Large Deviations and Importance Sampling for Queuing Systems with Discontinuous Service Policies

by

Kevin Z. Leder

Sc.M., Applied Mathematics, Brown University, 2005
B.E., Applied Mathematics, University of Colorado, 2002

Thesis

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

May 2008

This dissertation by Kevin Z. Leder 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, Director

Recommended to the Graduate Council

Date ....
Paul Dupuis, Reader

Date ....
Henrik Hult, Reader

Approved by the Graduate Council

Date ....
Sheila Bonde
Dean of the Graduate School

The Vita of Kevin Z. Leder

Born on November 14, 1979 in Denver, Colorado.

Education

- Sc.M. Applied Mathematics, Brown University, 2005
- B.E. Applied Mathematics, University of Colorado, Boulder, Colorado, 2002

Publications

Articles in Refereed Journals:

1. Dupuis, P., Leder, K., Wang, H. 'Large deviations and importance sampling for a tandem network with slow-down.' Accepted to QUESTA.
2. Dupuis, P., Leder, K., Wang, H. 'On the Large Deviations Properties of the Weighted-Serve-the-Longest Queue Policy.' Accepted to Birkhauser series "Progress in Probability".
3. Dupuis, P., Leder, K., Wang, H. 'Importance sampling for sums of random variables with regularly varying tails.' ACM Transactions on Machine and Computer Simulation. Volume 17, No. 14, 2007.

Acknowledgments

I would like to begin by thanking my advisors Hui Wang and Paul Dupuis, their guidance has been essential to the completion of my thesis. I would also like to thank Henrik Hult for serving on my committee, as well as all of the helpful advice he has given me during my time at Brown. In addition I would like to thank Carl Nuzman and Phil Whiting of Bell Labs for all the helpful advice they gave me. My time at Brown has been greatly improved by the wonderful staff and faculty of the Division of Applied Math.

As always I thank my family for their loving support. Lastly I would like to thank my wife Jasmine, whose love and support helped me finish this thesis.

Contents

Acknowledgments **v**

1 Introduction **1**

1.1 Characterization of Rare Events via Large Deviations 2

1.2 Importance Sampling 4

1.2.1 Importance Sampling for Discrete Markov Processes in Continuous Time 6

1.2.2 History of Importance Sampling 8

1.3 Layout of Thesis 12

2 A Large Deviations Principle for the Weighted-Serve-the-Longest-Queue Policy **13**

2.1 Introduction 13

2.2 Problem Setup 16

2.3 The main result 18

2.3.1 The rate function 19

2.3.2 The main theorem 20

2.4 Proof of the main theorem 21

2.4.1 An approximation lemma 21

2.4.2 A stochastic control representation 22

2.4.3 Properties of the rate function 23

2.4.4 The construction of controls and the cost 26
2.4.5 Weak convergence analysis 28
2.4.6 Stability analysis 31
2.4.7 Analysis of the cost 40
2.5 Summary 41
2.6 Proof of Lemma 2.4.1. 42
2.6.1 Dividing the time interval 42
2.6.2 Construction of the approximating function 48
2.6.3 The analysis of the rate function 49

3 Importance Sampling and Large Deviations Decay Rate for a Buffer Overflow Event in the WSLQ Policy 55

3.1 Introduction 55
3.2 Problem Formulation 56
3.2.1 System Dynamics 58
3.3 Large Deviations 59
3.3.1 The exponential decay rate 59
3.4 The Isaacs Equation 61
3.5 Classical Subsolutions 65
3.5.1 Piecewise affine subsolutions 66
3.5.2 Mollification 67
3.5.3 Importance sampling algorithm 68
3.6 Asymptotic Optimality 70
3.7 Numerical Experiments 77
3.8 Proofs of Several Results 79
3.8.1 Proof of Lemma 3.4.1 79
3.8.2 Proof of Proposition 3.4.2 81
3.8.3 Proof of Lemma 3.5.2. 82

3.8.4 Proof of Lemma 3.6.1 85
3.8.5 Proof of Theorem 3.3.2 87

4 Importance Sampling and Large Deviations for a Buffer Overflow Event in a Tandem Queue with Server Slowdown 92

4.1 Introduction 92
4.2 The model setup 94
4.3 The system dynamics 95
4.4 A representation of the exponential decay rate 97
4.5 The Isaacs equation 101
4.6 Subsolutions and an importance sampling algorithm 104
4.6.1 Piecewise affine subsolutions 104
4.6.2 The mollification 108
4.6.3 The importance sampling algorithm 109
4.7 The main results 111
4.8 Numerical results 114
4.9 Proof of Lemma 4.6.1 117
4.10 Exponential bounds on T_N 118

List of Tables

3.1 Probability of buffer overflow in two dimensional case. 78

3.2 Probability of buffer overflow in 4 dimensional case. 78

3.3 Probability of buffer overflow in 6 dimensional case. 79

4.1 Probability of Buffer 2 Overflow, First Example 116

4.2 Probability of Buffer 2 Overflow, Second Example 116

List of Figures

2.1 WSLQ system 16

2.2 System dynamics for $d = 2$. 18

4.1 The system dynamics 94

4.2 Notation for the State Space Partition 97

4.3 The gradients of the subsolution 106

4.4 The subsolution 108

Chapter 1

Introduction

In many real world systems there are rare but critical events, whose occurrence can be catastrophic. In order to protect against these events it is necessary to have some estimate for their likelihood. It is very often the case that the only way to obtain these estimates is via computer simulation. A naive approach to estimating unknown probabilities via simulation is the standard Monte Carlo algorithm, however it will turn out that this approach has some major defects.

The standard Monte Carlo algorithm is best explained via an example. Suppose for the random variable X we are interested in estimating $P(X \in A)$ for some set A in the range of X. First one chooses a large integer N and then using a pseudo-random number generator creates a vector $(X_1, \ldots, X_N)$ of independent random variables all with the same distribution as X. For future reference a sequence of independent identically distributed random variables will be denoted by i.i.d. Then the quantity

$$\hat{p}_N(A) = \frac{1}{N}\sum_{j=1}^{N} 1_A(X_j)$$

will serve as an estimator of the probability $P(X \in A)$. Note that the estimator is unbiased i.e. $E[\hat{p}_N(A)] = P(X \in A)$. Another useful property of the standard Monte Carlo estimator is that it is strongly consistent, i.e. that $\hat{p}_N(A)$ converges almost surely to $P(X \in A)$ as $N \to \infty$. Unfortunately the above two properties do not guarantee good performance when estimating the probability of rare events.

It is obvious that when trying to estimate a small quantity the relative error should be more important than the absolute error. The relative error of the Monte Carlo estimator is given by

$$\mathrm{RE}(\hat{p}_N(A)) = \frac{\text{std. deviation of } \hat{p}_N(A)}{\text{mean of } \hat{p}_N(A)} = \sqrt{\frac{1 - P(X \in A)}{NP(X \in A)}} \approx \frac{1}{\sqrt{NP(X \in A)}}.$$

Therefore in order to control the size of the relative error it is necessary that N is at least as big as, if not bigger than, the inverse of the probability of interest. With this in mind it is readily apparent that Monte Carlo is not a computationally feasible approach to estimating the probability of rare events.

1.1 Characterization of Rare Events via Large Deviations

Before continuing in the discussion of estimation of rare events it is necessary to establish a mathematical definition of rare events. The work in this thesis is concerned with events $\{A_n\}_{n \geq 1}$ in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that satisfy the following exponential decay for some $\gamma > 0$

$$\lim_{n \to \infty} -\frac{1}{n} \log \mathbb{P}(A_n) = \gamma. \tag{1.1.1}$$

Thus if we are interested in estimating $\mathbb{P}(X \in A) \approx 0$ we can think of the probability as part of an infinite sequence of probabilities going to 0 exponentially fast. It should be noted that it is possible for rare events to decay with subexponential decay rates, see e.g. [10].

Finding the exponential decay rate γ requires a fair amount of work in itself. We say a sequence of random variables $\{X^n\}$ taking values in a Polish space $\mathcal{X}$ satisfy a *large deviations principle* (LDP) with good rate function $I : \mathcal{X} \to [0, \infty]$ if the following two conditions hold

1. I has compact level sets;

2. $-\inf_{x\in A^0} I(x) \leq \liminf_n \mathbb{P}(X^n \in A) \leq \limsup_n \mathbb{P}(X^n \in A) \leq -\inf_{x\in\bar{A}} I(x)$,

where A is any Borel set contained in $\mathcal{X}$, and A^0 and $\bar{A}$ respectively denote the interior and closure of the set A with respect to the metric of the Polish space. The function I gives information about the exponential rate of decay for sets in the state space of the random variables $\{X^n\}$. Thus finding the exponential decay rate is reduced to minimizing the rate function I over the set of interest A. Of course this can itself be a serious chore in many cases, e.g. when the random variables X^n take values in some infinite dimensional space.

Thus before discussing effective rare event estimation algorithms two tasks must be performed. First it is necessary to find the correct rate function for the rare events of interest. Second the minimization of the rate function over the relevant region performed. Solving these issues has a long history going back several decades to Cramer's work estimating the probability of ruin for insurance companies receiving claims according to a stochastic process.

Remark 1.1.1. A very useful piece of qualitative information available from the LDP is how the rare event occurs. In particular consider a closed set $A \subset \mathcal{X}$ such that $\inf_{x\in A^0} I(x) = \inf_{x\in A} I(x) \in (0, \infty)$ and let $x^* = \text{arginf}_{x\in A} I(x)$ be the unique minimizer, then

$$\lim_{n\to\infty} \mathbb{P}\left(d(X^n, x^*) > \delta | X^n \in A\right) = 0.$$

Therefore given that the rare event $\{X^n \in A\}$ occurs, we know that the random variable is very close to the minimizer x^* with high probability. Thus information about the conditional distribution of X^n given the rare event occurs is available from the LDP.

1.2 Importance Sampling

As seen earlier the standard Monte Carlo algorithm fails as an estimation algorithm when dealing with small probabilities. The basic reason for its failure is that it will take a very large number (order of the inverse of the probability) of replications to see an occurrence of the rare event. The idea of importance sampling is to sample the random variable according to a new distribution so that the rare event is no longer so rare.

Suppose that were interested in estimating $p_n \doteq \mathbb{P}(X^n \in A)$ for integer values of n. The creation of an importance sampling estimator begins with the choice of a sampling measure $\mathbb{Q}_n$ that satisfies the property $\mathbb{P} \ll \mathbb{Q}_n$. Then the estimator is formed by averaging independent replications of

$$\hat{p}_n = \frac{d\mathbb{P}}{d\mathbb{Q}_n} 1_A(X^n), \tag{1.2.1}$$

where $d\mathbb{P}/d\mathbb{Q}_n$ is the Radon-Nikodym derivative of the original measure with respect to the sampling measure. Note that due to the Radon-Nikodym derivative all importance sampling estimators are unbiased. Also when studying the properties of the importance sampling estimator it suffices to study the properties of the single replicate $\hat{p}_n$.

Clearly the problem with the standard Monte Carlo estimator is that its variance is too large relative to the probability of interest. Thus the goal of importance sampling is to choose a sampling measure $\mathbb{Q}_n$ that creates an estimator with sufficiently small variance, and since all importance sampling measures are unbiased it suffices to minimize the second moment. Ideally it would be nice to find the minimizing argument in the following

$$\min_{\mathbb{Q}_n} E^{\mathbb{Q}_n}\left[\hat{p}_n^2\right].$$

In fact the minimizing sampling measure is $\mathbb{Q}_n(\cdot) = \mathbb{P}(\cdot|X^n \in A)$, which will produce an unbiased and zero variance estimator, but unfortunately use of this sampling measure requires knowledge of the unknown quantity $\hat{p}_n$. This result does tells us two things. First searching for the minimizing change of measure for fixed n is not going to work. Second that if a sampling measure that effectively approximates the above conditional measure can be constructed then it should work well. Of course the latter statement will depend on how well the sampling measure approximates the conditional measure.

From the previous paragraph it is clear that choosing the optimal measure for each fixed integer n is not a promising approach to the problem of finding importance sampling changes of measure. Thus the problem is to find a sequence of sampling measures $\mathbb{Q}_n$ that satisfy some asymptotic optimality. As mentioned earlier (1.1.1), we are interested in probabilities that have a positive exponential decay rate γ. Therefore, Jensen's inequality and the unbiased property show that for any sequence of sampling measures $\mathbb{Q}_n$

$$\liminf_n \frac{1}{n} \log E^{\mathbb{Q}}\left[\hat{p}_n^2\right] \geq \liminf_n \frac{1}{n} \log\left(E^{\mathbb{Q}}\hat{p}_n\right)^2 = 2\liminf_n \frac{1}{n}\log p_n = -2\gamma.$$

In other words the fastest possible exponential decay rate for an importance sampling estimator is twice the exponential decay rate of the probabilities. From this result comes the following criteria which will be used throughout much of the rest of the thesis.

Definition 1.2.1. *A sequence of sampling measures $\mathbb{Q}_n$ and the associated importance sampling estimator $\hat{p}_n$ are said to be* ***asymptotically optimal*** *for the sequence of unknown probabilities p_n if*

1. $\lim_{n\to\infty} \frac{1}{n} \log p_n = -\gamma$,

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

When trying to estimate rare event probabilities that decay exponentially fast using standard Monte Carlo the number of replications needed to achieve a given level of relative error will grow exponentially with the scaling parameter n. When performing the same estimation using *asymptotically optimal* importance sampling estimators the number of replications needed to achieve a given level of relative error will now grow subexponentially, which leads to extreme reductions in computation time.

1.2.1 Importance Sampling for Discrete Markov Processes in Continuous Time

All of the work in this thesis will look at Markov processes that have a discrete set of possible jumps that can be made at any time. All the rare events considered in this thesis can be classified as buffer overflow during a busy cycle. A more abstract characterization of this event is that there is a Markov process Q that starts at the origin and a set B a positive distance from the origin, and the dynamics of the process Q push it back towards the origin. Let $T_1, T_2, \ldots$ be the jump times of Q with the convention $T_0 = 0$, and define the hitting times

$$N^0 \doteq \inf\{j \geq 1 : Q(T_j) = 0\} \tag{1.2.2}$$

$$N^n \doteq \inf\{j \geq 1 : Q(T_j) \in nB\}. \tag{1.2.3}$$

Then the rare events probabilities studied in this thesis can be written as

$$\mathbb{P}_0\left(N^n < N^0\right).$$

In order to discuss importance sampling it is first necessary to give a precise mathematical characterization of the Markov jump process Q. For each possible jump direction v and point x in the state space let $r(x,v)$ be the jump rate in the direction v. Also define the total intensity at the point x by

$$R(x) = \sum_{v \in \mathbb{V}} r(x,v),$$

where $\mathbb{V}$ is the space of possible jumps. Then the process Q is characterized by the stochastic transition kernel

$$\begin{aligned} \Theta\left[dt, v|v\right] &\doteq \mathbb{P}\left\{T_{j+1} - T_j \in dt, Q(T_{j+1}) = nx + v | Q(T_j) = nx\right\} \\ &= r(x,v)e^{-R(x)t}dt, \end{aligned} \tag{1.2.4}$$

where $n \in \mathbb{N}$ and $x \in \mathbb{R}^d_+$ are such that $nx \in \mathbb{Z}^d_+$. The idea of importance sampling in these situations is to consider the process Q under a new measure $\mathbb{Q}$ such that its transition kernel now has the form

$$\mathbb{Q}\left\{T_{j+1} - T_j \in dt, Q(T_{j+1}) = nx + v | Q(T_j) = n\right\} = \bar{\Theta}\left[dt, v|x\right].$$

In order to define the importance sampling estimator it is necessary to define the sojourn times $s_j = T_j - T_{j-1}$ and the increments $v_j = Q(T_j) - Q(T_{j-1})$. Then a single sample of the importance sampling estimator is defined as

$$\hat{p}_n \doteq 1_{\{N^n < N^0\}} \prod_{j=1}^{N^n} \frac{\Theta\left[ds_j, v_j | Q(T_{j-1})/n\right]}{\bar{\Theta}^n\left[ds_j, v_j | Q(T_{j-1})/n\right]}. \tag{1.2.5}$$

Remark 1.2.2. The likelihood ratio of $\hat{p}_n$ is with respect to continuous-time sample paths. However, most importance sampling literature uses the likelihood ratio defined on the embedded discrete-time sample paths, or more precisely, a conditional expectation of the continuous-time likelihood ratio. To illustrate, consider the case where

$$\bar{\Theta}^n[dt, v|x] = \bar{r}(x,v)e^{-\bar{R}(x)t}dt, \quad \text{where } \bar{R}(x) = \sum_{v\in\mathbb{V}} \bar{r}(x,v). \tag{1.2.6}$$

Under this transition kernel, the process Q is again a continuous time Markov process with $\bar{r}(x,v)$ as the jump intensity from state nx to $nx+v$. Then the importance sampling estimator based on the embedded discrete time Markov chain $Z = \{Z(i) = Q(T_i) : i \geq 0\}$ is

$$\bar{p}_n \doteq 1_{\{N^n<N^0\}} \prod_{j=0}^{N^n} \frac{r[Z(j)/n, \triangle_{j+1}]/R(Z(j)/n)}{\bar{r}[Z(j)/n, \triangle_{j+1}]/\bar{R}(Z(j)/n)}.$$

Recalling the definitions of Θ and $\bar{\Theta}$ in equations (1.2.4) and (1.2.6)

$$E^{\mathbb{Q}}[\hat{p}_n|Q(T_1),\ldots,Q(T_N)] = \bar{p}_n. \tag{1.2.7}$$

This implies that $\bar{p}_n$ is unbiased and its second moment is at most that of $\hat{p}_n$. It is important to distinguish these two importance sampling estimators. $\hat{p}_n$ is more suitable for analysis, and $\bar{p}_n$ is easier for implementation. We should analyze the change of measure using $\hat{p}_n$, and use its discrete-time counterpart $\bar{p}_n$ for implementation. Note that if one can establish the asymptotic optimality of $\hat{p}_n$, then $\bar{p}_n$ is automatically asymptotically optimal.

1.2.2 History of Importance Sampling

The use of importance sampling dates back to at least the 1950's. The first major theoretical result in the field was the 1976 paper [26] by David Siegmund. In this work he considered estimating the probability of ruin via an importance sampling estimator. More specifically if $\{Y_i\}$ is an i.i.d sequence with negative mean and

$$S_n = \sum_{i=1}^{n} Y_i$$

is the associated random walk the goal was to find an importance sampling change of measure for estimating

$$\mathbb{P}\left(\sup_{n \geq 1} S_n > u\right)$$

for large u. Siegmund's idea was to sample a new random walk whose increments were an exponential twist of the original increments.

In order to better illustrate the connection with the Large Deviations theory we will look at the change of measure Siegmund proposed for estimating ruin probabilities in finite time. Let $\{Y_i\}$ be a sequence of i.i.d random variables which induce the measure μ on $\mathbb{R}$. The problem is to estimate the probability

$$p_n = \mathbb{P}\left(\frac{S_n}{n} > a\right), \quad \text{where } E(Y_i) < a$$

for large n. Siegmund's idea was to sample a random walk $\bar{S}_n$ with increments $\bar{Y}_i$ whose distribution on $\mathbb{R}$ under $\mathbb{P}$ is given by an exponential twist of the original distribution

$$\nu(dy) = \exp\left(\theta_a y - H(\theta_a)\right) \mu(dy)$$

where $H(\theta) = \log E e^{\theta Y}$. A question that still remains is what does θ_a mean? From Cramer's Theorem we know that $\{S_n/n\}$ satisfies an LDP with rate function

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

Therefore we know that

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\left(\frac{S_n}{n}>a\right)=-\inf_{\beta>a}L(\beta).$$

A simple result of convex analysis shows that the above infimum is achieved at $\beta=a$ and that $L(a)=\langle\theta_a,a\rangle-H(\theta_a)$ where $H'(\theta_a)=a$. Note that

$$E\left[\bar{Y}_i\right]=e^{-H(\theta_a)}E\left[Ye^{\theta_a Y}\right]=H'[\theta_a]=a.$$

The importance sampling estimator for p_n is obtained by averaging indepedent replications of

$$\hat{p}_n=1_{\{\bar{S}_n>na\}}e^{n\left(H(\theta_a)-\theta_a\bar{S}_n/n\right)}.$$

Thus Siegmund's importance sampling change of measure performs an exponential twist on the increments so that their new mean is a. Note that this point is the minimizing point for the rate function L on the set $[a,\infty)$. This is an encouraging result because we saw earlier in Remark 1.1.1 that if a rare event occurs then the most likely way for it to occur is via the minimizer of the rate function. So in some sense this change of measure should be a good approximation to the conditional distribution $\mathbb{P}(\cdot|S_n>na)$. It in fact does turn out that $\hat{p}_n$ is an asymptotically optimal importance sampling change of measure.

Due to the success and ease of implementation of Siegmund's approach, a general heuristic developed in the simulation community that a good method for generating importance sampling changes of measures is to perform an exponential twist of the original measure. Furthermore the direction and magnitude of this twist should be motivated by the exponential decay rate of the probability of interest. It should also be noted that under the standard heuristic all increments are treated identically, e.g. under the sampling measure the sequence discussed above $\{\bar{Y}_i\}$ is an i.i.d. sequence.

Unfortunately the standard heuristic was seen to be woefully inadequate in the work [18]. In this work the authors consider the problem of obtaining effective changes of measure for estimating

$$\mathbb{P}(|S_n| > na) \text{ where } S_n = \sum_{j=1}^{n} Y_j$$

and $\{Y_j\}$ is an i.i.d. sequence. Under very simple conditions on the increments the authors of [18] show that an importance sampling estimator based on the standard heuristic will in fact perform worse than standard Monte Carlo.

In [14] the failure of the standard heuristic is carefully analyzed. In that work it is shown that in the setting of Cramer's theorem one can only expect the standard heuristic to produce usable IS changes of measure if the set the target set A is convex. Thus it is clear that the standard heuristic will provide good changes of measure for estimating $\mathbb{P}(S_n > na)$ but not for estimating $\mathbb{P}(|S_n| > na)$. In addition the authors show that it is in general necessary to consider state-dependent changes of measure. For example if estimating $\mathbb{P}(S_n \in nA)$ one should sample the increment Y_k, $1 < k \leq n$ according to the state dependent distribution

$$\nu_j^n(dy) = f(S_j/n, y)\mu(dy).$$

The authors of [14] show that it is necessary to consider state dependent changes of measure by converting the analysis of the second moment of the importance sampling estimator to a differential game. One player of the game is the person creating a change of measure for the importance sampling estimator. The other player is introduced from a variational representation, see [5]. Therefore using a non-adaptive change of measure is equivalent to showing your opponent your strategy at the beginning of the game, and then never adjusting the strategy as the game evolves. Clearly most situations will favor an adaptive strategy that changes with the moves of the opponent.

Most importantly a method for the construction of IS changes of measure is given in the work [14]. In the work it is shown that asymptotically optimal IS changes of measure can be built using the gradient of the solution to an Isaacs equation. The Isaacs equation is a PDE that the value of the game should in some sense satisfy.

Unfortunately it is difficult to find solutions to these Isaacs equations. However in the paper [15] it was shown that efficient changes of measure can be built from the gradients of subsolutions to Isaacs equations. This result was a significant improvement that allowed for easy creation of efficient and easily implemented IS changes of measure, the power of these methods is shown in the works [13], [11], [9].

1.3 Layout of Thesis

The first chapter of the thesis is concerned with establishing an LDP for the queue length process of a single queue following the weighted-serve-the-longest (WSLQ) policy. The second chapter is interested in two topics. One topic is finding the explicit form of the exponential decay rate of the probability of a buffer overflow event for a single server following the WSLQ policy. Showing this result requires the main result from the first chapter. The other topic of the second chapter is the construction of asymptotically optimal changes of measure for estimating these buffer overflow probabilities. The third chapter considers a buffer overflow event in a two node tandem queuing network, where the downstream server is protected by a server-slowdown mechanism in the upstream queue. The explicit form of the exponential decay rate for this buffer overflow is found, and an asymptotically optimal importance sampling change of measure is constructed.

Chapter 2

A Large Deviations Principle for the Weighted-Serve-the-Longest-Queue Policy

2.1 Introduction

Consider a single server that must serve multiple queues of customers from different classes. A common service discipline in this situation is the serve-the-longest-queue policy, in which the longest queue is given priority. In this paper we will consider a natural generalization of this discipline, namely, the *weighted-serve-the-longest-queue* (WSLQ) policy. Under WSLQ, each queue length is multiplied by a constant to determine a "score" for that queue, and the queue with the largest score is granted priority. Such service policies are more appropriate than serve-the-longest-queue policy when the different arrival queues or customer classes have different requirements or statistical properties. For example, if there is a finite queueing capacity to be split among the different classes, one may want to choose the partition and the weighting constants in order to optimize a certain performance measure.

Because WSLQ is a frequently proposed discipline for queueing models in communication problems, a large deviations analysis of this protocol is useful [28]. However, service policies such as WSLQ are not smooth functions of the system state and lead to multidimensional stochastic processes with *discontinuous statistics*. In general, large deviation properties of processes with discontinuous statistics are hard to analyze [1, 17, 19, 21]. This is especially true when the discontinuities appear on the interior of the state space, rather than the boundary. In fact, very general results with an explicit identification of the rate function only exist for the case where two regions of smooth statistical behavior are separated by an interface of codimension one [4]. For the WSLQ policy, the large deviation analysis has been limited to special, two-dimensional cases [25]. Large deviations for a weighted-serve-the-largest-workload policy are treated in [24], but [24] considers a different class of stochastic processes.

An important contribution of the results in this chapter is that they show that a complete large deviation analysis of WSLQ is possible without any symmetry or dimensional assumptions. Given the intrinsic difficulties in models with discontinuous statistics, it is worthwhile to explain what makes such an analysis possible for WSLQ. To this end, we recall the main difficulty in the large deviation analysis of systems with discontinuous statistics. A large deviation upper bound can often be established using the results in [7], which assumes little regularity on the statistical behavior of the underlying processes. However, this upper bound is generally *not* tight, even for the very simple situation of two regions of constant statistical behavior separated by a hyperplane of codimension one [6].

The reason for this gap is most easily identified by considering the corresponding lower bound. When proving a large deviation lower bound, it is necessary to analyze the probability that the process closely follows or tracks a constant velocity trajectory that lies on the interface of two or more regions of smooth statistical behavior. For this one has to consider all changes of measure in these different regions that lead to the desired tracking behavior. The thorny issue is how to characterize such changes of measure. In the case of two regions [6], this can be done in a satisfactory fashion and it turns out that the large deviation rate function is a modified version of the upper bound in [7]. The modification is made to explicitly include certain "stability about the interface" conditions, and part of the reason that everything works out nicely in the setup of [6] is that these stability conditions can be easily characterized. However, the analogous characterization of stability is not known for more elaborate settings such as WSLQ, where the regions of constant statistical behavior are defined according to the partition of the state space by a finite number of hyperplanes of codimension one. Therefore, at present there is no general theory that subsumes WSLQ as a special case.

However, a key observation that makes possible a large deviations analysis for WSLQ is that for this model, the required stability conditions are *implicitly* and *automatically* built into the upper bound rate function of [7]. More precisely, it can be shown that in the lower bound analysis one can restrict, a priori, to a class of changes of measure for which the stability conditions are automatically implied. Thus while it is true that the upper bound of [7] is not tight in general, it is so in this case due to the structural properties of WSLQ policy.

The study of the large deviation properties of WSLQ is partly motivated by the problem of estimating buffer overflow (rare event) probabilities for stable WSLQ systems using importance sampling. It turns out that the simple form of the large deviation local rate function as exhibited in (2.3.1) and (2.3.2) is essential toward constructing simple and asymptotically optimal importance sampling schemes using a game theoretic approach. These results will be discussed in great detail in the next chapter.

This chapter is organized as follows. In Section 2, we introduce the single server system with WSLQ policy. In Section 3, we state the main result, whose proof is presented in Section 4. Collected in the appendices are a lengthy technical part of the proof that involves the approximation of continuous trajectories, as well as miscellaneous results.

2.2 Problem Setup

Consider a server with d customer classes, where customers of class i arrive according to a Poisson process with rate λ_i and are buffered at queue i for $i = 1, \ldots, d$. The service time for a customer of class i is exponentially distributed with rate μ_i.

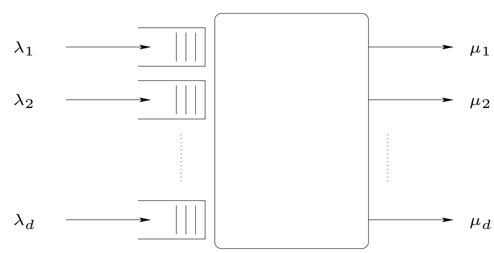

Figure 2.1: WSLQ system

The service policy is determined according to the WSLQ discipline that can be described as follows. Let c_i be the weight associated with class i. If the size of queue i is q_i, then the "score" of queue i is defined as $c_i q_i$ and service priority will be given to the queue with the maximal score. When there are multiple queues with the maximal score, the assignment of priority among these queues can be arbitrary – the choice is indeed non-essential and will lead to the same rate function. We adopt the convention that when there are ties, the priority will be given to the queue with the largest index.

The system state at time t is the vector of queue lengths and is denoted by $Q(t) \doteq (Q_1(t), \ldots, Q_d(t))$. Then Q is a continuous time pure jump Markov process whose possible jumps belong to the set

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

where e_j represents the jth unit vector. For $v \in \mathbb{V}$, let $r(x; v)$ denote the jump intensity of process Q from state x to state $x + v$. Under the WSLQ discipline, these jump intensities are as follows. For $x = (x_1, \ldots, x_d) \in \mathbb{R}_+^d$ and $x \neq 0$, let $\pi(x)$ denote the indices of queues that have the maximal score, that is,

$$\pi(x) \doteq \left\{ 1 \leq i \leq d : c_i x_i = \max_j c_j x_j \right\}.$$

Then

$$r(x; v) = \begin{cases} \lambda_i, & \text{if } v = e_i \text{ and } i = 1, \ldots, d, \\ \mu_i, & \text{if } v = -e_i \text{ where } i = \max \pi(x), \\ 0\,, & \text{otherwise.} \end{cases}$$

For $x = 0$, there is no service and the jump intensities are

$$r(0; v) = \begin{cases} \lambda_i, & \text{if } v = e_i \text{ and } i = 1, \ldots, d, \\ 0\,, & \text{otherwise.} \end{cases}$$

We also set

$$\pi(0) \doteq \{0, 1, 2, \ldots, d\}.$$

An illustrative figure for the case of two queues in given below.

Remark 2.2.1. It is not difficult to see that the system dynamics have constant statistical behavior in the regions where $\pi(\cdot)$ is constant. Discontinuity occurs when $\pi(\cdot)$ changes, and every x with $|\pi(x)| \geq 2$ (i.e., when there is a tie) is indeed a discontinuous point. Therefore, we have various discontinuity interfaces with different dimensions. For example, for every subset $A \subset \{1, 2, \ldots, d\}$ with $|A| \geq 2$ or $A =$

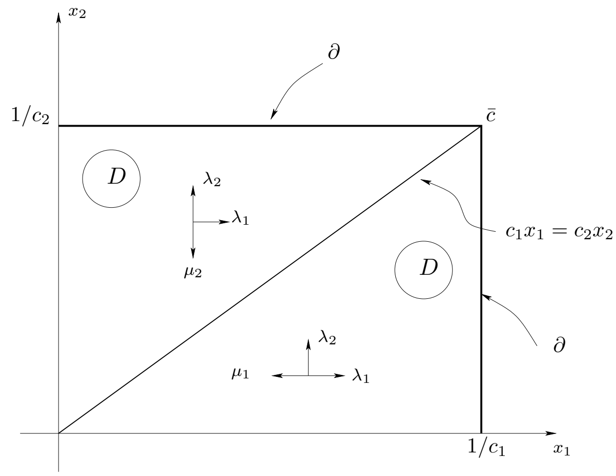

Figure 2.2: System dynamics for $d = 2$.

$\{0, 1, 2, \ldots, d\}$, the set $\{x \in \mathbb{R}_+^d : \pi(x) = A\}$ defines an interface with dimension $d - |A| + 1$.

Remark 2.2.2. The definition of $\pi(0)$ is introduced to cope with the discontinuous dynamics at the origin. Note that with this definition, $\pi(x)$ can only be $\{0, 1, 2, \ldots, d\}$ if $x = 0$ and a subset of $\{1, 2, \ldots, d\}$ if $x \neq 0$.

Remark 2.2.3. A useful observation is that π is *upper semicontinuous* as a set-valued function. That is, for any $x \in \mathbb{R}_+^d$, $\pi(y) \subset \pi(x)$ for all y in a small neighborhood of x.

2.3 The main result

In order to state a large deviation principle on path space, we fix arbitrarily $T > 0$, and for each $n \in \mathbb{N}$ let $\{X^n(t) : t \in [0, T]\}$ be the scaled process defined by

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

Then X^n is a continuous time Markov process with generator

$$\mathcal{L}^n f(x) = n \sum_{v \in \mathbb{V}} r(x;v) \left[f\left(x + v/n\right) - f(x) \right].$$

The processes $\{X^n\}$ live in the space of càdlàg functions $\mathcal{D}([0,T] : \mathbb{R}^d)$, which is endowed with the Skorohod metric and thus a Polish space.

2.3.1 The rate function

For each $i = 1, \ldots, d$, let $H^{(i)}$ be the convex function given by

$$H^{(i)}(\alpha) \doteq \mu_i(e^{-\alpha_i} - 1) + \sum_{j=1}^{d} \lambda_j (e^{\alpha_j} - 1),$$

for all $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{R}^d$. We also define for $i = 0$,

$$H^{(0)}(\alpha) \doteq \sum_{j=1}^{d} \lambda_j (e^{\alpha_j} - 1)$$

for $\alpha \in \mathbb{R}^d$.

For each non-empty subset $A \subset \{1, \ldots, d\}$ or $A = \{0, 1, \ldots, d\}$, let L^A be the Legendre transform of $\max_{i \in A} H^{(i)}$, that is,

$$L^A(\beta) \doteq \sup_{\alpha \in \mathbb{R}^d} \left[\langle \alpha, \beta \rangle - \max_{i \in A} H^{(i)}(\alpha) \right]$$

for each $\beta \in \mathbb{R}^d$. Clearly, L^A is convex and non-negative. When A is a singleton $\{i\}$, we simply write L^A as $L^{(i)}$. A useful representation of L^A [5, Corollary D.4.3] is

$$L^A(\beta) = \inf \sum_{i \in A} \rho^{(i)} L^{(i)}(\beta^{(i)}), \tag{2.3.1}$$

where the infimum is taken over all $\{(\rho^{(i)}, \beta^{(i)}) : i \in A\}$ such that

$$\rho^{(i)} \geq 0, \quad \sum_{i \in A} \rho^{(i)} = 1, \quad \sum_{i \in A} \rho^{(i)} \beta^{(i)} = \beta. \tag{2.3.2}$$

Furthermore, for $x \in \mathbb{R}^d_+$ let $L(x, \beta) \doteq L^{\pi(x)}(\beta)$.

Now we can define the process level rate function. For $x \in \mathbb{R}^d_+$, define $I_x : \mathcal{D}([0,T] : \mathbb{R}^d) \to [0, \infty]$ by

$$I_x(\psi) \doteq \int_0^T L(\psi(t), \dot{\psi}(t))dt \tag{2.3.3}$$

if $\psi(0) = x$, ψ is absolutely continuous, and $\psi(t) \in \mathbb{R}^d_+$ for all $t \in [0, T]$. Otherwise set $I_x(\psi) \doteq \infty$. The family of rate functions $\{I_x : x \in \mathbb{R}^d_+\}$ has compact level sets on compacts in the sense that, for every $M \geq 0$ and every compact set $C \in \mathbb{R}^d_+$, the set

$$\cup_{x \in C} \left\{\psi \in \mathcal{D}([0,T] : \mathbb{R}^d) : I_x(\psi) \leq M\right\}$$

is compact [7, Theorem 1.1].

2.3.2 The main theorem

The main result of this paper can be stated as follows. Let E_{x_n} denote the expectation conditioned on $X^n(0) = x_n$.

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

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

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

2.4 Proof of the main theorem

Throughout the rest of the paper, we will assume without loss of generality that $T = 1$. We only need to show the uniform Laplace principle which automatically implies the large deviation principle [5, Theorem 1.2.3]. The uniform Laplace principle upper bound is implied by the large deviation upper bound [7, Theorem 1.1] and an argument similar to [5, Theorem 1.2.1]. Therefore, it is only necessary to prove the uniform Laplace principle lower bound. That is,

$$\limsup_{n\to\infty} -\frac{1}{n} \log E_{x_n} \{\exp [-nh(X^n)]\} \leq \inf_{\psi\in\mathcal{D}([0,1]:\mathbb{R}^d)} \{I_x(\psi) + h(\psi)\}. \tag{2.4.1}$$

One can a priori restrict to ψ such that $I_x(\psi) < \infty$ since the inequality holds trivially otherwise. Note that such ψ's are necessarily absolutely continuous by the definition of I_x.

2.4.1 An approximation lemma

A very important step in the proof of (2.4.1) is the following approximation lemma. A function ψ^* has property $\mathcal{P}$ if there exists a $k \in \mathbb{N}$ and $0 = t_0 < t_1 < \cdots < t_{k-1} < t_k = 1$ such that on each interval $(t_{i-1}, t_i), i = 1, \ldots, k$, $\dot{\psi}^*$ and $\pi(\psi^*)$ are both constant. Then define $\mathcal{N}$ be the collection of functions $\psi^* \in \mathcal{C}([0,1] : \mathbb{R}^d_+)$ that have property $\mathcal{P}$.

Lemma 2.4.1. *Given any $\psi \in \mathcal{D}([0,1] : \mathbb{R}^d)$ such that $I_x(\psi) < \infty$ and any $\delta > 0$, there exists $\psi^* \in \mathcal{N}$ such that $\|\psi - \psi^*\|_\infty < \delta$ and $I_x(\psi^*) \leq I_x(\psi) + \delta$.*

The proof of Lemma 2.4.1 is lengthy and technical, and is deferred to Appendix 2.6.

We claim that, in order to show inequality (2.4.1), it suffices to show that for any

$x_n \to x$ and $\psi^* \in \mathcal{N}$,

$$\limsup_{n\to\infty} -\frac{1}{n} \log E_{x_n} \left\{ \exp\left[-nh(X^n)\right]\right\} \leq I_x(\psi^*) + h(\psi^*). \tag{2.4.2}$$

This reduction follows easily from Lemma 2.4.1 and the continuity of h. We omit the details.

2.4.2 A stochastic control representation

The proof of (2.4.2) uses the weak convergence approach, and is based on the following formula. Define the function ℓ by

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

with the convention $0\ell(0/0) \doteq 0$. Let $\bar{r}(x,t;v)$ be non-negative, uniformly bounded, and piecewise constant in t, and also satisfy $\bar{r}(x,t;v) = 0$ whenever $r(x;v) = 0$. Let $\bar{X}^n$ be the non-stationary jump Markov process with generator

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

Then

$$\begin{aligned} & -\frac{1}{n} \log E_x \left\{ \exp\left[-nh(X^n)\right]\right\} \qquad (2.4.3) \\ \leq \; & \inf_{\bar{r}} E_x \left[\int_0^1 \sum_{v\in\mathbb{V}} r(\bar{X}^n(t);v)\, \ell\left(\frac{\bar{r}(\bar{X}^n(t),t;v)}{r(\bar{X}^n(t);v)}\right) dt + h(\bar{X}^n) \right]. \end{aligned}$$

In this inequality $\bar{r}$ can be viewed as a control and $\bar{X}^n$ a controlled process.

The proof of (2.4.3) follows from the relative entropy representation formula for exponential integrals. Let P be a probability measure and let h be a bounded continuous function on $\mathcal{D}([0,1]:\mathbb{R}^d)$. For another probability measure Q on $\mathcal{D}([0,1]:\mathbb{R}^d)$ let $R(Q\,\|P)$ denote the relative entropy of Q with respect to P. Then [5]

$$-\frac{1}{n}\log\int_{\mathcal{D}([0,1]:\mathbb{R}^d)} e^{-nh}dP = \inf\left[\frac{1}{n}R(Q\,\|P) + \int_{\mathcal{D}([0,1]:\mathbb{R}^d)} hdQ\right],$$

where the infimum is over all Q. If one restricts Q to the probability measures induced by the class of Markov processes described above, one obtains an inequality. Finally, we substitute the explicit form of the Radon-Nikodym derivative dQ/dP (as in [25, Theorem B.6]) into the definition of $R(Q\,\|P)$ to arrive at the inequality in (2.4.3).

Thanks to the control representation (2.4.3), the Laplace principle lower bound (2.4.2) for $\psi^* \in \mathcal{N}$ follows if one can, for an arbitrarily fixed $\varepsilon > 0$, construct a control (abusing the notation) $\bar{r} = \bar{r}^\varepsilon$ such that

$$\begin{aligned}\limsup_{n\to\infty} E_{x_n}\left[\int_0^1 \sum_{v\in\mathbb{V}} r(\bar{X}^n(t);v)\ell\left(\frac{\bar{r}(\bar{X}^n(t),t;v)}{r(\bar{X}^n(t);v)}\right)dt + h(\bar{X}^n)\right] \\ \leq \; I_x(\psi^*) + h(\psi^*) + \varepsilon.\end{aligned} \tag{2.4.4}$$

The details of the construction will be carried out in the next section.

2.4.3 Properties of the rate function

We give some useful representation formulae for the local rate functions L^A.

Lemma 2.4.2. *Given $\beta \in \mathbb{R}^d$, the following representation for $L^{(i)}(\beta)$ holds.*

1. *For each $i = 1, 2, \ldots, d$,*

$$L^{(i)}(\beta) = \inf\left\{\mu_i\ell\left(\frac{\bar{\mu}_i}{\mu_i}\right) + \sum_{j=1}^d \lambda_j\ell\left(\frac{\bar{\lambda}_j}{\lambda_j}\right) : -\bar{\mu}_i e_i + \sum_{j=1}^d \bar{\lambda}_j e_j = \beta\right\}.$$

2. *For $i = 0$,*

$$L^{(0)}(\beta) = \inf \left\{ \sum_{j=1}^{d} \lambda_j \ell \left(\frac{\bar{\lambda}_j}{\lambda_j} \right) : \sum_{j=1}^{d} \bar{\lambda}_j e_j = \beta \right\}.$$

In every case the infimum is attained.

Note that for every $\lambda > 0$ and $v \in \mathbb{R}^d$, the Legendre transform of the convex function

$$h(\alpha) \doteq \lambda \left(e^{\langle \alpha, v \rangle} - 1 \right)$$

is

$$h^*(\beta) \doteq \begin{cases} \lambda \ell(\bar{\lambda}/\lambda), & \text{if } \beta = \bar{\lambda} v \text{ for some } \bar{\lambda} \in \mathbb{R}, \\ \infty \quad , & \text{otherwise} \end{cases}$$

for every $\beta \in \mathbb{R}^d$. The proof of this claim is straightforward computation and we omit the details. Now the representation for $L^{(i)}$ follows directly from [5, Corollary D.4.2]. The attainability of the infimum is elementary. ■

Lemma 2.4.3. *Given $\beta \in \mathbb{R}^d$, we have the following representation for $L^A(\beta)$.*

1. *Assume $A \in \{1, 2, \ldots, d\}$ is non-empty. Then*

$$L^A(\beta) = \inf \left[\sum_{i \in A} \rho^{(i)} \mu_i \ell \left(\frac{\bar{\mu}_i}{\mu_i} \right) + \sum_{j=1}^{d} \lambda_j \ell \left(\frac{\bar{\lambda}_j}{\lambda_j} \right) \right],$$

where the infimum is taken over all collections of $(\rho^{(i)}, \bar{\mu}_i, \bar{\lambda}_j)$ such that

$$\rho^{(i)} \geq 0, \quad \sum_{i \in A} \rho^{(i)} = 1, \quad -\sum_{i \in A} \rho^{(i)} \bar{\mu}_i e_i + \sum_{j=1}^{d} \bar{\lambda}_j e_j = \beta. \tag{2.4.5}$$

2. *For $A = \{0, 1, 2, \ldots, d\}$, we have*

$$L^A(\beta) = \inf \left[\sum_{i=1}^{d} \rho^{(i)} \mu_i \ell \left(\frac{\bar{\mu}_i}{\mu_i} \right) + \sum_{j=1}^{d} \lambda_j \ell \left(\frac{\bar{\lambda}_j}{\lambda_j} \right) \right],$$

where the infimum is taken over all collections of $(\rho^{(i)}, \bar{\mu}_i, \bar{\lambda}_j)$ *such that*

$$\rho^{(i)} \geq 0, \quad \sum_{i=0}^{d} \rho^{(i)} = 1, \quad -\sum_{i=1}^{d} \rho^{(i)} \bar{\mu}_i e_i + \sum_{j=1}^{d} \bar{\lambda}_j e_j = \beta.$$

We only present the proof for Part 1. The proof for Part 2 is similar and thus omitted. Thanks to Lemma 2.4.2 and equations (2.3.1)-(2.3.2), we have

$$L^A(\beta) = \inf \sum_{i \in A} \rho^{(i)} \left\{ \mu_i \ell \left(\frac{\bar{\mu}_i^{(i)}}{\mu_i} \right) + \sum_{j=1}^{d} \lambda_j \ell \left(\frac{\bar{\lambda}_j^{(i)}}{\lambda_j} \right) \right\},$$

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

$$\rho^{(i)} \geq 0, \quad \sum_{i \in A} \rho^{(i)} = 1, \quad \sum_{i \in A} \rho^{(i)} \left[-\bar{\mu}_i^{(i)} e_i + \sum_{j=1}^{d} \bar{\lambda}_j^{(i)} e_j \right] = \beta. \tag{2.4.6}$$

Abusing the notation a bit, write $\bar{\mu}_i = \bar{\mu}_i^{(i)}$ for $i \in A$, and let $\bar{\lambda}_j \doteq \sum_{i \in A} \rho^{(i)} \bar{\lambda}_j^{(i)}$ for $j = 1, 2, \ldots, d$. Thanks to (2.4.6), the collection $(\rho^{(i)}, \bar{\mu}_i, \bar{\lambda}_j)$ satisfies the constraints (2.4.5). Observing that, by convexity of ℓ,

$$\sum_{i \in A} \rho^{(i)} \sum_{j=1}^{d} \lambda_j \ell \left(\frac{\bar{\lambda}_j^{(i)}}{\lambda_j} \right) = \sum_{j=1}^{d} \lambda_j \sum_{i \in A} \rho^{(i)} \ell \left(\frac{\bar{\lambda}_j^{(i)}}{\lambda_j} \right) \geq \sum_{j=1}^{d} \lambda_j \ell \left(\frac{\bar{\lambda}_j}{\lambda_j} \right),$$

with equality if $\lambda_j^{(i)} = \lambda_k^{(i)}$ for every j, k. The first part of Lemma 2.4.3 now follows readily. ■

Remark 2.4.4. The representation of L^A in Lemma 2.4.3 remains valid if we further constrain $\rho^{(i)}$, $\bar{\mu}_i$, and $\bar{\lambda}_i$ to be strictly positive for every $i \in A$. This is an easy consequence of the fact that ℓ is finite and continuous on the interval $[0, \infty)$. We omit the details.

Remark 2.4.5. Given a non-empty subset $A \subsetneq \{1, 2, \ldots, d\}$ and $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$, $L^A(\beta)$ is finite if and only if $\beta_j \geq 0$ for all $j \notin A$. For $A = \{1, 2, \ldots, d\}$ or

$\{0,1,2,\ldots,d\}$, $L^A(\beta)$ is finite for every $\beta \in \mathbb{R}^d$.

2.4.4 The construction of controls and the cost

Fix $\varepsilon > 0$. We will use ψ^* to construct a control $\bar{r} = \bar{r}^\varepsilon$ based on the representation of the local rate function as in Lemma 2.4.3. For notational simplicity, we drop the superscript ε.

Since $\psi^* \in \mathcal{N}$, there exist $0 = t_0 < t_1 < \cdots < t_K = 1$ such that for every k there are β_k and A_k such that $\dot{\psi}^*(t) \equiv \beta_k$ and $\pi(\psi^*(t)) \equiv A_k$ for all $t \in (t_k, t_{k+1})$. We start by defining a suitable collection of $\{(\rho_k^{(i)}, \bar{\mu}_{i,k}, \bar{\lambda}_{j,k}) : 0 \le i \le d, 1 \le j \le d\}$. We consider the following two cases.

Case 1. Suppose $A_k = \{0,1,2,\ldots,d\}$. Lemma 2.4.3 and Remark 2.4.4 imply the existence of a collection $\{(\rho_k^{(i)}, \bar{\mu}_{i,k}, \bar{\lambda}_{j,k}) : 0 \le i \le d, 1 \le j \le d\}$ such that $\bar{\mu}_{i,k} > 0$, $\bar{\lambda}_{i,k} > 0$ for all i and

$$\rho_k^{(i)} > 0, \quad \sum_{i=0}^{d} \rho_k^{(i)} = 1, \quad -\sum_{i=1}^{d} \rho_k^{(i)} \bar{\mu}_{i,k} e_i + \sum_{j=1}^{d} \bar{\lambda}_{j,k} e_j = \beta_k, \tag{2.4.7}$$

$$\sum_{i=1}^{d} \rho_k^{(i)} \mu_i \ell\left(\frac{\bar{\mu}_{i,k}}{\mu_i}\right) + \sum_{j=1}^{d} \lambda_j \ell\left(\frac{\bar{\lambda}_{j,k}}{\lambda_j}\right) \le L^{A_k}(\beta_k) + \varepsilon. \tag{2.4.8}$$

Case 2. Suppose $A_k \subset \{1,2,\ldots,d\}$. According to Lemma 2.4.3 and Remark 2.4.4, for each k there exist a collection $\{(\rho_k^{(i)}, \bar{\mu}_{i,k}, \bar{\lambda}_{j,k}) : i \in A_k, 1 \le j \le d\}$ such that $\bar{\mu}_{i,k} > 0$, $\bar{\lambda}_{i,k} > 0$ for all $i \in A_k$ and

$$\rho_k^{(i)} > 0, \quad \sum_{i \in A_k} \rho_k^{(i)} = 1, \quad -\sum_{i \in A_k} \rho_k^{(i)} \bar{\mu}_{i,k} e_i + \sum_{j=1}^{d} \bar{\lambda}_{j,k} e_j = \beta_k, \tag{2.4.9}$$

$$\sum_{i\in A_k} \rho_k^{(i)} \mu_i \ell\left(\frac{\bar{\mu}_{i,k}}{\mu_i}\right) + \sum_{j=1}^{d} \lambda_j \ell\left(\frac{\bar{\lambda}_{j,k}}{\lambda_j}\right) \le L^{A_k}(\beta_k) + \varepsilon. \tag{2.4.10}$$

We extend the definition by letting $\rho_k^{(i)} \doteq 0$, $\bar{\mu}_{i,k} \doteq \mu_i$ for $i \notin A_k$ and $i \neq 0$, and letting $\rho_k^{(0)} \doteq 0$.

The control $\bar{r}$ is defined as follows. For $t \in [t_k, t_{k+1})$, let

$$\bar{r}(x,t;v) = \begin{cases} \bar{\lambda}_{j,k}, & \text{if } v = e_j \text{ and } j = 1,\ldots,d, \\ \bar{\mu}_{j,k}, & \text{if } v = -e_j \text{ where } j = \max \pi(x) \text{ and } x \neq 0, \\ 0\;, & \text{otherwise.} \end{cases}$$

Thus, on time interval $[t_k, t_{k+1})$, the system has arrival rates $\{\bar{\lambda}_{1,k},\ldots,\bar{\lambda}_{d,k}\}$ and service rates $\{\bar{\mu}_{1,k},\ldots,\bar{\mu}_{d,k}\}$ under this control $\bar{r}$. The corresponding running cost for $t \in [t_k, t_{k+1})$ is

$$\begin{aligned} &\sum_{v\in\mathbb{V}} r(\bar{X}^n(t);v)\ell\left(\frac{\bar{r}(\bar{X}^n(t),t;v)}{r(\bar{X}^n(t);v)}\right) \qquad (2.4.11) \\ &= \sum_{j=1}^{d} \mu_j \ell\left(\frac{\bar{\mu}_{j,k}}{\mu_j}\right) 1_{\{\max\pi(\bar{X}^n(t))=j,\bar{X}^n(t)\neq 0\}} + \sum_{j=1}^{d} \lambda_j \ell\left(\frac{\bar{\lambda}_{j,k}}{\lambda_j}\right) \\ &= \sum_{i\in A_k, i\neq 0} \mu_i \ell\left(\frac{\bar{\mu}_{i,k}}{\mu_i}\right) 1_{\{\max\pi(\bar{X}^n(t))=i,\bar{X}^n(t)\neq 0\}} + \sum_{j=1}^{d} \lambda_j \ell\left(\frac{\bar{\lambda}_{j,k}}{\lambda_j}\right), \end{aligned}$$

here the last equality holds since $\bar{\mu}_{j,k} = \mu_j$ for $j \notin A_k$ and $\ell(1) = 0$.

For future use, we also define for each $i = 1,\ldots,d$,

$$\beta_k^{(i)} \doteq -\bar{\mu}_{i,k} e_i + \sum_{j=1}^{d} \bar{\lambda}_{j,k} e_j, \tag{2.4.12}$$

which is the law of large number limit of the velocity of the controlled process if queue of class i is served. Analogously, we also define (when none of the queues are being served)

$$\beta_k^{(0)} \doteq \sum_{j=1}^{d} \bar{\lambda}_{j,k} e_j. \tag{2.4.13}$$

2.4.5 Weak convergence analysis

In this section we characterize the limit processes. Below are a few definitions. For each j, define random measures $\{\gamma_j^n\}$ on $[0,1]$ by

$$\begin{aligned}
\gamma_j^n\{B\} &\doteq \int_B 1_{\{\max \pi(\bar{X}^n(t))=j, \bar{X}^n(t)\neq 0\}} dt, \qquad j=1,2,\ldots,d, \\
\gamma_0^n\{B\} &\doteq \int_B 1_{\{\bar{X}^n(t)=0\}} dt,
\end{aligned}$$

for Borel subsets $B \subset [0,1]$, and denote $\gamma^n \doteq (\gamma_0^n, \gamma_1^n, \ldots, \gamma_d^n)$. We also define the stochastic processes

$$S^n(t) \doteq x_n + \sum_{j=0}^{d} \left[\sum_{k=0}^{\kappa(t)-1} \beta_k^{(j)} \gamma_j^n\{[t_k, t_{k+1})\} + \beta_{\kappa(t)}^{(j)} \gamma_j^n\{[t_{\kappa(t)}, t)\} \right],$$

where $\kappa(t) = \max\{0 \le k \le K : t_k \le t\}$.

Proposition 2.4.6. *Given any subsequence of $(\gamma^n, S^n, \bar{X}^n)$, there exist a subsubsequence, a collection of random measures $\gamma \doteq (\gamma_0, \gamma_1, \ldots, \gamma_d)$ on $[0,1]$, and a continuous process $\bar{X}$ such that*

(a) *The subsubsequence converges in distribution to $(\gamma, \bar{X}, \bar{X})$.*

(b) *With probability one, γ_j is absolutely continuous with respect to the Lebesgue measure on $[0,1]$, and its density, denoted by h_j, satisfies*

$$\sum_{j=0}^{d} h_j(t) = \sum_{j \in \pi(\bar{X}(t))} h_j(t) = 1$$

for almost every t.

(c) *With probability one, the process $\bar{X}$ satisfies*

$$\bar{X}(t) = x + \sum_{j=0}^{d} \left[\sum_{k=0}^{\kappa(t)-1} \beta_k^{(j)} \gamma_j\{[t_k, t_{k+1})\} + \beta_{\kappa(t)}^{(j)} \gamma_j\{[t_{\kappa(t)}, t)\} \right]$$

for every t. Therefore, $\bar{X}$ is absolutely continuous with derivative

$$\frac{d\bar{X}(t)}{dt} = \sum_{j=0}^{d} \beta_{\kappa(t)}^{(j)} h_j(t).$$

The family of random measures $\{\gamma_j^n\}$ is contained in the compact set of subprobability measures on $[0,1]$ and therefore tight. Furthermore, since $\{S^n\}$ is uniformly Lipschitz continuous, it takes values in a compact subset of $\mathcal{C}([0,1] : \mathbb{R}^d)$, and therefore is also tight. We also observe that for every $\varepsilon > 0$,

$$\lim_{n \to \infty} P(\|\bar{X}^n - S^n\|_\infty > \varepsilon) = 0, \tag{2.4.14}$$

which in turn implies that $\{\bar{X}^n\}$ is tight. Equation (2.4.14) is trivial since $\bar{X}^n - S^n$ is a martingale (e.g., [25, Appendix B.2]), whence the process $\|\bar{X}^n - S^n\|^2$ is a submartingale. Therefore, by the submartingale inequality and the uniform boundedness of the jump intensity $\bar{r}$

$$P\left(\sup_{0 \le t \le 1} \|\bar{X}^n(t) - S^n(t)\| > \varepsilon \right) \le \frac{2}{\varepsilon^2} E\left[\|\bar{X}^n(1) - S^n(1)\|^2 \right] \to 0.$$

It follows that there exists a subsubsequence that converges weakly to say (γ, S, S), with $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_d)$. By the Skorohod representation theorem, we assume without loss of generality that the convergence is almost sure convergence, and everything is defined on some probability space, say $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{P})$.

Since the $\{\gamma_j^n\}$ are absolutely continuous with respect to Lebesgue measure on $[0,1]$ with uniformly bounded densities (i.e., Radon-Nikodým derivatives) in both n

and t, the limit γ_j is also absolutely continuous. Furthermore, if we define the process $\bar{X}$ as in (c), the above consideration yields that, for every $t \in [0,1]$, $S^n(t)$ converges to $\bar{X}(t)$ almost surely. Therefore, with probability one, $S(t) = \bar{X}(t)$ for all these rational $t \in [0,1]$. Since both S and $\bar{X}$ are continuous, $S = \bar{X}$ with probability one.

It remains to show the two equalities of (b). Since $\sum_{j=0}^{d} \gamma_j^n$ equals Lebesgue measure for every n, we have $\sum_{j=0}^{d} h_j(t) = 1$ for almost every t. The proof of the second equality is similar to that of [5, Theorem 7.4.4(c)]. Consider an $\omega \in \bar{\Omega}$ such that $\bar{X}(t,\omega)$ is a continuous function of $t \in [0,1]$, $\gamma^n(\omega) \Rightarrow \gamma(\omega)$, and such that $\bar{X}^n(\cdot,\omega)$ converges to $\bar{X}(\cdot,\omega)$ in the Skorohod metric (whence also in sup-norm since $\bar{X}(\cdot,\omega)$ is continuous [5, Theorem A.6.5]). By the upper semicontinuity of $\pi(\cdot)$ [Remark 2.2.3], it follows that for any $t \in (0,1)$ and $A \subset \{0,1,\ldots,d\}$ such that $\pi(\bar{X}(t,\omega)) \subset A$, there exist an open interval (a,b) containing t and $N \in \mathbb{N}$ such that $\pi(\bar{X}^n(s,\omega)) \subset A$ for all $n \geq N$ and $s \in (a,b)$. Therefore $\sum_{j \notin A} \gamma_j^n(\omega)\{(a,b)\} = 0$ for all $n \geq N$. Taking the limit as $n \to \infty$ we have $\sum_{j \notin A} \gamma_j(\omega)\{(a,b)\} = 0$, which in turn implies that

$$\sum_{j \notin A} \gamma_j(\omega)\{t \in (0,1) : \pi(\bar{X}(t,\omega)) \subset A\} = 0,$$

or equivalently,

$$\int_0^1 \sum_{j \notin A} h_j(t,\omega) 1_{\{\pi(\bar{X}(t,\omega)) \subset A\}} \, dt = 0.$$

We claim that this implies the desired equality at ω. Otherwise, there exists a subset $D \subset (0,1)$ with positive Lebesgue measure such that for every $t \in D$,

$$\sum_{j \notin \pi(\bar{X}(t,\omega))} h_j(t,\omega) > 0.$$

Since $\pi(\bar{X}(t,\omega))$ can only take finitely many possible values, there exists a subset (abusing the notation) $A \subset \{0, 1, \ldots, d\}$ such that the set

$$\bar{D} \doteq \{t \in D : \pi(\bar{X}(t,\omega)) = A\}$$

has positive Lebesgue measure. It follows that

$$\begin{aligned}
\int_0^1 \sum_{j \notin A} h_j(t,\omega) 1_{\{\pi(\bar{X}(t,\omega)) \subset A\}}\, dt &\geq \int_{\bar{D}} \sum_{j \notin A} h_j(t,\omega) 1_{\{\pi(\bar{X}(t,\omega)) \subset A\}}\, dt \\
&= \int_{\bar{D}} \sum_{j \notin \pi(\bar{X}(t,\omega))} h_j(t,\omega)\, dt \\
&> 0,
\end{aligned}$$

a contradiction. This completes the proof. ∎

2.4.6 Stability analysis

In this section we prove a key lemma in the analysis that identifies the weak limit $\bar{X}$ as ψ^*. The proof uses the implied "stability about the interface" in a crucial way.

We discuss the main idea behind this stability property before giving the detailed proof. For the large deviation analysis, it is important to analyze the probability that the process tracks a segment of trajectory that lies on an interface, say $\{x : \pi(x) = A\}$, with a constant velocity, say β. To this end, it is natural to use the change of measure induced by β through the local rate function L as described in Section 4.4. However, for general systems, this very natural construction does *not* guarantee that $\bar{X}$ will follow or track ψ, and certain "stability about the interface" conditions have to be added for this to happen.

For the WSLQ system, a stability condition is not explicitly needed since it is implicitly and automatically built into the upper bound local rate function L. To be more precise, denote the arrival and service rates under the change of measure by $\bar{\lambda}_i$

and $\bar{\mu}_i$ respectively. Then for the tracking behavior to take place it is required that the proportion of time that the process $\bar{X}$ spent in the region $\{x : \max \pi(x) = i\}$ equals $\rho^{(i)}$, where $\{\rho^{(i)}\}$ is the (strictly positive) solution to the system of equations

$$c_i(\bar{\lambda}_i - \rho^{(i)}\bar{\mu}_i) = c_j(\bar{\lambda}_j - \rho^{(j)}\bar{\mu}_j), \quad i, j \in A, \text{ and } \sum_{i \in A} \rho^{(i)} = 1.$$

Thanks to Proposition 2.4.6, such proportions are characterized by $\{h_i(t)\}$ where

$$\sum_{i \in \pi(\bar{X}(t))} h_i(t) = 1.$$

Therefore, we would like to show that $\pi(\bar{X}(t)) \equiv A$ and $h_i(t) \equiv \rho^{(i)}$ for $i \in A$. This can be shown, and the argument is based on the simple fact that for any non-empty subset $B \subset \{1, 2, \ldots, d\}$ and any b, the solution $\{x_i : i \in B\}$ to the system of equations

$$c_i(\bar{\lambda}_i - x_i\bar{\mu}_i) = c_j(\bar{\lambda}_j - x_j\bar{\mu}_j), \quad i, j \in B, \text{ and } \sum_{i \in B} x_i = b, \tag{2.4.15}$$

is unique and component-wise strictly increasing with respect to b.

For example, suppose on some time interval (a, b) that $\pi(\bar{X}(t)) \equiv B$ where B is a strict subset of A. It is not difficult to see that $\{h_i(t) : i \in B\}$ is a solution to equation (2.4.15) because, thanks to Proposition 2.4.6,

$$\frac{d}{dt}(\bar{X}(t))_i = \bar{\lambda}_i - h_i(t)\bar{\mu}_i, \quad \text{for } i \in B.$$

Since

$$\sum_{i \in B} \rho^{(i)} = 1 - \sum_{i \in A \setminus B} \rho^{(i)} < 1,$$

the monotonicity implies $h_i(t) > \rho^{(i)}$ for all $i \in B$. This in turn yields [see the proof of (2.4.25) for details] that for $i \in B$ and $j \in A \setminus B$,

$$\frac{d}{dt}\left[c_i((\bar{X}(t))_i - c_j((\bar{X}(t))_j\right] < 0. \tag{2.4.16}$$

Thus the differences between weighted queue lengths grows smaller, and the state is "pushed" towards the interface A. This derivation can be used to prove by contradiction that $A \subset \pi(\bar{X}(t))$. The other direction $\pi(\bar{X}(t)) \subset A$ can be shown similarly. Once $\pi(\bar{X}(t)) = A$ is shown, $h_i(t) = \rho^{(i)}$ follows immediately from the uniqueness of the solution to equation (2.4.15).

We want to point out that in general such stability conditions are equivalent to the existence of certain Lyapunov functions, and for the case of WSLQ, inequality (2.4.16) indicates that the difference of weighted queue lengths is a Lyapunov function.

Lemma 2.4.7. *Let $(\gamma, \bar{X}, \bar{X})$ be a limit of any weakly converging subsubsequence $(\gamma^n, S^n, \bar{X}^n)$ as in Proposition 2.4.6. Then with probability one, $\bar{X}(t) = \psi^*(t)$ for every $t \in [0,1]$, and for each j,*

$$h_j(t) = \sum_{k=0}^{K-1} \rho_k^{(j)} 1_{(t_k,t_{k+1})}(t)$$

for almost every $t \in [0,1]$.

The proof is by induction. By definition $\bar{X}(0) = x = \psi^*(0)$. Assume that $\bar{X}(t) = \psi^*(t)$ for all $t \in [0, t_k]$. The goal is to show that $\bar{X}(t) = \psi^*(t)$ for all $t \in [0, t_{k+1}]$. Define $A_k = \pi(\psi^*(t))$, $t \in (t_k, t_{k+1})$ and $A = \pi(\psi^*(t_k))$. Note that $A_k \subset A$ thanks to the continuity of ψ^* and the upper semicontinuity of π. Define the random time

$$\tau_k = \inf\left\{t > t_k : \pi(\bar{X}(t)) \not\subset A\right\}.$$

Since $\pi(\bar{X}(t_k)) = \pi(\psi^*(t_k)) = A$ and $\bar{X}$ is continuous, the upper semicontinuity of π implies that $\tau_k > t_k$ and $\pi(\bar{X}(\tau_k)) \not\subset A$. We claim that it suffices to show $\bar{X}(t) = \psi^*(t)$

and $h_j(t) = \rho_k^{(j)}$ for all $t \in (t_k, t_{k+1} \wedge \tau_k)$. Indeed, if this is the case, we must have $\tau_k \geq t_{k+1}$ with probability one, since otherwise by continuity $\bar{X}(\tau_k) = \psi^*(\tau_k)$ and thus $\pi(\bar{X}(\tau_k)) = A_k \subset A$, a contradiction.

The proof of $\bar{X}(t) = \psi^*(t)$ and $h_j(t) = \rho_k^{(j)}$ for all $t \in (t_k, t_{k+1} \wedge \tau_k)$ proceeds in three steps.

Step 1 For every $t \in (t_k, t_{k+1} \wedge \tau_k)$, either $A_k \subset \pi(\bar{X}(t))$ or $\pi(\bar{X}(t))$ is a strict subset of A_k.

Step 2 For every $t \in (t_k, t_{k+1} \wedge \tau_k)$, $\pi(\bar{X}(t)) = A_k$.

Step 3 For every $t \in (t_k, t_{k+1} \wedge \tau_k)$, $\bar{X}(t) = \psi^*(t)$ and $h_j(t) = \rho_k^{(j)}$.

Note that, for $t \in (t_k, t_{k+1} \wedge \tau_k)$, the definition of τ_k implies $\pi(\bar{X}(t)) \subset A$. It follows from (2.4.12), (2.4.13), and Proposition 2.4.6 that,

$$\frac{d}{dt}\bar{X}(t) = \sum_{i \in \pi(\bar{X}(t))} \beta_k^{(i)} h_i(t) = \sum_{j=1}^{d} \bar{\lambda}_{j,k} e_j - \sum_{i \in \pi(\bar{X}(t)),\ i \neq 0} \bar{\mu}_{i,k} h_i(t) e_i. \tag{2.4.17}$$

To verify Step 1, we assume that A_k is a strict subset of A and a strict subset of $\{1, 2, \ldots, d\}$ as well, since otherwise the claim is trivial. It follows from the definitions of π, A, and A_k that for every $i \in A_k$ and $j \in A \setminus A_k$ such that $j \neq 0$,

$$c_i \left(\psi^*(t_k)\right)_i - c_j \left(\psi^*(t_k)\right)_j = 0$$

and for $t \in (t_k, t_{k+1})$

$$c_i \left(\psi^*(t)\right)_i - c_j \left(\psi^*(t)\right)_j > 0.$$

Since $\dot{\psi}^*(t) \equiv \beta_k$ for $t \in (t_k, t_{k+1})$, the last display and (2.4.9) yield

$$0 < c_i(\beta_k)_i - c_j(\beta_k)_j = c_i(\bar{\lambda}_{i,k} - \rho_k^{(i)} \bar{\mu}_{i,k}) - c_j \bar{\lambda}_{j,k}. \tag{2.4.18}$$

Note that Step 1 amounts to claiming that

$$\begin{aligned}\pi(\bar{X}(t)) \text{ cannot be written as } B \cup C, \text{ where } B \text{ is a strict}\\ \text{subset of } A_k \text{ and } C \subset A \setminus A_k \text{ is non-empty.}\end{aligned} \tag{2.4.19}$$

Indeed, note that $\pi(\bar{X}(t))$ can always be written as $\pi(\bar{X}(t)) = \bar{B} \cup \bar{C}$ where $\bar{B} \subset A_k$ and $\bar{C} \cap A_k = \emptyset$. The sets $\bar{B}, \bar{C}$ are uniquely determined by

$$\bar{B} = \pi(\bar{X}(t)) \cap A_k, \qquad \bar{C} = \pi(\bar{X}(t)) \cap (A \setminus A_k).$$

Then Step 1 amounts to claiming that either $\bar{B} = A_k$, or $\bar{B}$ is a strict subset of A_k and $\bar{C} = \emptyset$. This is clearly equivalent to (2.4.19).

We will prove (2.4.19) by contradiction and assume that there exists an $s \in (t_k, t_{k+1} \wedge \tau_k)$ such that $\pi(\bar{X}(s))$ can be written as such a union $B \cup C$. Note that C must contain at least one non-zero element, since otherwise $C = \{0\}$ and by Remark 2.2.2 $B \cup C = \{0, 1, 2, \ldots, d\}$ or $B = \{1, 2, \ldots, d\}$, which contradicts the assumption that B is a strict subset of A_k. Let

$$\bar{t} \doteq \sup \left\{ t \leq s : \pi(\bar{X}(t)) \cap (A_k \setminus B) \neq \emptyset \right\}.$$

We claim that $\bar{t} \in [t_k, s)$ and $\pi(\bar{X}(\bar{t})) \cap (A_k \setminus B) \neq \emptyset$. Indeed, $\bar{t} \geq t_k$ is trivial since $\pi(\bar{X}(t_k)) = A \supset A_k$ and $A_k \setminus B$ is non-empty. Thanks to the upper semicontinuity of π and the continuity of $\bar{X}$, there exists a small neighborhood of s such that for any t in this small neighborhood $\pi(\bar{X}(t)) \subset \pi(\bar{X}(s)) = B \cup C$. It follows readily that $\bar{t} < s$. An analogous use of upper semicontinuity shows that $\pi(\bar{X}(\bar{t})) \cap (A_k \setminus B) \neq \emptyset$.

Fix $i \in \pi(\bar{X}(\bar{t})) \cap (A_k \setminus B)$. Note that $i \notin \pi(\bar{X}(t))$ for $t \in (\bar{t}, s)$. Thus, for every $j \in C$ such that $j \neq 0$ and $t \in (\bar{t}, s)$, it follows from equations (2.4.17) and (2.4.18) that

$$\begin{aligned}\frac{d}{dt}\left[c_i((\bar{X}(t))_i - c_j((\bar{X}(t))_j\right] &= c_i\bar{\lambda}_{i,k} - c_j(\bar{\lambda}_{j,k} - \bar{\mu}_{j,k}h_j(t)) \\ &= [c_i(\bar{\lambda}_{i,k} - \rho_k^{(i)}\bar{\mu}_{i,k}) - c_j\bar{\lambda}_{j,k}] \\ &\qquad + [c_i\rho_k^{(i)}\bar{\mu}_{i,k} + c_j\bar{\mu}_{j,k}h_j(t)] \\ &> 0.\end{aligned}$$

Therefore,

$$0 \geq c_i((\bar{X}(s))_i - c_j((\bar{X}(s))_j > c_i((\bar{X}(\bar{t}))_i - c_j((\bar{X}(\bar{t}))_j \geq 0,$$

here the first inequality holds since $j \in C \subset \pi(\bar{X}(s))$ and the last inequality is due to $i \in \pi(\bar{X}(\bar{t}))$. This is a contradiction. Thus (2.4.19) holds and Step 1 is completed.

We now show Step 2. Thanks to Step 1, it suffices to show that $|\pi(\bar{X}(t))| \geq |A_k|$ for $t \in (t_k, t_{k+1} \wedge \tau_k)$. Let $t^* \in (t_k, t_{k+1} \wedge \tau_k)$ be such that

$$|\pi(\bar{X}(t^*))| = \min\{|\pi(\bar{X}(t))| : t \in (t_k, t_{k+1} \wedge \tau_k)\},$$

and assume that $|\pi(\bar{X}(t^*))| < |A_k|$ for the purpose of getting a contradiction. It follows from Step 1 that $B \doteq \pi(\bar{X}(t^*))$ is a strict subset of A_k. Note that $B \subset \{1, 2, \ldots, d\}$ (otherwise $B = \{0, 1, \ldots, d\}$ by Remark 2.2.2, a clear contradiction). Thanks to the upper semicontinuity of π and the continuity of $\bar{X}$, there exists an open interval containing t^* such that $\pi(\bar{X}(s)) \subset \pi(\bar{X}(t^*))$ for all s in this interval. We will assume (a, b) to be the largest of such intervals. By the definition of t^*, $\pi(\bar{X}(s)) = \pi(\bar{X}(t^*)) = B$ for every $s \in (a, b) \cap (t_k, t_{k+1} \wedge \tau_k)$. Obviously $a \geq t_k$ and $\pi(\bar{X}(a))$ contains B as a strict subset by the upper semicontinuity of π. Therefore, for $s \in (a,b)$, $c_i(\bar{X}(s))_i = c_j(\bar{X}(s))_j$ if $i,j \in B$. However, (2.4.17) yields

$$\frac{d}{ds}(\bar{X}(s))_j = \begin{cases} \bar{\lambda}_{j,k} & \text{if } j \in A_k \setminus B,\ j \neq 0, \\ \bar{\lambda}_{j,k} - h_j(s)\bar{\mu}_{j,k} & \text{if } j \in B. \end{cases} \tag{2.4.20}$$

It follows that

$$c_i(\bar{\lambda}_{i,k} - h_i(s)\bar{\mu}_{i,k}) = c_j(\bar{\lambda}_{j,k} - h_j(s)\bar{\mu}_{j,k}) \ \text{ for } \ i,\ j \in B. \tag{2.4.21}$$

In addition, thanks to Proposition 2.4.6(b),

$$\sum_{i \in B} h_i(s) = 1. \tag{2.4.22}$$

Solving the system of equations (2.4.21)-(2.4.22), we obtain the unique solution

$$h_i(s) = \frac{\bar{\lambda}_{i,k}}{\bar{\mu}_{i,k}} - \frac{1}{c_i\bar{\mu}_{i,k}} \cdot \left(\sum_{j \in B} \frac{1}{c_j\bar{\mu}_{j,k}} \right)^{-1} \cdot \left(-1 + \sum_{j \in B} \frac{\bar{\lambda}_{j,k}}{\bar{\mu}_{j,k}} \right).$$

On the other hand, since $\pi(\psi^*(s)) \equiv A_k \supset B$ and $\dot{\psi}^*(s) \equiv \beta_k$ for $s \in (a,b)$, we have $c_i(\beta_k)_i = c_j(\beta_k)_j$ for every $i,j \in B$. Invoking (2.4.7) and (2.4.9), we arrive at

$$c_i(\bar{\lambda}_{i,k} - \rho_k^{(i)}\bar{\mu}_{i,k}) = c_j(\bar{\lambda}_{j,k} - \rho_k^{(j)}\bar{\mu}_{j,k}) \ \text{ for } \ i,\ j \in B, \tag{2.4.23}$$

and that

$$b \doteq \sum_{i \in B} \rho_k^{(i)} = 1 - \sum_{i \in A_k \setminus B} \rho_k^{(i)} < 1. \tag{2.4.24}$$

Similarly, one can solve (2.4.23)-(2.4.24) to uniquely determine $\{\rho_k^{(i)}\}$ for $i \in B$:

$$\rho_k^{(i)} = \frac{\bar{\lambda}_{i,k}}{\bar{\mu}_{i,k}} - \frac{1}{c_i\bar{\mu}_{i,k}} \cdot \left(\sum_{j \in B} \frac{1}{c_j\bar{\mu}_{j,k}} \right)^{-1} \cdot \left(-b + \sum_{j \in B} \frac{\bar{\lambda}_{j,k}}{\bar{\mu}_{j,k}} \right).$$

Since $b < 1$ it follows that $h_i(s) > \rho_k^{(i)}$ for all $i \in B$ and $s \in (a, b)$.

Assume for now that there exists $j \in A_k \setminus B$ such that $j \neq 0$. Then for $i \in B$ and $s \in (a, b)$ we have, thanks to (2.4.20) and (2.4.23),

$$\begin{aligned} \frac{d}{ds}\left[c_i(\bar{X}(s))_i - c_j(\bar{X}(s))_j\right] &= c_i\left[\bar{\lambda}_{i,k} - h_i(s)\bar{\mu}_{i,k}\right] - c_j\bar{\lambda}_{j,k} \\ &< c_i\left[\bar{\lambda}_{i,k} - \rho_k^{(i)}\bar{\mu}_{i,k}\right] - c_j\bar{\lambda}_{j,k} \\ &= -c_j\rho_k^{(j)}\bar{\mu}_{j,k} \\ &< 0. \end{aligned} \tag{2.4.25}$$

However, since $\pi(\bar{X}(a)) \neq B$, Step 1 implies that $\pi(\bar{X}(a)) \cap (A_k \setminus B) \neq \emptyset$. Pick any j in this set. It follows from (2.4.25) that for any $i \in B$,

$$0 \geq c_i(\bar{X}(a))_i - c_j(\bar{X}(a))_j > c_i(\bar{X}(t^*))_i - c_j(\bar{X}(t^*))_j.$$

This contradicts that $i \in B = \pi(\bar{X}(t^*))$. Therefore, we have $|\pi(\bar{X}(t))| \geq |A_k|$. It follows from Step 1 that $A_k \subset \pi(\bar{X}(t))$, for $t \in (t_k, t_{k+1} \wedge \tau_k)$.

It remains to consider the case where $A_k \setminus B = \{0\}$. In this case we necessarily have $B = \{1, 2, \ldots, d\}$ and $A_k = \{0, 1, \ldots, d\} = A$. It follows that $\psi^*(t) \equiv 0$ for $t \in (t_k, t_{k+1})$ and thus $\beta_k = 0$. Then for any $i \in B$ and $s \in (a, b)$ we have, thanks to (2.4.20) and (2.4.7),

$$\frac{d}{ds}\left[c_i(\bar{X}(s))_i\right] = c_i\left[\bar{\lambda}_{i,k} - h_i(s)\bar{\mu}_{i,k}\right] < c_i\left[\bar{\lambda}_{i,k} - \rho_k^{(i)}\bar{\mu}_{i,k}\right] = (\beta_k)_i = 0.$$

However, since $\pi(X(a))$ contains B as a strict subset, we must have $\pi(X(a)) = \{0, 1, \ldots, d\}$ or $X(a) = 0$. It follows then

$$0 = c_i(\bar{X}(a))_i > c_i(\bar{X}(t^*))_i.$$

This contradicts the non-negativity of $\bar{X}$, and again we have $|\pi(\bar{X}(t))| \geq |A_k|$. It follows from Step 1 that $A_k \subset \pi(\bar{X}(t))$, for $t \in (t_k, t_{k+1} \wedge \tau_k)$.

If $A = A_k$ then we finish the proof of Step 2 since $\pi(\bar{X}(t)) \subset A$. Consider the case when A_k is a strict subset of A, which in turn implies that $A_k \subset \{1, 2, \ldots, d\}$. Following an argument analogous to that leading to equations (2.4.21)–(2.4.22) and (2.4.23)–(2.4.24), we have, for every $i, j \in A_k$ and almost every $t \in (t_k, t_{k+1} \wedge \tau_k)$,

$$c_i(\bar{\lambda}_{i,k} - h_i(t)\bar{\mu}_{i,k}) = c_j(\bar{\lambda}_{j,k} - h_j(t)\bar{\mu}_{j,k}), \quad \sum_{j \in A_k} h_j(t) \leq 1. \tag{2.4.26}$$

$$c_i(\bar{\lambda}_{i,k} - \rho_k^{(i)}\bar{\mu}_{i,k}) = c_j(\bar{\lambda}_{j,k} - \rho_k^{(j)}\bar{\mu}_{j,k}), \quad \sum_{i \in A_k} \rho_k^{(i)} = 1. \tag{2.4.27}$$

One can solve these two equations like before to obtain $h_i(t) \leq \rho_k^{(i)}$ for every $i \in A_k$ and $t \in (t_k, t_{k+1} \wedge \tau_k)$, whence

$$\frac{d}{dt}\left[c_i(\bar{X}(t))_i - c_i(\psi^*(t))_i\right] = c_i(\bar{\lambda}_{i,k} - h_i(t)\bar{\mu}_{i,k}) - c_i(\bar{\lambda}_{i,k} - \rho_k^{(i)}\bar{\mu}_{i,k}) \geq 0.$$

It follows that for any $i \in A_k$,

$$c_i(\bar{X}(t))_i - c_i(\psi^*(t))_i \geq c_i(\bar{X}(t_k))_i - c_i(\psi^*(t_k))_i = 0, \tag{2.4.28}$$

or $(\bar{X}(t))_i \geq (\psi^*(t))_i$. However, since $\pi((\psi^*(t)) \equiv A_k \subset \{1, 2, \ldots, d\}$, we must have $\bar{X}(t) \neq 0$ and thus $\pi(\bar{X}(t)) \subset \{1, 2, \ldots, d\}$.

On the other hand, for any $j \in A \setminus A_k$ and $j \neq 0$, since $(\dot{\psi}^*(t))_j = (\beta_k)_j = \bar{\lambda}_{j,k}$ thanks to (2.4.9), we have

$$\frac{d}{dt}\left[c_j(\bar{X}(t))_j - c_j(\psi^*(t))_j\right] = c_j(\bar{\lambda}_{j,k} - h_j(t)\bar{\mu}_{j,k}) - c_j\bar{\lambda}_{j,k} \leq 0.$$

It follows, since $\bar{X}(t_k) = \psi^*(t_k)$ and (2.4.28), that for every $t \in (t_k, t_{k+1} \wedge \tau_k)$ and

$i \in A_k$,

$$c_j(\bar{X}(t))_j - c_j(\psi^*(t))_j \le 0 \le c_i(\bar{X}(t))_i - c_i(\psi^*(t))_i.$$

But $(\psi^*(t))_i > (\psi^*(t))_j$ by definition of A_k, and therefore $(\bar{X}(t))_i > (\bar{X}(t))_j$. It follows that $j \notin \pi(\bar{X}(t))$. Therefore $\pi(\bar{X}(t)) \equiv A_k$ and we finish Step 2.

The proof of Step 3 is simple. Assume first $A_k \subset \{1, 2, \ldots, d\}$. Since $\pi(\bar{X}(t)) = A_k$ we know that $\{h_i(t)\}$ satisfies the equation (2.4.26) except now $\sum_{j \in A_k} h_j(t) = 1$. Compared with equation (2.4.27), it follows easily that $h_j(t) = \rho_k^{(j)}$. Note that equations (2.4.9) and (2.4.12) imply

$$\sum_{i \in A_k} \beta_k^{(i)} \rho_k^{(i)} = \beta_k,$$

whence

$$\frac{d}{dt}\bar{X}(t) = \sum_{i \in A_k} \beta_k^{(i)} h_i(t) = \sum_{i \in A_k} \beta_k^{(i)} \rho_k^{(i)} = \beta_k = \dot{\psi}^*(t),$$

which implies $\bar{X}(t) = \dot{\psi}^*(t)$ for all $t \in (t_k, t_{k+1} \wedge \tau_k)$.

For the case where $A_k = \{0, 1, \ldots, d\}$, we must have $\bar{X}(t) = \psi^*(t) = 0$. It is not difficult to see that equations (2.4.26) and (2.4.27) reduce to

$$c_i(\bar{\lambda}_{i,k} - h_i(t)\bar{\mu}_{i,k}) = 0 = c_i(\bar{\lambda}_{i,k} - \rho_k^{(i)}\bar{\mu}_{i,k}), \qquad i = 1, 2, \ldots, d.$$

Thus $h_i(t) = \rho_k^{(i)}$ for all $i = 1, 2, \ldots, d$, and also $h_0(t) = \rho_k^{(0)}$ since $\sum_{i=0}^{d} h_i(t) = \sum_{i=0}^{d} \rho_k^{(i)} = 1$.

■

2.4.7 Analysis of the cost

In this section, we prove the Laplace lower bound, or inequality (2.4.4).

Thanks to (2.4.11) and (2.4.10),

$$\begin{aligned}
&\lim_{n\to\infty} E_{x_n}\left[\int_{t_k}^{t_{k+1}} \sum_{v\in\mathbb{V}} r(\bar{X}^n(t);v)\ell\left(\frac{\bar{r}(\bar{X}^n(t),t;v)}{r(\bar{X}^n(t);v)}\right)dt\right] \\
&= \lim_{n\to\infty} E_{x_n}\left[\int_{t_k}^{t_{k+1}} \sum_{i\in A_k, i\neq 0} \mu_i \ell\left(\frac{\bar{\mu}_{i,k}}{\mu_i}\right)\gamma_i^n(dt) + \sum_{j=1}^{d}\lambda_j\ell\left(\frac{\bar{\lambda}_{j,k}}{\lambda_j}\right)dt\right] \\
&= \int_{t_k}^{t_{k+1}}\left[\sum_{i\in A_k, i\neq 0} \mu_i \ell\left(\frac{\bar{\mu}_{i,k}}{\mu_i}\right)h_i(t) + \sum_{j=1}^{d}\lambda_j\ell\left(\frac{\bar{\lambda}_{j,k}}{\lambda_j}\right)\right]dt \\
&= (t_{k+1}-t_k)\left[\sum_{i\in A_k, i\neq 0} \mu_i \ell\left(\frac{\bar{\mu}_{i,k}}{\mu_i}\right)\rho_k^{(i)} + \sum_{j=1}^{d}\lambda_j\ell\left(\frac{\bar{\lambda}_{j,k}}{\lambda_j}\right)\right] \\
&\leq (t_{k+1}-t_k)\cdot[L^{A_k}(\beta_k)+\varepsilon] \\
&= \int_{t_k}^{t_{k+1}} L(\psi^*(t),\dot{\psi}^*(t))dt + (t_{k+1}-t_k)\varepsilon.
\end{aligned}$$

Summing over k and observing

$$\lim_{n\to\infty} E_{x_n}h(\bar{X}^n) = E_x h(\bar{X}) = h(\psi^*),$$

(2.4.4) follows readily. This completes the proof of inequality (2.4.4) and also the proof of Theorem 2.3.1. ■

2.5 Summary

In this chapter we analyze the large deviation properties for a class of systems with discontinuous dynamics, namely, the WSLQ policy for a network with multiple queues and a single server. A key observation that allows us to derive explicitly the large deviation rate function on path space is that the stability condition about the interface is automatically implied in the formulation of a general large deviation upper bound.

This is not an unprecedented situation, and indeed analogous results are proved in [2] for processes that model Jackson networks and head-of-the-line processor sharing. These problems also feature discontinuous statistics, though the discontinuities appear on the boundary of the state space rather than the interior, and whence alternative techniques based on Skorohod mapping can be used. With the inclusion of the WSLQ policy there are now a number of physically meaningful models with discontinuous statistics for which the upper bound of [7] is tight. An intriguing question that will be investigated elsewhere is whether there is a common property of these models that can be easily identified and recognized.

2.6 Proof of Lemma 2.4.1.

Before getting into the details of the construction, it is worth pointing out that the main difficulty in the proof is dealing with the situation where the function ψ may spiral into (or away from) a low dimensional interface while hitting higher dimensional interfaces infinitely many times in the process. More precisely, there can be a time t and a sequence $t_n \uparrow t$ (or $t_n \downarrow t$) such that $|\pi(\psi(t_n))| < |\pi(\psi(t))|$, but with the sets $\{\pi(\psi(t_n))\}$ different for successive n. The problem is how to approximate ψ on a small neighborhood of t with a small cost, which is made difficult by the fact that the domain of finiteness of $L(x, \cdot)$ is not continuous in x.

Throughout this section we assume that ψ is Lipschitz continuous. This is without loss of generality, since for a given continuous function ψ and any $\delta > 0$, there exists a Lipschitz continuous function ζ such that $\|\zeta - \psi\|_\infty \leq \delta$ and $I_x(\zeta) \leq I_x(\psi) + \delta$. The proof of this claim is based on a time-rescaling argument very much analogous to that of [5, Lemma 6.5.3], and we omit the details.

2.6.1 Dividing the time interval

The approximation requires a suitable division of the time interval $[0, 1]$. The following results are useful in proving the main result of this section, namely, Lemma 2.6.4.

Note that whenever we say two intervals are "non-overlapping", it means that the two intervals cannot have common interiors but may have a same endpoint.

Lemma 2.6.1. *Consider a non-empty closed interval* $[a, b]$ *and assume*

$$k \doteq \max\{|\pi(\psi(t))| : t \in [a, b]\} \geq 2.$$

Then for arbitrary $\sigma > 0$*, there exists a finite collection of non-overlapping intervals* $\{[\alpha_j, \beta_j]\}$ *such that*

1. $\alpha_j, \beta_j \in [a, b]$ *and* $0 \leq \beta_j - \alpha_j \leq \sigma$ *for every* j*.*
2. $|\pi(\psi(t))| \leq k - 1$ *for every* $t \in [a, b] \setminus \cup_j [\alpha_j, \beta_j]$*.*

Lemma 2.6.2. *Consider a non-empty interval* (a, b) *such that for some* $k \geq 2$*,* $|\pi(\psi(t))| \leq k$ *for* $t \in (a, b)$ *and* $|\pi(\psi(a))| \wedge |\pi(\psi(b))| \geq k + 1$*. Then for arbitrary* $\sigma > 0$ *and* $\varepsilon > 0$*, there exist a finite collection of non-overlapping intervals* $\{[\alpha_j, \beta_j]\}$ *and* $\varepsilon_1, \varepsilon_2 \in [0, \varepsilon)$ *such that*

1. $\alpha_j, \beta_j \in [a + \varepsilon_1, b - \varepsilon_2]$ *and* $0 \leq \beta_j - \alpha_j \leq \sigma$ *for each* j*,*
2. *for* $t \in (\alpha_j, \beta_j)$*,* $\pi(\psi(t)) \subset \pi(\psi(\alpha_j)) = \pi(\psi(\beta_j))$*,*
3. $|\pi(\psi(\alpha_j))| = |\pi(\psi(\beta_j))| = k$ *for each* j*,*
4. *for any* $t \in (a, a + \varepsilon_1]$*,* $\pi(\psi(t))$ *is a strict subset of* $\pi(\psi(a))$ *and* $|\pi(\psi(t))| \leq k \leq |\pi(\psi(a + \varepsilon_1))|$*,*
5. *for any* $t \in [b - \varepsilon_2, b)$*,* $\pi(\psi(t))$ *is a strict subset of* $\pi(\psi(b))$ *and* $|\pi(\psi(t))| \leq k \leq |\pi(\psi(b - \varepsilon_2))|$*,*
6. $|\pi(\psi(t))| \leq k - 1$ *for every* $t \notin \cup_j [\alpha_j, \beta_j] \cup (a, a + \varepsilon_1] \cup [b - \varepsilon_2, b)$*.*

Lemma 2.6.3. *Consider a non-empty interval $[a, b)$ such that for some $2 \le k \le d$, $|\pi(\psi(t))| \le k$ for $t \in [a, b)$ and $|\pi(\psi(b))| \ge k+1$. Then for arbitrary $\sigma > 0$ and $\varepsilon > 0$, there exist a finite collection of non-overlapping intervals $\{[\alpha_j, \beta_j]\}$ and $\varepsilon_2 \in [0, \varepsilon)$ such that*

1. *$\alpha_j, \beta_j \in [a, b - \varepsilon_2]$ and $0 \le \beta_j - \alpha_j \le \sigma$ for each j,*
2. *for $t \in (\alpha_j, \beta_j)$, $\pi(\psi(t)) \subset \pi(\psi(\alpha_j)) = \pi(\psi(\beta_j))$,*
3. *$|\pi(\psi(\alpha_j))| = |\pi(\psi(\beta_j))| = k$ for each j,*
4. *for every $t \in [b - \varepsilon_2, b)$, $\pi(\psi(t))$ is a strict subset of $\pi(\psi(b))$ and $|\pi(\psi(t))| \le k \le |\pi(\psi(b - \varepsilon_2))|$,*
5. *$|\pi(\psi(t))| \le k - 1$ for every $t \notin \cup_j [\alpha_j, \beta_j] \cup [b - \varepsilon_2, b)$.*

Symmetric results hold for a non-empty interval $(a, b]$ such that for some $k \ge 2$, $|\pi(\psi(t))| \le k$ for $t \in (a, b]$ and $|\pi(\psi(a))| \ge k + 1$.

We will only provide the details of the proof for Lemma 2.6.2. The proofs for Lemma 2.6.1 and Lemma 2.6.3 are very similar (indeed, simpler versions of the proof for Lemma 2.6.2), and thus omitted.

If the set $\{t \in (a, b) : |\pi(\psi(t))| = k\}$ is empty, then the claim holds trivially since we can let $\varepsilon_1 = \varepsilon_2 = 0$ and $\{[\alpha_j, \beta_j]\} \doteq \emptyset$. Assume from now on that this set is non-empty. We first define ε_1 and ε_2. Let

$$\bar{a} \doteq \inf\{t \in (a, b) : |\pi(\psi(t))| = k\}, \quad \bar{b} \doteq \sup\{t \in (a, b) : |\pi(\psi(t))| = k\}.$$

If $\bar{a} = a$, then the upper semicontinuity of π implies that there exists $0 < \varepsilon_1 < \varepsilon$ such that $|\pi(\psi(a + \varepsilon_1))| = k$ and $\pi(\psi(t))$ is a subset (whence a strict subset) of $\pi(\psi(a))$ for every $t \in (a, a + \varepsilon_1]$. If $\bar{a} > a$ then we let $\varepsilon_1 = 0$. ε_2 is defined in a completely analogous fashion. It is easy to see that parts 4 and 5 of the claim are satisfied. We will define α_j and β_j recursively.

1. Let $\alpha_1 \doteq a + \varepsilon_1$ if $\bar{a} = a$ and $\alpha_1 \doteq \bar{a}$ if $\bar{a} > a$. Clearly, in either case $\alpha_1 \in [a + \varepsilon_1, b - \varepsilon_2]$. Moreover, we have $|\pi(\psi(\alpha_1))| = k$. Indeed, when $\bar{a} = a$ it follows from the definition, and when $\bar{a} > a$, it follows from a simple argument by contradiction, thanks to the upper semicontinuity of π and the assumption that $|\pi(\psi(t))| \leq k$ for every $t \in (a, b)$.

2. Suppose now $\alpha_j \in [a + \varepsilon_1, b - \varepsilon_2]$ is given such that $|\pi(\psi(\alpha_j))| = k$. We wish to define β_j. Let

$$\begin{aligned} t_j &\doteq \inf\left\{t \in (\alpha_j, b) : \pi(\psi(t)) \not\subset \pi(\psi(\alpha_j))\right\}, \\ s_j &\doteq \sup\{t \in [\alpha_j, t_j \wedge (\alpha_j + \sigma)) : \pi(\psi(t)) = \pi(\psi(\alpha_j))\}, \\ \beta_j &\doteq s_j \wedge (b - \varepsilon_2). \end{aligned}$$

   It is easy to check that $\pi(\psi(s_j)) = \pi(\psi(\alpha_j))$. We claim that $\pi(\psi(\beta_j)) = \pi(\psi(\alpha_j))$. This is trivial when $\beta_j = s_j$. Assume for now that $\beta_j < s_j$. Then we must have $s_j > b - \varepsilon_2$ and $\beta_j = b - \varepsilon_2$. This could happen only if $\varepsilon_2 > 0$, in which case $|\pi(\psi(b - \varepsilon_2))| = k$, or $|\pi(\psi(\beta_j))| = k$. But $\pi(\psi(\beta_j)) \subset \pi(\psi(\alpha_j))$ since $\beta_j < s_j \leq t_j$. Thus we must have $\pi(\psi(\alpha_j)) = \pi(\psi(\beta_j))$. It is clear now that parts 1, 2, and 3 of the lemma holds.

3. If $\beta_j = \bar{b}$ (when $\bar{b} < b$) or $\beta_j = b - \varepsilon_2$ (when $\bar{b} = b$), we stop. Otherwise, define $\alpha_{j+1} \doteq \inf\{t \in (\beta_j, b) : |\pi(\psi(t))| = k\}$. Then $\alpha_{j+1} \in [a + \varepsilon_1, b - \varepsilon_2]$ and $|\pi(\psi(\alpha_{j+1}))| = k$. Now repeat step 2.

Note that part 6 holds by the construction. It only remains to show that the construction will terminate in finitely many steps. Observe that $\alpha_{j+1} \geq t_j \wedge (\alpha_j + \sigma)$, since by definition, for every $t \in (\beta_j, t_j \wedge (\alpha_j + \sigma))$, $\pi(\psi(t))$ is a strict subset of $\pi(\psi(\alpha_j))$ and whence $|\pi(\psi(t))| < k$. Therefore, $\alpha_{j+1} - \alpha_j \geq (t_j - \alpha_j) \wedge \sigma$, and it suffices to show that $t_j - \alpha_j$ is uniformly bounded away from 0.

We should first consider the case $k \leq d-1$. Observe that in the construction we indeed have $\alpha_j, \beta_j \in \bar{I} \doteq [(a+\varepsilon_1) \vee \bar{a}, (b-\varepsilon_2) \wedge \bar{b}] \subset (a,b)$. Define

$$c \doteq \inf \left\{ \max_{i=1,\ldots,d} c_i(\psi(t))_i - \max_{i \notin \pi(\psi(t))} c_i(\psi(t))_i : t \in \bar{I}, |\pi(\psi(t))| = k \right\}.$$

We claim that $c > 0$. If this is not the case, there exists a sequence of $\{t_n\} \subset \bar{I}$ such that

$$\max_{i=1,\ldots,d} c_i(\psi(t_n))_i - \max_{i \notin \pi(\psi(t_n))} c_i(\psi(t_n))_i \downarrow 0.$$

One can find a subsequence of $\{t_n\}$, still denoted by $\{t_n\}$, such that $t_n \to t^* \in I$ and $\pi(\psi(t_n)) \equiv B$ for some $B \subset \{1, 2, \ldots, d\}$ with $|B| = k$. Thanks to the upper semicontinuity of π, $B \subset \pi(\psi(t^*))$. But $|\pi(\psi(t^*))| \leq k$, thus $B = \pi(\psi(t^*))$. However, for every $j \in B$,

$$\begin{aligned} 0 &= \lim_n \left[\max_{i=1,\ldots,d} c_i(\psi(t_n))_i - \max_{i \notin \pi(\psi(t_n))} c_i(\psi(t_n))_i \right] \\ &= \lim_n \left[c_j(\psi(t_n))_j - \max_{i \notin B} c_i(\psi(t_n))_i \right] \\ &= c_j(\psi(t^*))_j - \max_{i \notin B} c_i(\psi(t^*))_i \\ &> 0, \end{aligned}$$

a contradiction. Therefore $c > 0$.

Now, by the definition of t_j, there exists $l \notin \pi(\psi(\alpha_j))$ such that $l \in \pi(\psi(t_j))$. Therefore, for every $i \in \pi(\psi(\alpha_j))$,

$$c_i(\psi(t_j))_i - c_l(\psi(t_j))_l \leq 0.$$

But $c_i(\psi(\alpha_j))_i - c_l(\psi(\alpha_j))_l \geq c$ and ψ is Lipschitz continuous. It follows readily that

$t_j - \alpha_j$ is uniformly bounded away from 0.

The case $k = d$ can be treated in a completely analogous fashion with

$$c \doteq \inf \left\{ \max_{i=1,\ldots,d} c_i(\psi(t))_i : t \in \bar{I}, |\pi(\psi(t))| = d \right\}.$$

We omit the details. ■

The next lemma is the main result of this section, from which one can construct the approximating function in $\mathcal{N}$. The intervals $\{E_i\}$ in the lemma are indeed introduced to take care of the "spiraling" problem.

Lemma 2.6.4. *Consider the interval* $[0,1]$. *Given* $\sigma > 0$, *there exist two finite collections of non-overlapping intervals* $\{I_j\}$ *and* $\{E_i\}$ *such that*

1. $I_j \subset [0,1]$ *is of the form* $[a_j, b_j]$ *with* $0 \le b_j - a_j \le \sigma$ *for each* j,
2. $E_i \subset [0,1]$ *is either of the form* $[z_i, d_i)$ *or* $(z_i, d_i]$, *and* $\sum_i (d_i - z_i) \le \sigma$,
3. $|\pi(\psi(t))| = 1$ *for all* $t \notin (\cup_j I_j) \cup (\cup_i E_i)$,
4. *for every* j *and every* $t \in (a_j, b_j)$, $\pi(\psi(t)) \subset \pi(\psi(a_j)) = \pi(\psi(b_j))$,
5. *for every* j, $|\pi(\psi(a_j))| = |\pi(\psi(b_j))| \ge 2$,
6. *for all* $t \in E_i = (z_i, d_i]$, $\pi(\psi(t))$ *is a strict subset of* $\pi(\psi(z_i))$ *and* $|\pi(\psi(t))| \le |\pi(\psi(d_i))|$,
7. *for all* $t \in E_i = [z_i, d_i)$, $\pi(\psi(t))$ *is a strict subset of* $\pi(\psi(d_i))$ *and* $|\pi(\psi(t))| \le |\pi(\psi(z_i))|$.

Let $N \doteq \max\{|\pi(\psi(t))| : t \in [0,1]\}$. If $N = 1$ then the claim holds trivially. Assume from now on that $N \ge 2$. The construction can be done in N steps.

Step 1. Apply Lemma 2.6.1 to the interval $[0,1]$ with $k = N$. This will produce a collection of closed intervals $\{I_j^{(1)}\}$. Let $U_1 \doteq [0,1] \setminus \cup_j I_j^{(1)}$.

Step 2. Note that U_1 is the union of finitely many non-overlapping intervals. These intervals are of two types. They are either open intervals that satisfy the conditions of Lemma 2.6.2 with $k = N - 1$, or intervals of type $[0, a)$ or $(b, 1]$ that satisfy the conditions of Lemma 2.6.3. Apply the corresponding lemma to these intervals to generate a collection of closed intervals $\{I_j^{(2)}\}$ and intervals $\{E_i^{(2)}\}$ of type $[a, b)$ or $(b, a]$. Since the number of intervals in $\{E_i^{(2)}\}$ is at most twice the number of intervals in U_1, one can choose ε in Lemma 2.6.2 and Lemma 2.6.3 so small that the total Lebesgue measure of $\cup_i E_i^{(2)}$ is bounded from above by σ/N. Define $U_2 \doteq U_1 \setminus ((\cup_j I_j^{(2)}) \cup (\cup_i E_i^{(2)}))$.

Step m. For $3 \leq m \leq N - 1$, the procedure of Step m is just like Step 2, except U_1 is replaced by U_{m-1}.

Step N. Note that U_{N-1} consists of intervals on which $\pi(\psi(\cdot))$ is a singleton. We stop the construction.

Let $\{I_j\} \doteq \cup_l \{I_j^{(l)}\}$ and $\{E_i\} \doteq \cup_l \{E_j^{(l)}\}$, and it is not difficult to see that $\{I_j\}$ and $\{E_i\}$ have the required properties. ∎

2.6.2 Construction of the approximating function

In this section we construct an approximating function of ψ parameterized by σ. Note that the set $[0, 1] \setminus ((\cup_j I_j) \cup (\cup_i E_i))$ is the union of finitely many non-overlapping intervals, and we will denote these intervals by $\{G_k\}$. Without loss of generality assume that the length of each interval G_k is bounded above by σ (if necessary, we can divide G_k into the union of several smaller intervals each with a length less than σ).

Define a piece-wise constant function u^σ on interval $[0, 1]$ as follows. Given any interval $D \in \{I_j\} \cup \{E_i\} \cup \{G_k\}$ with positive length, define for every $t \in D$,

$$u^\sigma(t) \doteq \frac{1}{\text{length of } D} \int_D \dot{\psi}(s)\, ds.$$

Then the candidate approximating function will be

$$\psi^\sigma(t) \doteq \psi(0) + \int_0^t u^\sigma(s) ds.$$

Note that $\psi^\sigma(t)$ coincides with $\psi(t)$ at those t that are end points of intervals $\{I_j\}$, $\{E_i\}$, and $\{G_k\}$. Furthermore, ψ^σ is affine on each of these intervals.

We claim that $\psi^\sigma \in \mathcal{N}$. All we need is to show is that $\pi(\psi^\sigma(t))$ remain unchanged in the interior of each interval. This is immediate from the following result, whose proof is simple and straightforward, and thus omitted.

Lemma 2.6.5. *Let ϕ be an affine function on interval $[a, b]$. If $\pi(\phi(a)) \cap \pi(\phi(b)) \neq \emptyset$, then for every $t \in (a, b)$ we have $\pi(\phi(t)) = \pi(\phi(a)) \cap \pi(\phi(b))$.*

2.6.3 The analysis of the rate function

Since the length of any interval in $\{I_j\}$, $\{E_i\}$, and $\{G_k\}$ is bounded from above by σ, $\lim_{\sigma\to 0} \|\psi^\sigma - \psi\|_\infty = 0$. Therefore it only remains to show that given $\delta > 0$ we have $I_x(\psi^\sigma) \leq I_x(\psi) + \delta$ for σ small enough. Clearly we can assume $I_x(\psi) < \infty$ hereafter.

The analysis for intervals in $\{I_j\}$ and $\{G_k\}$ is simple. Consider an interval $D \in \{I_j\} \cup \{G_k\}$, and denote the interior of D by (a, b) [D itself could be $[a, b]$, (a, b), $(a, b]$, or $[a, b)$]. Since $\psi(a) = \psi^\sigma(a)$ and $\psi(b) = \psi^\sigma(b)$ by construction, it follows from Lemma 2.6.4 and Lemma 2.6.5 that

$$\pi(\psi(t)) \subset \pi(\psi(a)) \cap \pi(\psi(b)) = \pi(\psi^\sigma(a)) \cap \pi(\psi^\sigma(b)) = \pi(\psi^\sigma(t)) \doteq B.$$

for every $t \in (a, b)$. Note that $L^B(\beta) \leq L^A(\beta)$ for every β whenever $A \subset B$, thanks to Lemma 2.4.3. Therefore,

$$\int_a^b L(\psi(t), \dot{\psi}(t))\, dt = \int_a^b L^{\pi(\psi(t))}(\dot{\psi}(t))\, dt \geq \int_a^b L^B(\dot{\psi}(t))\, dt.$$

Furthermore, for $t \in (a, b)$, the construction of ψ^σ implies that

$$\psi^\sigma(t) \equiv \frac{\psi(b) - \psi(a)}{b - a} \doteq v.$$

Thanks to the convexity of L^B and Jensen's inequality, we arrive at

$$\int_a^b L^B(\dot{\psi}(t))\, dt \geq (b-a)L^B(v) = \int_a^b L(\psi^\sigma(t), \psi^\sigma(t))dt,$$

whence

$$\int_a^b L(\psi(t), \dot{\psi}(t))\, dt \geq \int_a^b L(\psi^\sigma(t), \psi^\sigma(t))dt. \tag{2.6.1}$$

for any σ.

Now we turn to analyze the intervals in $\{E_i\}$. The next two lemmas claim that over any interval, say $D \in \{E_i\}$, $\dot{\psi}^\sigma$ will always be in the domain of finiteness of the local rate function L. The first lemma considers the case where $D = (a, b]$ and the second considers the case where D is of form $[a, b)$.

Lemma 2.6.6. *Suppose that $I_x(\psi) < \infty$ and there exists an interval $(a, b] \subset [0, 1]$ such that $\pi(\psi(t))$ is a strict subset of $\pi(\psi(a))$ for $t \in (a, b]$. Then $(\psi(a))_j \leq (\psi(b))_j$ for all $j \in \{1, \ldots, d\}$.*

For notational simplicity, we assume $c_1 = c_2 = \cdots = c_d$. The case of general $(c_1, \ldots, c_d)$ is shown in exactly the same fashion with $(\psi(t))_j$ replaced by $c_j(\psi(t))_j$, and whence omitted.

Let $B \doteq \pi(\psi(a))$. For $j \notin B$, we have $j \notin \pi(\psi(t))$ and thus $\dot{\psi}(t) \geq 0$ for almost every $t \in (a, b]$, thanks to Remark 2.4.5 and the assumption that $I_x(\psi)$ is finite.

Therefore for $j \notin B$ the claim holds. Now we consider those $j \in B$. Fix arbitrarily $\varepsilon > 0$ and let

$$t^* \doteq \inf\{t > a : (\psi(t))_j < (\psi(a))_j - \varepsilon \text{ for some } j \in B\} \wedge b.$$

It suffices to show that $t^* = b$ always holds. Indeed, if this is the case, we have $(\psi(b))_j \geq (\psi(a))_j - \varepsilon$ for every $j \in B$. Since ε is arbitrary, we arrive at the desired inequality for $j \in B$.

We will argue by contradiction and assume $t^* < b$. It follows that $(\psi(t^*))_j \geq (\psi(a))_j - \varepsilon$ for every $j \in B$, and there exist $j^* \in B$ and a sequence $t_n \downarrow t^*$ such that

$$(\psi(t_n))_{j^*} < (\psi(a))_{j^*} - \varepsilon. \tag{2.6.2}$$

In particular, $(\psi(t^*))_{j^*} = (\psi(a))_{j^*} - \varepsilon$.

We claim that $(\psi(t^*))_j = (\psi(a))_j - \varepsilon$ holds for every $j \in B$. If this is not the case, then there exists (abusing the notation) $j \in B$ such that

$$(\psi(t^*))_j > (\psi(a))_j - \varepsilon.$$

Since $j, j^* \in B = \pi(\psi(a))$ we have $(\psi(a))_j = (\psi(a))_{j^*}$, and whence $(\psi(t^*))_j > (\psi(t^*))_{j^*}$. Therefore we can find a small interval, say $[t^*, t^* + \delta)$ such that for any t in this interval we have $(\psi(t))_j > (\psi(t))_{j^*}$, which in turn implies $j^* \notin \pi(\psi(t))$. Thanks to the finiteness of $I_x(\psi)$ and Remark 2.4.5, $(\dot{\psi}(t))_{j^*} \geq 0$ for almost every $t \in [t^*, t^* + \delta)$. In particular, $(\psi(t))_{j^*} \geq (\psi(t^*))_{j^*} = (\psi(a))_{j^*} - \varepsilon$ for all $t \in [t^*, t^* + \delta)$, which contradicts equation (2.6.2) for large n. Therefore $(\psi(t^*))_j = (\psi(a))_j - \varepsilon$ holds for every $j \in B$.

Since $B = \pi(\psi(a))$, it follows that $(\psi(t^*))_j$ takes the same value for every $j \in B$. This contradicts the assumption that $\pi(\psi(t^*))$ is a strict subset of B. We complete the proof. ∎

Lemma 2.6.7. *Suppose that $I_x(\psi) < \infty$ and there exists an interval $[a, b) \subset [0, 1]$ such that $\pi(\psi(t))$ is a strict subset of $\pi(\psi(b))$ for every $t \in [a, b)$. Let $B \doteq \pi(\psi(a))$. If $|\pi(\psi(t))| \leq |B|$ for every $t \in [a, b)$, then $(\psi(a))_j \leq (\psi(b))_j$ for all $j \notin B$ and $j \neq 0$.*

As in the proof of Lemma 2.6.6 we assume without loss of generality that $c_1 = c_2 = \cdots = c_d = 1$. Let $A \doteq \pi(\psi(b))$. For $j \notin A$, $(\dot{\psi}(t))_j \geq 0$ for almost every $t \in [a, b)$ since $j \notin \pi(\psi(t))$. Thus the claim holds for $j \notin A$. It remains to show for those $j \in A \setminus B$ such that $j \neq 0$. We will argue by contradiction and assume that there is a $j^* \in A \setminus B$ such that $(\psi(a))_{j^*} > (\psi(b))_{j^*}$.

For each $i \in B$ let $t_i \doteq \inf \{t > a : (\psi(t))_i \leq (\psi(a))_{j^*}\}$. Note that $(\psi(a))_i > (\psi(a))_{j^*}$ since $i \in B = \pi(\psi(a))$ and $j^* \notin B$, whence $t_i > a$. Similarly, since $j^* \in A = \pi(\psi(b))$, $(\psi(b))_i \leq (\psi(b))_{j^*} < (\psi(a))_{j^*}$, whence $t_i < b$. It follows that $(\psi(t_i))_i = (\psi(a))_{j^*}$.

We claim that $i \in \pi(\psi(t_i))$ for all $i \in B$. Otherwise, there exists a small positive number δ such that $i \notin \pi(\psi(t))$ for $t \in (t_i - \delta, t_i + \delta)$. Thanks to the assumption that $I_x(\psi)$ is finite, $(\dot{\psi}(t))_i \geq 0$ for almost every $t \in (t_i - \delta, t_i + \delta)$. In particular, $(\psi(t_i - \delta))_i \leq (\psi(t_i))_i = (\psi(a))_{j^*}$, which contradicts the definition of t_i. Therefore, $i \in \pi(\psi(t_i))$ for all $i \in B$. An immediate consequence is that for any $i, k \in B$, $(\psi(t_i))_k \leq (\psi(t_i))_i = (\psi(a))_{j^*}$, whence $t_k \leq t_i$ by definition. Therefore, for all $i, k \in B$, $t_i = t_k \doteq t^*$ and $B \subset \pi(\psi(t^*))$. However, since $|\pi(\psi(t^*))| \leq |B|$ by assumption, we have necessarily $B = \pi(\psi(t^*))$.

Since $j^* \notin B = \pi(\psi(t^*))$, we have $(\psi(t^*))_{j^*} < (\psi(a))_{j^*}$. Define

$$s \doteq \sup \{t \in [a, t^*) : (\psi(t))_{j^*} \geq (\psi(a))_{j^*}\} \wedge t^*.$$

Clearly $s \in [a, t^*)$ and $(\psi(s))_{j^*} = (\psi(a))_{j^*}$. By the definitions of s and t_i and that

$t^* \equiv t_i$ for all $i \in B$, we have, for any $t \in (s, t^*)$ and any $i \in B$,

$$(\psi(t))_{j^*} < (\psi(a))_{j^*} < (\psi(t))_i.$$

Therefore, $j^* \notin \pi(\psi(t))$ and whence $(\dot{\psi}(t))_{j^*} \geq 0$ for almost every $t \in (s, t^*)$. In particular, $(\psi(a))_{j^*} = (\psi(s))_{j^*} \leq (\psi(t^*))_{j^*} < (\psi(a))_{j^*}$, a contradiction. We complete the proof. ■

We are now ready to show that $I_x(\psi^\sigma) \leq I_x(\psi) + \delta$ for σ small enough. Let M be the Lipschitz constant for ψ, and define

$$C \doteq \max_{A \subset \{1,\ldots,d\}} \sup\{L^A(\beta) : \beta \in \mathrm{dom}(L^A), \|\beta\| \leq M\}.$$

That C is finite follows easily from Lemma 2.4.3 and the definition of ℓ.

We now analyze the intervals in $\{E_i\}$. If $E_i = [z_i, d_i)$, then by Lemma 2.6.5 and Lemma 2.6.4, $\pi(\psi^\sigma(t)) = \pi(\psi(z_i)) \doteq B$ for all $t \in (z_i, d_i)$. Since

$$\dot{\psi}^\sigma(t) \equiv \frac{\psi(d_i) - \psi(z_i)}{d_i - z_i} \doteq v,$$

it follows from Lemma 2.6.7 and Remark 2.4.5 that $v \in \mathrm{dom}(L^B)$. Therefore

$$\int_{z_i}^{d_i} L(\psi^\sigma(t), \dot{\psi}^\sigma(t))\, dt = (d_i - z_i) L^B(v) \leq C(d_i - z_i). \tag{2.6.3}$$

The same inequality holds for the case $E_i = (z_i, d_i]$, where Lemma 2.6.6 is invoked in place of Lemma 2.6.7.

By (2.6.1), (2.6.3), the non-negativity of L, and Lemma 2.6.4, we have

$$\begin{aligned}
\int_0^1 L(\psi^\sigma(t), \dot{\psi}^\sigma(t))\, dt &\leq \int_0^1 L(\psi(t), \dot{\psi}(t))\, dt + C \sum_i (d_i - z_i) \\
&\leq \int_0^1 L(\psi(t), \dot{\psi}(t))\, dt + C\sigma.
\end{aligned}$$

Choose $\sigma < \delta/C$ and we complete the proof. ■

Chapter 3

Importance Sampling and Large Deviations Decay Rate for a Buffer Overflow Event in the WSLQ Policy

3.1 Introduction

In the previous chapter it was shown that the queue length process in a weighted-serve-the-longest system satisfies a large deviations principle with the local rate function $L(\cdot, \cdot)$ (see section 2.3.1). The focus in this section will be on estimating the probability of one particular event.

The present chapter concentrates on a class of individual buffer overflow probabilities and resolves two questions: (1) the explicit identification of the exponential decay rate of such overflow probabilities; (2) the construction of asymptotically optimal importance sampling schemes for fast simulation of such rare events. These two questions turn out to be intimately connected in the following sense. To build efficient importance sampling schemes, we use the game/subsolution approach [15] and construct suitable subsolutions to the associated Isaacs equations. Not only do these subsolutions lead to asymptotically optimal importance sampling schemes, they also play an essential role in explicitly identifying the exponential decay rate of the overflow probabilities via a verification argument.

The serve-the-longest policy has been studied in the context of wireless communications, see for example [3, 28] and their references. There are differences between the stochastic processes used in the wireless literature and the processes used in our work, the most significant of which is that the dynamics in the wireless context are not Markovian, but rather modulated by an exogenous Markov process. However the methods used to deal with the discontinuous interfaces in this thesis can be combined with those presented in [15] for Markov modulated rates to analyze such models.

This chapter is organized as follows. In Section 2 the model dynamics and problem formulation are described. Section 3 presents the large deviations result regarding the overflow probabilities. A brief review of importance sampling is conducted in section 4, and the associated Isaacs equation is formally derived in Section 5. A classical subsolution to the Isaacs equation and the corresponding importance sampling scheme are constructed in Section 8. In Section 7 we derive the exponential decay rate of the overflow probabilities and establish the asymptotic optimality of the said importance sampling scheme. Numerical results are presented in Section 8. Many technical proofs are collected in section 9.

3.2 Problem Formulation

The system model consists of a single server and d classes of customers. Customers of class i, buffered at queue i, arrive according to a Poisson process with rate λ_i. The service time for a class i customer is exponentially distributed with rate μ_i. Upon the completion of service, the customer leaves the system. We will assume that the system satisfies the stability condition

$$\sum_{j=1}^{d} \frac{\lambda_j}{\mu_j} < 1. \tag{3.2.1}$$

The service policy is determined according to the following WSLQ discipline. Let c_i be the weight associated with class i. If the size of queue i is q_i, then the weighted length for queue i is defined as $c_i q_i$. Under the WSLQ policy, the queue with the maximal weighted length is assigned the service priority. When there are multiple queues with the maximal weighted length, the assignment of priority among these queues is non-essential and can be arbitrary. We adopt the convention that when there are ties, the priority will be given to the queue with the largest index.

Denote the system state at time t by $Q(t) \doteq (Q_1(t), \ldots, Q_d(t))$, where Q_i is the size of queue i. The process Q is a continuous time Markov jump process defined on some probability space, say $(\Omega, \mathcal{F}, \mathbb{P})$. The quantities of interest are the buffer overflow probabilities

$$\begin{aligned} p_n \;\doteq\; & \mathbb{P}\{\text{the process } Q \text{ exits domain } nD \text{ before returning} \\ & \qquad \text{to the origin, starting from the origin}\}, \end{aligned} \tag{3.2.2}$$

where n is a large integer and D is the domain

$$D \doteq \{x = (x_1, \ldots, x_d) \in \mathbb{R}_+^d : c_i x_i \leq 1 \text{ for all } i\}. \tag{3.2.3}$$

We also define the vector

$$\bar{c} = (1/c_1, 1/c_2, \ldots, 1/c_d)\,, \tag{3.2.4}$$

which is a corner of D, and denote the boundary of D in $\mathbb{R}_+^d$ by

$$\partial \doteq \partial D = \{x \in \mathbb{R}_+^d : x \in D, c_i x_i = 1 \text{ for some } i\}.$$

See Figure 2 for illustration of the case of dimension $d = 2$.

Throughout the analysis, it is often convenient to consider the scaled process

$$X^n(t) \doteq Q(nt)/n.$$

Note that we can now rewrite the overflow probability as

$$\begin{aligned} p_n &= \mathbb{P}\{X^n \text{ exits domain } D \text{ before returning to the origin,} \\ &\qquad \text{starting at the origin}\}. \end{aligned}$$

3.2.1 System Dynamics

Of course the queue length process Q will be identical to the queue length process described in the previous chapter. However for the analysis of importance sampling changes of measure it is more useful to characterize the process by its stochastic transition kernel $\Theta[dt, v|x]$. To be more precise, define the total intensity by

$$R(x) \doteq \sum_{v \in \mathbb{V}} r(x, v), \tag{3.2.1}$$

and let $T_1, T_2, \ldots$ be the random times at which the process Q jumps with convention $T_0 = 0$. Then for any $x \in \mathbb{R}_+^d$ and $n \in \mathbb{N}$ such that $nx \in \mathbb{Z}_+^d$,

$$\begin{aligned} \Theta[dt, v|x] &\doteq \mathbb{P}\{T_{j+1} - T_j \in dt, Q(T_{j+1}) = nx + v | Q(T_j) = nx\} \\ &= r(x, v) e^{-R(x)t} dt. \end{aligned} \tag{3.2.2}$$

3.3 Large Deviations

One goal of this paper is to explicitly identify the exponential decay rate of p_n. Our approach is to reduce it to a calculus of variation problem using the sample path large deviations principle established in the previous chapter. In order to explicitly solve the calculus of variation problem, we will use a verification argument based on essentially the same subsolution that we will use to build asymptotically optimal importance sampling schemes.

To ease notation, we will also denote by $\mathbb{A}$ the collection of non-empty subsets of $\{1, 2, \ldots, d\}$. That is

$$\mathbb{A} = \{A : A \subset \{1, 2 \ldots, d\}, A \text{ is non-empty}\}. \qquad (3.3.1)$$

It was seen in chapter 2 that the scaled queue length process satisfies an LDP with local rate function $L(\cdot, \cdot)$. With this knowledge it is possible to find the explicit form of the exponential decay rate for the buffer overflow probabilities.

3.3.1 The exponential decay rate

In this section we first identify a collection of important roots associated with the Hamiltonians $H^{(i)}$, $i = 1, \ldots, d$. The motivation for such roots will be discussed in Remark 3.3.4. These roots will be used to construct subsolutions and identify the large deviation rate for p_n.

For each $A \in \mathbb{A}$, thanks to the stability condition (3.2.1), there exists a unique constant, say z^A, such that

$$0 < z^A < \min_{i \in A} \mu_i, \qquad \sum_{j \in A} \frac{\lambda_j}{\mu_j - z^A} = 1. \qquad (3.3.2)$$

Define the vector $\alpha^A \in \mathbb{R}^d$ by

$$\alpha_i^A \doteq \begin{cases} \log\left(1 - \dfrac{z^A}{\mu_i}\right) & \text{if } i \in A, \\ 0 & \text{otherwise.} \end{cases} \tag{3.3.3}$$

We have the following result, whose proof is elementary and thus omitted.

Lemma 3.3.1. *For each $A \in \mathbb{A}$, α^A is the unique vector that satisfies*

1. $\alpha_i^A = 0$ *for* $i \notin A$,
2. $\alpha_i^A < 0$ *for* $i \in A$,
3. $H^{(i)}(-\alpha^A) = 0$ *for all* $i \in A$.
4. $H^{(i)}(-\alpha^A) > 0$ *for all* $i \notin A$.

Recall the definition of $\bar{c}$ as in (3.2.4) and define

$$\gamma \doteq \min\left\{\langle -\alpha^A, \bar{c}\rangle : A \in \mathbb{A}\right\}. \tag{3.3.4}$$

The following result identifies γ as the value of the calculus of variation problem that is associated with the large deviations of $\{p_n\}$. The proof is technical and deferred to the appendix.

Theorem 3.3.2. *We have the representation*

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

where the infimum is taken over all absolutely continuous functions $\phi : \mathbb{R}_+ \to \mathbb{R}_+^d$ such that $\phi(0) = 0$ and $\tau \doteq \inf\{t \geq 0 : \phi(t) \in \partial\} < \infty$.

Remark 3.3.3. It will be shown later on that γ is the exponential decay rate of $\{p_n\}$ [Theorem 3.6.3], that is,

$$\gamma = -\lim_n \frac{1}{n} \log p_n.$$

Remark 3.3.4. Large deviations rate functions are closely related to control problems and thus are also related to the solutions of appropriate Hamiltonian-Jacobi-Bellman (HJB) equations [5]. Due to the homogeneity of the state dynamics, it is natural to conjecture that the large deviation optimal trajectory leaving the domain will be a straight line, and the solution to the HJB equation is piecewise affine. If the optimal trajectory leaves the domain through interface $\{x : \pi(x) = A\}$, then it corresponds to an affine piece whose gradient, say v, should satisfy the HJB equation $H^{(i)}(v) = 0$ for all $i \in A$. Furthermore, v should satisfy $\langle v, e_i \rangle = 0$ for all $i \notin A$ since the value function to the corresponding control problem will take the same value 0 on the boundary $\{x : \pi(x) = A\} \cap \partial$. Solving these equations about v yields $v = -\alpha^A$.

3.4 The Isaacs Equation

The connection between importance sampling and differential games/Isaacs equations was explored in [15]. Since we are now dealing with continuous time processes as opposed to the discrete time setup in [15], we will formally derive the associated Isaacs equation in some detail. While the derivation is formal, the analysis of importance sampling schemes that are based on subsolutions to the equation will be rigorous.

Let $\mathcal{R}$ be the collection of functions mapping $\mathbb{V}$ to $[0, \infty)$ that are not identically zero. Each element in $\mathcal{R}$ represents a set of jump intensities. We will restrict our attention to those alternative stochastic transition kernels that take the form

$$\bar{\Theta}^n[dt, v|x] = \bar{r}(x, v)e^{-\bar{R}(x)t}\, dt, \quad \bar{r}(x, \cdot) \in \mathcal{R},\ \bar{R}(x) = \sum_{v \in \mathbb{V}} \bar{r}(x, v). \tag{3.4.1}$$

Consider the stochastic control problem of minimizing the second moment of $\hat{p}_n$ (see equation (1.2.5) in the introduction) among such transition kernels, and define the value function

$$V_n(x) \doteq \inf_{\bar{\Theta}^n} E^{\mathbb{P}}[\hat{p}_n | Q(0) = nx]. \tag{3.4.2}$$

Clearly $V_n(x)$ satisfies the boundary condition $V_n(x) = 1$ for all $x \notin D$.

Note that for $\bar{\Theta}^n$ defined as in (3.4.1), one can rewrite the likelihood ratio term in the importance sampling estimator $\hat{p}_n$, recall (1.2.5) from the introduction, as

$$\exp\left\{ \int_0^{T_{N^n}} [\bar{R}(Q(s)/n) - R(Q(s)/n)]ds + \sum_{j=1}^{N^n} \log \frac{r(Q(T_{j-1})/n, v_j)}{\bar{r}(Q(T_{j-1})/n, v_j)} \right\}.$$

Therefore, by the dynamic programming principle, the value function V_n should satisfy the dynamic programming equation (DPE)

$$V_n(x) = \inf_{\bar{r} \in \mathcal{R}} \int_0^\infty \sum_{v \in \mathbb{V}} e^{[\bar{R} - R(x)]t} \cdot \frac{r(x,v)}{\bar{r}(v)} \cdot V_n\left(x + \frac{v}{n}\right) \cdot e^{-R(x)t} r(x,v) dt.$$

Here instead of the fixed state dependent rates we minimize over $\bar{r}(v)$ and let $\bar{R} = \sum_{v \in \mathbb{V}} \bar{r}(v)$. In order to write down a limit partial differential equation, we introduce the following result, which is indeed a special case of the relative entropy representation for exponential integrals [5]. The proof is deferred to the appendix. Note that the function $\ell : \mathbb{R} \to \mathbb{R}_+ \cup \{\infty\}$ is defined as

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

with convention $0\ell(0/0) = 0$ and $0\ell(x/0) = \infty$ if $x \neq 0$.

Lemma 3.4.1. *For any $\theta \in \mathcal{R}$ let*

$$s[\theta] \doteq \sum_{v \in \mathbb{V}} \theta(v).$$

Then for any $\theta \in \mathcal{R}$, any constant c and function $h : \mathbb{V} \to \mathbb{R}$

$$\begin{aligned} &-\log \int_0^\infty \sum_{v \in \mathbb{V}} e^{-[ct+h(v)]} e^{-s[\theta]t} \theta(v) dt \\ &= \inf_{\hat{\theta} \in \mathcal{R}} \frac{1}{s[\hat{\theta}]} \left[c + \sum_{v \in \mathbb{V}} h(v)\hat{\theta}(v) + \sum_{v \in \mathbb{V}} \theta(v) \ell \left(\frac{\hat{\theta}(v)}{\theta(v)} \right) \right]. \end{aligned}$$

We combine Lemma 3.4.1 with the DPE of V_n and a transform motivated by the large deviation scaling

$$W_n(x) \doteq -\frac{1}{n} \log V_n(x)$$

to obtain

$$\begin{aligned} nW_n(x) &= \sup_{\bar{r} \in \mathcal{R}} \inf_{\hat{r} \in \mathcal{R}} \frac{1}{s[\hat{r}]} \left[R(x) - \bar{R} + \sum_{v \in \mathbb{V}} \hat{r}(v) \log \frac{\bar{r}(v)}{r(x,v)} \right. \\ &\qquad \left. + \sum_{v \in \mathbb{V}} nW_n \left(x + \frac{v}{n} \right) \hat{r}(v) + \sum_{v \in \mathbb{V}} r(x,v) \ell \left(\frac{\hat{r}(v)}{r(x,v)} \right) \right]. \end{aligned}$$

To formally obtain a limit PDE, we assume that

$$W_n(x) \to W(x), \quad n\left[W_n(x+v/n) - W_n(x)\right] \to \langle DW(x), v \rangle. \tag{3.4.3}$$

Observing that

$$R(x) - \bar{R} + \sum_{v \in \mathbb{V}} \hat{r}(v) \log \frac{\bar{r}(v)}{r(x,v)} = \sum_{v \in \mathbb{V}} r(x,v) \ell \left(\frac{\hat{r}(v)}{r(x,v)} \right) - \sum_{v \in \mathbb{V}} \bar{r}(v) \ell \left(\frac{\hat{r}(v)}{\bar{r}(v)} \right),$$

the limit function W satisfies the equation

$$\begin{aligned} 0 \quad = \quad & \sup_{\bar{r}\in\mathcal{R}} \inf_{\hat{r}\in\mathcal{R}} \frac{1}{s[\hat{r}]} \left[\sum_{v\in\mathbb{V}} \langle DW(x), v\rangle \hat{r}(v) + 2\sum_{v\in\mathbb{V}} r(x,v)\ell\left(\frac{\hat{r}(v)}{r(x,v)}\right) \right. \\ & \left. - \sum_{v\in\mathbb{V}} \bar{r}(v)\ell\left(\frac{\hat{r}(v)}{\bar{r}(v)}\right)\right]. \end{aligned} \tag{3.4.4}$$

Define for every $x \in \mathbb{R}_+$ and $\alpha \in \mathbb{R}^d$

$$\begin{aligned} \mathbb{H}(x,\alpha) \quad = \quad & \sup_{\bar{r}\in\mathcal{R}} \inf_{\hat{r}\in\mathcal{R}} \left[\sum_{v\in\mathbb{V}} \hat{r}(v)\langle v, \alpha\rangle + 2\sum_{v\in\mathbb{V}} r(x,v)\ell\left(\frac{\hat{r}(v)}{r(x,v)}\right) \right. \\ & \left. - \sum_{v\in\mathbb{V}} \bar{r}(v)\ell\left(\frac{\hat{r}(v)}{\bar{r}(v)}\right)\right]. \end{aligned} \tag{3.4.5}$$

It is not difficult to argue by the existence of saddle points for $\mathbb{H}$ [Proposition 3.4.2] that the equation (3.4.4) for W can be reduced to

$$\mathbb{H}(x, DW(x)) = 0.$$

This is the Isaacs equation associated with a differential game of two players. It should be mentioned that the $\bar{r}$-player chooses the change of measure used in the importance sampling scheme and the $\hat{r}$-player is an artificial player introduced by the representation formula.

Proposition 3.4.2. *For any $x \in \mathbb{R}^d_+ \setminus \{0\}$ and $\alpha \in \mathbb{R}^d$,*

1. $\mathbb{H}(x,\alpha) = \mathbb{H}_j(\alpha) \doteq -2H^{(j)}(-\alpha/2)$*, where* $j = \max\{i : i \in \pi(x)\}$.

2. *The saddle point is given by*

$$\bar{r}^*(x,\alpha)[v] = \hat{r}^*(x,\alpha)[v] = r(x,v)e^{-\langle\alpha,v\rangle/2}.$$

Remark 3.4.3. The Isaacs equation $\mathbb{H}(x, DW(x)) = 0$ should not be interpreted literally at the points of discontinuities, that is, points x such that $|\pi(x)| > 1$. This is easy to understand from the above formal analysis since the second part of the approximation (3.4.3) will not hold at such x. Instead, a natural interpretation at such points is [4]

$$\lim_{\varepsilon \downarrow 0} \inf_{\{y \in \mathbb{R}_+^d : \|y-x\| \leq \varepsilon\}} \mathbb{H}(y, DW(x)) = 0,$$

which amounts to

$$\min_{j \in \pi(x)} \mathbb{H}_j(DW(x)) = 0. \tag{3.4.6}$$

The intuition for this interpretation is that the behavior of the controlled state process on a discontinuity interface is the asymptotic characterization for the collective behavior of the process around the neighborhood of x. This observation is also essential in the construction of classical subsolutions to the Isaacs equation [that is, replace $\mathbb{H}(x, DW) = 0$ by $\mathbb{H}(x, DW) \geq 0$], since, owing to the infimum operation in (3.4.6), the subsolution property at x will imply that it holds, at least approximately, for all nearby points of x in the prelimit. See Lemma 3.5.3 for the precise statement.

3.5 Classical Subsolutions

A continuously differentiable function $\bar{W} : \mathbb{R}_+^d \to \mathbb{R}$ is a *classical subsolution* to the Isaacs equation if

1. $\mathbb{H}(x, D\bar{W}(x)) \geq 0$ for all $x \in D$
2. $\bar{W}(x) \leq 0$ for all $x \notin D$.

Classical subsolutions are the means by which the explicit representation for γ in Theorem 3.3.2 is shown and asymptotically efficient importance sampling schemes are built. We will construct the subsolutions by mollifying piecewise affine subsolutions as in [15]. In order to achieve asymptotic optimality it is necessary that the value of the subsolutions at the origin is maximal, i.e., $\bar{W}(0) = 2\gamma$.

3.5.1 Piecewise affine subsolutions

Recall that $\mathbb{A}$ is the collection of all non-empty subsets of $\{1, 2, \ldots, d\}$, the definition of the roots $\{\alpha^A\}$ in (3.3.2) and (3.3.3), and the definition of γ in (3.3.4). For each $A \in \mathbb{A}$ define

$$\bar{\alpha}^A = \frac{\gamma}{\langle -\alpha^A, \bar{c} \rangle} \cdot \alpha^A. \tag{3.5.1}$$

Note that $\gamma \leq \langle -\alpha^A, \bar{c} \rangle$ for all $A \in \mathbb{A}$. Thanks to Lemma 3.3.1, the convexity of $H^{(i)}$, and that $H^{(i)}(0) = 0$, we have

Lemma 3.5.1. *For every $A \in \mathbb{A}$,*

1. $\bar{\alpha}_i^A = 0$ *for* $i \notin A$,
2. $\bar{\alpha}_i^A < 0$ *for* $i \in A$,
3. $H^{(i)}(-\bar{\alpha}^A) \leq 0$ *for all* $i \in A$,
4. $\langle -\bar{\alpha}^A, \bar{c} \rangle = \gamma$.

Also define a sequence of positive constants $\{C_1, C_2, \ldots, C_d\}$ by

$$C_1 = 0, \; C_{k+1} = a_k^* \cdot (1 + C_k), \tag{3.5.2}$$

where

$$a_k^* = 1 \vee \max \left\{ \frac{\bar{\alpha}_i^B}{\bar{\alpha}_i^A} : i \in A \subset B, |A| = k, |B| = k+1 \right\} \tag{3.5.3}$$

for $1 \leq k \leq d-1$. Fix an arbitrary $\delta > 0$. Define the affine function

$$W_A^\delta(x) \doteq \langle 2\bar{\alpha}^A, x \rangle + 2\gamma - C_{|A|}\delta. \tag{3.5.4}$$

for each $A \in \mathbb{A}$. Their pointwise minimum is denoted by

$$W^\delta(x) \doteq \min\{W_A^\delta(x) : A \in \mathbb{A}\}. \tag{3.5.5}$$

Note that W^δ is a continuous piecewise affine function.

Lemma 3.5.2. *The function W^δ satisfies $W^\delta(x) \leq 0$ for all $x \notin D$. Furthermore, given any $x \in \mathbb{R}^d_+ \setminus \{0\}$, if $W^\delta_A(x) - W^\delta(x) < \delta$ then $\pi(x) \subset A$.*

The proof of Lemma 3.5.2 is deferred to the appendix. This lemma implies that W^δ is a piecewise affine subsolution, i.e., $\mathbb{H}(x, DW^\delta(x)) \geq 0$ at all those x where $DW^\delta(x)$ is well defined. Indeed, for any such x, there is a unique $A^* \in \mathbb{A}$ that attains the minimum in (3.5.5) and whence $DW^\delta(x) = 2\bar{\alpha}^{A^*}$. Let $j = \max\{i : i \in \pi(x)\}$. It follows from Lemma 3.5.2 that $j \in \pi(x) \subset A^*$. It is now immediate from Lemma 3.5.2 and Proposition 3.4.2 that

$$\mathbb{H}(x, DW^\delta(x)) = \mathbb{H}_j(2\bar{\alpha}^{A^*}) = -2H^{(j)}(-\bar{\alpha}^{A^*}) \geq 0.$$

3.5.2 Mollification

The construction of asymptotically optimal importance sampling schemes requires smooth subsolutions with bounded second derivatives. To this end, we consider a mollified version of W^δ using *exponential weighting* [13, 15]. Let ε be an arbitrary small positive number, and define

$$W^{\varepsilon,\delta}(x) \doteq -\varepsilon \log \sum_{A \in \mathbb{A}} \exp\left\{-\frac{1}{\varepsilon} W^\delta_A(x)\right\}.$$

The function $W^{\varepsilon,\delta}$ is smooth with

$$DW^{\varepsilon,\delta}(x) = \sum_{A \in \mathbb{A}} 2\rho^{\varepsilon,\delta}_A(x)\bar{\alpha}^A, \quad \rho^{\varepsilon,\delta}_A(x) \doteq \frac{\exp\left\{-W^\delta_A(x)/\varepsilon\right\}}{\sum_{B \in \mathbb{A}} \exp\left\{-W^\delta_B(x)/\varepsilon\right\}}.$$

Furthermore, for every x,

$$-\varepsilon \log |\mathbb{A}| \leq W^{\varepsilon,\delta}(x) - W^\delta(x) \leq 0. \tag{3.5.6}$$

The following Lemma 3.5.3 shows that $W^{\varepsilon,\delta}$ is approximately a subsolution.

Lemma 3.5.3. *The function* $W^{\varepsilon,\delta}$ *satisfies* $W^{\varepsilon,\delta}(x) \leq 0$ *for all* $x \notin D$*. Furthermore, for any every* $x \in \mathbb{R}^d_+ \setminus \{0\}$ *and* $j \in \pi(x)$*,*

$$\mathbb{H}_j(DW^{\varepsilon,\delta}(x)) \geq \sum_{A\in\mathbb{A}} \rho^{\varepsilon,\delta}_A(x) \cdot \mathbb{H}_j(2\bar{\alpha}^A) \geq -Ke^{-\delta/\varepsilon},$$

where K *is a positive constant that only depends on the system parameters. In particular,*

$$\mathbb{H}(x, DW^{\varepsilon,\delta}(x)) \geq -Ke^{-\delta/\varepsilon}.$$

Proof. We first note that $W^{\varepsilon,\delta}(x) \leq 0$ for all $x \notin D$ is trivial from Lemma 3.5.2 and inequality (3.5.6). Now let $x \in \mathbb{R}^d_+ \setminus \{0\}$ and $j \in \pi(x)$. Thanks to Proposition 3.4.2 and the concavity of $\mathbb{H}_j$,

$$\mathbb{H}_j(DW^{\varepsilon,\delta}(x)) \geq \sum_{A\in\mathbb{A}} \rho^{\varepsilon,\delta}_A(x) \cdot \mathbb{H}_j(2\bar{\alpha}^A).$$

For all those $A \in \mathbb{A}$ such that $W^\delta_A(x) - W^\delta(x) < \delta$, Lemma 3.5.2 implies $\pi(x) \subset A$, in particular $j \in A$, and therefore $\mathbb{H}_j(2\bar{\alpha}^A) = -2H^{(j)}(-\bar{\alpha}^A) \geq 0$ by Lemma 3.5.1. On the other hand, for those $A \in \mathbb{A}$ such that $W^\delta_A(x) - W^\delta(x) \geq \delta$, it is not difficult to see that $\rho^{\varepsilon,\delta}_A(x) \leq \exp\{-\delta/\varepsilon\}$. ∎

Remark 3.5.4. In general, the parameters ε and δ can depend on n, and we will write them as ε_n and δ_n.

3.5.3 Importance sampling algorithm

For each $A \in \mathbb{A}$, the vector $2\bar{\alpha}^A$ through Proposition 3.4.2 defines a new set of jump intensities of form

$$\bar{r}^*(x, 2\bar{\alpha}^A)[v] = r(x, v)e^{-\langle\bar{\alpha}^A, v\rangle}.$$

The change of measure used in our state-dependent importance sampling scheme corresponds to a stochastic transition kernel of form

$$\bar{\Theta}^n[dt, v|x] = \bar{r}^n(x, v)e^{-t\bar{R}^n(x)}dt,$$

where for $x \neq 0$

$$\bar{r}^n(x, v) \doteq \sum_{A \in \mathbb{A}} \rho_A^{\varepsilon_n, \delta_n}(x)\bar{r}^*(x, 2\bar{\alpha}^A)[v], \qquad \bar{R}^n(x) \doteq \sum_{v \in \mathbb{V}} \bar{r}^n(x, v).$$

The values at $x = 0$ are unimportant, and for simplicity we use the original jump intensities, that is,

$$\bar{r}^n(0, v) \doteq r(0, v), \qquad \bar{R}^n(0) \doteq \sum_{v \in \mathbb{V}} r(0, v).$$

The corresponding importance sampling estimator $\hat{p}_n$ is just as defined in equation (1.2.5) in the introduction. Under this change of measure, the state process Q is again a continuous time Markov jump process with jump intensity $\bar{r}^n(x, v)$ from state nx to $nx + v$ whenever $nx \in \mathbb{Z}_+^d$. Note that the calculation of $\bar{\Theta}^n$ is simple since $\rho_A^{\varepsilon_n, \delta_n}$ and $\bar{r}^*$ can be easily obtained.

Remark 3.5.5. The estimator $\hat{p}_n$ described here is not exactly the one that will be used in the numerical experiments. Indeed, we will use a closely related, easier-to-implement, discrete time version of $\hat{p}_n$, say $\bar{p}_n$. The precise definition of $\bar{p}_n$ can be found in Section 3.7. It turns out that $\bar{p}_n$ is the expectation of $\hat{p}_n$ conditioned on a suitable σ-algebra, and therefore $\bar{p}_n$ is unbiased and has smaller second moment than $\hat{p}_n$. The motivation for focusing so far on $\hat{p}_n$ is that it is more suitable for the asymptotic analysis.

3.6 Asymptotic Optimality

In this section we show that the importance sampling estimator $\hat{p}_n$ is asymptotically optimal under suitable conditions. We will need the following result, which essentially says that the return time to the origin is exponentially bounded. Recall that $\{T_1, T_2, \ldots\}$ are the random jump times of the process Q and T_{N^0} [recall the definition of N^0 from the introduction (1.2.2)] is the first time the process returns to origin.

Lemma 3.6.1. *There exist positive constants c and k such that for every $q \in \mathbb{Z}_+^d$*

$$\log E^{\mathbb{P}}[\exp\{cT_{N^0}\}|Q(0) = q] \leq k(1 + \|q\|).$$

The proof is deferred to the appendix.

Proposition 3.6.2. *Suppose that $\delta_n \to 0$, $\varepsilon_n/\delta_n \to 0$, and $n\varepsilon_n \to \infty$. Then for any sequence $\{q_n\} \subset \mathbb{Z}_+^d$ such that $q_n/n \to 0$, we have the upper bound on the second moment of the importance sampling estimator*

$$\limsup_n \frac{1}{n} \log E^{\mathbb{P}}[\hat{p}_n|Q(0) = q_n] \leq -2\gamma.$$

Proof. Let $x_n = q_n/n$. Then $x_n \to 0$. Throughout the proof, to ease notation, we denote $W(x) \doteq W^{\varepsilon_n,\delta_n}(x)$, $\rho_A(x) \doteq \rho_A^{\varepsilon_n,\delta_n}(x)$, and $E_{x_n}^{\mathbb{P}}[\cdot] \doteq E^{\mathbb{P}}[\cdot|Q(0) = nx_n]$. Furthermore, note that all we need to show is that the upper bound holds for $x_n \neq 0$ since we can write

$$E_0^{\mathbb{P}}[\hat{p}_n] = \sum_{i=1}^{d} \frac{\lambda_i}{\lambda_1 + \cdots + \lambda_d} E^{\mathbb{P}}[\hat{p}_n|Q(0) = e_i].$$

Let $J(t) = \min\{j : T_j \geq t\}$ and define the following process

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

which is in fact a generalization of the Radon-Nikodym derivative i.e., $\hat{p}_n = \hat{p}_n(T_{N^n})$. The next step in the proof will be to show that the process

$$M^n(t) \doteq e^{-\beta_n t - snW(Q(t)/n)/2}\hat{p}_n^{s-1}(t) \tag{3.6.1}$$

is a supermartingale. The term β_n is a positive constant that will be defined later. Note that since the process $M^n(\cdot)$ is non-negative it suffices to show that the process is a local supermartingale. In order to show that the process is a local supermartingale it suffices to show that

$$\mathfrak{L} \doteq \lim_{t\to 0}\frac{E^{\mathbb{P}}\left[M^n(t)|Q(0) = nx\right] - M^n(0)}{t} \tag{3.6.2}$$

is non-positive for x and n such that $nx \in \mathbb{Z}_+^d$. A simple calculation shows

$$\begin{aligned}\mathfrak{L} &= M^n(0)\left[(s-1)\left(\bar{R}^n(x) - R(x)\right) - \beta_n - R(x)\right]\\ &\quad + \sum_{v\in\mathbb{V}} r(x,v)\left(\frac{r(x,v)}{\bar{r}^n(x,v)}\right)^{s-1} e^{-snW(x+\frac{v}{n})2}.\end{aligned}$$

It will in fact be easier to analyze the quantity $\bar{\mathfrak{L}} \doteq \mathfrak{L}/M^n(0)$, which can be written

$$\begin{aligned}\bar{\mathfrak{L}} &= (s-1)\left(\bar{R}^n(x) - R(x)\right) - \beta_n\\ &\quad + \sum_{v\in\mathbb{V}} r(x,v)\left[e^{-sn\left(W(x+\frac{v}{n}) - W(x)\right)/2}\left(\frac{r(x,v)}{\bar{r}^n(x,v)}\right)^{s-1} - 1\right]\end{aligned}$$

Note that by construction the function W has a bounded second derivative. In particular there exists a constant $C < \infty$ such that the entries of the Hessian matrix

$D^2W(x)$ are bounded by C/ε_n for every x. A Taylor expansion then gives

$$\left|\langle DW(x), v\rangle - n\left(W(x+\frac{v}{n}) - W(x)\right)\right| \le \frac{\bar{C}}{n\varepsilon_n}$$

where $\bar{C}$ is a positive constant that only depends on the system parameters. Using the fact that $|DW|$ and $r(x,v)/\bar{r}^n(x,v)$ are uniformly bounded the above inequality can be combined with the equation for $\bar{\mathfrak{L}}$ to get the inequality

$$\begin{aligned} \bar{\mathfrak{L}} \;\le\; & (s-1)\left(\bar{R}^n(x) - R(x)\right) - \beta_n + \bar{K}(e^{\bar{C}/n\varepsilon_n} - 1) \\ & + \sum_{v\in\mathbb{V}} r(x,v)\left[e^{-s\langle DW(x),v\rangle/2}\left(\frac{r(x,v)}{\bar{r}^n(x,v)}\right)^{s-1} - 1\right], \end{aligned} \tag{3.6.3}$$

where $\bar{K}$ is a positive constant that only depends on the system parameters.

Next note that by definition

$$\langle DW(x), v\rangle = 2\sum_{A\in\mathbb{A}} \rho_A(x)\langle\bar{\alpha}^A, v\rangle$$

and

$$\bar{r}^n(x,v) = \sum_{A\in\mathbb{A}} \rho_A(x)\bar{r}^*(x, 2\bar{\alpha}^A)[v] = r(x,v)\sum_{A\in\mathbb{A}} \rho_A(x)e^{-\langle\bar{\alpha}^A, v\rangle}.$$

It follows from the definition of $\bar{r}^n(x,v)$ in the above display that

$$e^{-s\langle DW(x),v\rangle/2}\left(\frac{r(x,v)}{\bar{r}^n(x,v)}\right)^{s-1} = \exp\left[-s\langle DW(x),v\rangle/2 - (s-1)\log\sum_{A\in\mathbb{A}} \rho_A(x)e^{\langle-\bar{\alpha}^A, v\rangle}\right].$$

The convexity of $-\log$ and the definition of $DW(x)$ give the following inequality and equality

$$\begin{aligned} e^{-s\langle DW(x),v\rangle/2}\left(\frac{r(x,v)}{\bar{r}^n(x,v)}\right)^{s-1} & \le \exp\left[-s\langle DW(x),v\rangle/2 - (s-1)\sum_{A\in\mathbb{A}} \rho_A(x)\langle-\bar{\alpha}^A, v\rangle\right] \\ & = e^{-\langle DW(x),v\rangle/2} \end{aligned}$$

Plugging the previous inequality into the inequality (3.6.3), and using the result of Proposition 3.4.2 and Lemma 3.5.3 we arrive at the following

$$
\begin{aligned}
\bar{\mathcal{L}} &\leq (s-1)\left(\bar{R}^n(x) - R(x)\right) - \beta_n + K(e^{\bar{C}/n\varepsilon_n} - 1) + \sum_{v\in\mathbb{V}} r(x,v)\left[e^{-\langle DW(x),v\rangle/2} - 1\right] \\
&\leq (s-1)\left(\bar{R}^n(x) - R(x)\right) - \beta_n + \bar{K}(e^{\bar{C}/n\varepsilon_n} - 1) - \mathbb{H}(x, DW(x))/2 \\
&\leq (s-1)\left(\bar{R}^n(x) - R(x)\right) - \beta_n + \bar{K}(e^{\bar{C}/n\varepsilon_n} - 1) + \frac{K}{2}e^{-\delta_n/\epsilon_n}. \qquad (3.6.4)
\end{aligned}
$$

Lemma 3.5.3 can be used again to derive the following result

$$
\begin{aligned}
\bar{R}^n(x) - R(x) &= \sum_{v\in\mathbb{V}} r(x,v)\left(\sum_{A\in\mathbb{A}} \rho_A(x) e^{\langle -\bar{\alpha}^A, v\rangle} - 1\right) \\
&= \sum_{A\in\mathbb{A}} \rho_A(x)\left(\sum_{v\in\mathbb{V}} r(x,v)\left(e^{\langle -\bar{\alpha}^A, v\rangle} - 1\right)\right) \\
&= -\sum_{A\in\mathbb{A}} \rho_A(x)\mathbb{H}_j(2\bar{\alpha}^A) \leq K e^{-\delta_n/\epsilon_n}
\end{aligned}
$$

Combining the previous inequality with the result in (3.6.4) it follows that setting

$$\beta_n = 2Ke^{-\delta_n/\varepsilon_n} + \bar{K}\left(e^{\bar{C}/2n\varepsilon_n} - 1\right)$$

gives $\bar{\mathcal{L}} \leq 0$. Thus the process $M^n(\cdot)$ is a supermartingale.

Consequently, the optional sampling theorem implies that, with $N \doteq N^n \wedge N^0$,

$$E^{\mathbb{P}}_{x_n}[M^n(T_N)] \leq M^n(0) = e^{-snW(x_n)/2}.$$

Observing that $Q(T_N)/n \notin D$ on the set $\{N^n < N^0\}$, while $\hat{p}_n = 0$ on the set $\{N^n \geq N^0\}$, and that the boundary condition $W(x) \leq 0$ holds for all $x \notin D$, it follows that w.p.1

$$M^n(T_N) \geq e^{-\beta_n T_N} \hat{p}_n^{s-1}.$$

Therefore,

$$E^{\mathbb{P}}_{x_n}\left[e^{-\beta_n T_N} \hat{p}_n^{s-1}\right] \leq e^{-snW(x_n)/2}.$$

Using Holder's inequality

$$\begin{aligned} E^{\mathbb{P}}_{x_n}[\hat{p}_n] &\leq \left(E^{\mathbb{P}}_{x_n}\left[e^{-\beta_n T_N}\hat{p}_n^{s-1}\right]\right)^{\frac{1}{s-1}} \left(E^{\mathbb{P}}_{x_n}\left[e^{\frac{\beta_n}{s-2}T_N} 1_{\{N^n<N^0\}}\right]\right)^{\frac{s-2}{s-1}} \\ &\leq e^{-\frac{s}{2(s-1)}nW(x_n)} \left(E^{\mathbb{P}}_{x_n}\left[e^{\frac{\beta_n}{s-2}T_{N^0}}\right]\right)^{\frac{s-2}{s-1}}. \end{aligned}$$

Note that the above inequality is true for all $s \in (2,3]$. In particular, let c be the constant given by Lemma 3.6.1, and let $s = s_n$ where

$$\frac{\beta_n}{s_n - 2} = c \quad \text{or} \quad s_n = 2 + \frac{\beta_n}{c}.$$

Note that by assumption $\beta_n \to 0$, thus $s_n \in (2,3]$ for n large enough. Hence by Lemma 3.6.1

$$E^{\mathbb{P}}_{x_n}[\hat{p}_n] \leq e^{-\frac{s_n}{2(s_n-1)}nW(x_n)} \left(E^{\mathbb{P}}_{x_n}\left[e^{cT_{N^0}}\right]\right)^{\frac{s_n-2}{s_n-1}} \leq e^{-\frac{s_n}{2(s_n-1)}nW(x_n)} e^{\frac{k(s_n-2)}{s_n-1}(1+\|x_n\|)}.$$

Therefore,

$$\frac{1}{n}\log E^{\mathbb{P}}_{x_n}[\hat{p}_n] \leq -\frac{s_n}{2(s_n-1)}W(x_n) + \frac{1}{n}\frac{k(s_n-2)}{s_n-1}(1+\|x_n\|)$$

Note that $s_n \to 2$ and $\|x_n\| \to 0$ as $n \to \infty$, and that

$$W^n(0) \geq 2\gamma - \log|\mathbb{A}| \cdot \varepsilon_n - C_d\delta_n.$$

Now letting $n \to 0$ we have

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

This completes the proof. ∎

Theorem 3.6.3. *The exponential decay rate of* $\{p_n\}$ *is* γ*, that is,*

$$\lim_{n} \frac{1}{n} \log p_n = -\gamma.$$

Proof. It follows [(1.2.5) from the introduction], the unbiasedness of $\hat{p}_n$, and Jensen's inequality that $E^{\mathbb{P}}_0[\hat{p}_n] \geq p_n^2$. Therefore, by Proposition 3.6.2 the upper bound

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

holds. It remains to show the lower bound. Thanks to Theorem 3.3.2, it suffices to show that

$$\liminf_{n} \frac{1}{n} \log p_n \geq -\int_0^{\tau} L(\phi(t), \dot{\phi}(t))\, dt \tag{3.6.5}$$

for any absolutely continuous $\phi : \mathbb{R}_+ \to \mathbb{R}^d_+$ such that $\phi(0) = 0$ and $\tau \doteq \inf\{t \geq 0 : \phi(t) \in \partial\} < \infty$. Note that due to the non-negativity of L it suffices to consider ϕ such that $\phi(t) \neq 0$ for all $t \in (0, \tau]$. For any $T \geq 0$ denote by $\mathcal{D}[0, T]$ the collection of cadlag functions on $[0, T]$ taking values in $\mathbb{R}^d_+$, equipped with the Skorohod topology. Let $\mathbf{1} \doteq (1, 1, \ldots, 1)$ and

$$\varphi(t) \doteq \begin{cases} \phi(t) & \text{if } t \leq \tau, \\ \phi(\tau) + (t - \tau) \cdot \mathbf{1} & \text{if } t > \tau. \end{cases}$$

For any $\varepsilon, \delta > 0$, define

$$B^{\varepsilon}(\varphi;\delta) \doteq \{\psi \in \mathcal{D}[0,\tau] : \|\psi(t) - \varphi(t+\varepsilon)\| < \delta \text{ for all } 0 \le t \le \tau\}.$$

Observing that $\dot{\varphi}(t) = \mathbf{1}$ for $t > \tau$ and that $L^{A}(\mathbf{1})$ is finite for any $A \subset \{1, \ldots, d\}$, it follows that

$$\lim_{\varepsilon \to 0} \int_{\varepsilon}^{\tau+\varepsilon} L(\varphi(t), \dot{\varphi}(t))\, dt = \int_{0}^{\tau} L(\phi(t), \dot{\phi}(t))\, dt. \tag{3.6.6}$$

For every ε, there exists a $\delta = \delta(\varepsilon)$ such that for every $\psi \in B^{\varepsilon}(\varphi;\delta)$, $\psi(t) \neq 0$ for all $t \in [0,\tau]$ and $\psi(\tau) \notin D$. Now for each n define

$$x_n \doteq \frac{1}{n} \lfloor n\phi(\varepsilon) \rfloor,$$

where the integer part is applied component-wise. Then

$$p_n \geq \mathbb{P}_0(X^n \text{ hits } x_n \text{ before } 0)\mathbb{P}_{x_n}(X^n \text{ exits } D \text{ before hitting } 0).$$

However, note that $B^{\varepsilon}(\varphi;\delta)$ is open under the Skorohod topology [25, Appendix A] since φ is continuous, and that $\psi(t) \doteq \varphi(t+\varepsilon)$ for $t \in [0,\tau]$ is an element of $B^{\varepsilon}(\varphi;\delta)$. Moreover $x_n \to \phi(\varepsilon)$, and $\{X^n\}$ satisfies the sample path large deviation principle with local rate function L [12]. It follows that

$$\begin{aligned}
&\liminf_{n} \frac{1}{n} \log \mathbb{P}_{x_n}(X^n \text{ exits } D \text{ before hitting } 0) \qquad (3.6.7)\\
&\quad \geq \liminf_{n} \frac{1}{n} \log \mathbb{P}_{x_n}(X^n \in B^{\varepsilon}(\varphi;\delta)) \\
&\quad \geq -\inf\left\{ \int_0^{\tau} L(\psi(t), \dot{\psi}(t))\, dt : \psi \in B^{\varepsilon}(\varphi;\delta) \right\} \\
&\quad \geq -\int_0^{\tau} L(\varphi(t+\varepsilon), \dot{\varphi}(t+\varepsilon))\, dt.
\end{aligned}$$

Furthermore, it is easy to see that

$$\mathbb{P}_0(\text{hitting } x_n \text{ before } 0) \geq \prod_{i=1}^{d} \left(\frac{\lambda_i}{\sum_{j=1}^{d} (\lambda_i + \mu_i)} \right)^{(nx_n)_i}.$$

Therefore

$$\begin{aligned} & \liminf_n \frac{1}{n} \log \mathbb{P}_0(\text{hitting } x_n \text{ before } 0) \qquad (3.6.8) \\ \geq & \sum_{i=1}^{d} \lim \frac{1}{n} (nx_n)_i \log \frac{\lambda_i}{\sum_{j=1}^{d} (\lambda_i + \mu_i)} \\ = & \sum_{i=1}^{d} (\phi(\varepsilon))_i \log \frac{\lambda_i}{\sum_{j=1}^{d} (\lambda_i + \mu_i)}. \end{aligned}$$

Combining (3.6.6), (3.6.7), and (3.6.8), and letting $\varepsilon \to 0$, we arrive at the desired inequality (3.6.5). ■

The following result is immediate from Proposition 3.6.2 and Theorem 3.6.3.

Corollary 3.6.4. *Suppose that $\delta_n \to 0$, $\varepsilon_n/\delta_n \to 0$, and $n\varepsilon_n \to \infty$. Then the importance sampling estimator $\hat{p}_n$ is asymptotically optimal. That is,*

$$\lim_n \frac{1}{n} \log \mathbb{E}_0^{\mathbb{P}}[\hat{p}_n] = -2\gamma.$$

3.7 Numerical Experiments

For ease of computation the numerical results are obtained using the embedded discrete time Markov chain $Z = \{Z(j) = Q(T_j), j \geq 0\}$. This means that given that the current state $Z(j)$ is nx, the next jump Δ_{j+1} is chosen according to

$$\mathbb{Q}(\Delta_{j+1} = v) = \bar{r}^n(x, v)/\bar{R}^n(x).$$

The actual importance sampling estimator is then

$$\bar{p}_n \doteq 1_{\{Z(N)\notin D\}} \prod_{j=0}^{N-1} \frac{r(Z(j)/n, \Delta_{j+1})/R(Z(j)/n)}{\bar{r}^n(Z(j)/n, \Delta_{j+1})/\bar{R}^n(Z(j)/n)}, \tag{3.7.1}$$

where

$$N = \inf\{k \geq 1 : Q(T_k) \in \partial \text{ or } Q(T_k) = 0\}.$$

By referring to the definition (1.2.5) found in the introduction, a simple calculation shows that

$$E^{\mathbb{Q}}[\hat{p}_n | Q(T_1), \ldots, Q(T_N)] = \bar{p}_n.$$

Therefore the asymptotic optimality of $\bar{p}_n$ follows from the asymptotic optimality of $\hat{p}_n$.

In the tables below we present results from three different systems in 2, 4, and 6 dimensions. For each system we consider $n = 20, 50$, and 80. All results are achieved using $20,000$ samples.

Table 3.1: Probability of buffer overflow in two dimensional case.

	$n = 20$	$n = 50$	$n = 80$
Theoretical value	1.90×10^{-5}	4.36×10^{-13}	8.31×10^{-21}
Estimate	1.94×10^{-5}	4.36×10^{-13}	8.18×10^{-21}
Std. Err.	0.04×10^{-5}	0.12×10^{-13}	0.32×10^{-21}
95% C.I.	$[1.86, 2.02] \times 10^{-5}$	$[4.12, 4.59] \times 10^{-13}$	$[7.55, 8.81] \times 10^{-32}$

Parameter values: $(\lambda_1, \lambda_2) = (1, 2), (\mu_1, \mu_2) = (3, 4),$ and $(c_1, c_2) = (1/2, 1)$

Table 3.2: Probability of buffer overflow in 4 dimensional case.

	$n = 20$	$n = 50$	$n = 80$
Theoretical value	5.62×10^{-9}	1.54×10^{-14}	7.01×10^{-23}
Estimate	5.52×10^{-9}	1.51×10^{-14}	6.91×10^{-23}
Std. Err.	0.27×10^{-9}	0.09×10^{-14}	0.33×10^{-23}
95% C.I.	$[4.99, 6.04] \times 10^{-9}$	$[1.33, 1.69] \times 10^{-14}$	$[6.26, 7.56] \times 10^{-23}$

Parameter values:
$(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (1, 2, 2, 4), (\mu_1, \mu_2, \mu_3, \mu_4) = (5, 12, 10, 15),$ and $(c_1, c_2, c_3, c_4) = (1/2, 1, 1, 1)$

Table 3.3: Probability of buffer overflow in 6 dimensional case.

	$n = 20$	$n = 50$	$n = 80$
Theoretical value	2.1×10^{-8}	4.2×10^{-13}	3.7×10^{-20}
Estimate	2.21×10^{-8}	4.21×10^{-13}	3.75×10^{-20}
Std. Err.	0.08×10^{-8}	0.15×10^{-13}	0.14×10^{-20}
95% C.I.	$[2.05, 2.37] \times 10^{-8}$	$[3.91, 4.50] \times 10^{-13}$	$[3.48, 4.02] \times 10^{-20}$

Parameter values: $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6) = (1, 2, 2, 3, 1, 8), (\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6) = (10, 15, 16, 16, 15, 24)$, and $(c_1, c_2, c_3, c_4, c_5, c_6) = (1/2, 1, 1, 1, 1, 1/3)$

For all cases the values of the constants $C_{|A|}$ are determined using the formulas in (3.5.2) and (3.5.3). The constants δ_n and ε_n are determined by the formulas $\varepsilon_n = \frac{1}{5 \log n}$ and $\delta_n = \varepsilon_n \log \varepsilon_n$. The results are robust with respect to the choice of the parameters ε_n and δ_n. For example $\varepsilon_n = c/\sqrt{n}$ yields similar quality results for varying values of c.

For the two and four dimensional systems the exact values are obtained by solving the linear system that arises from a first step analysis. This approach is computationally infeasible for the six dimensional system, thus for that system the exact value is obtained by running the importance sampling algorithm for 2 million iterations.

3.8 Proofs of Several Results

3.8.1 Proof of Lemma 3.4.1

We will consider two cases separately.

Case 1. $c + s[\theta] \leq 0$. It is not difficult to see that the integral on the left-hand-side (LHS) is ∞. Therefore, it suffices to show that the infimum on the right-hand-side (RHS) is $-\infty$. Fix an arbitrary $\varepsilon > 0$ and consider $\hat{\theta}_\varepsilon \in \mathcal{R}$ such that

$$\hat{\theta}_\varepsilon(v) \doteq \varepsilon e^{-h(v)} \theta(v).$$

Then simple algebra yields

$$\text{RHS} \;=\; \frac{c+s[\theta]}{\varepsilon\sum_{v\in\mathbb{V}} e^{-h(v)}\theta(v)} + \log\varepsilon - 1.$$

Letting $\varepsilon \to 0$, since $c+s[\theta] \leq 0$, we conclude that the infimum on the RHS does equal $-\infty$.

Case 2. $c+s[\theta] > 0$. Abusing notation, for each $\theta \in \mathcal{R}$, we define the probability measure induced by θ on $\mathbb{R}_+ \times \mathbb{V}$ by

$$\mathbb{P}_\theta(dt, v) = e^{-s[\theta]t}\theta(v)dt.$$

Then we can write

$$\text{LHS} = -\log \int_{\mathbb{R}_+} \sum_{v\in\mathbb{V}} e^{-[ct+h(v)]}\mathbb{P}_\theta(dt, v).$$

Further abusing the notation, let $\mathcal{P}$ denote the collection of all probability measures on $\mathbb{R}_+ \times \mathbb{V}$. Then by a slight extension of [5, Proposition 1.4.2], namely [15, Lemma A.1], we have

$$\text{LHS} = \inf_{\mathbb{Q}\in\mathcal{P}} \left[\int_{\mathbb{R}_+} \sum_{v\in\mathbb{V}} [ct+h(v)]\mathbb{Q}(dt, v) + R(\mathbb{Q}\|\mathbb{P}_\theta) \right],$$

with the minimizing $\mathbb{Q}^*$ given by

$$\frac{d\mathbb{Q}^*}{d\mathbb{P}_\theta} = ke^{-ct-h(v)},$$

where k is the normalizing constant. It is not difficult to see that the minimizing $\mathbb{Q}^*$

is an element of $\{\mathbb{P}_{\hat{\theta}} : \hat{\theta} \in \mathcal{R}\}$. Therefore,

$$\text{LHS} = \inf_{\hat{\theta}\in\mathcal{R}} \left[\int_{\mathbb{R}_+} \sum_{v\in\mathbb{V}} [ct + h(v)] \mathbb{P}_{\hat{\theta}}(dt, v) + R(\mathbb{P}_{\hat{\theta}} \| \mathbb{P}_{\theta}) \right].$$

But it is straightforward to verify that

$$\int_{\mathbb{R}_+} \sum_{v\in\mathbb{V}} [ct + h(v)] \mathbb{P}_{\hat{\theta}}(dt, v) = \frac{1}{s[\hat{\theta}]} \left[c + \sum_{v\in\mathbb{V}} h(v)\hat{\theta}(v) \right].$$

and

$$R(\mathbb{P}_{\hat{\theta}} \| \mathbb{P}_{\theta}) = \frac{1}{s[\hat{\theta}]} \sum_{v\in\mathbb{V}} \theta(v) \ell\left(\frac{\hat{\theta}(v)}{\theta(v)}\right).$$

∎

3.8.2 Proof of Proposition 3.4.2

We will prove a stronger version of Proposition 3.4.2, which will be useful in the asymptotic analysis of the importance sampling estimators. Fix an arbitrary $s \geq 1$. We define, for any $x \in \mathbb{R}^d_+$ and $\bar{r}, \hat{r} \in \mathcal{R}$,

$$\begin{aligned} L_s(x, \alpha; \bar{r}, \hat{r}) \quad &\doteq \quad \frac{s}{2} \sum_{v\in\mathbb{V}} \hat{r}(v)\langle \alpha, v\rangle + s \sum_{v\in\mathbb{V}} r(x,v)\ell\left(\frac{\hat{r}(v)}{r(x,v)}\right) \\ &\quad - (s-1) \sum_{v\in\mathbb{V}} \bar{r}(v) \ell\left(\frac{\hat{r}(v)}{\bar{r}(v)}\right) \end{aligned} \tag{3.8.1}$$

and

$$\mathbb{H}_s(x, \alpha) = \sup_{\bar{r}\in\mathcal{R}} \inf_{\hat{r}\in\mathcal{R}} L_s(x, \alpha; \bar{r}, \hat{r}).$$

Note that $\mathbb{H}$ is the special case of $\mathbb{H}_s$ with $s = 2$. The following result subsumes Proposition 3.4.2.

Lemma 3.8.1. *Assume that $s \geq 1$. For any $x \in \mathbb{R}^d_+ \setminus \{0\}$ and $\alpha \in \mathbb{R}^d$,*

1. $\mathbb{H}_s(x, \alpha) = -sH^{(j)}(-\alpha/2) = s\mathbb{H}_j(\alpha)/2$ *where* $j = \max\{i : i \in \pi(x)\}$. *In*

particular, for every x, $\mathbb{H}_s(x,\cdot)$ is concave.

2. *The saddle point of L_s is independent of s and takes the form*

$$\bar{r}^*(x,\alpha)[v] = \hat{r}^*(x,\alpha)[v] = r(x,v)e^{-\langle \alpha, v\rangle/2}.$$

Proof. We first argue that $(\bar{r}^*(x,\alpha), \hat{r}^*(x,\alpha))$ is a saddle point, that is,

$$L_s(x,\alpha;\bar{r},\hat{r}^*(x,\alpha)) \le L_s(x,\alpha;\bar{r}^*(x,\alpha),\hat{r}^*(x,\alpha)) \le L_s(x,\alpha;\bar{r}^*(x,\alpha),\hat{r})$$

for all $\bar{r},\hat{r} \in \mathcal{R}$. The first inequality is trivial due to that $s \ge 1$, the non-negativity of $\bar{r}$ and function ℓ, and that $\ell(z) = 0$ if and only if $z = 1$. As for the second inequality, straightforward calculation yields that

$$\begin{aligned} L_s(x,\alpha;\bar{r}^*(x,\alpha),\hat{r}) &= \frac{1}{2}\sum_{v\in\mathbb{V}} \hat{r}(v)\langle \alpha, v\rangle + \sum_{v\in\mathbb{V}} r(x,v)\ell\left(\frac{\hat{r}(v)}{r(x,v)}\right) \\ &\quad - (s-1)H^{(j)}(-\alpha/2). \end{aligned}$$

By elementary calculus, the right-hand-side is minimized at $\hat{r} = \hat{r}^*(x,v)$. Therefore, the second inequality holds, and $(\bar{r}^*(x,\alpha), \hat{r}^*(x,\alpha))$ is a saddle point. In particular,

$$\mathbb{H}_s(x,\alpha) = L_s(x,\alpha;\bar{r}^*(x,\alpha),\hat{r}^*(x,\alpha)) = -sH^{(j)}(-\alpha/2) = s\mathbb{H}_j(\alpha)/2.$$

This completes the proof. ∎

3.8.3 Proof of Lemma 3.5.2.

Fix an arbitrary $x \in \mathbb{R}^d_+ \setminus \{0\}$. Let $w^* \doteq \max\{c_i x_i : i = 1,\ldots,d\}$ and

$$x^* \doteq w^*\bar{c} = (w^*/c_1, w^*/c_2, \ldots, w^*/c_d)\,.$$

Then $x_i - x_i^* \leq 0$ for all i with equality if and only if $i \in \pi(x)$. Moreover, for any nonempty subset $F \subset \{1, \ldots, d\}$, the definition (3.5.1) and Lemma 3.5.1 imply that

$$\begin{aligned}
\langle 2\bar{\alpha}^F, x\rangle &= \langle 2\bar{\alpha}^F, x^*\rangle + \langle 2\bar{\alpha}^F, x - x^*\rangle \\
&= 2w^*\langle 2\bar{\alpha}^F, \bar{c}\rangle + \sum_{i \notin \pi(x)} 2\bar{\alpha}_i^F(x_i - x_i^*) \\
&= -2w^*\gamma + \sum_{i \in F\setminus\pi(x)} 2\bar{\alpha}_i^F(x_i - x_i^*).
\end{aligned}$$

It follows that for any $A \subset \{1, \ldots, d\}$

$$\begin{aligned}
W_A^\delta(x) - W_F^\delta(x) &= \sum_{i \in A\setminus\pi(x)} 2\bar{\alpha}_i^A(x_i - x_i^*) - \sum_{i \in F\setminus\pi(x)} 2\bar{\alpha}_i^F(x_i - x_i^*) \\
&\quad + (C_{|F|} - C_{|A|})\delta.
\end{aligned} \tag{3.8.2}$$

Since $W_F^\delta(x) \geq W^\delta(x)$ for any F, we have by assumption

$$W_A^\delta(x) - W_F^\delta(x) \leq W_A^\delta(x) - W^\delta(x) < \delta.$$

In particular, the above inequality holds for $F = \pi(x)$ and $F = A \cup \pi(x)$ and (3.8.2) then becomes

$$\delta > \sum_{i \in A\setminus\pi(x)} 2\bar{\alpha}_i^A(x_i - x_i^*) + (C_{|\pi(x)|} - C_{|A|})\delta, \tag{3.8.3}$$

$$\delta > \sum_{i \in A\setminus\pi(x)} 2(\bar{\alpha}_i^A - \bar{\alpha}_i^{A\cup\pi(x)})(x_i - x_i^*) + (C_{|A\cup\pi(x)|} - C_{|A|})\delta. \tag{3.8.4}$$

We now argue by contradiction that $|A \cup \pi(x)| = |A|$ or $\pi(x) \subset A$. To this end, assume that $|A| = m$ and $|A \cup \pi(x)| = m + j$ with $j > 0$. It follows easily from the definition of $\{a_k^*\}$ that for every $i \in A$

$$\frac{\bar{\alpha}_i^{A\cup\pi(x)}}{\bar{\alpha}_i^A} \leq a_m^* a_{m+1}^* \cdots a_{m+j-1}^*.$$

Note that $a_k^* \geq 1$ for all k, and $\bar{\alpha}_i^A < 0$ and $x_i - x_i^* < 0$ for all $i \in A \setminus \pi(x)$. Therefore, it follows from (3.8.4) and (3.8.3), that

$$\begin{aligned} (C_{|A\cup\pi(x)|} - C_{|A|} - 1)\delta &< \sum_{i\in A\setminus\pi(x)} 2(\bar{\alpha}_i^{A\cup\pi(x)} - \bar{\alpha}_i^A)(x_i - x_i^*) \\ &\leq (a_m^* a_{m+1}^* \cdots a_{m+j-1}^* - 1) \sum_{i\in A\setminus\pi(x)} 2\bar{\alpha}_i^A (x_i - x_i^*) \\ &< (a_m^* a_{m+1}^* \cdots a_{m+j-1}^* - 1)(C_{|A|} - C_{|\pi(x)|} + 1)\delta \\ &\leq (a_m^* a_{m+1}^* \cdots a_{m+j-1}^* - 1)(C_{|A|} + 1)\delta, \end{aligned}$$

or

$$C_{m+j} - C_m - 1 < (a_m^* a_{m+1}^* \cdots a_{m+j-1}^* - 1)(C_m + 1),$$

or

$$C_{m+j} < a_m^* a_{m+1}^* \cdots a_{m+j-1}^* (C_m + 1).$$

But this is impossible since

$$\begin{aligned} C_{m+j} &= a_{m+j-1}^*(1 + C_{m+j-1}) \\ &> a_{m+j-1}^* C_{m+j-1} \\ &= a_{m+j-1}^* a_{m+j-2}^* (1 + C_{m-j+2}) \\ &> \cdots \\ &= a_m^* a_{m+1}^* \cdots a_{m+j-1}^* (C_m + 1), \end{aligned}$$

a contradiction. Thus $j = 0$ and $|A \cup \pi(x)| = |A|$, and whence $\pi(x) \subset A$.

For every $x \notin D$, observe that $x_i \geq 1/c_i$ for all $i \in \pi(x)$. Therefore, by Lemma 3.5.1 and the definition of W^δ,

$$W^\delta(x) \leq W^\delta_{\pi(x)}(x) \leq \langle 2\bar{\alpha}^{\pi(x)}, x\rangle + 2\gamma \leq \langle 2\bar{\alpha}^{\pi(x)}, \bar{c}\rangle + 2\gamma = 0.$$

We complete the proof. ■

3.8.4 Proof of Lemma 3.6.1

Assume for now that $Q(0) = q \neq 0$. To ease notation, unless specified, we simply denote

$$E^{\mathbb{P}}[\cdot] \doteq E^{\mathbb{P}}[\cdot|Q(0) = q].$$

We first show that T_{N^0} is finite with probability one. Define a vector

$$d = (1/\mu_1, 1/\mu_2, \ldots, 1/\mu_d)$$

and let

$$\varepsilon \doteq 1 - \sum_{i=1}^{d} \frac{\lambda_i}{\mu_i} > 0.$$

Consider the process $Y = \{Y(t)\}$ where

$$Y(t) \doteq \langle Q(t), d\rangle + \varepsilon t.$$

It is easy to check that the compensation process for $\{\langle Q(t), d\rangle\}$ is

$$\int_0^t \left[\sum_{j=1}^{d} \lambda_j \langle d, e_j\rangle - \sum_{j=1}^{d} \mu_j \langle d, -e_j\rangle 1_{\{j=\max \pi(Q(t))\}} \right] ds = -\varepsilon t.$$

Therefore, Y is a local martingale. Since Y is a non-negative, it is a supermartingale. In particular, by the optional sampling theorem

$$\varepsilon E^{\mathbb{P}}[T_{N^0}] \leq E^{\mathbb{P}}[Y(T_{N^0})] \leq Y(0) = \langle q, d\rangle.$$

This implies that T_{N^0} has finite expectation. In particular, it is finite with probability one.

The proof for the exponential bound is based on a similar argument and the existence of a *strict* smooth subsolution. Fix an arbitrary $\nu \in (0,1)$ and for ease of notation let $\alpha^* = \bar{\alpha}^{\{1,\ldots,d\}}$. We claim that there exists a $c > 0$ such that the process $M = \{M(t) : t \geq 0\}$ with

$$M(t) = \exp\{ct - \langle Q(t), \nu\alpha^* \rangle\} \tag{3.8.5}$$

is a supermartingale. Indeed, by the generalized Itô formula the process

$$M(t) - \int_0^t M(s) h(Q(s))\, ds$$

is a local martingale with

$$h(z) \doteq c + \sum_{v \in \mathbb{V}} r(z,v) \left(\exp\{-\langle \nu\alpha^*, v \rangle\} - 1\right) = c + H^{(i)}(-\nu\alpha^*)$$

with $i = \max\{\pi(z)\}$. It follows from Lemma 3.5.1 and the strict convexity of $H^{(i)}$ that

$$H^{(i)}(-\nu\alpha^*) < 0$$

for all i. Therefore there exists a $c > 0$ such that $h(z) \leq 0$, which in turn implies that M is a local supermartingale. But M is nonnegative, whence a true supermartingale. By the optional sampling theorem,

$$\exp\{-\langle q, \nu\alpha^* \rangle\} \geq E^{\mathbb{P}}[M(T_{N^0})] = E^{\mathbb{P}}[\exp\{cT_{N^0}\}].$$

In particular, this implies that there exists some $k_1 > 0$ such that for any $q \neq 0$,

$$\log E^{\mathbb{P}}[\exp\{cT_{N^0}\} | Q(0) = q] \leq k_1 \|q\|.$$

It remains to show for the case where $Q(0) = 0$. By conditioning on the first jump of Q, it follows that

$$E^{\mathbb{P}}[\exp\{cT_{N^0}\}|Q(0) = 0] \quad = \quad \sum_{i=1}^{d} \frac{\lambda_i}{\Lambda(\Lambda + c)} E^{\mathbb{P}}[\exp\{cT_{N^0}\}|Q(0) = e_i] < \infty,$$

where $\Lambda \doteq \lambda_1 + \cdots + \lambda_d$. Letting

$$k \doteq \max\{k_1, \log E^{\mathbb{P}}[\exp\{cT_{N^0}\}|Q(0) = 0]\},$$

we complete the proof. ∎

3.8.5 Proof of Theorem 3.3.2

The proof of Theorem 3.3.2 is essentially a verification argument, utilizing the smooth subsolution $W^{\varepsilon,\delta}$.

We first argue that, for any absolutely continuous function $\phi : \mathbb{R}_+ \to \mathbb{R}_+^d$ with $\phi(0) = 0$ and $\tau = \inf\{t \geq 0 : \phi(t) \in \partial\} < \infty$, the inequality

$$\gamma \leq \int_0^\tau L(\phi(t), \dot{\phi}(t))\, dt \tag{3.8.6}$$

holds. To this end, fix arbitrarily $\varepsilon, \delta > 0$, and to ease notation denote $W = W^{\varepsilon,\delta}$. Then by the definition of L,

$$L(\phi(t), \dot{\phi}(t)) = L^{\pi(\phi(t))}(\dot{\phi}(t)) = \sup_{\alpha \in \mathbb{R}^d} \left[\langle \alpha, \dot{\phi}(t) \rangle - \max_{i \in \pi(\phi(t))} H^{(i)}(\alpha) \right].$$

In particular, thanks to Lemma 3.5.3 and that $\mathbb{H}_i(\alpha) = -2H^{(i)}(-\alpha/2)$,

$$\begin{aligned} L(\phi(t), \dot{\phi}(t)) \quad &\geq \quad \langle -DW(\phi(t))/2, \dot{\phi}(t) \rangle + \min_{i \in \pi(\phi(t))} \mathbb{H}_i(DW(\phi(t)))/2 \\ &\geq \quad \langle -DW(\phi(t))/2, \dot{\phi}(t) \rangle - Ke^{-\delta/\varepsilon}/2. \end{aligned}$$

Integrating both sides from $t = 0$ to $t = \tau$, it follows that

$$\int_0^\tau L(\phi(t), \dot{\phi}(t))\, dt \geq \frac{1}{2} \left[W(\phi(0)) - W(\phi(\tau)) - Ke^{-\delta/\varepsilon}\tau \right].$$

Observe that the boundary condition $W(x) \leq 0$ holds for all $x \in \partial$, and that by inequality (3.5.6)

$$W(\phi(0)) = W(0) \geq W^\delta(0) - \varepsilon \log |\mathbb{A}| = 2\gamma - \varepsilon \log |\mathbb{A}| - C_d\delta.$$

Therefore,

$$\int_0^\tau L(\phi(t), \dot{\phi}(t))\, dt \geq \gamma - \frac{1}{2} \left(\varepsilon \log |\mathbb{A}| + C_d\delta + Ke^{-\delta/\varepsilon}\tau \right).$$

Letting $\varepsilon, \delta \to 0$ but $\delta/\varepsilon \to \infty$, we arrive at the desired inequality (3.8.6).

In order to show the other direction, it suffices to construct an absolutely continuous function $\phi^* : \mathbb{R}_+ \to \mathbb{R}^d_+$ such that $\phi^*(0) = 0$ and $\tau^* \doteq \inf\{t \geq 0 : \phi^*(t) \in \partial\} < \infty$ with

$$\int_0^{\tau^*} L(\phi^*(t), \dot{\phi}^*(t))\, dt = \gamma. \tag{3.8.7}$$

This ϕ^* is actually an optimal trajectory for the calculus of variation problem. To this end, fix arbitrarily

$$A^* \in \operatorname{argmin} \left\{ \langle -\alpha^A, \bar{c} \rangle : A \in \mathbb{A} \right\}.$$

Consider the set valued function $F^* : \mathbb{R}^d_+ \to \mathbb{R}^d$ defined by

$$F^*(x) \doteq \text{the convex hull of } \{DH^{(i)}(-\alpha^{A^*}) : i \in \pi(x)\}$$

for every $x \in \mathbb{R}^d_+$. Clearly F^* is upper semicontinuous since π is so. It follows the classical theory of differential inclusions that there exists an absolutely continuous solution to

$$\dot{\phi}(t) \in F^*(\phi(t)), \qquad \phi(0) = 0. \tag{3.8.8}$$

With ϕ^* denoting this solution, we will argue that ϕ^* satisfies all the desired properties. Note that by definition, we can write for almost every t

$$\dot{\phi}^*(t) = \sum_{i \in \pi(\phi^*(t))} \rho_i^*(t) DH^{(i)}(-\alpha^{A^*}) \tag{3.8.9}$$

where $\rho_i^*(t) \geq 0$ and

$$\sum_{i \in \pi(\phi^*(t))} \rho_i^*(t) = 1. \tag{3.8.10}$$

Since

$$DH^{(i)}(-\alpha^{A^*}) = -\mu_i e^{\alpha_i^{A^*}} e_i + \sum_{j=1}^{d} \lambda_j e^{-\alpha_j^{A^*}} e_j,$$

it follows that

$$\dot{\phi}^*(t) = \sum_{j=1}^{d} \lambda_j e^{-\alpha_j^{A^*}} e_j - \sum_{i \in \pi(\phi^*(t))} \rho_i^*(t) \mu_i e^{\alpha_i^{A^*}} e_i. \tag{3.8.11}$$

Define for each $t \geq 0$,

$$\begin{aligned} h(t) &\doteq \min_{1 \leq i \leq d} c_i(\phi^*(t))_i, \\ I(t) &\doteq \operatorname{argmin}_{1 \leq i \leq d} c_i(\phi^*(t))_i. \end{aligned}$$

Obviously h is absolutely continuous. We claim that there exists a constant $a > 0$ such that

$$\dot{h}(t) \geq a \tag{3.8.12}$$

for almost every t. We first assume that $I(t) \neq \{1, \ldots, d\}$. It follows that $I(t) \cap$

$\pi(\phi^*(t)) = \emptyset$. Therefore for any $j \in I(t)$ (3.8.11) implies that

$$(\dot{\phi}^*(t))_j = \lambda_j e^{-\alpha_j^{A^*}}$$

Note that there exists a small neighborhood of t, say U, such that $I(s) \subset I(t)$ for every $s \in U$. The claim (3.8.12) follows readily. It remains to show for the case where $I(t) = \{1, \ldots, d\}$. In this case, $I(t) = \pi(\phi^*(t)) = \{1, \ldots, d\}$. Since the Lebesgue measure of the set

$$\Big\{t \geq 0 : \pi(\phi^*(t)) = \{1, \ldots, d\}, \pi(\dot{\phi}^*(t)) \neq \{1, \ldots, d\}\Big\}$$

is zero [5, Theorem A.6.3], we can further assume that $\pi(\dot{\phi}^*(t)) = \{1, \ldots, d\}$. In other words, there a exist constant b such that

$$b = c_i(\dot{\phi}^*(t))_i, \qquad i = 1, \ldots, d.$$

Thanks to (3.8.11) and (3.3.3),

$$\frac{b}{c_j} = (\dot{\phi}^*(t))_j = \begin{cases} \dfrac{\lambda_j \mu_j}{\mu_j - z^{A^*}} - \rho_j^*(t)(\mu_j - z^{A^*}) & \text{if } j \in A^*, \\ \lambda_j - \rho_j^*(t)\mu_j & \text{if } j \notin A^*. \end{cases}$$

Since the summation of $\{\rho_j^*(t) : j = 1, \ldots, d\}$ is one, it is not difficult to solve for b to obtain that

$$b = \left[\sum_{j \notin A^*} \frac{1}{c_j \mu_j} + \sum_{j \in A^*} \frac{1}{c_j(\mu_j - z^{A^*})}\right]^{-1} \cdot \left[\sum_{j \notin A^*} \frac{\lambda_j}{\mu_j} + \sum_{j \in A^*} \frac{\lambda_j \mu_j}{(\mu_j - z^{A^*})^2} - 1\right].$$

However, by the definition of z^{A^*} of (3.3.2),

$$\sum_{j \in A^*} \frac{\lambda_j \mu_j}{(\mu_j - z^{A^*})^2} - 1 = \sum_{j \in A^*} \frac{\lambda_j z^{A^*}}{\mu_j - z^{A^*}} > 0.$$

It follows that $b > 0$ and the claim (3.8.12) again holds.

The inequality (3.8.12) implies that the trajectory ϕ^* stays in the positive orthant $\mathbb{R}^d_+$ and the corresponding exit time τ^* is finite. It remains to show the equality (3.8.7). Note that by (2.3.1), (3.8.9), and (3.8.10),

$$L(\phi^*(t), \dot{\phi}^*(t)) \leq \sum_{i \in \pi(\phi^*(t))} \rho_i^*(t) L^{(i)}(DH^{(i)}(-\alpha^{A^*})).$$

By Lemma 3.3.1 and the conjugacy of $-\alpha^{A^*}$ and $DH^{(i)}(-\alpha^{A^*})$,

$$\begin{aligned} L^{(i)}(DH^{(i)}(-\alpha^{A^*})) &= \langle -\alpha^{A^*}, DH^{(i)}(-\alpha^{A^*}) \rangle - H^{(i)}(-\alpha^{A^*}) \\ &\leq \langle -\alpha^{A^*}, DH^{(i)}(-\alpha^{A^*}) \rangle \end{aligned}$$

Therefore

$$L(\phi^*(t), \dot{\phi}^*(t)) \leq \langle -\alpha^{A^*}, \dot{\phi}^*(t) \rangle.$$

Integrating both sides from 0 to τ^* and using $\phi^*(0) = 0$, we have

$$\int_0^{\tau^*} L(\phi^*(t), \dot{\phi}^*(t)) dt \leq \langle -\alpha^{A^*}, \phi^*(\tau^*) \rangle \leq \langle -\alpha^{A^*}, \bar{c} \rangle = \gamma.$$

Since the inequality (3.8.6) holds for every ϕ, in particular for ϕ^*, the equality (3.8.7) follows readily. ∎

Chapter 4

Importance Sampling and Large Deviations for a Buffer Overflow Event in a Tandem Queue with Server Slowdown

4.1 Introduction

The previous two chapters were concerned with a single multiplexed server. However in most applications one must consider a network of servers. This chapter is concerned with the simplest networks: a two-node tandem network. There is a slight complication added to this model, we consider a two-node tandem network where the downstream server is protected in such a way that the upstream server will drop its service rate whenever the downstream queue size exceeds a prespecified slow-down threshold. This type of queueing model was introduced in [27], and has potential applications in manufacturing and Ethernet design. It will be seen shortly that this slowdown mechanism will introduce discontinuities in the interior of the state space.

This chapter studies the overflow probability of the downstream queue during a busy cycle, assuming the system is stable. Little is known about the large deviation properties of this rare event and how to design efficient importance sampling schemes for simulation. To our best knowledge, the only existing work is [23], where the authors proposed heuristically the exponential decay rate of the overflow probability and the most likely path leading to overflow. [23] also suggested an importance sampling scheme based on the most likely path, which turned out to be efficient only in certain cases.

The goal of this chapter is to provide a construction of asymptotically optimal importance sampling schemes for general parameter values and also a rigorous analysis of the variational problem for the large deviations rate. The analysis relies on a recently developed game/subsolution approach toward importance sampling [13, 15], and utilizes the techniques in [5, 12] to overcome the difficulties introduced by the discontinuous dynamics at the slow-down threshold. We should remark that the current paper analyzes only the most relevant case from [23], in which the second queue is the more likely bottleneck without slowdown, and the first queue is the more likely bottleneck with slowdown. However, the other cases can be dealt with in an analogous fashion. The approach can also be extended to networks with Markov modulated arrival/service rates [15].

This chapter is organized as follows. In Sections 2 and 3 we describe the model and system dynamics. Section 4 states the large deviations result for the rare event probability of interest. Sections 5 through 7 are concerned with the Isaacs equation, construction of subsolutions, and the optimality of the corresponding dynamic importance sampling algorithms. The main result is stated in Section 8, numerical results are given in Section 9, and some technical proofs are collected in Section 10.

4.2 The model setup

We consider a variant of the standard two-node tandem Jackson network. Suppose that the arrival process is Poisson with rate λ, and the downstream service times are exponentially distributed with rate μ_2. The distribution of the upstream service times is as follows. Let $Q = \{(Q_1(t), Q_2(t)) : t \geq 0\}$ denote the system state, that is, $Q_1(t)$ is the length of the upstream queue at time t and $Q_2(t)$ is that of the downstream queue. Let n be the overflow level of the downstream queue, and θn the slow-down threshold [assuming $\theta \in (0,1)$] for the upstream queue. Then the upstream service time distribution is exponential with rate μ_1 if $Q_2 < \theta n$ and exponential with a smaller rate ν_1 if $Q_2 \geq \theta n$. The probability of interest is the overflow probability

$$p_n \doteq \mathbb{P}\{Q_2 = n \text{ before } Q = (0,0), \text{ after starting from } Q = (1,0)\}.$$

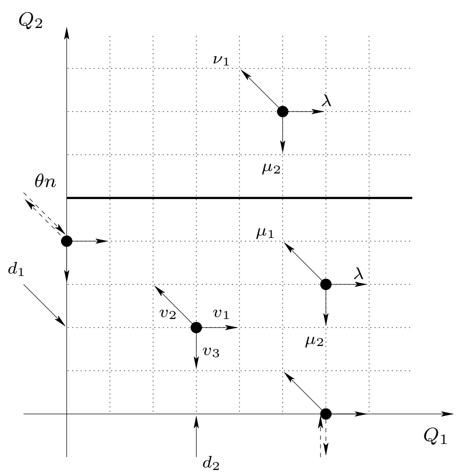

Figure 4.1: The system dynamics

To illustrate the main idea of game/subsolution approach, we restrict the analysis to the case where

$$\lambda < \nu_1 < \mu_2 \leq \mu_1. \tag{4.2.1}$$

Thus the downstream server is more likely to be the bottleneck when the system is below the slow-down threshold, while the upstream server becomes the bottleneck once the slow-down threshold is breached. When the slowdown mechanism is viewed as a control which tries to protect the downstream buffer from overflow, this is the most relevant case.

Remark 4.2.1. The paper [23] also considered two-node tandem Jackson networks without any server slow-down. For such a model the probability of interest p_n is a special case of one that appears in [13, Section 4.3.1] with $B_1 = \infty$ and $B_2 = 1$. One can use the subsolution there to build asymptotically optimal importance schemes for both $\mu_1 \leq \mu_2$ and $\mu_1 > \mu_2$.

4.3 The system dynamics

The state process Q is a continuous time pure jump Markov process defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In the interior of the state space the possible jumps of Q belong to the set

$$\mathbb{V} = \{v_1 = (1,0),\ v_2 = (-1,1),\ v_3 = (0,-1)\}.$$

To describe the discontinuous dynamics on the boundary, we allow the process Q to make fictitious jumps of size v_{i+1} on the boundary $\{Q_i = 0\}$, but they have to be accounted for by pushing back the state along the direction of constraints

$$d_i = -v_{i+1},$$

so that the queue sizes remain non-negative [see Figure 4.1]. For every $x = (x_1, x_2) \in \mathbb{R}_+^2$ and $v \in \mathbb{V}$ let

$$\pi[x, v] \doteq \begin{cases} 0, & \text{if } x_i = 0 \text{ and } v = v_{i+1} \text{ for some } i = 1, 2, \\ v, & \text{otherwise.} \end{cases} \tag{4.3.1}$$

The required projection on the boundary of the state space will be handled by π. We also define the jump intensity function $r(x, v)$ by

$$r(x, v_1) = \lambda, \quad r(x, v_2) = \begin{cases} \mu_1, & \text{if } x_2 < \theta \\ \nu_1, & \text{if } x_2 \geq \theta \end{cases}, \quad r(x, v_3) = \mu_2,$$

and the total intensity function by

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

Let $\{T_1, T_2, \ldots\}$ be the random jump times of the process Q with the convention $T_0 = 0$. The dynamics of Q are determined by the stochastic transition kernel $\Theta[\cdot|\cdot]$ on $\mathbb{R}_+ \times \mathbb{V}$ given $\mathbb{R}_+^2$, where

$$\begin{aligned} \Theta[dt, v|x] &\doteq \mathbb{P}\{T_{j+1} - T_j \in dt, Q(T_{j+1}) = nx + \pi[x, v] \,|\, Q(T_j) = nx\} \\ &= r(x, v) e^{-R(x)t} dt. \end{aligned} \tag{4.3.2}$$

In other words, the possible jumps of the process Q at state nx are $\{\pi[x, v] : v \in \mathbb{V}\}$ and the jump intensity from nx to $nx + \pi[x, v]$ is $r(x, v)$.

Notation. We collect here some useful notation [see Figure 4.2].

$$
\begin{aligned}
S &= \{(x_1, x_2) : x_1 > 0, \theta < x_2 < 1\}. \\
D &= \{(x_1, x_2) : x_1 > 0, 0 < x_2 < \theta\}. \\
\partial_1 &= \{(x_1, x_2) : x_1 = 0, x_2 > 0\}. \\
\partial_2 &= \{(x_1, x_2) : x_1 > 0, x_2 = 0\}. \\
\partial_e &= \{(x_1, x_2) : x_2 = 1\}.
\end{aligned}
$$

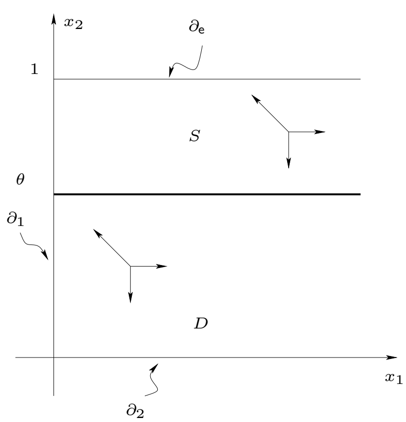

Figure 4.2: Notation for the State Space Partition

4.4 A representation of the exponential decay rate

The approach to be developed below is based on relations between large deviation properties and certain nonlinear partial differential equations (PDE). It will turn out that we will not need *solutions* to these PDE, but only *subsolutions*, and the PDE are simple enough that these subsolutions can be explicitly constructed. The PDE are described in terms of so-called Hamiltonians, with one for each region of different statistical behavior. As noted in the remark that follows, one can heuristically derive each Hamiltonian from the corresponding dynamics of the queueing process.

Given an arbitrary $\alpha = (\alpha_1, \alpha_2) \in \mathbb{R}^2$, the relevant Hamiltonians are defined as follows.

1. Below the slow-down threshold [region D],

$$H(\alpha) \doteq \lambda(e^{\alpha_1} - 1) + \mu_1(e^{\alpha_2 - \alpha_1} - 1) + \mu_2(e^{-\alpha_2} - 1).$$

2. Above the slow-down threshold [region S],

$$H_s(\alpha) \doteq \lambda(e^{\alpha_1} - 1) + \nu_1(e^{\alpha_2 - \alpha_1} - 1) + \mu_2(e^{-\alpha_2} - 1).$$

3. On the boundary ∂_1,

$$H_{\partial_1}(\alpha) \doteq \lambda(e^{\alpha_1} - 1) + \mu_2(e^{-\alpha_2} - 1).$$

4. On the boundary ∂_2,

$$H_{\partial_2}(\alpha) \doteq \lambda(e^{\alpha_1} - 1) + \mu_1(e^{\alpha_2 - \alpha_1} - 1).$$

Remark 4.4.1. We will (heuristically) show how the function H occurs in a PDE characterization of p_n. In a similar manner one can relate the Hamiltonians to the optimal possible performance in importance sampling (see [14]). We introduce the scaled process $X^n(t) = Q(nt)/n$, and for each $x \in D$ define

$$p_n(x) = \mathbb{P}_x \left\{ X_2^n \text{ exceeds 1 before } X^n = 0 \right\}.$$

By conditioning on the value of the first jump (given that the process starts at x), it follows that

$$(\lambda + \mu_1 + \mu_2) p_n(x) = \lambda p_n\left(x + \frac{v_1}{n}\right) + \mu_1 p_n\left(x + \frac{v_2}{n}\right) + \mu_2 p_n\left(x + \frac{v_3}{n}\right).$$

If the probabilities $p_n(x)$ decay at an exponential rate then one would expect $W_n(x) = -\frac{1}{n}\log p_n(x)$ to converge to some function $W(x)$. By rewriting the previous display in terms of W_n we see

$$\begin{aligned} 0 \;=\;& \lambda\left[\exp\left[-n\left(W_n(x+v_1/n)-W_n(x)\right)\right]-1\right] \\ &+\mu_1\left[\exp\left[-n\left(W_n(x+v_2/n)-W_n(x)\right)\right]-1\right] \\ &+\mu_2\left[\exp\left[-n\left(W_n(x+v_3/n)-W_n(x)\right)\right]-1\right]. \end{aligned}$$

For a smooth function W let DW denote its gradient. If each term of the form $n\left(W_n(x+v_i/n)-W_n(x)\right)$ also converges to $\langle DW(x), v_i\rangle$, then W should satisfy the PDE

$$\lambda\left(e^{-\langle DW,v_1\rangle}-1\right)+\mu_1\left(e^{-\langle DW,v_2\rangle}-1\right)+\mu_2\left(e^{-\langle DW,v_3\rangle}-1\right)=0.$$

This is just $H(-DW(x)) = 0$, and the other Hamiltonians can be derived similarly. The definition in a form similar to a moment generating function is motivated by standard notation in large deviation theory, and is responsible for the minus sign. ■

Denote by L, L_s, L_{∂_1}, and L_{∂_2} the Legendre transforms of H, H_s, H_{∂_1}, and H_{∂_2}, respectively. That is, for example,

$$L(\beta) = \sup_{\alpha\in\mathbb{R}^2}\left[\langle\alpha,\beta\rangle - H(\alpha)\right].$$

We continue by defining the so-called *local rate function* associated with the scaled processes $X^n(t) = Q(nt)/n, t \geq 0$. To ease exposition, we use the notation $L_1 \oplus \cdots \oplus L_d$ to denote the inf-convolution of convex functions $\{L_1, L_2, \ldots, L_d\}$, that is, for any $\beta \in \mathbb{R}^2$

$$[L_1\oplus\cdots\oplus L_d](\beta) \doteq \inf\left\{\sum_{i=1}^d \rho_i L_i(\beta_i) : \rho_i \geq 0, \sum_{i=1}^d \rho_i = 1, \sum_{i=1}^d \rho_i\beta_i = \beta\right\}.$$

The local rate function, denoted by $L(x, \beta)$, is defined as follows.

1. For $x \in S$, $L(x, \beta) \doteq L_s(\beta)$, and for $x \in D$, $L(x, \beta) \doteq L(\beta)$.

2. For $x \in \partial_2$, $L(x, \beta) = [L \oplus L_{\partial_2}](\beta)$ if $\beta_2 \geq 0$ and ∞ if $\beta_2 < 0$.

3. For $x \in \partial_1$,

$$L(x, \beta) \doteq \begin{cases} [L_s \oplus L_{\partial_1}](\beta), & \text{if } x_2 > \theta, \beta_1 \geq 0, \\ [L \oplus L_s \oplus L_{\partial_1}](\beta), & \text{if } x_2 = \theta, \beta_1 \geq 0, \\ [L \oplus L_{\partial_1}](\beta), & \text{if } x_2 < \theta, \beta_1 \geq 0, \\ \infty, & \text{if } \beta_1 < 0. \end{cases}$$

4. For $x = (0, 0)$, $L(x, \beta) = [L \oplus L_{\partial_1} \oplus L_{\partial_2}](\beta)$ if $\beta_1, \beta_2 \geq 0$ and ∞ otherwise.

5. For $x \in \{(x_1, \theta) : x_1 > 0\}$, $L(x, \beta) \doteq [L \oplus L_s](\beta)$.

Theorem 4.4.2.

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

where the infimum is taken over all absolutely continuous functions $\phi : [0, \infty) \to \mathbb{R}_+^2$ *such that*

$$\phi(0) = 0, \quad \tau \doteq \inf\{t \geq 0 : [\phi(t)]_2 = 1\} < \infty.$$

The main difficulties in the proof of Theorem 4.4.2 are due to the discontinuous dynamics of the prelimit process X^n. The discontinuities come in two forms: those along the interface between the region D and the slowdown region S, and those along the boundary of the state space due to non-negativity constraints on queue length. The first type of discontinuity is well understood in this particular setting (two regions of continuous behavior separated by a single smooth interface). See for example [5, Chapter 7] where the model discussed is a discrete time random walk, and [12, Theorem 3.1] for an application in a continuous time setting. In a very general setting (as in [5, Chapter 7]), the local rate function appears to be more complicated than it does here, since additional "stability-about-the-interface" constraints not present in the inf-convolution must be added. However, owing to the structure of the server slowdown dynamics these constraints are implicitly contained in the inf-convolution formula.

The discontinuities along the boundary can be dealt with by considering an unconstrained process which is mapped to the constrained process via a Skorokhod mapping Γ. It is easily shown that the mapping Γ is Lipschitz [8]. With the continuity of Γ the proof is now a simple application of the Contraction Principle.

4.5 The Isaacs equation

It has been established that importance sampling is closely related to differential games and that subsolutions to the corresponding Isaacs equation (a nonlinear PDE) can be used for constructing efficient importance sampling schemes [13, 15]. The Isaacs equation can be stated in terms of the Hamiltonians introduced in Section 4.

Define the function ℓ by

$$\ell(x) = \begin{cases} x \log x - x + 1, & \text{if } x \geq 0, \\ \infty \quad , & \text{if } x < 0. \end{cases} \tag{4.5.1}$$

Let $\mathcal{R}$ be the collection of functions from $\mathbb{V}$ to $(0, \infty)$, and for every $\hat{r} \in \mathcal{R}$ define

$$\mathbb{F}(\hat{r}) \doteq \sum_{v \in \mathbb{V}} \hat{r}[v] \cdot v.$$

For each $\alpha = (\alpha_1, \alpha_2) \in \mathbb{R}^2$, the direct continuous time analogue of the discrete time

Hamiltonian used in [13] is

$$\sup_{\bar{r}\in\mathcal{R}}\inf_{\hat{r}\in\mathcal{R}}\left[\langle\alpha,\mathbb{F}(\hat{r})\rangle+\sum_{v\in\mathbb{V}}\left(\hat{r}[v]\log\left(\frac{\bar{r}[v]}{r(x,v)}\right)+r(x,v)-\bar{r}[v]\right)\right.$$
$$\left.+\sum_{v\in\mathbb{V}}r(x,v)\ell\left(\frac{\hat{r}[v]}{r(x,v)}\right)\right].$$

We will find it convenient to rewrite the Hamiltonian in the form

$$\mathbb{H}(x,\alpha)\doteq\sup_{\bar{r}\in\mathcal{R}}\inf_{\hat{r}\in\mathcal{R}}\left[\langle\alpha,\mathbb{F}(\hat{r})\rangle+2\sum_{v\in\mathbb{V}}r(x,v)\ell\left(\frac{\hat{r}[v]}{r(x,v)}\right)-\sum_{v\in\mathbb{V}}\bar{r}[v]\ell\left(\frac{\hat{r}[v]}{\bar{r}[v]}\right)\right].$$

Recall that DW denotes the gradient for a smooth function W. A (classical) solution to the Isaacs equation is a continuously differentiable function $W:\mathbb{R}_+\times[0,1]\to\mathbb{R}$ satisfying

1. $\mathbb{H}(x,DW(x))=0$ for $x\in(0,\infty)\times(0,1)$,
2. $\langle DW(x),d_i\rangle=0$ for $x\in\partial_i$,
3. $W(x)=0$ for $x\in\partial_e$.

A formal derivation analogous to the one given in Section 4 shows that this PDE should characterize the optimal decay rate among all importance sampling schemes. In general one would not expect classical solutions to exist, and thus one should work with a weaker notion of solution. However, we will not need or use solutions, but rather only certain subsolutions whose regularity properties will be specified when they are introduced.

We recall the functions H and H_s introduced in Section 4. Abusing the notation, we define for every α

$$\mathbb{H}_s(\alpha)\doteq-2H_s(-\alpha/2),\quad\mathbb{H}(\alpha)\doteq-2H(-\alpha/2).$$

We have the following result, whose proof is straightforward and thus omitted.

Proposition 4.5.1. *Fix an arbitrary $\alpha = (\alpha_1, \alpha_2) \in \mathbb{R}^2$. Then for $x = (x_1, x_2)$ the Hamiltonian $\mathbb{H}(x, \alpha)$ satisfies the following properties.*

1. *If $x_2 \geq \theta$, $\mathbb{H}(x, \alpha) = \mathbb{H}_s(\alpha)$ with the saddle point*

$$\bar{r}_s^*(\alpha) = \hat{r}_s^*(\alpha) = \left(\lambda e^{-\alpha_1/2}, \nu_1 e^{(\alpha_1 - \alpha_2)/2}, \mu_2 e^{\alpha_2/2}\right).$$

2. *If $x_2 < \theta$, $\mathbb{H}(x, \alpha) = \mathbb{H}(\alpha)$ with the saddle point*

$$\bar{r}^*(\alpha) = \hat{r}^*(\alpha) = \left(\lambda e^{-\alpha_1/2}, \mu_1 e^{(\alpha_1 - \alpha_2)/2}, \mu_2 e^{\alpha_2/2}\right).$$

Remark 4.5.2. The Isaacs equation is associated with a differential game where the $\bar{r}$-player represents the choice of change of measure and the $\hat{r}$-player is introduced through a representation formula for the large deviation rate of decay. The explicit formula for the $\bar{r}$-component of the saddle point given in Proposition 4.5.1 is particularly useful in the construction of importance sampling schemes. Roughly speaking, for a given subsolution $\bar{W}$, depending on the state of the process, we use $\bar{r}_s^*(D\bar{W})$ or $\bar{r}^*(D\bar{W})$ as the alternative jump intensity function for simulation.

Remark 4.5.3. In general one should be more careful with the definition of the Isaacs equation on the interface $\{(x_1, x_2) : x_2 = \theta\}$, and in fact if we were considering a more general class of rare event estimation problems one would use $(\mathbb{H}_s \wedge \mathbb{H})(DW) = 0$ on the interface. This observation is crucial for the construction of subsolutions that lead to asymptotically optimal importance sampling schemes, but only when the most likely path (i.e. a minimizer to the variational problem in Theorem 4.4.2) leading to the rare event will spend non-trivial time on the interface (i.e., the Lebesgue measure of the time the minimizer is on the interface is positive). In the present setting, however, it will turn out that the most likely path does not spend positive time on the interface.

4.6 Subsolutions and an importance sampling algorithm

A *classical subsolution* to the Isaacs equation is a continuously differentiable function $\bar{W} : \mathbb{R}_+ \times [0,1] \to \mathbb{R}$ such that

1. $\mathbb{H}(x, D\bar{W}(x)) \geq 0$ for all $x \in (0,\infty) \times (0,1)$,
2. $\langle D\bar{W}(x), d_i \rangle \geq 0$ for all $x \in \partial_i$,
3. $\bar{W}(x) \leq 0$ for all $x \in \partial_e$.

Classical subsolutions can often be constructed from a mollification of piecewise affine subsolutions, see [13, 15] and the many examples therein.

We will associate importance sampling schemes to subsolutions. The performance of the scheme will be measured by the value of the subsolution at 0, with larger values of $\bar{W}(0)$ indicating better performance. Note that for any subsolution $\bar{W}(0)$ is bounded above by the value of the solution at the origin.

4.6.1 Piecewise affine subsolutions

The goal of this section is to construct a piecewise affine subsolution whose value at the origin is maximal. Consider the variational problem introduced in the statement of Theorem 4.4.2, but with a general initial condition $x \in [0,1] \times \mathbb{R}_+$:

$$V(x) = \inf\left\{ \int_0^\tau L(\phi(t), \dot{\phi}(t))dt : \phi(0) = x, [\phi(\tau)]_2 = 1, \tau \geq 0 \right\}. \tag{4.6.1}$$

It can be shown formally by a dynamic programming argument that $\mathbb{H}(x, 2DV(x)) = 0$, which suggests that the large deviations most likely path might be useful in constructing subsolutions. In [23, Proposition 10] the authors propose that the most likely path to overflow will travel along the boundary ∂_1 from $(0,0)$ to $(0,\theta)$ pushing against the boundary during this leg of the journey. What is meant by "pushing against the boundary" is that in the absence of non-negativity constraints on queue length, the (large deviation) change of jump rates would lead to a trajectory whose velocity has a negative first component. This velocity is projected back along the direction d_1 at no cost. From $(0,\theta)$ to $(0,1)$ the path will also travel along ∂_1 but this time gliding along the boundary. What is meant here is that the change of rates produces exactly velocity zero in the first component, and no projection is needed. If the conjecture of [23] is correct, these statements give important information on the gradient of the solution in a neighborhood of the optimal trajectory. If the subsolution we construct is to be close to the solution at $x = 0$, then it must be close all along the optimal trajectory. Hence we obtain necessary conditions on the gradient of the subsolution along the optimal trajectory.

We next formally derive constraints on the gradient of the function V along the proposed optimal trajectory. We say "formal" since it is not known if V is sufficiently smooth to justify all the calculations. However, our goal is to simply motivate the construction of a particular subsolution. As noted above, in region D the optimal trajectory pushes into the boundary and is returned along the direction d_1 at no cost. This implies that the value function V should be constant along this direction. We thus obtain two constraints on $\alpha^{[1]} = 2DV(x)$:

$$\mathbb{H}(\alpha^{[1]}) = 0, \quad \langle d_1, \alpha^{[1]} \rangle = 0.$$

In region S $2DV(x)$ should satisfy the corresponding part of the PDE that applies on the interior of the domain. At the same time, the linking of the optimal velocity with the gradient via dynamic programming implies that $\alpha^{[0]} = 2DV(x)$ should be dual (in the sense of convex duality) to a point whose first component is zero. Thus we obtain the constraints

$$\mathbb{H}_s(\alpha^{[0]}) = 0, \quad \frac{\partial \mathbb{H}_s}{\partial \alpha_1}(\alpha^{[0]}) = 0. \tag{4.6.2}$$

There are two roots with this property, and the correct (i.e., useful) root also satisfies $\alpha_1^{[0]}, \alpha_2^{[0]} < 0$ [see Lemma 4.6.1].

This identifies constraints on the gradient of a subsolution in the neighborhood of the optimal trajectory. In order to satisfy the boundary condition along the boundary ∂_2 we need the gradients $\alpha^{[2]} = 2\log(\mu_2/\lambda)(-1, 0)$, and $\alpha^{[3]} = (0, 0)$. As we will see, these choices allow the construction of a function which satisfies the subsolution property at all points in the domain and with a nearly optimal value at zero. Figure 4.3 shows how these vectors relate to the Hamiltonians.

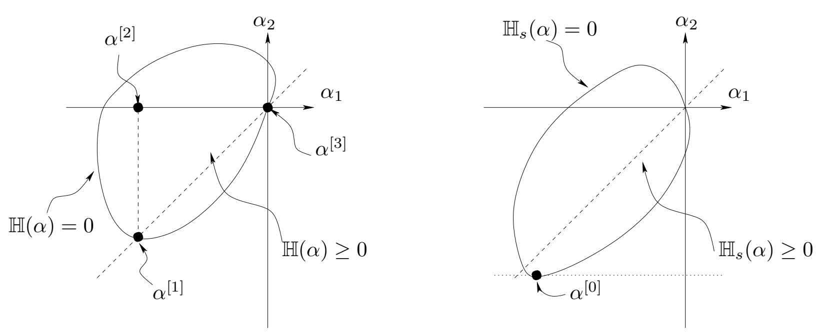

Figure 4.3: The gradients of the subsolution

In terms of the problem data we find

$$\alpha^{[1]} = 2\log(\mu_2/\lambda)(-1, -1), \alpha^{[2]} = 2\log(\mu_2/\lambda)(-1, 0), \text{ and} \alpha^{[3]} = (0, 0).$$

The existence of $\alpha^{[0]}$ is dealt with in the following lemma.

Lemma 4.6.1. *There exists a unique $z > 1$ that satisfies the equation*

$$2\sqrt{\lambda\nu_1 z} + \frac{\mu_2}{z} = \lambda + \nu_1 + \mu_2. \tag{4.6.3}$$

Furthermore, the unique solution $\alpha^{[0]}$ to equation (4.6.2) with $\alpha_1^{[0]}, \alpha_2^{[0]} < 0$ can be expressed as

$$\alpha^{[0]} = (\alpha_1^{[0]}, \alpha_2^{[0]}) = (-\log(\nu_1 z/\lambda), -2\log z)\,. \tag{4.6.4}$$

We also have

$$\mathbb{H}_s(\alpha^{[1]}) = 0, \qquad \alpha_2^{[0]} < \alpha_1^{[1]} < \alpha_1^{[0]}, \qquad \langle \alpha^{[0]}, d_1\rangle \geq 0.$$

The proof is deferred to an appendix. Note that the equation (4.6.3) for z is the same as equation (31) in [23] with z replaced by z^{-1}. We also let

$$\gamma = (1-\theta)\log z + \theta\log(\mu_2/\lambda). \tag{4.6.5}$$

The first term is the cost of the optimal trajectory between $(0,\theta)$ and $(0,1)$, the second term is the cost of the trajectory traveling between $(0,0)$ and $(0,\theta)$. Fix an arbitrary small $\delta > 0$, and define the affine functions

$$\begin{aligned}
\bar{W}_0^\delta(x) &= \langle x, \alpha^{[0]}\rangle + 2\log z,\\
\bar{W}_1^\delta(x) &= \langle x, \alpha^{[1]}\rangle + 2\gamma,\\
\bar{W}_2^\delta(x) &= \langle x, \alpha^{[2]}\rangle + 2\gamma - \delta,\\
\bar{W}_3^\delta(x) &= \langle x, \alpha^{[3]}\rangle + 2\gamma - 2\delta.
\end{aligned}$$

Let $\bar{W}^\delta \doteq \bar{W}_0^\delta \wedge \bar{W}_1^\delta \wedge \bar{W}_2^\delta \wedge \bar{W}_3^\delta$. Figure 4.4 depicts the gradients of $\bar{W}^\delta$ in different regions of the state space, with $\bar{W}^\delta = \bar{W}_i^\delta$ in region G_i.

Lemma 4.6.1 and a simple computation show that $\bar{W}^\delta$ indeed defines a piecewise

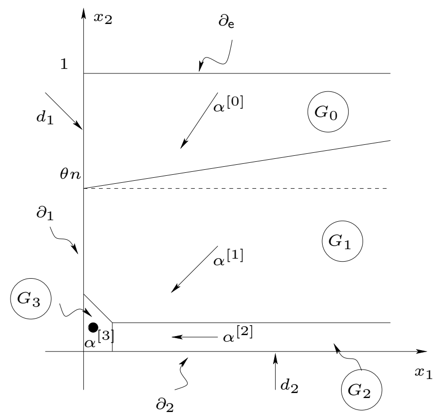

Figure 4.4: The subsolution

affine subsolution, that is, $\bar{W}^\delta$ satisfies the definition of classical subsolution in the regions where it is smooth.

Remark 4.6.2. In Remark 4.5.3 it was mentioned that the subsolution would be more complicated if the optimal trajectory spent significant amount of time on the interface. In particular this would require that an additional affine function $\bar{W}_4^\delta$ be constructed such that $(\mathbb{H} \wedge \mathbb{H}_s)(D\bar{W}_4^\delta) \geq 0$. In addition it would be necessary that $\bar{W}^\delta = \bar{W}_4^\delta$ for x in an $O(\delta)$ strip around the interface.

4.6.2 The mollification

In the proof of asymptotic optimality of the estimator $\hat{p}_n$ it is essential that the subsolutions we work with have a bounded second derivative. As far as the authors know this boundedness condition is indispensable. Since the function $\bar{W}^\delta$ is not smooth, this problem is resolved by mollifying the function. As we will see, the implementation of the algorithm corresponding to the mollified function is only slightly more computationally demanding than it would be for the corresponding un-mollified version.

A natural mollification technique for this problem is *exponential weighting.* See [13, 15] for a discussion of the benefits of this method and a comparison with other standard mollifications. Let ε be a small positive number, and define

$$W^{\varepsilon,\delta}(x) \doteq -\varepsilon \log \sum_{i=0}^{3} \exp\left\{-\frac{1}{\varepsilon}\bar{W}_i^{\delta}(x)\right\}.$$

The function $W^{\varepsilon,\delta}$ is continuously differentiable, and

$$DW^{\varepsilon,\delta}(x) = \sum_{i=0}^{3} \rho_i^{\varepsilon,\delta}(x)\alpha^{[i]}, \quad \rho_i^{\varepsilon,\delta}(x) \doteq \frac{\exp\left\{-\bar{W}_i^{\delta}(x)/\varepsilon\right\}}{\sum_{k=0}^{3}\exp\left\{-\bar{W}_k^{\delta}(x)/\varepsilon\right\}}.$$

Note that for every x, $\{\rho_i^{\varepsilon,\delta}(x) : i = 0,1,2,3\}$ forms a probability vector since

$$\rho_i^{\varepsilon,\delta}(x) > 0, \qquad \sum_{i=0}^{3} \rho_i^{\varepsilon,\delta}(x) = 1.$$

4.6.3 The importance sampling algorithm

For each $i = 0,1,2,3$, let

$$\bar{r}_s^{[i]} \doteq \bar{r}_s^*(\alpha^{[i]}), \quad \bar{r}^{[i]} \doteq \bar{r}^*(\alpha^{[i]}). \tag{4.6.6}$$

Explicit formulas for $\{\bar{r}_s^{[i]}, \bar{r}^{[i]}\}$ can be found in Remark 4.6.3. For each i, also define the jump intensity functions $\{\bar{r}_i(x,v) : v \in \mathbb{V}\}$ and total intensity function $\bar{R}_i(x)$ by

$$\bar{r}_i(x,v_k) \doteq \begin{cases} \left[\bar{r}_s^{[i]}\right]_k & \text{if } x_2 \geq \theta \\ \left[\bar{r}^{[i]}\right]_k & \text{if } x_2 < \theta \end{cases}, \qquad \bar{R}_i(x) \doteq \sum_{k=0}^{3} \bar{r}_i(x,v_k).$$

The corresponding stochastic transition kernel on $\mathbb{R}_+ \times \mathbb{V}$ given $\mathbb{R}_+ \times [0,1]$, denoted by $\bar{\Theta}^{[i]}$, is

$$\bar{\Theta}^{[i]}[dt, v|x] \doteq \bar{r}_i(x,v)e^{-\bar{R}_i(x)t}dt.$$

That is, the stochastic transition kernel $\bar{\Theta}^{[i]}$ corresponds to a continuous time Markov process whose jump intensity from nx to $nx + \pi[x,v]$ is $\bar{r}_i(x,v)$.

The dynamic importance sampling scheme corresponding to $W^{\varepsilon,\delta}$ uses a mixture of $\{\bar{\Theta}^{[i]}\}$:

$$\bar{\Theta}^n[\cdot|x] = \bar{\Theta}^{\varepsilon,\delta}[\cdot|x] = \sum_{i=0}^{3} \rho_i^{\varepsilon,\delta}(x)\bar{\Theta}^{[i]}[\cdot|x],$$

and the corresponding importance sampling estimator $\hat{p}_n$ is just as defined in [(1.2.5) from the introduction]. In other words, the importance sampling simulates the process Q in the following fashion. Given that the current state of Q is nx, the algorithm selects a random index I taking values in $\{0, 1, 2, 3\}$ according to the weights $\{\rho_i^{\varepsilon,\delta}(x) : i = 0, 1, 2, 3\}$. Then the algorithm simulates the sojourn time according to the exponential distribution with rate $\bar{R}_I(x)$ and the jump size $\pi[x, v]$ with probability $\bar{r}_I(x, v)/\bar{R}_I(x)$.

This algorithm is slightly different from the algorithm presented in the weighted-serve-the-longest(WSLQ) example (discussed in subsection 3.5.3). In that case the change of measure used preserves the Markov property of the queue length process. Using a similar change of measure for this example would change the queue length process to a Markov process with jump rates and total intensity given by

$$\bar{r}(x, v) = \sum_{i=0}^{3} \rho_i^{\varepsilon,\delta}(x)\bar{r}_i(x, v) \text{ and } \bar{R}(x) = \sum_{k=0}^{3} \bar{r}(x, v).$$

The performance of the two methods are nearly identical, but in order to be consistent with the algorithm used in [13] we will utilize the transition kernel $\bar{\Theta}^n[\cdot|x]$.

In general one can allow ε, δ to depend on n, and denote them by ε_n, δ_n. The corresponding stochastic transition kernel will be $\bar{\Theta}^n[\cdot|x] = \bar{\Theta}^{\varepsilon_n,\delta_n}[\cdot|x]$, and the corresponding importance sampling estimator will still be denoted by $\hat{p}_n$ when no confusion is incurred.

Remark 4.6.3. Using Proposition 4.5.1, Lemma 4.6.1 and (4.6.6) it follows that

$$\bar{r}_s^{[0]} = (\sqrt{\lambda\nu_1 z}, \sqrt{\lambda\nu_1 z}, \mu_2/z), \quad \bar{r}^{[0]} = (\sqrt{\lambda\nu_1 z}, \mu_1\sqrt{\lambda z/\nu_1}, \mu_2/z),$$

$$\bar{r}_s^{[1]} = (\mu_2, \nu_1, \lambda), \quad \bar{r}^{[1]} = (\mu_2, \mu_1, \lambda),$$

$$\bar{r}_s^{[2]} = (\mu_2, \lambda\nu_1/\mu_2, \mu_2), \quad \bar{r}^{[2]} = (\mu_2, \lambda\mu_1/\mu_2, \mu_2),$$

$$\bar{r}_s^{[3]} = (\lambda, \nu_1, \mu_2), \quad \bar{r}^{[3]} = (\lambda, \mu_1, \mu_2).$$

4.7 The main results

We recall the definition of γ in (4.6.5).

Theorem 4.7.1. *Suppose $\delta_n \to 0$, $\varepsilon_n/\delta_n \to 0$ and $n\varepsilon_n \to \infty$. Then the corresponding importance sampling estimator $\hat{p}_n$ satisfies*

$$\liminf_n -\frac{1}{n} \log [\text{2nd moment of } \hat{p}_n] \geq 2\gamma.$$

This theorem can be shown by a verification argument analogous to [13, Theorem 3.18]. Since by Lemma 4.6.1 $\mathbb{H}(\alpha^{[1]}) = \mathbb{H}_s(\alpha^{[1]}) = 0$, the subsolution inequality holds on both sides of the slow-down interface (see Figure 4). Hence the only major distinction is that in [13] the domain is assumed to be compact so that the exit time can be shown to have bounded moment generating function in a small neighborhood of origin. This compactness condition turns out to be unnecessary, and a proof of this fact is given in Section 4.10.

Theorem 4.7.2. *We have the large deviation result*

$$\lim_n -\frac{1}{n} \log p_n = \gamma.$$

Proof. Since $E^{\mathbb{Q}}[\hat{p}_n^2] \geq (E^{\mathbb{Q}}[\hat{p}_n])^2 = p_n^2$, Theorem 4.7.1 implies the large deviations upper bound

$$\liminf_n -\frac{1}{n} \log p_n \geq \gamma. \tag{4.7.1}$$

Note that this upper bound can also be shown by applying a direct verification argument [16] to the control representation of the large deviations rate function [see Theorem 4.4.2], using the classical subsolution $W^{\varepsilon_n,\delta_n}/2$.

We now consider the lower bound. By Theorem 4.4.2, we only need to show that there exists an absolutely continuous function $\phi^* : \mathbb{R}_+ \to \mathbb{R}_+^2$ such that

$$\int_0^\tau L(\phi^*(t), \dot{\phi}^*(t))\, dt \leq \gamma, \qquad \text{where } \tau \doteq \inf\{t \geq 0 : [\phi^*(t)]_2 = 1\} < \infty.$$

Theorems 4.4.2 and 4.7.1 imply that if such a ϕ^* exists it will be an optimal path.

The construction of ϕ^* is based on the change of measure determined by $\bar{r}_s^{[0]}$ and $\bar{r}^{[1]}$ [see Remark 4.6.3]. In other words, let

$$\beta_s^* \doteq \sum_{i=1}^3 \left[\bar{r}_s^{[0]}\right]_i v_i = (0, \sqrt{\lambda\nu_1 z} - \mu_2/z),$$

and

$$\beta^* \doteq \sum_{i=1}^3 \left[\bar{r}^{[1]}\right]_i v_i = (\mu_2 - \mu_1, \mu_1 - \lambda).$$

Since $[\beta^*]_1 = \mu_2 - \mu_1 \leq 0$, we further define, abusing the notation, $\pi[\beta^*]$ to be the projection of β^* onto ∂_1 along the direction of constraint $d_1 = (1, -1)$. Thus

$$\pi[\beta^*] = \beta^* - [\beta^*]_1\, d_1 = (0, \mu_2 - \lambda).$$

To ease exposition, write $\beta_s^* \doteq (0, u_s)$ and $\pi[\beta^*] \doteq (0, u)$, where

$$u_s \doteq \sqrt{\lambda\nu_1 z} - \mu_2/z, \quad u \doteq \mu_2 - \lambda.$$

Note that $u_s > 0$ thanks to (4.9.1) and $u > 0$ since $\mu_2 > \lambda$. Define $\phi^*(t)$ such that

$$\dot{\phi}^*(t) = \begin{cases} \pi[\beta^*], & \text{if } 0 \le t < \theta u^{-1}, \\ \beta_s^*, & \text{if } \theta u^{-1} \le t \le \theta u^{-1} + (1-\theta) u_s^{-1}. \end{cases}$$

Then ϕ^* is a vertical path passing through $(0,0) \to (0,\theta) \to (0,1)$, and by the definition of L [see Section 4.4],

$$\int_0^\tau L(\phi^*(t), \dot{\phi}^*(t))\, dt = \theta u^{-1} \cdot [L \oplus L_{\partial_1}](\pi[\beta^*]) + (1-\theta) u_s^{-1} \cdot [L_s \oplus L_{\partial_1}](\beta_s^*).$$

However, note that L_s and H_s are conjugate functions and β_s^* is conjugate to $-\alpha^{[0]}/2$, since by direct computation

$$D_\alpha H_s(-\alpha^{[0]}/2) = (0, u_s) = \beta_s^*.$$

Therefore, observing $H_s(-\alpha^{[0]}/2) = -\mathbb{H}_s(\alpha^{[0]})/2 = 0$, we have

$$[L_s \oplus L_{\partial_1}](\beta_s^*) \le L_s(\beta_s^*) = \langle -\alpha^{[0]}/2, \beta_s^* \rangle - H_s(-\alpha^{[0]}/2) = u_s \log z.$$

The local rate functions L, L_s, L_{∂_1}, L_{∂_2}, and their inf-convolutions have standard representations in terms of the ℓ function as defined in (4.5.1) [12, 25]. In particular,

$$\begin{aligned} [L \oplus L_{\partial_1}](\beta) \;\; = \;\; \inf \Bigg\{ & \lambda \ell\left(\frac{\bar{\lambda}}{\lambda}\right) + \bar{\rho}\mu_1 \ell\left(\frac{\bar{\mu}_1}{\mu_1}\right) + \mu_2 \ell\left(\frac{\bar{\mu}_2}{\mu_2}\right) : \\ & \bar{\rho} \in [0,1],\ \bar{\lambda} v_1 + \bar{\rho}\bar{\mu}_1 v_2 + \bar{\mu}_2 v_3 = \beta \Bigg\}. \end{aligned}$$

Since

$$\pi[\beta^*] = \bar{\lambda} v_1 + \bar{\rho}\bar{\mu}_1 v_2 + \bar{\mu}_2 v_3$$

where

$$\bar{\rho} \doteq \mu_2/\mu_1 \in [0,1], \quad \bar{\lambda} = \mu_2, \quad \bar{\mu}_1 = \mu_1, \quad \bar{\mu}_2 = \lambda,$$

it follows that

$$\begin{aligned}[L \oplus L_{\partial_1}](\pi[\beta^*]) &\leq \bar{\rho}\mu_1 \ell\left(\frac{\bar{\mu}_1}{\mu_1}\right) + \lambda \ell\left(\frac{\bar{\lambda}}{\lambda}\right) + \mu_2 \ell\left(\frac{\bar{\mu}_2}{\mu_2}\right) \\ &= \lambda \ell\left(\frac{\mu_2}{\lambda}\right) + \mu_2 \ell\left(\frac{\lambda}{\mu_2}\right) \\ &= u \log(\mu_2/\lambda).\end{aligned}$$

Therefore,

$$\int_0^\tau L(\phi^*(t), \dot{\phi}^*(t))\, dt \leq \theta \log(\mu_2/\lambda) + (1-\theta)\log z = \gamma.$$

This completes the proof. ∎

Corollary 4.7.3. *Suppose $\delta_n \to 0$, $\varepsilon_n/\delta_n \to 0$ and $n\varepsilon_n \to \infty$. Then the corresponding importance sampling estimator $\hat{p}_n$ is asymptotically optimal.*

Remark 4.7.4. Analogous to [13, Theorem 3.6], a near asymptotic optimality result can be shown for the case where $\varepsilon_n \equiv \varepsilon$ and $\delta_n \equiv \delta$. It also suggests that a good strategy is to set $\delta_n = -\varepsilon_n \log \varepsilon_n$ [13, Remark 3.7].

Remark 4.7.5. Our results show that the heuristically derived exponential decay rate of p_n and the limit most likely path to overflow in [23] are indeed correct.

4.8 Numerical results

As discussed in Remark 1.2.2, the numerical implementation in this section is carried out using the embedded discrete-time Markov chain $Z = \{Z(j) = Q(T_j) : j \geq 0\}$. That is, given that the current state of $Z(j)$ is nx, the algorithm selects a random index I taking values in $\{0,1,2,3\}$ according to the weights $\{\rho_i^{\varepsilon,\delta}(x) : i = 0,1,2,3\}$.

Then the algorithm generates $Z(j+1) = Z(j) + \pi[x, \triangle_{j+1}]$ with $\mathbb{Q}(\triangle_{j+1} = v) = \bar{r}_I(x,v)/\bar{R}_I(x)$. The corresponding importance sampling estimator is

$$\bar{p}_n \doteq 1_{\{Z_2(N)=n\}} \prod_{j=0}^{N-1} \frac{r[Z(j)/n, \triangle_{j+1}]/R(Z(j)/n)}{\sum_{i=0}^{3} \rho_i^{\varepsilon,\delta}(Z(j)/n)\bar{r}_i[Z(j)/n, \triangle_{j+1}]/\bar{R}_i(Z(j)/n)},$$

where

$$N \doteq \inf\{k \geq 0 : Z_2(k) = n \text{ or } Z(k) = 0\}.$$

As noted in the introduction

$$E^{\mathbb{Q}}\left[\hat{p}_n | Q(T_1), \ldots, Q(T_N)\right] = \bar{p}_n.$$

Therefore, since $\hat{p}_n$ is asymptotically optimal, it follows that $\bar{p}_n$ is also asymptotically optimal.

We present simulations for the two cases $(\lambda, \mu_1, \mu_2, \nu_1) = (0.1, 0.7, 0.2, 0.15)$ and $(0.3, 0.36, 0.34, 0.32)$, with $n = 20, 50, 100$. These cases are distinguished in that, according to [23], a variant on the standard "open loop" approach to importance sampling can be expected to be efficient for the first but not the second. Numerical results are also presented in [23] using such a scheme. A comparison of the numerical data supports the statements of [23]. Indeed, both algorithms perform well on the data of Table 1 with the state independent scheme producing slightly smaller confidence intervals, while only the state dependent algorithm appears to be asymptotically optimal for the data of Table 2.

For each case, we let the mollification parameter take the form $\varepsilon_n \doteq c/\sqrt{n}$ for a constant c, and set $\delta_n \doteq -\varepsilon_n \log \varepsilon_n$ as suggested by Remark 4.7.4. Asymptotic optimality of $\bar{p}_n$ follows from Corollary 4.7.3 and the discussion in the preceding paragraph.

The constant c is determined so that when $n = 20$, $\delta_n \approx 0.05\gamma$, where γ is the corresponding large deviation rate [see (4.6.5)]. It follows that the value of the piecewise affine subsolution $\bar{W}^{\delta_n}$ at the origin is $\bar{W}^{\delta_n}(0,0) = 2\gamma - 2\delta_n \approx 0.95 \cdot 2\gamma$ when $n = 20$. We wish to point out that the algorithm is fairly robust in that different small values of c yield similar simulation results. As in [23], a sample size of one million is used for each estimate, and the slowdown threshold $\theta = 0.8$.

The theoretical values were obtained by finding that for the related question of what is the probability that either queue 1 has n customers or queue 2 has m customers before emptying. The theoretical value for the finite state space problem can be found by a first step analysis, and then solving the resulting numerical system. This problem is then solved for increasing values of m, until the overflow probability given is constant for at least 4 significant figures.

Table 4.1: Probability of Buffer 2 Overflow, First Example

	$n = 20$	$n = 50$	$n = 100$
Theoretical value	3.80×10^{-7}	1.27×10^{-16}	3.55×10^{-32}
Estimate	3.82×10^{-7}	1.27×10^{-16}	3.53×10^{-32}
Std. Err.	0.01×10^{-7}	0.01×10^{-16}	0.07×10^{-32}
95% C.I.	$[3.80, 3.84] \times 10^{-7}$	$[1.25, 1.29] \times 10^{-16}$	$[3.36, 3.67] \times 10^{-32}$

Parameter values: $(\lambda, \mu_1, \mu_2, \nu_1) = (0.1, 0.7, 0.2, 0.15)$, $c = 0.03$. Sample size 1 million.

Table 4.2: Probability of Buffer 2 Overflow, Second Example

	$n = 20$	$n = 50$	$n = 100$
Theoretical value	5.63×10^{-2}	1.19×10^{-3}	1.63×10^{-6}
Estimate	5.62×10^{-2}	1.18×10^{-3}	1.61×10^{-6}
Std. Err.	0.03×10^{-2}	0.01×10^{-3}	0.02×10^{-6}
95% C.I.	$[5.56, 5.68] \times 10^{-2}$	$[1.16, 1.20] \times 10^{-3}$	$[1.57, 1.65] \times 10^{-6}$

Table 2. $(\lambda, \mu_1, \mu_2, \nu_1) = (0.3, 0.36, 0.34, 0.32)$, $c = 0.004$. Sample size 1 million.

4.9 Proof of Lemma 4.6.1

We first show the existence and uniqueness of z. Define $h : [1, \infty) \to \mathbb{R}$ by

$$h(y) \doteq 2\sqrt{\lambda \nu_1 y} + \frac{\mu_2}{y} - (\lambda + \nu_1 + \mu_2).$$

Let $y^* \doteq [\mu_2^2/(\lambda\nu_1)]^{1/3}$. Note that $y^* > 1$ thanks to (4.2.1). Since

$$h'(y) = \frac{1}{y^2}\left[y\sqrt{\lambda\nu_1 y} - \mu_2\right],$$

it follows that $h'(y) < 0$ for $1 \leq y < y^*$ and $h'(y) > 0$ for $y > y^*$. Furthermore,

$$h(1) = 2\sqrt{\lambda\nu_1} - \lambda - \nu_1 = -(\sqrt{\lambda} - \sqrt{\nu_1})^2 < 0, \quad h(\infty) = \infty.$$

Therefore there exists a unique $z > 1$ such that $h(z) = 0$. Indeed, we must have

$$z > y^* = \left[\mu_2^2/(\lambda\nu_1)\right]^{1/3} \tag{4.9.1}$$

and that $h(y) < 0$ if and only if $1 \leq y < z$. A discussion of the above material can also be found in [20]. We now consider equation (4.6.2). Since $\mathbb{H}_s(\alpha) = -2H_s(-\alpha/2)$,

$$\mathbb{H}_s(\alpha) = -2\left[\lambda(e^{-\alpha_1/2} - 1) + \nu_1(e^{(\alpha_1-\alpha_2)/2} - 1) + \mu_2(e^{\alpha_2/2} - 1)\right],$$

$$\frac{\partial \mathbb{H}_s}{\partial \alpha_1}(\alpha) = \lambda e^{-\alpha_1/2} - \nu_1 e^{(\alpha_1-\alpha_2)/2}.$$

Simple algebra yields that equation (4.6.2) amounts to

$$h(e^{-\alpha_2/2}) = 0, \quad \alpha_1 = -\log(\nu_1 e^{-\alpha_2/2}/\lambda).$$

Equation (4.6.4) follows immediately.

Finally, $\mathbb{H}_s(\alpha^{[1]}) = 0$ is just simple calculation, and thus omitted. It remains to show $\alpha_2^{[0]} < \alpha_1^{[1]} < \alpha_1^{[0]}$, from which $\langle \alpha^{[0]}, d_1 \rangle \geq 0$ follows immediately since $d_1 = (1, -1)$. Plugging the formulae for $\alpha^{[0]}$ and $\alpha^{[1]}$, the inequalities reduce to

$$\frac{\mu_2}{\lambda} < z < \frac{\mu_2^2}{\lambda \nu_1}.$$

But $h(\mu_2/\lambda) = -(\sqrt{\mu_2} - \sqrt{\nu_1})^2 < 0$ and

$$h\left(\frac{\mu_2^2}{\lambda \nu_1}\right) = \frac{1}{\mu_2}(\mu_2 - \nu_1)(\mu_2 - \lambda) > 0.$$

This completes the proof. ∎

4.10 Exponential bounds on T_N

We first define the stopping time

$$N \doteq \inf \{k \geq 1 : Q_2(T_k) = n \text{ or } Q(T_k) = (0,0)\}. \tag{4.10.1}$$

In the proof of Theorem 4.7.1 one needs to establish an exponential bound on the exit time T_N, namely, there exists a strictly positive constance c such that

$$\limsup_n \frac{1}{n} \log E_0[\exp\{cT_N\}] < \infty. \tag{4.10.2}$$

The proof in [13, Proposition A.1], which deals with a similar problem for the tandem queue networks, is not applicable here since it assumes that the domain is compact. However, this compactness assumption is not necessary and the goal of this appendix is to show a slightly stronger result, namely, there exists a strictly positive constant c such that

$$E_0[\exp\{cT_{N_0}\}] < \infty, \tag{4.10.3}$$

where T_{N_0} is the first time the process returns to origin, that is,

$$N_0 \doteq \inf\{k \geq 1 : Q(T_k) = 0\}.$$

Note that N_0 is independent of n and $T_N \leq T_{N_0}$. Therefore the exponential bound (4.10.2) is clearly implied by (4.10.3).

In order to show (4.10.3), we first observe the following. Consider a standard tandem queue network with no server slowdown and with arrival rate λ and consecutive service rates ν_1 and μ_2. For this network, similarly to T_{N_0}, define τ_0 the first time the state process return to origin. It is not difficult to show, by a pathwise argument, that τ_0 *stochastically dominates* T_{N_0}, that is,

$$P(\tau_0 \geq t) \geq P(T_{N_0} \geq t) \tag{4.10.4}$$

for all $t \geq 0$. Indeed, consider a Poisson arrival process with arrival rate λ, and suppose that the i-th arrival is associated with a random vector $U_i \doteq (U_i^{(1)}, U_i^{(2)}, W_i^{(1)})$. The vectors $\{U_i\}$ are iid, and $U_i^{(1)}$, $U_i^{(2)}$, and $W_i^{(1)}$ are all exponentially distributed with respective rates ν_1, μ_2, and μ_1. Furthermore, $U_i^{(1)} \geq W_i^{(1)}$ and $U_i^{(2)}$ is independent of $(U_i^{(1)}, W_i^{(1)})$. Such vectors always exist since exponential distribution with rate ν_1 stochastically dominates exponential distribution with rate μ_1. Consider the following two scenarios: (1) The i-th arrival always uses $U_i^{(1)}$ as the service time at node 1; (2) The i-the arrival uses $U_i^{(1)}$ as the service time at node 1 if the second queue is larger than θn and uses $W_i^{(1)}$ otherwise. In both cases, the service time for the i-th arrival at node 2 is assumed to be $U_i^{(2)}$. Clearly, the first scenario yields the standard tandem network with arrival λ and service rates ν_1 and μ_2, while the second scenario yields the tandem queue with server slowdown as studied in this paper. Since $U_i^{(1)} \geq W_i^{(1)}$ for every i, it is clear that the return time to the origin for the first scenario is pathwise bounded from below by that of the second scenario. This implies the stochastic dominance (4.10.4).

Thanks to the stochastic dominance (4.10.4), it suffices to show that there exists a strictly positive constant c such that $E_0[\exp\{c\tau_0\}] < \infty$. To this end, we consider the discrete time embedded Markov chain of the standard tandem network and (abusing notation) let $\{Z(k) \in \mathbb{Z}_+^2, k = 0, 1, \ldots\}$ denote the queue lengths at the transition epochs of the network. We also, without loss of generality, assume $\lambda + \nu_1 + \mu_2 = 1$. Let $\{Y(k)\}$ be an iid sequence of random vectors taking values in $\mathbb{V}$ with

$$P(Y(k) = v_1) = \lambda, \quad P(Y(k) = v_2) = \nu_1, \quad P(Y(k) = v_3) = \mu_2.$$

Then the dynamics of Z can be represented by the evolution

$$Z(k+1) = Z(k) + \pi[Z(k), Y(k+1)].$$

Define

$$\sigma_0 \doteq \inf\{k \geq 1 : Z(k) = 0\}.$$

Note that τ_0 can be written as

$$\tau_0 = \sum_{i=1}^{\sigma_0} s_i$$

where $\{s_i\}$ denote the time elapse between transition epochs in the continuous time model, and $\{s_i\}$ are iid exponential random variables with rate $\lambda+\nu_1+\mu_2$, independent of σ_0. Therefore, it suffices to show that there exists a strictly positive constant c such that

$$E_0[\exp\{c\sigma_0\}] < \infty.$$

Indeed, we have a stronger result [22].

Lemma 4.10.1. *There exist a constant $c > 0$ such that $E_z[\exp\{c\sigma_0\}]$ is finite for all $z \in \mathbb{Z}_+^2$.*

Proof. Define the stopping time $\hat{\sigma}_0 \doteq \inf\{k \geq 0 : Z(k) = 0\}$. Note that $\sigma_0 = \hat{\sigma}_0$ as long as $Z(0) \neq 0$ and $\hat{\sigma}_0 = 0$ if $Z(0) = 0$. Define $h(z) \doteq E_z[\hat{\sigma}_0]$ for each $z \in \mathbb{Z}_+^d$. Since the network is positive recurrent by assumption $\lambda < \nu_1 \leq \mu_2$, $h(z)$ is finite for every z. Define the process

$$S(k) \doteq \begin{cases} h(Z(k)) & \text{if } k \leq \hat{\sigma}_0, \\ \hat{\sigma}_0 - k & \text{if } k > \hat{\sigma}_0. \end{cases}$$

There are two key properties of S [13, Lemmas A.2, A.4, and A.5].

(i) Let $\{\mathcal{F}_k = \sigma(Z(0), Y(1), \ldots, Y(k))\}$ be the filtration. Then

$$E_z[S(k+1) - S(k) | \mathcal{F}_k] = -1$$

for all $z \in \mathbb{Z}_+^2$ and all $k \geq 0$.

(ii) The increments of the process S are *uniformly* bounded, say by $M < \infty$.

Note that there exists a $\varepsilon_0 > 0$ such that, for all $|x| \leq \varepsilon_0$,

$$e^x \leq 1 + x + x^2.$$

Define $\bar{\varepsilon} \doteq \min\{\varepsilon_0/M, 1/(2M^2)\}$. Then for every $k \geq 0$, thanks to property (ii) and that $\bar{\varepsilon}M \leq \varepsilon_0$,

$$\begin{aligned} e^{\bar{\varepsilon}(S(k+1)-S(k))} &\leq 1 + \bar{\varepsilon}(S(k+1) - S(k)) + \bar{\varepsilon}^2(S(k+1) - S(k))^2 \\ &\leq 1 + \bar{\varepsilon}(S(k+1) - S(k)) + \bar{\varepsilon}^2 M^2. \end{aligned}$$

But by property (i) and that $\bar{\varepsilon}^2 M^2 \leq \bar{\varepsilon}/2$, it follows that

$$E\left[e^{\bar{\varepsilon}(S(k+1)-S(k))} \middle| \mathcal{F}_k \right] \leq 1 - \bar{\varepsilon}/2.$$

This is equivalent to that the process $\{(1 - \bar{\varepsilon}/2)^{-k} e^{\bar{\varepsilon}S(k)}, \mathcal{F}_k\}$ is a supermartingale.

In particular, given $Z(0) = z \neq 0$, the Optional Sampling Theorem, $\hat{\sigma}_0 = \sigma_0$, and $h(0) = 0$ imply that

$$e^{\bar{\varepsilon}h(z)} = e^{\bar{\varepsilon}S(0)} \geq E_z\left[(1 - \bar{\varepsilon}/2)^{-\sigma_0}\right].$$

Letting $c \doteq -\log(1 - \bar{\varepsilon}/2) > 0$, we have

$$E_z\left[e^{c\sigma_0}\right] < \infty$$

for all $z \in \mathbb{Z}_+^2$ and $z \neq 0$. For $Z(0) = 0$, we only need to note that since the first jump away from 0 must be to e_1, there is $C < \infty$ such that

$$E_0\left[e^{c\sigma_0}\right] = CE_{e_1}\left[e^{c\sigma_0}\right] < \infty.$$

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

[3] D. Bertsimas and I. Paschalidis. Asymptotic buffer overflow probabilities in multiclass multiplexers: An optimal control approach. *IEEE Transactions on Automatic Control*, 43:315–335, 1998.

[4] P. Dupuis and R. S. Ellis. The large deviation principle for a general class of queueing systems, I. *Trans. Amer. Math. Soc.*, 347:2689–2751, 1996.

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

[6] P. Dupuis and R.S. Ellis. Large deviations for Markov processes with discontinuous statistics, II: Random walks. *Probab. Th. Rel. Fields*, 91:153–194, 1992.

[7] 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.

[8] P. Dupuis and H. Ishii. On Lipschitz continuity of the solution mapping to the Skorokhod problem, with applications. *Stochastics*, 35:31–62, 1991.

[9] P. Dupuis, K. Leder, and H. Wang. Importance sampling for the weighted serve the longest queueing policy.

[10] P. Dupuis, K. Leder, and H. Wang. Importance sampling for sums of random variables with regularly varying tails. *ACM Trans. Modeling Comp. Simulation*, 17, 2007.

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

[12] P. Dupuis, K. Leder, and H. Wang. On the large deviations of the weighted-serve-the-longest-queue policy. *Progress in Probability*, 2008.

[13] P. Dupuis, A. Sezer, and H. Wang. Dynamic importance sampling for queueing networks. *Ann. Appl. Prob.*, pages 1306–1346, 2007.

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

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

[16] W. H. Fleming and R. Rishel. *Deterministic and Stochastic Optimal Control*. Springer, Berlin, 1975.

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

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

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

[20] D. Kroese, W. Scheinhardt, and P. Taylor. Spectral properties of the tandem jackson network, seen as quasi-birth-and-death process. *The Annals of Applied Probability*, 14:2057–2089, 2004.

[21] K. Majewski. Large deviations of the steady-state distribution of reflected processes with applications to queueing systems. *Queueing Systems*, 29:351–381, 1998.

[22] S.P. Meyn and D. Down. Stability of generalized jackson networks. *Annals of Applied Probability*, 4:124–148, 1994.

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

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

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

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

[27] N.D. van Foreest, M.R.H. Mandjes, J.C.W. van Ommeren, and W.R.W. Scheinhardt. A tandem queue with server slow-down and blocking. *Stochastic Models*, 21:695–724, 2005.

[28] L. Ying, R. Srikant, A. Eryilmaz, and G.E. Dullerud. A large deviation analysis of scheduling in wireless networks. *Preprint*, 2005.