G_N(t, \phi)$ defined in (4.22), i.e.

$$G_N(t, \phi) = \inf_{(\xi, \eta) \in A(\phi)} \left\{ \int R(\xi(\cdot|\mathbf{x}) \parallel \Lambda(\cdot|\mathbf{x}, t)) \xi_1(d\mathbf{x}) + \int R(\eta(\cdot|n) \parallel \Theta(\cdot|n)) \eta_1(dn) \right\},$$

is convex in the second variable.

5.4 Proof of the Upper Bound (2.3)

We now proceed to the proof of the large deviation lower bound (2.3):

$$\begin{aligned} \limsup_{m \rightarrow \infty} \hat{W}_N^m &\leq \inf_{\ell \in \mathcal{P}(\{1, \dots, N\})} \{I_N(\ell) + \langle F^N, \ell \rangle\} \\ &= \inf_{\ell \in \mathcal{P}(\{1, \dots, N\})} \inf_{\phi: \ell = \int_{[0,1]} \phi(t) dt} \left\{ \int_0^1 G_N(t, \phi(t)) dt + \langle F^N, \ell \rangle \right\}. \end{aligned}$$

Proof of (2.3). The infimum in (2.3) is over all measurable ϕ . Since such ϕ are not regular enough for the construction of a nearly optimal control, we first show how one can, without loss, assume greater regularity. To simplify notation, we omit the N dependence from I_N and G_N .

Step 1: Existence of a near infimizer, ℓ^ and $\phi^*(t)$, to the right hand side of (2.3).*

Given $\varepsilon > 0$, choose $\ell^* \in \mathcal{P}(\{1, \dots, N\})$ such that

$$I(\ell^*) + \langle F^N, \ell^* \rangle \leq \inf_{\ell \in \mathcal{P}(\{1, \dots, N\})} \{I(\ell) + \langle F^N, \ell \rangle\} + \varepsilon < \infty.$$

For such ℓ^* and ε there is $\phi^*(t)$, $t \in [0, 1]$ with $\ell^* = \int_0^1 \phi^*(s) ds$ satisfying

$$\int_0^1 G(t, \phi^*(t)) dt \leq I(\ell^*) + \varepsilon.$$

Step 2: Discretization of $\phi^(t)$ in time, $\phi_\kappa(t)$, without increasing the cumulative rate.*

Now let $L(t) = \int_0^t \phi^*(s) ds$, so that $L(1) = \ell^*$. For $\kappa \in \mathbb{N}$ and $t \in [0, 1]$ define a piecewise linear approximation to $L(t)$ via

$$\phi_\kappa(t) = \kappa \int_{i/\kappa}^{(i+1)/\kappa} \phi^*(s) ds \text{ if } t \in [i/\kappa, (i+1)/\kappa], i \in \{0, 1, \dots, \kappa-1\}$$

and

$$L_\kappa(t) = \int_0^t \phi_\kappa(s) ds.$$

Note that $L_\kappa(1) = \int_0^1 \phi_\kappa(s) ds = \int_0^1 \phi^*(s) ds = \ell^*$. The discrete form of $\phi_\kappa(t)$ will be used, and we denote $\phi_\kappa(i/\kappa) = \phi_i^\kappa$. The convexity of G in the second variable proved in Proposition 5.9 yields

$$\begin{aligned} \int_0^1 G(t, \dot{L}_\kappa(t)) dt &= \frac{1}{\kappa} \sum_{i=0}^{\kappa-1} G\left(t, \kappa \int_{i/\kappa}^{(i+1)/\kappa} \phi^*(s) ds\right) dt \\ &\leq \sum_{i=0}^{\kappa-1} \int_{i/\kappa}^{(i+1)/\kappa} G(t, \phi^*(t)) dt \\ &= \int_0^1 G(t, \phi^*(t)) dt. \end{aligned}$$

For the given $\varepsilon > 0$ and $\kappa \in \mathbb{N}$, it follows that

$$\int_0^1 G(t, \dot{L}_\kappa(t)) dt + \langle F^N, L_\kappa(1) \rangle \leq \inf_{\ell \in \mathcal{P}(\{1, \dots, N\})} \{I(\ell) + \langle F^N, \ell \rangle\} + 2\varepsilon. \quad (5.13)$$

Step 3: Discrete sum approximation to the cumulative rate $\int_0^1 G(t, \dot{L}_\kappa(t)) dt$. Construction of a sequence of near infimum ergodic admissible controls $(\bar{\xi}_i^\kappa, \bar{\eta}_{i+1}^\kappa)$.

For convenience in the presentation, we assume that the infimum in the definition of G is achieved. If not true, then we could alternatively work with a quantity that is within ε of the infimum as in step 1. Denote the infimizer in (4.22) at $(i/\kappa, \phi_i^\kappa)$ by $(\xi_i^\kappa, \eta_{i+1}^\kappa)$. Under Condition 1.4 the measures $\Lambda(\cdot | \mathbf{x}, t)$ for different values of t are mutually absolutely continuous with uniformly bounded and continuous Radon-Nikodym derivatives. Thus for all $t \in [i/\kappa, (i+1)/\kappa)$

$$\begin{aligned} &\left| G(t, \dot{L}_\kappa(t)) - G(i/\kappa, \phi_i^\kappa) \right| \\ &\leq \left| \int [R(\xi_i^\kappa(\cdot | \mathbf{x}) \|\Lambda(\cdot | \mathbf{x}, t)) - R(\xi_i^\kappa(\cdot | \mathbf{x}) \|\Lambda(\cdot | \mathbf{x}, i/\kappa))] \xi_{i,1}^\kappa(d\mathbf{x}) \right| \\ &< \varepsilon_\kappa, \end{aligned}$$

where $\varepsilon_\kappa \rightarrow 0$ as $\kappa \rightarrow \infty$, and so we can approximate $\int_0^1 G(t, \dot{L}_\kappa(t)) dt$ by the sum $\frac{1}{\kappa} \sum_{i=0}^{\kappa-1} G(i/\kappa, \phi_i^\kappa)$. Choose κ large enough that $\varepsilon_\kappa < \varepsilon$.

Furthermore, given any $\delta > 0$, Lemma 5.6 implies that for each $(\xi_i^\kappa, \eta_{i+1}^\kappa) \in A(\phi_i^\kappa)$ there exist $(\bar{\xi}_i^\kappa, \bar{\eta}_{i+1}^\kappa) \in A(\bar{\phi}_i^\kappa)$ with $\|\bar{\phi}_i^\kappa - \phi_i^\kappa\|_v \leq \delta$, and such that the associated Markov process is geometrically ergodic with bounded local transportation cost and

$$\begin{aligned} & \int R(\bar{\xi}_i^\kappa(\cdot | \mathbf{x}) \|\Lambda(\cdot | \mathbf{x}, i/\kappa)) \bar{\xi}_{i,1}^\kappa(d\mathbf{x}) + \int R(\bar{\eta}_{i+1}^\kappa(\cdot | n) \|\Theta(\cdot | n)) \bar{\phi}_i^\kappa(dn) \\ & \leq G(i/\kappa, \phi_i^\kappa) + \delta. \end{aligned}$$

If $\bar{\ell} \doteq \frac{1}{\kappa} \sum_{i=0}^{\kappa-1} \bar{\phi}_i^\kappa$, then $\|\ell^* - \bar{\ell}\|_v \leq \delta$. Note that by construction, $\bar{\ell}(0) = 0$. Thus, we can choose $\delta \in (0, \varepsilon)$ small enough and use the boundedness of F^N to obtain

$$\langle F^N, \bar{\ell} \rangle \leq \langle F^N, \ell^* \rangle + \varepsilon.$$

Combining the above inequalities with (5.13), we have

$$\frac{1}{\kappa} \sum_{i=0}^{\kappa-1} G(i/\kappa, \bar{\phi}_i^\kappa) + \langle F^N, \bar{\ell} \rangle \leq \inf_{\ell \in \mathcal{P}(\{1, \dots, N\})} \{I(\ell) + \langle F^N, \ell \rangle\} + 5\varepsilon. \quad (5.14)$$

Step 4: Reconstruction of controlled random walks from $(\bar{\xi}_i^\kappa, \bar{\eta}_{i+1}^\kappa)$.

For each $i \in \{0, 1, \dots, \kappa-1\}$, $\bar{\xi}_i^\kappa(\cdot | \mathbf{x})$ is a probability measure on the location of N particles after transportation. Writing this as a product of conditional distributions, we have the decomposition

$$\bar{\xi}_i^\kappa(dy_1, \dots, dy_N | \mathbf{x}) = \prod_{n=1}^N \tau_{i,n}^\kappa(dy_n | \mathbf{x}, y_1, \dots, y_{n-1}).$$

For notational convenience we write $\tau_{i,n}^\kappa(dy_n | \mathbf{x}, y_1, \dots, y_{n-1}) = \tau_{i,n}^\kappa(dy_n | \mathbf{x})$, since the dependence on the previously generated particles is not particularly important.

By construction of the ergodic control in (5.9),

$$\bar{\xi}_i^\kappa = (1 - a) \xi_i^\kappa + a \tilde{\pi} \otimes \tilde{\Lambda},$$

where a is chosen according to the parameter δ described in Step 3. Following a similar discussion to that in Part 2 of Lemma 5.6, we obtain the inequality

$$\bar{\xi}_i^\kappa (B | \mathbf{x}) \geq \frac{a}{f_i^\kappa(\mathbf{x})} \tilde{\Lambda}(B | \mathbf{x}, i/\kappa),$$

where $f_i^\kappa(\mathbf{x}) = (d\bar{\xi}_{i,1}^\kappa/d\tilde{\pi})(\mathbf{x})$, B is any Borel subset of $\mathcal{T}_2$, and $\tilde{\pi}$ is the invariant distribution associated with $\tilde{\Lambda}(\cdot | \cdot, i/\kappa)$ and η_{i+1}^κ . Let $\gamma_i^\kappa(\cdot | \cdot) = \gamma(\cdot | \cdot, i/\kappa)$. By the definition of $\tilde{\Lambda}$, we then have $d\tau_{i,n}^\kappa/d\gamma_i^\kappa > 0$ for all i, n . Hence, Remark 5.8 applies. That is, for each single particle transition kernel $\tau_{i,n}^\kappa$, one can construct a sequence of conditional distributions $\{\nu_{i,n,k}^\kappa\} \subset \mathcal{P}(\mathbb{R})$ associated with random walks $\bar{Z}_{i,n,k+1}^\kappa = \bar{Z}_{i,n,k}^\kappa + \bar{\theta}_{i,n,k}^\kappa$, $\bar{Z}_{i,n,0}^\kappa = \bar{Z}_{i-1,n,\bar{\sigma}_{i-1,n}^\kappa}^\kappa$ and stopping times $\bar{\sigma}_{i,n}^\kappa = \min\{k \geq 0 : \bar{Z}_{i,n,k}^\kappa \geq 1\}$ satisfying $\bar{Z}_{i,n,\bar{\sigma}_{i,n}^\kappa}^\kappa \sim \tau_{i,n}^\kappa(\cdot)$, such that the relative entropy cost is preserved, i.e.,

$$R(\tau_{i,n}^\kappa(\cdot | \mathbf{x}) \parallel \gamma(\cdot | x_n, i/\kappa)) = \mathbb{E} \left[\sum_{k=0}^{\bar{\sigma}_{i,n}^\kappa - 1} R(\nu_{i,n,k}^\kappa(\cdot) \parallel \mu(\cdot | i/\kappa)) \right]. \quad (5.15)$$

The random walks are constructed sequentially with the increment distributions

$$\bar{\mathbb{P}}\{\bar{\theta}_{i,n,k}^\kappa \in \cdot | \mathbf{x}, \bar{Z}_{i,n,0}^\kappa, \dots, \bar{Z}_{i,n,k}^\kappa\} = \nu_{i,n,k}^\kappa(\cdot | \mathbf{x}, \bar{Z}_{i,n,0}^\kappa, \dots, \bar{Z}_{i,n,k}^\kappa).$$

Again for simplicity we write $\nu_{i,n,k}^\kappa(\cdot | \mathbf{x}, \bar{Z}_{i,n,0}^\kappa, \dots, \bar{Z}_{i,n,k}^\kappa) = \nu_{i,n,k}^\kappa(\cdot | \mathbf{x})$.

Step 5: Introduction of large deviation parameter and construction of the prelimit controls $(\hat{\xi}_j^m, \hat{\eta}_j^m)$ for the final proof of (2.3).

Now we introduce a large deviation parameter $m \in \mathbb{N}$. We first elaborate on the construction of transportation controls which are later denoted as $\hat{\xi}_j^m$. For $n \in \{1, \dots, N\}$, $i \in \{0, 1, \dots, \kappa - 1\}$ and $j/m \in [i/\kappa, (i+1)/\kappa)$ define the controlled

random walk

$$\hat{Z}_{j,n,k+1}^m = \hat{Z}_{j,n,k}^m + \hat{\theta}_{j,n,k}^m, \quad \hat{Z}_{j,n,0}^m = \hat{Z}_{j-1,n,\hat{\sigma}_{j-1,n}^m}^m$$

with increment distributions

$$\hat{\mathbb{P}} \left\{ \hat{\theta}_{j,n,k}^m \in \cdot \mid \mathbf{x}, \hat{Z}_{j,n,0}^m, \dots, \hat{Z}_{j,n,k}^m \right\} = \begin{cases} \nu_{i,n,k}^\kappa(\cdot \mid \mathbf{x}) & k \leq \hat{\sigma}_{j,n}^m(b) \\ \mu(\cdot \mid \hat{Z}_{j,n,k}^m/m + j/m) & \text{otherwise} \end{cases},$$

where $\hat{\sigma}_{j,n}^m(b) = \min\{k \geq 0 : \hat{Z}_{j,n,k}^m \leq j(b-1) \text{ or } \hat{Z}_{j,n,k}^m \geq 1\}$ with $b \in (0, 1)$. This means that we give up trying to control the process if it is sufficiently negative, and so the control used differs slightly from the control based completely on $\bar{\xi}_i^\kappa$, but as we will see momentarily this is needed to control the cost. We let $\hat{\xi}_j^m$ denote the corresponding transportation kernel over N particles. Also, the corresponding resampling controls are defined as $\hat{\eta}_j^m(\cdot \mid n) = \eta_{i+1}^\kappa(\cdot \mid n)$ for $j/m \in [i/\kappa, (i+1)/\kappa)$.

As noted in (3.12), given $b < 1$ there exists a $c_b < 1$ such that for $b^2 t \leq s \leq t/b^2$

$$c_b \mu(dy \mid t) \leq \mu(dy \mid s) \leq \frac{1}{c_b} \mu(dy \mid t).$$

Thus, for m big enough such that $(jb/m, (j+1)/m) \subset (ib^2/\kappa, i/(b^2\kappa))$ and $k \leq \hat{\sigma}_{j,n}^m(b)$, we have the bound

$$R \left(\nu_{i,n,k}^\kappa(\cdot \mid \mathbf{x}) \parallel \mu \left(\cdot \mid \hat{Z}_{j,n,k}^m/m + j/m \right) \right) \leq R \left(\nu_{i,n,k}^\kappa(\cdot \mid \mathbf{x}) \parallel \mu(\cdot \mid i/\kappa) \right) - \log c_b. \quad (5.16)$$

Hence one can bound the prelimit transportation cost according to

$$\begin{aligned}
\bar{\mathbb{E}} R\left(\hat{\xi}_j^m(\cdot|\mathbf{x}) \parallel \Lambda_j^m(\cdot|\mathbf{x})\right) &\leq \bar{\mathbb{E}} \sum_{n=1}^N \sum_{k=0}^{\hat{\sigma}_{j,n}^m(b)-1} R\left(\nu_{i,n,k}^\kappa(\cdot|\mathbf{x}) \parallel \mu\left(\cdot \mid \hat{Z}_{j,n,k}^m/m + j/m\right)\right) \\
&\leq \bar{\mathbb{E}} \sum_{n=1}^N \sum_{k=0}^{\hat{\sigma}_{j,n}^m(b)-1} R\left(\nu_{i,n,k}^\kappa(\cdot|\mathbf{x}) \parallel \mu(\cdot|i/\kappa)\right) \\
&\quad - N \bar{\mathbb{E}} \hat{\sigma}_{j,n}^m(b) \log c_b \\
&\leq \bar{\mathbb{E}} \sum_{n=1}^N \sum_{k=0}^{\hat{\sigma}_{j,n}^m(0)-1} R\left(\nu_{i,n,k}^\kappa(\cdot|\mathbf{x}) \parallel \mu(\cdot|i/\kappa)\right) + \varepsilon_b^m \\
&= \bar{\mathbb{E}} R\left(\bar{\xi}_i^\kappa(\cdot|\mathbf{x}) \parallel \Lambda(\cdot|\mathbf{x}, i/\kappa)\right) + \varepsilon_b^m, \tag{5.17}
\end{aligned}$$

where $\varepsilon_b^m \rightarrow 0$ as $m \rightarrow \infty$. The first inequality is an application of Chain Rule (Theorem 0.8) and Lemma 5.7(a). Note that the controlled increments after time $\hat{\sigma}_{j,n}^m(b)$ will have no contribute to the relative entropy cost, since, by construction, the control follows the law of underlying random walk. The second inequality is obtained from (5.16). The third inequality is true because $\sup \bar{\mathbb{E}} \hat{\sigma}_{j,n}^m(b) < \infty$ is guaranteed by the finiteness of relative entropy. The last equality is due to (5.15) and the fact that $\hat{\sigma}_{j,n}^m(0)$ corresponds to $\bar{\sigma}_{i,n}^\kappa$.

It is useful to review where we are in the construction at this stage. Starting with a control that is nearly optimal for the limit, we have constructed a control that can be applied to the increments of individual particles at the prelimit level, and which leads to nearly the same cost. However, to do so we have had to make a small change in the distribution of the processes at the threshold crossings. Specifically, in order to control the costs it was necessary to suspend active control of particles once that had moved far enough to the left, switch back to the underlying distribution, and rely on the law of large numbers for the underlying distribution to finish the construction. This lead to a small change in the distribution of the transition kernel whose effect will vanish as $m \rightarrow \infty$. However, it means that we cannot directly apply the ergodic theorem as would be traditionally done in the large deviations analysis

at this point in the proof, since we no longer have a stationary ergodic process, but only one whose nonstationary transition kernels are in some sense uniformly close to those of a geometrically ergodic Markov process. To prove the convergence of the controlled distributions on the number of success particles, we will need the following result, whose proof is presented at the end of this chapter.

Proposition 5.10. *Consider $j \in [i/\kappa, (i+1)/\kappa)$. Let $\hat{p}_j^m(\mathbf{x}, d\mathbf{y})$ and $\bar{p}_i^\kappa(\mathbf{x}, d\mathbf{y})$ be the transition kernels associated with $(\hat{\xi}_j^m, \bar{\eta}_{i+1}^\kappa)$ and $(\bar{\xi}_i^\kappa, \bar{\eta}_{i+1}^\kappa)$ respectively. Denote the empirical measure of $\{\hat{\mathbf{X}}_j^m, i/\kappa \leq j/m \leq (i+1)/\kappa\}$ over this interval by $\hat{J}_i^m$ and the invariant distribution of $\bar{p}_i^\kappa$ by $\bar{\lambda}_i^\kappa$. Then*

$$\left\| \hat{J}_i^m - \bar{\lambda}_i^\kappa \right\|_v \rightarrow 0$$

in probability as $m \rightarrow \infty$.

Step 6: Completion of the proof of (2.3).

With Proposition 5.10 in hand, we are now ready to complete the proof of (2.3). The starting point is the representation (3.8). It is easy to check that Steps 1-5 provide an admissible control sequence, and hence gives an upper bound on $\hat{W}_N^m$. Using such controls we define the controlled sequences $(\hat{\mathbf{X}}_j^m, \hat{\mathbf{Y}}_j^m)$. For $i \in \{0, \dots, \kappa - 1\}$, let

$$m_i = \sum_{j=0}^{m-1} \mathbf{1}_{[i/\kappa, (i+1)/\kappa)} \frac{j}{m}$$

be the number of stages within the continuous time interval $[i/\kappa, (i+1)/\kappa)$, and

$$i_j = \sum_{k=0}^{i-1} m_k + j$$

be the index of the j th element at within interval $[i/\kappa, (i+1)/\kappa)$. With these

definitions, we can write

$$\begin{aligned}
& \hat{\mathbb{E}} \left\{ \frac{1}{m} \sum_{j=0}^{m-1} \left[R \left(\hat{\xi}_j^m \left(\cdot \mid \hat{\mathbf{X}}_j^m \right) \parallel \Lambda_j^m \left(\cdot \mid \hat{\mathbf{X}}_j^m \right) \right) + R \left(\hat{\eta}_{j+1}^m \left(\cdot \mid \hat{N}_{j+1}^m \right) \parallel \Theta \left(\cdot \mid \hat{N}_{j+1}^m \right) \right) \right] \right\} \\
&= \sum_{i=0}^{\kappa-1} \frac{m_i}{m} \hat{\mathbb{E}} \left\{ \frac{1}{m_i} \sum_{j=0}^{m_i-1} \left[R \left(\hat{\xi}_{i_j}^m \left(\cdot \mid \hat{\mathbf{X}}_{i_j}^m \right) \parallel \Lambda_{i_j}^m \left(\cdot \mid \hat{\mathbf{X}}_{i_j}^m \right) \right) \right. \right. \\
&\quad \left. \left. + R \left(\hat{\eta}_{i_{j+1}}^\kappa \left(\cdot \mid \hat{N}_{i_{j+1}}^m \right) \parallel \Theta \left(\cdot \mid \hat{N}_{i_{j+1}}^m \right) \right) \right] \right\} \\
&\leq \sum_{i=0}^{\kappa-1} \frac{m_i}{m} \hat{\mathbb{E}} \left\{ \frac{1}{m_i} \sum_{j=0}^{m_i-1} \left[R \left(\bar{\xi}_i^\kappa \left(\cdot \mid \hat{\mathbf{X}}_{i_j}^m \right) \parallel \Lambda \left(\cdot \mid \hat{\mathbf{X}}_{i_j}^m, i/\kappa \right) \right) \right. \right. \\
&\quad \left. \left. + R \left(\bar{\eta}_{i+1}^\kappa \left(\cdot \mid \hat{N}_{i_{j+1}}^m \right) \parallel \Theta \left(\cdot \mid \hat{N}_{i_{j+1}}^m \right) \right) \right] \right\} + \varepsilon_b^m.
\end{aligned}$$

The equality is just due to definitions, while the inequality follows from (5.17).

Since κ is fixed, $m_i/m \rightarrow 1/\kappa$ as $m \rightarrow \infty$, and it will suffice to obtain upper bounds on each term of the form

$$\begin{aligned}
& \limsup_{m_i \rightarrow \infty} \hat{\mathbb{E}} \left\{ \frac{1}{m_i} \sum_{j=0}^{m_i-1} \left[R \left(\bar{\xi}_i^\kappa \left(\cdot \mid \hat{\mathbf{X}}_{i_j}^m \right) \parallel \Lambda \left(\cdot \mid \hat{\mathbf{X}}_{i_j}^m, i/\kappa \right) \right) \right. \right. \\
&\quad \left. \left. + R \left(\bar{\eta}_{i+1}^\kappa \left(\cdot \mid \hat{N}_{i_{j+1}}^m \right) \parallel \Theta \left(\cdot \mid \hat{N}_{i_{j+1}}^m \right) \right) \right] \right\}. \tag{5.18}
\end{aligned}$$

For each i , the chain $\hat{\mathbf{X}}_{i_j}^m$ is obtained from the transition kernel $\hat{\xi}_{i_j}^m$, which is uniformly close to $\bar{\xi}_i^\kappa$. It follows from part 3 of Lemma 5.6 and Proposition 5.10 that the limit superior of the first term is bounded above by $\int R(\bar{\xi}_i^\kappa(\cdot \mid \mathbf{x}) \parallel \Lambda(\cdot \mid \mathbf{x}, i/\kappa)) \bar{\xi}_{i,1}^\kappa(d\mathbf{x}) + \varepsilon$. A similar but simpler argument can be applied to the second term to obtain the combined upper bound

$$\int R(\bar{\xi}_i^\kappa(\cdot \mid \mathbf{x}) \parallel \Lambda(\cdot \mid \mathbf{x}, i/\kappa)) \bar{\xi}_{i,1}^\kappa(d\mathbf{x}) + \int R(\bar{\eta}_{i+1}^\kappa(\cdot \mid n) \parallel \Theta(\cdot \mid n)) \bar{\phi}_i^\kappa(dn) + 2\varepsilon$$

for (5.18), where $\bar{\xi}_{i,1}^\kappa = \bar{\lambda}_i^\kappa$.

To complete the argument, we will have to compute an upper bound for $\hat{\mathbb{E}}\langle F^N, \hat{L}^m \rangle$, where $\hat{L}^m = 1/m \sum_{j=1}^m \delta_{\hat{N}_j^m}$. Since the transportation measures $\hat{\xi}_j^m$ and $\bar{\xi}_i^\kappa$ are close in total variation norm for $j \in [i/\kappa, (i+1)/\kappa)$ by construction, together with Proposition 5.10 and the definitions of $\hat{L}_j^m$ and $\bar{\ell}$ for sufficiently large m

$$\hat{\mathbb{E}}\langle F^N, \hat{L}^m \rangle \leq \langle F^N, \bar{\ell} \rangle + \varepsilon.$$

We now combine the last four displays and (5.14) to obtain

$$\begin{aligned} & \limsup_{m \rightarrow \infty} \hat{W}_N^m \\ & \leq \frac{1}{\kappa} \int R(\bar{\xi}_i^\kappa(\cdot | \mathbf{x}) \|\Lambda(\cdot | \mathbf{x}, i/\kappa)) \bar{\xi}_{i,1}^\kappa(d\mathbf{x}) + \int R(\bar{\eta}_{i+1}^\kappa(\cdot | n) \|\Theta(\cdot | n)) \bar{\phi}_i^\kappa(dn) \\ & \quad + \langle F^N, \bar{\ell} \rangle + 3\varepsilon + \limsup_{b \rightarrow 1} \limsup_{m \rightarrow \infty} \varepsilon_b^m \\ & \leq \frac{1}{\kappa} \sum_{i=0}^{\kappa-1} G(i/\kappa, \bar{\phi}_i^\kappa) + \langle F^N, \bar{\ell} \rangle + 4\varepsilon \\ & \leq \inf_{\ell \in \mathcal{P}(\{1, \dots, N\})} \{I(\ell) + \langle F^N, \ell \rangle\} + 9\varepsilon. \end{aligned}$$

Since ε is arbitrary, we have proved the upper bound (2.3).

5.5 Appendix

We finish this chapter by giving the proofs of Lemma 5.7, Proposition 5.9, Proposition 5.10 and Lemma 5.6(3).

Lemma 5.7. *Let $\alpha_j^m(\cdot | x)$ be the rescaled single particle transition kernel defined in (3.6), and let $x \in [0, 1)$.*

1. *Given a sequence of controlled increments $\{\nu_{j,i}^m, i = 0, 1, \dots\}$, define the process $\bar{Z}_{j,i}^m$ by $\bar{Z}_{j,0}^m = j + x$ and $\bar{Z}_{j,i+1}^m = \bar{Z}_{j,i}^m + \bar{\theta}_{j,i}^m$, where $\bar{\theta}_{j,i}^m$ has the conditional distribution $\nu_{j,i}^m$. Let $\bar{\sigma}_j^m = \min\{i \geq 0 : \bar{Z}_{j,i}^m \notin (-1, j+1)\}$ and τ_j^m be the*

distribution of $\bar{Z}_{j,\bar{\sigma}_j^m}^m$. Then

$$\bar{\mathbb{E}} \left[\sum_{i=0}^{\bar{\sigma}_j^m-1} R(\nu_{j,i}^m \parallel \mu) \right] \geq R(\tau_j^m \parallel \alpha_j^m).$$

2. Conversely, given a controlled transition kernel $\tau(\cdot|x)$ that satisfies $R(\tau \parallel \alpha_j^m) < \infty$ and $d\tau(\cdot|x)/d\alpha_j^m(\cdot|x) \geq c$ for some $c > 0$, one can construct controlled increments $\{\nu_{j,i}^m\}$ and a process $\{\bar{Z}_{j,i}^m\}$, such that $\bar{Z}_{j,\bar{\sigma}_j^m}^m$ has distribution τ , and

$$R(\tau \parallel \alpha_j^m) = \bar{\mathbb{E}} \left[\sum_{i=0}^{\bar{\sigma}_j^m-1} R(\nu_{j,i}^m \parallel \mu) \right].$$

Proof. Part 1. For fixed $T < \infty$, consider a modification of $\nu_{j,i}^m$, such that if $\bar{\sigma}_j^m < T$, then we redefine $\nu_{j,i}^m = \mu$ for $\bar{\sigma}_j^m < i \leq T-1$. With this modified definition of the control, which does not change the distribution of the hitting time on the set $\{\bar{\sigma}_j^m < T\}$, we have

$$\begin{aligned} \bar{\mathbb{E}} \left[\sum_{i=0}^{\bar{\sigma}_j^m-1} R(\nu_{j,i}^m \parallel \mu) \right] &\geq \bar{\mathbb{E}} \left[\sum_{i=0}^{(\bar{\sigma}_j^m \wedge T)-1} R(\nu_{j,i}^m \parallel \mu) \right] \\ &= R\left(\mathbb{Q}^T|_{\{0,\dots,T-1\}} \parallel \mathbb{P}|_{\{0,\dots,T-1\}}\right), \end{aligned} \tag{5.19}$$

where $\mathbb{Q}^T$ denotes the modified joint measure on the space of increments, $\mathbb{Q}^T|_{\{0,\dots,T-1\}}$ denotes the restriction to the first T coordinates in the underlying product space, and $\mathbb{P}|_{\{0,\dots,T-1\}}$ denotes product measure with marginal μ .

We now consider the disjoint, finite partition of the space $\mathbb{R}^T = A_j^T \cup B_j^T \cup D_j^T$,

where

$$\begin{aligned} A_j^T &= \left\{ \omega : \bar{Z}_{j, \sigma_j^m}^m \geq j+1, \bar{\sigma}_j^m < T \right\}, \\ B_j^T &= \left\{ \omega : \bar{Z}_{j, \sigma_j^m}^m \leq -1, \bar{\sigma}_j^m < T \right\}, \\ D_j^T &= \left\{ \omega : \bar{\sigma}_j^m \geq T \right\}. \end{aligned}$$

Using the approximation property of relative entropy via sums over finite measurable partitions Lemma 0.7(g), we obtain

$$R\left(\mathbb{Q}^T|_{\{0, \dots, T-1\}} \parallel \mathbb{P}|_{\{0, \dots, T-1\}}\right) \geq R\left(\tau_j^{m, T} \parallel \alpha_j^{m, T}\right), \quad (5.20)$$

where $\tau_j^{m, T}, \alpha_j^{m, T} \in \mathcal{P}(\{(-\infty, -1] \cup [j+1, \infty), \Sigma\})$ are the measures induced by $\mathbb{Q}^T|_{\{0, \dots, T-1\}}$ and $\mathbb{P}|_{\{0, \dots, T-1\}}$, and Σ corresponds to the event $\sigma_j^m \geq T$.

We know that $\alpha_j^{m, T}(\Sigma) \rightarrow 0$ as $T \rightarrow \infty$. Let $\bar{\alpha}_j^m$ denote the extension of α_j^m to $\{(-\infty, -1] \cup [j+1, \infty), \Sigma\}$ with $\bar{\alpha}_j^m(\Sigma) = 0$. Let $\bar{\tau}_j^m$ denote the limit of $\tau_j^{m, T}$, which must exist by monotonicity. By the lower semi-continuity of relative entropy,

$$\liminf_{T \rightarrow \infty} R\left(\tau_j^{m, T} \parallel \alpha_j^{m, T}\right) \geq R\left(\bar{\tau}_j^m \parallel \bar{\alpha}_j^m\right). \quad (5.21)$$

If $\bar{\tau}_j^m(\Sigma) > 0$ then $R(\bar{\tau}_j^m \parallel \bar{\alpha}_j^m) = \infty$, and if $\bar{\tau}_j^m(\Sigma) = 0$ then $R(\bar{\tau}_j^m \parallel \bar{\alpha}_j^m) = R(\tau_j^m \parallel \alpha_j^m)$. The last sentence and (5.19), (5.20) and (5.20) imply

$$\bar{\mathbb{E}} \left[\sum_{i=0}^{\bar{\sigma}_j^m - 1} R(\nu_{j,i}^m \parallel \mu) \right] \geq R(\tau_j^m \parallel \alpha_j^m).$$

Part 2. Define

$$U(x; y) = -\log \left(\frac{d\tau(\cdot | x)}{d\alpha_j^m(\cdot | x)}(y) \right).$$

Since by assumption $U(x; y)$ is bounded from above, according to [9, Proposition

4.5.1]

$$-\log \int e^{-U(x;y)} \alpha_j^m(dy|x) = \inf_{\beta: R(\beta \parallel \alpha_j^m) < \infty} \left\{ R(\beta \parallel \alpha_j^m) + \int U(x;y) \beta(dy) \right\}.$$

Since $R(\tau \parallel \alpha_j^m) < \infty$, by the definition of U

$$\tau(dy|x) = e^{-U(x;y)} \alpha_j^m(dy|x) \quad (5.22)$$

achieves the infimum in the variational formula, i.e.,

$$-\log \int e^{-U(x;y)} \alpha_j^m(dy|x) = R(\tau \parallel \alpha_j^m) + \int U(x;y) \tau(dy|x).$$

We next consider the analogous relations on the space of increments of the process. We have

$$-\log \int e^{-U\left(x; \bar{Z}_{j, \bar{\sigma}_j^m}^m(\omega)\right)} \mathbb{P}(d\omega) = \inf_{\mathbb{Q}} \left\{ R(\mathbb{Q} \parallel \mathbb{P}) + \int U\left(x; \bar{Z}_{j, \bar{\sigma}_j^m}^m(\omega)\right) \mathbb{Q}(d\omega) \right\},$$

where $\bar{\sigma}_j^m$ denotes the first hitting time of the process and $\mathbb{P}$ is the underlying product probability with marginal μ . The definition of α_j^m implies that for any integrable function K

$$\int K\left(\bar{Z}_{j, \bar{\sigma}_j^m}^m(\omega)\right) e^{-U\left(x; \bar{Z}_{j, \bar{\sigma}_j^m}^m(\omega)\right)} \mathbb{P}(d\omega) = \int K(y) e^{-U(x;y)} \alpha_j^m(dy|x).$$

Also, the minimum is achieved at

$$\mathbb{Q}(d\omega) = e^{-U\left(x; \bar{Z}_{j, \bar{\sigma}_j^m}^m(\omega)\right)} \mathbb{P}(d\omega). \quad (5.23)$$

Since the first hitting probability α_j^m is induced by $\mathbb{P}$, (5.22) and (5.23) imply

$$\begin{aligned} \int U(x; \bar{Z}_{j, \bar{\sigma}_j^m}^m(\omega)) \mathbb{Q}(d\omega) &= \int U(x; \bar{Z}_{j, \bar{\sigma}_j^m}^m(\omega)) e^{-U(x; \bar{Z}_{j, \bar{\sigma}_j^m}^m(\omega))} \mathbb{P}(d\omega) \\ &= \int U(x; y) e^{-U(x; y)} \alpha_j^m(dy|x) \\ &= \int U(x; y) \tau(dy|x). \end{aligned}$$

Hence, by the chain rule

$$R(\tau \parallel \alpha_j^m) = R(\mathbb{Q} \parallel \mathbb{P}) = \mathbb{E}^{\mathbb{Q}} \left[\sum_{i=0}^{\infty} R(\nu_{j,i}^m \parallel \mu) \right] \geq \mathbb{E}^{\mathbb{Q}} \left[\sum_{i=0}^{\bar{\sigma}_j^m-1} R(\nu_{j,i}^m \parallel \mu) \right],$$

where $\{\nu_{j,i}^m\}$ are the controls defined by factoring $\mathbb{Q}$.

Combining this with the bound proved in part 1, we have constructed controls and a controlled process $\{\bar{Z}_{j,i}^m\}$ satisfying

$$\mathbb{E}^{\mathbb{Q}} \left[\sum_{i=0}^{\bar{\sigma}_j^m-1} R(\nu_{j,i}^m \parallel \mu) \right] = R(\tau \parallel \alpha_j^m),$$

where τ is the distribution induced by the stopped, controlled process. Although here we have constructed the control on the canonical space $\mathbb{R}^\infty$, it can be applied in the general setting by making the obvious identifications. $\square$

Proposition 5.9. *The local rate function $G_N(t, \phi)$ defined in (4.22) is convex in the second variable.*

Proof. By (4.21),

$$\begin{aligned} R(\xi \parallel \xi_1 \otimes \Lambda) &= \int \xi(d\mathbf{x}, d\mathbf{y}) \eta(d\boldsymbol{\sigma} | g_{\text{num}}(\mathbf{y})) \log \frac{d\xi(\cdot | \mathbf{x})}{d\Lambda(\cdot | \mathbf{x}, t)}(\mathbf{y}), \\ R(\eta \parallel \eta_1 \otimes \Theta) &= \int \xi(d\mathbf{x}, d\mathbf{y}) \eta(d\boldsymbol{\sigma} | g_{\text{num}}(\mathbf{y})) \log \frac{d\eta(\cdot | g_{\text{num}}(\mathbf{y}))}{d\Theta(\cdot | g_{\text{num}}(\mathbf{y}))}(\boldsymbol{\sigma}). \end{aligned}$$

Summing these two equalities, we have

$$\begin{aligned} G_N(t, \phi) &= \inf_{(\xi, \eta) \in A(\phi)} \int \xi(d\mathbf{x}, d\mathbf{y}) \eta(d\boldsymbol{\sigma} | g_{\text{num}}(\mathbf{y})) \log \frac{d\xi(\cdot | \mathbf{x})}{d\Lambda(\cdot | \mathbf{x}, t)}(\mathbf{y}) \frac{d\eta(\cdot | g_{\text{num}}(\mathbf{y}))}{d\Theta(\cdot | g_{\text{num}}(\mathbf{y}))}(\boldsymbol{\sigma}). \end{aligned}$$

Let

$$\Psi(d\mathbf{y}, d\boldsymbol{\sigma} | \mathbf{x}, t) = \Theta(d\boldsymbol{\sigma} | g_{\text{num}}(\mathbf{y})) \Lambda(d\mathbf{y} | \mathbf{x}, t),$$

and let p be the transition kernel associated with (Λ, Θ) . Denote the corresponding controlled versions by $\bar{\Psi}$, $\bar{p}$. Then the local rate function G can be written as

$$\begin{aligned} G_N(t, \phi) &= \inf \left\{ \inf_{\xi_1 = \xi_1 \bar{p}} \int \xi_1(d\mathbf{x}) R(\bar{\Psi}(\cdot | \mathbf{x}) \| \Psi(\cdot | \mathbf{x}, t)) : \phi(\cdot) = \int_{\{g_{\text{num}}(\mathbf{y}) = \cdot\}} [\xi_1 \otimes \bar{\Psi}]_2(d\mathbf{y}) \right\}. \end{aligned}$$

For $\phi_1, \phi_2 \in \mathcal{P}(\{1, \dots, N\})$, let $(\xi_1^i, \bar{\Psi}^i)$, $i = 1, 2$ be within $\varepsilon > 0$ of the infimum. For $\beta \in (0, 1)$, the convex combination $\beta \xi_1^1 \otimes \bar{\Psi}^1 + (1 - \beta) \xi_1^2 \otimes \bar{\Psi}^1$ can be decomposed into $\xi_1 \otimes \bar{\Psi}$, where $(\xi_1, \bar{\Psi})$ are defined on the same space as $(\xi_1^i, \bar{\Psi}^i)$. Then $\xi_1 = \beta \xi_1^1 + (1 - \beta) \xi_1^2$ and

$$\int \mathbf{1}_{\{g_{\text{num}}(\mathbf{y}) = n\}} [\xi_1 \otimes \bar{\Psi}]_2(d\mathbf{y}) = \beta \phi_1(\{n\}) + (1 - \beta) \phi_2(\{n\}),$$

Furthermore,

$$\begin{aligned} & \int \delta_{\{h(g(\mathbf{y}), \boldsymbol{\sigma})\}}(\cdot) \bar{\Psi}(d\mathbf{y}, d\boldsymbol{\sigma} | \mathbf{x}) \xi_1(d\mathbf{x}) \\ &= \beta \int \delta_{\{h(g(\mathbf{y}), \boldsymbol{\sigma})\}}(\cdot) \bar{\Psi}^1(d\mathbf{y}, d\boldsymbol{\sigma} | \mathbf{x}) \xi_1^1(d\mathbf{x}) \\ & \quad + (1 - \beta) \int \delta_{\{h(g(\mathbf{y}), \boldsymbol{\sigma})\}}(\cdot) \bar{\Psi}^2(d\mathbf{y}, d\boldsymbol{\sigma} | \mathbf{x}) \xi_1^2(d\mathbf{x}) \\ &= \beta \xi_1^1(d\mathbf{x}) + (1 - \beta) \xi_1^2(d\mathbf{x}) \\ &= \xi_1(d\mathbf{x}). \end{aligned}$$

The convexity of relative entropy then implies

$$\begin{aligned}
& \beta G_N(t, \phi_1) + (1 - \beta) G_N(t, \phi_2) + 2\varepsilon \\
& \geq \beta R(\xi_1^1 \otimes \bar{\Psi}^1 \parallel \xi_1^1 \otimes \Psi) + (1 - \beta) R(\xi_1^2 \otimes \bar{\Psi}^1 \parallel \xi_1^2 \otimes \Psi) \\
& \geq R(\beta \xi_1^1 \otimes \bar{\Psi}^1 + (1 - \beta) \xi_1^2 \otimes \bar{\Psi}^1 \parallel (\beta \xi_1^1 + (1 - \beta) \xi_1^2) \otimes \Psi) \\
& = R(\xi_1 \otimes \bar{\Psi} \parallel \xi_1 \otimes \Psi) \\
& \geq G_N(t, \beta \phi_1 + (1 - \beta) \phi_2).
\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, convexity follows. $\square$

Proposition 5.10. *Consider $j \in [i/\kappa, (i+1)/\kappa)$. Let $\hat{p}_j^m(\mathbf{x}, d\mathbf{y})$ and $\bar{p}_i^\kappa(\mathbf{x}, d\mathbf{y})$ be the transition kernels associated with $(\hat{\xi}_j^m, \bar{\eta}_{i+1}^\kappa)$ and $(\bar{\xi}_i^\kappa, \bar{\eta}_{i+1}^\kappa)$ respectively. Denote the empirical measure of $\{\hat{\mathbf{X}}_j^m, i/\kappa \leq j/m \leq (i+1)/\kappa\}$ over this interval by $\hat{J}_i^m$ and the invariant distribution of $\bar{p}_i^\kappa$ by $\bar{\lambda}_i^\kappa$. Then*

$$\left\| \hat{J}_i^m - \bar{\lambda}_i^\kappa \right\|_v \rightarrow 0$$

in probability as $m \rightarrow \infty$.

Proof. Since $\{\hat{\mathbf{X}}_j^m\}$ is tight, given $\eta > 0$, there is compact $S_1 \subset \mathcal{T}_1$ such that $\mathbb{P}\{\hat{\mathbf{X}}_j^m \notin S_1\} \leq \eta$ for all $j \leq m$, including the index that first exceeds i/κ . Using standard results from stochastic stability, there is a compact set $S_2 \subset \mathcal{T}_1$ such that, after starting in S_1 , the probability of escaping from S_2 is also less than η .

To simplify the presentation, we denote the initial value of $j \in [i/\kappa, (i+1)/\kappa)$ by zero and the final value by m , and write $\hat{J}^m$ for $\hat{J}_i^m$. By the geometric ergodicity of $\bar{p}_i^\kappa$ and the compactness of S_2

$$\left\| \bar{p}_i^{\kappa, (k)}(\mathbf{x}, \cdot) - \bar{\lambda}_i^\kappa(\cdot) \right\|_v \rightarrow 0$$

uniformly in $\mathbf{x} \in S_2$, where $\bar{p}_i^{\kappa, (k)}$ is the k -fold convolution of $\bar{p}_i^\kappa$. Let $\delta > 0$, and assume that k is large enough that for such $\mathbf{x}$

$$\left\| \bar{p}_i^{\kappa, (k)}(\mathbf{x}, \cdot) - \bar{\lambda}_i^\kappa(\cdot) \right\|_v < \delta. \quad (5.24)$$

Let $\hat{p}_{j,\ell}^m$ denote convolution of the transition kernels $\hat{p}_{jk+\ell}^m, \hat{p}_{jk+\ell+1}^m, \dots, \hat{p}_{(j+1)k+\ell-1}^m$, so that if $\hat{\mathbf{X}}_{kj+\ell}^m = \mathbf{x}$ then the distribution of $\hat{\mathbf{X}}_{k(j+1)+\ell}^m$ is given by $\hat{p}_{j,\ell}^m(\mathbf{x}, \cdot)$. Let $\zeta > 0$ be given. For m sufficiently large we know that

$$\left\| \hat{p}_j^m(\mathbf{x}, \cdot) - \bar{p}_i^\kappa(\mathbf{x}, \cdot) \right\|_v < \zeta. \quad (5.25)$$

Iterating the (5.25) gives

$$\left\| \hat{p}_{j,\ell}^m(\mathbf{x}, \cdot) - \bar{p}_i^{\kappa, (k)}(\mathbf{x}, \cdot) \right\|_v \leq k\zeta$$

for all ℓ and all $\mathbf{x} \in S_2$, and thus

$$\left\| \hat{p}_{j,\ell}^m(\mathbf{x}, \cdot) - \bar{\lambda}_i^\kappa(\cdot) \right\|_v < k\zeta + \delta \quad (5.26)$$

for all ℓ and all $\mathbf{x} \in S_2$.

Let $k \in \mathbb{N}$ be given. We will break the empirical measure $\hat{J}^m$ into k sums, and it follows from the fact that k will be fixed as $m \rightarrow \infty$ that we can consider just those m of the form $rk, r \in \mathbb{N}$. We then write

$$\hat{J}^m = \frac{1}{m} \sum_{j=0}^{m-1} \delta_{\hat{\mathbf{X}}_j^m} = \frac{1}{m} \left[\sum_{\ell=0}^{k-1} \sum_{j=0}^{[m/k]-1} \delta_{\hat{\mathbf{X}}_{kj+\ell}^m} \right] = \frac{1}{k} \sum_{\ell=0}^{k-1} \frac{k}{m} \sum_{j=0}^{[m/k]-1} \delta_{\hat{\mathbf{X}}_{kj+\ell}^m} \doteq \frac{1}{k} \sum_{\ell=0}^{k-1} \hat{J}_\ell^m,$$

where the last equality defines the $\hat{J}_\ell^m$.

Now fix $\ell \in \{0, 1, \dots, k-1\}$ and consider any bound measurable function $f :$

$\mathcal{T}_1 \rightarrow \mathbb{R}$ with $\|f\|_\infty \leq 1$. As usual, we have that

$$f\left(\hat{\mathbf{X}}_{k(j+1)+\ell}^m\right) - \int f(\mathbf{y}) \hat{p}_{j,\ell}^m\left(\hat{\mathbf{X}}_{jk+\ell}^m, d\mathbf{y}\right)$$

is a martingale difference sequence, and thus by (5.26)

$$\left[f\left(\hat{\mathbf{X}}_{k(j+1)+\ell}^m\right) - \int f(\mathbf{y}) \bar{\lambda}_i^\kappa(d\mathbf{y}) \right] 1_{\{\hat{\mathbf{X}}_{jk+\ell}^m \in S_2\}} = \varepsilon_{1,j}^m + \varepsilon_{2,j}^m,$$

where $|\varepsilon_{1,j}^m| \leq k\zeta + \delta$ and $\varepsilon_{2,j}^m$ has conditional mean zero and uniformly (in m and j) bounded second moment. It follows using a standard calculation that

$$\mathbb{P} \left\{ \bigcup_{\ell=0}^{k-1} \left\{ \left| \int f(\mathbf{x}) \hat{J}_\ell^m(d\mathbf{x}) - \int f(\mathbf{y}) \bar{\lambda}_i^\kappa(d\mathbf{y}) \right| \geq 2[k\zeta + \delta] \right\} \right\} \leq k \frac{k}{4[k\zeta + \delta]^2 m} + 2\eta.$$

Letting first $m \rightarrow \infty$ shows that weak limits of the $\hat{J}_\ell^m$ are all within $2[k\zeta + \delta]$ of $\bar{\lambda}_i^\kappa$ in total variation norm save on a set of probability no more than 2η . We now send $\zeta \downarrow 0$, then $\delta \downarrow 0$, and finally $\eta \downarrow 0$ to complete the proof. $\square$

We are now ready to proof Part 3 of Lemma 5.6.

Proof. Let $h(\mathbf{x}, \mathbf{y}) = (d\bar{\Lambda}/d\Lambda)(\mathbf{x}, \mathbf{y})$, where the Radon-Nikodym derivative is in $\mathbf{y}$. For $k \in \mathbb{N}$ define

$$S^k(\mathbf{x}) = 1 - \int_{\mathcal{T}_2} h^k(\mathbf{x}, \mathbf{y}) \Lambda(d\mathbf{y} | \mathbf{x}),$$

where $h^k(\mathbf{x}, \mathbf{y}) = k \wedge h(\mathbf{x}, \mathbf{y})$. Also, let

$$\bar{\Lambda}^k(A | \mathbf{x}) = \int_A h^k(\mathbf{x}, \mathbf{y}) \Lambda(d\mathbf{y} | \mathbf{x}) + S^k(\mathbf{x}) \tilde{\Lambda}(A | \mathbf{x})$$

for any Borel set A of $\mathcal{T}_2$. Obviously, $(\bar{\xi}^k, \bar{\eta})$ is also admissible and the associated transition kernel is equivalent to $\bar{p}(\mathbf{x}, \cdot)$. Hence this transition kernel is also geomet-

rically ergodic. Furthermore, the relative entropy

$$R\left(\bar{\Lambda}^k(\cdot|\mathbf{x})\|\Lambda(\cdot|\mathbf{x})\right) = \int_{\mathcal{T}_2} \log h^k(\mathbf{x}, \mathbf{y}) \bar{\Lambda}^k(d\mathbf{y}|\mathbf{x}) + S^k(\mathbf{x}) R\left(\tilde{\Lambda}(\cdot|\mathbf{x})\|\Lambda(\cdot|\mathbf{x})\right)$$

is uniformly bounded as a function of $\mathbf{x}$. Since $S^k(\mathbf{x}) \rightarrow 0$ as $k \rightarrow \infty$, the Dominated Convergence Theorem implies (5.8).

It is easy to check that the construction implies $\|\bar{\Lambda}^k(\cdot|\mathbf{x}) - \bar{\Lambda}(\cdot|\mathbf{x})\|_v \rightarrow 0$. It then follows from Proposition 5.10 that the first marginals of $\bar{\xi}^k$ converges to $\bar{\xi}_1$ in total variation, i.e. $\|\bar{\xi}_1^k - \bar{\xi}_1\|_v \rightarrow 0$. $\square$

Chapter 6

Asymptotics of the Decay Rate as $N \rightarrow \infty$

In this chapter, we assume for simplicity that the increments $\{\theta_i, i \in \mathbb{N}\}$ are iid with $\theta_i \sim \mu(dy)$ and evaluate the asymptotics of I_N as $N \rightarrow \infty$. The analogous result holds for the general model used previously, but with more involved constructions and notation. Under this assumption, it follows that the rate function of interest coincides with the local rate function. That is, for $\mu \in \mathcal{P}(\{1, \dots, N\})$, the rate function can be written as

$$I_N(\mu) = \inf_{\{\bar{\xi}, \bar{\eta}\} \in A(\mu)} [R(\bar{\xi} \parallel \bar{\xi}_1 \otimes \Lambda) + R(\bar{\eta} \parallel \bar{\eta}_1 \otimes \Theta)]$$

where set $A(\mu) = \{\text{feasible } (\bar{\xi}, \bar{\eta}) \text{ with } \bar{\eta}_1 = \mu\}$. As established in Theorem 2.2, the decay rate for the second moment has the expression

$$V^N = \inf \left[R(\bar{\xi} \parallel \bar{\xi}_1 \otimes \Lambda) + R(\bar{\eta} \parallel \bar{\eta}_1 \otimes \Theta) - 2 \int \log \left(\frac{n}{N} \right) \bar{\eta}_1(dn) \right].$$

Throughout this chapter we assume that all the conditions of Theorem 2.2 are in force, and adapt the notation to reflect that fact that the local rate function no longer

depends on t . Recalling that $\gamma(dy|x)$ is the transition kernel for a single particle, it follows from Condition 1.4 and Condition 5.1 that there exists a unique pair of measures $(\pi, \kappa) \in \mathcal{P}([0, \infty)) \times \mathcal{P}(\{\Delta\} \cup [0, \infty))$ satisfying

$$\kappa = [\pi \otimes \gamma]_2, \text{ and } \pi(\cdot) = \frac{\kappa(\cdot)}{1 - \kappa(\Delta)}. \quad (6.1)$$

In other words, $\kappa(\Delta)$ is the stationary mass absorbed by Δ under γ , and if we renormalize the mass that remains then π is the invariant probability for the resulting transition kernel.

It is easy to check that the large deviation rate defined in (1.5) is exactly $\gamma = -\log(1 - \kappa(\Delta))$, and so the best possible rate of convergence for the Monte Carlo scheme is $-2\log(1 - \kappa(\Delta))$. By Jensen's inequality $V^N \leq -2\log(1 - \kappa(\Delta))$, and it is not hard to show that the inequality is in general strict. The following result shows that this rate is achieved when the number of particles tends to infinity.

Theorem 6.1. $V^N \rightarrow -2\log(1 - \kappa(\Delta))$ as $N \rightarrow \infty$.

We will study the asymptotics of V^N using many of the same weak convergence arguments used previously. We first recall the following factorization of the underlying transportation and resampling distributions:

$$\begin{aligned} \Lambda(dy|\mathbf{x}) &= \otimes_{i=1}^N \gamma(dy_i|x_i) \\ &= \otimes_{i=1}^N [\mathbf{1}_{[0,1]}(x_i)\alpha(dy_i|x_i) + \mathbf{1}_{[1,\infty)}(x_i)\delta_{x_i}(dy_i)] \\ \Theta(d\sigma|n) &= \otimes_{i=1}^n \delta_i(d\sigma_i) \otimes_{i=n+1}^N U_n(d\sigma_i). \end{aligned}$$

As in Lemma 4.5 we can factor the controls according to

$$\begin{aligned}
\bar{\xi}(d\mathbf{x}, d\mathbf{y}) &= \otimes_{i=1}^N \bar{\xi}^i(dx_i, dy_i | x_1, y_1, \dots, x_{i-1}, y_{i-1}) = \otimes_{i=1}^N \bar{\xi}^i(dx_i, dy_i) \\
\bar{\xi}_1(d\mathbf{x}) &= \otimes_{i=1}^N \bar{\xi}_1^i(dx_i | x_1, \dots, x_{i-1}) = \otimes_{i=1}^N \bar{\xi}_1^i(dx_i) \\
\bar{\eta}(d\boldsymbol{\sigma} | n) &= \otimes_{i=n+1}^N \bar{\eta}^i(d\sigma_i | n, \sigma_1, \dots, \sigma_{i-1}) \otimes_{i=1}^n \delta_i(d\sigma_i) \\
&= \otimes_{i=n+1}^N \bar{\eta}^i(d\sigma_i | n) \otimes_{i=1}^n \delta_i(d\sigma_i),
\end{aligned}$$

where the right hand side indicates the variables for which dependence is suppressed in the notation. With an abuse of notation we sometimes write $\bar{\eta}^i(d\sigma_i | n)$ for $\delta_i(d\sigma_i)$ even when $i \leq n$. Note that the controls all depend on N , but we also omit this dependence in the notation. The chain rule for relative entropy and Jensen's inequality then yields

$$\begin{aligned}
R(\bar{\xi} \| \bar{\xi}_1 \otimes \Lambda) &= \bar{\mathbb{E}} \left[\sum_{i=1}^N R(\bar{\xi}^i(dx, dy) \| \bar{\xi}_1^i(dx) \otimes \gamma(dy | x)) \right] \\
&\geq N \bar{\mathbb{E}} \left[R \left(\frac{1}{N} \sum_{i=1}^N \bar{\xi}^i(dx, dy) \left\| \frac{1}{N} \sum_{i=1}^N \bar{\xi}_1^i(dx) \otimes \gamma(dy | x) \right\| \right) \right]
\end{aligned}$$

where $\bar{\mathbb{E}}$ indicates integration with respect to the suppressed variables. Since the total cost is bounded for all N , we conclude that

$$R \left(\frac{1}{N} \sum_{i=1}^N \bar{\xi}^i(dx, dy) \left\| \frac{1}{N} \sum_{i=1}^N \bar{\xi}_1^i(dx) \otimes \gamma(dy | x) \right\| \right) \rightarrow 0$$

in probability. Tightness follows as in Chapter 4, and hence for any subsequence there exists a subsubsequence such that

$$\frac{1}{N} \sum_{i=1}^N \bar{\xi}_1^i \rightarrow \bar{\pi}$$

in probability. Thus

$$\frac{1}{N} \sum_{i=1}^N \bar{\xi}^i \rightarrow \bar{\pi} \otimes \gamma. \quad (6.2)$$

6.1 Proof of Theorem 6.1

For any feasible $(\bar{\xi}, \bar{\eta})$ controls, let $\{(\bar{X}_i^N, \bar{Y}_i^N), i = 1, 2, \dots, N\}$ be processes with the given joint distributions. Define $\bar{\mathcal{F}}_i^N$ to be the σ -field generated by $\{\bar{X}_k^N, k = 1, 2, \dots, i-1\}$ and $\bar{\mathcal{G}}_i^N$ generated by $\{(\bar{X}_k^N, \bar{Y}_k^N), k = 1, 2, \dots, i-1\}$. Then for any bounded continuous function f

$$\begin{aligned} & \left\{ f(\bar{X}_i) - \int f(x) \bar{\xi}_1^i(dx), i = 1, 2, \dots, N \right\} \\ & \left\{ f(\bar{Y}_i) - \int f(y) \bar{\xi}^i(dx, dy), i = 1, 2, \dots, N \right\} \end{aligned}$$

are martingale difference sequences, with respect to $\bar{\mathcal{F}}_i^N$ and $\bar{\mathcal{G}}_i^N$, respectively. This implies

$$\begin{aligned} & \left(\frac{1}{N} \sum_{i=1}^N \mathbf{1}_{\bar{X}_i}, \frac{1}{N} \sum_{i=1}^N \bar{\xi}_1^i \right) \rightarrow (\bar{\pi}, \bar{\pi}) \\ & \left(\frac{1}{N} \sum_{i=1}^N \mathbf{1}_{\bar{Y}_i}, \left[\frac{1}{N} \sum_{i=1}^N \bar{\xi}^i \right]_2 \right) \rightarrow (\bar{\kappa}, \bar{\kappa}). \end{aligned}$$

in distribution, and by (6.2) $\bar{\kappa}$ satisfies

$$\bar{\kappa} = [\bar{\pi} \otimes \gamma]_2. \quad (6.3)$$

Interpreting $g_{\text{loc}}(\mathbf{y})$ as vector valued with components $g_{\text{loc},i}(\mathbf{y})$ that identify the locations from which the resampling controls can select, we have

$$\mathbb{E} \left[f(\bar{X}_i) - \int f(g_{\text{loc},\sigma}(\bar{\mathbf{Y}}^N)) \bar{\eta}^i(d\sigma | g_{\text{num}}(\bar{\mathbf{Y}}^N)) \middle| \bar{\mathbf{Y}}^N, \bar{X}_1, \dots, \bar{X}_{i-1} \right] = 0.$$

This implies the bound

$$\begin{aligned}
& \bar{\mathbb{E}} \left[\left(\frac{1}{N} \sum_{i=1}^N f(\bar{X}_i) - \frac{1}{N} \sum_{i=1}^N \int f(g_{\text{loc},\sigma}(\bar{\mathbf{Y}}^N)) \bar{\eta}^i(d\sigma | g_{\text{num}}(\bar{\mathbf{Y}}^N)) \right)^2 \right] \\
&= \bar{\mathbb{E}} \left[\frac{1}{N^2} \sum_{i=1}^N \left(f(\bar{X}_i) - \int f(g_{\text{loc},\sigma}(\bar{\mathbf{Y}}^N)) \bar{\eta}^i(d\sigma | g_{\text{num}}(\bar{\mathbf{Y}}^N)) \right)^2 \right] \\
&\leq 2 \|f\|_\infty^2 / N \rightarrow 0.
\end{aligned} \tag{6.4}$$

If we can show that

$$\frac{1}{N} \sum_{i=1}^N \int \delta_{g_{\text{loc},\sigma}(\bar{\mathbf{Y}}^N)}(dx) \bar{\eta}^i(d\sigma | g_{\text{num}}(\bar{\mathbf{Y}}^N)) \rightarrow \frac{\bar{\kappa}(dx)}{1 - \bar{\kappa}(\Delta)}, \tag{6.5}$$

in probability then (6.4) implies

$$\frac{1}{N} \sum_{i=1}^N \mathbf{1}_{\bar{X}_i} \rightarrow \frac{\bar{\kappa}(dx)}{1 - \bar{\kappa}(\Delta)} = \bar{\pi}.$$

By uniqueness of the invariant distribution under the renormalized transition kernel and (6.1), we conclude

$$(\bar{\pi}, \bar{\kappa}) = (\pi, \kappa).$$

Then Theorem 6.1 will follow from the fact that by Fatou's Lemma

$$\begin{aligned}
& \liminf_{N \rightarrow \infty} V^N \\
&\geq \liminf_{N \rightarrow \infty} \left[R(\bar{\xi} \| \bar{\xi}_1 \otimes \Lambda) + R(\bar{\eta} \| \bar{\eta}_1 \otimes \Theta) - \bar{\mathbb{E}} 2 \log \left(\frac{1}{N} \sum_{i=1}^N \mathbf{1}_{[0,\infty)}(\bar{Y}_i) \right) \right] \\
&\geq -2 \log(1 - \kappa(\Delta)).
\end{aligned}$$

Note that in fact we will have shown that along the minimizing sequence

$$R(\bar{\xi} \| \bar{\xi}_1 \otimes \Lambda) + R(\bar{\eta} \| \bar{\eta}_1 \otimes \Theta) \rightarrow 0.$$

It remains to show that (6.5) is true. This is a consequence of the finiteness of the rate. In particular,

$$\begin{aligned} & R(\bar{\eta} \parallel \bar{\eta}_1 \otimes \Theta) \\ & \geq N \bar{\mathbb{E}} \left[\int R \left(\frac{1}{N} \left[\sum_{i=1}^N \bar{\eta}^i(d\sigma | n) \right] \right) \parallel \frac{1}{N} \left[(N-n)U_n(d\sigma) + \sum_{i=1}^n \delta_i(d\sigma) \right] \right) \bar{\eta}_1(dn) \right] \end{aligned}$$

for all N . Therefore

$$\frac{1}{N} \sum_{i=1}^N \int \delta_{g_{\text{loc},\sigma}(\bar{\mathbf{Y}}^N)}(\cdot) \bar{\eta}^i(d\sigma | g_{\text{num}}(\bar{\mathbf{Y}}^N))$$

has the same limit in probability as

$$\frac{1}{N} \left[\sum_{i=1}^{g_{\text{num}}(\bar{\mathbf{Y}}^N)} \int \delta_{g_{\text{loc},\sigma}(\bar{\mathbf{Y}}^N)}(\cdot) \delta_i(d\sigma) + [N - g_{\text{num}}(\bar{\mathbf{Y}}^N)] \int \delta_{g_{\text{loc},\sigma}(\bar{\mathbf{Y}}^N)}(\cdot) U_{g_{\text{num}}(\bar{\mathbf{Y}}^N)}(d\sigma) \right].$$

This completes the proof of Theorem 6.1.

Part III

Numerical Results

Chapter 7

Computational Results

In this chapter, we present some computational results for interacting particle scheme under a simple one dimensional random walk setting.

Problem Setup and Notations

Let $\{X_j, j = 0, 1, \dots\}$ be a sequence of i.i.d. random variables taking values in a discrete state space $\{-1, 1, 2\}$ with probability $\{0.6, 0.3, 0.1\}$ respectively. For each fixed $n \in \mathbb{N}$, let S_n denote the sum $X_0 + X_1 + \dots + X_{n-1}$. Let $m \in \mathbb{N}$ be the large deviation parameter, i.e., the number of thresholds. Define the stopping time $\sigma^m = \min\{n : S_n \geq m, \text{ or } S_n \leq -1\}$. Then the quantity of interest defined in (1.3) becomes

$$p^m = \mathbb{P}\{S_{\sigma^m} \geq m \mid S_0 = 0\}.$$

The importance sampling estimates of p^m for different values of m are given in Table 7.1 with sample size 100,000,000. For simplicity, we take these approximations as the true values of p^m and compare the results from IPS scheme to them.

Throughout this chapter, we use N as the number of particles of the IPS estimator and K as the sample size. $\{\hat{p}_i, i = 0, \dots, K-1\}$ represents the collection of K number

m	p^m	m	p^m
10	3.709502e-002	60	1.617500e-005
20	6.997822e-003	70	3.710000e-006
30	1.503110e-003	80	7.680000e-007
40	3.320130e-004	90	1.560000e-007
50	7.368103e-005	100	3.996866e-008

Table 7.1: estimates of p^m for different m using Importance Sampling

of IPS samples and

$$\frac{1}{K} \sum_{i=0}^{K-1} \hat{p}_i$$

is the IPS estimate of p^m .

The Large Deviation Rate of the Second Moment

In Figure 7.1, we compare the large deviation rate of the second moment of the IPS estimators for different m and N . In each case, the sample size K is taken to be 10,000. The large deviation rate of the second moment is approximated by computing

$$-\log \left[\frac{1}{K} \sum_{i=0}^{K-1} \hat{p}_i^2 \right].$$

The black dashed line represents the asymptotic optimal rate $-2 \log p^m$ which is approximately an affine function of m for large m . Here p^m are evaluated by important sampling. The theoretical results from previous chapter indicate that IPS scheme is asymptotic suboptimal and the large deviation rate of its second moment is a function of N . However, as the number of particles N goes to ∞ , IPS scheme becomes asymptotic optimal.

As one can see in Figure 7.1, the large deviation rate of IPS converges to the asymptotic optimal rate as N grows, which supports the theoretical results.

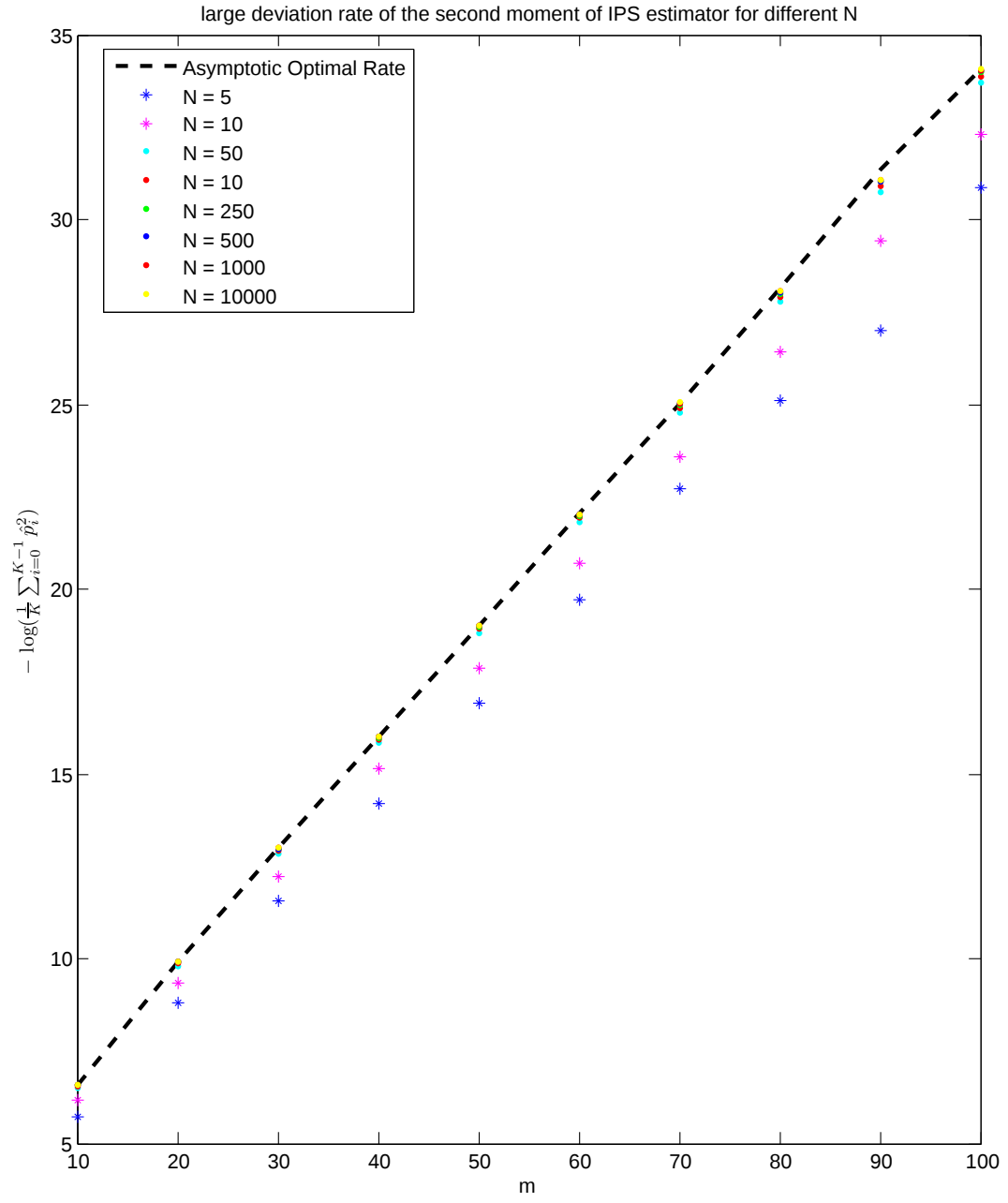


Figure 7.1: large deviation rate of the second moment of the interacting particle system estimator for different m and N

Trade-off Between N and K

Another interesting question to ask related to IPS estimator is the trade-off between the number of particles N and number of samples K . We illustrate the problem in two aspects:

1. The empirical standard error:

$$\sigma_K^N = \sqrt{\frac{1}{K-1} \sum_{i=0}^{K-1} (\hat{p}_i - \hat{p})^2}$$

where $\hat{p}$ is the sample mean of IPS estimates.

2. The per sample standard error:

$$\sqrt{\text{var} \left(\frac{1}{K} \sum_{i=0}^{K-1} \hat{p}_i \right)} = \sqrt{\frac{\text{var}(\hat{p}_i)}{K}} \approx \frac{\sigma_K^N}{\sqrt{K}}.$$

To begin with, we fixed the computational efforts. That is, we let the total number of simulated paths to be a constant, i.e. $N \times K = \text{constant}$. The constant is taken to be 1,000,000 in our simulation.

Figure 7.2 and 7.3 plot the decaying of the empirical standard error against the increasing of number of particles N for $m = 10$ and $m = 50$ respectively. Note that the sample size K is also varying since we keep the total computational effort fixed. N ranges from 5 to 100 with increment of size 1.

Figure 7.4 compared the changes of per sample standard error to the changes N . The total range of N is within $[5, 10000]$. The top left graph plots the changes of $\sigma_K^N/\sqrt{K}$ for $N \in \{5, 6, \dots, 100\}$. As one can see, there is a significant per sample standard error reduction for N less than 40. The number of particles in the top left graph, which is part of the bottom graph, ranges from 5 to 1,000 with increment of size 5. It shows that there is no per sample standard error reduction for N greater

than 100. The bottom graph shows the overall trends of $\sigma_K^N/\sqrt{K}$.

In conclusion, the computational results suggest that the number of particles N should be chosen to be around 40 for our problem.

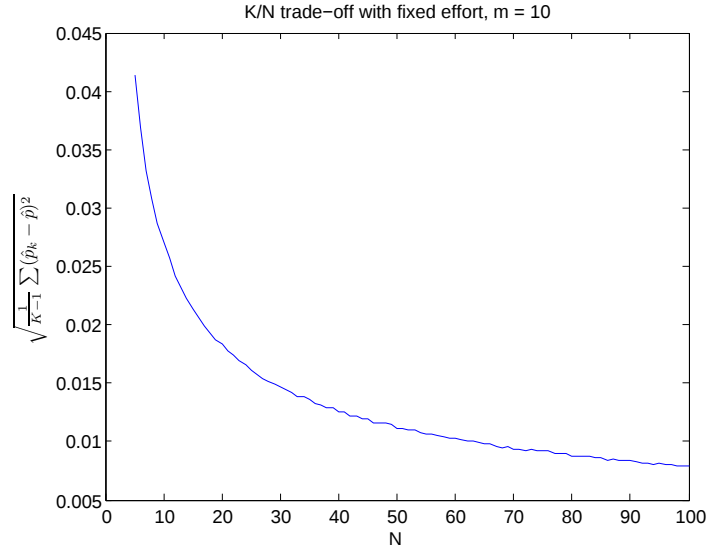


Figure 7.2: trade-off between N and K with fixed effort, $N \times K = 1,000,000$, $m = 10$

General Results

In this section, we present some basic results of IPS scheme under the i.i.d random walk setting with $m = 10$ and 50. Figure 7.5 and 7.6 plot the changes of sample error

$$\frac{1}{K} \sum_{i=0}^{K-1} \hat{p}_i - p^m$$

for fixed sample size $K = 10,000$, where p^m is the true value obtain from importance sampling. It appears that the sample error is asymptotic independent of N for large N which is what we expected.

For fixed sample size $K = 10,000$, Figure 7.7 and 7.8 show that the empirical

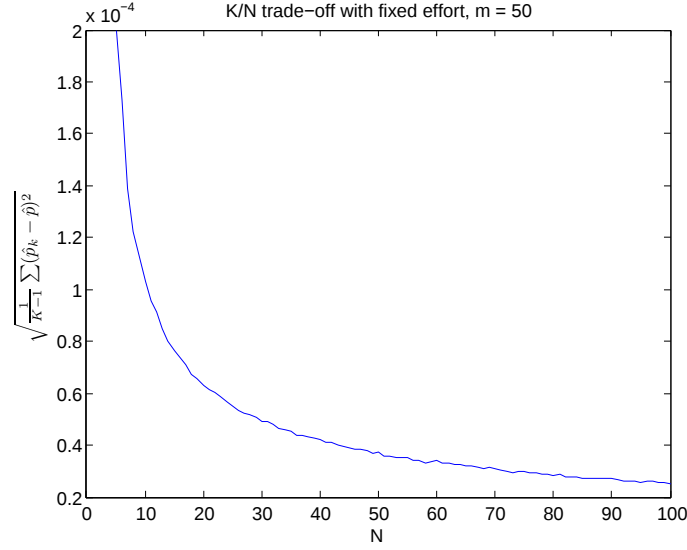


Figure 7.3: trade-off between N and K with fixed effort, $N \times K = 1,000,000$, $m = 50$

standard error decreases as particle size N grows.

Figure 7.9 plots the changes of empirical standard error with different sample size K while N is fixed. As expected, the estimates are roughly a constant.

We predict the rate of decay of the second moment of IPS scheme in N is of order $1/N$. If true, the statistic

$$\frac{N}{m} \log \frac{\frac{1}{K} \sum_{i=0}^{K-1} \hat{p}_i^2}{(p^m)^2} \quad (7.1)$$

should be a constant of N . However, such statistic requires m to be comparably large to N . In Figure 7.10, we can't not see such trend because m is too small. In Figure 7.11, one can see that for N less than 1,000, the quantity (7.1) is roughly a constant. However, since sample error is a constant of K , for fixed K (7.1) amplifies such error by size N . Hence, when N grows large in Figure 7.11, we see a increasing oscillation of (7.1). One of our future work will be theoretically proving such decay rate in N .

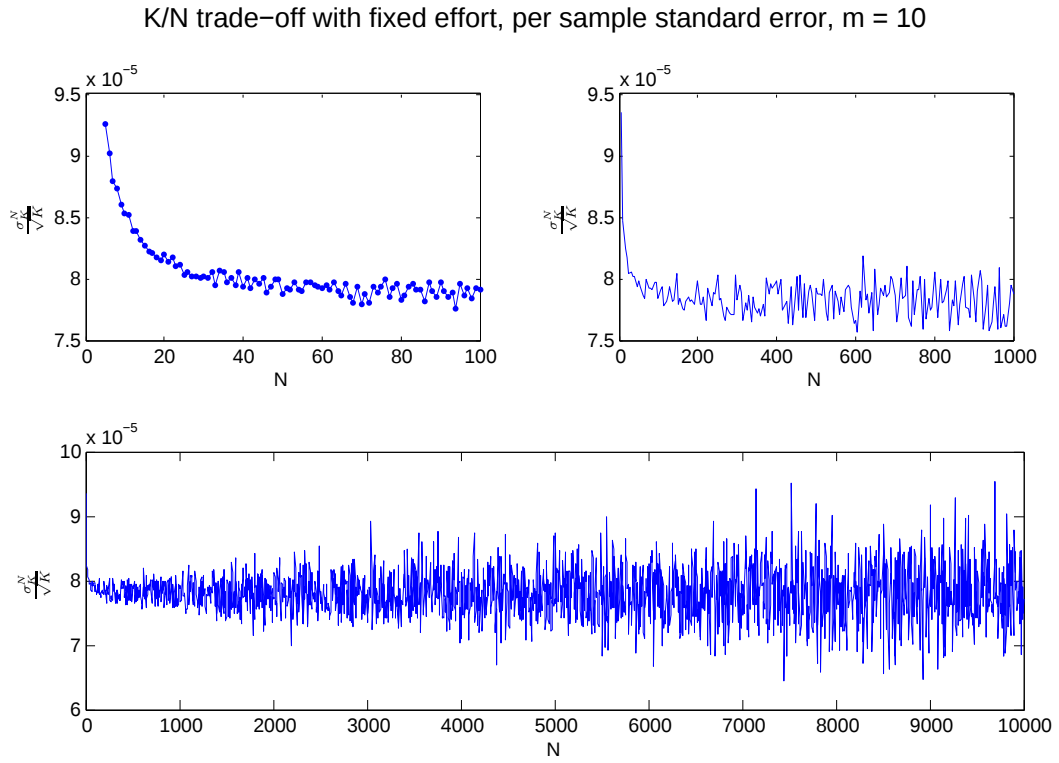
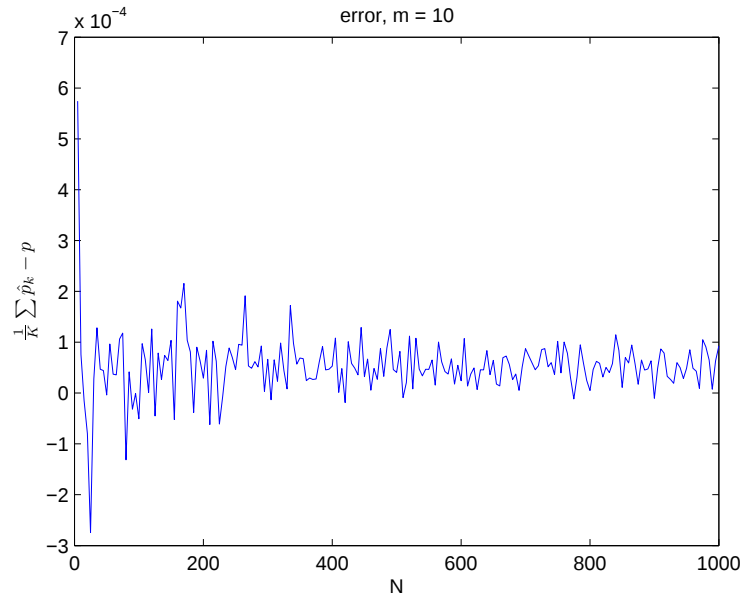
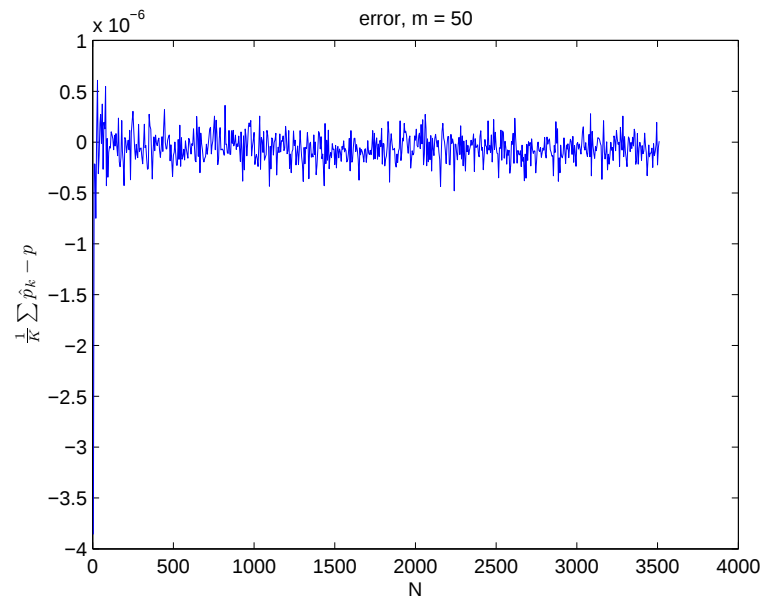


Figure 7.4: trade-off between N and K with fixed effort, $N \times K = 1,000,000$, $m = 10$

Figure 7.5: sample error, $m = 10$, fixed K Figure 7.6: sample error, $m = 50$, fixed K

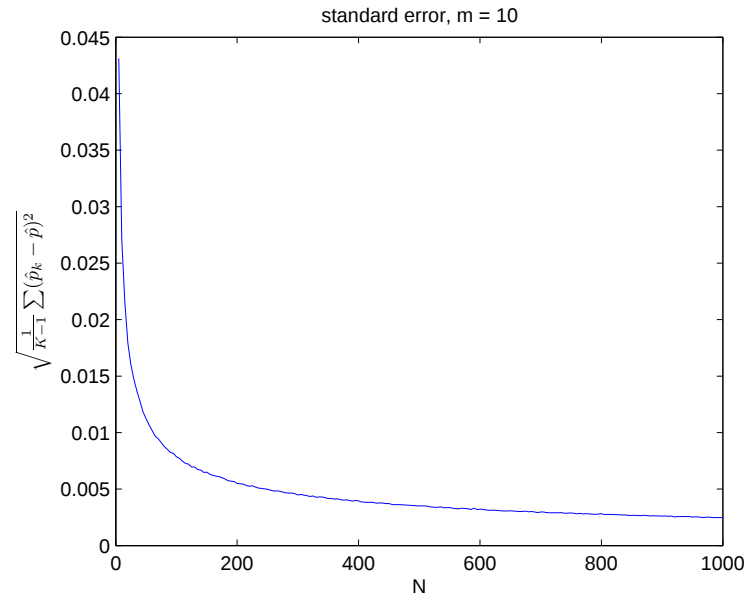


Figure 7.7: standard error, $m = 10$, fixed K

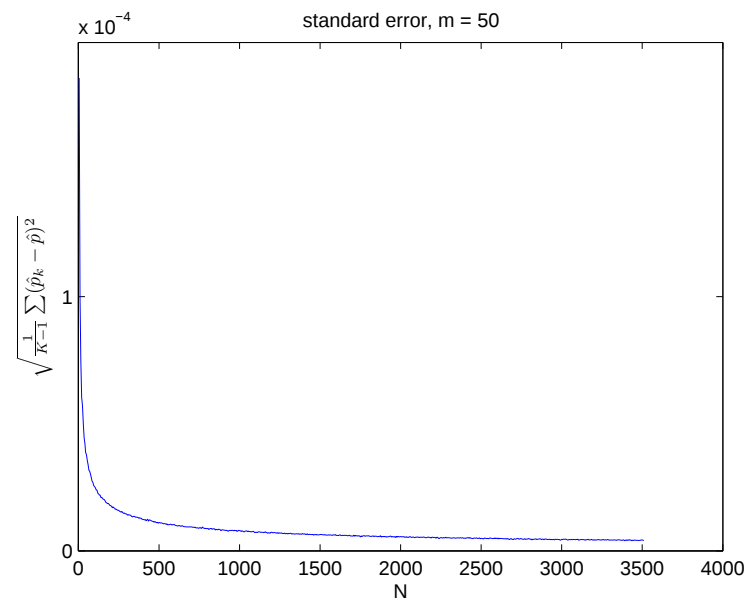


Figure 7.8: standard error, $m = 50$, fixed K

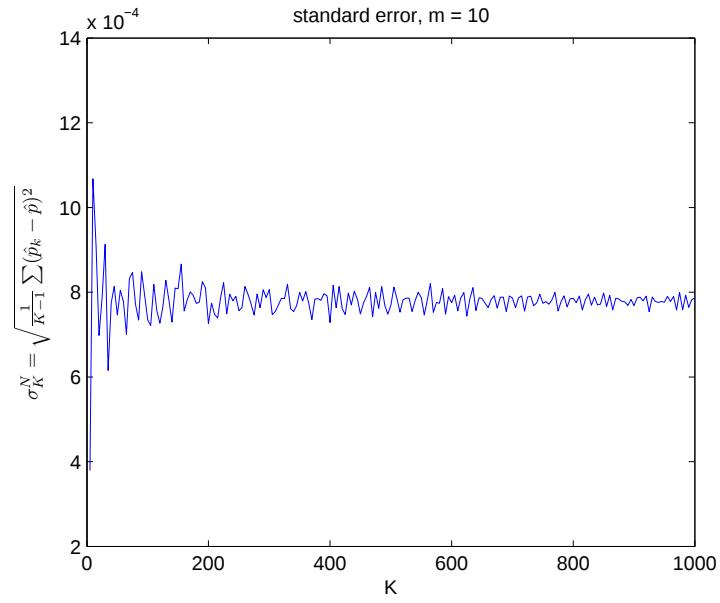


Figure 7.9: standard error, $m = 10$, fixed N

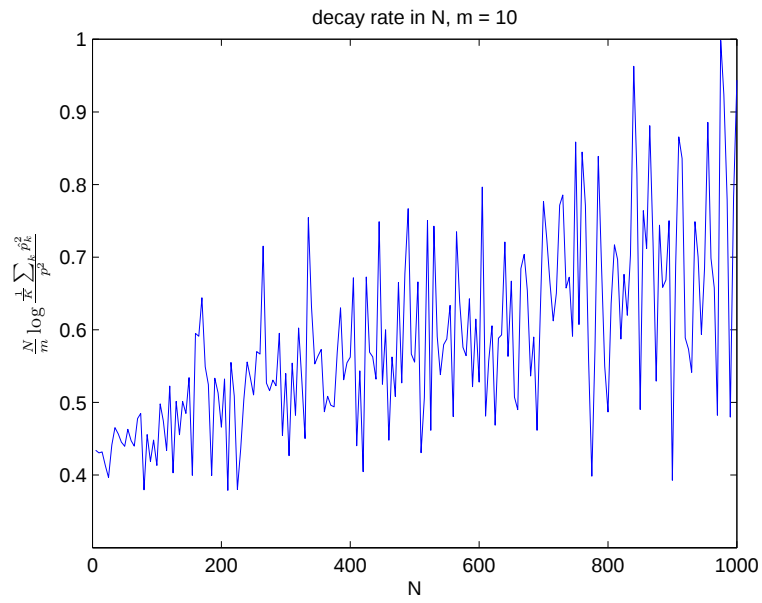


Figure 7.10: decay rate in N , $m = 10$, fixed K

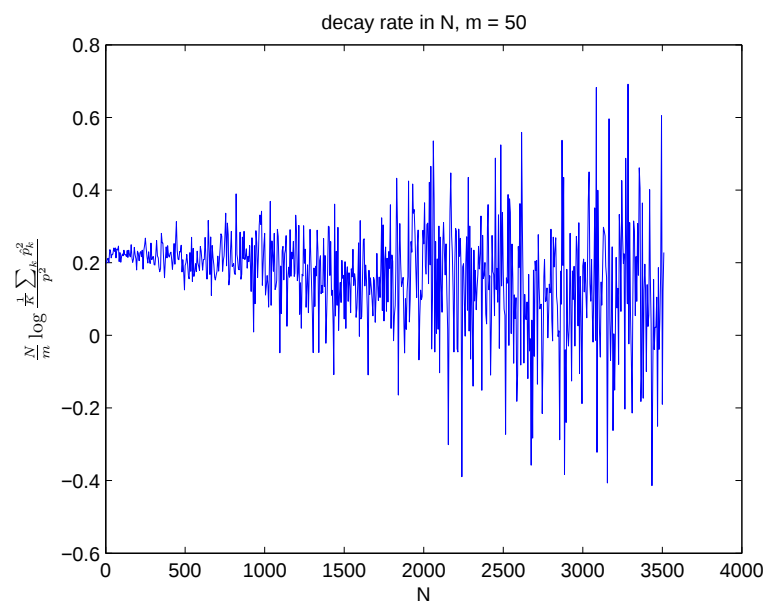


Figure 7.11: decay rate in N, $m = 50$, fixed K

Bibliography

- [1] L. Breiman. *Probability Theory*. Addison-Wesley, Reading, Mass., 1968.
- [2] T. Dean and P. Dupuis. Splitting for rare event simulation: A large deviations approach to design and analysis. *Stoch. Proc. Appl.*, 119:562–587, 2009.
- [3] T. Dean and P. Dupuis. The design and analysis of a generalized RESTART/DPR algorithm for rare event simulation. *Math. of Operations Research*, page to appear, 2010.
- [4] P. Del Moral and J. Garnier. Genealogical particle analysis of rare events. *Annals of Applied Probability*, 15:2496–2534, 2005.
- [5] P. Del Moral and P. Lezaud. Branching and interacting particle interpretations of rare event probabilities. In *Stochastic Hybrid Systems, Lecture Notes in Control and Information Science*, volume 337. Springer, Berlin, 2006.
- [6] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Jones and Bartlett, Boston, 1993.
- [7] 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.
- [8] M.D. Donsker and S.R.S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, II. *Comm. Pure Appl. Math.*, 28:279–301, 1975.

- [9] P. Dupuis and R. S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. John Wiley & Sons, New York, 1997.
- [10] P. Dupuis and H. Wang. Importance sampling, large deviations, and differential games. *Stoch. and Stoch. Reports.*, 76:481–508, 2004.
- [11] P. Dupuis and H. Wang. Dynamic importance sampling for uniformly recurrent Markov chains. *Ann. Appl. Prob.*, 15:1–38, 2005.
- [12] P. Dupuis and H. Wang. Subsolutions of an Isaacs equation and efficient schemes for importance sampling. *Math. Oper. Res.*, 32:1–35, 2007.
- [13] P. Dupuis and O. Zeitouni. A nonstandard form of the rate function for the occupation measure of a Markov chain. *Stoch. Proc. and Their Appl.*, 61:249–261, 1996.
- [14] M.J.J. Garvels. *The Splitting Method in Rare Event Simulation*. PhD thesis, University of Twente, The Netherlands, 2000.
- [15] M.J.J. Garvels and D.P. Kroese. *A comparison of RESTART implementation*. IEEE Publishers, Washington, DC, 1998.
- [16] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. A large deviations perspective on the efficiency of multilevel splitting. *IEEE Trans. Automat. Control*, 43:1666–1679, 1998.
- [17] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. Multilevel splitting for estimating rare event probabilities. *Operations Research*, 47:585–600, 1999.
- [18] Z. Haraszti and J.K. Townsend. The theory of direct probability redistribution and its application to rare event simulation. *ACM Transactions on Modelling and Computer Simulation*, 9:105–140, 1999.

- [19] H. Kahn and T.E. Harris. Estimation of particle transmission by random sampling. *National Bureau of Standards Applied Mathematics Series*, 12:27–30, 1951.
- [20] H. J. Kushner. *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, volume 3 of *Systems and control*. Birkhaeuser, Boston, 1990.
- [21] S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability, Second Edition*. Cambridge University Press, New York, 2009.
- [22] M. Villen-Altamirano and J. Villen-Altamirano. RESTART: A method for accelerating rare event simulations. In *Proc. of the 13th International Teletraffic Congress, Queueing, Performance and Control in ATM*, pages 71–76, Elsevier, Amsterdam, 1991.
- [23] M. Villen-Altamirano and J. Villen-Altamirano. RESTART: A straightforward method for fast simulation of rare events. In *Proc. of the 1994 Winter Simulation Conference*, pages 282–289, 1994.
- [24] M. Villen-Altamirano and J. Villen-Altamirano. Analysis of RESTART simulation: Theoretical basis and sensitivity study. *European Transactions on Telecommunications*, 13:373–385, 2002.

Index

- Δ , 23
- Λ_j^m , 33
- Θ , 33
- $\alpha_j^m(A|x)$, 36
- $\alpha_j^m(dy|x)$, 31
- $\bar{\Lambda}_j^m$, 35
- $\bar{\Theta}_j^m$, 35
- $\bar{\eta}^m(C_1 \times C_2 \times B)$, 43
- $\bar{\eta}^m(dn, d\sigma|t)$, 43
- $\bar{\xi}^m(\mathbf{A}_1 \times \mathbf{A}_2 \times B)$, 43
- $\bar{\xi}^m(d\mathbf{x}, d\mathbf{y}|t)$, 43
- $\check{\alpha}_j^m(dy|x)$, 32
- $\gamma(A|v, t)$, 24
- $\bar{\sigma}_j^m$, 35
- σ , 34
- μ , 18
- ρ , 19
- σ^m , 18
- σ_j^m , 32
- τ_C , 5
- $\theta_i(x)$, 17
- $\tilde{\Lambda}_j^m$, 63
- $\tilde{\gamma}(A|x, t)$, 61
- $\tilde{\phi}(\{n\}|t)$, 70
- $\tilde{\pi}(d\mathbf{x}|t)$, 65
- $G_N(t, \phi)$, 58, 76
- $I(\nu)$, 59
- $A(\phi)$, 57
- C_j^m , 20
- F^N , 28
- $H(x, \alpha)$, 18
- $L(x, \beta)$, 19
- L^m , 22
- N , 20
- N_j^m , 20
- U_n , 33
- $X_{j,n}^m$, 20, **33**
- $Y_{j,n}^m$, 20
- Z_i^m , 18
- $\bar{L}^m$, 44
- $\bar{N}_j^m$, 37
- $\bar{\mathbf{X}}_j^m$, 42
- $\hat{p}^m$, 21
- $\mathbb{P}^m$, 34
- $\mathbb{Q}^m$, 34
- $\bar{\mathbf{Y}}_j^m$, 35

$\mathcal{S}_j^m$, 20

$\mathcal{T}_1$, 63

$\mathcal{T}_2$, 63

$\mathcal{X}$, 34

$\mathcal{Y}$, 34

$\mathcal{Z}$, 34

$\tilde{p}(\mathbf{x}, \cdot)$, 63

$c(\mathbf{x}, t)$, 63

$g = (g_{\text{num}}, g_{\text{loc}})$, 36

$g_j^m = (g_{\text{num},j}^m, g_{\text{loc},j}^m)$, 33

h , 33

m , 19

p^m , 18

$q^m(C_1 \times C_2 \times B)$, 44

$q^m(dn, d\boldsymbol{\sigma} | t)$, 43

$r^m(\mathbf{A}_1 \times \mathbf{A}_2 \times B)$, 44

$r^m(d\mathbf{x}, d\mathbf{y} | t)$, 43