Large Deviations for a Feed-forward Network

&

Importance Sampling for a Single Server Priority Queue

by

Leila Setayeshgar

B.Sc., Sharif University of Technology; Tehran, Iran, 1998

M.S., Northeastern University; Boston, MA, 2000

M.S., California Institute of Technology; Pasadena, CA, 2004

M.S., Brown University; Providence, RI, 2008

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

This dissertation by Leila Setayeshgar 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_____ _____

Hui Wang, Ph.D., Advisor

Recommended to the Graduate Council

Date_____ _____

Jose H. Blanchet, Ph.D., Reader

Date_____ _____

Justin Holmer, Ph.D., Reader

Date_____ _____

Kavita Ramanan, Ph.D., Reader

Approved by the Graduate Council

Date_____ _____

Peter Weber, Ph.D., Dean of the Graduate School

Vitae

Education

- B.Sc., Sharif University of Technology; Tehran, Iran, 1998
- M.S., Northeastern University; Boston, MA, 2000
- M.S., California Institute of Technology; Pasadena, CA, 2004
- M.S., Brown University; Providence, RI, 2008
- Ph.D., Brown University; Providence, RI, 2012

Dedication

To my *parents*

Contents

Dedication viii

1 Introduction **1**

1.1 Background and Motivation 1

1.2 Overview 4

1.3 Outline 4

2 Large Deviations for a Feed-forward Network **5**

2.1 Overview 5

2.2 The Model and System Dynamics 9

2.3 Large Deviations Analysis 11

2.3.1 Hamiltonians and Rate Functions 12

2.3.2 Sample Path Large Deviations 13

2.3.3 The Main Theorem 13

2.4 A Case Study: The 2-dimensional Network 28

2.4.1 Three Important Roots of the Hamiltonians 29

2.4.2 The Exponential Decay Rate of p_n 34

2.5 Summary 41

3 Importance Sampling for a Single Server Priority Queue **42**

3.1 Overview 42

3.2 System Model and Dynamics 44

3.3 The Rare Event 45

3.4 Review of Large Deviations Results 46

3.4.1 System Hamiltonians 47

3.4.2 The Exponential Decay Rate of p_n 47
3.5 Importance Sampling 49
3.5.1 Asymptotic Optimality 49
3.5.2 Classical Subsolution Approach 50
3.5.3 Piecewise Constant Change of Measure 51
3.5.4 The Importance Sampling Estimator 53
3.5.5 The Verification Argument 55
3.6 Numerical Results 61
3.7 Summary 62

A Collection of Proofs (Chapter II) 64

A.1 Proof of Lemma 2.3.2 65
A.2 Proof of Lemma 2.3.3 68
A.3 Proof of Lemma 2.4.2 70

B Collection of Proofs (Chapter III) 73

B.1 Proof of Lemma 3.5.2 74
B.2 Proof of Lemma 3.5.3 75

List of Figures

2.1 Feed-forward network with preemptive priority service policy 10

2.2 System Dynamics for $d = 2$ 11

2.3 Representative limit sample path ϕ 30

2.4 Geometry and trajectory (I)..... 31

2.5 Geometry and trajectory (II) 33

3.1 A feed-forward network with priority service policy..... 44

3.2 Discontinuous dynamics..... 45

3.3 Roots of the Hamiltonians..... 48

3.4 An example of $\alpha^{[1]}$ and $\alpha^{[2]}$: case 1 with $\alpha_1^* > \hat{\alpha}_1$ 53

CHAPTER I:

Introduction

1.1 Background and Motivation

Theory of large deviations has its roots in actuarial science when F. Esscher (in the 1930s) became interested in finding the rare event probability that the total earnings in an insurance company exceeds the total claim. He modeled the claims as independent random variables and associated a distribution to them. The probability of interest, then became the estimation of the *tail probabilities* of sums of independent random variables. This marked the origin of the theory of large deviations. Over the years, other scientists including H. Cramér made further contributions to the subject, but a formal definition was given by S. R. S. Varadhan in the 1960s. In short, theory of large deviations takes the *Central Limit Theorem* one step further. To make this precise, consider a sequence of *i.i.d* random variables $\{X_k\}$ with mean zero and unit variance. By the *Law of Large Numbers* it follows that

$$S_n = \frac{1}{n}\sum_{k=1}^{n} X_k$$

converges to zero with probability one. Thus, for any $\delta > 0$,

$$P(|S_n| \geq \delta) \to 0 \qquad as\ n \to \infty, \tag{1.1}$$

i.e. for n sufficiently large, $(|S_n| \geq \delta)$ is a rare event. On the other hand, the Central Limit Theorem (CLT) asserts that as n approaches infinity, the random variable $\sqrt{n}S_n$ converges in *distribution* to that of a normal $N(0,1)$, therefore for any $A \subseteq \mathbb{R}$

$$P(\sqrt{n}S_n \in A) \to \frac{1}{\sqrt{2\pi}} \int_A e^{-x^2/2} dx \qquad as\ n \to \infty. \tag{1.2}$$

Let $A \doteq \mathbb{R} \setminus (-\delta\sqrt{n}, \delta\sqrt{n})$, then

$$\text{L.H.S.} \doteq P(\sqrt{n}S_n \in A) = P(|S_n| \geq \delta).$$

On the other hand, by direct calculation

$$\text{R.H.S.} \doteq \frac{1}{\sqrt{2\pi}} \int_A e^{-x^2/2} dx = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\delta\sqrt{n}}^{\delta\sqrt{n}} e^{-x^2/2} dx \approx e^{-n\delta^2/2},$$

thus applying (1.2) *naively* yields

$$\frac{1}{n} \log P(|S_n| \geq \delta) \to -\frac{\delta^2}{2} \qquad as\ n \to \infty. \tag{1.3}$$

Note that (1.1) and (1.2) hold true as long as $\{X_k\}$ are i.i.d with zero mean and unit variance, *regardless* of the distribution of each X_k; however, Cramér's theorem asserts that the value of the limit in (1.3) *does* depend on the distribution of $\{X_k\}$ – i.e., the limit is not unique for all distributions. Therefore, in the case of a rare event where the set A depends on "n", using CLT to replace $\sqrt{n}S_n$ by its distributional limit, is a naive approach; instead, one should appeal to the "Large Deviation Limit" of $n^{-1}\log P(|S_n| \geq \delta)$.

Large deviations can also answer questions regarding the sample paths of stochastic processes. For example, let $X^n(t)$ be a family of processes with a deterministic limit as $n \to \infty$, then large deviations is able to identify the rate of this convergence. Therefore, through sample path large deviations one can identify the exponential decay rate of the probability of a rare event. While large deviations can successfully identify this rate of convergence, the estimation of the "exact" probability of interest belongs to the subject of stochastic simulation.

Stochastic simulation goes back to the work of Ulam, and Von Neuman (among others), who coined the term "Monte Carlo" in Los Alamos in order to construct better atomic bombs. Monte Carlo, which is particularly useful for simulating systems with many coupled degrees of freedom such as fluids, exploits the Law of Large Numbers in order to approximate expectations. Note that many quantities of interest such as probabilities, integrals, and summations can be cast as expectations. The consequence of this is that "probabilities" can be approximated by the Monte Carlo method; however, the natural question to ask is whether Monte Carlo is also efficient when dealing with rare events. The answer is *not* affirmative. The reason for which is that the *relative error* (which is of the order of the inverse of the probability of the rare event) renders the simulation computationally inefficient. "Importance Sampling" – a general variance reduction technique – (which dates back to the 1950s and is illustrated best in the work of D. Siegmund) is a remedy to this situation. An importance sampling scheme generates samples from a new probability distribution under which the rare event is no longer rare, i.e.

$$\text{rare event (under old measure)} \ \to \text{common event (under new measure)}.$$

However, there are many changes of measure that can turn a rare event into a common event. The choosing of the measure which reduces the variance most, is

the core of the importance sampling technique and is referred to as "Asymptotic Optimality" (This will be made precise in Section 3.5.5).

1.2 Overview

In this thesis, we consider a feed-forward network with a single server station serving jobs with multiple levels of priority. The service discipline is preemptive in that the server always serves a job with the current highest priority level. For this system with discontinuous dynamics, we show that the family of scaled state processes satisfy the sample path large deviations principle using a weak convergence argument. In the special case where the jobs have two different levels of priority, we explicitly identify the exponential decay rate of the probability a rare event, namely, the total population overflow associated to the feed-forward network. We then use importance sampling – a variance reduction technique – efficient for rare event probabilities to simulate the exact probability of interest.

1.3 Outline

This thesis is organized as follows. In the second Chapter we establish a large deviation principle for the family of scaled state processes, where we employ the weak convergence approach. The identification of the exponential decay rate of the probability of the rare event of interest (i.e., the total population overflow) associated to the 2-dimensional network is also performed in this chapter. The third Chapter is concerned with simulating the exact probability of interest via importance sampling. The two main contributions are proposing an importance sampling estimator for evaluating the probability of interest, and verifying that this estimator is in fact asymptotically optimal. The chapter concludes with numerical simulations that confirm the asymptotic optimality of our schemes. A collection of proofs is presented in the Appendices.

CHAPTER II:

Large Deviations for a Feed-forward Network

2.1 Overview

We consider a single server station with multiple classes of exogenous jobs, where each class is assigned a priority level. The service discipline is preemptive in that the server always serves a job with the current highest level of priority. Jobs with the same priority level are served under the first-in-first-out policy. This model is probably the simplest feed-forward network with preemptive priority discipline [5]. Yet, it still captures the source of difficulty in the analysis of such systems, namely, the discontinuous dynamics due to the preemptive service policy.

Theory of large deviations is concerned with the asymptotic behavior of tails of sequences of probability distributions. Let S be a Polish space (i.e., a complete, separable, metric space) equipped with the Borel σ-algebra and $\{X^n\}$ a sequence of S-valued random variables. A lower semicontinuous function $I : S \to [0, \infty]$ with compact level sets is said to be a large deviation *upper bound rate function* if for

every closed subset F of S

$$\limsup_n \frac{1}{n} \log P(X^n \in F) \leq - \inf_{x \in F} I(x).$$

Similarly, I is said to be a large deviation *lower bound rate function* if for every open subset G of S

$$\liminf_n \frac{1}{n} \log P(X^n \in G) \geq - \inf_{x \in G} I(x).$$

If I is both an upper and a lower bound rate function, then $\{X^n\}$ satisfies the large deviation principle with rate function I.

Large deviations analysis for stable stochastic systems with continuous dynamics has been a classical topic in probability theory [15]. However, the general methodologies and techniques therein cannot be applied to models with discontinuous dynamics that arise naturally in a variety of applications (notably queueing networks). In the last two decades, research on the large deviations properties of such models has become more and more popular and many interesting results have been obtained [4, 6, 14, 20, 21]. With minor regularity conditions, it is possible to establish an explicit large deviation upper bound rate function [29] for stochastic systems with very general discontinuous dynamics. However, this upper bound rate function is *not* a lower bound rate function in general [1, 17]. The reason for this gap lies in the so called "stability-about-the-interface" condition. To give an intuitive explanation, let us consider a simple model of random walk in $\mathbb{R}^d$ where the dynamics are constant in the two half spaces $\Lambda_1 = \{x \in \mathbb{R}^d : x_1 \leq 0\}$ and $\Lambda_2 = \{x \in \mathbb{R}^d : x_1 > 0\}$. Denote by L_i the large deviation local rate function for the dynamics in the region Λ_i, $i = 1, 2$. The upper bound rate function suggested by [29] on the interface $\Sigma = \{x \in \mathbb{R}^d : x_1 = 0\}$ is the inf-convolution of L_1 and L_2. That is, for every $x \in \Sigma$

and $\beta \in \mathbb{R}^d$

$$L(x;\beta) = \inf\left[\rho_1 L_1(\nu) + \rho_2 L_2(\theta)\right], \tag{2.1}$$

where the infimum is taken over all quadruples $(\nu, \theta, \rho_1, \rho_2)$ such that

$$\nu \in \mathbb{R}^d,\ \theta \in \mathbb{R}^d,\ \rho_1 \geq 0,\ \rho_2 \geq 0,\ \rho_1 + \rho_2 = 1,\ \rho_1\nu + \rho_2\theta = \beta. \tag{2.2}$$

This upper bound rate function L is not a lower bound rate function in general. Indeed, it is shown in [28, Chapter 7] that the large deviation rate function is defined exactly as in (2.1) but with extra constraints (i.e., the "stability-about-the-interface" condition)

$$\nu_1 \geq 0, \quad \theta_1 \leq 0$$

in (2.2). The reason for these extra constraints is that in order to prove a large deviation lower bound, one needs to analyze the cost associated with a piece of trajectory that travels on the interface Σ. This is usually achieved by a change of measure argument so that the state process closely tracks the trajectory under the new probability distribution. The vital role of this stability-about-the-interface condition is to characterize all those changes of measures that lead to the desired tracking behavior; see [28, Chapter 7] for more details.

The current Chapter consists of two parts. In the first part we establish the sample path large deviation principle for the feed-forward network under consideration. It turns out that the "stability-about-the-interface" condition is *implicitly* built into the upper bound rate function [29]. Consequently, the upper bound rate function is indeed the rate function. Similar results have been obtained in [3], whose analysis uses the techniques of the Skorokhod Problem and therefore do not apply here. We

also wish to point out that [18] can be applied to the current system to establish a sample path large deviation principle. However, in [18] the rate function is only implicitly defined in terms of the convergence parameters of the transform semigroup. Furthermore, we use a different approach based on weak convergence, which seems to be very powerful especially in dealing with discontinuous dynamics; see also [10].

The simple form of the upper bound rate function (or the rate function) allows one to characterize through partial differential equations the asymptotic behavior of various types of buffer overflow probabilities. In the second part, we illustrate this connection by explicitly identifying the exponential decay rate of the total population overflow probabilities when the exogenous jobs have two levels of priority. The form of the decay rate is motivated by examining the geometry of the zero levels sets of the system Hamiltonians, and then rigorously verified by constructing suitable subsolutions to the related partial differential equation.

This Chapter is partly motivated by the problem of estimating various buffer overflow probabilities for feed-forward networks via importance sampling. It serves as a starting point towards large deviation analysis for more complicated networks with preemptive priority service disciplines. The analysis suggests that it may not be uncommon for the "stability-about-the-interface" condition to hold automatically for physically meaningful systems; see also [10]. This leads to the interesting open question of establishing a general sufficient condition to recognize such systems.

This Chapter is organized as follows. In Section 2.2 the model setup and system dynamics are introduced. The large deviation analysis of the scaled state process is performed in Section 2.3. In Section 2.4 we specialize to the two-dimensional case and explicitly identify the exponential decay rate of the total population overflow probabilities. A brief summary is given in Section 2.5. Some of the technical proofs

are deferred to appendices.

Remark on Notation: Unless otherwise specified, we will adopt the following notation.

1. If x is a vector, then x_i denotes its i-th component.
2. If β_i is a vector, then $[\beta_i]_k$ denotes its k-th component.
3. e_i denotes the vector with the i-th component 1 and 0 otherwise.
4. The sup-norm is denoted by $\|\cdot\|_\infty$. For example, say $f(x,t)$ is a function on $\mathbb{R}^d \times [0,T]$. Then
$$\|f\|_\infty = \sup_{(x,t)\in\mathbb{R}^d\times[0,T]} |f(x,t)|.$$
5. A collection of random variables that take values in a Polish space S is said to be tight if the probability measures that these random variables induce on S are tight.
6. At times random variables and stochastic processes will be defined on different probability spaces. This happens, for example, when the Skorohod Representation Theorem is invoked. To ease exposition, we will use the same notation E to denote the expectation on all these different probability spaces.

2.2 The Model and System Dynamics

Consider a single server station serving d classes of exogenous jobs. Jobs of class i, $i = 1,\ldots,d$ arrive according to a Poisson process with rate $\lambda_i > 0$, and are buffered at queue i. The service time for a class i job is exponentially distributed with rate $\mu_i > 0$. The arrival processes and service times are assumed to be mutually

independent. The system adopts a service discipline such that a job of class i has preemptive priority over a job of class j whenever $i > j$, and the server always serves a job with the current highest level of priority. Jobs with the same priority level are served according to the first-in-first-out policy. See Figure 2.1. The state process

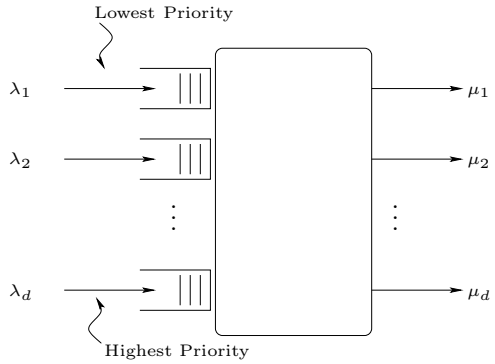

Figure 2.1: Feed-forward network with preemptive priority service policy

$Q = \{(Q_1(t), \ldots, Q_d(t)) : t \geq 0\}$ is a d-dimensional process where $Q_i(t)$ denotes the queue size of class i job at time t. It is a continuous time pure jump Markov process defined on some probability space, say, $(\Omega, \mathcal{F}, \mathbb{P})$. Define $\Pi(x)$ to be the index of the non-empty queue with the highest priority at state $x = (x_1, \ldots, x_d) \in \mathbb{R}^d_+$, that is,

$$\Pi(x) = \max\{i : x_i > 0\} \quad \text{with convention } \Pi(0) = 0. \tag{2.3}$$

Note that the mapping Π is *lower semicontinuous*. Under the preemptive service policy, the set of all possible jumps of Q is

$$\mathbb{V} = \{\pm e_1, \ldots, \pm e_d\},$$

and the jump intensity from state x to state $x + v$ is defined as

$$r(x, v) = \begin{cases} \lambda_i & \text{if } v = e_i, \\ \mu_i & \text{if } v = -e_i \text{ and } i = \Pi(x) \geq 1, \\ 0 & \text{otherwise.} \end{cases}$$

The dynamics of the system are discontinuous at the interface $\{x : \Pi(x) = i\}$ for each $0 \leq i \leq d-1$. Thus there are in total d interfaces of discontinuity whose dimensions range from 0 to $d-1$. These interfaces are also boundaries of the state space.

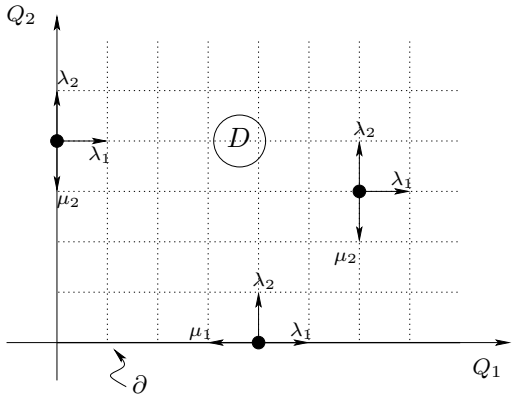

Figure 2.2: System Dynamics for $d = 2$

2.3 Large Deviations Analysis

In this section we study the sample path large deviation properties of the state process Q. To this end, we define the scaled state process

$$X^n(t) = \frac{1}{n} Q(nt).$$

Our goal is to show that the family of processes $\{X^n(t) : t \in [0,T], n \in \mathbb{N}\}$ (which are again continuous time, pure jump Markov processes) satisfy the large deviation principle with "some" rate function. In order achieve this goal, we need the following definitions.

2.3.1 Hamiltonians and Rate Functions

For every $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{R}^d$, we define

$$H_0(\alpha) = \sum_{k=1}^{d} \lambda_k (e^{\alpha_k} - 1), \tag{2.4}$$

$$H_i(\alpha) = \mu_i (e^{-\alpha_i} - 1) + \sum_{k=1}^{d} \lambda_k (e^{\alpha_k} - 1), \quad 1 \le i \le d. \tag{2.5}$$

The functions $H_0, H_1, \ldots, H_d$ are all strictly convex, and H_i corresponds to the Hamiltonian in the region $\{x \in \mathbb{R}^d_+ : \Pi(x) = i\}$. These Hamiltonians are closely related to the log of the moment generating functions of the infinitesimal increments of the process Q. Therefore, they play an important role in the PDE approach to the large deviation analysis [9].

For each i denote by L_i the Legendre transform of H_i, that is, for each $\beta \in \mathbb{R}^d$,

$$L_i(\beta) = \sup_{\alpha \in \mathbb{R}^d} [\langle \alpha, \beta \rangle - H_i(\alpha)]$$

Define $\oplus$ as the *inf-convolution* operator and $\bar{L}_i$ as the inf-convolution of L_i, L_{i+1}, ..., L_d. That is, for every $\beta \in \mathbb{R}^d$,

$$\begin{aligned} \bar{L}_i(\beta) &= (L_i \oplus L_{i+1} \oplus \cdots \oplus L_d)(\beta) \\ &= \inf \left\{ \sum_{j=i}^{d} \rho_j L_j(\beta_j) : \beta_j \in \mathbb{R}^d, \rho_j \ge 0, \sum_{j=i}^{d} \rho_j = 1, \sum_{j=i}^{d} \rho_j \beta_j = \beta \right\}. \end{aligned} \tag{2.6}$$

The local rate function, denoted by $L(x, \beta)$ for every $x \in \mathbb{R}^d_+$ and $\beta \in \mathbb{R}^d$, is defined as

$$L(x, \beta) = \bar{L}_{\Pi(x)}(\beta).$$

Note that the Legendre transform and inf-convolution of convex functions are still convex. Thus the local rate function $L(x, \cdot)$ is convex for every $x \in \mathbb{R}^d_+$.

2.3.2 Sample Path Large Deviations

Fix an arbitrary time $T > 0$. The sample paths $\{X^n(t) : t \in [0,T]\}$ live in the Polish space of cadlag functions $\mathcal{D}([0,T] : \mathbb{R}^d)$ endowed with the Skorohod metric. For each $x \in \mathbb{R}^d_+$, define the rate function $I_x : \mathcal{D}([0,T] : \mathbb{R}^d) \to [0,\infty]$ by

$$I_x(\phi) = \int_0^T L(\phi(t), \dot{\phi}(t))\, dt$$

if $\phi(0) = x$, $\phi(t) \in \mathbb{R}^d_+$ for all t, and ϕ is absolutely continuous, and set $I_x(\phi) = \infty$ otherwise. It was established in [29] that the rate function $\{I_x : x \in \mathbb{R}^d_+\}$ is an upper bound rate function and has compact level sets on compacts in the sense that the set

$$\cup_{x \in C}\{\phi : I_x(\phi) \leq M\}$$

is compact for every $M \geq 0$ and compact set $C \in \mathbb{R}^d_+$.

2.3.3 The Main Theorem

Recall that the large deviation principle and the Laplace principle are equivalent for probability measures on a Polish space [28, Theorem 1.2.1 and Theorem 1.2.3]. Let E_{x_n} denote the expectation conditional on $X^n(0) = x_n$.

Theorem 2.3.1. *The processes $\{X^n(t) : t \in [0,T]\}$ satisfy the uniform Laplace principle principle with rate functions $\{I_x : x \in \mathbb{R}^d_+\}$. That is, for any sequence $\{x_n\} \subseteq \mathbb{R}^d_+$ such that $x_n \to x$ and any bounded continuous function $h : \mathcal{D}([0,T] : \mathbb{R}^d) \to \mathbb{R}$, we have*

$$\lim_{n\to\infty} -\frac{1}{n} \log E_{x_n}\{\exp[-nh(X^n)]\} = \inf_{\phi \in \mathcal{D}([0,T]:\mathbb{R}^d_+)} \{I_x(\phi) + h(\phi)\}.$$

Therefore, $\{X^n(t) : t \in [0,T]\}$ with $X^n(0) = x \in \mathbb{R}^d_+$ satisfy the large deviation principle with rate function I_x.

Proof. Throughout the proof we will assume without loss of generality that $T = 1$. The uniform Laplace principle upper bound is implied by the uniform large deviation upper bound [29, Theorem 1.1] through an argument analogous to [28, Theorem 1.2.1]. Therefore, it suffices to show the uniform Laplace principle lower bound. That is to show

$$\liminf_n \frac{1}{n} \log E_{x_n}\{\exp[-nh(X^n)]\} \geq - \inf_{\phi \in \mathcal{D}([0,1]:\mathbb{R}^d)} \{I_x(\phi) + h(\phi)\}. \tag{2.7}$$

Since the above inequality holds trivially if $I_x(\phi) = \infty$, we can a priori assume that $I_x(\phi)$ is finite, which dictates that ϕ is absolutely continuous. For the convenience of the reader, we divide the proof into four steps. In Step 1, an alternative representation for the left hand side of (2.7) is established, which turns the analysis of the lower bound (2.7) into that of a stochastic control problem. The construction of nearly optimal controls is given in Step 2. The analysis of the limit controlled process is carried out in Step 3 via the weak convergence approach. The desired lower bound (2.7) is finally established in Step 4.

Step 1: Stochastic Control Representation

In order to prove (2.7), it suffices to show

$$\liminf_n \frac{1}{n} \log E_{x_n}\{\exp[-nh(X^n)]\} \geq -[I_x(\phi) + h(\phi)] \tag{2.8}$$

for every ϕ. Denote by P_n the probability measure induced by X^n on the Polish space $\mathcal{D}([0,1] : \mathbb{R}^d)$. Then by the relative entropy representation of exponential integrals [28, Section 1.4]

$$-\frac{1}{n} \log E_{x_n}\{\exp[-nh(X^n)]\} = \inf \left[\frac{1}{n} R(Q\|P_n) + \int_{\mathcal{D}([0,1]:\mathbb{R}^d)} h \, dQ \right]$$

where the infimum is taken over all probability measures Q on $\mathcal{D}([0,1]:\mathbb{R}^d)$. Now consider those probability measures induced by jump Markov processes $\bar{X}^n$ with initial condition $\bar{X}_n(0) = x_n$ and generator $\bar{\mathcal{L}}^n$ such that

$$\bar{\mathcal{L}}^n f(x,t) = n \sum_{v \in \mathbb{V}} \bar{r}(x,t;v)[f(x+v/n) - f(x)]. \tag{2.9}$$

Here $\bar{r}(x,t;v)$ is non-negative and uniformly bounded, and also satisfies $\bar{r}(x,t;v) = 0$ whenever $r(x;v) = 0$ [in other words, $\bar{r}(x,t;v)$ (which can be viewed as the *control*) is the new jump intensity from state x to $x+v$ at time t, and $\bar{X}^n$ (which can be viewed as the *controlled process*) is the scaled version of the jump Markov process with $\bar{r}(x,t;v)$, as the jump intensity] . If we restrict the infimum to such probability measures, for which the explicit evaluation of the relative entropy $R(\cdot\|P_n)$ is available [22, Theorem B.6], we arrive at the inequality

$$\begin{aligned} &-\frac{1}{n}\log E_{x_n}\{\exp[-nh(X^n)]\} \\ &\qquad \leq \inf_{\bar{r}} E_{x_n}\left[\int_0^1 \sum_{v\in\mathbb{V}} r(\bar{X}^n(t);v)\ell\left(\frac{\bar{r}(\bar{X}^n(t),t;v)}{r(\bar{X}^n(t);v)}\right)\,dt + h(\bar{X}^n)\right], \end{aligned}$$

where ℓ is defined by

$$\ell(x) = \begin{cases} x\log x - x + 1 & \text{if } x \geq 0, \\ \infty & \text{if } x < 0, \end{cases}$$

with the convention $0 \cdot \ell(0/0) = 0$. Therefore, in order to prove (2.8), it suffices to construct, for an arbitrarily fixed positive constant ε, an alternative jump intensity function $\bar{r}$ (dependent on ε) such that

$$\begin{aligned} \limsup_n E_{x_n}\left[\int_0^1 \sum_{v\in\mathbb{V}} r(\bar{X}^n(t);v)\ell\left(\frac{\bar{r}(\bar{X}^n(t),t;v)}{r(\bar{X}^n(t);v)}\right)\,dt + h(\bar{X}^n)\right] \\ \leq I(\phi) + h(\phi) + \varepsilon. \end{aligned} \tag{2.10}$$

Since proving (2.10) for all ϕ is not feasible, we restrict ϕ to a more analytically tractable class $\mathcal{N}$, which consists of those absolutely continuous functions $\phi^* : [0,1] \to \mathbb{R}_+^d$ such that there exists a positive integer K and a partition $0 = t_0 < t_1 < \cdots < t_{k-1} < t_K = 1$ where on each open interval (t_{i-1}, t_i), $i = 1, \ldots, K$, both $\dot{\phi}^*$ and $\Pi(\phi^*)$ take constant values. The following lemma states that any trajectory ϕ with finite cost can be approximated by a trajectory in class $\mathcal{N}$. The proof of this lemma is deferred to Appendix A.1.

Lemma 2.3.2. *Given any $\phi \in \mathcal{D}([0,1] : \mathbb{R}^d)$ such that $I_x(\phi) < \infty$ and any $\delta > 0$, there exists a $\phi^* \in \mathcal{N}$ such that $\|\phi - \phi^*\|_\infty < \delta$ and $I_x(\phi^*) \leq I_x(\phi)$.*

Due to this lemma and the continuity of h, it is easy to see that in order to show (2.7) one only needs to prove

$$\begin{aligned} \limsup_n E_{x_n} \left[\int_0^1 \sum_{v \in \mathbb{V}} r(\bar{X}^n(t); v) \ell \left(\frac{\bar{r}(\bar{X}^n(t), t; v)}{r(\bar{X}^n(t); v)} \right) dt + h(\bar{X}^n) \right] \\ \leq \quad I(\phi^*) + h(\phi^*) + \varepsilon. \end{aligned} \tag{2.11}$$

We now set forth to prove this inequality.

Step 2: Construction of Optimal Controls ($\bar{r}$)

The construction of $\bar{r}$ is based on the representation of the rate function $\bar{L}_i$ (which is the infimum of the associated *running cost*, and can be viewed as the *energy* of a particle) in terms of the function ℓ. More precisely, we have the following lemma, whose proof is very similar to [10, Section 4.3]. For the sake of completeness, we include the proof in Appendix A.2.

Lemma 2.3.3. *Given $\beta \in \mathbb{R}^d$ and $i = 0, 1, \ldots, d$, we have the representation*

$$\bar{L}_i(\beta) = \inf \left[\sum_{k=1}^{d} \rho_k \mu_k \ell \left(\frac{\bar{\mu}_k}{\mu_k} \right) + \sum_{k=1}^{d} \lambda_k \ell \left(\frac{\bar{\lambda}_k}{\lambda_k} \right) \right]$$

where the infimum is taken over strictly positive constants $\{\bar{\mu}_k, \bar{\lambda}_k : k \geq 1\}$ and strictly positive constants $\{\rho_k : k \geq i\}$ with $\rho_k = 0$ for $k < i$ such that

$$\sum_{k=i}^{d} \rho_k = 1, \quad -\sum_{k=1}^{d} \rho_k \bar{\mu}_k e_k + \sum_{k=1}^{d} \bar{\lambda}_k e_k = \beta.$$

Furthermore, $\bar{L}_i(\beta)$ is finite if and only if $\beta_k \geq 0$ for all $k < i$.

Since $\phi^* \in \mathcal{N}$, there exists a partition $0 = t_0 < t_1 < \cdots < t_{K-1} < t_K = 1$ such that on the open interval (t_j, t_{j+1}), both $\dot{\phi}^*(t)$ and $\Pi(\phi^*(t))$ take constant values, say $\dot{\phi}^*(t) = \beta_j$ and $\Pi(\phi^*(t)) = I_j$. Due to Lemma 2.3.3 we can define a collection $\{\rho_k^j, \bar{\mu}_k^j, \bar{\lambda}_k^j\}_{k \geq 0}$ (where the superscript denotes the time index) such that

1. For $k < I_j$, $\rho_k^j = 0$, $\bar{\mu}_k^j = \mu_k$, and $\bar{\lambda}_k^j = \lambda_k$; Note that the definitions of $\bar{\mu}_k^j$ and $\bar{\lambda}_k^j$ can be arbitrary since the limit process does not spend any meaningful amount of time on the interface $\{x \in \mathbb{R}_+^d : \Pi(x) = k\}$.

2. For $k \geq I_j$, ρ_k^j, $\bar{\mu}_k^j$ and $\bar{\lambda}_k^j$ are all strictly positive and satisfy

$$\sum_{k=I_j}^{d} \rho_k^j = 1, \quad -\sum_{k=1}^{d} \rho_k^j \bar{\mu}_k^j e_k + \sum_{k=1}^{d} \bar{\lambda}_k^j e_k = \beta_j, \tag{2.12}$$

$$\sum_{k=1}^{d} \rho_k^j \mu_k \ell \left(\frac{\bar{\mu}_k^j}{\mu_k} \right) + \sum_{k=1}^{d} \lambda_k \ell \left(\frac{\bar{\lambda}_k^j}{\lambda_k} \right) \leq \bar{L}_{I_j}(\beta_j) + \varepsilon \tag{2.13}$$

We can now define the alternative jump intensity $\bar{r}$ as follows. For every $t \in [t_j, t_{j+1})$,

let

$$\bar{r}(x,t;v) = \begin{cases} \bar{\lambda}_k^j & \text{if } v = e_k, \\ \bar{\mu}_k^j & \text{if } v = -e_k \text{ and } \Pi(x) = k \geq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{2.14}$$

The function $\bar{r}$ defines a jump process $\bar{X}^n$, given the initial condition $\bar{X}_n(0) = x_n$. We also introduce the notation

$$\beta^{j,0} = \sum_{k=1}^{d} \bar{\lambda}_k^j e_k; \quad \beta^{j,i} = -\bar{\mu}_i^j e_i + \sum_{k=1}^{d} \bar{\lambda}_k^j e_k, \ i = 1, \ldots, d. \tag{2.15}$$

where the second superscript represents the surface of discontinuity. It is trivial from definitions (2.14) and (2.15) that for every $t \in [t_j, t_{j+1})$,

$$\beta^{j,\Pi(x)} = \sum_{v \in \mathbb{V}} \bar{r}(x,t;v) \cdot v. \tag{2.16}$$

In other words, $\{\beta^{j,\Pi(x)}\}$ corresponds to the law of large number limit of the velocity of the process $\bar{X}^n$ at state x. It can be viewed as the *average* velocity.

Remark 2.3.4. The probability measures induced by $\bar{X}^n$ and X^n are absolutely continuous with respect to each other. This is because for any given jump size v, the corresponding jump intensities $\bar{r}(x,t;v)$ and $r(x;v)$ are either both zero or both strictly positive.

Step 3: Weak convergence analysis of the limit process

The goal of this step is to argue that $\{\bar{X}^n\}$ converges in distribution to ϕ^*. We first show that $\{\bar{X}^n\}$ is tight and thus has a subsequence converging in distribution, and then identify the weak limit to be ϕ^*. The proof of tightness is standard. It is in the identification of the weak limit, that the structure of the model, namely the

stability-about-the-interface condition, plays a crucial role; see Remark 2.3.6.

For each n, we define a collection of random measures $\gamma^n = (\gamma_0^n, \gamma_1^n, \ldots, \gamma_d^n)$ on $[0,1]$ where for every $k = 0, 1, \ldots, d$ and every Borel set $B \subset [0,1]$

$$\gamma_k^n(B) = \int_B 1_{\{\Pi(\bar{X}^n(t))=k\}}\, dt.$$

Each γ_k^n is a random variable taking values in the Polish space of sub-probability measures on the interval $[0,1]$, equipped with the topology of weak convergence.

Lemma 2.3.5. *Given any subsequence of $(\gamma^n, \bar{X}^n)$, there exists a subsubsequence and a collection of random measures $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_d)$ on $[0,1]$ such that*

1. *the subsubsequence converges in distribution to (γ, ϕ^*);*
2. *with probability one, γ_k is absolutely continuous with respect to the Lebesgue measure on $[0,1]$, and its density, say h_k, satisfies for almost every t*

$$h_k(t) = \sum_{j=0}^{K-1} \rho_k^j 1_{(t_j, t_{j+1})}(t). \tag{2.17}$$

Proof. To simplify notation, the subsequence is still denoted by $(\gamma^n, \bar{X}^n)$. We first argue that it is tight. The family of random measures $\{\gamma_k^n\}$ is contained in the set of all sub-probability measures on $[0,1]$. Since $[0,1]$ is compact, this set is compact as well. This proves the tightness of $\{\gamma^n\}$.

In order to show the tightness of $\{\bar{X}^n\}$, we introduce an auxiliary process S^n. Loosely speaking, it is the "average" of the process $\bar{X}^n$:

$$S^n(t) = x_n + \sum_{k=0}^{d} \left[\sum_{j=0}^{I(t)-1} \beta^{j,k} \gamma_k^n\{[t_j, t_{j+1})\} + \beta^{I(t),k} \gamma_k^n\{[t_{I(t)}, t)\} \right],$$

where $I(t) = \max\{j : t_j \leq t\}$. Since every random measure γ_k^n is absolutely continuous with respect to the Lebesgue measure on $[0,1]$ with the density or the Radon-Nikodým derivative uniformly bounded by one, $\{S^n\}$ is uniformly Lipschitz continuous. It follows that $\{S^n\}$ takes values in a compact subset of $\mathcal{C}([0,1] : \mathbb{R}^d)$ by the Arzélà-Ascoli Theorem, which in turn implies the tightness of $\{S^n\}$.

It suffices now to show that $\|\bar{X}^n - S^n\|_\infty$ converges to 0 in probability (and therefore $\{\bar{X}^n\}$ is tight). To this end, we introduce the process

$$Z^n(t) = n\bar{X}^n\left(\frac{t}{n}\right), \qquad 0 \leq t \leq n.$$

Note that $\bar{X}^n$ is a scaled version of Z^n. Since the generator of $\bar{X}^n$ takes the form (2.9), it is clear that the generator of Z^n, denoted by $\mathcal{L}^n$, is such that

$$\mathcal{L}^n f(z,t) = \sum_{v \in \mathbb{V}} \bar{r}\left(\frac{z}{n}, \frac{t}{n}; v\right)[f(z+v) - f(z)].$$

In other words, Z^n is a pure jump Markov process whose jump intensity (for a jump of size v) at state $Z^n = z$ and time t is

$$\lambda_n(z,t;v) = \bar{r}\left(\frac{z}{n}, \frac{t}{n}; v\right). \tag{2.18}$$

For every $v \in \mathbb{V}$, denote by $Y^{n,v}$ the counting process for jumps of size v associated with the process Z^n. That is,

$$Y^{n,v}(t) = \text{Number of jumps of size } v \text{ up until time } t \text{ for the process } Z^n.$$

It is clear that for every $t \in [0,1]$

$$Z^n(t) = Z^n(0) + \sum_{v \in \mathbb{V}} Y^{n,v}(t) \cdot v = nx_n + \sum_{v \in \mathbb{V}} Y^{n,v}(t) \cdot v, \tag{2.19}$$

and the instantaneous intensity function for $Y^{n,v}$ is $\lambda_n(Z^n(t), t; v)$; see also [22, Appendix B] for a more detailed discussion on counting processes.

We can now rewrite S^n in terms of the intensity function λ_n. Recalling the definitions of S^n and $\{\gamma_k^n\}$, and that $I(s) = j$ if $s \in [t_j, t_{j+1})$ and $I(s) = I(t)$ if $s \in [t_{I(t)}, t)$, we have

$$\begin{aligned} S^n(t) &= x_n + \sum_{k=0}^{d} \left[\sum_{j=0}^{I(t)-1} \beta^{j,k} \int_{t_j}^{t_{j+1}} 1_{\{\Pi(\bar{X}^n(s))=k\}}\, ds \right. \\ &\qquad \left. + \beta^{I(t),k} \int_{t_{I(t)}}^{t} 1_{\{\Pi(\bar{X}^n(s))=k\}}\, ds \right] \\ &= x_n + \sum_{k=0}^{d} \int_0^t \beta^{I(s),k} \cdot 1_{\{\Pi(\bar{X}^n(s))=k\}}\, ds \\ &= x_n + \int_0^t \beta^{I(s),\Pi(\bar{X}^n(s))}\, ds. \end{aligned}$$

Thanks to (2.16) and (2.18), it follows that

$$\begin{aligned} S^n(t) &= x_n + \int_0^t \sum_{v \in \mathbb{V}} \bar{r}(\bar{X}^n(s), s; v) \cdot v\, ds \\ &= x_n + \int_0^t \sum_{v \in \mathbb{V}} \lambda_n(Z^n(ns), ns; v) \cdot v\, ds \\ &= x_n + \frac{1}{n} \int_0^{nt} \sum_{v \in \mathbb{V}} \lambda_n(Z^n(s), s; v) \cdot v\, ds. \end{aligned}$$

Combined with equation (2.19), we have

$$\begin{aligned}\bar{X}^n(t) - S^n(t) &= \frac{1}{n}Z^n(nt) - S^n(t) \\ &= \frac{1}{n}\sum_{v\in\mathbb{V}}\left[Y^{n,v}(nt) - \int_0^{nt}\lambda_n(Z^n(s), s; v)\,ds\right]\cdot v.\end{aligned}$$

It is now clear that $\bar{X}^n - S^n$ is a martingale since λ_n is the intensity of $Y^{n,v}$ [13, Lemma 2.3.2]. Therefore, it follows from Doob's maximal inequality that for every fixed $\varepsilon > 0$

$$\mathbb{P}_{x_n}\left(\sup_{t\in[0,1]}\|\bar{X}^n(t) - S^n(t)\| > \varepsilon\right) \le \frac{1}{\varepsilon^2}E_{x_n}\|\bar{X}^n(1) - S^n(1)\|^2.$$

Thanks to (2.18) and the definition (2.14) of $\bar{r}$, λ_n is uniformly bounded by $\|r\|_\infty$. Therefore, for some constant C [13, Theorem 2.5.3]

$$\begin{aligned}E_{x_n}\|\bar{X}^n(1) - S^n(1)\|^2 &\le \frac{C}{n^2}\sum_{v\in\mathbb{V}}E_{x_n}\left[Y^{n,v}(n) - \int_0^n\lambda_n(Z^n(s), s; v)\,ds\right]^2 \\ &= \frac{C}{n^2}\sum_{v\in\mathbb{V}}E_{x_n}\int_0^n\lambda_n(Z^n(s), s; v)\,ds \\ &\le \frac{C\cdot 2d\|\bar{r}\|_\infty}{n}.\end{aligned}$$

The right hand side of the above inequality converges to 0 as n tends to infinity. Therefore $\|\bar{X}_n - S_n\|_\infty$ converges to 0 in probability and $\{\bar{X}_n\}$ is tight.

By Prohorov's Theorem [12, Chapter 3] there exists a subsubsequence, still denoted by $(\gamma^n, \bar{X}^n)$, that converges in distribution to say $(\gamma, \bar{X})$ where $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_d)$. Note that $\bar{X}$ is continuous since it also is the weak limit of S^n. By the Skorohod Representation Theorem [28, Theorem A.3.9], we can assume that the convergence is almost sure convergence when everything is defined on some probability space, say $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{\mathbb{P}})$. Again since $\{\gamma_k^n\}$ is absolutely continuous with respect to the Lebesgue

measure on $[0,1]$ with the density uniformly bounded by one, the limit γ_k also enjoys the same property. Furthermore, it follows that for every t, $S^n(t)$ converges almost surely to

$$S(t) = x + \sum_{k=0}^{d} \left[\sum_{j=0}^{I(t)-1} \beta^{j,k} \gamma_k\{[t_j, t_{j+1})\} + \beta^{I(t),k} \gamma_k\{[t_{I(t)}, t)\} \right].$$

Therefore, $S(t) = \bar{X}(t)$ almost surely for every t. Since both S and $\bar{X}$ are continuous, $S = \bar{X}$ with probability one. In particular, if we denote by h_k the density of γ_k,

$$\frac{d\bar{X}(t)}{dt} = \sum_{k=0}^{d} \beta^{I(t),k} h_k(t) \tag{2.20}$$

for almost every t.

It remains to show (2.17) and that $\bar{X} = \phi^*$. In doing so, we first establish a useful property of $\{h_k\}$, namely, that with probability one

$$\sum_{k=0}^{d} h_k(t) = 1 = \sum_{k=\Pi(\bar{X}(t))}^{d} h_k(t) \tag{2.21}$$

for almost every $t \in [0,1]$. The first equality is trivial since $\sum_{k=0}^{d} \gamma_k^n$ equals the Lebesgue measure on $[0,1]$ for every n. The second equality follows from a standard argument [28, Theorem 7.4.4(c)]. Note that for almost every $\omega \in \Omega$, $\bar{X}(t,\omega)$ is continuous with respect to t, $\gamma^n(\omega)$ converges weakly to $\gamma(\omega)$, and $\bar{X}^n(\cdot,\omega)$ converges to $\bar{X}(\cdot,\omega)$ in the Skorohod metric. Fix arbitrarily such an ω. Define $A_i = \{t \in [0,1] : \Pi(\bar{X}(t,\omega)) = i\}$. Since $\bar{X}^n(\cdot,\omega)$ also converges to $\bar{X}(\cdot,\omega)$ in the sup-norm [22, Theorem A.6.5] and Π is lower-semicontinuous, it follows that for every $t \in A_i$ there exists an open interval (a_t, b_t) containing t and an $N \in \mathbb{N}$ such that $\Pi(\bar{X}^n(s,\omega)) \geq i$ for all $s \in (a_t, b_t)$ and $n \geq N$. Therefore, $\sum_{k<i} \gamma_k^n(\omega)\{(a_t, b_t)\} = 0$ for all $n \geq N$. Letting $n \to \infty$ it follows that $\sum_{k<i} \gamma_k(\omega)\{(a_t, b_t)\} = 0$ for every $t \in A_i$. Since

$A_i \subseteq \cup_{t\in A_i}(a_t, b_t)$, there exists a countable subcover [19, Page 49, Lindelöf Theorem] that is, there exists $\{t_j\} \subseteq A_i$ such that

$$A_i \subseteq \cup_j (a_{t_j}, b_{t_j}).$$

It follows from the countable sub-additivity of measures that $\sum_{k<i} \gamma_k(\omega)\{A_i\} = 0$. Therefore,

$$0 = \sum_{i=0}^{d} \sum_{k<i} \gamma_k(\omega)\{A_i\} = \int_0^1 \sum_{k=0}^{\Pi(\bar{X}(t))-1} h_k(t,\omega)\, dt.$$

This completes the proof (2.21). Combining (2.21), (2.20), and (2.15) we obtain the following identity

$$\frac{d\bar{X}(t)}{dt} = \sum_{k=0}^{d} \beta^{j,k} h_k(t) = \sum_{k=1}^{d} \bar{\lambda}_k^j e_k - \sum_{k=\max\{\Pi(\bar{X}(t)),1\}}^{d} \bar{\mu}_k^j h_k(t) e_k, \tag{2.22}$$

for almost every $t \in (t_j, t_{j+1})$.

We will now use induction to argue (2.17). It is trivial that (2.17) holds for almost every $t \in [0, t_j]$ with $j = 0$ since $t_0 = 0$. Assume that (2.17) holds for almost every $t \in [0, t_j]$. The goal is to show that it holds for almost every $t \in [0, t_{j+1}]$, or equivalently, $h_k(t) = \rho_k^j$ for almost every $t \in (t_j, t_{j+1})$.

It is not difficult to verify that $\bar{X}(t) = \phi^*(t)$ for all $t \in [0, t_j]$. Indeed, by the induction hypothesis that (2.17) holds for almost every $t \in [0, t_j]$, and equations

(2.12), (2.15), we have

$$\begin{aligned}\frac{d\bar{X}(t)}{dt} &= \sum_{k=0}^{d} \beta^{I(t),k} \sum_{j=0}^{K-1} \rho_k^j 1_{(t_j,t_{j+1})}(t) \\ &= \sum_{j=0}^{K-1} \sum_{k=0}^{d} \rho_k^j \beta^{j,k} 1_{(t_j,t_{j+1})}(t) \\ &= \sum_{j=0}^{K-1} \beta_j 1_{(t_j,t_{j+1})}(t) \\ &= \frac{d\phi^*(t)}{dt}. \end{aligned} \tag{2.23}$$

Therefore, since $\bar{X}(0) = x = \phi^*(0)$, $\bar{X}(t) = \phi^*(t)$ for every $t \in [0, t_j]$. In particular, $\bar{X}(t_j) = \phi^*(t_j)$.

Define $I_j = \Pi(\phi^*(t))$ and $\beta_j = \dot{\phi}^*(t)$ for every $t \in (t_j, t_{j+1})$. Observing that $[\beta_j]_k = 0$ for all $I_j < k \leq d$, one can uniquely determine the value of $\{\rho_k^j\}$ based on the definition of $\{\rho_k^j\}$ and equation (2.12), namely,

$$\rho_k^j = \begin{cases} \bar{\lambda}_k^j / \bar{\mu}_k^j & \text{if } I_j < k \leq d, \\ 1 - \sum_{k=I_j+1}^{d} \bar{\lambda}_k^j / \bar{\mu}_k^j & \text{if } k = I_j, \\ 0 & \text{if } k < I_j. \end{cases} \tag{2.24}$$

We also note that the lower semicontinuity of Π implies $I_j \geq \Pi(\phi^*(t_j)) = \Pi(\bar{X}(t_j))$.

The key step in this inductive argument is to prove that $\Pi(\bar{X}(t)) = I_j$ for every $t \in (t_j, t_{j+1})$. To this end, note that $\Pi(\bar{X}(t))$ can only take finitely many possible values, hence the maximum of $\Pi(\bar{X}(t))$ on (t_j, t_{j+1}) must be attained at some $t^* \in (t_j, t_{j+1})$. Since Π is lower semicontinuous, there exists an open interval that is contained in (t_j, t_{j+1}), such that for all t in this interval, $\Pi(\bar{X}(t)) \geq \Pi(\bar{X}(t^*))$. Denote by $(a, b) \subseteq (t_j, t_{j+1})$ the largest of such intervals. By the definition of t^*,

$\Pi(\bar{X}(t)) = \Pi(\bar{X}(t^*)) = i$ (say) for every $t \in (a, b)$. It follows from (2.22) that on the interval (a, b),

$$\frac{d\bar{X}(t)}{dt} = \sum_{k=1}^{d} \bar{\lambda}_k^j e_k - \sum_{k=\max\{i,1\}}^{d} \bar{\mu}_k^j h_k(t) e_k. \tag{2.25}$$

Furthermore, since clearly $[d\bar{X}(t)/dt]_k = 0$ for all $k > i$ and $t \in (a, b)$, one can directly compute h_k from (2.25) and (2.21) to obtain a formula analogous to (2.24):

$$h_k(t) = \begin{cases} \bar{\lambda}_k^j/\bar{\mu}_k^j & \text{if } i < k \le d, \\ 1 - \sum_{k=i+1}^{d} \bar{\lambda}_k^j/\bar{\mu}_k^j & \text{if } k = i, \\ 0 & \text{if } k < i, \end{cases} \tag{2.26}$$

for almost every $t \in (a, b)$.

We will argue by contradiction that $i \le I_j$. Assume otherwise, namely, $i > I_j$. Then by comparing (2.24) and (2.26) it follows easily that $h_i(t) > \rho_i^j$ and thus

$$\left[\frac{d\bar{X}(t)}{dt}\right]_i = \bar{\lambda}_i^j - \bar{\mu}_i^j h_i(t) < \bar{\lambda}_i^j - \bar{\mu}_i^j \rho_i^j = 0.$$

This implies that $[\bar{X}(a)]_i > [\bar{X}(t^*)]_i > 0$, or $\Pi(\bar{X}(a)) \ge i > I_j$. Recall that $I_j \ge \Pi(\bar{X}(t_j))$. Therefore, $a \ne t_j$ and thus we must have that $a > t_j$. By the lower semicontinuity of Π there exists a small $\eta > 0$ such that $a - \eta > t_j$ and $\Pi(\bar{X}(t)) \ge i = \Pi(\bar{X}(t^*))$ for every $t \in (a-\eta, a]$. Therefore, $(a-\eta, b) \subseteq (t_j, t_{j+1})$ is an interval on which $\Pi(\bar{X}(t)) \ge \Pi(\bar{X}(t^*))$. This contradicts the maximality of the interval (a, b). Therefore $i \le I_j$ and hence

$$\Pi(\bar{X}(t)) \le I_j, \quad \text{for all } t \in (t_j, t_{j+1}). \tag{2.27}$$

In order to show the reverse inequality, we exclude the trivial case by assuming

$I_j \geq 1$. Note that (2.27) implies $[d\bar{X}(t)/dt]_k = 0$ for all $k > I_j$. Thanks to (2.22), this is equivalent to $h_k(t) = \bar{\lambda}_k^j/\bar{\mu}_k^j = \rho_k^j$ for all $k > I_j$. It follows that

$$h_{I_j}(t) \leq 1 - \sum_{k=I_j+1}^{d} h_k(t) = 1 - \sum_{k=I_j+1}^{d} \rho_k^j = \rho_{I_j}^j. \tag{2.28}$$

which in turn implies that for every $t \in (t_j, t_{j+1})$

$$\frac{d[\bar{X}(t) - \phi^*(t)]_{I_j}}{dt} = [\bar{\lambda}_{I_j}^j - \bar{\mu}_{I_j}^j h_{I_j}(t)] - [\bar{\lambda}_{I_j}^j - \bar{\mu}_{I_j}^j \rho_{I_j}^j(t)] \geq 0.$$

Since $\bar{X}(t_j) = \phi^*(t_j)$, we have $[\bar{X}(t)]_{I_j} \geq [\phi^*(t)]_{I_j} > 0$, or $\Pi(\bar{X}(t)) \geq I_j$ for all $t \in (t_j, t_{j+1})$. Therefore, taking (2.27) into consideration we arrive at

$$\Pi(\bar{X}(t)) = I_j = \Pi(\phi^*(t))$$

on the interval (t_j, t_{j+1}).

The desired equality $h_k(t) = \rho_k^j$ for every $t \in (t_i, t_{j+1})$ is now trivial. Indeed, the two formulas (2.24) and (2.26) are identical when $i = \Pi(\bar{X}(t)) = I_j$. This completes the proof of (2.17).

It remains to show that $\bar{X}(t) = \phi^*(t)$ for all $t \in [0, 1]$. This can be done by repeating the steps in (2.23) for every $t \in (0, 1)$. The proof of Lemma 2.3.5 is now complete.

Step 4: Analysis of the cost

Along the convergent subsubsequence (still denoted by $(\gamma^n, \bar{X}^n)$), Lemma 2.3.5 and (2.13) imply that

$$\begin{aligned}
&\lim_n E_{x_n}\left[\int_0^1 \sum_{v\in\mathbb{V}} r(\bar{X}^n(t);v)\ell\left(\frac{\bar{r}(\bar{X}^n(t),t;v)}{r(\bar{X}^n(t);v)}\right)\,dt + h(\bar{X}^n)\right] \\
&= \lim_n E_{x_n}\sum_{j=0}^{K-1}\left[\int_{t_j}^{t_{j+1}}\sum_{k=1}^{d}\lambda_k\ell\left(\frac{\bar{\lambda}_k^j}{\lambda_k}\right)dt + \sum_{k=1}^{d}\mu_k\ell\left(\frac{\bar{\mu}_k^j}{\mu_k}\right)\gamma_k^n(dt)\right] + h(\phi^*) \\
&= \sum_{j=0}^{K-1}\left[\int_{t_j}^{t_{j+1}}\sum_{k=1}^{d}\lambda_k\ell\left(\frac{\bar{\lambda}_k^j}{\lambda_k}\right)dt + \sum_{k=1}^{d}\rho_k^j\mu_k\ell\left(\frac{\bar{\mu}_k^j}{\mu_k}\right)dt\right] + h(\phi^*) \\
&\le \sum_{j=0}^{K-1}\int_{t_j}^{t_{j+1}}[\bar{L}_{I_j}(\beta_j)+\varepsilon]dt + h(\phi^*) \\
&= \int_0^1 L(\phi^*(t),\dot{\phi}^*(t))dt + \varepsilon + h(\phi^*).
\end{aligned}$$

This completes the proof of (2.11) as well as the proof of Theorem 2.3.1. ■

Remark 2.3.6. The *stability-about-the-interface condition* manifests itself in the monotonicity of the value $h_k(t)$ with respect to the value of $\Pi(\bar{X}(t))$; see (2.26). Loosely speaking, this *monotonicity* property implies that the change of measure (control) defined by the upper bound rate function automatically pushes the trajectory back to the discontinuous interface if it ever wanders off. This guarantees the desired tracking behavior.

2.4 A Case Study: The 2-dimensional Network

In this section we illustrate in the context of an example how to explicitly identify the exponential decay rate of a rare event of interest. Consider the case where $d = 2$

in the original model. The probability of interest is

$$\begin{aligned} p_n &= \mathbb{P}\{\text{total population } Q_1 + Q_2 \text{ reaches } n \text{ before coming back to } 0, \\ &\qquad \text{starting from } Q = (0,0)\}. \end{aligned}$$

Under the assumption that the stability condition holds, that is,

$$\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} < 1, \tag{2.29}$$

the total population overflow is a rare event when n is large.

The exponential decay rate of p_n can be explicitly identified in terms of the appropriate roots of the Hamiltonians H_1 and H_2. Note that H_2 is the Hamiltonian in the interior of the state space, whereas H_1 is the Hamiltonian on the boundary $\partial = \{x : \Pi(x) = 1\} = \{x = (x_1, x_2) : x_2 = 0, x_1 > 0\}$. For this reason, we simplify the notation and denote

$$H = H_2, \quad H_\partial = H_1.$$

Sometimes H and H_∂ are referred to as the interior and the boundary Hamiltonians, respectively. Similarly, the rate functions L_2 and L_1 will be replaced by L and L_∂, correspondingly. We will proceed heuristically for now to show the form of the decay rate of p_n, which is closely connected to the geometry of the zero-level sets of H and H_∂.

2.4.1 Three Important Roots of the Hamiltonians

The quantity of interest p_n is just the probability of the scaled process X^n reaching the exit boundary $\partial_e = \{x = (x_1, x_2) : x_i \geq 0, x_1 + x_2 = 1\}$ before coming back to

the origin, starting from the origin itself. Thanks to Theorem 3.4.1, it is reasonable to expect that the exponential decay rate of p_n equals the value of the calculus of variations problem

$$\inf \int_0^\tau L(\phi(t), \dot{\phi}(t))\, dt,$$

where the infimum is taken over all absolutely continuous functions $\phi : [0, \infty) \to \mathbb{R}_+^2$ and $\tau \geq 0$ such that $\phi(0) = 0$ and $\phi(\tau) \in \partial_e$. It is not difficult to see that an optimal trajectory ϕ^*, if it exists, should be a straight line due to the convexity of the local rate function and the homogeneity of the system dynamics. See Figure 2.3.

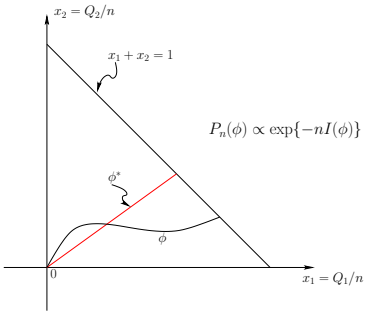

Figure 2.3: Representative limit sample path ϕ

In order to solve the aforementioned calculus of variations problem, we recast it into a control problem. To this end, we slightly expand this variational problem to a general initial condition $\phi(0) = x$ and denote the corresponding infimum by $V(x)$. Notice that the exponential decay rate of p_n is in fact $V(0)$. Recall that the optimal trajectory ϕ^* is a straight line, which either travels through the interior of the state space or along the boundary ∂. The value function V is different in each of these two cases. We will discuss them separately.

If the optimal trajectory travels through the interior of the state space, then the dynamic programming principle implies that the value function V satisfies the

Hamiltonian-Jacobi-Bellman (HJB) equation

$$0 = \inf_{\beta}[L(\beta) + \langle \nabla V(x), \beta \rangle] = -H(-\nabla V(x)).$$

Furthermore, the boundary condition $V(x) = 0$ for $x \in \partial_e$ should hold. This suggests that $\nabla V(x) = -\alpha^*$ where $H(\alpha^*) = 0$ and that α^* is orthogonal to ∂_e, or equivalently $\alpha_1^* = \alpha_2^*$. In this case, the exponential decay rate of p_n is just α_1^*, and the optimal trajectory leaves the domain in a straight line with slope $\beta^* = \nabla H(\alpha^*)$ (β^* is the minimizer in the HJB equation). We wish to make an important cautionary comment, namely, that the geometry of the zero-level set of H has to be taken into consideration in order for these heuristics to determine a possible optimal trajectory. For illustration, consider the following two scenarios (see Figure 2.4). In both cases, $\bar{\alpha}$ denotes the point on the level set $\{H = 0\}$ with the maximal first component, whence $\nabla H(\bar{\alpha}) = ae_1$ for some non-negative constant a. In case (a) the 45° line intersects with the level set at point α^* which is above $\bar{\alpha}$. The corresponding $\beta^* = \nabla H(\alpha^*)$ has non-negative components. Therefore, the root α^* determines a candidate optimal trajectory $\phi^*(t) = \beta^* t$ that lives in the non-negative orthant and hits ∂_e in finite time. In contrast, in case (b) the 45° line intersects with the level set at point α^* which is below $\bar{\alpha}$ and $\beta^* = \nabla H(\alpha^*)$ has a negative second component. It is clear that this root α^* does *not* associate with any physically meaningful trajectory since $\phi^*(t) = \beta^* t$ will not live in the non-negative orthant.

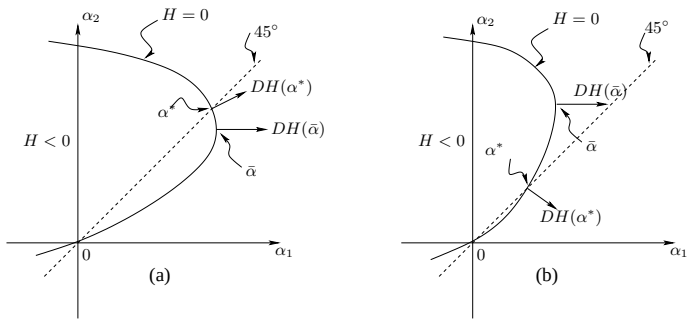

Figure 2.4: Geometry and trajectory (I)

In the case where the optimal trajectory travels along the boundary ∂, it may represent two different types of prelimit behavior: (1) the trajectory really "pushes into" the boundary if it is the limit of the prelimit sample paths that constantly switch residence between the interior and the boundary ∂; (2) the trajectory barely "touches" or "glides" along the boundary ∂ if it is the limit of those prelimit sample paths that live very close to the boundary ∂. The way to determine the trajectory also differs. For case (1), it is expected that along the boundary ∂ both the interior and the boundary HJB equations will be satisfied. That is,

$$-H(-\nabla V(x)) = 0, \quad -H_\partial(-\nabla V(x)) = 0.$$

This suggests that $-\nabla V(x) = \hat{\alpha}$ where $H(\hat{\alpha}) = H_\partial(\hat{\alpha}) = 0$. The exponential decay rate of p_n is therefore $\langle \hat{\alpha}, e_1 \rangle = \hat{\alpha}_1$. The corresponding trajectory is $\phi^*(t) = \beta^* t$ where

$$\beta^* = (\beta_1^*, 0) = \rho_1 \nabla H(\hat{\alpha}) + \rho_2 \nabla H_\partial(\hat{\alpha})$$

for some non-negative constants ρ_1, ρ_2 such that $\rho_1 + \rho_2 = 1$. The physical meaning of this identity is fairly clear: ρ_1 and ρ_2 are respectively the limit fraction of time that the prelimit sample paths spend in the interior and on the boundary ∂, whereas $\nabla H(\hat{\alpha})$ and $\nabla H_\partial(\hat{\alpha})$ are respectively the limit velocity of the prelimit sample paths in the interior and on the boundary ∂. For case (2), when the limit optimal trajectory glides along the boundary ∂, we expect that only the interior HJB equation $-H(-\nabla V(x)) = 0$ will be satisfied. Hence $\nabla V = -\hat{\alpha}$ where $H(\hat{\alpha}) = 0$ and the corresponding $\beta^* = \nabla H(\hat{\alpha})$ is a horizontal vector. The exponential decay rate is thus $\langle \hat{\alpha}, e_1 \rangle = \hat{\alpha}_1$ and the corresponding trajectory is $\phi^*(t) = \beta^* t$.

Again, when this heuristic is used to determine a possible optimal trajectory, the geometry of the zero-level sets of H and H_∂ has to be incorporated. For illustration,

consider the following two scenarios (see Figure 2.5). As before, $\bar{\alpha}$ denotes the point on the level set $\{H = 0\}$ with the maximal first component. In case (a), the intersection of the two zero-level sets, $\hat{\alpha}$, is below $\bar{\alpha}$. The corresponding β^* does determine a possible optimal trajectory $\phi^*(t) = \beta^* t$, which "pushes into" the boundary ∂. In case (b), however, $\hat{\alpha}$ is above $\bar{\alpha}$. In this case, since both $\nabla H(\hat{\alpha})$ and $\nabla H_\partial(\hat{\alpha})$ have positive second components, none of their convex combinations will yield a horizontal velocity β^*. Therefore, this root $\hat{\alpha}$ does *not* represent any meaningful trajectory traveling along the boundary ∂. Indeed, the root that will determine such a trajectory is $\bar{\alpha}$. It corresponds to a trajectory that "glides" along the boundary ∂ with velocity $\beta^* = \nabla H(\bar{\alpha})$, a horizontal vector.

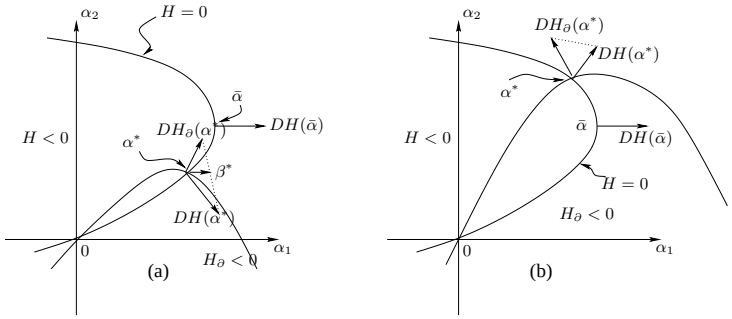

Figure 2.5: Geometry and trajectory (II)

It is now clear that the roots α^*, $\bar{\alpha}$, $\hat{\alpha}$ are crucial in the identification of the exponential decay rate of p_n. They can be explicitly calculated and we summarize the result in the following lemma. Its proof is straightforward but tedious, and thus omitted. To ease notation, from now on we let

$$\theta_1 = \frac{\lambda_1}{\mu_2},\ \theta_2 = \frac{\lambda_2}{\mu_2},\ \theta_3 = \frac{\mu_1}{\mu_2}, \tag{2.30}$$

and define the constant

$$z = \frac{(\theta_1 + \theta_2 + \theta_3 - 1) + \sqrt{(\theta_1 + \theta_2 + \theta_3 - 1)^2 + 4\theta_1(1 - \theta_3)}}{2\theta_3}. \tag{2.31}$$

Lemma 2.4.1. *The constant z satisfies $\max\{0, 1 - 1/\theta_3\} < z < 1$. Define vectors α^*, $\bar{\alpha}$, and $\hat{\alpha}$ to be*

$$\begin{aligned} \alpha^* &= -\log[\theta_1 + \theta_2] \cdot (1, 1), \\ \hat{\alpha} &= (-\log z, -\log[1 - \theta_3 + \theta_3 z]), \\ \bar{\alpha} &= \left(\log\left[1 + (1 - \sqrt{\theta_2})^2/\theta_1\right], -\log\sqrt{\theta_2}\right). \end{aligned}$$

Then $H(\alpha^*) = 0$, $H(\hat{\alpha}) = H_\partial(\hat{\alpha}) = 0$, and $H(\bar{\alpha}) = 0 = \langle \nabla H(\bar{\alpha}), e_2 \rangle$. Furthermore, for any α such that $H(\alpha) = 0$, the inequality $\alpha_1 \leq \bar{\alpha}_1$ holds, with equality if and only if $\alpha = \bar{\alpha}$.

2.4.2 The Exponential Decay Rate of p_n

It is now intuitively clear what the exponential decay rate of p_n should be. For example, if $\alpha_2^* > \bar{\alpha}_2$ and $\hat{\alpha}_2 < \bar{\alpha}_2$, then it corresponds to case (a) in Figure 2.4 and case (a) in Figure 2.5. Therefore, α^* determines a trajectory leaving the domain through the interior with cost α_1^*, whereas $\hat{\alpha}$ determines a trajectory leaving the domain by "pushing into" the boundary ∂ with cost $\hat{\alpha}_1$. The optimal trajectory should be the one with a smaller cost and the minimal cost is $\min(\alpha_1^*, \hat{\alpha}_1^*)$. More generally, we have Theorem 2.4.3, which can be shown by constructing suitable subsolutions and invoking Lemma 2.4.2 below.

Lemma 2.4.2. *Suppose that $W : \mathbb{R}_+^2 \to \mathbb{R}$ is a twice continuously differentiable function satisfying*

$$\begin{aligned} -H(-\nabla W(x)) &\geq 0, \text{ for } x = (x_1, x_2) \in \mathbb{R}_+^2 \text{ such that } x_2 > 0 \\ -H_\partial(-\nabla W(x)) &\geq 0, \text{ for } x \in \partial \\ W(x) &\leq 0, \text{ for } x \in \partial_e \end{aligned}$$

Then

$$\liminf_n -\frac{1}{n}\log p_n \geq W(0).$$

The function W is called a *classical subsolution* to the related partial differential equation. This lemma can be shown by a verification argument and its proof is deferred to Appendix A.3.

Theorem 2.4.3. *The exponential decay rate of p_n is*

$$\lim_n -\frac{1}{n}\log p_n = \begin{cases} \min(\alpha_1^*, \hat{\alpha}_1) & \text{if } \alpha_2^* > \bar{\alpha}_2,\ \hat{\alpha}_2 < \bar{\alpha}_2 \\ \alpha_1^* & \text{if } \alpha_2^* > \bar{\alpha}_2,\ \hat{\alpha}_2 \geq \bar{\alpha}_2 \\ \hat{\alpha}_1 & \text{if } \alpha_2^* \leq \bar{\alpha}_2,\ \hat{\alpha}_2 < \bar{\alpha}_2 \\ \bar{\alpha}_1 & \text{if } \alpha_2^* \leq \bar{\alpha}_2,\ \hat{\alpha}_2 \geq \bar{\alpha}_2 \end{cases}.$$

Proof. We only give the details for the case where $\alpha_1^* > \bar{\alpha}_2$ and $\hat{\alpha}_2 < \bar{\alpha}_2$. The proof for other cases is similar and thus omitted. Let $\gamma = \min(\alpha_1^*, \hat{\alpha}_1)$. We first show the upper bound

$$\liminf_n -\frac{1}{n}\log p_n \geq \gamma. \tag{2.32}$$

Thanks to Lemma 2.4.2, it suffices to construct a sequence of subsolutions whose values at the origin approach γ. To this end, we define two vectors

$$v^* = \frac{\gamma}{\alpha_1^*}\alpha^* = \gamma \cdot (1,1), \quad \hat{v} = \frac{\gamma}{\hat{\alpha}_1}\hat{\alpha} = \gamma \cdot \left(1, \frac{\hat{\alpha}_2}{\hat{\alpha}_1}\right).$$

Since $H(\alpha^*) = H(\hat{\alpha}) = H(0) = 0$ and $H_\partial(\hat{\alpha}) = H_\partial(0) = 0$, it follows from the convexity of H and H_∂ that

$$H(v^*) \leq 0, \quad H(\hat{v}) \leq 0, \quad H_\partial(\hat{v}) \leq 0.$$

We claim that $v_2^* > \hat{v}_2$. Indeed, letting $\nu = (0, \bar{\alpha}_2)$ where by Lemma 2.4.1 $\bar{\alpha}_2 = -\log\sqrt{\theta_2}$, straightforward calculation yields

$$H(\nu) = \lambda_2(1/\sqrt{\theta_2} - 1) + \mu_2(\sqrt{\theta_2} - 1) = -(\sqrt{\lambda_2} - \sqrt{\mu_2})^2 < 0.$$

Therefore, it follows from the strict convexity of H and $H(\bar{\alpha}) = 0$ that $H(s\bar{\alpha} + (1 - s)\nu) \leq 0$ for all $s \in [0, 1]$. This in turn implies that

$$\frac{\hat{\alpha}_2}{\hat{\alpha}_1} < \frac{\bar{\alpha}_2}{\bar{\alpha}_1},$$

since otherwise $s^* = \bar{\alpha}_2/\bar{\alpha}_1 \cdot \hat{\alpha}_1/\hat{\alpha}_2 \in [0, 1]$, and the convexity of H implies that [note that $\hat{\alpha}_2 < \bar{\alpha}_2$ by assumption]

$$H(\hat{\alpha}) < \frac{\hat{\alpha}_2}{\bar{\alpha}_2} H\left(s^*\bar{\alpha} + (1 - s^*)\nu\right) + \left(1 - \frac{\hat{\alpha}_2}{\bar{\alpha}_2}\right) H(0) \leq 0.$$

The above inequality is impossible since $H(\hat{\alpha}) = 0$. Observing that $\bar{\alpha}_2 < \alpha_1^*$ by assumption, and $\bar{\alpha}_1 > \alpha_1^*$ by Lemma 2.4.1, we have

$$\hat{v}_2 = \frac{\hat{\alpha}_2}{\hat{\alpha}_1}\gamma < \frac{\bar{\alpha}_2}{\bar{\alpha}_1}\gamma < \frac{\alpha_1^*}{\alpha_1^*}\gamma = \gamma = v_2^*.$$

Now fix an arbitrary small positive number δ and define a piecewise affine function on $x \in \mathbb{R}^2$ by

$$W^\delta(x) = \min\{\langle -v^*, x\rangle,\ \langle -\hat{v}, x\rangle - \delta\} = \begin{cases} \langle -v^*, x\rangle & \text{if } x_2 > b\delta, \\ \langle -\hat{v}, x\rangle - \delta & \text{otherwise,} \end{cases}$$

where $b = (v_2^* - \hat{v}_2)^{-1} > 0$. Let $W^{\varepsilon,\delta}$ be the classical mollification of W^δ [35, Section

7.2], namely,

$$W^{\varepsilon,\delta}(x) = \int_{\mathbb{R}^2} \rho(y) W^{\delta}(x + \varepsilon y) dy,$$

where ρ is a smooth symmetric kernel defined by

$$\rho(y) = \begin{cases} c \exp\{1/(\|y\|^2 - 1)\} & \text{if } \|y\| \le 1, \\ 0 & \text{if } \|y\| \ge 1, \end{cases} \qquad \int_{\mathbb{R}^2} \rho(y)\, dy = 1,$$

Assuming that the mollification parameter $\varepsilon < b\delta$, we now argue that

$$W(x) = W^{\varepsilon,\delta}(x) + \gamma - \|v^*\| \cdot \varepsilon$$

is a classical subsolution. Indeed, for $x \in \mathbb{R}^2_+$, it is not difficult to see that

$$\nabla W(x) = -a(x)v^* - (1 - a(x))\hat{v}, \quad a(x) = \int_{\{y:\ \varepsilon y_2 > b\delta - x_2\}} \rho(y)\, dy.$$

Therefore, by the convexity of H and that $a(x) \in [0, 1]$

$$-H(-\nabla W(x)) \ge -[a(x)H(v^*) + (1 - a(x))H(\hat{v})] \ge 0.$$

On the other hand, for every $x = (x_1, x_2) \in \mathbb{R}^2$ such that $x_2 < b\delta - \varepsilon$, we have $\{y : \varepsilon y_2 > b\delta - x_2\} \subset \{y : \|y\| > 1\}$. Hence $a(x) = 0$ and $\nabla W(x) = -\hat{v}$. In particular, for every $x \in \partial$

$$-H_{\partial}(-\nabla W(x)) = -H_{\partial}(\hat{v}) \ge 0.$$

Finally for every $x \in \partial_e$, since $W^{\delta}(x) \le \langle -v^*, x \rangle = -\gamma$ and W^{δ} is Lipschitz continuous with $\|v^*\|$ as a Lipschitz constant (note that $\|v^*\| \ge \|\hat{v}\|$), it follows that

$$W(x) \le \int_{\mathbb{R}^2} \rho(y)\|v^*\| \cdot \varepsilon \|y\|\, dy - \|v^*\| \cdot \varepsilon \le \int_{\mathbb{R}^2} \rho(y)\|v^*\| \cdot \varepsilon dy - \|v^*\| \cdot \varepsilon = 0.$$

Applying Lemma 2.4.2, we arrive at

$$\liminf_n -\frac{1}{n}\log p_n \geq W(0) = \gamma - \delta - \|v^*\|\varepsilon \geq \gamma - (1 + b\|v^*\|)\delta$$

for all $\delta > 0$. Letting δ tend to 0, we complete the proof of the upper bound (2.32).

It remains to show the lower bound

$$\limsup_n -\frac{1}{n}\log p_n \leq \gamma.$$

We first observe that the sample path large deviation principle (i.e., Theorem 2.3.1) implies that

$$\limsup_n -\frac{1}{n}\log p_n \leq \inf \int_0^\tau L(\phi(t), \dot{\phi}(t))\, dt$$

where the infimum is taken over all absolutely continuous sample paths $\phi : [0, \infty) \to \mathbb{R}^2_+$ such that $\phi(0) = 0$, $\phi(\tau) \in \partial_e$. The proof of this inequality is standard and almost verbatim to that of equation (8.5) in [31] – the only major difference is that "$L^A(\mathbf{1})$ is finite for any $A \subset \{1, 2, \ldots, d\}$" should be replaced by "$\bar{L}_i(\mathbf{1})$ is finite for any $i = 0, 1, \ldots, d$".

In view of the above discussion, it suffices to construct a sample path ϕ^* with hitting time τ^* such that

$$\int_0^{\tau^*} L(\phi^*(t), \dot{\phi}^*(t))\, dt \leq \gamma.$$

We consider the following two scenarios.

Case 1: $\alpha_1^* \leq \hat{\alpha}_1$. Define $\beta^* = \nabla H(\alpha^*) = (\lambda_1 e^{\alpha_1^*}, \lambda_2 e^{\alpha_2^*} - \mu_2 e^{-\alpha_2^*})$. That is, β^* is the conjugate of α^* through the convex duality of H and L. Clearly $\beta_1^* > 0$. Since

$\alpha_2^* = \alpha_1^* > \bar{\alpha}_2 = -\log\sqrt{\theta_2}$ [Lemma 2.4.1], it follows that

$$\beta_2^* > \lambda_2 e^{\bar{\alpha}_2} - \mu_2 e^{-\bar{\alpha}_2} = \sqrt{\lambda_2\mu_2} - \sqrt{\lambda_2\mu_2} = 0. \tag{2.33}$$

Thus the trajectory $\phi^*(t) = \beta^* t$ lives in the positive orthant and τ^*, defined as the first hitting time to ∂_e, is finite. It follows from the definition of $L(\cdot,\cdot)$ and the conjugacy of β^* and α^* that for every $t > 0$

$$L(\phi^*(t), \dot{\phi}^*(t)) = L(\dot{\phi}^*(t)) = L(\beta^*) = \langle \alpha^*, \beta^* \rangle - H(\alpha^*) = \langle \alpha^*, \beta^* \rangle.$$

Therefore,

$$\int_0^{\tau^*} L(\phi^*(t), \dot{\phi}^*(t))\, dt = \int_0^{\tau^*} \langle \alpha^*, \beta^* \rangle\, dt = \langle \alpha^*, \beta^* \tau^* \rangle.$$

Since $\alpha_1^* = \alpha_2^*$ and $\beta^*\tau^* \in \partial_e$, we have that $\langle \alpha^*, \beta^*\tau^* \rangle = \alpha_1^* = \gamma$.

Case 2: $\alpha_1^* > \hat{\alpha}_1$. Define $\bar{\beta} = \nabla H(\hat{\alpha})$ and $\hat{\beta} = \nabla H_\partial(\hat{\alpha})$. Thus $\bar{\beta}$ and $\hat{\alpha}$ are conjugate through the convex duality of H and L, while $\hat{\beta}$ and $\hat{\alpha}$ are conjugate through the convex duality of H_∂ and L_∂. By direct calculation

$$\bar{\beta} = (\lambda_1 e^{\hat{\alpha}_1}, \lambda_2 e^{\hat{\alpha}_2} - \mu_2 e^{-\hat{\alpha}_2}), \quad \hat{\beta} = (\lambda_1 e^{\hat{\alpha}_1} - \mu_1 e^{-\hat{\alpha}_1}, \lambda_2 e^{\hat{\alpha}_2}).$$

Since $\hat{\alpha}_2 < \bar{\alpha}_2$, it follows that $\bar{\beta}_2 < 0$ by an argument analogous to (2.33). Define $\rho_1 = \hat{\beta}_2(\hat{\beta}_2 - \bar{\beta}_2)^{-1}$ and $\rho_2 = -\bar{\beta}_2(\hat{\beta}_2 - \bar{\beta}_2)^{-1}$. Then ρ_1, ρ_2 are both non-negative, $\rho_1 + \rho_2 = 1$, and

$$\beta^* = \rho_1\bar{\beta} + \rho_2\hat{\beta} = (\beta_1^*, 0).$$

We claim that $\beta_1^* > 0$. Indeed, since H and L are both strictly convex and $L(\beta) = 0$

if and only if $\beta = \nabla H(0)$, it follows from the conjugacy of $\bar{\beta}$ and $\hat{\alpha}$ that

$$\langle \hat{\alpha}, \bar{\beta} \rangle = \langle \hat{\alpha}, \bar{\beta} \rangle - H(\hat{\alpha}) = L(\bar{\beta}) > 0. \tag{2.34}$$

Similarly

$$\langle \hat{\alpha}, \hat{\beta} \rangle = \langle \hat{\alpha}, \hat{\beta} \rangle - H_{\partial}(\hat{\alpha}) = L_{\partial}(\bar{\beta}) > 0. \tag{2.35}$$

Therefore,

$$\beta_1^* \hat{\alpha}_1 = \langle \beta^*, \hat{\alpha} \rangle = \rho_1 \langle \hat{\alpha}, \bar{\beta} \rangle + \rho_2 \langle \hat{\alpha}, \hat{\beta} \rangle > 0,$$

which in turn implies that $\beta_1^* > 0$. Define the trajectory $\phi^*(t) = \beta^* t$ and let τ^* be the first hitting time to the exit boundary ∂_e. The trajectory travels along the boundary ∂ and the hitting time τ^* is finite. Furthermore,

$$\int_0^{\tau^*} L(\phi^*(t), \dot{\phi}^*(t))\, dt = \int_0^{\tau^*} (L \oplus L_{\partial})(\beta^*) dt = \tau^* \cdot (L \oplus L_{\partial})(\beta^*).$$

However, by the definition of inf-convolution, (2.34) and (2.35)

$$(L \oplus L_{\partial})(\beta^*) \leq \rho_1 L(\bar{\beta}) + \rho_2 L_{\partial}(\hat{\beta}) = \rho_1 \langle \hat{\alpha}, \bar{\beta} \rangle + \rho_2 \langle \hat{\alpha}, \hat{\beta} \rangle = \langle \hat{\alpha}, \beta^* \rangle.$$

It follows that

$$\int_0^{\tau^*} L(\phi^*(t), \dot{\phi}^*(t))\, dt \leq \langle \hat{\alpha}, \beta^* \tau^* \rangle = \hat{\alpha}_1 = \gamma.$$

This completes the proof. ■

Remark 2.4.4. The proof actually shows that the decay rate γ equals the value of the calculus of variation problem

$$\gamma = \inf \int_0^{\tau} L(\phi(t), \dot{\phi}(t))\, dt$$

and that the trajectory ϕ^* is indeed a minimizing trajectory.

2.5 Summary

This chapter uses a weak convergence approach to establish the sample path large deviation principle for a single server system with preemptive priority service policy. The difficulty of the analysis is due to the *discontinuity* of the system dynamics. It is shown that the general upper bound rate function [29] is indeed *tight* since the stability-about-the-interface condition is automatically built into the upper bound rate function. This simple form of the rate function proves to be useful when one studies the asymptotic behavior of various buffer overflow probabilities. For illustration, in the two dimensional case the exponential decay rate of the total population overflow probabilities is explicitly identified. This is achieved by studying the geometry of the zero level sets of the system Hamiltonians and by constructing appropriate subsolutions to the related partial differential equation.

CHAPTER III:

Importance Sampling for a Single Server Priority Queue

3.1 Overview

Importance sampling is a variance reduction technique, especially effective in Monte Carlo simulation of rare event probabilities [23, 24, 25, 37, 38, 39, 40, 41, 42, 44, 45] and the references therein. The basic idea of importance sampling is to simulate the dynamics of the system under an alternative probability distribution (i.e., change of measure). This introduces a bias in the Monte Carlo estimator, which will be rectified by multiplying the outcome with the appropriate likelihood ratio.

It is clear by now that unless in very simple settings, state-dependent change of measure is needed to achieve efficiency [26, 36]. This is particularly true in the context of queueing networks where the system dynamics are in general high-dimensional and discontinuous.

The systematic study of state-dependent importance sampling schemes originated from [32]. The key observation is that importance sampling is closely connected to

a small noise stochastic game and the corresponding limit Isaacs equation. Moreover, in order to construct efficient importance sampling schemes it suffices to build appropriate *classical subsolutions* to the said Isaacs equation [33, 34].

A commonly used strategy is to first build a non-classical piecewise affine subsolution and then mollify it in an appropriate fashion in order to obtain a classical subsolution. This approach is particularly effective for systems with piecewise homogeneous dynamics such as queueing networks. However, a drawback is that it will introduce a mollification parameter that complicates the change of measure, and the determination of its value is more of an art than a science.

In this Chapter we restrict our attention to the two-dimensional case of the feedforward network introduced in the previous Chapter, and demonstrate that for a system where the form of discontinuity is simple, one does not need this mollification in order to construct efficient importance sampling schemes. The asymptotic optimality is established via a verification argument, where we construct a suitable subsolution. Compared with the previous proofs involving classical subsolutions [33, 34], the construction is more subtle in the sense that we don't have the extra mollification parameter to control the second derivative of the subsolution.

The outline of this Chapter is as follows. In Sections 3.2 the system model and dynamics are described. The rare event of interest is introduced in Section 3.3 and the corresponding large deviations results are stated in Section 3.4. Section 3.5 is devoted to importance sampling and the proof of asymptotic optimality. Numerical results are presented in Section 3.6. A brief summary is given in Section 3.7. For ease of exposition, some of the technical proofs are deferred to an appendix.

Remark 3.1.1. Throughout the Chapter, if x is a vector, then x_j denotes its j-th coordinate. We also adopt the standard notation that e_i represents the vector with

the i-th coordinate 1 and 0 otherwise.

3.2 System Model and Dynamics

The system consists of a single server station serving two classes of exogenous jobs. Jobs of class i arrive according to a Poisson process with rate λ_i, and are buffered at queue i. The service time for a class i job is exponentially distributed with rate μ_i, $i = 1, 2$. The arrival processes and the service times are assumed to be independent. The service discipline for the system is such that a job of class 2 has preemptive priority over a job of class 1. Jobs with the same priority level are served according to the first-in-first-out policy. See Figure 3.1.

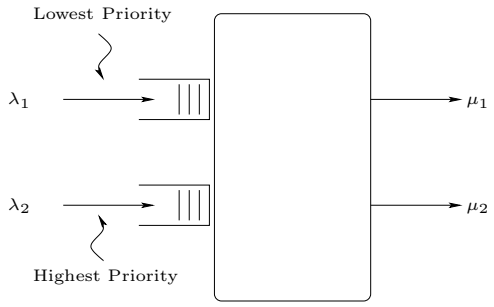

Figure 3.1: A feed-forward network with priority service policy

We denote by $Q_i(t)$ the queue size of class i job at time t. Then the state process $Q = \{(Q_1(t), Q_2(t)) : t \geq 0\}$ is a two dimensional continuous time pure jump Markov process, defined on some probability space say $(\Omega, \mathbb{F}, P)$. We define a lower-semicontinuous mapping Π to denote the index of the non-empty queue with the highest priority at state $x = (x_1, x_2)$, that is,

$$\Pi(x) \doteq \max\{i : x_i > 0\} \quad \text{with convention } \Pi(0) = 0.$$

Under the preemptive service policy, the set of all possible jumps of Q is

$$\mathbb{V} = \{\pm e_1, \pm e_2\}$$

and the jump intensity from state x to state $x + v$ is

$$r(x, v) \doteq \begin{cases} \lambda_i & \text{if } v = e_i \\ \mu_i & \text{if } v = -e_i \text{ and } i = \Pi(x) \\ 0 & \text{otherwise} \end{cases}$$

for all $x \in \mathbb{R}^2_+$ and $v \in \mathbb{V}$.

Due to the priority service policy, the system dynamics is *discontinuous* at the interface $\{Q_2 = 0\}$ and the origin. Note that since no jumps in the interior of the state space attempt to leave the non-negative orthant through $\{Q_1 = 0\}$, no discontinuity is present on this boundary. See Figure 3.2.

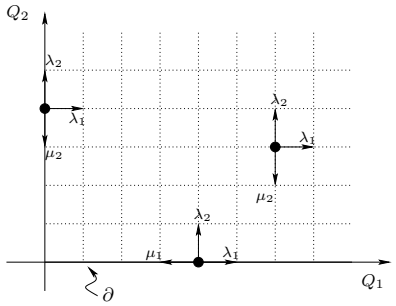

Figure 3.2: Discontinuous dynamics

3.3 The Rare Event

Throughout the Chapter, we assume that the stability condition holds, i.e.,

$$\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} < 1.$$

Under this assumption, the total population overflow

$$A_n \doteq \{\text{total population } Q_1 + Q_2 \text{ reaches } n \text{ before going back to 0,} \\ \text{starting from } Q(0) = (0,0)\}$$

is a rare event when n is large. The present Chapter is interested in efficient Monte Carlo importance sampling schemes for estimating the rare event probability

$$p_n \doteq P(A_n).$$

3.4 Review of Large Deviations Results

In order to construct efficient importance sampling algorithms, it is essential to understand the large deviation asymptotics of $\{p_n\}$. In the previous Chapter, the sample path large deviations for a feedforward network with priority service policy was established, and the exponential decay rate of p_n was also explicitly identified. The purpose of this Section is to briefly review these results.

For a system with discontinuous dynamics, there exists a simple, sample path large deviation upper bound local rate function, whose value at a given point can be identified as the inf-convolution of the neighboring local rate functions at that point [29]. This upper bound in general is not tight since it does not explicitly take into account the "stability-about-the-interface" condition [28, Chapter 7]. However, for many physically meaningful systems, this stability condition is implicitly built into the upper bound rate function [30, 31]. Thus the upper bound rate function is indeed the true sample path large deviation rate function. With this simple form of rate function at hand, one can show through partial differential equations the asymptotic

behavior of various buffer overflow probabilities. This is done through examining the geometry of the zero level sets of system Hamiltonians and by constructing suitable subsolutions to the aforementioned partial differential equation.

3.4.1 System Hamiltonians

For every $\alpha = (\alpha_1, \alpha_2) \in \mathbb{R}^2$, we define

$$\begin{aligned} H(\alpha) &\doteq \lambda_1(e^{\alpha_1} - 1) + \lambda_2(e^{\alpha_2} - 1) + \mu_2(e^{-\alpha_2} - 1), \\ H_\partial(\alpha) &\doteq \lambda_1(e^{\alpha_1} - 1) + \lambda_2(e^{\alpha_2} - 1) + \mu_1(e^{-\alpha_1} - 1). \end{aligned}$$

The functions H and H_∂ are both strictly convex, where H corresponds to the Hamiltonian in the interior of the state space $\{x_2 > 0\}$ and H_∂ the Hamiltonian on the boundary $\partial \doteq \{x_2 = 0\}$. Since these Hamiltonians are closely related to the log of the moment generating functions of the infinitesimal increments of the process Q, they play an important role in the large deviations analysis.

3.4.2 The Exponential Decay Rate of p_n

The exponential decay rate of the rare event probability p_n can be explicitly identified in terms of the roots of the Hamiltonians H and H_∂ [43]. There are three roots of particular importance: (i) α^*, the intersection of $\{H = 0\}$ and the 45-degree line; (ii) $\hat{\alpha}$, the intersection of $\{H = 0\}$ and $\{H_\partial = 0\}$; (iii) $\bar{\alpha}$, the right-most point on $\{H = 0\}$. See Figure 3.3. Each of these three roots corresponds to a possible asymptotically most likely path leading to the rare event (a straight line by convexity of the local rate function). More precisely, α^* corresponds to a path in the interior

of the state space, $\hat{\alpha}$ a path that "pushes into" the boundary ∂, and $\bar{\alpha}$ a path that "glides" along the boundary ∂.

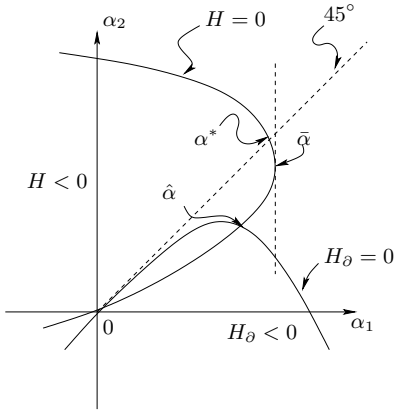

Figure 3.3: Roots of the Hamiltonians

As shown in Theorem 3.4.1, these three roots can be explicitly solved. For notational convenience, we define

$$\theta_1 \doteq \frac{\lambda_1}{\mu_2}, \quad \theta_2 \doteq \frac{\lambda_2}{\mu_2}, \quad \theta_3 \doteq \frac{\mu_1}{\mu_2},$$

and

$$z \doteq \frac{(\theta_1 + \theta_2 + \theta_3 - 1) + \sqrt{(\theta_1 + \theta_2 + \theta_3 - 1)^2 + 4\theta_1(1 - \theta_3)}}{2\theta_3}.$$

Theorem 3.4.1. *The constant z satisfies $\max\{0, 1-1/\theta_3\} < z < 1$. The exponential decay rate of p_n is:*

$$\gamma \doteq -\lim_n \frac{1}{n} \log p_n = \begin{cases} \min(\alpha_1^*, \hat{\alpha}_1) & \text{if } \alpha_2^* > \bar{\alpha}_2, \hat{\alpha}_2 < \bar{\alpha}_2 \\ \alpha_1^* & \text{if } \alpha_2^* > \bar{\alpha}_2, \hat{\alpha}_2 \geq \bar{\alpha}_2 \\ \hat{\alpha}_1 & \text{if } \alpha_2^* \leq \bar{\alpha}_2, \hat{\alpha}_2 < \bar{\alpha}_2 \\ \bar{\alpha}_1 & \text{if } \alpha_2^* \leq \bar{\alpha}_2, \hat{\alpha}_2 \geq \bar{\alpha}_2 \end{cases},$$

where

$$\begin{aligned}
\alpha^* &\doteq (-\log[\theta_1+\theta_2], -\log[\theta_1+\theta_2]),\\
\hat{\alpha} &\doteq (-\log z, -\log[1-\theta_3+\theta_3 z]),\\
\bar{\alpha} &\doteq \left(\log\left[1+(1-\sqrt{\theta_2})^2/\theta_1\right], -\log\sqrt{\theta_2}\right).
\end{aligned}$$

A detailed proof can be found in [43].

3.5 Importance Sampling

As mentioned before, importance sampling is a variance reduction technique for Monte Carlo simulation, particularly effective in the context of rare event simulation. The idea is to generate samples from an alternative probability distribution under which the event is no longer rare. However, one has to select this change of measure appropriately in order to achieve the so called *asymptotic optimality* [38, 45]. In this Chapter, we are interested in building asymptotically optimal importance sampling schemes for estimating p_n when n is large.

3.5.1 Asymptotic Optimality

Recall that A_n is the rare event of buffer overflow, and $p_n \doteq P(A_n)$. An importance sampling scheme generates samples from a new probability measure, say Q_n, such that $P \ll Q_n$. The unbiased importance sampling estimator, denoted by $\hat{p}_n$, is then given by:

$$\hat{p}_n = 1_{A_n}\frac{dP}{dQ_n},$$

where dP/dQ_n denotes the Radon-Nikodym derivative or the likelihood ratio. The goal here is to choose Q_n such that the variance, or the second moment of $\hat{p}_n$, is minimized. By Jensen's inequality, and the large deviations properties of p_n (Theorem 3.4.1) a lower bound follows:

$$\liminf_n \frac{1}{n} \log E^{Q_n}[\hat{p}_n^2] \geq \liminf_n \frac{2}{n} \log E^{Q_n}[\hat{p}_n] = \liminf_n \frac{2}{n} \log p_n = -2\gamma.$$

An importance sampling scheme is said to be asymptotically optimal if this lower bound is *achieved*, i.e., if

$$\limsup_n \frac{1}{n} \log E^{Q_n}[\hat{p}_n^2] \leq -2\gamma.$$

We would like to make the following straightforward but useful observation

$$E^{Q_n}[\hat{p}_n^2] = E^{Q_n}\left[\hat{p}_n \frac{dP}{dQ_n}\right] = E^P[\hat{p}_n],$$

and hence asymptotic optimality amounts to

$$\limsup_n \frac{1}{n} \log E^P[\hat{p}_n] \leq -2\gamma.$$

3.5.2 Classical Subsolution Approach

For our model, an importance sampling change of measure can be described by an alternative jump intensity function say $\bar{r}(x, v)$. In other words, under the new probability measure, the jump intensity for the process Q to make a jump of size v at state $q \in \mathbb{Z}_+^2$ is $\bar{r}(x, v)$ where $x = q/n \in \mathbb{R}_+^2$. The key question is how $\bar{r}$ should be related to the original jump intensity function $r(x, v)$.

In a series of papers [32, 33, 34], it has been established that efficient state-dependent importance sampling schemes can be constructed from *classical subsolutions* to a related Hamilton-Jacobi-Bellman (HJB) equation. When applied to our setting, a classical subsolution is a continuously differentiable function $W : \mathbb{R}^2_+ \to \mathbb{R}$ that satisfies the following three properties

$$\begin{aligned} -H(-\nabla W(x)) &\geq 0 \quad \text{if } x_1 \geq 0, x_2 > 0 && (3.1)\\ -H_\partial(-\nabla W(x)) &\geq 0 \quad \text{if } x_1 \geq 0, x_2 = 0 && (3.2)\\ W(x) &\leq 0 \quad \text{if } x_i \geq 0, x_1 + x_2 = 1. && (3.3) \end{aligned}$$

The change of measure associated with W takes the form [33]

$$\bar{r}(x, v) = r(x, v) \cdot e^{-\langle \nabla W(x), v \rangle} \tag{3.4}$$

for all $x \in \mathbb{R}^2_+$ and $v \in \mathbb{V}$. In [33] it was also shown that the exponential decay rate of the second moment of the corresponding importance sampling estimator is at least $2W(0)$. Therefore to achieve asymptotic optimality, one should construct a classical subsolution W whose value at the origin is as close to γ as possible.

3.5.3 Piecewise Constant Change of Measure

For queueing networks or systems with piecewise homogeneous dynamics, classical subsolutions are often constructed by mollifying suitable piecewise affine subsolutions [33, 34]. This requires an additional small mollification parameter, whose value is often chosen by experience and/or trial-and-error. In general, it seems that this mollification is necessary in order to control the variance of the importance sampling estimator.

In this work, we show that for the current model, this mollification is *not* needed. This is due to the fact that a simpler *piecewise constant* change of measure that is asymptotically optimal can be directly constructed. In what follows, we identify this change of measure (denoted by Q_n), for the estimation of p_n.

Consider a triple $(\alpha^{[1]}, \alpha^{[2]}, \delta_n)$, where $\alpha^{[1]}$ and $\alpha^{[2]}$ are two-dimensional vectors and δ_n a non-negative real number. The values of $\alpha^{[1]}$ and $\alpha^{[2]}$ are essential in determining the new jump intensity function $\bar{r}$. In light of Theorem 3.4.1, taking into account (3.1) and (3.2), we choose $\alpha^{[1]}$ and $\alpha^{[2]}$ as follows (see Figure 3.4).

1. If $\alpha_2^* > \bar{\alpha}_2$ and $\hat{\alpha}_2 < \bar{\alpha}_2$, then:

$$\alpha^{[1]} \doteq \frac{\min\{\alpha_1^*, \hat{\alpha}_1\}}{\alpha_1^*} \cdot \alpha^*, \quad \alpha^{[2]} \doteq \frac{\min\{\alpha_1^*, \hat{\alpha}_1\}}{\hat{\alpha}_1} \cdot \hat{\alpha}.$$

2. If $\alpha_2^* > \bar{\alpha}_2$ and $\hat{\alpha}_2 \geq \bar{\alpha}_2$, then:

$$\alpha^{[1]} \doteq \alpha^*, \quad \alpha^{[2]} \doteq \frac{\alpha_1^*}{\bar{\alpha}_1} \bar{\alpha}.$$

3. If $\alpha_2^* \leq \bar{\alpha}_2$ and $\hat{\alpha}_2 < \bar{\alpha}_2$, then:

$$\alpha^{[1]} \doteq \frac{\hat{\alpha}_1}{\bar{\alpha}_1} \bar{\alpha}, \quad \alpha^{[2]} \doteq \hat{\alpha}.$$

4. If $\alpha_2^* \leq \bar{\alpha}_2$ and $\hat{\alpha}_2 \geq \bar{\alpha}_2$, then

$$\alpha^{[1]} \doteq \bar{\alpha}, \quad \alpha^{[2]} \doteq \bar{\alpha}.$$

δ_n is a small parameter that determines the thickness of what we call a "boundary layer". By a boundary layer we mean a thin stripe where x_2 is between 0 and δ_n;

see Figure 3.4.

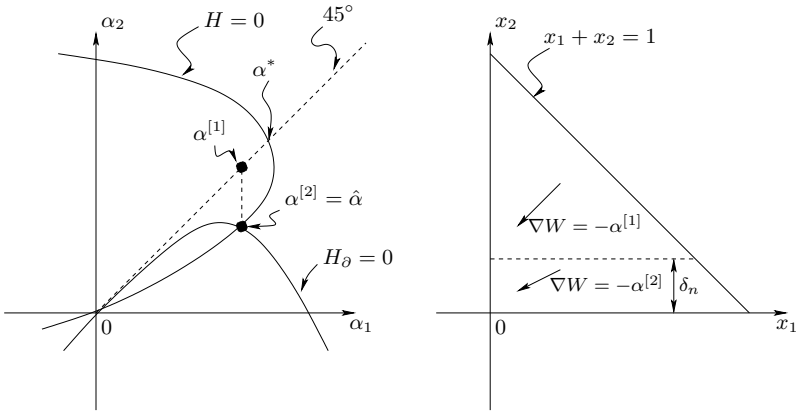

Figure 3.4: An example of $\alpha^{[1]}$ and $\alpha^{[2]}$: case 1 with $\alpha_1^* > \hat{\alpha}_1$

Finally, the change of measure Q_n associated with this triple can be characterized by the new jump intensity function $\bar{r}$, where for $x \in \mathbb{R}^2_+ \setminus \{0\}$

$$\bar{r}(x,v) \doteq \begin{cases} r(x,v) \cdot e^{\langle \alpha^{[1]}, v \rangle} & \text{if } x_2 \geq \delta_n, \\ r(x,v) \cdot e^{\langle \alpha^{[2]}, v \rangle} & \text{if } 0 \leq x_2 < \delta_n, \end{cases}$$

and $\bar{r}(0,v) \doteq r(0,v)$ for all $v \in \mathbb{V}$. Compared with the change of measure formula (3.4), $\bar{r}$ corresponds to a piecewise affine subsolution whose gradient is $-\alpha^{[1]}$ when $x_2 \geq \delta_n$ and $-\alpha^{[2]}$ when $x_2 < \delta_n$.

3.5.4 The Importance Sampling Estimator

In order to identify the importance sampling estimator, we introduce the following notation. Denote the total jump intensities by

$$R(x) \doteq \sum_{v \in \mathbb{V}} r(x,v), \qquad \bar{R}(x) \doteq \sum_{v \in \mathbb{V}} \bar{r}(x,v).$$

Let $\{T_1, T_2, \ldots\}$ be the jump times of the state process Q with convention $T_0 = 0$. Let $s_j \doteq T_j - T_{j-1}$ (sojourn times) and $v_j \doteq Q(T_j) - Q(T_{j-1})$ (jump sizes). Define

$$N \doteq \inf\{k \geq 1 : Q_1(T_k) + Q_2(T_k) = n \text{ or } 0\}.$$

and the scaled state process

$$X^n(t) \doteq \frac{1}{n} Q(t).$$

Then an unbiased importance sampling estimator is

$$\hat{p}_n = 1_{A_n} \cdot \prod_{j=1}^{N} \frac{r(X^n(T_{j-1}), v_j) e^{-R(X^n(T_{j-1}))s_j}}{\bar{r}(X^n(T_{j-1}), v_j) e^{-\bar{R}(X^n(T_{j-1}))s_j}}. \tag{3.5}$$

Note that the likelihood ratio in the definition of $\hat{p}_n$ is with respect to the continuous time sample paths. One can also identify another unbiased importance sampling estimator based on the embedded discrete time Markov chains $\{X^n(T_j) : j \geq 0\}$, which is

$$\bar{p}_n \doteq 1_{A_n} \cdot \prod_{j=1}^{N} \frac{r(X^n(T_{j-1}), v_j)/R(X^n(T_{j-1}))}{\bar{r}(X^n(T_{j-1}), v_j)/\bar{R}(X^n(T_{j-1}))}. \tag{3.6}$$

In other words, $\bar{p}_n$ is obtained by integrating out $\{s_j\}$ in the definition of $\hat{p}_n$, or more precisely,

$$E^{Q_n}[\hat{p}_n | X^n(T_1), X^n(T_2), \ldots, X^n(T_N)] = \bar{p}_n.$$

In the next section, we will show that $\hat{p}_n$ is asymptotically optimal, which implies the asymptotic optimality of $\bar{p}_n$. This is because the second moment of $\bar{p}_n$ does exceed that of $\hat{p}_n$.

The reason for introducing $\hat{p}_n$ is its suitability for asymptotic analysis. However, for the purpose of numerical simulation, we use $\bar{p}_n$ due to its convenience.

3.5.5 The Verification Argument

In this section, we establish the asymptotic optimality of the importance sampling estimator $\hat{p}_n$ (Theorem 3.5.1) under mild conditions through a verification argument. The key part in the proof is the construction of a suitable subsolution and a related supermartingale. The difference from the original proofs involving classical subsolutions [33, 34] is that we don't have the luxury to directly control the second derivative of the subsolution by adjusting the mollification parameter. As a consequence, the construction of a good subsolution is more involved.

Theorem 3.5.1 (Verification Argument)**.** *Suppose that $\delta_n \to 0$ and $n\delta_n \to \infty$. Then the importance sampling estimator $\hat{p}_n$ defined in (3.5) is asymptotically optimal.*

Proof. By the definition of asymptotic optimality in Section 3.5.1, it suffices to show that

$$\limsup_n \frac{1}{n} \log E^P[\hat{p}_n] \leq -2\gamma.$$

To ease notation, we drop the super-index P, with the understanding that all the calculation will be done under the original probability measure P. Recall the scaled state process $X^n(t) \doteq Q(t)/n$, and denote by $E_{x^n}[\cdot] \doteq E[\cdot|X^n(0) = x^n]$. Furthermore, observe that by conditioning on the first jump [note the first jump can only be e_1 or e_2 with probability $\lambda_1/(\lambda_1+\lambda_2)$ and $\lambda_2/(\lambda_1+\lambda_2)$ respectively]

$$E[\hat{p}_n] = \sum_{i=1}^{2} \frac{\lambda_i}{\lambda_1+\lambda_2} E[\hat{p}_n|Q(0) = e_i] = \sum_{i=1}^{2} \frac{\lambda_i}{\lambda_1+\lambda_2} E_{e_i/n}[\hat{p}_n].$$

Therefore it suffices to show that

$$\limsup_n \frac{1}{n} \log E_{x^n}[\hat{p}_n] \leq -2\gamma,$$

where $x^n \to 0$ and $x^n \neq 0$.

We will only prove this upper bound for $\bar{\alpha}_2^* > \bar{\alpha}_2$ and $\hat{\alpha}_2 < \bar{\alpha}_2$, that is, the first case in Section 3.5.3. The proof for the other scenarios is indeed a simpler version of this and thus omitted. Note that in this case the exponential decay rate $\gamma = \min\{\alpha_1^*, \hat{\alpha}_1\}$.

As mentioned before, the proof involves the construction of a subsolution. Let $B \doteq \alpha_2^{[1]} - \alpha_2^{[2]}$, which is strictly positive (Figure 4 depicts this fact and we omit the trivial but tedious calculations). Define

$$\begin{aligned} W_1(x) &\doteq -\langle \alpha^{[1]}, x \rangle + \gamma, \\ W_2(x) &\doteq -\langle \alpha^{[2]}, x \rangle + \gamma - B\delta_n. \end{aligned}$$

According to the definition of $\alpha^{[i]}$ (Section 3.5.3), observe $\alpha_1^{[1]} = \alpha_1^{[2]} = \gamma$. It is easy to verify that W_1 and W_2 are two affine functions that tie on the horizontal line $\{x_2 = \delta_n\}$. Furthermore,

$$\bar{W}(x) \doteq \min\{W_1(x), W_2(x)\} = \begin{cases} W_1(x) & \text{if } x_2 \geq \delta_n, \\ W_2(x) & \text{if } x_2 \leq \delta_n. \end{cases} \tag{3.7}$$

Fix an arbitrary $\varepsilon_n > 0$ such that $\varepsilon_n/\delta_n \in (0,1)$ but bounded away from 0 and 1. The specific choice of ε_n is not important and therefore for simplicity, we will set $\varepsilon_n = \delta_n/2$ throughout. For the purpose of exploiting Itô formula later on, we consider a classical mollification of $\bar{W}$ as follows:

$$W(x) \doteq \int_{\mathbb{R}^2} \rho(y) \bar{W}\left(x + \varepsilon_n y\right) dy - \|\alpha^{[1]}\| \varepsilon_n,$$

where ρ is the classical smooth symmetric mollifier given by

$$\rho(y) = \begin{cases} c \exp\{1/(\|y\|^2 - 1)\} & \text{if } \|y\| \leq 1, \\ 0 & \text{if } \|y\| \geq 1, \end{cases}$$

and the normalizing constant c is chosen such that

$$\int_{\mathbb{R}^2} \rho(y)\, dy = 1.$$

The following lemma establishes some useful properties of W which are essential in the future analysis. Its proof is deferred to Appendix B.

Lemma 3.5.2 (Subsolution Properties)**.** *The function W is smooth and has the following properties.*

1. *The gradient of W is a convex combination of $\alpha^{[1]}$ and $\alpha^{[2]}$:*

$$\nabla W(x) = -\theta(x)\alpha^{[1]} - (1 - \theta(x))\alpha^{[2]},$$

where

$$\theta(x) \doteq \int_{\{y=(y_1,y_2):\ \varepsilon_n y_2 > \delta_n - x_2\}} \rho(y)\, dy \in [0, 1].$$

2. *W satisfies the boundary inequality*

$$W(x) \leq 0, \qquad \text{if } x_1 + x_2 \geq 1$$

3. *The value at the origin $W(0, 0) = \gamma - B\delta_n$.*

4. *Each element of the Hessian matrix $\nabla^2 W(x)$ is uniformly bounded by C/ε_n for some constant C independent of n.*

Finally for any $\theta^* \in (0,1)$ we also define

$$W_{\theta^*}(x_1, x_2) = W(x_1, x_2 - \varepsilon_n x_2^*), \tag{3.8}$$

where $x_2^* = x_2^*(\theta^*)$ is uniquely determined by the equation

$$\theta^* = \int_{\{y=(y_1,y_2):\ y_2>x_2^*\}} \rho(y)\, dy. \tag{3.9}$$

Clearly if $\theta^* = 1/2$ then $x_2^* = 0$ (by the symmetry of ρ) and $W_{\theta^*} = W$.

Following the notation of Section 3.5.4, let $J(t) \doteq \inf\{j : T_j \geq t\}$ and

$$Y_n(t) \doteq \exp\left\{ \int_0^t [\bar{R}(X^n(s)) - R(X^n(s))]\, ds + \sum_{j=1}^{J(t)} \log \frac{r(X^n(T_{j-1}), v_j)}{\bar{r}(X^n(T_{j-1}), v_j)} \right\}.$$

Recalling the definition of $\hat{p}_n$ in (3.5), it is not difficult to see that

$$\hat{p}_n = 1_{A_n} \cdot Y_n(T_N). \tag{3.10}$$

Let b and θ^* be two constants independent of n, whose values will be specified later in Lemma 3.5.3. Pick arbitrarily $q \in (1,2)$. Consider the non-negative process

$$M^n(t) \doteq \exp\left\{-bnW_{\theta^*}(X^n(t))\right\} Y_n^q(t),$$

It follows from generalized Itô formula [31, Appendix A.6] that the process

$$M^n(t) + \int_0^t M^n(s) h_n(X^n(s))\, ds$$

is a local martingale, where

$$h_n(x) \doteq q[R(x) - \bar{R}(x)] + \sum_{v \in \mathbb{V}} r(x,v) \left[1 - e^{-bn(W_{\theta^*}(x+v/n) - W_{\theta^*}(x))} \left(\frac{r(x,v)}{\bar{r}(x,v)} \right)^q \right].$$

It follows from Lemma 3.5.2 Part 4, definition (3.8) of W_{θ^*}, and Taylor expansion that for some constant $\bar{C}$ independent of n

$$|\langle DW_{\theta^*}(x), v \rangle - n(W_{\theta^*}(x+v/n) - W_{\theta^*}(x))| \leq \frac{\bar{C}}{n\varepsilon_n} = \frac{2\bar{C}}{n\delta_n},$$

which in turn implies that for some constant K independent of n

$$h_n(x) \geq \bar{h}_n(x) - K[e^{2b\bar{C}/(n\delta_n)} - 1] \tag{3.11}$$

where

$$\bar{h}_n(x) \doteq q[R(x) - \bar{R}(x)] + \sum_{v \in \mathbb{V}} r(x,v) \left[1 - e^{-b\langle \nabla W_{\theta^*}(x), v \rangle} \left(\frac{r(x,v)}{\bar{r}(x,v)} \right)^q \right].$$

The following lemma is the key observation in this verification argument. Its proof is fairly technical and is deferred to Appendix B.

Lemma 3.5.3. *For every $0 < \varepsilon < 1$, there exists $b \in (2-\varepsilon, 2]$ such that for $\theta^* = 1/b$ and all $q > 1$ but sufficiently close to 1, $\{\bar{h}_n(x) : x \in \mathbb{R}^2_+, x \neq 0\}$ is uniformly bounded from below by a strictly positive constant that is independent of n (may depend on b, θ^*, q).*

Fix the triple (b, θ^*, q). Since $n\delta_n \to \infty$, the preceding lemma and inequality (3.11) imply that $h_n(x) \geq 0$ for every $x \in \mathbb{R}^2_+ \setminus \{0\}$ when n is large enough. It follows that M^n is a non-negative local supermartingale, and hence a true supermartingale,

up until the hitting time T_N. In particular, by the optional sampling theorem

$$E_{x^n}[M^n(T_N)] \leq M^n(0) = e^{-bnW_{\theta^*}(x^n)}. \tag{3.12}$$

Since $\theta^* = 1/b \geq 1/2$, it follows immediately from definition (3.9) that $x_2^* \leq 0$. Therefore for $x_1 + x_2 = 1$, $x_1 + x_2 - \varepsilon_n x_2^* \geq 1$, and thus Lemma 3.5.2 Part 2 yields

$$W_{\theta^*}(x_1, x_2) = W(x_1, x_2 - \varepsilon_n x_2^*) \leq 0.$$

Note that on the set A_n, $X^n(T_N)$ lies on the line $\{x_1 + x_2 = 1\}$. It follows from non-negativity of M^n and the preceding discussion that

$$M^n(T_N) \geq 1_{A_n} e^{-bnW_{\theta^*}(X^n(T_N))} Y_n^q(T_N) \geq 1_{A_n} Y_n^q(T_N),$$

and hence by (3.12)

$$E_{x^n}[1_{A_n} Y_n^q(T_N)] \leq E_{x^n}[M^n(T_N)] \leq e^{-bnW_{\theta^*}(x^n)}.$$

Using Hölder's inequality, we arrive at

$$E_{x^n}[\hat{p}_n] = E_{x^n}[1_{A_n} Y_n(T_N)] \leq (E_{x^n}[1_{A_n} Y_n^q(T_N)])^{1/q} \leq e^{-bnW_{\theta^*}(x^n)/q}. \tag{3.13}$$

Observe that Lemma 3.5.2 Part 1 implies the Lipschitz continuity of W and hence

$$\lim_n |W_{\theta^*}(x^n) - W(0)| = \lim_n |W(x_1^n, x_2^n - \varepsilon_n x_2^*) - W(0,0)| = 0.$$

However, owing to Lemma 3.5.2 Part 3, $W(0) = \gamma - B\delta_n$. Therefore

$$\lim_n W_{\theta^*}(x^n) = \lim_n W(0) = \gamma.$$

Combined with (3.13), taking logarithm and dividing by n on both sides, it follows that

$$\limsup_n \frac{1}{n} \log E_{x^n}[\hat{p}_n] \leq -b\gamma/q.$$

Now letting $q \to 1$ and then $b \to 2$ we have

$$\limsup_n \frac{1}{n} \log E_{x^n}[\hat{p}_n] \leq -2\gamma,$$

and the proof is complete. □

3.6 Numerical Results

As mentioned in Section 3.5.4, the importance sampling estimator, $\bar{p}_n$, based on the embedded discrete time Markov chain, is used in the simulation. Each estimate is based on a sample size of 10000. The empirical relative errors of the estimates, defined as below

$$\text{empirical relative error} \doteq \frac{\text{standard error of the estimate}}{\text{estimate}}$$

are very small in all these numerical simulations, indicating that the importance sampling estimators are highly accurate. The parameters in the i-th table correspond to the i-th scenario in Section 3.5.3, for each $i = 1, \ldots, 4$. We set the parameter $\delta_n \doteq 1/\sqrt{n}$ for all these cases. The simulation results are quite robust for different choices of δ_n. For example, we run the simulation for $\delta_n = 1/\log n$ and the results are almost identical.

	$n = 20$	$n = 50$	$n = 100$
Estimate	2.01×10^{-4}	2.46×10^{-9}	1.64×10^{-17}
Relative Error	2.5%	2.5%	2.6%

Table 1. $\lambda_1 = 0.1$, $\lambda_2 = 0.2$, $\mu_1 = 0.2$, $\mu_2 = 0.8$

	$n = 20$	$n = 50$	$n = 100$
Estimate	1.65×10^{-8}	1.93×10^{-20}	2.35×10^{-40}
Relative Error	1.2%	1.4%	1.6%

Table 2. $\lambda_1 = 0.2$, $\lambda_2 = 0.2$, $\mu_1 = 1.0$, $\mu_2 = 1.0$

	$n = 20$	$n = 50$	$n = 100$
Estimate	1.10×10^{-3}	2.85×10^{-7}	3.04×10^{-13}
Relative Error	1.8%	1.9%	1.9%

Table 3. $\lambda_1 = 0.2$, $\lambda_2 = 0.2$, $\mu_1 = 0.6$, $\mu_2 = 0.5$

	$n = 20$	$n = 50$	$n = 100$
Estimate	4.81×10^{-7}	3.08×10^{-16}	2.13×10^{-31}
Relative Error	1.5%	1.8%	2.0%

Table 4. $\lambda_1 = 0.2$, $\lambda_2 = 0.2$, $\mu_1 = 1.0$, $\mu_2 = 0.8$

3.7 Summary

We constructed a piecewise constant change of measure that led to the asymptotic optimality of importance sampling schemes for the total population overflow of a single server queue with priority service policy. The main feature of the queue was the *discontinuity* of its dynamics. The construction of the importance sampling schemes was based on an analysis of a closely related *Hamilton-Jacobi-Bellman equation* and its subsolutions. Finally, a verification argument proved the asymptotic optimality of

the schemes. Numerical simulations were also shown. It is our belief that the results of this paper can be extended to systems with simple structures of discontinuity.

Appendix A

Collection of Proofs (Chapter II)

A.1 Proof of Lemma 2.3.2

Given an arbitrary $\delta > 0$, we need to show that there exists a $\phi^* \in \mathcal{N}$ such that $\|\phi - \phi^*\|_\infty \leq \delta$ and $I_x(\phi^*) \leq I_x(\phi)$. The idea is to approximate ϕ by suitable linear interpolations. We introduce the following notation. Denote by $[\![a, b]\!]$ an interval with end points a and b. The interval can be of any type [open, closed, or half open half closed]. We first introduce the following lemma.

Lemma A.1.1. *Given an arbitrary interval $[\![a, b]\!]$ and any $\sigma > 0$, there exists a finite partition*

$$[\![a, b]\!] = \cup_j [\![\alpha_j, \beta_j]\!]$$

such that for each j

1. $0 \leq \beta_j - \alpha_j \leq \sigma$;
2. $\Pi(\phi(t)) \geq \max\{\Pi(\phi(\alpha_j)), \Pi(\phi(\beta_j))\}$ *for every* $t \in (\alpha_j, \beta_j)$.

Proof: Let $k^* = \min\{\Pi(\phi(t)) : t \in [\![a, b]\!]\}$. Note that the minimum is always attained since $\Pi(\cdot)$ can only take values from $\{0, 1, \ldots, d\}$. We will prove the lemma by backward induction on k^*. The claim is trivial in the case $k^* = d$. Indeed, in order to satisfy Part 1 one can partition the interval $[\![a, b]\!]$ into subintervals of equal length with the length of each subinterval at most σ, while Part 2 holds automatically.

Assume that the lemma holds for $k^* = k + 1, \ldots, d$. We would like to show that it is also valid when $k^* = k$. To ease exposition we assume that $[\![a, b]\!] = [a, b]$ is a closed interval. The proof for other cases is almost verbatim and thus omitted.

It suffices to show that there exists a finite collection of closed intervals $\{[\bar{a}_i, \bar{b}_i]\}$ with non-overlapping interiors such that $0 \leq \bar{b}_i - \bar{a}_i \leq \sigma$, $\Pi(\phi(\bar{a}_i)) = \Pi(\phi(\bar{b}_i)) =$

$k^* = k$, and

$$\min\{\Pi(\phi(t)) : t \in [a,b] \setminus \cup_i [\bar{a}_i, \bar{b}_i]\} \geq k+1.$$

Indeed in this case, by the induction hypothesis, the set $[a,b] \setminus \cup_i [\bar{a}_i, \bar{b}_i]$ which is the union of a finite number of intervals, can be partitioned in a way that Parts 1 and 2 are satisfied. Adding to this partition the collection of closed intervals $\{[\bar{a}_i, \bar{b}_i]\}$, we obtain a desired partition of $[a,b]$ (note that Part 2 is satisfied for interval $[\bar{a}_i, \bar{b}_i]$ by the definition of k^*).

The values of $\bar{a}_i, \bar{b}_i$ are defined recursively as follows. Let

$$\bar{a} = \inf\{t \in [a,b] : \Pi(\phi(t)) = k\}; \quad \bar{b} = \sup\{t \in [a,b] : \Pi(\phi(t)) = k\}.$$

Thanks to the lower semicontinuity of Π, $\Pi(\phi(\bar{a})) = \Pi(\phi(\bar{b})) = k$. Define

$$\begin{aligned} \bar{a}_1 &= \bar{a} \\ \bar{b}_1 &= \sup\{t \in [\bar{a}_1, (\bar{a}_1 + \sigma) \wedge b] : \Pi(\phi(t)) = k\}, \end{aligned}$$

and for $i \geq 1$,

$$\begin{aligned} \bar{a}_{i+1} &= \inf\{t \in [\bar{a}_i + \sigma, b] : \Pi(\phi(t)) = k\} \\ \bar{b}_{i+1} &= \sup\{t \in [\bar{a}_{i+1}, (\bar{a}_{i+1} + \sigma) \wedge b] : \Pi(\phi(t)) = k\}. \end{aligned}$$

The recursion will end if $\bar{b}_N = \bar{b}$ for some N. It is clear that N is finite since $\bar{a}_{i+1} - \bar{a}_i \geq \sigma$. Furthermore, the collection $\{[\bar{a}_i, \bar{b}_i] : i = 1, 2, \ldots, N\}$ clearly has the desired property. This completes the proof. ■

Since $I_x(\phi) < \infty$, ϕ is absolutely continuous and hence uniformly continuous on

$[0,1]$. Therefore, there exists $\sigma > 0$ such that for $s, t \in [0,1]$,

$$|\phi(s) - \phi(t)| \leq \delta, \quad \text{if } |s-t| \leq \sigma.$$

Let $[0,1] = \cup_j [\![\alpha_j, \beta_j]\!]$ be the partition in Lemma A.1.1 with the given σ. Define ϕ^* as the linear interpolation of ϕ from this partition. That is, for every j and every $t \in (\alpha_j, \beta_j)$

$$\dot{\phi}^*(t) = \frac{\phi(\beta_j) - \phi(\alpha_j)}{\beta_j - \alpha_j},$$

and $\phi^*(t) = \phi(t)$ if $t = \alpha_j$ or β_j for some j. Clearly ϕ^* is absolutely continuous and $\|\phi^* - \phi\|_\infty \leq \delta$. It remains to show that $I_x(\phi^*) \leq I_x(\phi)$. Note that for every $t \in (\alpha_j, \beta_j)$,

$$\Pi(\phi^*(t)) = \max\{\Pi(\phi(\alpha_j)), \Pi(\phi(\beta_j))\} \leq \Pi(\phi(t)).$$

Observing that the rate functions $\{\bar{L}_i\}$ are monotonically non-decreasing in that $\bar{L}_0 \leq \bar{L}_1 \leq \cdots \leq \bar{L}_d$, we have

$$\int_{\alpha_j}^{\beta_j} L(\phi(t)), \dot{\phi}(t)) \, dt = \int_{\alpha_j}^{\beta_j} \bar{L}_{\Pi(\phi(t))}(\dot{\phi}(t)) \, dt \geq \int_{\alpha_j}^{\beta_j} \bar{L}_{\Pi(\phi^*(t))}(\dot{\phi}(t)) \, dt.$$

Thanks to the convexity of $\{\bar{L}_i\}$ and Jensen's inequality, it follows that

$$\int_{\alpha_j}^{\beta_j} L(\phi(t)), \dot{\phi}(t)) \, dt \geq (\beta_j - \alpha_j) \bar{L}_{\Pi(\phi^*(t))}(\dot{\phi}^*(t)) \, dt = \int_{\alpha_j}^{\beta_j} L(\phi^*(t)), \dot{\phi}^*(t)) \, dt.$$

This completes the proof. ■

A.2 Proof of Lemma 2.3.3

For any given $\lambda > 0$ and $v \in \mathbb{R}^d$, straightforward calculation yields that the Legendre transform of the convex function $h(\alpha) = \lambda[e^{\langle \alpha, v\rangle} - 1]$ is

$$\begin{aligned} h^*(\beta) &= \sup_{\alpha \in \mathbb{R}^d} [\langle \alpha, \beta \rangle - h(\alpha)] \\ &= \begin{cases} \lambda \ell(\bar{\lambda}/\lambda) & \text{if } \beta = \bar{\lambda} v \text{ for some } \bar{\lambda} \in \mathbb{R}, \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

for every $\beta \in \mathbb{R}^d$. It is now an immediate consequence of [28, Corollary D.4.2] that L_i, the Legendre transform of H_i, has the following alternative representation. That is, for every $\beta \in \mathbb{R}^d$,

$$L_0(\beta) = \inf \left[\sum_{k=1}^{d} \lambda_k \ell \left(\frac{\bar{\lambda}_k}{\lambda_k} \right) : \sum_{k=1}^{d} \bar{\lambda}_k e_k = \beta \right] \tag{A.1}$$

and for $i = 1, \ldots, d$,

$$L_i(\beta) = \inf \left[\mu_i \ell \left(\frac{\bar{\mu}_i}{\mu_i} \right) + \sum_{k=1}^{d} \lambda_k \ell \left(\frac{\bar{\lambda}_k}{\lambda_k} \right) : -\bar{\mu}_i e_i + \sum_{k=1}^{d} \bar{\lambda}_k e_k = \beta \right]. \tag{A.2}$$

We are now in a position to prove the alternative representation for $\bar{L}_i$. Without loss of generality, we assume that $i = 0$. The proof for $i \geq 1$ is similar and thus omitted. Thanks to the definition of $\bar{L}_i$ (2.6) and equations (A.1)-(A.2), we have

$$\begin{aligned} \bar{L}_0(\beta) &= \inf \left\{ \rho_0 \sum_{k=1}^{d} \lambda_k \left(\frac{\bar{\lambda}_k^{(0)}}{\lambda_k} \right) + \sum_{i=1}^{d} \rho_i \left[\mu_i \ell \left(\frac{\bar{\mu}_i^{(i)}}{\mu_i} \right) + \sum_{k=1}^{d} \lambda_k \ell \left(\frac{\bar{\lambda}_k^{(i)}}{\lambda_k} \right) \right] \right\} \\ &= \inf \left\{ \sum_{i=1}^{d} \rho_i \mu_i \ell \left(\frac{\bar{\mu}_i^{(i)}}{\mu_i} \right) + \sum_{i=0}^{d} \rho_i \sum_{k=1}^{d} \lambda_k \ell \left(\frac{\bar{\lambda}_k^{(i)}}{\lambda_k} \right) \right\}, \end{aligned}$$

where the infimum is taken over all $(\rho_i, \bar{\mu}_i^{(i)}, \bar{\lambda}_k^{(i)})$ such that

$$\rho_i \geq 0,\ \sum_{i=0}^{d} \rho_i = 1,\ \rho_0 \sum_{k=1}^{d} \bar{\lambda}_k^{(0)} e_k + \sum_{i=1}^{d} \rho_i \left[-\bar{\mu}_i^{(i)} e_i + \sum_{k=1}^{d} \bar{\lambda}_k^{(i)} e_k \right] = \beta. \tag{A.3}$$

Abusing the notation a bit, write for $k = 1, \ldots, d$,

$$\bar{\mu}_k = \bar{\mu}_k^{(k)}, \qquad \bar{\lambda}_k = \sum_{i=0}^{d} \rho_i \bar{\lambda}_k^{(i)}.$$

Then the constraints (A.3) become

$$\rho_i \geq 0, \quad \sum_{i=0}^{d} \rho_i = 1, \quad -\sum_{k=1}^{d} \rho_k \bar{\mu}_k e_k + \sum_{k=1}^{d} \bar{\lambda}_k e_k = \beta,$$

which are exactly the constraints in the statement of Lemma 2.3.3. Observe that, by the convexity of ℓ,

$$\sum_{i=0}^{d} \rho_i \sum_{k=1}^{d} \lambda_k \ell\left(\frac{\bar{\lambda}_k^{(i)}}{\lambda_k}\right) = \sum_{k=1}^{d} \lambda_k \sum_{i=0}^{d} \rho_i \ell\left(\frac{\bar{\lambda}_k^{(i)}}{\lambda_k}\right) \geq \sum_{k=1}^{d} \lambda_k \ell\left(\frac{\bar{\lambda}_k}{\lambda_k}\right),$$

with equality if $\bar{\lambda}_k^{(i)} = \bar{\lambda}_k^{(j)} = \bar{\lambda}_k$ for every i, j. Furthermore, one can restrict the parameters $\{\bar{\lambda}_k, \bar{\mu}_k : 1 \leq k \leq d\}$ and $\{\rho_k : i \leq k \leq d\}$ to be strictly positive. This is because ℓ is finite and continuous on $[0, \infty)$. The representation for $\bar{L}_i$ now follows readily.

It remains to show that $\bar{L}_i(\beta)$ is finite if and only if $\beta_k \geq 0$ for all $k < i$. This is trivial since the set of $(\rho_k, \bar{\lambda}_k, \bar{\mu}_k)$ that satisfies the constraints is non-empty if and only if $\beta_k \geq 0$ for all $k < i$. Thus completes the proof. ■

A.3 Proof of Lemma 2.4.2

Consider the discrete embedded Markov chain of the state process Q and denote by $\{Z(k) \in \mathbb{Z}_+^2 : k = 0, 1, 2, \ldots\}$ the queue lengths at the transition epochs of the network. Since the process Q starts at the origin, the initial state of the Markov chain Z is $Z(0) = 0$.

We claim that for all k, n, and $z = (z_1, z_2) \in \mathbb{Z}_2^+$ such that $z_1 + z_2 \leq n$

$$E\left[e^{-n[W(Z(k+1)/n)-W(Z(k)/n)]}\middle|\, Z(k) = z, Z(k-1), \ldots, Z(0)\right] \leq e^{M/n} \tag{A.4}$$

for some constant M. We will only show this inequality for the case when $z_2 > 0$. The case where $z_2 = 0$ is similar and thus omitted. Let $x = z/n$ and without loss of generality assume $\lambda_1 + \lambda_2 + \mu_2 = 1$. Since z_2 is strictly positive, $Z(k+1)$ can only take values in the set $\{z + e_1, z + e_2, z - e_2\}$ with respective probabilities $\{\lambda_1, \lambda_2, \mu_2\}$. Therefore, the conditional expectation on the left hand side of (A.4) equals

$$\lambda_1 e^{-n[W(x+e_1/n)-W(x)]} + \lambda_2 e^{-n[W(x+e_2/n)-W(x)]} + \mu_2 e^{-n[W(x-e_2/n)-W(x)]}.$$

Since W is twice continuously differentiable, every component of the Hessian matrix $\nabla^2 W(x)$ is uniformly bounded on the compact set $\{x = (x_1, x_2) : x_i \geq 0, x_1 + x_2 \leq 1\}$. Then by Taylor's expansion we have that

$$|\langle \nabla W(x), v\rangle - n[W(x + v/n) - W(x)]| \leq \frac{M}{n}\|v\|^2$$

for every vector v and some constant M. Therefore, the conditional expectation is

bounded from above by

$$e^{M/n}\left[\lambda_1 e^{-\langle \nabla W(x), e_1\rangle} + \lambda_2 e^{-\langle \nabla W(x), e_2\rangle} + \mu_2 e^{-\langle \nabla W(x), -e_2\rangle}\right].$$

Observe that the sum in the square bracket is exactly $1 + H(-\nabla W(x))$, which is bounded from above by 1, owing to the subsolution property of W. This completes the proof of inequality (A.4).

Fix an arbitrary positive integer n. Define T_n to be the first hitting time to the exit boundary ∂_e:

$$T_n = \inf\{k \geq 0 : Z_1(k) + Z_2(k) = n\}.$$

Define a non-negative process

$$Y^n(k) = e^{-Mk/n - nW(Z(k)/n)}, \quad k = 0, 1, 2, \ldots.$$

It follows from inequality (A.4) that the stopped process $\{Y^n(k \wedge T_n)\}$ is a supermartingale with respect to the natural filtration generated by Z. Let T_0 be the return time to the origin:

$$T_0 = \inf\{k \geq 1 : Z(k) = 0\}.$$

Owing to the Optional Sampling Theorem and the non-negativity of Y^n, we have

$$E[Y^n(T_0 \wedge T_n)] \leq E[Y^n(0)] = e^{-nW(0)}.$$

Furthermore, by the fact that $W(x) \leq 0$ for every $x \in \partial_e$,

$$Y^n(T_0 \wedge T_n) \geq Y^n(T_n)1_{\{T_n < T_0\}} \geq e^{-MT_n/n}1_{\{T_n < T_0\}},$$

and thus

$$E[e^{-MT_n/n}1_{\{T_n<T_0\}}] \leq e^{-nW(0)}.$$

Since the system is exponentially ergodic, there exists a constant $c > 0$ such that $E[e^{cT_0}]$ is finite [2, Lemma 6.3]. Applying Hölder's inequality and observing that any power of an indicator function is still itself, we arrive at

$$\begin{aligned} p_n &= E[1_{\{T_n<T_0\}}] \\ &\leq \left(E[e^{-MT_n/n}1_{\{T_n<T_0\}}]\right)^{\frac{cn}{M+cn}} \left(E[e^{cT_n}1_{\{T_n<T_0\}}]\right)^{\frac{M}{M+cn}}, \\ &\leq e^{-nW(0)\cdot\frac{cn}{M+cn}} \cdot \left(E[e^{cT_0}]\right)^{\frac{M}{M+cn}}. \end{aligned}$$

Taking logarithm on both sides, it follows easily that

$$\liminf_n -\frac{1}{n}\log p_n \geq W(0).$$

This completes the proof. ■

Appendix B

Collection of Proofs (Chapter III)

B.1 Proof of Lemma 3.5.2

It is a standard result that W is a smooth function [35, Section 7.2]. To compute its gradient, we observe that $\bar{W}$ is Lipschitz continuous and thus a standard application of Lebesgue's Dominated Convergence Theorem [27] implies that

$$\nabla W(x) = \int_{\mathbb{R}^2} \rho(y) \nabla \bar{W}(x + \varepsilon_n y) dy.$$

Owing to equation (3.7),

$$\nabla \bar{W}(x) = \begin{cases} -\alpha^{[1]} & \text{if } x_2 > \delta_n, \\ -\alpha^{[2]} & \text{if } x_2 < \delta_n, \end{cases}$$

thus Part 1 of the lemma follows readily. As for Part 2, observe that $\alpha_1^{[1]} = \alpha_2^{[1]} = \gamma$ and therefore for $x = (x_1, x_2)$ such that $x_1 + x_2 \geq 1$

$$\bar{W}(x) \leq W_1(x) = -\langle \alpha^{[1]}, x \rangle + \gamma \leq -\gamma + \gamma = 0.$$

It is easy to check that $\bar{W}$ is Lipschitz continuous with $\|\alpha^{[1]}\|$ as a Lipschitz constant since $\|\alpha^{[1]}\| > \|\alpha^{[2]}\|$, it follows that

$$\bar{W}(x + \varepsilon_n y) \leq \bar{W}(x) + \|\alpha^{[1]}\| \cdot \varepsilon_n \|y\| \leq \|\alpha^{[1]}\| \cdot \varepsilon_n \|y\|$$

and thus

$$\begin{aligned} W(x) &\leq \int_{\mathbb{R}^2} \rho(y) \|\alpha^{[1]}\| \cdot \varepsilon_n \|y\| \, dy - \|\alpha^{[1]}\| \varepsilon_n \\ &\leq \int_{\mathbb{R}^2} \rho(y) \|\alpha^{[1]}\| \cdot \varepsilon_n \, dy - \|\alpha^{[1]}\| \varepsilon_n \\ &= 0. \end{aligned}$$

This proves Part 2 of the lemma. Part 3 is immediate from equation (3.7).

It remains to show Part 4. Owing to Part 1, we only need to show that the gradient of $\theta(x)$ is uniformly bounded by C/ε_n for some constant C independent of n. Clearly, by definition $\partial\theta/\partial x_1 = 0$ and

$$\frac{\partial\theta}{\partial x_2} = \lim_{h\downarrow 0}\frac{1}{h}\int_{\{y=(y_1,y_2):\delta_n - x_2 - h \le \varepsilon_n y_2 < \delta_n - x_2\}} \rho(y)dy.$$

Since ρ is bounded and supported on the unit disc, the integral of the right-hand-side can be at most

$$\|\rho\|_\infty \cdot \text{Area}(\{y=(y_1,y_2) : \delta_n - x_2 - h \le \varepsilon_n y_2 < \delta_n - x_2\} \cap \{\|y\|^2 \le 1\}),$$

which is bounded from above by $\|\rho\|_\infty \cdot 2h/\varepsilon_n$. Therefore

$$\left|\frac{\partial\theta}{\partial x_2}\right| \le \frac{2\|\rho\|_\infty}{\varepsilon_n}.$$

This completes the proof.

B.2 Proof of Lemma 3.5.3

We would like to point out the following trivial lemma, whose proof is straightforward by the strict convexity of H and H_∂, and that $H(0) = H_\partial(0) = 0$.

Lemma B.2.1. *Suppose that α is a non-zero, two-dimensional vector such that $H(\alpha) = 0$. Then*

$$H(k\alpha) < 0$$

for all $k \in (0,1)$. The same result holds if H is replaced by H_∂.

To prove that $\bar{h}_n(x)$ is uniformly bounded from below by a strictly positive constant, we consider the following two scenarios separately.

Case 1: $\alpha_1^* \neq \hat{\alpha}_1$. Without loss of generality, assume $\alpha_1^* < \hat{\alpha}_1$. The proof for the other direction is almost verbatim and thus omitted. In this case, by definition in Section 3.5.3, $\alpha^{[1]} = \alpha^*$ and $\alpha^{[2]} = k\hat{\alpha}$ where $k \in (0,1)$. Note that $H(\alpha^{[2]}) < 0$ and $H_\partial(\alpha^{[2]}) < 0$ by Lemma B.2.1. Let $b = 2$ and $\theta^* = 1/2$. In this case, $x_2^* = 0$ and $W_{\theta^*} = W$. For $x = (x_1, x_2) \in \mathbb{R}_+^2$ such that $x_2 \geq \delta_n$, plugging in the definition of $\bar{r}$ and $\bar{R}$, we have

$$\begin{aligned}\bar{h}_n(x) &\doteq q\sum_{v\in\mathbb{V}} r(x,v)\left[1 - e^{\langle\alpha^*, v\rangle}\right] + \sum_{v\in\mathbb{V}} r(x,v)\left[1 - e^{-2\langle\nabla W(x), v\rangle - q\langle\alpha^*, v\rangle}\right] \\ &= -qH(\alpha^*) - H(-2\nabla W(x) - q\alpha^*).\end{aligned}$$

According to Part 1 of Lemma 3.5.2 and observing $H(\alpha^*) = 0$, we arrive at

$$\bar{h}_n(x) = -H(2\theta(x)\alpha^* + 2(1-\theta(x))\alpha^{[2]} - q\alpha^*).$$

If we regard the right-hand-side as a function of $\theta = \theta(x)$ and denote it by $F(\theta)$, then F is concave since H is convex. Note that for $x_2 \geq \delta_n$, $\theta = \theta(x) \in [1/2, 1]$. If $\theta = 1$, then for $q \in (1,2)$ the right-hand-side is

$$F(1) = -H((2-q)\alpha^*) > 0,$$

again owing to Lemma B.2.1 and that $0 < 2-q < 1$. For $\theta = 1/2$, the right-hand-side equals

$$F(1/2) = -H((1-q)\alpha^* + \alpha^{[2]}).$$

This term is again strictly positive for any q which is sufficiently close to 1, owing to

the continuity of H and that $H(\alpha^{[2]}) < 0$. Since F is concave, it is easy to see that

$$\min_{\theta \in [1/2,1]} F(\theta) = \min\{F(1/2), F(1)\}.$$

Therefore, for all $x = (x_1, x_2)$ such that $x_2 \geq \delta_n$ and all $q > 1$ but sufficiently close to 1, F or $\bar{h}_n$ is uniformly bounded from below by a positive constant independent of n .

One can similarly argue for $x = (x_1, x_2) \in \mathbb{R}^2_+$ such that $0 < x_2 < \delta_n$. Indeed, for such x analogous calculation yields

$$\bar{h}_n(x) = -qH(\alpha^{[2]}) - H(2\theta(x)\alpha^* + 2(1-\theta(x))\alpha^{[2]} - q\alpha^{[2]}),$$

and $\theta(x) \in [0, 1/2]$. For $\theta(x) = 1/2$, the right-hand-side equals

$$-qH(\alpha^{[2]}) - H(\alpha^* + (1-q)\alpha^{[2]}).$$

Owing to the fact that $H(\alpha^{[2]}) < 0$, $H(\alpha^*) = 0$, and the continuity of H, this term is strictly positive when q is close to 1. For $\theta(x) = 0$, the right-hand-side equals

$$-qH(\alpha^{[2]}) - H((2-q)\alpha^{[2]}),$$

which is again strictly positive for $q \in (1, 2)$ since so is each summand [by Lemma B.2.1]. The uniform positivity of $\bar{h}_n$ for x such that $0 < x_2 < \delta_n$ follows analogously.

It remains to show the claim for $x = (x_1, x_2) \in \mathbb{R}^2_+$ such that $x_2 = 0$. For those x, it is easy to check that $\theta(x) = 0$ from definition [Lemma 3.5.2 Part 1] and that

$\varepsilon_n = \delta_n/2 < \delta_n$, which in turn implies that

$$\bar{h}_n(x) = -qH_\partial(\alpha^{[2]}) - H_\partial(2\alpha^{[2]} - q\alpha^{[2]}).$$

This is strictly positive for $q \in (1,2)$ since each summand is strictly positive by Lemma B.2.1, and therefore the proof is complete.

Case 2: $\alpha_1^* = \hat{\alpha}_1$. In this case $\alpha^{[1]} = \alpha^*$ and $\alpha^{[2]} = \hat{\alpha}$. For any $\varepsilon \in (0,1)$ we can select $b \in (2-\varepsilon, 2)$ such that

$$-H(\alpha^* + (b-2)\hat{\alpha}) > 0. \tag{B.1}$$

This is always possible since $H(\alpha^*) = 0$ and $\langle \nabla H(\alpha^*), \hat{\alpha} \rangle > 0$ [we omit the straightforward but tedious technical proof that both $\nabla H(\alpha^*)$ and $\hat{\alpha}$ live in the positive orthant, which is fairly obvious from Figure 3]. Define $\theta^* \doteq 1/b$, then $\theta^* > 1/2$. We will also assume that $x_2^* \geq -1$. This assumption is without loss of generality — If necessary, one can make b very close to 1 so that θ^* is very close to $1/2$. Then $x_2^* = x_2^*(\theta^*)$ is very close to 0 and $x_2^* \geq -1$ holds automatically. The rest of the proof is similar to the previous case, and thus we only give outlines.

For $x = (x_1, x_2) \in \mathbb{R}_+^2$ such that $x_2 \geq \delta_n$, it follows from straightforward calculations that

$$\begin{aligned} \bar{h}_n(x) &= q \sum_{v \in \mathbb{V}} r(x,v) \left[1 - e^{\langle \alpha^*, v \rangle}\right] + \sum_{v \in \mathbb{V}} r(x,v) \left[1 - e^{-b\langle \nabla W_{\theta^*}(x), v \rangle - q\langle \alpha^*, v \rangle}\right] \\ &= -qH(\alpha^*) - H(-b\nabla W_{\theta^*}(x) - q\alpha^*). \end{aligned}$$

Since $H(\alpha^*) = 0$, it follows from Lemma 3.5.2 Part 1 and the definition of W_{θ^*} (3.8)

$$\bar{h}_n(x) = -H(b\theta\alpha^* + b(1-\theta)\hat{\alpha} - q\alpha^*), \tag{B.2}$$

where $\theta = \theta(x_1, x_2 - \varepsilon_n x_2^*)$. Note that for $x_2 = \delta_n$, by definition (3.9) and Lemma 3.5.2 Part 1

$$\begin{aligned}
\theta(x_1, \delta_n - \varepsilon_n x_2^*) &= \int_{\{y=(y_1,y_2):\varepsilon_n y_2 > \delta_n - (\delta_n - \varepsilon_n x_2^*)\}} \rho(y)dy \\
&= \int_{\{y=(y_1,y_2):y_2 > x_2^*\}} \rho(y)dy \\
&= \theta^*.
\end{aligned}$$

It follows that for $x_2 \geq \delta_n$, $\theta \in [\theta^*, 1]$. Again the right-hand-side of (B.2) as a function of θ, is concave. It takes value $-H((b-q)\alpha^*)$ at $\theta = 1$, which is strictly positive for $q \in (1, b)$, thanks to Lemma B.2.1 and that $0 < b - q < 1$. At $\theta = \theta^*$, it equals $-H(\alpha^* + (b-1)\hat{\alpha} - q\alpha^*)$. This is strictly positive for $q > 1$ but sufficiently close to 1, since H is continuous and $-H((b-1)\hat{\alpha}) > 0$ [again thanks to Lemma B.2.1 and $0 < b - 1 < 1$]. The uniform positivity of $\bar{h}_n(x)$ for those x follows by an analogous argument using concavity with respect to θ.

Now let us consider those $x = (x_1, x_2) \in \mathbb{R}_+^2$ such that $0 < x_2 < \delta_n$. Similarly, we have

$$\bar{h}_n(x) = -H(b\theta\alpha^* + b(1-\theta)\hat{\alpha} - q\hat{\alpha}),$$

where $\theta = \theta(x_1, x_2 - \varepsilon_n x_2^*) \in [0, \theta^*]$. For $\theta = \theta^* = 1/b$, the right-hand-side equals

$$-H(\alpha^* + (b-1)\hat{\alpha} - q\hat{\alpha}).$$

This is strictly positive for q sufficiently close to 1, owing to inequality (B.1) and

the continuity of H. For $\theta = 0$, the right-hand-side equals $-H((b-q)\hat{\alpha}) > 0$ for $q \in (1, b)$. The uniform positivity again follows in the same fashion.

It remains to show for those $x = (x_1, x_2) \in \mathbb{R}^2_+$ such that $x_2 = 0$. Straightforward calculation yields that

$$\bar{h}_n(x) = -H_\partial(b\theta\alpha^* + b(1-\theta)\hat{\alpha} - q\hat{\alpha}),$$

where $\theta = \theta(x_1, x_2 - \varepsilon_n x_2^*) = \theta(x_1, -\varepsilon_n x_2^*)$. Note that by Lemma 3.5.2 Part 1 and $\varepsilon_n = \delta_n/2$,

$$\begin{aligned} \theta(x_1, -\varepsilon_n x_2^*) &= \int_{\{y=(y_1,y_2):\varepsilon_n y_2 > \delta_n + \varepsilon_n x_2^*\}} \rho(y)dy \\ &= \int_{\{y=(y_1,y_2):y_2 > 2 + x_2^*\}} \rho(y)dy \\ &= 0. \end{aligned}$$

The last equality is because $2 + x_2^* \geq 1$ and ρ is supported in the unit disc. It follows that

$$\bar{h}_n(x) = -H_\partial((b-q)\hat{\alpha}) > 0$$

for $q \in (1, b)$, thanks to Lemma B.2.1. This completes the proof.

Bibliography

[1] M. Alanyali and B. Hajek. On large deviations of Markov processes with discontinuous statistics. *Ann. Appl. Probab.*, 8:45–66, 1998.

[2] W.J. Anderson. *Continuous Time Markov Chains*. Springer-Verlag, New York, 1991.

[3] R. Atar and P. Dupuis. Large deviations and queueing networks: methods for rate function identification. *Stoch. Proc. and Their Appl.*, 84:255–296, 1999.

[4] A.A. Borovkov and A.A. Mogulskii. Large deviations for Markov chains in the positive quadrant. *Russian Math. Surveys*, 56:803-916, 2001.

[5] H. Chen and D. Yao. *Fundamentals of Queueing Networks*. Springer, New York, 2001.

[6] T.S Chiang and S.J. Sheu. Large deviation of diffusion processes with discontinuous drift and their occupation times. *Ann. Probab.*, 28:140–165, 2000.

[7] P. Dupuis and R. S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. John Wiley & Sons, New York, 1997.

[8] P. Dupuis, R.S. Ellis, and A. Weiss. Large deviations for Markov processes with discontinuous statistics, I: General upper bounds. *Ann. of Probab.*, 19:1280–1297, 1991.

[9] P. Dupuis, H. Ishii, and H. M. Soner. A viscosity solution approach to the asymptotic analysis of queueing systems. *Ann. Probab.*, 18:226–255, 1990.

[10] P. Dupuis, K. Leder, and H. Wang. On the large deviations properties of the weighted-serve-the-longest-queue policy. In V. Sidoravicius and M.E. Vares, editors, *In and Out of Equilibrium 2*. Birkhauser, New York, 2008.

[11] P. Dupuis, K. Leder, and H. Wang. Importance sampling for weighted serve-the-longest-queue. *Math. of Operations Research*, 34:642–660, 2009.

[12] S.N. Ethier and T.G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley, New York, 1986.

[13] T.R. Fleming and D.R. Harrington. *Counting Processes and Survival Analysis*. John Wiley & Sons, New York, 1991.

[14] R. Foley and D. McDonald. Join the shortest queue: stability and exact asymptotics. *Ann. Appl. Probab.*, 11:569–607, 2001.

[15] M. I. Freidlin and A. D. Wentzell. *Random Perturbations of Dynamical Systems*. Springer-Verlag, New York, 1984.

[16] D. Gilbarg and N. S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Springer–Verlag, Berlin, second edition, 1983.

[17] I. Ignatiouk-Robert. Sample path large deviations and convergence parameters. *Ann. Appl. Prob.*, 11:1292–1329, 2001.

[18] I. Ignatiouk-Robert. Large deviations for processes with discontinuous statistics. *Ann. Probab.*, 33:1479–1508, 2005.

[19] J.L. Kelly. *General Topology*. Springer, New York, 1951.

[20] A. Puhalskii and A. Vladimirov. A large deviation principle for join the shortest queue. *Math. Oper. Res.*, 32:700–710, 2007.

[21] K. Ramanan and S. Stolyar. Largest weighted delay first scheduling: Large deviations and optimality. *The Annals of Applied Probab.*, 11:1–49, 2001.

[22] A. Shwartz and A. Weiss. *Large Deviations for Performance Analysis: Queues, Communication and Computing*. Chapman and Hall, New York, 1995.

[23] S. Asmussen. *Ruin Probabilities*. World Scientific, Singapore, 2000.

[24] J.H. Blanchet and P. Glynn. Efficient rare-event simulation for the maximum of heavy-tailed random walks. *Ann. Appl. Prob.*, 18:1351–1378, 2008.

[25] J.H. Blanchet, P. Glynn, and J.C. Liu. Fluid heuristics, Lyapunov bounds and efficient importance sampling for a heavy-tailed G/G/1 queue. *QUESTA*, 57:99–113, 2007.

[26] P.T. De Boer. Analysis of state-independent importance sampling measures for the two-node tandem queue. *ACM Trans. Modeling Comp. Simulation*, 16:225–250, 2006.

[27] K.-L. Chung. *A Course in Probability Theory*. Academic Press, New York, second edition, 1974.

[28] P. Dupuis and R. S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. John Wiley & Sons, New York, 1997.

[29] P. Dupuis, R.S. Ellis, and A. Weiss. Large deviations for Markov processes with discontinuous statistics, I: General upper bounds. *Annals of Probability*, 19:1280–1297, 1991.

[30] P. Dupuis, K. Leder, and H. Wang. Large deviations and importance sampling for a tandem network with slow-down. *QUESTA*, 57:71–83, 2007.

[31] P. Dupuis, K. Leder, and H. Wang. Importance sampling for weighted serve-the-longest-queue. *Math. of Operations Research*, 34:642–660, 2009.

[32] P. Dupuis and H. Wang. Importance sampling, large deviations, and differential games. *Stoch. and Stoch. Reports.*, 76:481–508, 2004.

[33] P. Dupuis and H. Wang. Subsolutions of an Isaacs equation and efficient schemes for importance sampling. *Math. Oper. Res.*, 32:1–35, 2007.

[34] P. Dupuis and H. Wang. Importance sampling for Jackson networks. *Queueing Systems*, 62:113–157, 2009.

[35] D. Gilbarg and N. S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Springer–Verlag, Berlin, second edition, 1983.

[36] P. Glasserman and Y. Wang. Counter examples in importance sampling for large deviations probabilities. *Ann. Appl. Prob.*, 7:731–746, 1997.

[37] P.W. Glynn and D.L. Iglehart. Simulation methods for queues: an overview. *Queueing Systems: Theory and Applications*, 3:221–256, 1988.

[38] P. Heidelberger. Fast simulation of rare events in queueing and reliability models. *ACM Trans. Modeling Comp. Simulation*, 4:43–85, 1995.

[39] S. Juneja and V. Nicola. Efficient simulation of buffer overflow probabilities in Jackson networks with feedback. *ACM Transactions on Modelling and Computer Simulation*, 15:281–315, 2005.

[40] D.I. Miretskiy, W.R.W. Scheinhardt, and M.R.H. Mandjes. Efficient simulation of a tandem queue with server slowdown. *Simulation*, 83:751–767, 2007.

[41] S. Parekh and J. Walrand. A quick simulation method for excessive backlogs in networks of queues. *IEEE. Trans. Autom. Control*, 34:54–66, 1989.

[42] J.S. Sadowsky. On Monte Carlo estimation of large deviations probabilities. *Ann. Appl. Prob.*, 6:399–422, 1996.

[43] L. Setayeshgar and H. Wang. Large deviations for a feed-forward network. *Submitted*, 2010.

[44] P. Shahsbuddin. Importance sampling for the simulation of highly reliable Markovian systems. *Management Sciences*, 40:333–352, 1994.

[45] D. Siegmund. Importance sampling in the Monte Carlo study of sequential tests. *Ann. Statist.*, 4:673–684, 1976.