September 1996 UILU-ENG-96-2223

University of Illinois at Urbana-Champaign

Analysis of Some Simple Policies for Dynamic Resource Allocation

Murat Alanyali

Coordinated Science Laboratory
1308 West Main Street, Urbana, IL 61801

REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 1996	Technical

4. TITLE AND SUBTITLE
Analysis of Some Simple Policies for Dynamic Resource Allocation

5. FUNDING NUMBERS

6. AUTHOR(S)
Alanyali, Murat

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(ES)
Coordinated Science Lab
University of Illinois
1308 W. Main St.
Urbana, IL 61801

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Science Foundation
4201 Wilson Blvd.
Arlington, VA 22230

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Complexity-performance trade-offs are investigated for dynamic resource allocation in load sharing networks with Erlang-type statistics. The emphasis is on the performance of simple allocation strategies that can be implemented on-line. The resource allocation problem is formulated as a stochastic optimal control problem. Variants of a simple least load routing policy are shown to lead to a fluid type limit and to be asymptotically optimal. Either finite capacity constraints or migration of load can be incorporated into the setup.

Three policies, namely optimal repacking, least load routing, and Bernoulli splitting, are examined in more detail. Large deviations principles are established for the three policies in a simple network of three consumer types and two resource locations and are used to identify the network overflow exponents. The overflow exponents for networks with arbitrary topologies are identified for optimal repacking and Bernoulli splitting policies, and conjectured for the least load routing policy.

Finally, a process-level large deviations principle is established for Markov processes in the Euclidean space with a discontinuity in the transition mechanism along a hyperplane.

14. SUBJECT TERMS	15. NUMBER IF PAGES
Dynamic resource allocation, load balancing, fluid equations, loss networks, least load routing, large deviations, Markov processes, discontinuous statistics	108
	16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

ANALYSIS OF SOME SIMPLE POLICIES FOR DYNAMIC RESOURCE ALLOCATION

BY

MURAT ALANYALI

B.S., Middle East Technical University, 1988
M.S., Bilkent University, 1990

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 1996

Urbana, Illinois

ANALYSIS OF SOME SIMPLE POLICIES FOR DYNAMIC RESOURCE ALLOCATION

Murat Alanyali, Ph.D.
Department of Electrical Engineering
University of Illinois at Urbana-Champaign, 1996
Bruce Hajek, Advisor

Complexity-performance trade-offs are investigated for dynamic resource allocation in load sharing networks with Erlang-type statistics. The emphasis is on the performance of simple allocation strategies that can be implemented on-line. The resource allocation problem is formulated as a stochastic optimal control problem. Variants of a simple least load routing policy are shown to lead to a fluid type limit and to be asymptotically optimal. Either finite capacity constraints or migration of load can be incorporated into the setup.

Three policies, namely optimal repacking, least load routing, and Bernoulli splitting, are examined in more detail. Large deviations principles are established for the three policies in a simple network of three consumer types and two resource locations and are used to identify the network overflow exponents. The overflow exponents for networks with arbitrary topologies are identified for optimal repacking and Bernoulli splitting policies, and conjectured for the least load routing policy.

Finally, a process-level large deviations principle is established for Markov processes in the Euclidean space with a discontinuity in the transition mechanism along a hyperplane. The transition mechanism of the process is assumed to be continuous on one closed half-space and also continuous on the complementary open half-space. A similar result was recently obtained by Dupuis and Ellis for lattice-valued Markov processes satisfying a mild communication/controllability condition. The proof presented here relies on the work of Blinovskii and Dobrushin, which in turn is based on an earlier work of Dupuis and Ellis.

DEDICATION

To the memory of Professor Bülent Kerim Altay.

ACKNOWLEDGMENTS

I would like to express my gratitude to Professor Bruce Hajek for his invaluable guidance and enthusiasm, which were essential for the realization of this thesis. I would also like to thank E. Arıkan, T. Başar, D. Burkholder, U. Einmahl, P.R. Kumar, U. Madhow, and K. Zeger for their teachings and suggestions and the Scientific and Technical Research Council of Turkey for funding the early stages of my studies. Finally, I am grateful to my officemates and my housemates for their comradeship and to my family for providing constant support and encouragement over the years.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Economic pressures and reliability considerations generally lead to communication networks with load sharing capabilities and highlight resource allocation as a fundamental issue in network design. The objective of resource allocation in such systems is oftentimes consumer satisfaction, which may translate to minimizing consumer blocking or achieving fairness by load balancing. Regardless of its objective, an essential aspect of a resource allocation policy is implementability: Practical considerations require allocation policies to have low complexity, require little information about the network state, and be robust to changes in the traffic parameters. This thesis concerns trade-offs implied by these requirements.

Two instances of resource allocation arise in transmission scheduling in wireless networks and dynamic routing in telephone switching networks. A wireless network consists of a number of base stations and users (as in Figure 1.1(a)). The users require communication channels that are available at the base stations, whereas each base station can serve the users within its geographical range. The resource allocation problem in this setting concerns the question of station selection. Similarly, in a telephone switching network, users demand channels that are available at the network links. In networks such as the multiparented network (see [1]) of Figure 1.1(b), there are multiple paths between pairs of users, and the resource allocation problem concerns path selection.

In this thesis the mathematical abstraction of a load sharing network is a triple (U, V, N). Here U is a finite set of *consumer types*, V is a finite set of *locations*, and $(N(u) \subset V : u \in U)$ is a set of *neighborhoods* (see Figure 1.2 for examples). A *demand* for this network is a vector $(\lambda(u) : u \in U)$ of positive numbers. In a dynamic setting $\lambda(u)$ denotes the arrival rate of *type u consumers*. Each consumer is served, starting immediately upon its arrival, for the duration of its *holding time*. The neighborhood $N(u)$ denotes the locations that are available to type u consumers, in the sense that each such consumer can be served only at a location within $N(u)$. Note that the abstraction of the wireless network of Figure 1.1(a) is the W network of Figure 1.2(b). An *allocation policy* is an

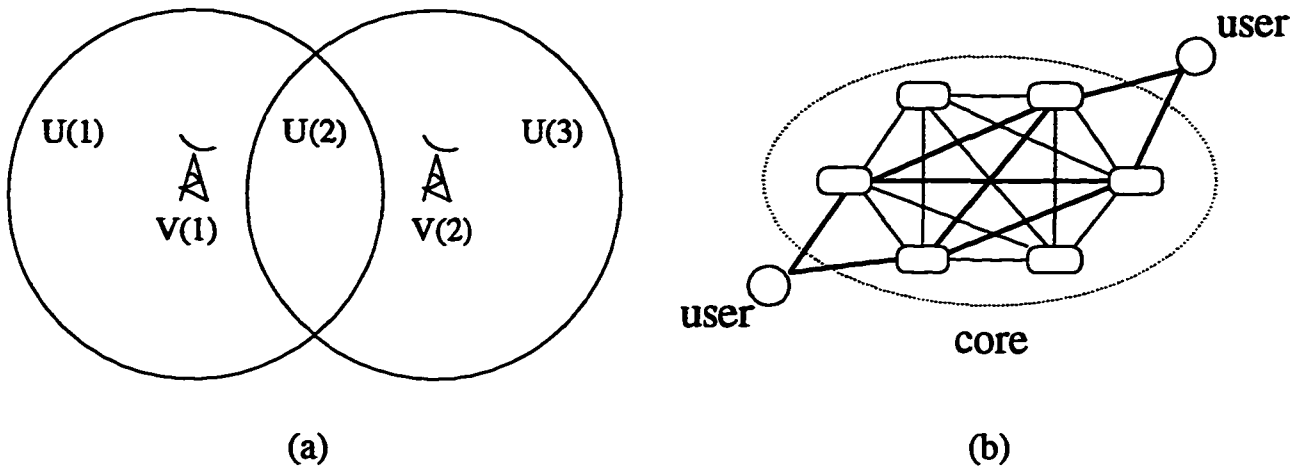


Figure 1.1 Two instances of resource allocation: (a) a typical wireless network with overlapping neighborhoods, and (b) a multiparented circuit switching network.

algorithm that assigns consumers to locations within their respective neighborhoods. The *load* at a location is the number of consumers at the location.

Load balancing is a possible guiding principle for resource allocation, whereby the load is allocated across locations as evenly as possible. There is a rich literature on load balancing, and both static and dynamic versions of the problem have been studied by numerous authors (e.g., [2], [3], [4], [5], [6], and references therein). It is well-known that load balancing can be an effective allocation strategy when the cost is convex (or the reward concave) as a function of the allocated loads. For example, $x(1)^2 + x(2)^2$ is minimized over probability vectors $(x(1), x(2))$ by $x(1) = x(2) = 1/2$. This is connected with the convexity of the function $f(x) = x^2$.

A brute force approach for dynamic load balancing is the *optimal repacking* (OR) policy under which consumers are continuously repacked to balance the load *at all times*. Nevertheless, in many applications, repacking of consumers is not acceptable due to operational reasons. Besides, computational complexity of this policy may be too high for large networks. As an alternative strategy, one can consider the nonrepacking *Bernoulli splitting* (BS) policy under which consumers are assigned to locations randomly, so as to balance the *mean* load in the network. BS is a simple policy; however, it exerts only open-loop control, and it is not robust with respect to the network demand. Another reasonable allocation strategy is the popular *least load routing* (LLR) policy whereby each arriving consumer is assigned to a location with the least load in the associated neighborhood. The following chapters concentrate on the comparison of the OR, BS, and LLR

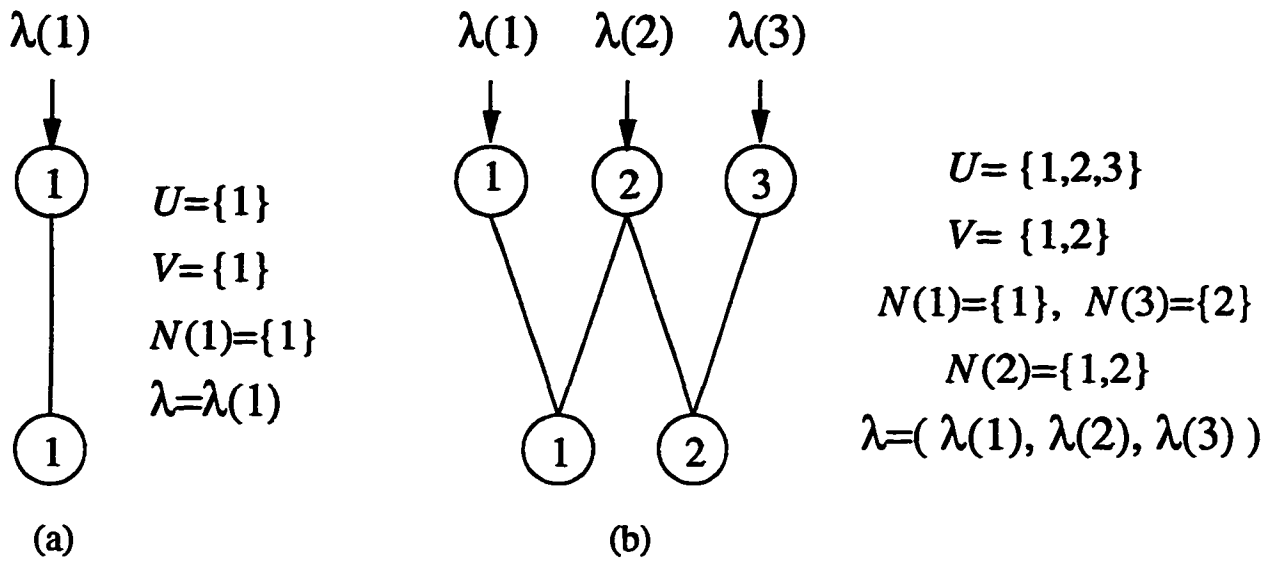


Figure 1.2 Two load sharing networks: (a) the single-location network, and (b) the W network.

policies based on certain estimates on their performances. The emphasis is on the analysis of the LLR policy.

Chapter 2 concerns the properties of the LLR policy implied by *fluid limit approximations*, which are deterministic weak limits of the network load, suitably normalized, for large values of demand. Under a stochastic model that incorporates mobility of consumers, but not finite capacity constraints on locations, the dynamic resource allocation problem is formulated as an optimal control problem with a long-term average convex cost. The fluid limit approximations of the network load under the LLR policy are obtained and used to establish the asymptotic optimality of LLR, in the sense of minimizing a suitably normalized version of the cost, for large values of network demand. Under a complementary model that incorporates capacity constraints on resources but not mobility of consumers, the problem is formulated as the minimization of blocking probability. Fluid limit approximations are obtained, and certain variants of the LLR policy are shown to be asymptotically optimal for large values of network demand. In the presence of both finite capacity constraints and mobility, it is shown that simple policies are not necessarily optimal, even in the asymptotic sense.

The fluid limit approximation provides a description of the *typical* behavior of the network load. Although this description proves to be fruitful, there are reasonable questions that demand a finer analysis. In particular, (1) the three policies, OR, BS, and LLR, have the same performance in the fluid scale, and (2) certain events of interest, such as network overflow, correspond to large

deviations of the network from its typical behavior, and hence are not described accurately by the fluid limit approximation. Although network overflow is a rare event, it has a strong impact on network performance; it is therefore desirable to estimate its probability and mode accurately. In Chapter 3 we employ large deviations theory to study network overflow in terms of *overflow exponents*, which provide estimates on the probability of overflow within a fixed, long time interval. The three policies are then ordered based on their overflow exponents.

Chapter 3 establishes an explicit large deviations principle (LDP) for the load process in the W network of Figure 1.2(b) under each policy of interest. Mobility of consumers is not incorporated. The LDP under the LLR policy entails a treatment of large deviations of Markov processes with discontinuous transition mechanisms, which is the subject of Chapter 4. Based on the obtained LDPs, the overflow exponents are identified for the three policies, and it is shown that the LLR policy performs as well as the OR policy for small enough capacities, whereas it performs significantly better than the BS policy for the whole range of capacities. The methods used in the analysis of the W network generalize easily to identify the overflow exponents for arbitrary networks under the OR and BS policies. This generalization appears difficult under the LLR policy, for which we conjecture the general form of the overflow exponents.

CHAPTER 2

FLUID SCALE ANALYSIS OF ALLOCATION POLICIES

2.1 Introduction

This chapter concentrates on load sharing networks in the dynamic setting and provides deterministic descriptions for the network behavior under certain allocation policies. These descriptions are then used to estimate the performance of the policies of interest, and certain simple policies are shown to have optimality properties.

To observe a typical behavior of the network load under the LLR policy, consider the load sharing network of Figure 1.2(b). Suppose that the network demand is $\lambda = (\gamma, \gamma, \gamma)$ so that the arrivals of consumers of each type form a Poisson process of rate γ. Each consumer remains in the network for an exponentially distributed amount of time, with unit mean. Finally, suppose that initially location 1 has zero load, whereas location 2 has load 3γ. Figures 2.1(a)-2.1(c) depict typical sample paths of the normalized load, defined as the load divided by γ, at the two locations for $\gamma = 1, 10, 100$, for the time interval $[0, 8]$. In the limit as γ goes to infinity, the normalized load converges to the deterministic trajectory depicted in Figure 2.1(d). Deterministic descriptions of this sort are commonly referred to as fluid limit approximations.

This chapter focuses on the optimality properties of the LLR allocation policy (and variants) implied by the corresponding fluid limits. The main results of the chapter are that for large values of network demand, (1) the LLR policy is asymptotically optimal in the sense of minimizing a long-term average quadratic cost (Theorem 2.3.1, and for a model including migration, Theorem 2.5.1), and (2) variants of the same policy achieve the minimum blocking probability in the case of locations with finite capacities (Theorem 2.4.1). The strategies of interest do not display the pathologies found in some finite capacity networks (as in [7]) so that optimality properties can be obtained via fluid limit analysis.

The outline of the rest of the chapter is as follows. Section 2.2 gives some preliminary results regarding *static* load balancing. Section 2.3 defines the basic dynamic model in which consumers

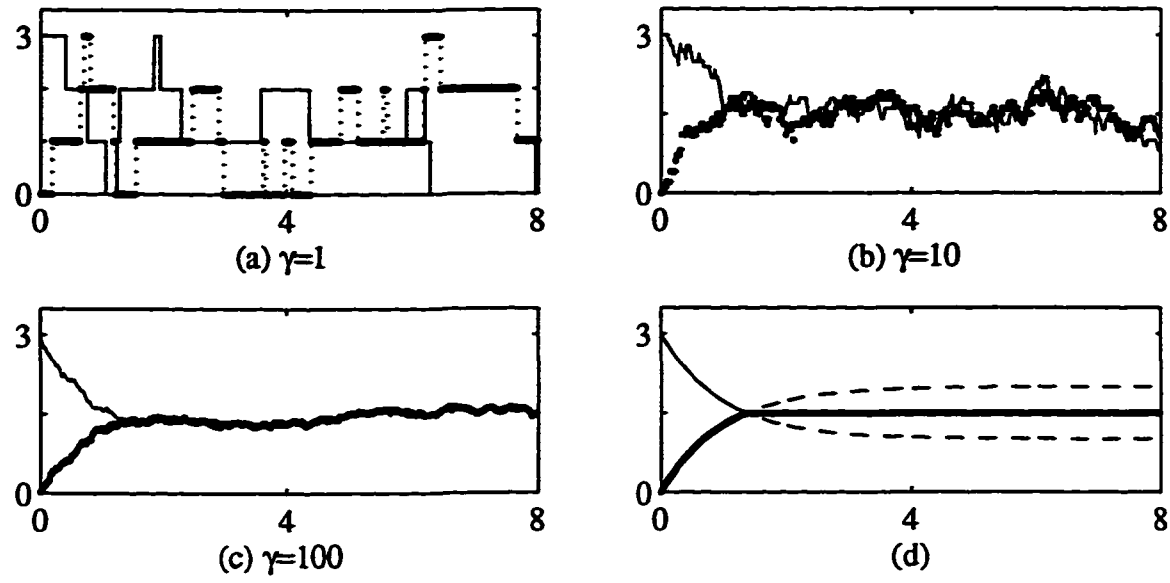


Figure 2.1 Load under the least load routing policy.

remain stationary in the network until departure, and locations have infinite capacities. The dynamic resource allocation problem is formulated as a stochastic optimal control problem with a long-term average quadratic cost, which is to be minimized within a set of practical controls. The results for the static load balancing policy are used to provide a lower bound on the performance of arbitrary controls. Section 2.3.1 describes the LLR policy. Sections 2.3.2 and 2.3.3 identify the fluid limit approximations of the network load under LLR as the solutions to certain integral equations with boundary constraints. This solution converges to an optimal point in equilibrium, and Section 2.3.4 exploits this fact to establish the asymptotic optimality of LLR. Section 2.4 considers the case in which locations have finite capacities, and the resource allocation problem is defined as the minimization of blocking probability. It is shown that a class of *least relative load routing* (LRLR) policies asymptotically achieve the smallest blocking probability for large arrival rates. A connection with trunk reservation policies is discussed. Section 2.5, which can be read independently of Section 2.4, generalizes the results of Section 2.3.4 by considering an infinite capacity network in which consumers can migrate. A summary of conclusions and final remarks are collected in Section 2.6.

2.2 Preliminaries: The Static Load Balancing Problem

This section concerns the *static* load balancing problem, which plays an important role in the discussion of the dynamic load balancing problem of Section 2.3. Thus, as a precursor to the arguments therein, we define the static load balancing problem and provide three lemmas that characterize its solutions.

Given a load sharing network (U, V, N), we say that an *assignment* a, given by $(a_{u,v} : u \in U, v \in V)$, is *admissible* if $a \geq 0$ and $a_{u,v} = 0$ whenever $v \notin N(u)$. Given a demand vector λ, an admissible assignment a *satisfies* demand λ if $\sum_v a_{u,v} = \lambda(u)$ for all $u \in U$. The *load* at location $v \in V$ corresponding to assignment a is given by $q(v) = \sum_u a_{u,v}$, and $q = (q(v) : v \in V)$ is called the *load vector.*

Let $\mathcal{A}_\lambda$ be the set of admissible assignments that satisfy demand λ. Let $\Phi : R^V \to R$ be a strictly convex, differentiable function which is symmetric in its arguments. The *static load balancing problem* (SLB) is defined as

$$SLB(\lambda, \Phi) \ : \ Minimize(\Phi(q) \ : \ a \in \mathcal{A}_\lambda).$$

The proofs of the following three lemmas can be found in Section 2.7.

Lemma 2.2.1 *There exists a solution to $SLB(\lambda, \Phi)$. An assignment $a \in \mathcal{A}_\lambda$ is a solution if and only if for all $u \in U$, and all $v \in N(u)$*

$$a_{u,v} = 0 \quad \textit{whenever} \quad q(v) > m_u(q), \tag{2.1}$$

where $m_u(q) = \min_{v \in N(u)} q(v)$. Furthermore, all such assignments yield the same load vector.

Lemma 2.2.2 *There exists a unique partition $\{V_1, V_2, \cdots, V_J\}$ of V and a unique partition $\{U_1, U_2, \cdots, U_J\}$ of U such that for any assignment a satisfying condition (2.1), and the corresponding load vector q,*

$$q(v) = q(v') \qquad v, v' \in V_i \qquad i = 1, 2, \cdots, J \tag{2.2}$$

$$q(v) < q(v') \quad v \in V_i, \ \ v' \in V_j \quad i < j \tag{2.3}$$

$$a_{u,v} = 0 \quad v \in V_i, \ \ u \in U_j \quad i > j \tag{2.4}$$

$$N(u) \cap V_i = \emptyset \qquad u \in U_j \qquad i < j. \tag{2.5}$$

Let the function $\Psi : R_+^U \to R$ denote the value of SLB as a function of the demand vector, i.e., $\Psi(\lambda) = \Phi(q)$, where q is the unique load vector corresponding to the solutions of $SLB(\lambda, \Phi)$. The following lemma holds:

Lemma 2.2.3 *The function Ψ is convex.*

2.3 The Basic Model

Given a load sharing network (U, V, N), a demand vector λ, and a positive number γ, consider the following stochastic description of the network dynamics: For each $u \in U$ consumers of type u arrive according to a Poisson process of rate $\gamma\lambda(u)$, the processes for different types of arrivals being independent. In this section we assume that each location has infinite capacity; therefore, the network can accommodate every consumer immediately. This assumption is relaxed in Section 2.4, in which finite capacities are imposed on the locations. Each consumer has a holding time that is exponentially distributed with unit mean, independent of the past history. In the basic model it is also assumed that consumers do not change their types until they depart from the system. This assumption is relaxed in Section 2.5, which introduces a model such that consumers can migrate in the sense that their types change.

Let $X_t(v)$ denote the load at location $v \in V$ at time t, and set $X_t = (X_t(v) : v \in V)$. The consumer arrival and departure times, together with the allocation policy and an initial condition, determine the load process $X = (X_t : t \geq 0)$. The performance measure for an allocation policy π is the long-term average cost J_γ^π, defined by

$$J_\gamma^\pi = \liminf_{T \to \infty} E\left[\frac{1}{T}\int_0^T \Phi(X_t)dt \,|X_0 = x_0\right],$$

where $\Phi(x) = \sum_v x^2(v)$ for $x \in R^V$. The *dynamic load balancing problem* is to determine the set of allocation policies that minimize J_γ^π.

By Lemma 2.2.1, a wide class of convex functions would be appropriate for the instantaneous cost in order to focus on load balancing, in the sense that certain optimality properties remain true for any such cost function. In this chapter we concentrate on the quadratic instantaneous cost for convenience.

Let $L_t(u)$ denote the number of type $u \in U$ consumers in the network at time t, and set $L_t = (L_t(u) : u \in U)$. Note that consumer arrivals and departures, and hence the process $L = (L_t : t \geq 0)$, are not affected by the assignment decisions. Therefore, the value of problem $SLB(L_t, \Phi)$ yields a lower bound on the instantaneous cost at time t under *any* allocation policy. The process $(L_t(u) : t \geq 0)$ for fixed u is an $M/M/\infty$ queue length process with load factor $\gamma\lambda(u)$; hence the equilibrium distribution of L is described by a vector $(L_\infty(u) : u \in U)$ of Poisson random variables with mean vector $\gamma\lambda$. This implies the following lower bound on the cost of general allocation policies:

$$\begin{aligned} J_\gamma^\pi &\geq \liminf_{T\to\infty} E\left[\frac{1}{T}\int_0^T \Psi(L_t)dt \,|X_0 = x_0\right] \\ &= E[\Psi(L_\infty)] \\ &\geq \Psi(\gamma\lambda). \end{aligned} \tag{2.6}$$

Here, the last inequality follows by Lemma 2.2.3 and Jensen's inequality.

Let $\Psi_I(L_t)$ denote the value of the problem $SLB(L_t, \Phi)$ under the additional constraint that the assignment a have integer coordinates. If repacking is allowed, then the optimal policy is clearly *optimal repacking* (OR), which continuously rearranges the consumers in the network so as to maintain $\Phi(X_t) = \Psi_I(L_t)$ at all times t and achieves $J_\gamma^{OR} = E[\Psi_I(L_\infty)]$. The OR policy can be implemented at a cost of $O(|V||U| + |V|^2)$ computations per consumer arrival and consumer departure, which may be impractical for large networks. Furthermore, frequent repacking of consumers may not be feasible due to operational constraints. These considerations lead to the study of nonrepacking policies that are much simpler in terms of computational complexity and information required about the network state.

2.3.1 Least load routing

This section describes the particular nonrepacking type allocation policy considered in the context of the basic model, namely the least load routing (LLR) policy. LLR is defined with the following assignment rule:

- When a type u consumer arrives, it is assigned to a location $v \in N(u)$ with the minimum load. If multiple locations achieve the minimum in $N(u)$, the consumer is assigned at random to one such location, each location having equal probability.

The LLR policy is a nonrepacking policy and costs $|N(u)|$ comparisons per consumer arrival of type $u \in U$. Another desirable feature of LLR is that it can be implemented in a distributed manner by using one independent assignment agent per consumer type. Each arrival can be assigned to a location based on partial information about the network state. Furthermore, LLR is robust with respect to the network demand. On the other hand, LLR is a myopic allocation policy and is not necessarily optimal for finite arrival rates. The LLR policy has been studied by a number of authors and has been shown to have a poor worst-case performance relative to the optimal nonrepacking policy (see [8]).

Under LLR, the load process X is Markov on the state space Z_+^V. For $v \in V$, define the operator $T_v : Z_+^V \to Z_+^V$ as

$$(T_v x)(v') = \begin{cases} x(v') + 1 & \text{if } v' = v, \\ x(v') & \text{else.} \end{cases}$$

Then the off-diagonal entries of the generator matrix of X are given by

$$Q(x,y) = \begin{cases} \sum_{u \in N^{-1}(v)} \gamma\lambda(u) \frac{I\{x(v)=m_u(x)\}}{\sum_{v' \in N(u)} I\{x(v')=m_u(x)\}} & \text{if } y = T_v x \\ x(v) & \text{if } y = T_v^{-1} x \\ 0 & \text{else,} \end{cases} \tag{2.7}$$

where $N^{-1}(v) = \{u \in U : v \in N(u)\}$.

In principle, given a load sharing network, one can compute the equilibrium distribution of X and thereby the cost incurred under the LLR policy. However, it is computationally intractable to obtain an expression for the cost of LLR for arbitrary networks through an expression for the equilibrium distribution. As an alternative approach, we study the network for large values of the parameter γ and, by obtaining fluid limit approximations, evaluate the performance of the LLR policy for arbitrary network topologies.

2.3.2 Convergence

This section addresses the weak convergence of the network load as γ tends to infinity. The main result, Lemma 2.3.3, characterizes the possible weak limits of the load process, properly normalized, via a semimartingale representation.

Let the *normalized load process* X^γ be defined as $X^\gamma = \gamma^{-1}X$, where X denotes the network load under the LLR policy. Assume the existence of a finite number K such that $E[\sum_v X_0^\gamma(v)] \leq K$

for all γ. Define

$$a_{u,v}(x) = \frac{I\{x(v) = m_u(x)\}}{\sum_{v' \in N(u)} I\{x(v') = m_u(x)\}} \lambda(u),$$

and

$$A^\gamma_{u,v}(t) = \int_0^t a_{u,v}(X^\gamma_s) ds. \tag{2.8}$$

For each $v \in V$, the drift of the process $X^\gamma(v)$ at time t, given that $X^\gamma_t = x$, is $\gamma^{-1} \sum_{y \in R^V} (y(v) - \gamma x(v))$ $Q(\gamma x, y)$, which by (2.7) is given by $(\sum_{u \in N^{-1}(v)} a_{u,v}(X^\gamma_t)) - X^\gamma_t(v)$. Therefore, the process $M^\gamma(v)$ defined implicitly by

$$X^\gamma_t(v) = X^\gamma_0(v) + M^\gamma_t(v) + \sum_{u \in N^{-1}(v)} A^\gamma_{u,v}(t) - \int_0^t X^\gamma_s(v) ds \tag{2.9}$$

is a local martingale with $M^\gamma_0(v) = 0$. (See Section 4.7.B and Problem 4.11.15 of [9].)

Lemma 2.3.1 *For $v \in V$, $M^\gamma(v)$ is a square integrable martingale, and*

$$E[(M^\gamma_t(v))^2] \leq \frac{1}{\gamma}(2t \sum_u \lambda(u) + K). \tag{2.10}$$

Proof. Let $\tau_n = \inf\{t : M^\gamma_t(v) \geq n\}$. Since the local martingale $M^\gamma(v)$ has jumps of size γ^{-1}, the process $M^\gamma_{t \wedge \tau_n}(v)$ is bounded and hence is a square integrable martingale. Thus,

$$\begin{aligned} E[(M^\gamma_{t\wedge\tau_n}(v))^2] &= E[[M^\gamma(v)]_{t\wedge\tau_n}] \\ &\leq \frac{1}{\gamma^2} E[\text{number of jumps of } X^\gamma(v) \text{ in } [0,t]] \\ &\leq \frac{1}{\gamma}(2t \sum_u \lambda(u) + K), \end{aligned}$$

where $[M^\gamma(v)]$ is the quadratic variation process of $M^\gamma(v)$. Fatou's Lemma implies (2.10). Finally (2.10) implies that $M^\gamma(v)$ over any finite interval is uniformly integrable; hence it is a martingale. □

Remark 2.3.1 *By Doob's L^2 inequality and Lemma 2.3.1,*

$$E[\sup_{0 \leq s \leq t} (M^\gamma_s(v))^2] \leq 4E\left[(M^\gamma_t(v))^2\right] = O(\gamma^{-1}).$$

Therefore, $M^\gamma(v) \Longrightarrow 0$ *for all* $v \in V$.

Lemma 2.3.2 *(Tightness) If the sequence* $(X_0^\gamma : \gamma > 0)$ *is tight, then the sequence of processes* $((X^\gamma, A^\gamma) : \gamma > 0)$ *is tight.*

Proof. By [9, Proposition 3.2.4], it suffices to show the tightness of (X^γ) and (A^γ) separately. Towards this end, the observation

$$\begin{aligned} A_{u,v}^\gamma(0) &= 0, \\ 0 \leq A_{u,v}^\gamma(t) - A_{u,v}^\gamma(s) &\leq (t-s)\lambda(u) \end{aligned}$$

for $t \geq s$, yields the tightness of (A^γ) ([9, Corollary 3.7.4]). This, along with Remark 2.3.1 and the representation (2.9), implies that to establish tightness of (X^γ), it suffices to establish the tightness of $(\int_0^\cdot X_s^\gamma(v)ds)$. Note that

$$\begin{aligned} E[\sup_{0\leq s\leq t} X_s^\gamma(v)] &\leq E[\frac{1}{\gamma}(\text{total number of arrivals in [0,t]}) + \sum_v X_0^\gamma(v)] \\ &\leq t\sum_u \lambda(u) + K. \end{aligned}$$

Thus, for $\eta > 0$, Markov's inequality yields that

$$P\left(\sup_{0\leq s\leq t} X_s^\gamma(v) > \frac{t\sum_u \lambda(u) + K}{\eta}\right) < \eta, \tag{2.11}$$

and the desired result follows by [9, Corollary 3.7.4]. □

Lemma 2.3.3 *(Convergence of Subsequences and Fluid Equations) If* $X_0^\gamma \Longrightarrow x_0$, *then every subsequence* $(X^{\gamma_n}, A^{\gamma_n})$ *has a further subsequence* $(X^{\gamma_{n_k}}, A^{\gamma_{n_k}})$ *such that* $(X^{\gamma_{n_k}}, A^{\gamma_{n_k}}) \Longrightarrow (x, A)$, *where* (x, A) *satisfies the following fluid equations :*

$$x_t(v) = x_0(v) + \sum_{u\in N^{-1}(v)} A_{u,v}(t) - \int_0^t x_s(v)ds \tag{2.12}$$

$$A_{u,v}(0) = 0, \;\; A_{u,v}(t) \text{ nondecreasing}, \;\; \sum_{v\in N(u)} A_{u,v}(t) = \lambda(u)t, \tag{2.13}$$

$$\int_0^t I\{x_s(v) > m_u(x_s)\}dA_{u,v}(s) = 0. \tag{2.14}$$

Proof. Let (γ_{n_k}) be a subsequence of $\gamma_n \to \infty$ such that $(X^{\gamma_{n_k}}, A^{\gamma_{n_k}})$ converges weakly. Let (x, A) denote the limit. By Skorokhod's theorem ([9, Theorem 3.1.8]), the processes can be constructed on the same probability space such that the convergence is almost everywhere. The limit x is continuous with probability one and the convergence is uniform on compact time sets ([9, Lemma 3.10.2] and [9, Lemma 3.10.1] respectively); therefore,

$$\lim_{n\to\infty} \int_0^t X^{\gamma_n}(s)ds = \int_0^t x(s)ds$$

with probability one. Since $M^{\gamma_n} \to 0$, (x, A) satisfies Equation (2.12) with the initial condition x_0. By (2.8), $(X^{\gamma_{n_k}}, A^{\gamma_{n_k}})$ satisfies conditions (2.13) and (2.14) for all k. The relation (2.13) defines a closed subset in the Skorokhod topology; hence it is satisfied by the limit A. Since

$$\int_0^t I\{X_s^\gamma(v) > m_u(X_s^\gamma)\}dA_{u,v}^\gamma(s) = 0,$$

it follows that

$$\int_0^t (X_s^\gamma(v) - m_u(X_s^\gamma)) \wedge 1 \; dA_{u,v}^\gamma(s) = 0. \tag{2.15}$$

By [10, Lemma 2.4], we can take a limit in (2.15) along the subsequence (γ_{n_k}) so that (x, A) satisfies (2.14). This establishes the lemma. □

2.3.3 The fluid limit

In this section we concentrate on the solutions to the fluid equations (2.12)-(2.14), existence of which is known due to Lemma 2.3.3. In particular, via a monotonicity argument, Lemma 2.3.5 establishes that there is a unique load trajectory that solves the fluid equations, and Lemma 2.3.7 identifies the limit point of this trajectory. We start with a remark.

Remark 2.3.2 *Equation (2.13) implies that $A_{u,v}$ has a density $a_{u,v}$ such that $\sum_{v\in N(u)} a_{u,v}(t) = \lambda(u)$ almost everywhere on the positive real line. Therefore, x and A are almost everywhere differentiable, and whenever the derivatives exist, $\dot{A}_{u,v}(t) = a_{u,v}(t)$, $\dot{x}_t(v) = (\sum_u a_{u,v}(t)) - x_t(v)$, and $I\{x_t(v) > m_u(x_t)\}a_{u,v}(t) = 0$.*

Lemma 2.3.4 *(Monotonicity) Suppose (x', A') and (x, A) are two solutions to the fluid equations (2.12)-(2.14) with $x'_0(v) \geq x_0(v)$ for all $v \in V$. Then $x'_t(v) \geq x_t(v)$ for all $v \in V$ and $t \geq 0$.*

Proof. To prove the claim by contradiction, suppose that the conclusion is false. Take $\epsilon > 0$ so that t_1 defined as follows is finite:

$$t_1 = \inf\{t \geq 0 : \ x_t(v) - x'_t(v) \geq \epsilon \text{ for some location } v \in V\}.$$

Since x'_t and x_t are continuous, the set F defined as follows is nonempty:

$$F = \{v \in V : \ x_{t_1}(v) = x'_{t_1}(v) + \epsilon\}.$$

Let $\epsilon' = \max\{x_{t_1}(v) - x'_{t_1}(v) : \ v \in F^c\}$ and ϵ_0, ϵ_1 be such that $\epsilon' < \epsilon_0 < \epsilon_1 < \epsilon$, and $\epsilon_1 > 0$. By the continuity of solutions, there exists t_0 with $0 \leq t_0 < t_1$ such that

$$x_s(v) - x'_s(v) \geq \epsilon_1 \text{ for } v \in F, \ s \in [t_0, t_1) \tag{2.16}$$

$$x_s(v) - x'_s(v) \leq \epsilon_0 \text{ for } v \in F^c, \ s \in [t_0, t_1). \tag{2.17}$$

Note that $\sum_{v \in F} x_{t_0}(v) - x'_{t_0}(v) < |F|\epsilon = \sum_{v \in F} x_{t_1}(v) - x'_{t_1}(v)$ so that

$$\sum_{v \in F} (x_{t_1}(v) - x_{t_0}(v)) > \sum_{v \in F} (x'_{t_1}(v) - x'_{t_0}(v)).$$

This, together with (2.12) and (2.16), implies the existence of a $u \in U$ such that

$$\int_{t_0}^{t_1} \sum_{v \in N(u) \cap F} a_{u,v}(s) ds > \int_{t_0}^{t_1} \sum_{v \in N(u) \cap F} a'_{u,v}(s) ds. \tag{2.18}$$

By Remark 2.3.2, for almost all $s \in [t_0, t_1)$ such that the integrand of the left-hand side of (2.18) is positive,

$$\min_{v \in N(u) \cap F} x_s(v) \leq \min_{v \in N(u) \cap F^c} x_s(v). \tag{2.19}$$

In view of (2.16) and (2.17), this implies that

$$\min_{v \in N(u) \cap F} x'_s(v) < \min_{v \in N(u) \cap F^c} x'_s(v). \tag{2.20}$$

Thus, for almost all such s, the integrand of the right-hand side of (2.18) equals $\lambda(u)$, which is an upper bound to the integrand of the left-hand side. This contradicts (2.18) and hence also the existence of t_1 for any $\epsilon > 0$. □

Lemma 2.3.5 *(Uniqueness of Load Trajectory) If (x, A) and (x', A') are two solutions to the fluid equations (2.12)-(2.14) with $x_0 = x'_0$, then $x_t = x'_t$ for all $t \geq 0$.*

Proof. Use Lemma 2.3.4 twice with $x'_0 \leq x_0$ and $x'_0 \geq x_0$. □

Remark 2.3.3 *Note that the fluid equations and the initial state x_0 do not necessarily determine A uniquely. For a simple illustration, suppose that $V = U = \{0, 1\}$, $N(u) = V$ for $u \in \{0, 1\}$, and $\lambda = (1, 1)$. Let $A_{u,v}(t) = (1/2)t$, and $\tilde{A}_{u,v}(t) = I\{u = v\}t$. Both (x, A) and $(x, \tilde{A})$ satisfy the fluid equations with $x_0(v) = 0$ and $x_t(v) = 1 - e^{-t}$ for all v.*

The uniqueness result of Lemma 2.3.5 removes the need to pass to a subsequence for the convergence of X^γ in Lemma 2.3.3.

Corollary 2.3.1 *If $X_0^\gamma \Longrightarrow x_0$, then $X^\gamma \Longrightarrow x$, where for some process A, (x, A) is a solution of the fluid equations (2.12)-(2.14) with the initial condition x_0.*

Let a be an assignment that solves the static problem $SLB(\lambda, \Phi)$ with the corresponding load q. It is easy to verify that $(q(1 - e^{-t}), at)$ is a solution to the fluid equations with zero initial state and that this solution converges to q exponentially fast as $t \to \infty$. The next two lemmas show that starting from *any* initial state x_t converges to q exponentially fast.

Lemma 2.3.6 *Let (U, V, N) be an arbitrary load sharing network. For any (x, A) that satisfies (2.12) and (2.13),*

$$\sum_v x_t(v) = \sum_v x_0(v)e^{-t} + \sum_u \lambda(u)(1 - e^{-t}).$$

In particular, $\lim_{t\to\infty} \sum_v x_t(v) = \sum_u \lambda(u)$ uniformly for all x_0 in bounded subsets of R^V.

Proof. Equations (2.12) and (2.13) yield the integral equation

$$\sum_v x_t(v) = \sum_v x_0(v) + t\sum_u \lambda(u) - \int_0^t \sum_v x_s(v)ds,$$

which yields the desired result. □

Lemma 2.3.7 *(Insensitivity to Initial State) Let (x, A) be a solution to the fluid equations with $x_0 \geq 0$. Then*

$$\|x_t - q\|_{sup} \leq e^{-t} \left(\|q\|_{sup} \vee \sum_v x_0(v) \right),$$

where $\|\cdot\|_{sup}$ denotes the supremum norm. In particular, $\lim_{t\to\infty} \|x_t - q\|_{sup} = 0$ uniformly for all x_0 in bounded subsets of R_+^V.

Proof. Let $v \in V$ be arbitrary. By Lemma 2.3.4, $x_t(v) \geq q(v)(1-e^{-t})$; thus $0 \leq x_t(v) - q(v)(1-e^{-t}) \leq \sum_{v'}(x_t(v') - q(v')(1-e^{-t}))$. By Lemma 2.3.6, $\sum_{v'}(x_t(v') - q(v')(1-e^{-t})) = \sum_{v'} x_0(v')e^{-t}$; therefore,

$$-q(v)e^{-t} \leq x_t(v) - q(v) \leq (-q(v) + \sum_{v'} x_0(v'))e^{-t}.$$

This establishes the lemma. □

2.3.4 Asymptotic optimality of least load routing

This section establishes the asymptotic optimality of LLR for the optimal control problem formulated in Section 2.3. In Section 2.3.3 it was shown that the finite dimensional distributions of the normalized load process converge as $\gamma \to \infty$, and the limit process converges to an optimal point q as $t \to \infty$. Lemma 2.3.9 establishes the convergence of the equilibrium distribution of the normalized load process to the deterministic distribution concentrated at q. These facts are used to prove Theorem 2.3.1 on the asymptotic optimality of LLR.

In what follows, P_μ denotes the distribution of the process X^γ when X_0^γ has distribution μ. Also, μ_0 is the deterministic distribution concentrated at the zero state, and μ_t^γ denotes the distribution of X_t^γ given $X_0^\gamma = 0$. We start with an auxiliary lemma.

Lemma 2.3.8 *Given $\epsilon > 0$, there exists a γ_ϵ such that whenever $\gamma > \gamma_\epsilon$,*

$$P_{\mu_0}\left(\sum_v X_t^\gamma(v) \leq \epsilon + \sum_u \lambda(u) \right) \geq 1 - \epsilon \quad \textit{for any } t \geq 0.$$

Proof. Starting from the zero state, the total load in the system at any time $t > 0$ is stochastically dominated by a Poisson random variable with mean $\gamma \sum_u \lambda(u)$. Chebychev's inequality yields the desired result. □

Lemma 2.3.9 *(Convergence of Equilibrium Distributions) Let q be the unique load vector corresponding to solutions of $SLB(\lambda, \Phi)$ and ν be the distribution of the equilibrium load X_∞^γ. Then for all $\epsilon > 0$,*

$$\lim_{\gamma \to \infty} \nu \left(\|X_\infty^\gamma - q\|_{sup} > \epsilon \right) = 0.$$

Proof. Let $\epsilon > 0$ be fixed. Note that $\nu(\|X_\infty^\gamma - q\|_{sup} > \epsilon) \leq \liminf_{T \to \infty} P_{\mu_0}(\|X_T^\gamma - q\|_{sup} > \epsilon)$ so that it suffices to show that

$$P_{\mu_0}(\|X_T^\gamma - q\|_{sup} > \epsilon) < \epsilon$$

for all sufficiently large T and γ. Towards this end, appeal to Lemma 2.3.7 to fix θ so that $\|x_t - q\|_{sup} < \epsilon/2$ whenever $t \geq \theta$, for all x such that $\sum_v x_0(v) < \epsilon + \sum_u \lambda(u)$. By the time homogeneous Markov property of X^γ,

$$P_{\mu_0}\left(\|X_T^\gamma - q\|_{sup} > \epsilon\right) = P_{\mu_{T-\theta}^\gamma}\left(\|X_\theta^\gamma - q\|_{sup} > \epsilon\right)$$

whenever $T \geq \theta$. Therefore, to prove the lemma, it suffices to establish the following claim: There exists a γ_ϵ' such that whenever $\gamma > \gamma_\epsilon'$,

$$P_{\mu_{T-\theta}^\gamma}\left(\|X_\theta^\gamma - q\|_{sup} > \epsilon\right) < \epsilon \quad \text{for all } T > \theta. \tag{2.21}$$

To argue by contradiction, suppose that the claim is false. Then one can construct a sequence $(\tilde{\mu}^\xi)$ with $\xi \to \infty$, such that $\tilde{\mu}^\xi = \mu_{t(\xi)}^\xi$ for some choice of $t(\xi) > 0$, and

$$P_{\tilde{\mu}^\xi}\left(\|X_\theta^\xi - q\|_{sup} > \epsilon\right) \geq \epsilon. \tag{2.22}$$

By Lemma 2.3.8, $(\tilde{\mu}^\xi)$ is tight; therefore, by Lemma 2.3.3, there exists a subsequence $\xi_n \to \infty$ such that if $X_0^{\xi_n} \sim P_{\tilde{\mu}^{\xi_n}}$ then $X^{\xi_n} \Longrightarrow x$, for some x as in Lemma 2.3.3. Hence there exists an n_ϵ such that

$$P_{\tilde{\mu}^{\xi_n}}\left(\|X_\theta^{\xi_n} - x_\theta\|_{sup} > \epsilon/2\right) < \epsilon/2 \tag{2.23}$$

whenever $n > n_\epsilon$. However, by the choice of θ and Lemma 2.3.8, for all ξ_n sufficiently large,

$$P_{\tilde{\mu}^{\xi_n}}\left(\|x_\theta - q\|_{sup} > \epsilon/2\right) < \epsilon/2. \tag{2.24}$$

Observations (2.23) and (2.24) contradict (2.22), hence proving (2.21), which establishes the lemma. □

Theorem 2.3.1 *(Asymptotic Optimality of LLR) Given an allocation policy π, let J_γ^π denote the cost under π when the network demand is $\gamma\lambda$. Then*

$$\liminf_{\gamma\to\infty} \gamma^{-2} J_\gamma^\pi \geq \lim_{\gamma\to\infty} \gamma^{-2} J_\gamma^{LLR} > 0.$$

Proof. Note that in equilibrium $\gamma X_\infty^\gamma(v)$ is stochastically dominated by a Poisson random variable with mean $\gamma \sum_u \lambda(u)$, for all $v \in V$. Consequently, $E[(X_\infty^\gamma(v))^p] = O(1)$ for all $p \in Z_+$. In particular, $((X_\infty^\gamma(v))^2 : \gamma \geq 0)$ is uniformly integrable. Thus, by Lemma 2.3.9,

$$\begin{aligned} \lim_{\gamma\to\infty} \gamma^{-2} J_\gamma^{LLR} &= \lim_{\gamma\to\infty} E\left[\sum_v (X_\infty^\gamma(v))^2\right] \\ &= E\left[\sum_v (q(v))^2\right] \\ &= \Psi(\lambda) > 0. \end{aligned}$$

Inequality (2.6) implies that for any allocation policy π,

$$\begin{aligned} \gamma^{-2} J_\gamma^\pi &\geq \gamma^{-2}\Psi(\gamma\lambda) \\ &= \Psi(\lambda) \end{aligned}$$

for all $\gamma > 0$. This proves the theorem. □

2.4 Finite Capacities

This section considers a variation of the basic model in which each location has a finite capacity. Namely, we assume that the load of a location cannot exceed its capacity, and arrivals to the congested neighborhoods are dropped. In this setting, a natural objective for the allocation policy is to minimize the percentage of consumers dropped in the system. We concentrate on a broad class of practical allocation policies, namely the *least* relative *load routing* policies, in which new

consumers are assigned to the location with the least relative load. The relative load of a location is defined by applying a normalization function to the actual load. Theorem 2.4.1 establishes that such policies asymptotically achieve the smallest loss probability for large arrival rates. We then provide stronger results on two members of this class, namely the *least ratio routing* (LRR) and the *maximum residual capacity routing* (MRCR) policies.

The use of a binary valued normalization function would model trunk reservation strategies, studied in a similar context by Hunt and Kurtz [7]. However, we require the normalization functions to be strictly increasing; thus, trunk reservation strategies are not covered in this chapter. In doing so, we avoid the pathologies associated with trunk reservations in heavy traffic, and can therefore establish optimality results. The drawback of our approach is that more feedback information about the network state is required to implement the allocation policies.

To describe the dynamic model of interest, let a *capacity vector* $\kappa = (\kappa(v) \ : \ v \in V)$ be a vector of positive numbers. Given a load sharing network (U, V, N), a capacity vector κ, and a load vector λ, consider the limiting regime of Section 2.3.2. Suppose that in each system indexed by γ, each location v has capacity $\lfloor \gamma\kappa(v) \rfloor$. A location is called *full* if its load and capacity are equal, and a consumer is *lost* if, upon its arrival, all of the locations in its neighborhood are full. Lost consumers cannot be assigned later; hence they are treated by the system as if they never arrived. The problem of interest is to find allocation policies that minimize the *loss probability* $P_\gamma(Loss)$, which is defined as

$$P_\gamma(Loss) \ = \ \liminf_{t\to\infty} \frac{E[\text{number of consumers lost in } [0,t]]}{t\gamma \sum_u \lambda(u)}.$$

In light of Section 2.3, we start with some definitions regarding an associated static problem. Given a capacity vector κ, define $\mathcal{B}_{\lambda,\kappa}$ as the set of admissible assignment vectors a such that $\sum_v a_{u,v} \leq \lambda(u)$ for all u, and $q(v) \leq \kappa(v)$ for all v, where q denotes the load vector determined by assignment a. The *static load packing problem* (SLP) is the simple assignment problem defined as

$$SLP(\lambda,\kappa) \ : \ Maximize(\sum_v q(v) \ : \ a \in \mathcal{B}_{\lambda,\kappa}).$$

Towards the end of characterizing certain solutions to SLP, we have the following definition:

Definition 2.4.1 *A function $f : R^V \to R^V$ is called a* normalization function *if for all $v \in V$, the real valued function $f(\cdot, v)$ has the following three properties:*

(*i*) $f(q,v)$ *depends on* q *only through* $q(v)$,

(*ii*) $f(q,v)$ *is a strictly increasing and continuously differentiable function of* $q(v)$, *such that* $\partial f(q,v)/\partial q(v) \geq \delta$ *for some* $\delta > 0$,

(*iii*) $f(q,v) = 0$ *when* $q(v) = \kappa(v)$.

Consider the following two conditions on a generic assignment a, where q denotes the load vector corresponding to a, and $m_u(f(q)) = \min_{v \in N(u)} f(q,v)$:

Condition 2.4.1 $a_{u,v} = 0$ *whenever* $f(q,v) > m_u(f(q))$.

Condition 2.4.2 $\sum_v a_{u,v} < \lambda(u)$ *only if* $f(q,v) = 0$ *for all* $v \in N(u)$.

Let the function $\Phi : R^V \to R$ be defined as $\Phi(q) = \sum_v \int_0^{q(v)} f(\tilde{q},v)d\tilde{q}(v)$. Note that Φ is convex, however, not necessarily symmetric in its arguments. The following three lemmas are proved in Section 2.7. The first lemma concerns a static load *balancing* problem, the second concerns a connection between static load balancing and load packing, and the third gives a sufficient condition for optimality in $SLP(\lambda,\kappa)$.

Lemma 2.4.1 *(Load Balancing) There exists a solution to* $SLB(\lambda,\Phi)$. *An admissible assignment* $\tilde{a}$ *which satisfies demand* λ *solves the* $SLB(\lambda,\Phi)$ *if and only if* $\tilde{a}$ *satisfies Condition 2.4.1 . Furthermore, all such assignments yield the same load vector.*

Lemma 2.4.2 *(Truncation) Let* $\tilde{a}$ *solve* $SLB(\lambda,\Phi)$ *with the corresponding load vector* $\tilde{q}$, *and let* a *be the assignment defined by*

$$a_{u,v} = \tilde{a}_{u,v}\left(1 \wedge \frac{\kappa(v)}{\tilde{q}(v)}\right).$$

Then $a \in \mathcal{B}_{\lambda,\kappa}$ *and* a *satisfies Conditions 2.4.1 and 2.4.2 with the corresponding load vector* $\tilde{q}(v) \wedge \kappa(v)$.

Lemma 2.4.3 *(Sufficiency) An assignment vector* $a \in \mathcal{B}_{\lambda,\kappa}$ *solves* $SLP(\lambda,\kappa)$ *if there exists a normalization function* f *such that both Conditions 2.4.1 and 2.4.2 hold. For a given normalization function, there exists an assignment* $a \in \mathcal{B}_{\lambda,\kappa}$ *that satisfies Conditions 2.4.1 and 2.4.2, and all such assignments yield the same load vector.*

To treat the lossy network in the context of the already existing theory, we introduce a new location v_L and define an extended load sharing network $(\hat{U}, \hat{V}, \hat{N})$ by $\hat{U} = U$, $\hat{V} = V \cup \{v_L\}$, and $\hat{N}(u) = N(u) \cup \{v_L\}$, where $\kappa(v_L) = \infty$. A load process $(X(v) : v \in V)$ corresponding to an allocation policy π can be extended to a load process $(X(v) : v \in \hat{V})$ on $(\hat{U}, \hat{V}, \hat{N})$ by letting $X_t(v_L)$ denote the number of blocked consumers that would have been in service at time t if they were not blocked. We continue to use X to denote the extended process, and let $X^\gamma(v) = \gamma^{-1}X(v)$ for all $v \in \hat{V}$.

The solution of $SLP(\lambda, \kappa)$ provides a lower bound for the loss probability of *any* allocation policy.

Lemma 2.4.4 *For any allocation policy π and any $\gamma > 0$,*

$$P_\gamma^\pi(Loss) \geq 1 - \frac{\sum_{v \in V} q(v)}{\sum_u \lambda(u)},$$

where q is the load vector corresponding to a solution of $SLP(\lambda, \kappa)$.

Proof. Consider the load process X on $(\hat{U}, \hat{V}, \hat{N})$, and assume, without loss of generality, that X starts with the zero initial state. Let

$$\begin{aligned} s_k &= \text{Holding time of the } k^{th} \text{ arrival to } v_L, \\ \beta(t) &= \text{Number of arrivals to } v_L \text{ in } [0, t). \end{aligned}$$

Note that $\beta(t)$ is the number of consumers lost by time t in the original system (U, V, N). By the construction of X,

$$\begin{aligned} E[\int_0^t \sum_{v \in \hat{V}} X_s(v)ds] - E[\int_0^t \sum_{v \in V} X_s(v))ds] &= E[\int_0^t X_s(v_L)ds] \\ &\leq E[\sum_{k=1}^{\beta(t)} s_k] \\ &= E[\beta(t)], \end{aligned} \tag{2.25}$$

where the last step follows by the independence of $(s_k : k \geq 1)$ and $\beta(t)$. For every s, $E[X_s^\gamma]$ is the load vector corresponding to an assignment in $\mathcal{B}_{\lambda,\kappa}$; therefore, by the choice of q, $\sum_{v \in V} E[X_s^\gamma(v)] \leq \sum_{v \in V} q(v)$ so that $E[\int_0^t \sum_{v \in V} X_s(v)ds] \leq t\gamma \sum_{v \in V} q(v)$. Rearranging (2.25) and observing that

$E[\int_0^t \sum_{v\in\hat{V}} X(s)ds] = \gamma \sum_u \lambda(u) \int_0^t (1-e^{-s})ds$ yield

$$\frac{E[\beta(t)]}{t\gamma} \geq (1 - \frac{1-e^{-t}}{t}) \sum_u \lambda(u) - \sum_{v\in V} q(v).$$

The result follows by dividing each side by $\sum_u \lambda(u)$ and letting $t \to \infty$. □

In the context of finite capacity constraints, we investigate a class of allocation policies which are named *least relative load routing* (LRLR) policies. Given a normalization function f, an LRLR policy is defined by the following assignment rule:

- Upon arrival, a consumer of type u is assigned to a location $v \in N(u)$ with the minimum *relative load*, $f(X^\gamma, v)$, provided that the minimum relative load is less than zero. Otherwise, all of the locations in $N(u)$ are full, and the consumer is lost.

Consider the extended load process $(X(v) : v \in \hat{V})$ under an LRLR policy. Intuitively, this process is lossless, and it is also governed by LRLR with the normalization function extended by defining $f(q, v_L) = 0^-$. The process X^γ is Markov and has the following representation:

$$X_t^\gamma(v) = X_0^\gamma(v) + M_t^\gamma(v) + \sum_{u\in\hat{N}^{-1}(v)} A_{u,v}^\gamma(t) - \int_0^t X_s^\gamma(v)ds,$$

where

$$A_{u,v}^\gamma(t) = \begin{cases} \int_0^t \frac{I\{f(X_s,v)=m_u(f(X_s))\}I\{f(X_s,v)<0\}}{\sum_{v'\in N(u)} I\{f(X_s,v')=m_u(f(X_s))\}}\lambda(u)ds & \text{if } v \in V \\ \int_0^t I\{m_u(f(X_s)) = 0\}\lambda(u)ds & \text{if } v = v_L, \end{cases}$$

and $M^\gamma(v)$ is a local martingale with $M_0^\gamma(v) = 0$. Given that (X_0^γ) is tight, the methods of Section 2.3.2 can be applied to establish the tightness of (X^γ, A^γ) and characterize the possible weak limits. Namely, the following lemma holds:

Lemma 2.4.5 *Suppose (X_0^γ) is tight. Then every subsequence $(X^{\gamma_n}, A^{\gamma_n})$ has a further subsequence $(X^{\gamma_{n_k}}, A^{\gamma_{n_k}})$ such that $(X^{\gamma_{n_k}}, A^{\gamma_{n_k}}) \Longrightarrow (x, A)$, where (x, A) satisfies the following fluid equations:*

$$x_t(v) = x_0(v) + \sum_{u\in\hat{N}^{-1}(v)} A_{u,v}(t) - \int_0^t x_s(v)ds, \tag{2.26}$$

$$A_{u,v}(0) = 0, \quad A_{u,v}(t) \text{ nondecreasing}, \tag{2.27}$$

$$0 \le x_t(v) \le \kappa(v), \quad \sum_{v \in \tilde{N}(u)} A_{u,v}(t) = \lambda(u)t, \tag{2.28}$$

$$\int_0^t I\{f(x_s, v) > m_u(f(x_s))\} dA_{u,v}(s) = 0 \quad v \neq v_L, \tag{2.29}$$

$$\int_0^t I\{m_u(f(x_s)) < 0\} dA_{u,v_L}(s) = 0. \tag{2.30}$$

Call t a *regular point* of a function $g : R \to R$ if g is differentiable at t, and let $\dot{g}_t$ denote the derivative of g at a regular point t. The following lemma is proved in Section 2.7.

Lemma 2.4.6 *Let g be an absolutely continuous function, and let $\alpha \in R$. Then $\dot{g}_t = 0$ for almost all t such that $g_t = \alpha$.*

Lemma 2.4.7 *(Monotonicity) Let (x', A') and (x, A) be two solutions to the fluid equations (2.26)-(2.30). Then,*

(i) If $x'_0(v) \ge x_0(v)$ for all $v \in V$, then $x'_t(v) \ge x_t(v)$ for all $v \in V$ and $t \ge 0$.

(ii) If in addition $x'_0(v_L) \ge x_0(v_L)$, then $x'_t(v_L) \ge x_t(v_L)$ for all $t \ge 0$.

Proof. For (i), the proof of Lemma 2.3.4 applies directly by replacing F^c by $F^c \cap V$ and using $f(x_s)$ and $f(x'_s)$ in place of x_s and x'_s in inequalities (2.19) and (2.20), respectively.

To prove (ii), for each t define

$$F_t = \{v \in V : x_t(v) = \kappa(v)\}, \quad F'_t = \{v \in V : x'_t(v) = \kappa(v)\}, \tag{2.31}$$

$$G_t = \{u \in U : N(u) \subset F_t\}, \quad G'_t = \{u \in U : N(u) \subset F'_t\}. \tag{2.32}$$

By part (i), $F_t \subset F'_t$, and $G_t \subset G'_t$ for all t. By Remark 2.3.2 applied to (2.26)-(2.30) and Lemma 2.4.6, for almost all t,

$$\begin{aligned} &(\sum_{u \in G_t} a_{u,v}(t)) - \kappa(v) = \dot{x}_t(v) = 0 \quad \text{for all } v \in F_t, \\ &(\sum_{u \in G'_t} a'_{u,v}(t)) - \kappa(v) = \dot{x'_t}(v) = 0 \quad \text{for all } v \in F'_t. \end{aligned} \tag{2.33}$$

Therefore, for almost all t,

$$\begin{aligned}
\dot{x}_t(v_L) &= \dot{x}_t(v_L) + \sum_{v \in F_t} \dot{x}_t(v) \\
&= \sum_{u \in G_t} \lambda(u) - (\sum_{v \in F_t} \kappa(v)) - x_t(v_L), \\
\dot{x'_t}(v_L) &= \dot{x'}_t(v_L) + \sum_{v \in F_t} \dot{x'}_t(v) \\
&\leq \sum_{u \in G_t} \lambda(u) - (\sum_{v \in F_t} \kappa(v)) - x'_t(v_L),
\end{aligned}$$

where the inequality follows by (2.33) and the definitions (2.31) and (2.32). Hence if $e_t = x'_t(v_L) - x_t(v_L)$, then $\dot{e}_t \geq -e_t$ for almost all t. This, along with the hypothesis $e_0 \geq 0$, proves (ii). □

Corollary 2.4.1 *(Uniqueness) If (x, A) and (x', A') are two solutions to the fluid equations (2.26)-(2.30) with $x_0 = x'_0$, then $x_t = x'_t$ for all $t \geq 0$.*

We now concentrate on the properties of the unique trajectory x that corresponds to the solutions of the fluid equations (2.26)-(2.30). The proof of the following lemma can be found in Section 2.7.

Lemma 2.4.8 *Let $g_t(i)$ be absolutely continuous, $i = 1, 2, \cdots, I$, and set $m_t = \min_i g_t(i)$. Then m is absolutely continuous, almost every t is a regular point for $g(1), \cdots, g(I), m$, and for all such t, $\dot{m}_t = \dot{g}_t(i)$ for all i such that $g_t(i) = m_t$.*

Note that by the continuous differentiability of f, there exists Δ such that $\frac{\partial f(q', v)}{\partial q'(v)} \leq \Delta$ whenever $0 \leq q'(v) \leq \kappa(v)$ for all $v \in V$. Let q be the optimal load vector corresponding to the assignments satisfying Conditions 2.4.1 and 2.4.2. Extend q to $\hat{V}$ by setting $q(v_L) = (\sum_u \lambda(u)) - \sum_{v \in V} q(v)$. The next two lemmas establish the convergence of x to the load vector q.

Lemma 2.4.9 *Given $\epsilon > 0$, there exists $\tau_0(\epsilon)$ such that for all $v \in \hat{V}$,*

$$x_t(v) \geq q(v) - \epsilon \tag{2.34}$$

whenever $t \geq \tau_0(\epsilon)$.

Proof. We first establish the inequality (2.34) for $v \in V$. Let $\{V_1, V_2, \cdots, V_J\}$ and $\{U_1, U_2, \cdots, U_J\}$ be the unique partitions of V and U, respectively, defined by Lemma 2.2.2 when condition (2.1) is replaced by Condition 2.4.1, and the vector q is replaced by $f(q)$ in (2.2) and (2.3). Let $j \in \{1, 2, \cdots, J\}$ and define

$$\begin{aligned} m_t^j &= \inf_{v \in \bigcup_{i=j}^J V_i} f(x_t, v), \\ F_t^j &= \left\{ v \in \bigcup_{i=j}^J V_i \,:\, f(x_t, v) = m_t^j \right\}, \\ N^*(F_t^j) &= \{u : N(u) \cap F_t^j \neq \emptyset \text{ and } N(u) \cap (\bigcup_{i=1}^{j-1} V_i) = \emptyset\}, \end{aligned}$$

with the understanding that $\bigcup_{i=1}^0 V_i = \emptyset$. Let f_j denote the value such that $f(q, v) = f_j$ for all $v \in V_j$. Assume that t is a regular point of m^j such that $m_t^j < f_j - \epsilon\delta$. Then by the explanations indicated in parentheses,

$$\begin{aligned} |F_t^j| \dot{m}_t^j &= \textstyle\sum_{v \in F_t^j} \dot{f}(x_t, v) && (\text{ Lemma 2.4.8 }) \\ &\geq \delta \textstyle\sum_{v \in F_t^j} \dot{x}_t(v) && (\text{ Definition 2.4.1 }) \\ &\geq \delta(\textstyle\sum_{u \in N^*(F_t^j)} \lambda(u) - \sum_{v \in F_t^j} x_t(v)) && (\text{ Definition of } F_t^j \text{ and the fluid equations }) \\ &\geq \delta(\textstyle\sum_{u \in N^*(F_t^j)} \lambda(u) - \sum_{v \in F_t^j} (q(v) - \epsilon\delta/\Delta)) && (\text{ Definition of } F_t^j \text{ }) \\ &\geq |F_t^j| \epsilon\delta^2/\Delta. && (\text{ Definition of } N^* \text{ and } q \text{ }) \end{aligned}$$

Therefore, if $t \geq \sup_v |f(0, v)|\Delta/\epsilon\delta^2$, then $m_t^j \geq f_j - \epsilon\delta$, so that $f(x_t, v) \geq f(q, v) - \epsilon\delta$ for $v \in V_j$, which in turn implies that $x_t(v) \geq q(v) - \epsilon$ for $v \in V_j$. Since j is arbitrary, (2.34) holds for all $t \geq \sup_v |f(0, v)|\Delta/\epsilon\delta^2$, and $v \in V$.

To complete the proof of the lemma, note that by Lemma 2.3.6, there exists a $\tau_J(\epsilon)$ such that $x_t(v_L) + \sum_{v \in V_J} x_t(v) \geq \sum_{u \in U_J} \lambda(u) - \epsilon$ for all $t \geq \tau_J(\epsilon)$. Therefore, for all such t, $x_t(v_L) \geq (\sum_{u \in U_J} \lambda(u) - \epsilon - \sum_{v \in V_J} \kappa(v))_+ \geq q(v_L) - \epsilon$. This proves (2.34) for $v = v_L$ and establishes the lemma with $\tau_0(\epsilon) = (\sup_v |f(0, v)|\Delta/\epsilon\delta^2) \vee \tau_J(\epsilon)$. □

Lemma 2.4.10 *(Insensitivity to Initial State) If (x, A) is a solution to the fluid equations (2.26)-(2.30), then $\lim_{t \to \infty} \|x_t - q\|_{sup} = 0$ uniformly for all x_0 in bounded subsets of $R_+^{\hat{V}}$.*

Proof. Fix $l > 0$ and let $\sum_{v\in\hat{V}} x_0(v) < l$. Given $\epsilon > 0$, set $\epsilon_0 = \epsilon/(|V|+2)$. Appealing to Lemma 2.3.6, let $\tau_1(l,\epsilon_0)$ be such that for all $t \geq \tau_1(l,\epsilon_0)$, $|\sum_{v\in\hat{V}} x_t(v) - \sum_u \lambda(u)| < \epsilon_0$, or equivalently $|\sum_{v\in\hat{V}}(x_t(v) - q(v))| < \epsilon_0$. If $t > \tau_0(\epsilon_0) \vee \tau_1(l,\epsilon_0)$, then Lemma 2.4.9 implies that

$$\inf_{v\in\hat{V}}(x_t(v) - q(v)) \geq -\epsilon_0 \geq -\epsilon.$$

This in turn implies that

$$\begin{aligned}\sup_{v\in\hat{V}}(x_t(v) - q(v)) &\leq \sum_{v\in\hat{V}}(x_t(v) - q(v) + \epsilon_0) \\ &= \sum_{v\in\hat{V}}(x_t(v) - q(v)) + (|V|+1)\epsilon_0 \\ &\leq (|V|+2)\epsilon_0 = \epsilon,\end{aligned}$$

which yields the desired result. □

Lemma 2.4.11 *(Convergence of Equilibrium Distributions) Let q be the unique load vector corresponding to assignments satisfying Conditions 2.4.1 and 2.4.2, and ν be the distribution of the equilibrium load X_∞^γ. Then for all $\epsilon > 0$,*

$$\lim_{\gamma\to\infty} \nu\left(\|X_\infty^\gamma - q\|_{sup} > \epsilon\right) = 0.$$

Proof. The proof of Lemma 2.3.9 applies directly by using Lemmas 2.4.5 and 2.4.10 in place of Lemmas 2.3.3 and 2.3.7, respectively. □

Theorem 2.4.1 *(Asymptotic Optimality of LRLR) Let $P_\gamma^\pi(Loss)$ denote the loss probability of an arbitrary allocation policy π and $P_\gamma^{LRLR}(Loss)$ denote the loss probability of the LRLR policy for some normalization function f. Then*

$$\liminf_{\gamma\to\infty} P_\gamma^\pi(Loss) \geq \lim_{\gamma\to\infty} P_\gamma^{LRLR}(Loss) = 1 - \frac{\sum_{v\in V} q(v)}{\sum_u \lambda(u)},$$

where q is the load vector corresponding to a solution of $SLP(\lambda,\kappa)$.

Proof. The collection $(X_\infty^\gamma : \gamma > 0)$ is dominated by normalized Poisson random variables and is uniformly integrable. Therefore,

$$\begin{aligned}
\lim_{\gamma\to\infty} P_\gamma^{LRLR}(Loss) &= \lim_{\gamma\to\infty} \liminf_{t\to\infty} \frac{E[\text{number of consumers lost in } [0,t]]}{t\gamma\sum_u \lambda(u)}, \\
&= \frac{1}{\sum_u \lambda(u)} \lim_{\gamma\to\infty} E[X_\infty^\gamma(v_L)], \\
&= \frac{q(v_L)}{\sum_u \lambda(u)}, \\
&= 1 - \frac{\sum_{v\in V} q(v)}{\sum_u \lambda(u)},
\end{aligned}$$

where the second step follows by Little's Theorem, and the third step is a consequence of Lemma 2.4.11 and the uniform integrability of $(X_\infty^\gamma : \gamma > 0)$. The theorem now follows by Lemma 2.4.4. □

Having proven the optimality properties of generic LRLR policies, we now focus on two particular elements of this class, namely the least ratio routing and the maximum residual capacity routing policies. In the next two sections, we obtain a stronger version of Lemma 2.4.10 for these policies and provide explicit solutions of the fluid equations (2.26)-(2.30) for certain initial conditions.

2.4.1 Least ratio routing

In this section we focus on a particular least relative load routing policy, namely the *least ratio routing* (LRR). The LRR policy is defined as the LRLR policy associated with the normalization function $f(q,v) = -1 + q(v)/\kappa(v)$. Note that LRR assigns each consumer to the location with the least *load-to-capacity ratio*, $X_t^\gamma(v)/\kappa(v)$.

Let $\tilde{a}$ and $\tilde{q}$ be defined as in Lemma 2.4.2. Define the trajectories a and z as follows:

$$a_{u,v}(t) = \begin{cases} \tilde{a}_{u,v} & v \in V \quad t < \ln(\tilde{q}(v)/(\tilde{q}(v) - \kappa(v))_+), \\ \tilde{a}_{u,v}\kappa(v)/\tilde{q}(v) & v \in V \quad t \geq \ln(\tilde{q}(v)/(\tilde{q}(v) - \kappa(v))_+), \\ \sum_{v\in V}(\tilde{a}_{u,v} - a_{u,v}(t)) & v = v_L, \end{cases}$$

$$z_t(v) = \begin{cases} \tilde{q}(v)(1 - e^{-t}) \wedge \kappa(v) & v \in V, \\ (1 - e^{-t})\sum_u \lambda(u) - \sum_{v\in V} z_t(v) & v = v_L. \end{cases}$$

It is straightforward to verify that $(z, \int_0^\cdot a(s)ds)$ solves the fluid equations (2.26)-(2.30) with the zero initial state.

Lemma 2.4.12 *(Convergence Rate of LRR) Let q be the unique load vector corresponding to assignments satisfying Conditions 2.4.1 and 2.4.2. If (x, A) is a solution to the fluid equations (2.26)-(2.30), then*

$$\sup_{v \in V} |x_t(v) - q(v)| \leq e^{-t} \left(\sup_{v \in V} q(v) \vee \sum_{v \in V} x_0(v) \right).$$

Proof. By Lemma 2.4.7, for all $v \in V$, $x_t(v) \geq z_t(v)$. Also,

$$x_t(v) - z_t(v) \leq \sum_{v \in V} (x_t(v) - z_t(v)) = e^{-t} \sum_{v \in V} x_0(v).$$

These inequalities, together with the fact that

$$q(v)(1 - e^{-t}) \leq z_t(v) \leq q(v)$$

for all $v \in V$, yield the desired result. □

2.4.2 Maximum residual capacity routing

The *maximum residual capacity routing* (MRCR) is defined as the LRLR policy associated with the normalization function $f(q, v) = q(v) - \kappa(v)$. Note that the MRCR policy assigns each consumer to the location with the maximum *residual capacity* defined as $\gamma\kappa(v) - X_t(v)$.

Let a be an optimal assignment satisfying Conditions 2.4.1 and 2.4.2, and let q be the load vector determined by a. Extend a to $(\hat{U}, \hat{V}, \hat{N})$ by setting $a_{u,v_L} = \lambda(u) - \sum_{v \in V} a_{u,v}$. Let z_0 denote the vector such that $z_0(v) = \kappa(v)$ for $v \in V$, and $0 \leq z_0(v_L) < \infty$. Direct verification yields that (z, A), defined by

$$\begin{aligned} z_t(v) &= z_0(v)e^{-t} + q(v)(1 - e^{-t}) \\ A_{u,v}(t) &= a_{u,v}t, \end{aligned}$$

is a solution to the fluid equations (2.26)-(2.30) starting with the initial state z_0.

Lemma 2.4.13 *(Convergence Rate of MRCR) Let q be the unique load vector corresponding to assignments satisfying Conditions 2.4.1 and 2.4.2. If (x, A) is a solution to the fluid equations*

(2.26)-(2.30) with an arbitrary initial state x_0, then

$$\sup_{v\in\hat{V}} |x_t(v) - q(v)| \leq e^{-t}\left((q(v_L) \vee x_0(v_L)) + \sum_{v\in V}\kappa(v)\right).$$

Proof. Let z_0 be an initial state vector defined as $z_0(v) = \kappa(v)$ for $v \in V$ and $z_0(v_L) = x_0(v_L)$, and let z denote the load trajectory starting with z_0. By Lemma 2.4.7, for all $v \in \hat{V}$,

$$\begin{aligned} x_t(v) &\leq z_t(v) \\ &= z_0(v)e^{-t} + q(v)(1-e^{-t}). \end{aligned}$$

On the other hand, for any $v \in \hat{V}$,

$$\begin{aligned} z_t(v) - x_t(v) &\leq \sum_{v\in\hat{V}}(z_t(v) - x_t(v)) \\ &= e^{-t}\sum_{v\in\hat{V}}(z_0(v) - x_0(v)) \\ &\leq e^{-t}\sum_{v\in V}\kappa(v). \end{aligned}$$

Therefore,

$$(z_0(v) - q(v) - \sum_{v\in V}\kappa(v))e^{-t} \leq x_t(v) - q(v) \leq (z_0(v) - q(v))e^{-t},$$

and the desired result follows. □

2.5 The Migration Model

In this section we consider locations with infinite capacities and generalize the basic model of Section 2.3 by allowing consumers to change their types while they are in the system and also by including type-dependent departure rates. Towards this end, Lemma 2.5.1 identifies the weak limits of the network process as solutions to certain fluid equations. An example shows that the fluid equations do not necessarily uniquely determine the transient behavior of the load; nevertheless, by Lemma 2.5.5, the limit point is unique. These facts are used to establish Theorem 2.5.1 on the asymptotic optimality of LLR.

The analytical description of the migration model involves a *routing matrix* R, such that $R = [r_{u,u'}]_{U\times U}$, where $r_{u,u'} \geq 0$ for $u \neq u'$, and $\sum_{u'\in U} r_{u,u'} \leq 0$ for all $u \in U$. Given a load sharing

network (U, V, N), an arrival rate vector λ, and a routing matrix R, consider the load balancing problem of Section 2.3. Suppose for all $u, u' \in U$ such that $u' \neq u$, each type u consumer transforms into a type u' consumer with rate $r_{u,u'}$ or departs from the system with rate $-\sum_{u' \in U} r_{u,u'}$. Each arrival is assigned to a location via the LLR policy. In addition, when a consumer changes its type, it is reassigned using LLR. Its location may or may not change. We assume that R is nonsingular, so that every consumer eventually departs from the system. Let $L_t(u)$ continue to denote the number of type u consumers in the network at time t. It can be verified by direct substitution that the equilibrium vector $(L_\infty(u) : u \in U)$ is a vector of independent Poisson random variables with mean vector $\gamma\rho$, where $\rho = -\lambda R^{-1}$, so that the normalized cost of any allocation policy is lower bounded by $\Psi(\rho)$. This section extends the analysis of Section 2.3 to the more general setting.

Let the *contribution* of type u to location v at time t, denoted by $C_t(u, v)$, be the number of type u consumers at location v at time t. Define $C_t^\gamma(u, v) = \gamma^{-1} C_t(u, v)$, and $L_t^\gamma(u) = \gamma^{-1} L_t(u)$. Note that $X_t^\gamma(v) = \sum_u C_t^\gamma(u, v)$, and $L_t^\gamma(u) = \sum_v C_t^\gamma(u, v)$. Under LLR, C is Markov on the state space $Z_+^{U \times V}$, and $C^\gamma(u, v)$ can be represented as

$$C_t^\gamma(u, v) = C_0^\gamma(u, v) + M_t^\gamma(u, v) + A_{u,v}^\gamma(t) + \int_0^t r_{u,u} C_s^\gamma(u, v) ds,$$

where

$$A_{u,v}^\gamma(t) = \int_0^t (\lambda(u) + \sum_{u' \neq u} L_s^\gamma(u') r_{u',u}) \frac{I\{X_s(v) = m_u(X_s)\}}{\sum_{v' \in N(u)} I\{X_s(v') = m_u(X_s)\}} ds,$$

and $M^\gamma(u, v)$ is a local martingale with $M_0^\gamma(u, v) = 0$. Given that (C_0^γ) is tight, the methods of Section 2.3.2 can be applied directly to establish the tightness of (C^γ, A^γ) and characterize the possible weak limits. Namely, the following lemma holds:

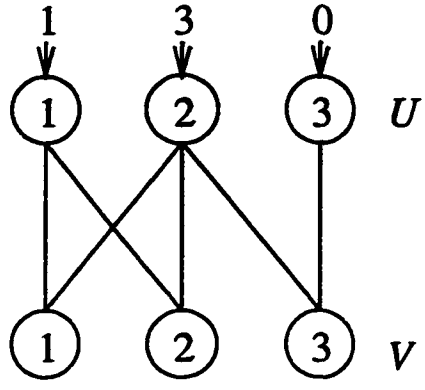
Lemma 2.5.1 *Suppose (C_0^γ) is tight. Then every subsequence $(C^{\gamma_n}, A^{\gamma_n})$ has a further subsequence $(C^{\gamma_{n_k}}, A^{\gamma_{n_k}})$ such that $(C^{\gamma_{n_k}}, A^{\gamma_{n_k}}) \Longrightarrow (c, A)$, where (c, A) satisfies the following fluid equations:*

$$c_t(u, v) = c_0(u, v) + A_{u,v}(t) + \int_0^t r_{u,u} c_s(u, v) ds \tag{2.35}$$

$$A_{u,v}(0) = 0, \ A_{u,v}(t) \text{ nondecreasing}, \ \sum_{v \in N(u)} A_{u,v}(t) = \lambda(u)t + \sum_{u' \neq u} \int_0^t l_s(u') r_{u',u} ds, \tag{2.36}$$

$$\int_0^t I\{x_s(v) > m_u(x_s)\} dA_{u,v}(s) = 0, \tag{2.37}$$

where $x_t(v) = \sum_{u \in N^{-1}(v)} c_t(u, v)$ and $l_t(u) = \sum_{v \in N(u)} c_t(u, v)$.



$$R = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & -7 \end{bmatrix}$$

Figure 2.2 An example to illustrate the nonuniqueness of the solutions to the fluid equations (2.35)-(2.37).

Table 2.1 Two assignment regimes for the network of Figure 2.2.

	$A_t(u,v)$			$c_t(u,v)$		
$u \setminus v$	1	2	3	1	2	3
1	$t/2$	$t/2$	0	$(1-e^{-t})/2$	$(1-e^{-t})/2$	0
2	$3t/2$	$3t/2$	0	$(1-e^{-3t})/2$	$(1-e^{-3t})/2$	0
3	0	0	0	0	0	$(.9495 \times 10^7)e^{-7t}$

	$\tilde{A}_t(u,v)$			$\tilde{c}_t(u,v)$		
$u \setminus v$	1	2	3	1	2	3
1	t	0	0	$1-e^{-t}$	0	0
2	$1-e^{-t}$	$3t-1+e^{-t}$	0	$(e^{-t}-e^{-3t})/2$	$1-(e^{-t}+e^{-3t})/2$	0
3	0	0	0	0	0	$(.9495 \times 10^7)e^{-7t}$

The following example shows, in contrast to Lemma 2.3.5, that the initial state c_0 and the fluid equations (2.35)-(2.37) do not necessarily determine a unique load trajectory x.

Example. (Type dependent departure rates, no routing) Consider the load sharing network and the routing matrix of Figure 2.2. Suppose $c_0(3,3) = .9495 \times 10^7$, and $c_0(u,v) = 0$ otherwise. Let $\tau = \ln 10$, and consider the two assignment regimes (c, A) and $(\tilde{c}, \tilde{A})$ for $t \in [0, \tau]$ as listed in Table 2.1. It is straightforward to verify that both (c, A) and $(\tilde{c}, \tilde{A})$ satisfy (2.35)-(2.37), and $x_t(v) = \tilde{x}_t(v)$ for $v \in V$ and $t \in [0, \tau]$. Note that $\tau = \inf\{t : x_t(1) = x_t(2) = x_t(3)\}$, and $x_\tau(1) = .9495$. Under both regimes, at time τ, the instantaneous rate of decrease of load at location 3 is 7(.9495), whereas the instantaneous rates of decrease of load at locations 1 and 2, respectively, are each upper bounded by 3(.9495). The difference is larger than the flow rate of type 2 arrivals; therefore, there exists a $\delta > 0$ such that (2.35)-(2.37) are not violated only if all type

2 arrivals are directed to location 3 during $[\tau, \tau+\delta]$. Under (c, A), type 1 arrivals can split evenly between locations 1 and 2 and maintain $x_t(1) = x_t(2)$ for $t \in [\tau, \tau+\delta]$. However, under $(\tilde{c}, \tilde{A})$, at time τ location 1 discharges at an instantaneous rate of $.9 + 3(.0495) = 1.0485$, and location 2 discharges at an instantaneous rate of $3(.9495) = 2.8485$. The difference between the discharge rates is greater than the flow rate of type 1 arrivals; therefore, there exists a $\delta' > 0$ such that all type 1 arrivals are directed to location 2, and $\tilde{x}_t(1) > \tilde{x}_t(2)$ for $t \in (\tau, \tau+\delta']$. Thus, c_0 together with the fluid equations (2.35)-(2.37) does not determine x uniquely. □

We now concentrate on the properties of the load trajectories corresponding to the solutions of the fluid equations. By the definition of the demand vector, the coordinates of ρ are strictly positive; however, the extension of the results to nonnegative ρ is trivial.

Lemma 2.5.2 *Given* $\bar{l}_0 \in R_+^U$, $\lim_{t\to\infty} \|l_t - \rho\|_{sup} = 0$ *uniformly for* $0 \le l_0 \le \bar{l}_0$.

Proof. By Equations (2.35)-(2.37), l_t satisfies

$$l_t = l_0 + \lambda t + \int_0^t R l_s ds,$$

which can be solved to yield

$$l_t = l_0 e^{Rt} + \rho(I - e^{Rt}),$$

where the exponential e^{Rt} of the matrix R can be defined by a power series. Since $e^{Rt} \to 0$ exponentially, l_t converges to ρ exponentially fast, uniformly for $l_0 \le \bar{l}_0$. This establishes the lemma. □

The following auxiliary lemma is proved in Section 2.7.

Lemma 2.5.3 *Suppose that* $a_i \le \bar{a}_i$ *for* $1 \le i \le J$, *and* $w_{min} = \min_{1\le i\le J} w_i$. *Then*

$$\sum_i a_i w_i \le \sum_i \bar{a}_i w_i + (\sum_i a_i - \sum_i \bar{a}_i) w_{min}.$$

Lemma 2.5.4 *Let* q *denote the unique load vector corresponding to the solutions of* $SLB(\rho, \Phi)$. *Given* $\bar{l}_0 \in R_+^U$ *and* $\epsilon > 0$, *there exists* $t_1(\bar{l}_0, \epsilon)$ *such that for all* $v \in V$,

$$x_t(v) \ge q(v) - \epsilon$$

whenever $t \geq t_1(\bar{l}_0, \epsilon)$, *and* $0 \leq l_0 \leq \bar{l}_0$.

Proof. Let $\{V_1, V_2, \cdots, V_J\}$ and $\{U_1, U_2, \cdots, U_J\}$ be the unique partitions of V and U, respectively, defined by Lemma 2.2.2 adapted to $SLB(\rho, \Phi)$. It is enough to show that

$$\inf_{v \in \bigcup_{i=j}^{J} V_i} x_t(v) \geq q_j - \epsilon \tag{2.38}$$

for all $j \in \{1, 2, \cdots, J\}$, and $t \geq t_1(\bar{l}_0, \epsilon)$, where q_j is the value such that $q(v) = q_j$ for all $v \in V_j$. Towards this end, for each $j \in \{1, 2, \cdots, J\}$ define

$$\begin{aligned}
m_t^j &= \inf_{v \in \bigcup_{i=j}^{J} V_i} x_t(v), \\
F_t^j &= \left\{ v \in \bigcup_{i=j}^{J} V_i \, : \, x_t(v) = m_t^j \right\}, \\
c_t(u, F_t^j) &= \sum_{v \in F_t^j} c_t(u, v), \\
N^*(F_t^j) &= \{u : N(u) \cap F_t^j \neq \emptyset \text{ and } N(u) \cap (\bigcup_{i=1}^{j-1} V_i) = \emptyset\},
\end{aligned}$$

with the understanding that $\bigcup_{i=1}^{0} V_i = \emptyset$.

Let $r_{min} = \min_u\{-r_{u,u}\}$ and $r_{max} = \max_u\{-r_{u,u}\}$. Given $\epsilon > 0$, let $\epsilon_0 < \epsilon r_{min}(2\sum_u \sum_{u'} |r_{u,u'}|)^{-1}$. Appeal to Lemma 2.5.2 to fix $t_0(\bar{l}_0, \epsilon_0)$ such that $\sup_u |l_t(u) - \rho(u)| < \epsilon_0$ for all $t \geq t_0(\bar{l}_0, \epsilon_0)$ whenever $0 \leq l_0 \leq \bar{l}_0$. To prove the lemma by induction on j, let $j = 1$, and choose $t > t_0(\bar{l}_0, \epsilon_0)$. Suppose that t is a regular point of m^1 and x, and that

$$m_t^1 < q_1 - \epsilon. \tag{2.39}$$

Then, by Lemma 2.4.8 and the fact that $N^*(F_t^1) = N^{-1}(F_t^1)$,

$$\begin{aligned}
|F_t^1| \dot{m}_t^1 &= \sum_{v \in F_t^1} \dot{x}_t(v) \\
&= \sum_{u \in N^*(F_t^1)} \lambda(u) + \sum_{u \in U} l_t(u) \sum_{u' \in N^*(F_t^1) \setminus \{u\}} r_{u,u'} - \sum_{u \in U} c_t(u, F_t^1)(-r_{u,u}) \\
&= \sum_{u \in N^*(F_t^1)} \lambda(u) + \sum_{u \in U} l_t(u) \sum_{u' \in N^*(F_t^1) \setminus \{u\}} r_{u,u'} - \sum_{u \in N^*(F_t^1)} c_t(u, F_t^1)(-r_{u,u}).
\end{aligned} \tag{2.40}$$

By the choice of t, $c_t(u, F_t^1) \leq \rho(u) + \epsilon_0$, so that Lemma 2.5.3 and (2.39) can be used to bound the third term on the right-hand side of (2.40) to obtain

$$\begin{aligned} |F_t^1|\dot{m}_t^1 \geq & \sum_{u \in N^*(F_t^1)} \lambda(u) + \sum_{u \in U} l_t(u) \sum_{u' \in N^*(F_t^1)\setminus\{u\}} r_{u,u'} - \sum_{u \in N^*(F_t^1)} (\rho(u) + \epsilon_0)(-r_{u,u}) \\ & - \left(|F_t^1|(q_1 - \epsilon) - \sum_{u \in N^*(F_t^1)} (\rho(u) + \epsilon_0) \right) r_{min}. \end{aligned}$$

Note that $\sum_{u \in N^*(F_t^1)} \rho(u) \geq |F_t^1| q_1$, and $l_t(u) \geq \rho(u) - \epsilon_0$. This, together with the identity $\rho(u)(-r_{u,u}) = \lambda(u) + \sum_{u' \neq u} \rho(u') r_{u',u}$ and the choice of ϵ_0, yields that

$$\begin{aligned} |F_t^1|\dot{m}_t^1 &\geq -\epsilon_0 \left(\sum_{u \in U} \sum_{u' \in N^*(F_t^1)\setminus\{u\}} r_{u,u'} + \sum_{u \in N^*(F_t^1)} (-r_{u,u}) \right) + \epsilon |F_t^1| r_{min} \\ &\geq \frac{\epsilon r_{min} |F_t^1|}{2}. \end{aligned}$$

Thus, for almost all regular points t of m such that $t > t_0(\bar{l}_0, \epsilon_0)$, $\dot{m}_t^1 \geq \epsilon r_{min}/2$ whenever $m_t^1 < q_1 - \epsilon$. Therefore, (2.38) holds for $j = 1$ for all $t \geq T^1(\bar{l}_0, \epsilon) = t_0(\bar{l}_0, \epsilon_0) + 2\sum_u (\rho(u) + \epsilon_0)/\epsilon r_{min}$.

As the induction hypothesis, fix $j \in \{2, 3, \cdots, J\}$, and suppose that given $\epsilon_0 > 0$, for each $p \in \{1, \cdots, j-1\}$ there exists a $T^p(\bar{l}_0, \epsilon_0)$ such that

$$m_t^p \geq q_p - \epsilon_0$$

whenever $t \geq T^p(\bar{l}_0, \epsilon_0)$.

Let $\epsilon_0 < \min\{\epsilon r_{min}(4(|U|+|V|)r_{max})^{-1}, \epsilon r_{min}(4\sum_u \sum_{u'} |r_{u,u'}|)^{-1}\}$, and choose $t > \max\{t_0(\bar{l}_0, \epsilon_0), T^1(\bar{l}_0, \epsilon_0), \cdots, T^{j-1}(\bar{l}_0, \epsilon_0)\}$. Suppose that t is a regular point of m^j and x, and $m_t^j < q_j - \epsilon$. Note that the methods used in the case $j = 1$ imply that

$$\begin{aligned} |F_t^j|\dot{m}_t^j &= \sum_{v \in F_t^j} \dot{x}_t(v) \\ &\geq \sum_{u \in N^*(F_t^j)} \lambda(u) + \sum_u l_t(u) \sum_{u' \in N^*(F_t^j)\setminus\{u\}} r_{u,u'} \\ &- \sum_{u \in N^*(F_t^j)} c_t(u, F_t^j)(-r_{u,u}) - \sum_{u \in N^*(F_t^j)^c} c_t(u, F_t^j)(-r_{u,u}) \end{aligned}$$

$$\geq -\epsilon_0 \left(\sum_u \sum_{u' \in N^*(F_t^1)\setminus\{u\}} r_{u,u'} + \sum_{u \in N^*(F_t^1)} (-r_{u,u}) \right) + \epsilon |F_t^1| r_{min} - r_{max} \sum_{u \in N^*(F_t^j)^c} c_t(u, F_t^j).$$

By (2.5), the choice of t, and the induction hypothesis,

$$\begin{aligned} \sum_{u \in N^*(F_t^j)^c} c_t(u, F_t^j) &< \sum_{u \in \bigcup_{i=1}^{j-1} U_i} (\rho(u) + \epsilon_0) - \sum_{v \in \bigcup_{i=1}^{j-1} V_i} (q_v - \epsilon_0) \\ &\leq \frac{\epsilon r_{min}}{4 r_{max}}. \end{aligned}$$

Hence by the choice of ϵ_0, $|F_t^j| \hat{m}_t^j \geq \epsilon r_{min} |F_t^j|/2$. Thus, there exists a $T^j(\bar{l}_0, \epsilon)$ such that $m_t^j \geq q_j - \epsilon$ for all $t \geq T^j(\bar{l}_0, \epsilon)$. This completes the induction step and the proof of the lemma with $t_1(\bar{l}_0, \epsilon) = T^J(\bar{l}_0, \epsilon)$. □

Lemma 2.5.5 *(Global Convergence) Let x be the load function corresponding to an arbitrary solution of the fluid equations (2.35)-(2.37), and q be the unique load vector corresponding to the solutions of $SLB(\rho, \Phi)$. Then $\lim_{t\to\infty} \|x_t - q\|_{sup} = 0$ uniformly for all l_0 in bounded subsets of R_+^U.*

Proof. Fix $\bar{l}_0 \in R_+^U$, and let $l_0 \leq \bar{l}_0$. Given $\epsilon > 0$, set $\epsilon_0 = \epsilon/(|V| + 1)$, and appealing to Lemma 2.5.2, let $t_0(\bar{l}_0, \epsilon_0)$ be such that $|\sum_{v \in V}(x_t(v) - q(v))| < \epsilon_0$ for all $t \geq t_0(\bar{l}_0, \epsilon_0)$. If $t > t_0(\bar{l}_0, \epsilon_0) \vee t_1(\bar{l}_0, \epsilon_0)$, then Lemma 2.5.4 implies that

$$\inf_{v \in V}(x_t(v) - q(v)) \geq -\epsilon_0 \geq -\epsilon,$$

which in turn implies

$$\begin{aligned} \sup_{v \in V}(x_t(v) - q(v)) &\leq \sum_{v \in V}(x_t(v) - q(v) + \epsilon_0) \\ &= \sum_{v \in V}(x_t(v) - q(v)) + |V|\epsilon_0 \\ &\leq (|V| + 1)\epsilon_0 = \epsilon. \end{aligned}$$

This proves the desired result. □

Let P_μ denote the distribution of the process C^γ when C_0^γ has distribution μ and μ_0 denote the deterministic distribution concentrated at the zero state. The proof of the following lemma is immediate:

Lemma 2.5.6 *Given $\epsilon > 0$, there exists a γ_ϵ such that whenever $\gamma > \gamma_\epsilon$,*

$$P_{\mu_0}\left(\sum_v X_t^\gamma(v) \leq \epsilon + \sum_u \rho(u)\right) \geq 1 - \epsilon \quad \textit{for any } t \geq 0.$$

Lemma 2.5.7 *(Convergence of Equilibrium Distributions) Let q be the unique load vector corresponding to solutions of $SLB(\rho, \Phi)$ and ν be the distribution of the equilibrium load X_∞^γ. Then for all $\epsilon > 0$*

$$\lim_{\gamma \to \infty} \nu\left(\|X_\infty^\gamma - q\|_{sup} > \epsilon\right) = 0.$$

Proof. The proof of Lemma 2.3.9 applies directly by redefining μ_t^γ as the distribution of C_t^γ and by using Lemmas 2.5.1, 2.5.5, and 2.5.6 in place of Lemmas 2.3.3, 2.3.7, and 2.3.8, respectively. □

Theorem 2.5.1 *(Asymptotic Optimality of LLR) Given an allocation policy π, let J_γ^π denote the cost under π when the network demand is $\gamma\lambda$. Then*

$$\liminf_{\gamma \to \infty} \gamma^{-2} J_\gamma^\pi \geq \lim_{\gamma \to \infty} \gamma^{-2} J_\gamma^{LLR} > 0.$$

Proof. The proof of Lemma 2.3.1 applies directly by using Lemma 2.5.7 in place of Lemma 2.3.9 and ρ in place of λ. □

2.6 Conclusions and Discussion

This chapter concentrates on the dynamic load balancing problem and studies the performance of practical allocation policies, namely LLR and the class LRLR. When there are no capacity constraints on the resources, LLR is shown to achieve asymptotically the most balanced load in the

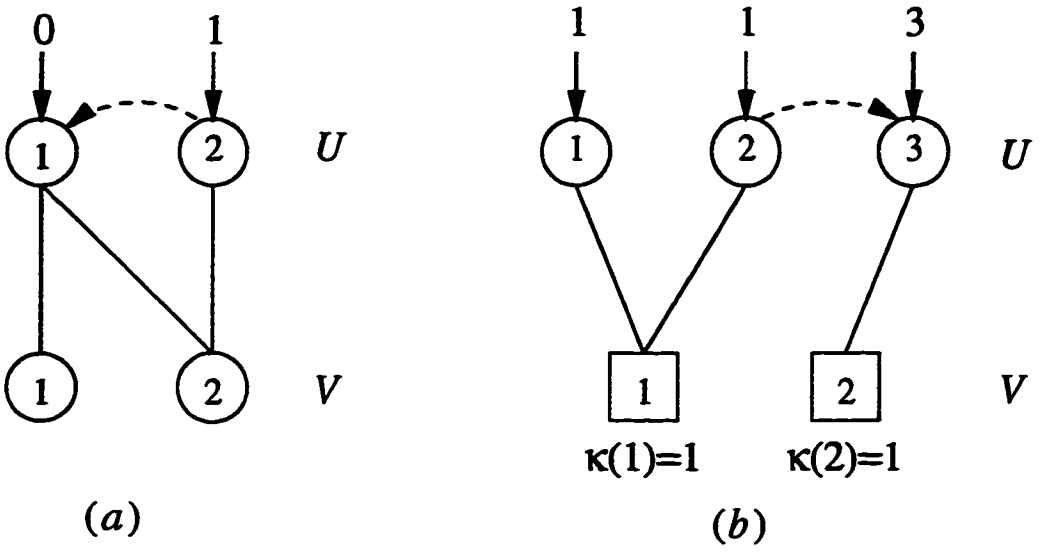


Figure 2.3 The counter examples for the optimality of (*a*) sticky LLR, (*b*) LRLR under finite capacities and migration.

sense of minimizing a wide class of long-term average costs. LLR is also robust to migration, provided that consumers are reassigned according to LLR whenever their types change. On the other hand, when the resources have finite capacities, LRLR policies asymptotically achieve the minimum possible loss probability. The desirable aspects of the considered policies are low computational complexity, decentralized implementation, and robustness to arrival and migration rates.

The reassignment of migrating consumers is important for the asymptotic optimality of LLR in the migration model. The network of Figure 2.3(a) is an example in which LLR is not optimal without reassignments. Let $\lambda = (0, 1)$ and the routing matrix be

$$R = \begin{bmatrix} -1 & 0 \\ 1 & -1 \end{bmatrix}.$$

Hence all consumers arrive as type 2 and migrate to become type 1 before leaving the system. Suppose that consumers are assigned using LLR upon arrival; however, they maintain their original locations even though they migrate. Then at any time t, all of the load in the network is at location 2; hence the limiting normalized cost of this policy is 4. A simple calculation yields that the LLR policy splits the load equally between the two locations, thus having a limiting normalized cost of 2.

Optimality properties of the policies discussed in this chapter do not necessarily persist in the case of finite capacities *and* migration. In particular, myopic policies, which accept a consumer whenever possible, may not be asymptotically optimal. As an example to illustrate this, consider the network of Figure 2.3(b) in heavy traffic. Let $\lambda = (1, 1, 3)$, $\kappa = (1, 1)$, and the routing matrix be

$$R = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}.$$

Hence type 2 arrivals first visit location 1 and then location 2 before exiting the system. Without loss of generality, assume that location 2 gives higher priority to exogenous arrivals, in the sense that an exogenous arrival blocks a migrated consumer that is already in location 2, provided that location 2 is full at the time of arrival. Since exogenous arrivals suffice to overload location 2, all type 2 arrivals are bound to be lost. Any myopic policy blocks half of type 1 arrivals and has a limiting consumer loss probability of 0.7. On the other hand, a policy that blocks type 2 arrivals regardless of the system state has a limiting consumer loss probability of 0.6. We therefore conclude that the optimal policies have considerably more complex structures under the more general case.

There are other asymptotically optimal policies that are not of repacking type. In particular, let the assignment a be a solution to $SLB(\lambda, \Phi)$, and consider the *Bernoulli splitting* (BS) policy under which each arriving type $u \in U$ consumer is assigned to location $v \in N(u)$ with probability $a_{u,v}/\lambda(u)$ independently of previous events. Under this policy and the basic model of Section 2.3, the equilibrium load at each location $v \in V$ is a Poisson random variable with mean $\gamma q(v)$, where q denotes the load vector corresponding to the assignment a. It is therefore easy to see that the BS policy achieves the optimal normalized cost in the limiting regime of interest. Similarly, if a is a solution to problem $SLP(\lambda, \kappa)$, then the BS policy asymptotically achieves the minimum loss probability in the case of finite capacity constraints. However, this policy explicitly uses the traffic parameters; therefore, it is not robust with respect to the network demand. Furthermore, it exerts only open-loop control; hence one expects a finer analysis to reveal the superiority of LLR to BS. This suggests a large deviations analysis to compare the policies that are optimal in the fluid scale, which is the approach taken in the following chapter.

2.7 Proofs of Lemmas

This section contains the deferred proofs from previous sections. We start with the proof of Lemma 2.2.1 by first giving an auxiliary result.

Lemma 2.7.1 *Let $\Phi : R^d \to R$ be strictly convex and differentiable, and Φ_v denote the v^{th} partial derivative of Φ. If Φ is symmetric in its arguments (i.e., $\Phi(x(1), \cdots, x(d)) = \Phi(x(p(1)), \cdots, x(p(d)))$ for any permutation p), then for all v, v'*

$$x(v) > x(v') \Longrightarrow \Phi_v(x) > \Phi_{v'}(x).$$

Proof. Since the conclusion involves only two arguments, we can assume without loss of generality that $d = 2$. For $(a, b) \in R^2$, define $g_{a,b}(\alpha) = \Phi(\alpha(a, b) + (1 - \alpha)(b, a))$. Then, $\dot{g}_{a,b}(\alpha) = (\Phi_1 - \Phi_2)(a - b)$, where the partial derivatives Φ_1 and Φ_2 are evaluated at $\alpha(a, b) + (1 - \alpha)(b, a)$. Note that by the strict convexity of Φ, $g_{a,b}$ is strictly convex for $a \neq b$. Also by the symmetry of Φ, $\dot{g}_{a,b}(\alpha)|_{\alpha=\frac{1}{2}} = 0$; therefore, for $a \neq b$, $\dot{g}_{a,b}(\alpha)|_{\alpha=1} > 0$. This implies $(\Phi_1(a, b) - \Phi_2(a, b))(a - b) > 0$, which proves the claim. □

Proof of Lemma 2.2.1. The problem $SLB(\lambda, \Phi)$ is a convex optimization problem on a compact and convex set; thus, there exists a solution.

To establish the second statement of the lemma, we argue by contradiction in each direction. In what follows, $\psi(a)$ denotes the value $\Phi(q)$ induced by the assignment a. First, let a satisfy (2.1), and suppose that a is not a solution to $SLB(\lambda, \Phi)$. Then there exists an admissible perturbation vector h such that

$$\sum_{v \in N(u)} h_{u,v} = 0 \quad \text{for all } u \in U \tag{2.41}$$

$$h_{u,v} \geq 0 \text{ whenever } a_{u,v} = 0, \quad h_{u,v} = 0 \text{ whenever } v \in N(u)^c \tag{2.42}$$

$$\lim_{\epsilon \searrow 0} \frac{\psi(a + \epsilon h) - \psi(a)}{\epsilon} = \sum_u \sum_{v \in N(u)} h_{u,v} \Phi_v(q) < 0. \tag{2.43}$$

By (2.42) we have that for all u and $v \in N(u)$

$$\begin{aligned} h_{u,v} < 0 &\Longrightarrow a_{u,v} > 0 \\ &\Longrightarrow q(v) \leq q(v') \quad \text{for all } v' \in N(u) \\ &\Longrightarrow \Phi_v(q) \leq \Phi_{v'}(q) \quad \text{for all } v' \in N(u) \end{aligned} \tag{2.44}$$

by using Lemma 2.7.1 in the last step. Now for $u \in U$ define

$$\Phi_u^-(q) = \max(\Phi_v(q) : h_{u,v} < 0) \quad , \; h_u^- = \textstyle\sum_{v:h_{u,v}<0} h_{u,v},$$
$$\Phi_u^+(q) = \min(\Phi_v(q) : h_{u,v} > 0) \quad , \; h_u^+ = \textstyle\sum_{v:h_{u,v}>0} h_{u,v}.$$

Then by (2.41) $h_u^+ = -h_u^-$, and by (2.44) $\Phi_u^+(q) \geq \Phi_u^-(q)$, and therefore for all $u \in U$,

$$\sum_u \sum_{v \in N(u)} h_{u,v}\Phi_v(q) \geq \sum_u h_u^+(\Phi_u^+(q) - \Phi_u^-(q)) \geq 0,$$

which contradicts (2.43).

To show that the converse also holds, suppose that a does not satisfy the condition (2.1). In particular, let u be such that for some $v, v' \in N(u)$,

$$a_{u,v} > 0 \quad \text{and} \quad q(v) > q(v').$$

Then by Lemma 2.7.1, $\Phi_v(q) - \Phi_{v'}(q) > 0$; thus, there exists a δ small enough such that it is possible to decrease $a_{u,v}$ and increase $a_{u,v'}$ by an amount δ without violating the constraints of $SLB(\lambda, \Phi)$ and obtain a smaller value for Φ. Therefore, a cannot be a solution.

Finally, by the strict convexity of Φ, there is a unique load vector corresponding to the solutions of $SLB(\lambda, \Phi)$. □

Proof of Lemma 2.2.2. It is straightforward to form the partition $\{V_1, V_2, \cdots, V_J\}$ that satisfies (2.2) and (2.3). This partition is unique since q is the same for all assignments a satisfying (2.1). Let a be an arbitrary assignment satisfying (2.1), and define the set of subsets $\{U_1, U_2, \cdots, U_J\}$ of U as

$$U_i = \{u \in U : a(u, v) > 0 \text{ for some } v \text{ in } V_i\}.$$

By (2.1), (2.4) and (2.5) hold; therefore, $\{U_1, U_2, \cdots, U_J\}$ is a partition of U.

It remains to show that *any* assignment satisfying (2.1) yields the same $\{U_1, U_2, \cdots, U_J\}$. Suppose $\tilde{a}$ is another such assignment which yields $\{\tilde{U}_1, \tilde{U}_2, \cdots, \tilde{U}_J\}$. To prove the claim by contradiction, assume that there exists a $u \in U$ such that $u \in U_j \cap \tilde{U}_i$, where $i < j$. Since $u \in U_j$, by (2.5) $N(u) \cap V_i = \phi$. Therefore, $\tilde{a}(u, v) = 0$ for all $v \in V_i$. This contradicts the assumption that $u \in \tilde{U}_i$ and establishes the uniqueness of $\{U_1, U_2, \cdots, U_J\}$. □

Proof of Lemma 2.2.3. For $i = 1, 2$, let $\lambda_i \in R_+^U$, and a_i solve $SLB(\lambda_i, \Phi)$ with the corresponding load vector q_i. Then for any $\alpha \in [0, 1]$,

$$\begin{aligned} \alpha\Psi(\lambda_1) + (1-\alpha)\Psi(\lambda_2) &= \alpha\Phi(q_1) + (1-\alpha)\Phi(q_2) \\ &\geq \Phi(\alpha q_1 + (1-\alpha)q_2) \\ &\geq \Psi(\alpha\lambda_1 + (1-\alpha)\lambda_2). \end{aligned}$$

The second step follows by the convexity of Φ. The third step follows by the definition of Ψ and the fact that $\alpha a_1 + (1-\alpha)a_2$ is an admissible assignment that satisfies $\alpha\lambda_1 + (1-\alpha)\lambda_2$ with the load vector $\alpha q_1 + (1-\alpha)q_2$. □

Proof of Lemma 2.4.1. The proof of Lemma 2.2.1 applies by replacing $q(v)$ with $f(q, v)$ and by noting that (2.44) follows by the definition of Φ. □

Proof of Lemma 2.4.2. It is straightforward to see that $a \in \mathcal{B}_{\lambda,\kappa}$ and that q is the load vector corresponding to a. To show that a satisfies Condition 2.4.1, suppose that $f(q, v) > m_u(f(q))$ for some u, v with $v \in N(u)$. Then by Lemma 2.4.1 there exists a $v' \in N(u)$ such that $f(q, v') < 0$. Since $q(v') = \tilde{q}(v')$ for all v' with $f(\tilde{q}, v') < 0$, $m_u(f(q)) = m_u(f(\tilde{q}))$, and therefore,

$$f(\tilde{q}, v) \geq f(q, v) > m_u(f(q)) = m_u(f(\tilde{q})). \tag{2.45}$$

Since $\tilde{a}_{u,v}$ satisfies Condition 2.4.1, (2.45) implies that $\tilde{a}_{u,v} = 0$; thus, $a_{u,v} = 0$, and Condition 2.4.1 is satisfied.

To show that Condition 2.4.2 is satisfied, note that if $\sum_v a_{u,v} < \lambda(u)$, then there exists a $v \in N(u)$ such that $\tilde{a}_{u,v} > 0$ and $\tilde{q}(v) > \kappa(v)$. This implies that $f(\tilde{q}, v') > 0$ for all $v' \in N(u)$ and hence that $f(q, v') = 0$ for all $v' \in N(u)$, establishing Condition 2.4.2. □

Proof of Lemma 2.4.3. Let f be a normalization function and a be an assignment satisfying Conditions 2.4.1 and 2.4.2 with f and the corresponding load vector q. To prove the optimality of a, argue by contradiction. If a is not a solution to $SLP(\lambda, \kappa)$, then there exists a perturbation vector h such that

$$\begin{aligned} h_{u,v} &= 0 \quad v \in N(u)^c, \\ h_{u,v} &\geq 0 \quad \text{if } a_{u,v} = 0, \end{aligned} \tag{2.46}$$

$$\sum_u h_{u,v} \leq \kappa(v) - q(v), \tag{2.47}$$

$$\sum_v h_{u,v} \leq \lambda(u) - \sum_v a_{u,v}, \tag{2.48}$$

$$\sum_u \sum_v h_{u,v} > 0. \tag{2.49}$$

We use induction to arrive at the desired contradiction. Let

$$\begin{aligned} U_0 &= \{u : \sum_v h_{u,v} > 0\}, \\ U_{j+1} &= U_j \cup \{u \notin U_j : h_{u,v} < 0 \text{ for some } v \in N(U_j)\}, \ j \geq 0. \end{aligned}$$

By inequality (2.49), U_0 is nonempty. Inequality (2.48), Condition 2.4.2, and the Definition 2.4.1 of the normalization function imply that

$$q(v) = \kappa(v) \text{ for all } v \in N(U_0). \tag{2.50}$$

By (2.47), $\sum_{v \in N(U_0)} \sum_u h_{u,v} \leq 0$, and by the definition of U_0, $\sum_{u \in U_0} \sum_{v \in N(U_0)} h_{u,v} > 0$; therefore, $U_1 \neq U_0$. If $u \in U_1 \setminus U_0$, then inequality (2.46) implies that $a_{u,v} > 0$ for some $v \in N(U_0)$. By Condition 2.4.1, $f(q,v) = 0$ for all $v \in N(u)$. This, along with (2.50), implies that

$$q(v) = \kappa(v) \text{ for all } v \in N(U_1).$$

As the induction hypothesis, assume that $q(v) = \kappa(v)$ for all $v \in N(U_k)$. By the definition of $(U_j : j \geq 0)$, $\sum_{u \in U_{k+1}} \sum_{v \in N(U_k)} h_{u,v} > 0$; therefore, the argument of the base case yields that

$$U_{k+1} \neq U_k \quad \text{and} \quad q(v) = \kappa(v) \quad \text{for all } v \in N(U_{k+1}).$$

This contradicts the finiteness of the network, and hence the existence of h, proving the optimality of a.

Lemma 2.4.2 establishes the existence of an assignment $a \in \mathcal{B}_{\lambda,\kappa}$ that satisfies Conditions 2.4.1 and 2.4.2. To prove the uniqueness of the load vector by contradiction, let a and $\tilde{a}$ be two such assignments with the corresponding load vectors q and $\tilde{q}$ such that $q \neq \tilde{q}$. Let

$$F = \{v : q(v) > \tilde{q}(v)\}.$$

If $v \in F$, then for any u,

$$\begin{aligned} a_{u,v} > 0 &\Rightarrow f(q,v) \leq f(q,v') \quad \text{for all } v' \in F^c \cap N(u) \\ &\Rightarrow f(\tilde{q},v) < f(\tilde{q},v') \quad \text{for all } v' \in F^c \cap N(u) \\ &\Rightarrow \sum_{v \in F} \tilde{a}_{u,v} = \lambda(u), \end{aligned}$$

where the second step follows by the strict monotonicity of f, and the third step follows by Condition 2.4.1. However, this implies that $\sum_{v \in F}(\tilde{a}_{u,v} - a_{u,v}) \geq 0$ for any $u \in U$, which contradicts the definition of F and proves the desired result. □

Proof of Lemma 2.4.6. By the absolute continuity of g, $\dot{g}$ exists almost everywhere ([11, Corollary 5.12]). Thus, it suffices to prove that the set $\{t : \dot{g}_t \text{ exists}, g_t = \alpha, \dot{g}_t \neq 0\}$ has Lebesgue measure zero. However, all of the points in this set are isolated; hence the set contains at most countably infinite elements and therefore has zero measure. □

Proof of Lemma 2.4.8. Since $g_t(i) \quad i = 1, 2, \cdots, I$ are absolutely continuous, so is m. Therefore, $g(1), \cdots, g(I), m$ are almost everywhere differentiable. Let t be a regular point of $g(1), \cdots, g(I), m$ and $\{i_1, \cdots, i_r\}$ be such that $g_t(i_1) = \cdots = g_t(i_r) = m_t$. Note that

$$\begin{aligned} \max_{1 \leq k \leq r} \dot{g}_t(i_k) &= \max_{1 \leq k \leq r} \lim_{\epsilon \searrow 0} \frac{g_t(i_k) - g_{t-\epsilon}(i_k)}{\epsilon} \\ &\leq \liminf_{\epsilon \searrow 0} \max_{1 \leq k \leq r} \frac{g_t(i_k) - g_{t-\epsilon}(i_k)}{\epsilon} \\ &\leq \liminf_{\epsilon \searrow 0} \frac{m_t - m_{t-\epsilon}}{\epsilon} \\ &= \dot{m}_t. \end{aligned}$$

Similarly,

$$\min_{1 \leq k \leq r} \dot{g}_t(i_r) \geq \limsup_{\epsilon \searrow 0} \frac{m_{t+\epsilon} - m_t}{\epsilon} = \dot{m}_t,$$

and the proof of the lemma is complete. □

Proof of Lemma 2.5.3.

$$\sum_i a_i w_i = \sum_i \bar{a}_i w_i + \sum_i (a_i - \bar{a}_i) w_i$$

$$\leq \sum_i \bar{a}_i w_i + \sum_i (a_i - \bar{a}_i) w_{min},$$

since $a_i \leq \bar{a}_i$ for all i. □

CHAPTER 3

ANALYSIS OF OVERFLOW

3.1 Introduction

This chapter concerns load sharing networks in the basic stochastic model of Section 2.3 and analyzes network overflow under the optimal repacking (OR), Bernoulli splitting (BS), and least load routing (LLR) allocation policies (see Sections 2.3 and 2.6 for definitions of OR and BS policies, respectively). Namely, given a load sharing network (U, V, N), a demand vector λ, and a positive number γ, it is assumed that consumers of type $u \in U$ arrive according to a Poisson process of rate $\gamma\lambda(u)$, the processes for different types of arrivals being independent. The holding time of each consumer is exponentially distributed with unit mean, independent of the past history. Given $\kappa \geq 0$, the *overflow time of a location* $v \in V$ is the first time that its load, $X_t(v)$, exceeds the designated capacity $\lfloor \gamma\kappa \rfloor$, and the *network overflow time* is the minimum over all v of the overflow times of location v.

Under allocation policy π, the *overflow exponent of the network*, $F^\pi(\kappa)$, and the *overflow exponent of a location* v, $F^\pi(v, \kappa)$, are defined as

$$\begin{aligned} F^\pi(\kappa) &= -\lim_{T\to\infty} \lim_{\gamma\to\infty} \gamma^{-1} \log P(\text{ Network overflow time} \leq T | X_0 = 0) \\ F^\pi(v,\kappa) &= -\lim_{T\to\infty} \lim_{\gamma\to\infty} \gamma^{-1} \log P(\text{ Overflow time of location } v \leq T | X_0 = 0). \end{aligned}$$

It is easy to see that $F^\pi(\kappa) = \min_{v\in V} F^\pi(v,\kappa)$ whenever the above quantities exist. A crude interpretation of the overflow exponent of the network is that for fixed but large T, $P($ Network overflow time $\leq T | X_0 = 0) \approx \exp(-\gamma F^\pi(\kappa))$. Note that larger values of the overflow exponent indicate larger overflow times. The approach taken in this chapter is to compare allocation policies based on the corresponding overflow exponents. The rest of this section states the main results. We start with the two basic networks of Figure 1.2.

The Single-Location Network. The *single-location network* of Figure 1.2(a) has been studied extensively in the context of Erlang's model for circuit switched traffic. In particular, the following theorem can be obtained by applying the results in [12, Section 12]. Details of the proof are given in Section 3.3, since the basic notation and concepts carry over to analysis of more general network topologies.

Theorem 3.1.1 *(Single Location) The overflow exponent of the single-location network exists and is given by* $H_{\lambda(1)}(0,\kappa)$, *where*

$$H_{\lambda(1)}(x,y) = \int_x^y \left(\log(\frac{z}{\lambda(1)})\right)_+ dz, \qquad y \geq x \geq 0.$$

Intuitively, for $x < y$, $H_{\lambda(1)}(x,y)$ is a measure of how unlikely it is for the normalized load, starting at level x, to reach level y within a fixed, long time interval. Note that $H_{\lambda(1)}(x,y) = 0$ for $0 \leq x < y \leq \lambda(1)$, since such transition is not a rare event in this case. The reader is referred to [12, Section 12] for large deviations exponents for transitions within *fixed* time duration. For simplicity, we concentrate here on long time intervals.

The W Network. In the *W network* of Figure 1.2(b) it is assumed without loss of generality that $\lambda(1) \geq \lambda(3)$. We first discuss two upper bounds on the network overflow time that apply to *any* allocation policy and then provide three theorems that identify the overflow exponents under the policies of interest. The proofs of the theorems are the subjects of subsequent sections.

Stochastic ordering arguments provide the two upper bounds on the overflow time of the W network under any allocation policy: *1) The Single-Location Bound:* The load at location 1 is stochastically larger than the load of a single-location network with demand $\gamma\lambda(1)$. Hence the overflow time of a single-location network with capacity $\lfloor\gamma\kappa\rfloor$ and demand $\gamma\lambda(1)$ dominates the overflow time of location 1, which in turn dominates the overflow time of the network. *2) Pooling Bound:* The network necessarily overflows if the total load exceeds $\lfloor 2\gamma\kappa\rfloor$. Thus, the overflow time of the network is dominated by the overflow time of a single-location network with capacity $\lfloor 2\gamma\kappa\rfloor$ and demand $\gamma(\lambda(1)+\lambda(2)+\lambda(3))$.

We now discuss the three policies, starting with some essential definitions. For real x, a, and b such that $a \leq b$, let $[x]_a^b$ denote the number in the interval $[a,b]$ that is closest to x. Let $q(1) = \lambda(1) + p\lambda(2)$ and $q(2) = \lambda(3) + (1-p)\lambda(2)$, where p is chosen to minimize $|q(1) - q(2)|$.

More explicitly,

$$q(1) = [(\lambda(1) + \lambda(2) + \lambda(3))/2]_{\lambda(1)}^{\lambda(1)+\lambda(2)} \quad \text{and} \quad q(2) = [(\lambda(1) + \lambda(2) + \lambda(3))/2]_{\lambda(3)}^{\lambda(3)+\lambda(2)},$$

and $p = [(\lambda(3) - \lambda(1) + \lambda(2))/2\lambda(2)]_0^1$. The assumption $\lambda(1) \geq \lambda(3)$ implies that $q(1) \geq q(2)$.

Consider first the BS policy under which each type 2 consumer is assigned to location 1 with probability p or to location 2 with probability $(1-p)$. The load at each location v behaves as in a single-location network with demand $\gamma q(v)$, independent of the other location. In turn the overflow exponent of each location can be obtained by appealing to the single-location result, and the overflow exponent of the network is equal to that of the more heavily loaded location 1.

Theorem 3.1.2 *(BS) For $v = 1, 2$, the overflow exponent of location v under the BS policy exists and is given by $F^{BS}(v, \kappa) = H_{q(v)}(0, \kappa)$. In particular, $F^{BS}(\kappa) = H_{q(1)}(0, \kappa)$.*

As for the BS policy, the network load under the OR policy can be represented in terms of single-location loads. Under the OR policy, overflow of location v implies that either the number of type $2v-1$ consumers exceeds $\lfloor \gamma\kappa \rfloor$, or the total number of consumers exceeds $2\lfloor \gamma\kappa \rfloor$. The following theorem holds:

Theorem 3.1.3 *(OR) For $v = 1, 2$, the overflow exponent of location v under the OR policy exists and is given by $F^{OR}(v, \kappa) = H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2\kappa) \wedge H_{\lambda(2v-1)}(0, \kappa)$. In particular,*

$$F^{OR}(\kappa) = F^{OR}(1, \kappa) = \begin{cases} H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2\kappa) & \text{if } \kappa \leq \kappa_o \\ H_{\lambda(1)}(0, \kappa) & \text{if } \kappa > \kappa_o, \end{cases}$$

where κ_o is the larger root of $\kappa_o = \kappa_o \log(\kappa_o \lambda(1)/q(1)q(2)) + \lambda(2) + \lambda(3)$.

The overflow exponents under the LLR policy are identified by the following theorem, for which we provide somewhat detailed comments. For simplicity, it is assumed that if the locations are equally loaded, an arriving type 2 consumer is assigned to location 1.

Theorem 3.1.4 *(LLR) For $v = 1, 2$, the overflow exponent of location v under the LLR policy exists and is given by*

$$F^{LLR}(v, \kappa) = \begin{cases} H_{q(1)}(0, \kappa) + H_{q(2)}(0, \kappa) & \text{if } \kappa \leq \kappa_*(v) \\ H_{q(1)}(0, \kappa_*(v)) + H_{q(2)}(0, \kappa_*(v)) + H_{\lambda(2v-1)}(\kappa_*(v), \kappa) & \text{if } \kappa > \kappa_*(v), \end{cases}$$

where $\kappa_*(v) = q(1)q(2)/\lambda(2v-1)$. *In particular,* $F^{LLR}(\kappa) = F^{LLR}(1,\kappa)$, *which can also be expressed as*

$$F^{LLR}(\kappa) = \begin{cases} H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2\kappa) & \text{if } \kappa \leq \kappa_*(1) \\ H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2\kappa_*(1)) + H_{\lambda(1)}(\kappa_*(1), \kappa) & \text{if } \kappa > \kappa_*(1). \end{cases}$$

Remark 3.1.1 *To see the equivalence of the two expressions for* $F^{LLR}(\kappa)$, *note that* (*i*) *if* $q(1) = q(2)$, *then* $H_{q(1)}(0,\kappa) + H_{q(2)}(0,\kappa) = H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2\kappa)$ *for all* κ *and* (*ii*) *if* $q(1) > q(2)$, *then* $\lambda(1) = q(1) > q(2) = \kappa_*(1)$; *hence* $H_{q(1)}(0,\kappa) + H_{q(2)}(0,\kappa) = H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2\kappa) = 0$ *whenever* $\kappa \leq \kappa_*(1)$, *so that* $F^{LLR}(\kappa) = F^{LLR}(1,\kappa) = H_{\lambda(1)}(0,\kappa)$ *for all* κ.

Here we give an intuitive explanation for the formulas appearing in Theorem 3.1.4. In this paragraph and in the following paragraph, the "load" at a location is understood to be the normalized load for some suitably large value of γ. Focusing first on location $v = 1$, refer to Figure 3.1, which pictures the extremal trajectories associated with Theorem 3.1.4 for $v = 1$. Consider first the case $q(1) = q(2)$. If $\kappa \leq q(1) = q(2)$, then $F^{LLR}(1,\kappa) = 0$, which is expected since overflow of location 1 is not a rare event for such κ. If $q(1) = q(2) < \kappa \leq \kappa_*(1)$, then overflow in location 1 typically occurs because the whole network becomes overloaded, and both locations maintain roughly equal loads. For larger values of κ, the most likely scenario is that first the loads at the two locations together build up to level $\kappa_*(1)$, and then the load at location 1 continues to grow to level κ. The given value of $\kappa_*(1)$ minimizes the expression for $F^{LLR}(1,\kappa)$. Finally, consider the case in which $q(1) > q(2)$. Then $F^{LLR}(1,\kappa) = H_{\lambda(1)}(0,\kappa)$ as explained in Remark 3.1.1, and the typical scenario for overflow of location 1 is that the load at location 1 reaches κ, while the load at location 2 relaxes towards its mean $q(2)$.

Now focusing on location 2, let us give an intuitive explanation for the expression for the overflow exponent $F^{LLR}(2,\kappa)$. Consult Figure 3.2. If $q(1) = q(2)$, the explanation is similar to that for $F^{LLR}(1,\kappa)$, so assume that $q(1) > q(2)$. If $\kappa \leq q(2)$, then $F^{LLR}(2,\kappa) = 0$, since for such κ network overflow is not a rare event. If $q(2) < \kappa \leq q(1)$, then the load at location 2 can grow to level κ, while, even without any large deviation occurring at location 1, all type 2 arrivals are assigned to location 2. Thus, it makes sense that $F^{LLR}(2,\kappa) = H_{q(2)}(0,\kappa)$ for such values of κ. Finally, if $\kappa > q(1) > q(2)$, as the load at location 2 begins to build beyond $q(2)$, the load at location 1 begins to build beyond $q(1)$, even though the two loads are not equal. In that way, all type 2 consumers are assigned to location 2, even after the load at location 2 exceeds $q(1)$. Eventually, the loads at

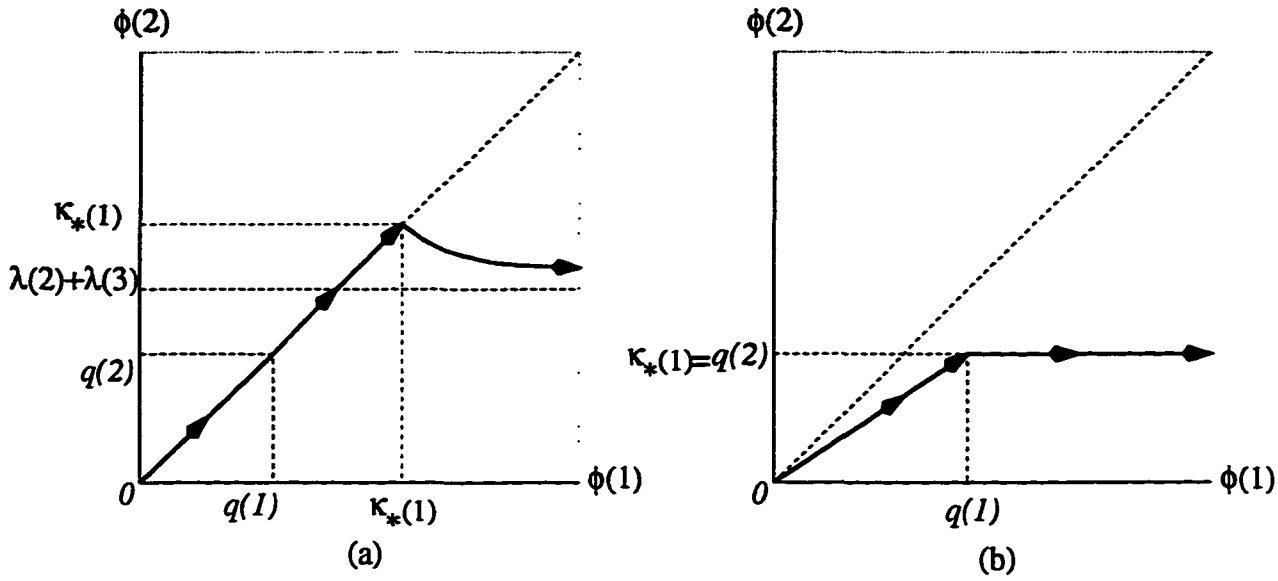


Figure 3.1 The most likely scenario for the overflow of location 1 of the W network under the LLR policy, for the cases (a) $q(1) = q(2)$, (b) $q(1) > q(2)$.

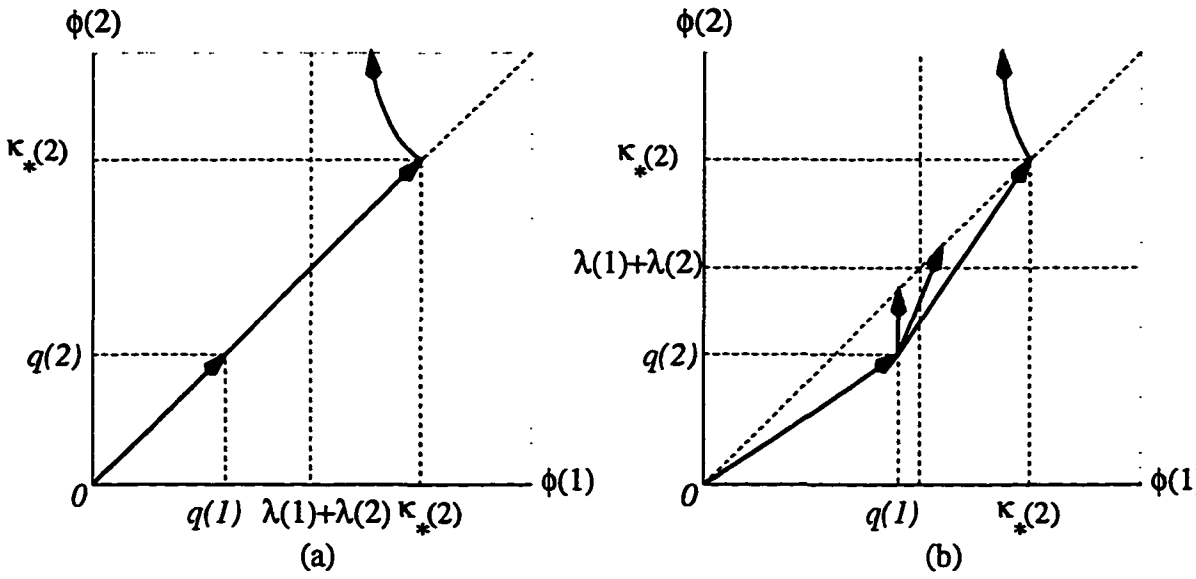


Figure 3.2 The most likely scenario for the overflow of location 2 of the W network under the LLR policy, for the cases (a) $q(1) = q(2)$, (b) $q(1) > q(2)$.

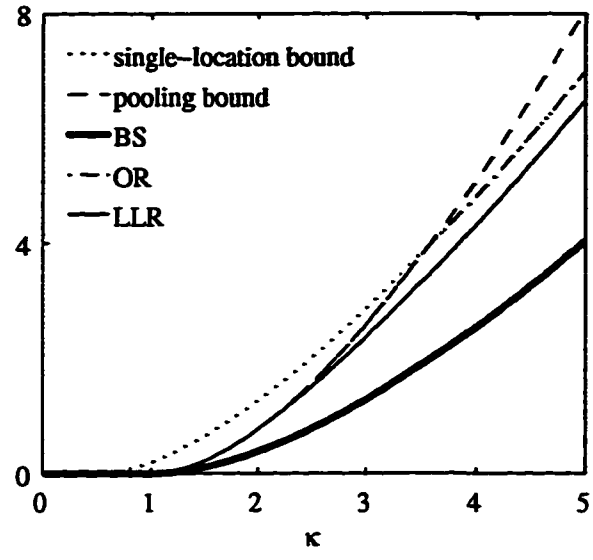


Figure 3.3 The network overflow exponents of the three policies, along with the single-location and pooling bounds, for $\alpha = 0.5$.

the two locations simultaneously become approximately equal to $\kappa \wedge \kappa_*(2)$. If $\kappa > \kappa_*(2)$, then the load at location 2 unilaterally continues to increase to level κ. It is interesting to note that the initial segments of the most likely trajectories depend on κ as κ ranges over $\kappa > q(1) > q(2)$, as illustrated by the multiple trajectories in Figure 3.2(b).

As a numerical example to compare the three policies, consider the W network with demand $\lambda = (1-\alpha, 2\alpha, 1-\alpha)$, where $0 \leq \alpha \leq 1$. The network overflow exponents under the three policies are plotted in Figure 3.3, along with the single-location and pooling upper bounds, for the case $\alpha = 0.5$. The OR policy employs the tightest possible control; hence the network overflow time under OR dominates the network overflow time under *any* allocation policy. Furthermore, in the W network, $F^{OR}(\kappa)$ is equal to the smaller of the single-location and pooling bounds. From a practical point of view, the OR policy has drawbacks such as high computational complexity and the required repacking of consumers. For the values of $\kappa \leq \kappa_*(1)$, the simple and nonrepacking LLR policy performs as well as any other policy, in the sense that $F^{LLR}(\kappa) = F^{OR}(\kappa)$. For larger values of κ, the nonrepacking nature of LLR reveals itself, and LLR is outperformed by OR. The BS policy only exerts open-loop control, and its performance is significantly worse than the LLR policy for the whole range of capacities as illustrated in Figure 3.3.

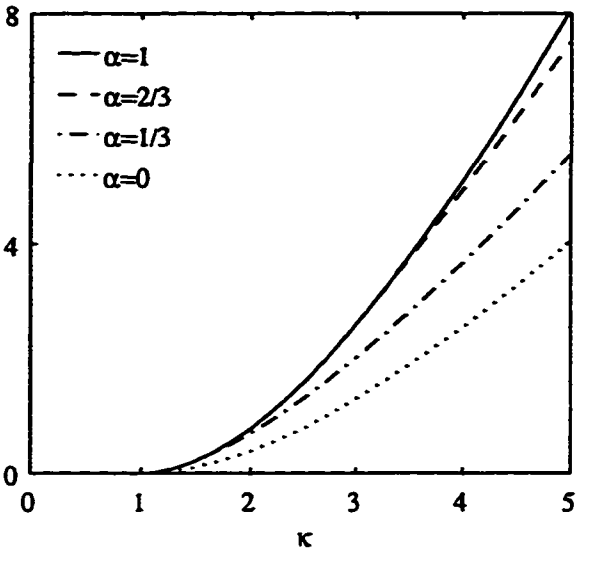


Figure 3.4 $F^{LLR}(\kappa)$ for several values of α.

Consider also the dependence of $F^{LLR}(\kappa)$ on α, illustrated in Figure 3.4. Larger values of α correspond to increased load sharing capability of the network for the same total demand, so it is not surprising that $F^{LLR}(\kappa)$ is increasing in α. Note that when $\alpha = 0$ and $\alpha = 1$, $F^{LLR}(\kappa)$ achieves the single-location and pooling bounds, respectively.

Networks with Arbitrary Topologies. We next consider the overflow exponents of networks with arbitrary topologies under the three policies. Define the function $\Phi : R^V \to R$ as $\Phi(x) = \sum_{v \in V}(x(v))^2$ for $x \in R^V$. Given a load sharing network (U, V, N) and a demand vector λ, let a denote an assignment that solves the problem $SLB(\lambda, \Phi)$ of Section 2.2, and let q denote the load vector corresponding to a.

The BS policy assigns each type u consumer to location $v \in N(u)$ with probability $a_{u,v}/\lambda(u)$ so that the load at each location v behaves as an independent single-location network load with demand $\gamma q(v)$. Straightforward adaptation of Theorem 3.1.2 (details are omitted) yields that the overflow exponent of location $v \in V$ exists and is given by $H_{q(v)}(0, \kappa)$.

To analyze the OR policy in general networks, let $L_t(u)$ continue to denote the number of type u consumers in the network at time t, and $L_t = (L_t(u) : u \in U)$. Note that the process L is a vector of independent single-location load processes with demand vector $\gamma\lambda$, and the network load at time t is determined by L_t only. The network overflow time is the first time t such that there is a subset F of locations such that $\sum_{u:N(u) \subset F} L_t(u) > \lfloor \gamma\kappa \rfloor |F|$. In turn Theorem 3.1.3 can be easily

generalized to show that the overflow exponent of the OR policy exists and can be expressed as $\min_{F \subset V} H_{\lambda(F)}(0, \kappa|F|)$, where $\lambda(F) = \sum_{u:N(u) \subset F} \lambda(u)$.

Except for simple network topologies such as the W network, the load process under the LLR policy has discontinuous statistics along complicated geometries. Due to this fact, establishing explicit large deviations principles for arbitrary load sharing networks appears difficult. Nevertheless, the form of the overflow exponents provided by Theorem 3.1.4, together with the extremal paths of Figures 3.1 and 3.2, suggests the following conjecture:

Conjecture 3.1.1 *For each $v \in V$ and $\kappa \geq 0$, $F^{LLR}(v, \kappa)$ can be identified as follows: Let S range over the set of set-valued functions of the form $S = (S(x) : 0 \leq x \leq \kappa)$, where $v \in S(x) \subset V$ for $0 \leq x \leq \kappa$, and $S(x) \subset S(x')$ for $x \geq x'$. Associated with each such S and $0 \leq x \leq \kappa$, define $R(x) = \{u \in U : N(u) \subset S(x) \cup \{v' : q(v') > x\}\}$, $N(x,u) = N(u) \cap S(x)$ for $u \in R(x)$, and $\lambda_x = (\lambda(u) : u \in R(x))$, and let $(q(v', x) : v' \in S(x))$ denote the unique load vector corresponding to the solutions of $SLB(\lambda_x, \Phi)$ on the subnetwork $(R(x), S(x), N(x))$. Then*

$$F^{LLR}(v, \kappa) = \inf_S \int_0^\kappa \sum_{v' \in S(x)} \left(\log(\frac{x}{q(v', x)}) \right)_+ dx.$$

The rest of the chapter is organized as follows: Section 3.2 consists of the basic definitions regarding the techniques employed in the analysis, namely the theory of large deviations. Theorem 3.1.1 regarding the single-location network is proved in Section 3.3, and Theorems 3.1.2-3.1.4 regarding the W network are proved in Section 3.4. A large deviations principle for the W network under the LLR policy is stated as a proposition in Section 3.4 and is proved in Section 3.5.

3.2 Definitions

Given a positive integer d, let R^d denote the d dimensional Euclidean space. A collection $\nu = (\nu(x) : x \in R^d)$ is called a *rate-measure field* if for each x, $\nu(x) = \nu(x, \cdot)$ is a positive Borel measure on R^d, and $\sup_x \nu(x, R^d) < \infty$. For each positive scalar γ, a right continuous Markov jump process $X^\gamma = (X_t^\gamma : t \geq 0)$ is said to be *generated* by the pair (γ, ν) if given its value at time t, the process X^γ jumps after a random time exponentially distributed with parameter $\gamma \nu(X_t^\gamma, R^d)$, and the jump size is a random variable Δ, where $\gamma\Delta$ has distribution $\nu(X_t^\gamma)/\nu(X_t^\gamma, R^d)$, independent of the past

history. The *polygonal interpolation* of the process X^γ, $\tilde{X}^\gamma$, is defined as

$$\tilde{X}_t^\gamma = \frac{t-\tau_k}{\tau_{k+1}-\tau_k} X_{\tau_{k+1}}^\gamma + \frac{\tau_{k+1}-t}{\tau_{k+1}-\tau_k} X_{\tau_k}^\gamma \qquad \tau_k \le t \le \tau_{k+1},$$

where τ_k is the k^{th} jump time of X^γ. Since X^γ has a finite number of jumps in bounded time intervals, $\tilde{X}^\gamma$ has sample paths in $C_{[0,\infty)}(R^d)$, the space of continuous functions $\phi : [0,\infty) \to R^d$ with the topology of uniform convergence on compact sets.

The following are some standard definitions of large deviations theory: Let P_x denote the probability measure governing the process $\tilde{X}^\gamma$ given $\tilde{X}_0^\gamma = x$, and suppose that $\tilde{X}_t^\gamma$ takes on values in $D \subset R^d$ for each $0 \le t \le T$. The sequence $(\tilde{X}^\gamma : \gamma > 0)$ is said to *satisfy the large deviations principle* in $C_{[0,T]}(D)$ with the rate function $\Gamma : C_{[0,T]}(D) \times D \to R_+ \cup \{+\infty\}$ if for each $x_0 \in D$, the function $\Gamma(\cdot, x_0)$ is lower semicontinuous, and for any sequence $(x^\gamma : \gamma > 0)$ such that $\lim_{\gamma\to\infty} x^\gamma = x_0$ and Borel measurable $S \subset C_{[0,T]}(D)$,

$$\begin{aligned} \limsup_{\gamma\to\infty} \gamma^{-1} \log P_{x^\gamma}\left((\tilde{X}_t^\gamma : 0 \le t \le T) \in S\right) &\le -\inf_{\phi\in\overline{S}} \Gamma(\phi, x_0) \\ \liminf_{\gamma\to\infty} \gamma^{-1} \log P_{x^\gamma}\left((\tilde{X}_t^\gamma : 0 \le t \le T) \in S\right) &\ge -\inf_{\phi\in S^o} \Gamma(\phi, x_0), \end{aligned}$$

where $\overline{S}$ and S^o denote the closure and the interior of S, respectively. The rate function Γ is called *good* if for each $x_0 \in D$ and $l \ge 0$, the level set $\{\phi \in C_{[0,T]}(D) : \Gamma(\phi, x_0) \le l\}$ is compact.

Remark 3.2.1 *If the sequence $(\tilde{X}^\gamma : \gamma > 0)$ satisfies the large deviations principle in $C_{[0,T]}(D)$ with the good rate function Γ, then the following inequalities hold for each $\phi \in C_{[0,T]}(D)$:*

$$\begin{aligned} \lim_{\delta\searrow 0} \limsup_{\gamma\to\infty} \gamma^{-1} \log \sup_{|x-\phi_0|<\delta} P_x\Big(\sup_{0\le t\le T} |\tilde{X}_t^\gamma - \phi_t| < \delta\Big) &\le -\Gamma(\phi, \phi_0) \\ \lim_{\delta\searrow 0} \lim_{\rho\searrow 0} \liminf_{\gamma\to\infty} \gamma^{-1} \log \inf_{|x-\phi_0|<\rho} P_x\Big(\sup_{0\le t\le T} |\tilde{X}_t^\gamma - \phi_t| < \delta\Big) &\ge -\Gamma(\phi, \phi_0). \end{aligned}$$

Two sequences $(\tilde{X}^\gamma : \gamma > 0)$ and $(\tilde{Y}^\gamma : \gamma > 0)$ on the same probability space are *exponentially equivalent* if for each $\delta > 0$, $\lim_{\gamma\to\infty} \gamma^{-1} \log P(\sup_{0\le t\le T} |\tilde{X}_t^\gamma - \tilde{Y}_t^\gamma| > \delta) = -\infty$. In this case, if $(\tilde{X}^\gamma : \gamma > 0)$ satisfies the large deviations principle with a good rate function Γ, then so does $(\tilde{Y}^\gamma : \gamma > 0)$ ([13, Theorem 4.2.13]).

3.3 The Single-Location Network

This section presents the proof of Theorem 3.1.1. The essential ingredient of the proof is Lemma 3.3.1, which establishes a large deviations principle for the load process. In view of Lemma 3.3.1, the proof of Theorem 3.1.1 hinges on the solution of a variational optimization problem that is provided by Lemma 3.3.4.

The normalized load process $X^\gamma = \gamma^{-1}X$ is a Markov jump process that takes values in R_+. It is generated by the pair (γ, ν), where for each $x \in R_+$, $\nu(x, \{1\}) = \lambda(1)$, $\nu(x, \{-1\}) = x$, and $\nu(x, \{1, -1\}^c) = 0$. Note that $\gamma\nu(x, \{1\})$ and $\gamma\nu(x, \{-1\})$ are the consumer arrival and departure rates, respectively, when the normalized load is x. The polygonal interpolation of the normalized load process, $\tilde{X}^\gamma$, satisfies a large deviations principle as identified by the following lemma. The lemma is a slight variation of the results in [12, Section 12], which assume a bounded state space, and it follows by taking $\lambda(2) = 0$ in Lemma 3.4.1. Its proof is therefore omitted.

Lemma 3.3.1 *The sequence $(\tilde{X}^\gamma : \gamma > 0)$ satisfies the large deviations principle in $C_{[0,T]}(R_+)$ with the good rate function $\Gamma_{\lambda(1)}$, where for each $\phi \in C_{[0,T]}(R_+)$ and $x \in R_+$,*

$$\Gamma_{\lambda(1)}(\phi, x) = \begin{cases} \int_0^T \Lambda_{\lambda(1)}(\phi_t, \dot{\phi}_t)dt & \text{if } \phi_0 = x \text{ and } \phi \text{ is absolutely continuous} \\ +\infty & \text{otherwise,} \end{cases}$$

and

$$\Lambda_{\lambda(1)}(\phi_t, \dot{\phi}_t) = \dot{\phi}_t \log\left(\frac{\dot{\phi}_t + \sqrt{\dot{\phi}_t^2 + 4\lambda(1)\phi_t}}{2\lambda(1)}\right) + \phi_t + \lambda(1) - \sqrt{\dot{\phi}_t^2 + 4\lambda(1)\phi_t}.$$

Remark 3.3.1 *For fixed $\lambda(1)$ and ϕ_t, the function $\Lambda_{\lambda(1)}(\phi_t, \dot{\phi}_t)$ is a strictly convex, nonnegative function of $\dot{\phi}_t$. Furthermore, $\Lambda_{\lambda(1)}(\phi_t, \dot{\phi}_t) = 0$ if and only if $\dot{\phi}_t = \lambda(1) - \phi_t$, in which case we say that ϕ* relaxes *under $\lambda(1)$.*

Remark 3.3.2 *Note that $\Lambda_{\lambda(1)}(\phi_t, \dot{\phi}_t) \geq \inf_{\alpha \geq 0} \Lambda_{\lambda(1)}(\phi_t, \alpha\dot{\phi}_t)/\alpha = \dot{\phi}_t \log(\phi_t/\lambda(1))$. The equality holds if and only if $\dot{\phi}_t = \phi_t - \lambda(1)$, in which case we say that ϕ* relaxes in reverse time *under $\lambda(1)$. Thus, for absolutely continuous ϕ,*

$$\int_0^T \Lambda_{\lambda(1)}(\phi_t, \dot{\phi}_t)dt \;\geq\; \int_0^T \dot{\phi}_t \log(\frac{\phi_t}{\lambda(1)})dt \;=\; \int_{\phi_0}^{\phi_T} \log(\frac{x}{\lambda(1)})dx,$$

and therefore, $\Gamma_{\lambda(1)}(\phi,\phi_0) \geq H_{\lambda(1)}(\phi_0,\phi_T)$ *for all* ϕ *such that* $\phi_T \geq \phi_0 \geq \lambda(1)$, *with equality if and only if* ϕ *relaxes in reverse time under* $\lambda(1)$.

Lemma 3.3.2 *For each* $x \geq 0$, $y \in R$, *and* $\epsilon > 0$, $\Lambda_{\lambda(1)}(x+\epsilon,y) \leq \Lambda_{\lambda(1)}(x,y)+\epsilon$.

Proof. The lemma follows by the fact that for all $x \geq 0$ and $y \in R$,

$$\frac{\partial \Lambda_{\lambda(1)}(x,y)}{\partial x} = 1 - \frac{2\lambda(1)}{y+\sqrt{y^2+4\lambda(1)x}} \leq 1.$$

□

Lemma 3.3.3 *For each* $0 \leq x \leq y$,

$$\inf\{\Gamma_{\lambda(1)}(\phi,x) : \phi_0 = x, \sup_{0\leq t\leq T} \phi_t > y\} = \inf\{\Gamma_{\lambda(1)}(\phi,x) : \phi_0 = x, \sup_{0\leq t\leq T} \phi_t \geq y\}.$$

Proof. To prove the lemma, it suffices to show that

$$\inf\{\Gamma_{\lambda(1)}(\phi,x) : \phi_0 = x, \sup_{0\leq t\leq T} \phi_t > y\} \leq \inf\{\Gamma_{\lambda(1)}(\phi,x) : \phi_0 = x, \sup_{0\leq t\leq T} \phi_t \geq y\}. \tag{3.1}$$

Fix $\epsilon > 0$. By the goodness of $\Gamma_{\lambda(1)}$ there exists a solution ϕ to the right-hand side of inequality (3.1), and clearly $\Gamma_{\lambda(1)}(\phi,x)$ is finite. Set $M = \sup_{0\leq t\leq T}\phi_t < \infty$, and choose B large enough so that $\inf_{0\leq z\leq M}\Lambda_{\lambda(1)}(z,\dot{\phi}_t) > \Gamma_{\lambda(1)}(\phi,x)/T$ whenever $\dot{\phi}_t > B$. Then the set $S = \{t \in [0,T] : \dot{\phi}_t \leq B\}$ has positive measure; therefore, $\xi \in C_{[0,T]}(R_+)$ defined by $\xi_0 = \phi_0$ and $\dot{\xi}_t = \dot{\phi}_t + \epsilon I\{t \in S\}$ satisfies $\sup_{0\leq t\leq T}\xi_t > y$. Lemma 3.3.2 and the fact that $\partial\Lambda_{\lambda(1)}(\phi_t,y)/\partial y$ is bounded on S imply that $\Gamma_{\lambda(1)}(\xi,x) \leq \Gamma_{\lambda(1)}(\phi,x) + f(\epsilon)$ for some $f(\epsilon) \to 0$ as $\epsilon \to 0$. The arbitrariness of $\epsilon > 0$ proves the lemma. □

Lemma 3.3.4 *For each* $0 \leq x \leq y$,

$$\inf\{\Gamma_{\lambda(1)}(\phi,x) : \phi_0 = x, \sup_{0\leq t\leq T} \phi_t \geq y,\ T \geq 0\} = H_{\lambda(1)}(x,y). \tag{3.2}$$

Proof. By the nonnegativity of $\Lambda_{\lambda(1)}$, it suffices to show that $\inf\{\Gamma_{\lambda(1)}(\phi,x) : \phi_0 = x,\ \phi_T = y, T \geq 0\} = H_{\lambda(1)}(x,y)$. Consider the following three cases:

Case 1: $x \leq y < \lambda(1)$. There exist a $T \geq 0$ and $\phi \in C_{[0,T]}(R_+)$ such that $\phi_0 = x$, $\phi_T = y$, and ϕ relaxes under $\lambda(1)$. By Remark 3.3.1, $\Gamma_{\lambda(1)}(\phi,x) = H_{\lambda(1)}(x,y) = 0$. The nonnegativity of $\Gamma_{\lambda(1)}$ implies equality (3.2).

Case 2: $\lambda(1) < x \leq y$. There exist $T \geq 0$ and a $\phi \in C_{[0,T]}(R_+)$ such that $\phi_0 = x$, $\phi_T = y$, and ϕ relaxes in reverse time under $\lambda(1)$. Remark 3.3.2 implies equality (3.2).

Case 3: $x \leq \lambda(1) \leq y$. Fix $\epsilon > 0$. Note that the nonnegativity of $\Lambda_{\lambda(1)}$ and Remark 3.3.2 imply that the left-hand side of (3.2) is bounded below by $H_{\lambda(1)}(y \wedge (\lambda(1)+\epsilon), y)$, and therefore by $H_{\lambda(1)}(\lambda(1), y) = H_{\lambda(1)}(x,y)$. The lemma is established by constructing $T \geq 0$ and $\phi \in C_{[0,T]}(R_+)$ such that $\Gamma_{\lambda(1)}(\phi,x)$ is arbitrarily close to $H_{\lambda(1)}(x,y)$: Set $\phi_0 = x$, and let ϕ relax under $\lambda(1)$ until it reaches level $x \vee (\lambda(1) - \epsilon)$; then satisfy $\dot{\phi}_t = (y \wedge (\lambda(1)+\epsilon) - x \vee (\lambda(1) - \epsilon))/\epsilon$ until it reaches level $y \wedge (\lambda(1) + \epsilon)$; and from then on, relax in reverse time under $\lambda(1)$ until time T such that $\phi_T = y$. By Remarks 3.3.1 and 3.3.2, $\Gamma_{\lambda(1)}(\phi,x) = H_{\lambda(1)}(y \wedge (\lambda(1)+\epsilon), y) + f(\epsilon)$ for some $f(\epsilon) \to 0$ as $\epsilon \to 0$. The arbitrariness of $\epsilon > 0$ and continuity of $H_{\lambda(1)}$ imply inequality (3.2). □

Proof of Theorem 3.1.1. The fact

$$\{\sup_{0\leq t\leq T} \tilde{X}^\gamma_t - \gamma^{-1} > \kappa\} \subset \{\text{ Overflow time} \leq T\} \subset \{\sup_{0\leq t\leq T} \tilde{X}^\gamma_t + \gamma^{-1} \geq \kappa\}$$

together with Lemma 3.3.1 and the exponential equivalence of $(\tilde{X}^\gamma - \gamma^{-1} : \gamma > 0)$ and $(\tilde{X}^\gamma + \gamma^{-1} : \gamma > 0)$ imply that

$$\begin{aligned}
\limsup_{\gamma\to\infty} \gamma^{-1} \log P(\text{ Overflow time} \leq T | X_0 = 0) &\leq -\inf\{\Gamma_{\lambda(1)}(\phi,0) : \sup_{0\leq t\leq T} \phi_t \geq \kappa\}, \\
\liminf_{\gamma\to\infty} \gamma^{-1} \log P(\text{ Overflow time} \leq T | X_0 = 0) &\geq -\inf\{\Gamma_{\lambda(1)}(\phi,0) : \sup_{0\leq t\leq T} \phi_t > \kappa\}.
\end{aligned}$$

By Lemma 3.3.3, $\lim_{\gamma\to\infty} \gamma^{-1} \log P(\text{ Overflow time} \leq T | X_0 = 0)$ exists; in turn Lemma 3.3.4 implies that $\lim_{T\to\infty} \lim_{\gamma\to\infty} \gamma^{-1} \log P(\text{ Overflow time} \leq T | X_0 = 0) = -H_{\lambda(1)}(0,\kappa)$. This establishes the theorem. □

3.4 The W Network

This section consists of the proofs of Theorems 3.1.2-3.1.4. We start each proof by establishing the large deviations principle satisfied by the network load, then formulate and solve a variational optimization problem that yields the desired conclusions via large deviations bounds.

3.4.1 Bernoulli splitting

The Bernoulli splitting (BS) policy is to assign each type 2 consumer to location 1 with probability $p = [(\lambda(3) - \lambda(1) + \lambda(2))/2\lambda(2)]_0^1$ or to location 2 with probability $1 - p$, independently of the past history. Thus, under the BS policy, $X(1)$ and $X(2)$ are independent single-location network loads with demands $\gamma q(1)$ and $\gamma q(2)$, respectively.

Let $\tilde{X}^\gamma$ denote the polygonal interpolation of the normalized load process X^γ. Note that the processes $\tilde{X}^\gamma(1)$ and $\tilde{X}^\gamma(2)$ are independent, and each satisfies a large deviations principle in the complete separable metric space $C_{[0,T]}(R_+)$ with a good rate function. Therefore, $\tilde{X}^\gamma = (\tilde{X}^\gamma(1), \tilde{X}^\gamma(2))$ satisfies a large deviations principle in the product space with the good rate function given by the sum of individual rate functions (see [13, Theorems 4.1.18 and 4.1.11, Lemma 1.2.18, and Exercise 4.1.10]).

Lemma 3.4.1 *The sequence* $(\tilde{X}^\gamma : \gamma > 0)$ *satisfies a large deviations principle in* $C_{[0,T]}(R_+^2)$ *with the good rate function* Γ^{BS}, *where for each* $\phi \in C_{[0,T]}(R_+^2)$ *and* $x \in R_+^2$, $\Gamma^{BS}(\phi, x) = \Gamma_{q(1)}(\phi(1), x(1)) + \Gamma_{q(2)}(\phi(2), x(2))$.

For $\kappa \geq 0$ and $v = 1, 2$, define the set

$$\Omega(v, \kappa) = \bigcup_{T \geq 0} \{ \phi \in C_{[0,T]}(R_+^2) : \phi_0 = 0, \sup_{0 \leq t \leq T} \phi_t(v) \geq \kappa \}.$$

The following lemma gives the solution of a variational optimization problem associated with the overflow of each location.

Lemma 3.4.2 *For each* $\kappa \geq 0$ *and* $v = 1, 2$, $\inf\{\Gamma^{BS}(\phi, 0) : \phi \in \Omega(v, \kappa)\} = H_{q(v)}(0, \kappa)$.

Proof. The same proof applies for both locations; therefore, only location $v = 1$ is considered. If $\phi \in \Omega(1, \kappa)$, then $\phi_\tau(1) = \kappa$ for some $\tau \geq 0$, so the definition of Γ^{BS} and Lemma 3.3.4 imply that $\Gamma^{BS}(\phi, 0) \geq H_{q(1)}(0, \kappa)$. Thus, $\inf\{\Gamma^{BS}(\phi, 0) : \phi \in \Omega(1, \kappa)\} \geq H_{q(1)}(0, \kappa)$. The proof is completed

by constructing a $\phi \in \Omega(1,\kappa)$ such that $\Gamma^{BS}(\phi,0)$ is arbitrarily close to $H_{q(1)}(0,\kappa)$: Fix $\epsilon > 0$ and appeal to Lemma 3.3.4 to choose $T \geq 0$ and $\phi(1) \in C_{[0,T]}(R_+)$ such that $\phi_0(1) = 0$, $\phi_T(1) = \kappa$, and $\Gamma_{q(1)}(\phi(1),0) \leq H_{q(1)}(0,\kappa) + \epsilon$. Let $\phi(2) \in C_{[0,T]}(R_+)$ be such that $\phi_0(2) = 0$ and $\phi(2)$ relaxes under $q(2)$. Note that $\phi = (\phi(1),\phi(2)) \in \Omega(1,\kappa)$ and $\Gamma^{BS}(\phi,0) \leq H_{q(1)}(0,\kappa) + \epsilon$. The lemma follows by the arbitrariness of $\epsilon > 0$. □

Proof of Theorem 3.1.2. The proof of Theorem 3.1.1, with Lemmas 3.4.1 and 3.4.2 in place of Lemmas 3.3.1 and 3.3.4, respectively, and an adaptation of Lemma 3.3.3, applied separately on each location v, establishes the existence and the desired form of $F^{BS}(v,\kappa)$. The fact that $H_{q(1)}(0,\kappa) \leq H_{q(2)}(0,\kappa)$ implies $F^{BS}(\kappa) = F^{BS}(1,\kappa)$. □

3.4.2 Optimal repacking

The optimal repacking (OR) policy is to continuously rearrange the consumers in the network so as to minimize the maximum load in the network subject to the neighborhood constraints. For each type $u \in U$, let $L_t(u)$ continue to denote the number of type u consumers in the network at time t. The processes $L(1)$, $L(2)$, and $L(3)$ are independent single-location network loads with demands $\gamma\lambda(1)$, $\gamma\lambda(2)$, and $\gamma\lambda(3)$, respectively. Under the OR policy, the value of the load at time t is determined by the value of the process $L = (L(1),L(2),L(3))$ at time t only. In particular, $|X_t - m(L_t)| < 1$, where the mapping $m : R_+^3 \to R_+^2$ is defined by the relations

$$\begin{aligned} m(L_t,1) &= [(L_t(1)+L_t(2)+L_t(3))/2]_{L_t(1)}^{L_t(1)+L_t(2)} \\ m(L_t,2) &= [(L_t(1)+L_t(2)+L_t(3))/2]_{L_t(3)}^{L_t(3)+L_t(2)}. \end{aligned}$$

The following lemma identifies the large deviations principle satisfied by the network load under the OR policy.

Lemma 3.4.3 *Define the mapping $\mathcal{M} : C_{[0,T]}(R_+^3) \to C_{[0,T]}(R_+^2)$ by $\mathcal{M}(\xi)_t = m(\xi_t)$, $0 \leq t \leq T$. The sequence $(\tilde{X}^\gamma : \gamma > 0)$ satisfies a large deviations principle in $C_{[0,T]}(R_+^2)$ with the good rate function Γ^{OR}, where for each $\phi \in C_{[0,T]}(R_+^2)$ and $x \in R_+^2$,*

$$\Gamma^{OR}(\phi,x) = \inf_{\substack{\xi \in C_{[0,T]}(R_+^3), \\ \mathcal{M}(\xi)=\phi,\ \mathcal{M}(\xi)_0 = x}} \{\Gamma_{\lambda(1)}(\xi(1),\xi_0(1)) + \Gamma_{\lambda(2)}(\xi(2),\xi_0(2)) + \Gamma_{\lambda(3)}(\xi(3),\xi_0(3))\}$$

with the understanding that the infimum over an empty set equals $+\infty$.

Proof. Let $\tilde{L}^\gamma$ denote the polygonal interpolation of the scaled process $\gamma^{-1}L$. Note that the sequences $(\tilde{X}^\gamma : \gamma > 0)$ and $(\mathcal{M}(\tilde{L}^\gamma) : \gamma > 0)$ are exponentially equivalent; thus, it suffices to establish the desired large deviations principle for $(\mathcal{M}(\tilde{L}^\gamma) : \gamma > 0)$. By Lemma 3.3.1, $\tilde{L}^\gamma(1)$, $\tilde{L}^\gamma(2)$, and $\tilde{L}^\gamma(3)$ satisfy large deviations principles in $C_{[0,T]}(R_+)$ with good rate functions $\Gamma_{\lambda(1)}$, $\Gamma_{\lambda(2)}$, and $\Gamma_{\lambda(3)}$, respectively. Therefore, the sequence $(\tilde{L}^\gamma : \gamma > 0)$ satisfies the large deviations principle in $C_{[0,T]}(R_+^3)$ with the good rate function $\hat{\Gamma}$, where for each $\xi \in C_{[0,T]}(R_+^3)$ and $x \in R_+^3$, $\hat{\Gamma}(\xi, x) = \Gamma_{\lambda(1)}(\xi(1), x(1)) + \Gamma_{\lambda(2)}(\xi(2), x(2)) + \Gamma_{\lambda(3)}(\xi(3), x(3))$. Continuity of the mapping $\mathcal{M}$ and the Contraction Principle ([13, Theorem 4.2.1]) imply the statement of the lemma. □

Lemma 3.4.4 *For any positive integer* d, $x \in R_+^d$, $y \in R^d$, *and positive* $\alpha \in R_+^d$

$$\sum_{u=1}^{d} \Lambda_{\alpha(u)}(x(u), y(u)) \geq \Lambda_{\sum_{u=1}^{d} \alpha(u)}(\sum_{u=1}^{d} x(u), \sum_{u=1}^{d} y(u)).$$

Proof. Note that $\sigma\Lambda_\alpha(x, y) = \Lambda_{\sigma\alpha}(\sigma x, \sigma y)$ for $\sigma > 0$ and that $\Lambda_{\sum_{u=1}^{d} \alpha(u)}(\cdot, \cdot)$ is convex on $R_+ \times R$ as can be verified by checking that the Hessian matrix is positive semidefinite. Therefore,

$$\begin{aligned}
\sum_{u=1}^{d} \Lambda_{\alpha(u)}(x(u), y(u)) &= \sum_{u=1}^{d} \frac{\alpha(u)}{\sum_{w=1}^{d} \alpha(w)} \Lambda_{\sum_{w=1}^{d} \alpha(w)} \left(\frac{\sum_{w=1}^{d} \alpha(w)}{\alpha(u)} x(u), \frac{\sum_{w=1}^{d} \alpha(w)}{\alpha(u)} y(u) \right) \\
&\geq \Lambda_{\sum_{u=1}^{d} \alpha(u)} \left(\sum_{u=1}^{d} x(u), \sum_{u=1}^{d} y(u) \right),
\end{aligned}$$

and the lemma follows. □

Lemma 3.4.5 *For each* $\kappa \geq 0$ *and* $v = 1, 2$,

$$\inf\{\Gamma^{OR}(\phi, 0) : \phi \in \Omega(v, \kappa)\} = H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2\kappa) \wedge H_{\lambda(2v-1)}(0, \kappa).$$

Proof. The same proof applies for both locations; therefore, only location $v = 1$ is considered. Let $\phi \in \Omega(1, \kappa)$ be absolutely continuous and $\xi \in C_{[0,T]}(R_+^3)$ be such that $\phi = \mathcal{M}(\xi)$. Define

$\tau = \inf\{t \geq 0 : \phi_t(1) = \kappa\}$. If $\phi_\tau(2) < \phi_\tau(1)$, then necessarily $\xi_\tau(1) = \kappa$; hence the definition of Γ^{OR} and Lemma 3.3.4 imply that $\Gamma^{OR}(\phi, 0) \geq H_{\lambda(1)}(0, \kappa)$. Otherwise,

$$\begin{aligned}
\Gamma^{OR}(\phi, 0) &\geq \int_0^\tau \left(\Lambda_{\lambda(1)}(\xi_t(1), \dot{\xi}_t(1)) + \Lambda_{\lambda(2)}(\xi_t(2), \dot{\xi}_t(2)) + \Lambda_{\lambda(3)}(\xi_t(3), \dot{\xi}_t(3))\right) dt \\
&\geq \int_0^\tau \Lambda_{\lambda(1)+\lambda(2)+\lambda(3)}(\xi_t(1) + \xi_t(2) + \xi_t(3), \dot{\xi}_t(1) + \dot{\xi}_t(2) + \dot{\xi}_t(3)) dt \qquad (3.3) \\
&\geq H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2\kappa), \qquad (3.4)
\end{aligned}$$

where inequality (3.3) follows by Lemma 3.4.4, and inequality (3.4) follows by Lemma 3.3.4 and the monotonicity of $H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, \cdot)$. Thus, $\inf\{\Gamma^{OR}(\phi, 0) : \phi \in \Omega(1, \kappa)\} \geq H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2\kappa) \wedge H_{\lambda(1)}(0, \kappa)$. The proof is completed by constructing a function $\phi \in \Omega(1, \kappa)$ such that $\Gamma^{OR}(\phi, 0)$ is arbitrarily close to the established lower bound. Fix $\epsilon > 0$ and consider the following two cases:

Case 1: $H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2\kappa) \leq H_{\lambda(1)}(0, \kappa)$. Appeal to Lemma 3.3.4 to choose $T \geq 0$ and $\xi \in C_{[0,T]}(R_+^3)$ such that $\xi_0 = 0$, $\xi_T = 2\kappa(\lambda(1), \lambda(2), \lambda(3))/(\lambda(1)+\lambda(2)+\lambda(3))$, and $\Gamma_{\lambda(u)}(\xi(u), 0) \leq H_{\lambda(u)}(0, \xi_T(u)) + \epsilon$ for each $1 \leq u \leq 3$. Then $\mathcal{M}(\xi) \in \Omega(1, \kappa)$ and $\Gamma^{OR}(\mathcal{M}(\xi), 0) \leq H_{\lambda(1)}(0, \xi_T(1)) + H_{\lambda(2)}(0, \xi_T(2)) + H_{\lambda(3)}(0, \xi_T(3)) + 3\epsilon = H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2\kappa) + 3\epsilon$.

Case 2: $H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2\kappa) > H_{\lambda(1)}(0, \kappa)$. Appeal to Lemma 3.3.4 to choose $T \geq 0$ and $\xi(1) \in C_{[0,T]}(R_+)$ such that $\xi_0(1) = 0$, $\xi_T(1) = \kappa$, and $\Gamma_{\lambda(1)}(\xi(1), 0) \leq H_{\lambda(1)}(0, \kappa) + \epsilon$. Let $\xi = (\xi(1), \xi(2), \xi(3))$, where $\xi_0(2) = \xi_0(3) = 0$, and $\xi(2)$ and $\xi(3)$ relax under $\lambda(2)$ and $\lambda(3)$, respectively. Then $\mathcal{M}(\xi) \in \Omega(1, \kappa)$ and $\Gamma^{OR}(\mathcal{M}(\xi), 0) \leq H_{\lambda(1)}(0, \kappa) + \epsilon$.

Set $\phi = \mathcal{M}(\xi)$. The lemma follows by the arbitrariness of $\epsilon > 0$. □

Proof of Theorem 3.1.3. The proof of Theorem 3.1.1, with Lemmas 3.4.3 and 3.4.5 in place of Lemmas 3.3.1 and 3.3.4, respectively, and an adaptation of Lemma 3.3.3, applied separately on each location v, establishes the existence and the desired form of $F^{OR}(v, \kappa)$. The fact that $H_{\lambda(1)}(0, \kappa) \leq H_{\lambda(3)}(0, \kappa)$ implies $F^{OR}(\kappa) = F^{OR}(1, \kappa)$. □

3.4.3 Least load routing

The least load routing (LLR) policy is to assign each new consumer to an admissible location that has the least load within its associated neighborhood. In the W network of Figure 1.2(b), we assume that when both locations have the same load the assignment decision is made in favor of location 1. The normalized load process, X^γ, is a Markov jump process that takes values in R_+^2. The process

X^γ has jumps of magnitude γ^{-1} in the four directions, $e_1^+ = (1,0)$, $e_2^+ = (0,1)$, $e_1^- = (-1,0)$, and $e_2^- = (0,-1)$, and is generated by the pair (γ, ν), where for each $x \in R_+^2$,

$$\begin{aligned}
\nu(x, \{e_1^-\}) &= x(1), \\
\nu(x, \{e_2^-\}) &= x(2), \\
\nu(x, \{e_1^+\}) &= \begin{cases} \lambda(1) + \lambda(2) & \text{if } x(1) \le x(2) \\ \lambda(1) & \text{if } x(1) > x(2), \end{cases} \\
\nu(x, \{e_2^+\}) &= \begin{cases} \lambda(3) & \text{if } x(1) \le x(2) \\ \lambda(3) + \lambda(2) & \text{if } x(1) > x(2). \end{cases}
\end{aligned}$$

Let $\bar{X}^\gamma$ denote the polygonal interpolation of X^γ. The following proposition establishes a large deviations principle for the network load under the LLR policy. The proof of the proposition can be found in Section 3.5.

Proposition 3.4.1 *The sequence $(\bar{X}^\gamma : \gamma > 0)$ satisfies the large deviations principle in $C_{[0,T]}(R_+^2)$ with the good rate function Γ^{LLR}, where for each $\phi \in C_{[0,T]}(R_+^2)$ and $x \in R_+^2$,*

$$\Gamma^{LLR}(\phi, x) = \begin{cases} \int_0^T \Lambda(\phi_t, \dot{\phi}_t) dt & \text{if } \phi_0 = x \text{ and } \phi \text{ is absolutely continuous} \\ +\infty & \text{otherwise,} \end{cases}$$

and Λ satisfies

$$\Lambda(\phi_t, \dot{\phi}_t) = \begin{cases} \Lambda_{\lambda(1)}(\phi_t(1), \dot{\phi}_t(1)) + \Lambda_{\lambda(2)+\lambda(3)}(\phi_t(2), \dot{\phi}_t(2)) & \text{if } \phi_t(1) > \phi_t(2) \\ \Lambda_{q(1)}(\phi_t(1), \dot{\phi}_t(1)) + \Lambda_{q(2)}(\phi_t(2), \dot{\phi}_t(2)) & \text{if } \phi_t(1) = \phi_t(2) \\ \Lambda_{\lambda(1)+\lambda(2)}(\phi_t(1), \dot{\phi}_t(1)) + \Lambda_{\lambda(3)}(\phi_t(2), \dot{\phi}_t(2)) & \text{if } \phi_t(1) < \phi_t(2). \end{cases}$$

The following three lemmas provide the solutions of the two variational optimization problems regarding the overflow of each location. In particular, Lemma 3.4.6 concerns location 1 and Lemma 3.4.8 concerns location 2. Lemma 3.4.7 provides an auxiliary result that is used in the proof of Lemma 3.4.8.

Lemma 3.4.6 *For each $\kappa \ge 0$,*

$$\inf\{\Gamma^{LLR}(\phi, 0) : \phi \in \Omega(1, \kappa)\} = H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2(\kappa_*(1) \wedge \kappa)) + H_{\lambda(1)}(\kappa_*(1) \wedge \kappa, \kappa).$$

Proof. Given absolutely continuous $\phi \in \Omega(1,\kappa)$, let $\tau = \inf\{t \geq 0 : \phi_t(1) = \kappa\}$ and $\tau' = \sup\{t \leq \tau : \phi_t(1) = \phi_t(2)\}$. By the nonnegativity of Λ,

$$\Gamma^{LLR}(\phi,0) \geq \int_0^{\tau'} \Lambda(\phi_t,\dot{\phi}_t)dt + \int_{\tau'}^{\tau} \Lambda(\phi_t,\dot{\phi}_t)dt.$$

Lemmas 3.4.4 and 3.3.4 can be used to bound the terms on the right-hand side as

$$\begin{aligned}
\int_0^{\tau'} \Lambda(\phi_t,\dot{\phi}_t)dt &\geq \int_0^{\tau'} \Lambda_{\lambda(1)+\lambda(2)+\lambda(3)}(\phi_t(1)+\phi_t(2),\dot{\phi}_t(1)+\dot{\phi}_t(2))dt \\
&\geq H_{\lambda(1)+\lambda(2)+\lambda(3)}(0,\phi_{\tau'}(1)+\phi_{\tau'}(2)), \\
\int_{\tau'}^{\tau} \Lambda(\phi_t,\dot{\phi}_t)dt &\geq \int_{\tau'}^{\tau} \Lambda_{\lambda(1)}(\phi_t(1),\dot{\phi}_t(1))dt \\
&\geq H_{\lambda(1)}(\phi_{\tau'}(1),\kappa).
\end{aligned}$$

This, along with the observation $\phi_{\tau'}(1) = \phi_{\tau'}(2)$, implies

$$\begin{aligned}
\inf\{\Gamma^{LLR}(\phi,0) : \phi \in \Omega(1,\kappa)\} &\geq \inf_{0\leq s\leq \kappa}\left\{ H_{\lambda(1)+\lambda(2)+\lambda(3)}(0,2s) + H_{\lambda(1)}(s,\kappa)\right\} \\
&= H_{\lambda(1)+\lambda(2)+\lambda(3)}(0,2(\kappa_*(1)\wedge\kappa)) + H_{\lambda(1)}(\kappa_*(1)\wedge\kappa,\kappa). \quad (3.5)
\end{aligned}$$

The proof is completed by constructing a function $\phi \in \Omega(1,\kappa)$ such that $\Gamma^{LLR}(\phi,0)$ is arbitrarily close to the right-hand side of inequality (3.5). Fix $\epsilon > 0$ and consider the following two cases:

Case 1: $q(1) = q(2)$. Appeal to Lemma 3.3.4 to choose $T \geq 0$ and $\phi(1) \in C_{[0,T]}(R_+)$ such that $\phi_0(1) = 0$, $\phi_T(1) = \kappa_*(1)\wedge\kappa$, and $\Gamma_{q(1)}(\phi(1),0) \leq H_{q(1)}(0,\kappa_*(1)\wedge\kappa) + \epsilon$. Set $\phi = (\phi(1),\phi(1))$. If $\kappa \leq \kappa_*(1)$, the construction is complete and $\Gamma^{LLR}(\phi,0) \leq H_{\lambda(1)+\lambda(2)+\lambda(3)}(0,2\kappa) + 2\epsilon$. Else, if $\kappa > \kappa_*(1)$, then for some small $\delta < \kappa - \kappa_*(1)$, extend ϕ further by setting $\dot{\phi}_t = (1,1)$ for $T \leq t \leq T+\delta$ (this insures that $\phi_{T+\delta} > q(1)$) and by letting $\phi(1)$ relax in reverse time under $\lambda(1)$ and $\phi(2)$ relax under $\lambda(2)+\lambda(3)$ for $T+\delta \leq t \leq T'$, where T' is such that $\phi_{T'}(1) = \kappa$. Note that $\phi_t(1) > \phi_t(2)$ for $T+\delta < t \leq T'$; hence δ can be chosen small enough so that $\Gamma^{LLR}(\phi,0) \leq H_{\lambda(1)+\lambda(2)+\lambda(3)}(0,2\kappa_*(1)) + H_{\lambda(1)}(\kappa_*(1),\kappa) + 3\epsilon$.

Case 2: $q(1) > q(2)$. Note that in this case Remark 3.1.1 implies $F^{LLR}(1,\kappa) = H_{\lambda(1)}(0,\kappa)$. Appeal to Lemma 3.3.4 to choose $T \geq 0$ and $\phi(1) \in C_{[0,T]}(R_+)$ so that $\phi_0(1) = 0$, $\phi_T(1) = \kappa$, and $\Gamma_{\lambda(1)}(\phi(1),0) \leq H_{\lambda(1)}(0,\kappa) + \epsilon$. Let $\phi(2) \in C_{[0,T]}(R_+)$ be such that $\phi_0(2) = 0$ and $\phi(2)$ relaxes

under $\lambda(2)+\lambda(3)$. Set $\phi = (\phi(1), \phi(2))$. Note that $\phi(1)$ can be constructed as in the proof of Lemma 3.3.4 so that $\phi_t(1) \geq \phi_t(2)$ for $0 \leq t \leq T$, and therefore, $\Gamma^{LLR}(\phi, 0) \leq H_{\lambda(1)}(0, \kappa) + \epsilon$.

Figure 3.1 sketches the function ϕ constructed above. The lemma follows by the arbitrariness of $\epsilon > 0$. □

Lemma 3.4.7 *For each $s \geq 0$ and absolutely continuous $\phi \in C_{[0,T]}(R_+^2)$ such that $\phi_0 = 0$ and $\phi_T = (s, s)$,*

$$\int_0^T \Lambda(\phi_t, \dot{\phi}_t)dt \geq H_{q(1)}(0, s) + H_{q(2)}(0, s).$$

Proof. It is convenient to use the representation

$$\int_0^T \Lambda(\phi_t, \dot{\phi}_t)dt = \int_0^T \left(\Lambda_{\rho(1,t)}(\phi_t(1), \dot{\phi}_t(1)) + \Lambda_{\rho(2,t)}(\phi_t(2), \dot{\phi}_t(2))\right) dt, \tag{3.6}$$

where

$$(\rho(1,t), \rho(2,t)) = \begin{cases} (\lambda(1), \lambda(2)+\lambda(3)) & \text{if } \phi_t(1) > \phi_t(2) \\ (q(1), q(2)) & \text{if } \phi_t(1) = \phi_t(2) \\ (\lambda(1)+\lambda(2), \lambda(3)) & \text{if } \phi_t(1) < \phi_t(2). \end{cases}$$

For each $v = 1, 2$, define $\tau(v, x) = \inf\{t \geq 0 : \phi_t(v) = x\}$ for $0 \leq x \leq s$, and define $\sigma_t(v) = I\{\dot{\phi}_t(v) > 0,\ \phi_t(v) \geq \phi_z(v),\ 0 \leq z \leq t\}$ and $\phi_t^*(v) = \sup_{0 \leq z \leq t} \phi_z(v)$ for $0 \leq t \leq T$. Note that the function $\phi^*(v)$ is absolutely continuous, and

$$\phi_t^*(v) = \phi_t(v),\ \dot{\phi}_t^*(v) = \dot{\phi}_t(v) \ \text{ and } \ \tau(v, \phi_t^*(v)) = t \ \text{ for almost all } t \text{ such that } \sigma_t(v) > 0. \tag{3.7}$$

Therefore, if $s \geq q(v)$, then

$$\begin{aligned}
\int_0^T \Lambda_{\rho(v,t)}(\phi_t(v), \dot{\phi}_t(v))dt &\geq \int_{\tau(v,q(v))}^{\tau(v,s)} \Lambda_{\rho(v,t)}(\phi_t(v), \dot{\phi}_t(v))dt \\
&\geq \int_{\tau(v,q(v))}^{\tau(v,s)} \Lambda_{\rho(v,t)}(\phi_t(v), \dot{\phi}_t(v))\sigma_t(v)dt \\
&= \int_{\tau(v,q(v))}^{\tau(v,s)} \Lambda_{\rho(v,\tau(v,\phi_t^*(v)))}(\phi_t^*(v), \dot{\phi}_t^*(v))\sigma_t(v)dt \qquad (3.8) \\
&\geq \int_{\tau(v,q(v))}^{\tau(v,s)} \dot{\phi}_t^*(v) \log\left(\frac{\phi_t^*(v)}{\rho(v, \tau(v, \phi_t^*(v)))}\right) \sigma_t(v)dt \qquad (3.9)
\end{aligned}$$

$$= \int_{\tau(v,q(v))}^{\tau(v,s)} \dot{\phi}_t^*(v) \log\left(\frac{\phi_t^*(v)}{\rho(v,\tau(v,\phi_t^*(v)))}\right) dt \tag{3.10}$$

$$= \int_{q(v)}^{s} \log\left(\frac{x}{\rho(v,\tau(v,x))}\right) dx. \tag{3.11}$$

Here equality (3.8) follows by the observation (3.7), inequality (3.9) is a consequence of Remark 3.3.2, equality (3.10) is implied by the fact that $\dot{\phi}_t^*(v)\sigma_t(v) = \dot{\phi}_t^*(v)$ for almost all t, and equality (3.11) follows by a change of variables. Inequality (3.11), together with representation (3.6) and the nonnegativity of $\Lambda_{\rho(1,t)}$ and $\Lambda_{\rho(2,t)}$, implies

$$\begin{aligned}\int_0^T \Lambda(\phi_t,\dot{\phi}_t)dt &\geq \int_{q(1)\wedge s}^{s} \log\left(\frac{x}{\rho(1,\tau(1,x))}\right) dx + \int_{q(2)\wedge s}^{s} \log\left(\frac{x}{\rho(2,\tau(2,x))}\right) dx \\ &= \int_{q(2)\wedge s}^{q(1)\wedge s} \log\left(\frac{x}{\rho(2,\tau(2,x))}\right) dx \\ &+ \int_{q(1)\wedge s}^{s} \log\left(\frac{x}{\rho(1,\tau(1,x))}\right) + \log\left(\frac{x}{\rho(2,\tau(2,x))}\right) dx. \end{aligned} \tag{3.12}$$

We complete the proof by obtaining appropriate lower bounds for each of the terms on the right-hand side of inequality (3.12). Note that if $\tau(1,x) = \tau(2,x)$, then $(\rho(1,\tau(1,x)),\rho(2,\tau(2,x))) = (q(1),q(2))$; else, if $\tau(1,x) < \tau(2,x)$, then $\rho(1,\tau(1,x)) = \lambda(1)$, and if $\tau(1,x) > \tau(2,x)$, then $\rho(2,\tau(2,x)) = \lambda(3)$. Therefore, $(\rho(1,\tau(1,x)),\rho(2,\tau(2,x)))$ takes values in the set

$$\{\,(q(1),q(2)),\ (\lambda(1),q(2)),\ (\lambda(1),\lambda(2)+\lambda(3)),\ (\lambda(1),\lambda(3)),\ (q(1),\lambda(3)),\ (\lambda(1)+\lambda(2),\lambda(3))\,\},$$

and a simple calculation yields

$$\log\left(\frac{x}{\rho(1,\tau(1,x))}\right) + \log\left(\frac{x}{\rho(2,\tau(2,x))}\right) \geq \log\left(\frac{x}{q(1)}\right) + \log\left(\frac{x}{q(2)}\right) \tag{3.13}$$

$$\log\left(\frac{x}{\rho(2,\tau(2,x))}\right) \geq \log\left(\frac{x}{\lambda(2)+\lambda(3)}\right). \tag{3.14}$$

Since $q(1) > q(2)$ implies $\lambda(2)+\lambda(3) = q(2)$, inequalities (3.13) and (3.14), together with inequality (3.12), imply that $\int_0^T \Lambda(\phi_t,\dot{\phi}_t)dt \geq H_{q(1)}(0,s) + H_{q(2)}(0,s)$. This establishes the lemma. □

Lemma 3.4.8 *For each* $\kappa \geq 0$,

$$\inf\{\Gamma^{LLR}(\phi,0) : \phi \in \Omega(2,\kappa)\} = H_{q(1)}(0,\kappa_*(2)\wedge\kappa) + H_{q(2)}(0,\kappa_*(2)\wedge\kappa) + H_{\lambda(3)}(\kappa_*(2)\wedge\kappa,\kappa).$$

Proof. Given absolutely continuous $\phi \in \Omega(2,\kappa)$, let $\tau = \inf\{t \geq 0 : \phi_t(2) = \kappa\}$ and $\tau' = \sup\{t \leq \tau : \phi_t(1) = \phi_t(2)\}$. By the nonnegativity of Λ,

$$\Gamma^{LLR}(\phi,0) \geq \int_0^{\tau'} \Lambda(\phi_t,\dot{\phi}_t)dt + \int_{\tau'}^{\tau} \Lambda(\phi_t,\dot{\phi}_t)dt.$$

Lemmas 3.4.7 and 3.3.4 can be used to bound the terms on the right-hand side as

$$\begin{aligned}
\int_0^{\tau'} \Lambda(\phi_t,\dot{\phi}_t)dt &\geq H_{q(1)}(0,\phi_{\tau'}(1)) + H_{q(2)}(0,\phi_{\tau'}(2)), \\
\int_{\tau'}^{\tau} \Lambda(\phi_t,\dot{\phi}_t)dt &\geq H_{\lambda(3)}(\phi_{\tau'}(2),\kappa) \wedge \left(H_{\lambda(1)}(\phi_{\tau'}(1),\kappa) + H_{\lambda(2)+\lambda(3)}(\phi_{\tau'}(2),\kappa)\right).
\end{aligned}$$

This, along with the observation $\phi_{\tau'}(1) = \phi_{\tau'}(2)$, implies

$$\begin{aligned}
\inf\{\Gamma^{LLR}(\phi,0) : \phi \in \Omega(2,\kappa)\} &\geq \inf_{0\leq s\leq \kappa}\Big\{ H_{q(1)}(0,s) + H_{q(2)}(0,s) \\
&\qquad + H_{\lambda(3)}(s,\kappa) \wedge \left(H_{\lambda(1)}(s,\kappa) + H_{\lambda(2)+\lambda(3)}(s,\kappa)\right)\Big\} \\
&= H_{q(1)}(0,\kappa_*(2)\wedge\kappa) + H_{q(2)}(0,\kappa_*(2)\wedge\kappa) + H_{\lambda(3)}(\kappa_*(2)\wedge\kappa,\kappa).
\end{aligned}$$

The proof is completed by constructing a function $\phi \in \Omega(2,\kappa)$ such that $\Gamma^{LLR}(\phi,0)$ is arbitrarily close to the right-hand side of the above inequality. Fix $0 < \epsilon < 1$ and consider the following three cases:

Case 1: $\kappa < q(2)$. Choose $T \geq 0$ and $\phi \in C_{[0,T]}(R_+^2)$ such that $\phi_0 = 0$ and $\phi(1)$ and $\phi(2)$ relax, respectively, under $q(1)$ and $q(2)$ so that $\phi_T = \kappa(q(1)/q(2),1)$. Note that $\phi_t(1) = (q(1)/q(2))\phi_t(2)$ for $0 \leq t \leq T$; therefore, $\Gamma^{LLR}(\phi,0) = 0$.

Case 2: $q(2) \leq \kappa \leq \kappa_*(2)$. Let $T > 0$ and $(\phi_t : 0 \leq t \leq T)$ be constructed as in Case 1 with $\kappa = (1-\epsilon)q(2)$. Extend ϕ by setting $\dot{\phi} = (\kappa \vee q(1),\kappa)$ for $T \leq t \leq T+\epsilon$. Note that $\phi_{T+\epsilon}$ is on the line segment with end points $(q(1),q(2))$ and $(\kappa \vee q(1),\kappa)$. If $\kappa = q(2)$, this completes the construction. Else, if $\kappa > q(2)$, extend ϕ further by letting $\phi(1)$ and $\phi(2)$ relax in reverse time, respectively, under $q(1)$ and $q(2)$ so that $\phi_{T'} = (\kappa \vee q(1),\kappa)$ for some time $T' > T+\epsilon$. Note that in this case $(\phi_t(1) - q(1))/(\phi_t(2) - q(2)) = (\kappa \vee q(1) - q(1))/(\kappa - q(2))$ for $T+\epsilon \leq t \leq T'$, so that ϕ traces out the line segment from $\phi_{T+\epsilon}$ to $\phi_{T'}$, and thus, $\phi_t(1) > \phi_t(2)$ for $0 < t < T'$. Therefore, $\Gamma^{LLR}(\phi,0) = H_{q(1)}(0,\kappa) + H_{q(2)}(0,\kappa) + f(\epsilon)$ for some $f(\epsilon) \to 0$ as $\epsilon \to 0$.

Case 3: $\kappa > \kappa_*(2)$. Let $T > 0$ and $(\phi_t : 0 \le t \le T)$ be constructed as in Case 2 with $\kappa = \kappa_*(2)$. Note that $\kappa_*(2) > q(1)$; thus, $\phi_T(1) = \phi_T(2) = \kappa_*(2)$. Extend ϕ by letting $\phi(1)$ relax under $\lambda(1)+\lambda(2)$ and $\phi(2)$ relax in reverse time under $\lambda(3)$ so that $\phi_{T'}(2) = \kappa$ at some time $T' > T$. Note that $(\phi_t(1) - q(1))(\phi_t(2) - q(2)) = (\kappa_*(2) - q(1))(\kappa_*(2) - q(2))$ and $\phi_t(2) > \phi_t(1)$ for $T < t \le T'$; therefore, $\Gamma^{LLR}(\phi, 0) = H_{q(1)}(0, \kappa_*(2)) + H_{q(2)}(0, \kappa_*(2)) + H_{\lambda(3)}(\kappa_*(2), \kappa) + f(\epsilon)$ for some $f(\epsilon) \to 0$ as $\epsilon \to 0$.

Figure 3.2 sketches the function ϕ constructed above. The arbitrariness of $\epsilon > 0$ establishes the lemma. □

Proof of Theorem 3.1.4. The proof of Theorem 3.1.1, with Proposition 3.4.1 and Lemma 3.4.6 (Lemma 3.4.8) in place of Lemmas 3.3.1 and 3.3.4, respectively, and an adaptation of Lemma 3.3.3, applied on location 1 (location 2), establishes the existence and the desired form of $F^{LLR}(1, \kappa)$ ($F^{LLR}(2, \kappa)$). Since $H_{\lambda(1)}(0, \kappa) \le H_{\lambda(3)}(0, \kappa)$ and $\kappa_*(v)$ minimizes the expression $H_{q(1)}(0, \kappa_*(v) \wedge \kappa) + H_{q(2)}(0, \kappa_*(v) \wedge \kappa) + H_{\lambda(2v-1)}(\kappa_*(v) \wedge \kappa, \kappa)$ for each $v = 1, 2$, it follows that $F^{LLR}(\kappa) = F^{LLR}(1, \kappa) \le F^{LLR}(2, \kappa)$. □

3.5 Large Deviations Principle for the W Network under Least Load Routing

This section proves Proposition 3.4.1, the large deviations principle satisfied by the normalized load process X^γ under the LLR policy. The proof entails an application of the theory of large deviations of Markov processes with discontinuous transition mechanisms (see Chapter 4, [14], [15], [12]). The transition mechanism of X^γ changes smoothly in the two open halves of R_+^2 separated by the hyperplane $\{x \in R^2 : x(1) = x(2)\}$ so that the process X^γ nearly conforms to the conditions of Theorem 4.2.1 of Chapter 4. The theorem does not apply directly, however, because for $v = 1, 2$ the log rates $\log(\nu(x, e_v^-))$ are neither bounded above (since $\nu(x, e_v^-) \to \infty$ as $x(v) \to \infty$) nor continuous and finite over R_+^2 (since $\nu(x, e_v^-) \to 0$ as $x(v) \to 0$). We therefore establish Lemma 3.4.1 by approximating X^γ by a sequence of auxiliary processes each of which conforms to the conditions of Theorem 4.2.1 and adapting the techniques used in [12, Section 12.6] for the one-dimensional Erlang model.

The outline of the proof is as follows: Lemma 3.5.2 identifies the large deviations principle satisfied by each auxiliary process. Lemma 3.5.6 establishes the goodness of the rate function Γ^{LLR}. Based on a coupling of the auxiliary processes with the load process, Lemmas 3.5.8 and 3.5.9 prove the large deviations upper and lower bounds. We start with the following lemma:

Lemma 3.5.1 *For $x_1, x_2 \geq 0$, $\sigma_1, \sigma_2 > 0$, $y \in R$, and $\beta \in (0,1)$,*

$$\inf_{\substack{y_1 \in R,\ y_2 \in R \\ \beta y_1 + (1-\beta) y_2 = y}} \{\ \beta \Lambda_{\sigma_1}(x_1, y_1) + (1-\beta)\Lambda_{\sigma_2}(x_2, y_2)\ \} = \Lambda_{\beta\sigma_1 + (1-\beta)\sigma_2}(\beta x_1 + (1-\beta)x_2, y). \quad (3.15)$$

There exists a unique solution to the left-hand side of (3.15) that satisfies

$$\frac{y_1 + \sqrt{(y_1)^2 + 4\sigma_1 x_1}}{2\sigma_1} = \frac{y_2 + \sqrt{(y_2)^2 + 4\sigma_2 x_2}}{2\sigma_2}. \quad (3.16)$$

Proof. The function $\beta\Lambda_{\sigma_1}(x_1, y_1) + (1-\beta)\Lambda_{\sigma_2}(x_2, (y - \beta y_1)/(1-\beta)) \to \infty$ as $|y_1| \to \infty$ and is strictly convex in y_1. Therefore, it achieves its minimum at a unique stationary point, which satisfies equality (3.16) with y_2 defined by $\beta y_1 + (1-\beta)y_2 = y$. The quantity on both sides of (3.16) is the nonnegative root of the equation $\sigma_v z^2 - y_v z - x_v = 0$ for each $v = 1, 2$. This quantity is therefore equal to the nonnegative root of the equation $(\beta\sigma_1 + (1-\beta)\sigma_2)z^2 - (\beta y_1 + (1-\beta)y_2)z - (\beta x_1 + (1-\beta)x_2) = 0$ so that

$$\frac{y_1 + \sqrt{(y_1)^2 + 4\sigma_1 x_1}}{2\sigma_1} = \frac{y_2 + \sqrt{(y_2)^2 + 4\sigma_2 x_2}}{2\sigma_2} = \frac{y + \sqrt{y^2 + 4(\beta\sigma_1 + (1-\beta)\sigma_2)(\beta x_1 + (1-\beta)x_2)}}{2(\beta\sigma_1 + (1-\beta)\sigma_2)}.$$

Equality (3.15) follows by direct substitution. □

Given $0 < \epsilon \leq 1$, let $Y^{\gamma,\epsilon}$ denote the Markov process generated by the pair (γ, ν^ϵ), where for each $x \in R^2$,

$$\nu^\epsilon(x, \{e_1^-\}) = [x(1)]_\epsilon^{1/\epsilon}, \quad \nu^\epsilon(x, \{e_1^+\}) = \begin{cases} \lambda(1) + \lambda(2) & \text{if } x(1) \leq x(2) \\ \lambda(1) & \text{if } x(1) > x(2), \end{cases}$$

$$\nu^\epsilon(x, \{e_2^-\}) = [x(2)]_\epsilon^{1/\epsilon}, \quad \nu^\epsilon(x, \{e_2^+\}) = \begin{cases} \lambda(3) & \text{if } x(1) \leq x(2) \\ \lambda(3) + \lambda(2) & \text{if } x(1) > x(2), \end{cases}$$

and let $\bar{Y}^{\gamma,\epsilon}$ denote the polygonal interpolation of $Y^{\gamma,\epsilon}$.

Lemma 3.5.2 *For each $0 < \epsilon \leq 1$, the sequence $(\tilde{Y}^{\gamma,\epsilon} : \gamma > 0)$ satisfies the large deviations principle in $C_{[0,T]}(R^2)$ with the good rate function Γ^ϵ, where for each $\phi \in C_{[0,T]}(R^2)$ and $x \in R^2$,*

$$\Gamma^\epsilon(\phi, x) = \begin{cases} \int_0^T \Lambda^\epsilon(\phi_t, \dot{\phi}_t)dt & \text{if } \phi_0 = x \text{ and } \phi \text{ is absolutely continuous} \\ +\infty & \text{otherwise,} \end{cases}$$

and $\Lambda^\epsilon(\phi_t, \dot{\phi}_t) = \Lambda([\phi_t]_\epsilon^{1/\epsilon}, \dot{\phi}_t)$.

Proof. Let $A^o = \{x \in R^2 : x(1) = x(2)\}$, $A^+ = \{x \in R^2 : x(1) < x(2)\}$, and $A^- = \{x \in R^2 : x(1) > x(2)\}$, and let the rate-measure fields $\nu^{+,\epsilon}$ and $\nu^{-,\epsilon}$ be defined as

$$\nu^{+,\epsilon}(x) = \begin{cases} \nu^\epsilon(x) & \text{if } x \in \overline{A^+} \\ \nu^\epsilon(x(1), x(1)) & \text{if } x \in A^- \end{cases} \qquad \nu^{-,\epsilon}(x) = \begin{cases} \nu^\epsilon(x) & \text{if } x \in A^- \\ \lim_{\delta \searrow 0} \nu^\epsilon(x(2)+\delta, x(2)-\delta) & \text{if } x \in \overline{A^+}. \end{cases}$$

Note that $\nu^{+,\epsilon}$, $\nu^{-,\epsilon}$, and $Y^{\gamma,\epsilon}$ satisfy the conditions of Theorem 4.2.1; therefore, the sequence $(\tilde{Y}^{\gamma,\epsilon} : \gamma > 0)$ satisfies a large deviations principle with the good rate function Γ^ϵ, where for $x, y \in R^2$

$$\begin{aligned} \Lambda^\epsilon(x,y) &= I\{\phi_t \in A^+\}\Lambda^{+,\epsilon}(x,y) + I\{\phi_t \in A^o\}\Lambda^{o,\epsilon}(x,y) + I\{\phi_t \in A^-\}\Lambda^{-,\epsilon}(x,y), \\ \Lambda^{\pm,\epsilon}(x,y) &= \sum_{v=1}^{2} \sup_{\zeta \in R} \left\{ \zeta y(v) - \left((e^\zeta - 1)\nu^{\pm,\epsilon}(x, e_v^+) + (e^{-\zeta} - 1)\nu^{\pm,\epsilon}(x, e_v^-) \right) \right\}, \qquad (3.17) \\ \Lambda^{o,\epsilon}(x,y) &= \inf_{\substack{0 \leq \beta \leq 1,\ y^+ \in \overline{A^-}, y^- \in \overline{A^+} \\ \beta y^+ + (1-\beta)y^- = y}} \{ \beta\Lambda^{+,\epsilon}(x, y^+) + (1-\beta)\Lambda^{-,\epsilon}(x, y^-) \}. \end{aligned}$$

Applying [12, Exercise 7.24] on each term on the right-hand side of equation (3.17) yields that

$$\begin{aligned} \Lambda^{+,\epsilon}(x,y) &= \Lambda_{\lambda(1)+\lambda(2)}([x(1)]_\epsilon^{1/\epsilon}, y(1)) + \Lambda_{\lambda(3)}([x(2)]_\epsilon^{1/\epsilon}, y(2)) \quad \text{for } x \in \overline{A^+}, \\ \Lambda^{-,\epsilon}(x,y) &= \Lambda_{\lambda(1)}([x(1)]_\epsilon^{1/\epsilon}, y(1)) + \Lambda_{\lambda(2)+\lambda(3)}([x(2)]_\epsilon^{1/\epsilon}, y(2)) \quad \text{for } x \in \overline{A^-}. \end{aligned}$$

To complete the proof of the lemma, it remains to evaluate $\Lambda^{o,\epsilon}(x,y)$ for $x \in A^o$. Note that for absolutely continuous ϕ, $\dot{\phi}_t \in A^o$ for almost all t such that $\phi_t \in A^o$; therefore, it suffices to consider the case when $y \in A^o$. Fix $x, y \in A^o$. For any $y^+, y^- \in R^2$ and $\beta \in [0,1]$ such that $\beta y^+ + (1-\beta)y^- = y$,

$$\beta\Lambda^{+,\epsilon}(x, y^+) \quad + \quad (1-\beta)\Lambda^{-,\epsilon}(x, y^-)$$

$$
\begin{aligned}
&= \beta\Lambda_{\lambda(1)+\lambda(2)}([x(1)]_\epsilon^{1/\epsilon}, y^+(1)) + \beta\Lambda_{\lambda(3)}([x(2)]_\epsilon^{1/\epsilon}, y^+(2)) \\
&+ (1-\beta)\Lambda_{\lambda(1)}([x(1)]_\epsilon^{1/\epsilon}, y^-(1)) + (1-\beta)\Lambda_{\lambda(2)+\lambda(3)}([x(2)]_\epsilon^{1/\epsilon}, y^-(2)) \qquad (3.18) \\
&\geq \Lambda_{\lambda(1)+\beta\lambda(2)}([x(1)]_\epsilon^{1/\epsilon}, y(1)) + \Lambda_{(1-\beta)\lambda(2)+\lambda(3)}([x(2)]_\epsilon^{1/\epsilon}, y(2)) \qquad (3.19) \\
&\geq \inf_{0\leq\beta'\leq 1}\{\, \Lambda_{\lambda(1)+\beta'\lambda(2)}([x(1)]_\epsilon^{1/\epsilon}, y(1)) + \Lambda_{(1-\beta')\lambda(2)+\lambda(3)}([x(2)]_\epsilon^{1/\epsilon}, y(2)) \,\} \qquad (3.20) \\
&= \Lambda_{q(1)}([x(1)]_\epsilon^{1/\epsilon}, y(1)) + \Lambda_{q(2)}([x(2)]_\epsilon^{1/\epsilon}, y(2)). \qquad (3.21)
\end{aligned}
$$

In the above argument inequality (3.19) follows by applying Lemma 3.5.1 separately on the first and third and the second and fourth terms on the right-hand side of equality (3.18). Since $x(1) = x(2)$ and $y(1) = y(2)$, each of the terms of the right-hand side of inequality (3.19) is the same convex function evaluated at $\lambda(1) + \beta\lambda(2)$ and $(1-\beta)\lambda(2) + \lambda(3)$, respectively. Equality (3.21) follows by straightforward minimization.

We next identify $\Lambda^{o,\epsilon}(x,y)$ with the right-hand side of inequality (3.21) by establishing the existence of $\beta \in [0,1]$, $y^+ \in \overline{A^-}$, and $y^- \in \overline{A^+}$ such that $\beta y^+ + (1-\beta)y^- = y$ and both inequalities (3.19) and (3.20) are satisfied with equality. Inequality (3.20) is satisfied with equality if $\beta = [(\lambda(3) - \lambda(1) + \lambda(2))/2\lambda(2)]_0^1$. If $\beta = 0$ ($\beta = 1$), then take $y^- = y$ ($y^+ = y$). Otherwise, by Lemma 3.5.1 there exist $y^+, y^- \in R^2$ such that $\beta y^+ + (1-\beta)y^- = y$, inequality (3.19) is satisfied with equality, and the following two conditions hold:

$$
\frac{y^-(1) + \sqrt{(y^-(1))^2 + 4\lambda(1)[x(1)]_\epsilon^{1/\epsilon}}}{2\lambda(1)} = \frac{y^+(1) + \sqrt{(y^+(1))^2 + 4(\lambda(1)+\lambda(2))[x(1)]_\epsilon^{1/\epsilon}}}{2(\lambda(1)+\lambda(2))}, \qquad (3.22)
$$

$$
\frac{y^+(2) + \sqrt{(y^+(2))^2 + 4\lambda(3)[x(2)]_\epsilon^{1/\epsilon}}}{2\lambda(3)} = \frac{y^-(2) + \sqrt{(y^-(2))^2 + 4(\lambda(2)+\lambda(3))[x(2)]_\epsilon^{1/\epsilon}}}{2(\lambda(2)+\lambda(3))}. \qquad (3.23)
$$

The left-hand side of equality (3.22) is increasing in $y^-(1)$ and decreasing in $\lambda(1)$; therefore, (3.22) and (3.23), respectively, imply that $y^+(1) \geq y^-(1)$ and $y^-(2) \geq y^+(2)$. This, together with the assumption that $\beta y^+ + (1-\beta)y^- \in A^o$, implies that $y^- \in \overline{A^+}$ and $y^+ \in \overline{A^-}$. The proof of the lemma is complete. □

For $x \in R_+$ and $\epsilon = 0$ set $[x]_\epsilon^{1/\epsilon} = x$, so that $\Gamma = \Gamma^\epsilon|_{\epsilon=0}$.

Lemma 3.5.3 *Let S be a finite set of positive numbers. For each $l \geq 0$ there exists an $M > 0$ such that for any absolutely continuous $\phi \in C_{[0,T]}(R_+)$ and $0 \leq \epsilon \leq 1$,*

$$\int_0^T \inf_{\sigma \in S} \Lambda_\sigma([\phi_t]_\epsilon^{1/\epsilon}, \dot{\phi}_t)dt \leq l \Longrightarrow \sup_{0 \leq t \leq T} (\phi_t - \phi_0) \leq M.$$

Proof. Examination of $\partial \Lambda_\sigma(x,y)/\partial x$ yields that $\Lambda_\sigma(x,y)$ is increasing in x whenever $y > \sigma$; thus, $\lim_{y \to \infty} \Lambda_\sigma(x,y)/y = \infty$ uniformly in $x \in R_+$ and $\sigma \in S$. Given $l \geq 0$, choose a constant $B(l)$ large enough so that $\inf_{\sigma \in S, x \in R_+} \Lambda_\sigma(x,y)/y \geq l$ whenever $y \geq B(l)$. For absolutely continuous $\phi \in C_{[0,T]}(R_+)$ and $0 \leq \tau \leq T$,

$$\begin{aligned}
\int_0^T \inf_{\sigma \in S} \Lambda_\sigma([\phi_t]_\epsilon^{1/\epsilon}, \dot{\phi}_t)dt &\geq \int_{\{t \in [0,\tau]: \dot{\phi}_t \geq B(l)\}} \inf_{\sigma \in S} \frac{\Lambda_\sigma([\phi_t]_\epsilon^{1/\epsilon}, \dot{\phi}_t)}{\dot{\phi}_t} \dot{\phi}_t dt \\
&\geq \int_{\{t \in [0,\tau]: \dot{\phi}_t \geq B(l)\}} l\dot{\phi}_t dt \\
&\geq \int_0^\tau l(\dot{\phi}_t - B(l))dt \\
&\geq l(\phi_\tau - \phi_0 - B(l)T).
\end{aligned}$$

Choosing $M = B(l)T + 1$ establishes the lemma. □

Lemma 3.5.4 *(Relative Compactness) For each $x_0 \in R_+^2$ and $l \geq 0$, the collection $C(l) = \cup_{0 \leq \epsilon \leq 1} \{\phi : \Gamma^\epsilon(\phi, x_0) \leq l\}$ is relatively compact in $C_{[0,T]}(R_+^2)$.*

Proof. If $\phi \in C(l)$, then ϕ is absolutely continuous, $\phi_0 = x_0$, and for some $0 \leq \epsilon \leq 1$,

$$\begin{aligned}
l &\geq \int_0^T \Lambda^\epsilon(\phi_t, \dot{\phi}_t)dt \\
&\geq \int_0^T \inf_{\sigma \in S} \Lambda_\sigma([\phi_t(1)]_\epsilon^{1/\epsilon}, \dot{\phi}_t(1))dt + \int_0^T \inf_{\sigma \in S} \Lambda_\sigma([\phi_t(2)]_\epsilon^{1/\epsilon}, \dot{\phi}_t(2))dt, \qquad (3.24)
\end{aligned}$$

where $S = \{\lambda(1), q(1), \lambda(1) + \lambda(2), \lambda(3), q(2), \lambda(2) + \lambda(3)\}$. Lemma 3.5.3, applied separately on the terms of the right-hand side of (3.24), implies the existence of a finite $M > 1$ such that $\sup_{0 \leq t \leq T} |\phi_t| \leq M$ for all $\phi \in C(l)$.

Fix $\delta > 0$. Choose a constant $B(\delta)$ large enough so that $\inf_{0 \leq x \leq M, \sigma \in S} \Lambda_\sigma(x,y)/|y| > 2l/\delta$ whenever $|y| > B(\delta)$. Let $((s_j, t_j) : j = 1, \cdots, J)$ be a finite collection of nonoverlapping intervals

in $[0,T]$, and set $D = \cup_j (s_j, t_j)$. Given $\phi \in C(l)$, let $0 \le \epsilon \le 1$ be such that $\Gamma^\epsilon(\phi, x_0) \le l$. Then

$$\begin{aligned}
\sum_{j=1}^{J} |\phi_{t_j} - \phi_{s_j}| &\le \int_D |\dot\phi_t| dt \\
&= \int_{D \cap \{t : |\dot\phi_t| > B(\delta)\}} \frac{|\dot\phi_t|}{\Lambda^\epsilon(\phi_t, \dot\phi_t)} \Lambda^\epsilon(\phi_t, \dot\phi_t) dt + \int_{D \cap \{t : |\dot\phi_t| \le B(\delta)\}} |\dot\phi_t| dt \\
&\le \frac{\delta}{2l} \int_{D \cap \{t : |\dot\phi_t| > B(\delta)\}} \Lambda^\epsilon(\phi_t, \dot\phi_t) dt + B(\delta) \sum_{j=1}^{J} |t_j - s_j| \\
&\le \frac{\delta}{2} + B(\delta) \sum_{j=1}^{J} |t_j - s_j|.
\end{aligned}$$

Thus, $\sum_{j=1}^{J} |\phi_{t_j} - \phi_{s_j}| \le \delta$ whenever $\sum_{j=1}^{J} |t_j - s_j| \le \delta / 2B(\delta)$, uniformly for all $\phi \in C(l)$. The Arzela-Ascoli Theorem implies the relative compactness of $C(l)$. □

Lemma 3.5.5 *(Lower Semicontinuity) Given $x_0 \in R_+^2$, the function $\Gamma^{LLR}(\cdot, x_0) : C_{[0,T]}(R_+^2) \to R_+ \cup \{+\infty\}$ is lower semicontinuous.*

Proof. Let $(\phi^m : m \ge 1)$ be a sequence such that $\phi^m \to \phi$ in $C_{[0,T]}(R_+^2)$. To prove the lemma, it suffices to show that $\Gamma^{LLR}(\phi, x_0) \le \liminf_{m\to\infty} \Gamma^{LLR}(\phi^m, x_0)$. Assume, without loss of generality, the existence of $l, k \ge 0$ such that $\Gamma^{LLR}(\phi^m, x_0) \le l$ for all $m \ge k$. The proof of Lemma 3.5.4 shows that the sequence $(\phi^m : m \ge k)$ is uniformly absolutely continuous; therefore, ϕ is absolutely continuous, $\phi_0 = x_0$, and by the explanations indicated in parentheses,

$$\begin{aligned}
\Gamma^{LLR}(\phi, x_0) &= \int_0^T \Lambda(\phi_t, \dot\phi_t) dt && \text{(Definition of } \Gamma^{LLR}) \\
&\le \liminf_{\epsilon \searrow 0} \int_0^T \Lambda^\epsilon(\phi_t, \dot\phi_t) dt && \text{(Fatou's Lemma)} \\
&\le \liminf_{\epsilon \searrow 0} \liminf_{m\to\infty} \int_0^T \Lambda^\epsilon(\phi_t^m, \dot\phi_t^m) dt && \text{(L.s.c. property of } \Gamma^\epsilon) \\
&\le \liminf_{\epsilon \searrow 0} \liminf_{m\to\infty} \left(\int_0^T \Lambda(\phi_t^m \wedge \tfrac{1}{\epsilon}, \dot\phi_t^m) dt + 2\epsilon T \right) && \text{(Lemma 3.3.2)} \\
&= \liminf_{m\to\infty} \Gamma^{LLR}(\phi^m, x_0). && (\sup_{0\le t\le T} |\phi_t^m| \text{ is bounded})
\end{aligned}$$

This establishes the lemma. □

Lemma 3.5.6 *(Goodness) Γ^{LLR} is a good rate function.*

Proof. Note that for each $l \ge 0$ the level set $\{\phi : \Gamma^{LLR}(\phi, x_0) \le l\}$ is contained in $C(l)$. Lemmas 3.5.4 and 3.5.5 imply, respectively, the relative compactness and closedness, and therefore the compactness, of the level set. □

Lemma 3.5.7 *For $x_0 \in R_+^2$ and closed $F \subset C_{[0,T]}(R_+^2)$,*

$$\inf_{\phi \in F} \Gamma^{LLR}(\phi, x_0) \leq \limsup_{\epsilon \searrow 0} \inf_{\phi \in F} \Gamma^{\epsilon}(\phi, x_0).$$

Proof. Without loss of generality, we may assume the existence of an $l \geq 0$ such that $\inf_{\phi \in F} \Gamma^{\epsilon}(\phi, x_0) \leq l$ for each $0 < \epsilon \leq 1$. For each such ϵ, appeal to the goodness of the rate function Γ^{ϵ} to choose a $\phi^{\epsilon} \in F$ such that $\Gamma^{\epsilon}(\phi^{\epsilon}, x_0) = \inf_{\phi \in F} \Gamma^{\epsilon}(\phi, x_0)$. By Lemma 3.5.4 the collection $(\phi^{\epsilon} : 0 < \epsilon \leq 1)$ is relatively compact; hence there exists a sequence $\epsilon_n \to 0$ and $\tilde{\phi} \in F$ such that $\phi^{\epsilon_n} \to \tilde{\phi}$. Define

$$\xi_t^{\epsilon_n} = \begin{cases} x_0 + (t,t) & 0 \leq t \leq \epsilon_n \\ (\epsilon_n, \epsilon_n) + \phi_{t-\epsilon_n}^{\epsilon_n} & \epsilon_n \leq t \leq T. \end{cases}$$

Note that $\xi^{\epsilon_n} \to \tilde{\phi}$, and

$$\begin{aligned} \int_0^T \Lambda(\xi_t^{\epsilon_n}, \dot{\xi}_t^{\epsilon_n}) dt &= \int_0^{\epsilon_n} \Lambda(\xi_t^{\epsilon_n}, (1,1)) dt + \int_{\epsilon_n}^T \Lambda^{\epsilon_n}((\epsilon_n, \epsilon_n) + \phi_{t-\epsilon_n}^{\epsilon_n}, \dot{\phi}_{t-\epsilon_n}^{\epsilon_n}) dt \\ &\leq \int_0^{\epsilon_n} (\Lambda(x_0, (1,1)) + 2\epsilon_n) dt + \int_0^T (\Lambda^{\epsilon_n}(\phi_t^{\epsilon_n}, \dot{\phi}_t^{\epsilon_n}) + 2\epsilon_n) dt, \end{aligned} \tag{3.25}$$

where the first step follows by the definition of Λ^{ϵ_n} and the construction of ξ^{ϵ_n}, and the second step follows by Lemma 3.3.2 and the nonnegativity of Λ^{ϵ_n}. Therefore, by the explanations indicated in parentheses,

$$\begin{aligned} \inf_{\phi \in F} \Gamma^{LLR}(\phi, x_0) &\leq \Gamma^{LLR}(\tilde{\phi}, x_0) && (\tilde{\phi} \in F) \\ &\leq \liminf_{n \to \infty} \Gamma^{LLR}(\xi^{\epsilon_n}, x_0) && \text{(Lemma 3.5.5)} \\ &\leq \liminf_{n \to \infty} \Gamma^{\epsilon_n}(\phi^{\epsilon_n}, x_0) && \text{(Inequality (3.25))} \\ &\leq \limsup_{\epsilon \searrow 0} \inf_{\phi \in F} \Gamma^{\epsilon}(\phi, x_0) && \text{(Definition of } \phi^{\epsilon}), \end{aligned}$$

and the lemma is established. □

For each $\gamma > 0$ construct X^{γ} on an appropriately extended probability space, and for each $0 < \epsilon \leq 1$ construct the process $Y^{\gamma,\epsilon}$ on the same space as follows: Let $N(1)$ and $N(2)$ be mutually independent Poisson processes, which are also independent of X^{γ}, each having rate $\gamma\epsilon$. Set $Y_0^{\gamma,\epsilon} \equiv X_0^{\gamma}$. Let $\tau = \inf\{t \geq 0 : X_t^{\gamma} \notin [0, 1/\epsilon]^2 \text{ or } N_t \neq 0\}$. At every time $t \leq \tau$ such that X^{γ} jumps, $Y^{\gamma,\epsilon}$ takes the same jump. In addition, for $v = 1, 2$, at every time $t \leq \tau$ such that $X_t^{\gamma}(v) \leq \epsilon$

and $N(v)$ jumps, $Y^{\gamma,\epsilon}(v)$ jumps down by γ^{-1} with probability $(\epsilon - X_t^\gamma(v))/\epsilon$. After time τ the construction is done so that $Y^{\gamma,\epsilon}$ is generated by the specified pair (γ, ν^ϵ).

Let $\tilde{X}^\gamma$ and $\tilde{Y}^{\gamma,\epsilon}$ denote the polygonal interpolations of X^γ and $Y^{\gamma,\epsilon}$, respectively.

Lemma 3.5.8 *(Upper Bound) For any closed $F \subset C_{[0,T]}(R_+^2)$, $x_0 \in R_+^2$, and sequence $(x^\gamma : \gamma > 0)$ such that $x^\gamma \to x_0$,*

$$\limsup_{\gamma\to\infty} \gamma^{-1} \log P_{x^\gamma}\left((\tilde{X}_t^\gamma : 0 \le t \le T) \in F\right) \le -\inf_{\phi\in F} \Gamma^{LLR}(\phi, x_0).$$

Proof. Note that for each $\gamma > 0$ and $0 < \epsilon \le 1$,

$$\begin{aligned}
P_{x^\gamma}\left((\tilde{Y}_t^{\gamma,\epsilon} : 0 \le t \le T) \in F\right) &\ge P_{x^\gamma}\left((\tilde{Y}_t^{\gamma,\epsilon} : 0 \le t \le T) \in F, \sup_{0\le t\le T} |\tilde{Y}_t^{\gamma,\epsilon}| < \frac{1}{\epsilon}, N_T = 0\right) \\
&= P_{x^\gamma}\left((\tilde{X}_t^{\gamma} : 0 \le t \le T) \in F, \sup_{0\le t\le T} |\tilde{X}_t^{\gamma}| < \frac{1}{\epsilon}, N_T = 0\right) \\
&= P_{x^\gamma}\left((\tilde{X}_t^{\gamma} : 0 \le t \le T) \in F, \sup_{0\le t\le T} |\tilde{X}_t^{\gamma}| < \frac{1}{\epsilon}\right) P_{x^\gamma}(N_T = 0) \\
&\ge \left(P_{x^\gamma}\left((\tilde{X}_t^{\gamma} : 0 \le t \le T) \in F\right) - P_{x^\gamma}(\sup_{0\le t\le T} |\tilde{X}_t^{\gamma}| \ge \frac{1}{\epsilon})\right) e^{-2\gamma\epsilon T},
\end{aligned}$$

where the second step follows by the construction of the processes X^γ and $Y^{\gamma,\epsilon}$, and the third step follows by the independence of X^γ and N. In view of the above inequality,

$$\begin{aligned}
&\limsup_{\gamma\to\infty} \gamma^{-1} \log P_{x^\gamma}\left((\tilde{X}_t^\gamma : 0 \le t \le T) \in F\right) \le \\
&\left(\limsup_{\gamma\to\infty} \gamma^{-1} \log P_{x^\gamma}\left((\tilde{Y}_t^{\gamma,\epsilon} : 0 \le t \le T) \in F\right) + 2\epsilon T\right) \bigvee \limsup_{\gamma\to\infty} \gamma^{-1} \log P_{x^\gamma}(\sup_{0\le t\le T} |\tilde{X}_t^{\gamma}| \ge \frac{1}{\epsilon}) \quad (3.26)
\end{aligned}$$

for each $0 < \epsilon \le 1$. Note that $\sup_{0\le t\le T} |X_t|$ is stochastically dominated by a Poisson random variable of mean $\gamma(\lambda(1)+\lambda(2)+\lambda(3))T$; hence the second term in the right-hand side of inequality (3.26) is arbitrarily large for small ϵ. Therefore,

$$\begin{aligned}
\limsup_{\gamma\to\infty} \gamma^{-1} \log P_{x^\gamma}\left((\tilde{X}_t^\gamma : 0 \le t \le T) \in F\right) &\le \liminf_{\epsilon\searrow 0} \limsup_{\gamma\to\infty} \gamma^{-1} \log P_{x^\gamma}\left((\tilde{Y}_t^{\gamma,\epsilon} : 0 \le t \le T) \in F\right) \\
&\le -\limsup_{\epsilon\searrow 0} \inf_{\phi\in F} \Gamma^\epsilon(\phi, x_0)
\end{aligned}$$

$$\leq \; -\inf_{\phi \in F} \Gamma^{LLR}(\phi, x_0),$$

where the second step is a consequence of Lemma 3.5.2, and the third step follows by Lemma 3.5.7. This establishes the lemma. □

Lemma 3.5.9 *(Lower Bound) For any open $G \subset C_{[0,T]}(R_+^2)$, $x_0 \in R_+^2$, and sequence $(x^\gamma : \gamma > 0)$ such that $x^\gamma \to x_0$,*

$$\liminf_{\gamma \to \infty} \gamma^{-1} \log P_{x^\gamma}\left((\tilde{X}_t^\gamma : 0 \leq t \leq T) \in G\right) \geq -\inf_{\phi \in G} \Gamma^{LLR}(\phi, x_0).$$

Proof. Fix $\epsilon > 0$ and $\phi \in G$. Without loss of generality, we may assume that ϕ is absolutely continuous and $\phi_0 = x_0$. Let $\delta > 0$ be small enough such that the open ball of radius 6δ around ϕ, $B(\phi, 6\delta)$, is contained in G. By Lemma 2.3.3 of Section 2.3.2, as $\gamma \to \infty$, the process $(X_t^\gamma : 0 \leq t \leq T)$ converges weakly to a Lipschitz continuous function $(x_t : 0 \leq t \leq T)$ that satisfies $x_t(1) \wedge x_t(2) > 0$ for $t > 0$. Let $d = \sup_{0 \leq t \leq T} |\dot{x}_t|$, and choose a positive $\sigma < \delta/d$ such that

$$|\phi_t - \phi_s| < \delta \;\text{ and }\; |x_t - x_s| < \delta \quad \text{whenever } |t - s| \leq \sigma.$$

Construct ξ as

$$\xi_0 = x_0, \qquad \dot{\xi}_t = \begin{cases} \dot{x}_t & 0 \leq t \leq \sigma \\ (2d, 2d) & \sigma \leq t \leq 2\sigma \\ \dot{\phi}_{t-2\sigma} & 2\sigma \leq t \leq T. \end{cases}$$

It can be easily verified that $|\xi_t - \phi_t| < 5\delta$ for $0 \leq t \leq T$, and $\xi_t(1) \wedge \xi_t(2) \geq x_\sigma(1) \wedge x_\sigma(2) > 0$ for $\sigma \leq t \leq T$. Choose positive $\eta < (x_\sigma(1) \wedge x_\sigma(2) \wedge \delta)/2$ small enough, and choose δ and σ smaller, if necessary, so that

$$\begin{aligned} \int_\sigma^T \Lambda^\eta(\xi_t, \dot{\xi}_t) dt \;&\leq\; \int_\sigma^{2\sigma} (\Lambda(x_\sigma, (2s, 2s)) + 4\delta + 2\eta) dt + \int_{2\sigma}^T (\Lambda(\phi_{t-2\sigma}, \dot{\phi}_{t-2\sigma}) + 6\delta + 2\eta) dt \\ &\leq\; \int_0^T \Lambda(\phi_t, \dot{\phi}_t) dt + \frac{\epsilon}{2}, \end{aligned}$$

where the first inequality follows by Lemma 3.3.2 and the fact that for $v = 1, 2$, $x_\sigma(v) \leq \xi_t(v) \leq x_\sigma(v) + 2\delta$ for $\sigma \leq t \leq 2\sigma$ and $\phi_{t-2\sigma}(v) \leq \xi_t(v) \leq \phi_{t-2\sigma}(v) + 3\delta$ for $2\sigma \leq t \leq T$. Finally, appeal to

the time-homogeneous Markov property of $Y^{\gamma,\eta}$ and Lemma 3.5.2 together with Remark 3.2.1 to choose $\rho < \eta$ small enough so that for large enough γ,

$$\inf_{|x-\xi_\sigma|<\rho} P_{x^\gamma}(\sup_{\sigma\le t\le T}|\tilde{Y}_t^{\gamma,\eta}-\xi_t|<\eta \mid \tilde{Y}_\sigma^{\gamma,\eta}=x) \ge \exp\left(-\gamma(\int_\sigma^T \Lambda^\eta(\xi_t,\dot{\xi}_t)dt+\frac{\epsilon}{2})\right).$$

For large enough γ,

$$\begin{aligned}
P_{x^\gamma}\left((\tilde{X}_t^\gamma : 0\le t\le T)\in G\right) &\ge P_{x^\gamma}\left((\tilde{X}_t^\gamma : 0\le t\le T)\in B(\phi,6\delta)\right)\\
&\ge P_{x^\gamma}\left((\tilde{X}_t^\gamma : 0\le t\le T)\in B(\xi,\eta)\right) \qquad (3.27)\\
&\ge P_{x^\gamma}(\sup_{0\le t\le\sigma}|\tilde{X}_t^\gamma-\xi_t|<\rho)\\
&\qquad \inf_{|x-\xi_\sigma|<\rho} P_{x^\gamma}(\sup_{\sigma\le t\le T}|\tilde{X}_t^\gamma-\xi_t|<\eta \mid \tilde{X}_\sigma^\gamma=x) \qquad (3.28)\\
&= P_{x^\gamma}(\sup_{0\le t\le\sigma}|\tilde{X}_t^\gamma-x_t|<\rho)\\
&\qquad \inf_{|x-\xi_\sigma|<\rho} P_{x^\gamma}(\sup_{\sigma\le t\le T}|\tilde{Y}_t^{\gamma,\eta}-\xi_t|<\eta \mid \tilde{Y}_\sigma^{\gamma,\eta}=x) \qquad (3.29)\\
&\ge \frac{1}{2}\exp\left(-\gamma(\int_\sigma^T \Lambda^\eta(\xi_t,\dot{\xi}_t)dt+\frac{\epsilon}{2})\right)\\
&\ge \frac{1}{2}\exp\left(-\gamma(\int_0^T \Lambda(\phi_t,\dot{\phi}_t)dt+\epsilon)\right).
\end{aligned}$$

In the above argument, inequality (3.27) follows by the fact that $B(\xi,\eta)\subset B(\xi,\delta)\subset B(\phi,6\delta)$, inequality (3.28) is a consequence of the choice of ρ and the Markov property of X^γ, and equality (3.29) is implied by the choice of η and the construction of $Y^{\gamma,\eta}$. The arbitrariness of $\epsilon > 0$ implies that

$$\liminf_{\gamma\to\infty}\gamma^{-1}\log P_{x^\gamma}\left((\tilde{X}_t^\gamma : 0\le t\le T)\in G\right) \ge -\Gamma^{LLR}(\phi,x_0),$$

and the arbitrariness of $\phi\in G$ establishes the lemma. □

CHAPTER 4

LARGE DEVIATIONS OF MARKOV PROCESSES WITH DISCONTINUOUS STATISTICS

4.1 Overview of Previous Work

This chapter establishes a large deviations principle (LDP) for a Markov random process in R^d with a discontinuity in the transition mechanism along a hyperplane. The transition mechanism of the process is assumed to be continuous on one closed half-space and also continuous on the complementary open half-space. The following paragraphs give an overview of the related work and identify the contribution of the present chapter. The formulation and proof of the main result are the subjects of subsequent sections.

In their paper, Dupuis and Ellis [16] established an explicit representation of the rate function in the case of a constant transition mechanism in the two half-spaces. The paper [16] proves an LDP for the process observed at a fixed point in time, though an underlying process-level LDP is implicit in the paper.

Subsequently, Blinovskii and Dobrushin [14] and Malyshev et al. [17] derived process-level LDPs for the case of a constant transition mechanism in each half-space, using different approaches. The work [17] is somewhat restrictive in that the first coordinate of the process is assumed to take values in a lattice, and when off the hyperplane, the process can step at most one unit towards the hyperplane at a time. This condition prevents jumps that strictly cross the hyperplane of discontinuity. On the other hand, the work [17] allows the process to have a different transition mechanism in each open half-space and on the hyperplane itself. The paper [14] does not rely on a lattice assumption; the jump distribution need not even be concentrated on a countable number of points, and jumps strictly crossing the boundary can occur.

The book by Shwartz and Weiss [12] establishes an LDP for a process on a half-space with a flat boundary that cannot be crossed. The transition mechanism can vary continuously on both the open half-space and on the hyperplane boundary. The method is lattice-based and also includes

the assumption of at most unit jumps towards the boundary. The model applies to processes with continuous transition mechanisms in two half-spaces separated by a hyperplane only if a symmetry condition holds. A somewhat different explicit representation for the rate function is given in [12], though as noted by Remark 4.2.1 below, it can be easily related to the expression of [16]. The paper [18] establishes a large deviations style upper bound, which is tight for the flat boundary process of [12], but which is not always tight for the case of two half-spaces separated by a boundary.

The paper by Dupuis and Ellis [15] establishes LDPs for Markov processes with transition probabilities that are continuous over facets generated by a finite number of hyperplanes. For example, two intersecting hyperplanes generate nine such facets. While in general the paper [15] does not identify the rate function explicitly, it does state an explicit integral representation for the case of a single hyperplane of discontinuity. The integrand in the representation given in [15] and [14] has the form established in the original paper [16]. The paper [15] assumes the processes are lattice-valued and satisfy a mild communication/controllability condition.

The LDP established in this chapter is based on an adaptation of the construction in [14]. Like [14], it therefore does not require lattice assumptions as posed in all previous papers with piecewise continuous transition mechanisms. The present chapter is restricted to the case of a single hyperplane of discontinuity. The method described in [15] accommodates the continuous variation of transition mechanisms throughout the proof, while it is not clear how to directly incorporate continuous variation of transition mechanisms in the approach of [14]. The tact taken in this chapter, therefore, is a two-step procedure: First, an LDP for a piecewise-constant transition mechanism is identified, and then the LDP is extended to cover a continuously varying transition mechanism within the half-spaces.

Another contribution of this chapter is to somewhat streamline the proof of [14] and to show that the method is appropriate in either continuous or discrete time.

4.2 Statement of the Main Result

Let A^o denote the hyperplane $\{x \in R^d : x(1) = 0\}$, and set $A^+ = \{x \in R^d : x(1) > 0\}$ and $A^- = \{x \in R^d : x(1) < 0\}$. Given two rate-measure fields ν^+ and ν^-, let Λ^+, Λ^-, and Λ^o be defined as follows:

$$M^{\pm}(x,\zeta) = \int_{R^d} (e^{z\zeta} - 1)\nu^{\pm}(x,dz), \qquad x,\zeta \in R^d$$

$$\Lambda^{\pm}(x,y) = \sup_{\zeta \in R^d} \{y\zeta - M^{\pm}(x,\zeta)\}, \qquad y \in R^d \tag{4.1}$$

$$\Lambda^{o}(x,y) = \inf_{\substack{0 \le \beta \le 1,\ y^+ \in \overline{A^-},\ y^- \in \overline{A^+} \\ \beta y^+ + (1-\beta) y^- = y}} \{\, \beta \Lambda^+(x,y^+) + (1-\beta)\Lambda^-(x,y^-) \,\}. \tag{4.2}$$

Consider the following conditions regarding rate-measure fields.

Condition 4.2.1 *(Boundedness) There exists a finite number m such that $\nu(x, R^d) \le m$ for all $x \in R^d$.*

Condition 4.2.2 *(Exponential Moments) For each $\zeta \in R^d$, there exists a finite number b such that $\int_{R^d}(e^{z\zeta} - 1)\nu(x,dz)/\nu(x,R^d) < b$ for all $x \in R^d$.*

Condition 4.2.3 *(Uniform Continuity) For each $x, x' \in R^d$, the measures $\nu(x)$ and $\nu(x')$ are equivalent. Furthermore, given a positive number ϵ, there exists a corresponding positive number δ such that $(1+\epsilon)^{-1} \le d\nu(x)/d\nu(x') \le (1+\epsilon)$ whenever $|x - x'| < \delta$.*

The main result of the chapter is the following theorem:

Theorem 4.2.1 *Let ν^+ and ν^- be two rate-measure fields on R^d, each of which satisfies Conditions 4.2.1-4.2.3, and $\nu^+(x, A^-) > 0$ and $\nu^-(x, A^+) > 0$ for some (equivalently all) $x \in R^d$. Let X^γ denote the Markov process generated by the pair (γ, ν), where ν is given by*

$$\nu(x) = \begin{cases} \nu^+(x) & \text{if } x \in \overline{A^+} \\ \nu^-(x) & \text{if } x \in A^-, \end{cases}$$

and let $\tilde{X}^\gamma$ denote the polygonal interpolation of X^γ. Then the sequence $(\tilde{X}^\gamma : \gamma > 0)$ satisfies the large deviations principle in $C_{[0,T]}(R^d)$ with the good rate function Γ, where for each $\phi \in C_{[0,T]}(R^d)$ and each $x_0 \in R^d$,

$$\Gamma(\phi, x_0) = \begin{cases} \int_0^T \Lambda(\phi_t, \dot{\phi}_t)dt & \text{if } \phi_0 = x_0 \text{ and } \phi \text{ is absolutely continuous} \\ +\infty & \text{otherwise,} \end{cases}$$

and Λ satisfies

$$\Lambda(\phi_t, \dot{\phi}_t) = I\{\phi_t \in A^+\}\Lambda^+(\phi_t, \dot{\phi}_t) + I\{\phi_t \in A^o\}\Lambda^o(\phi_t, \dot{\phi}_t) + I\{\phi_t \in A^-\}\Lambda^-(\phi_t, \dot{\phi}_t). \tag{4.3}$$

Remark 4.2.1 *(Alternative Representation of Λ^o) Let $n = (1, 0, \cdots, 0) \in R^d$, and define*

$$\tilde{\Lambda}^+(x,y) = \begin{cases} \Lambda^+(x,y) & \text{if } y \in \overline{A^-} \\ +\infty & \text{otherwise.} \end{cases}$$

Then

$$\Lambda^o(x,y) = \inf_{\substack{0 \le \beta \le 1,\ y^+ \in R^d,\ y^- \in R^d \\ \beta y^+ + (1-\beta) y^- = y}} \{ \beta \tilde{\Lambda}^+(x,y^+) + (1-\beta)\Lambda^-(x,y^-) \}.$$

It is easy to check that $\tilde{\Lambda}^+(x,\cdot)$ is the Legendre-Fenchel transform of $\inf_{\alpha \ge 0} M^+(x, \cdot - \alpha n)$; therefore, by [19, Theorem 16.5], $\Lambda^o(x,\cdot)$ is the Legendre-Fenchel transform of $\inf_{\alpha \ge 0} M^+(x, \cdot - \alpha n) \vee M^-(x,\cdot)$. In particular, for $y \in A^o$,

$$\Lambda^o(x,y) = \sup_{\zeta(2),\cdots,\zeta(d)} \left\{ y\zeta - \inf_{u \le v} M^-(x,(v,\zeta(2),\cdots,\zeta(d))) \vee M^+(x,(u,\zeta(2),\cdots,\zeta(d))) \right\}.$$

Note also that if $\nu^+(x, A^+) = 0$ (as in the case of a process in $\overline{A^-}$ with a flat boundary that cannot be crossed), then $\inf_{\alpha \ge 0} M^+(x, \zeta - \alpha n) = M^+(x,\zeta)$ and

$$\Lambda^o(x,y) = \sup_{\zeta \in R^d} \{ y\zeta - M^-(x,\zeta) \vee M^+(x,\zeta) \},$$

as found in [12].

The proof of Theorem 4.2.1 can be easily adapted to yield the following theorem for discrete-time Markov chains.

Theorem 4.2.2 *Let ν^+ and ν^- be two probability-measure fields on R^d, each of which satisfies Conditions 4.2.2-4.2.3, and $\nu^+(x, A^-) > 0$ and $\nu^-(x, A^+) > 0$ for some (equivalently all) $x \in R^d$. For $\gamma > 0$, let $(X_k^\gamma : k \in Z_+)$ denote the Markov chain such that given X_k^γ, the scaled increment $\gamma(X_{k+1}^\gamma - X_k^\gamma)$ has distribution $\nu(X_k^\gamma)$, where*

$$\nu(x) = \begin{cases} \nu^+(x) & \text{if } x \in \overline{A^+} \\ \nu^-(x) & \text{if } x \in A^-, \end{cases}$$

and let $\tilde{X}^\gamma$ denote the polygonal interpolation of the process $(X^\gamma_{\lfloor \gamma t \rfloor} : t \ge 0)$. Then the sequence $(\tilde{X}^\gamma : \gamma > 0)$ satisfies the large deviations principle in $C_{[0,T]}(R^d)$ with the good rate function Γ, where Λ^+, Λ^-, and Λ^o are defined by Equations (4.1)-(4.2) with $M^\pm(x,\zeta) = \log \int_{R^d} \exp(z\zeta)\nu^\pm(x,dz)$.

The proof of Theorem 4.2.1 is organized as follows. Section 4.3 contains some observations that are instrumental for the proof. Section 4.4 extends the work of Blinovskii and Dobrushin [14] to continuous-time random walks, hence proving the theorem in the special case of constant transition mechanisms in each half-space. In view of this, Sections 4.5 and 4.6 establish, respectively, the large deviations lower and upper bounds in the general case. Goodness of the rate function is shown in Section 4.7.

4.3 Preliminaries

This section contains preliminary results regarding the proof of Theorem 4.2.1. Lemma 4.3.2 establishes that the process X^γ and its polygonal interpolation $\tilde{X}^\gamma$ are close in a certain sense so that they have equivalent large deviation probabilities. The section concludes with Lemma 4.3.3 on the sensitivity of the rate function to variations of the rate measures.

Lemma 4.3.1 *Given $x \in R^d$ and $\gamma > 0$, let $\gamma\Delta_x$ have distribution $\nu(x)/\nu(x, R^d)$. Then for each $\delta > 0$, $\limsup_{\gamma\to\infty} \gamma^{-1} \log \sup_x P(|\Delta_x| \geq \delta) = -\infty$.*

Proof. If $|\Delta_x| \geq \delta$, then for some coordinate $1 \leq i \leq d$, $|\Delta_x(i)| \geq \delta/\sqrt{d}$. This, together with the union bound and Chernoff's inequality, implies that

$$\sup_x P(|\Delta_x| \geq \delta) \leq 2d \exp(-\alpha\gamma\delta/\sqrt{d}) \sup_{x, 1\leq i\leq d} E[\exp(\alpha\gamma\Delta_x(i))] \qquad \text{for each } \alpha \geq 0.$$

By Condition 4.2.2, $\sup_{x,1\leq i\leq d} E[\exp(\alpha\gamma\Delta_x(i))]$ is finite, and it does not depend on γ so that

$$\limsup_{\gamma\to\infty} \gamma^{-1} \log \sup_x P(|\Delta(x)| \geq \delta) \leq -\alpha\delta/\sqrt{d}.$$

The arbitrariness of $\alpha \geq 0$ yields the desired result. □

Lemma 4.3.2 *(Exponential Equivalence) For each $\delta > 0$,*

$$\limsup_{\gamma\to\infty} \gamma^{-1} \log \sup_x P_x(\sup_{0\leq t\leq T} |X_t^\gamma - \tilde{X}_t^\gamma| > \delta) = -\infty.$$

Proof. Let N_T^γ denote the number of jumps of X^γ in the interval $[0,T]$. Note that if $\sup_{0\le t\le T}|X_t^\gamma - -\tilde{X}_t^\gamma| > \delta$, then at least one of the first $N_T^\gamma + 1$ jumps of X^γ has size larger than δ. Therefore, for each $\gamma > 0$, $B > 0$, and $x \in R^d$,

$$\begin{aligned} P_x(\sup_{0\le t\le T}|X_t^\gamma - \tilde{X}_t^\gamma| > \delta) &\le P_x(N_T^\gamma \ge \gamma B) + P_x(\sup_{0\le t\le T}|X_t^\gamma - \tilde{X}_t^\gamma| > \delta, N_T^\gamma < \gamma B) \\ &\le P_x(N_T^\gamma \ge \gamma B) + (\gamma B + 1)\sup_{x'} P(|\Delta_{x'}| \ge \delta), \end{aligned}$$

where $\gamma\Delta_x$ has distribution $\nu(x)/\nu(x,R^d)$. By Condition 4.2.1, uniformly over all initial states, N_T^γ is stochastically dominated by a Poisson random variable with mean γmT. Therefore, given $K > 0$, B can be taken large enough so that

$$\begin{aligned} \limsup_{\gamma\to\infty}\gamma^{-1}\log\sup_x P_x(\sup_{0\le t\le T}|X_t^\gamma - \tilde{X}_t^\gamma| > \delta) &\le \left(\limsup_{\gamma\to\infty}\gamma^{-1}\log\sup_x P_x(N_T^\gamma \ge \gamma B)\right) \\ &\quad \bigvee \left(\limsup_{\gamma\to\infty}\gamma^{-1}\log\sup_x P(|\Delta_x| \ge \delta)\right) \\ &= \limsup_{\gamma\to\infty}\gamma^{-1}\log\sup_x P_x(N_T^\gamma \ge \gamma B) \qquad (4.4) \\ &\le -K, \end{aligned}$$

where (4.4) follows by Lemma 4.3.1. The arbitrariness of $K > 0$ proves the lemma. □

Given a Borel measure μ on R^d, define

$$\Lambda_\mu(y) = \sup_{\zeta\in R^d}\left\{y\zeta - \int_{R^d}(e^{z\zeta} - 1)\mu(dz)\right\}, \qquad y \in R^d.$$

Lemma 4.3.3 *If ν_0 and ν_1 are two positive, finite Borel measures on R^d such that $(1+\epsilon)^{-1} \le d\nu_0/d\nu_1 \le (1+\epsilon)$ for some $\epsilon > 0$, then for all $y \in R^d$,*

$$\Lambda_{\nu_0}(y) \ge (1+\epsilon)^{-1}\Lambda_{\nu_1}(y) - \epsilon e\nu_1(R^d).$$

Proof. Define $\chi(\epsilon) = \sup_{u\in R}\{(1+\epsilon)^2 e^u - e^{(1+\epsilon)u}\}$. Straightforward evaluation yields that $\chi(\epsilon) = \epsilon(1+\epsilon)^{(1+\epsilon)/\epsilon}$. For each $\zeta \in R^d$,

$$\int_{R^d}(e^{z\zeta} - 1)\nu_0(dz) = \int_{R^d} e^{z\zeta}\nu_0(dz) - \nu_0(R^d)$$

$$\leq (1+\epsilon)\int_{R^d} e^{z\zeta}\nu_1(dz) - (1+\epsilon)^{-1}\nu_1(R^d) \tag{4.5}$$

$$\leq (1+\epsilon)^{-1}\int_{R^d} e^{z\zeta(1+\epsilon)}\nu_1(dz) + (\chi(\epsilon)-1)(1+\epsilon)^{-1}\nu_1(R^d) \tag{4.6}$$

$$\leq (1+\epsilon)^{-1}\int_{R^d}(e^{z\zeta(1+\epsilon)}-1)\nu_1(dz) + \epsilon e\nu_1(R^d), \tag{4.7}$$

where inequality (4.5) is a consequence of the hypothesis, (4.6) is implied by the definition of $\chi(\epsilon)$, and (4.7) follows by the fact that $\chi(\epsilon)/(1+\epsilon) \leq \epsilon e$. This in turn implies that for any $y \in R^d$,

$$\begin{aligned}
\Lambda_{\nu_0}(y) &= \sup_{\zeta\in R^d}\left\{y\zeta - \int_{R^d}(e^{z\zeta}-1)\nu_0(dz)\right\} \\
&\geq \sup_{\zeta\in R^d}\left\{y\zeta - (1+\epsilon)^{-1}\int_{R^d}(e^{z\zeta(1+\epsilon)}-1)\nu_1(dz)\right\} - \epsilon e\nu_1(R^d) \\
&= (1+\epsilon)^{-1}\sup_{\zeta\in R^d}\left\{y\zeta(1+\epsilon) - \int_{R^d}(e^{z\zeta(1+\epsilon)}-1)\nu_1(dz)\right\} - \epsilon e\nu_1(R^d) \\
&= (1+\epsilon)^{-1}\Lambda_{\nu_1}(y) - \epsilon e\nu_1(R^d).
\end{aligned}$$

This proves the lemma. □

4.4 The Piecewise Homogeneous Case

This section establishes Theorem 4.2.1 for the case in which the two rate-measure fields are constant. The result, stated as Lemma 4.4.1 below, can be proved by adapting the proof of the analogous result for discrete-time piecewise-homogeneous random walks, as presented in [14]. Here, we outline the sufficient modifications of that proof, while at the same time pointing out how the argument in [14] can be somewhat streamlined.

Lemma 4.4.1 *If ν^+, ν^- and $\tilde{X}^\gamma$ satisfy the conditions of Theorem 4.2.1 with $\nu^+(x) \equiv \nu_o^+$ and $\nu^-(x) \equiv \nu_o^-$ for two fixed measures ν_o^+ and ν_o^-, then the sequence $(\tilde{X}^\gamma : \gamma > 0)$ satisfies the large deviations principle with the good rate function Γ.*

Let $(s_t^\pm : t \geq 0)$ denote a compound Poisson process with rate measure $\nu_o^\pm$ so that the probability distribution $P^\pm$ of the random variable $s_1^\pm$ is a compound Poisson probability distribution with a log moment generating function $G_P^\pm$ given by

$$G_P^\pm(\zeta) = \log\int_{R^d} e^{z\zeta}P^\pm(dz) = \int_{R^d}(e^{z\zeta}-1)\nu_o^\pm(dz) = M^\pm(\zeta).$$

(The first arguments of $M^{\pm}$ and $\Lambda^{\pm}$ are suppressed in this section since the rate-measure fields are constant.) Thus, $\Lambda^{\pm}$ is the Legendre-Fenchel transform (denoted $H_P^{\pm}$ in [14]) of $G_P^{\pm}$. The expression $\Gamma(\phi, x_0)$ of Theorem 4.2.1 is thus identical to the rate function $N(\phi)$ defined in [14]; hence we simply refer to [14] for the properties of $\Lambda^{\pm}$, Λ^o, and Γ. In particular, it is shown in [14] that Γ is a good rate function.

A representation of $\tilde{X}^\gamma$: The key to the proof in [14] is to combine two independent homogeneous random walks to produce a single, piecewise-homogeneous random walk. The continuous time process X^γ can similarly be constructed by combining the processes s^+ and s^-, as shown below. The random variables $s_t^{\pm}/t$ obey the Cramer Theorem as $t \to +\infty$, with the rate function $\Lambda^{\pm}$. Furthermore, the exponential tightness property and local large deviations properties hold exactly as for the discrete time case stated in Proposition 5.3 of [14].

We next define an "unscaled" process X so that the process X^γ has the same distribution as the process $((X_{\gamma t})/\gamma : t \geq 0)$. The process X is conveniently defined via a jump representation, using the following jump representations of $s^{\pm}$. Let $(J^{\pm}(k) : k \geq 1)$ be independent, identically distributed random variables with the probability distribution $\nu_o^{\pm}(\cdot)/\nu_o^{\pm}(R^d)$. Let $(U^{\pm}(k) : k \geq 1)$ be independent, exponentially distributed random variables with parameter $\nu_o^{\pm}(R^d)$. Also, for convenience, set $J^{\pm}(0) = U^{\pm}(0) = 0$. Then $s^{\pm}$ can be represented as

$$s_t^{\pm} = J^{\pm}(0) + \cdots + J^{\pm}(k) \quad \text{if} \quad U^{\pm}(0) + \cdots + U^{\pm}(k) \leq t < U^{\pm}(0) + \cdots + U^{\pm}(k+1).$$

Of course it is assumed that s^+ is independent of s^-. Given an initial state X_0, let X denote the Markov process for which the corresponding variables $(U(k) : k \geq 0)$ and $(J(k) : k \geq 0)$ are defined recursively as follows: $U(0) = 0, n^{\pm}(0) = 0, J(0) \equiv X_0$, and

$$\text{if } X_0 + J(1) + \cdots + J(k) \in \overline{A^+}, \text{ then } \begin{cases} n^+(k+1) = n^+(k) + 1 & U(k+1) = U^+(n^+(k+1)) \\ n^-(k+1) = n^-(k) & J(k+1) = J^+(n^+(k+1)), \end{cases}$$

$$\text{else if } X_0 + J(1) + \cdots + J(k) \in A^-, \text{ then } \begin{cases} n^+(k+1) = n^+(k) & U(k+1) = U^-(n^-(k+1)) \\ n^-(k+1) = n^-(k) + 1 & J(k+1) = J^-(n^-(k+1)). \end{cases}$$

Then X can be represented as

$$X_t = X_0 + J(1) + \cdots + J(k) \quad \text{if} \quad U(0) \cdots + U(k) \leq t < U(0) + \cdots + U(k+1).$$

Note that the process $((X_{\gamma t})/\gamma : t \geq 0)$ can be identified with the process X^γ as desired. Define $(\Theta_t : t \geq 0)$ as follows. If $U(0) + \cdots + U(k) \leq t < U(0) + \cdots + U(k+1)$, then

$$\Theta_t = \begin{cases} 1 & \text{if } X_0 + J(1) + \cdots + J(k) \in \overline{A^+} \\ 0 & \text{else.} \end{cases}$$

Intuitively, X evolves according to s^+ on the intervals in which $\Theta_t = 1$. In particular, let $\tau(t) = \int_0^t \Theta_s ds$. Then for $t \geq 0$,

$$X_t = X_0 + s^+_{\tau(t)} + s^-_{t-\tau(t)}. \tag{4.8}$$

Identify $\tilde{X}^\gamma$ as the polygonal interpolation of the scaled process $((X_{\gamma t})/\gamma : t \geq 0)$. Let S^+ and S^- denote the polygonal interpolations of $((s^+_{\gamma t})/\gamma : t \geq 0)$ and $((s^-_{\gamma t})/\gamma : t \geq 0)$, respectively. (For brevity we do not explicitly indicate the dependence of $S^\pm$ on γ.) It is useful to note that the relation (4.8) carries over to the scaled processes:

$$\tilde{X}^\gamma_t = y_0 + S^+_{\tau(t)} + S^-_{t-\tau(t)}, \tag{4.9}$$

where $y_0 = X_0/\gamma$, for all $t \geq 0$.

Given $\eta > 0$ and $T > 0$, define the events

$$K^\pm(\eta, T, \gamma) = \{|S^\pm_t - (s^\pm_{\gamma t})/\gamma| \leq \eta, 0 \leq t \leq T\}$$

and set $K(\eta, T, \gamma) = K^+(\eta, T, \gamma) \cap K^-(\eta, T, \gamma)$. Lemma 4.3.2 implies that the set $K(\eta, T, \gamma)^c$ is negligible for the purposes of proving large deviations principles, in the sense that

$$\lim_{\gamma \to \infty} \gamma^{-1} \log P[K(\eta, T, \gamma)^c] = -\infty.$$

Note that on the event $K(\eta, T, \gamma)$, $|\tilde{X}^\gamma_t - (X_{\gamma t})/\gamma| \leq \eta$ for $0 \leq t \leq T$.

Due to the analytic considerations in [14], the proof of the large deviations principle in continuous time can be reduced to proving upper and lower large deviations bounds for the events of the form $\mathcal{E}(\sigma, \delta, T) = \{|\tilde{X}^\gamma_t - \sigma_t| \leq \delta, 0 \leq t \leq T\}$, where $T > 0$ and $\delta > 0$ can be taken arbitrarily small. Here $\sigma_t = x_0 + tv$, where $v \in R^d$ and, with $x_1 = \sigma_T$, either $x_0, x_1 \in \overline{A^+}$ or $x_0, x_1 \in \overline{A^-}$. The key to proving these bounds is to bound the event $\mathcal{E}(\sigma, \delta, T)$ from inside and outside by simple events involving the process (S^+, S^-) and to appeal to the large deviations principle for (S^+, S^-).

This is essentially the same idea as in [14], translated for continuous time. Our proof is simplified somewhat in that (a) in the case of the upper bound, our proof makes better use of the large deviations principle for (S^+, S^-), which is common to both discrete and continuous time, and (b) we exploit the representation (4.9). These simplifications make the translation between discrete and continuous time more transparent.

Lower bound: The three lemmas that follow identify events involving (S^+, S^-) that are subsets of the event $\mathcal{E}(\sigma, \delta, T)$ whenever $x_0, x_1 \in \overline{A^+}$. The case $x_0, x_1 \in \overline{A^-}$ can be handled similarly. The large deviations lower bound for the process (S^+, S^-) can then be readily used to provide the required lower bound for $P[\mathcal{E}(\sigma, \delta, T)]$.

Lemma 4.4.2 *If $x_0, x_1 \in A^+$, then for δ small enough,*

$$\mathcal{E}(\sigma, \delta, T) \supset \{|y_0 - x_0| \leq \delta/2\} \cap \{|S_t^+ - tv| \leq \delta/2, 0 \leq t \leq T\}. \tag{4.10}$$

Proof. It is enough to note that for δ small enough, $\tilde{X}_t^\gamma = y_0 + S_t^+$ for $0 \leq t \leq T$ if the event on the right side of (4.10) is true. □

Corollary 3.2 of [14] states that there is a vector $b^- \in A^+$ such that $\Lambda^-(b^-) < +\infty$.

Lemma 4.4.3 *If $(x_0 \in A^+, x_1 \in A^o)$ or if $(x_0 \in A^o, x_1 \in A^+)$, then there exist $\eta = \eta(\delta) \to 0$ and $\kappa = \kappa(\delta) \to 0$ as $\delta \to 0$ so that*

$$\begin{aligned} \mathcal{E}(\sigma, \delta, T) \supset\ & \{|S_t^+ - tv| \leq \eta, 0 \leq t \leq T\} \cap \{|S_t^- - tb^-| \leq \eta, 0 \leq t \leq \kappa\} \\ & \cap \{|y_0 - x_0| \leq \eta\} \cap K(\eta, T, \gamma). \end{aligned} \tag{4.11}$$

Proof. Assume that $(x_0 \in A^o, x_1 \in A^+)$. Take $\kappa = 5\eta/b^-(1)$, where $\eta = \eta(\delta)$ is yet to be specified, and suppose that the event on the right side of (4.11) is true. We first prove the following claim: $T - \tau(T) < \kappa$. If this claim is false, let u denote the minimum positive value such that $u - \tau(u) = \kappa$. Then $\kappa \leq u \leq T$ and

$$\begin{aligned} \tilde{X}_u^\gamma(1) &= y_0(1) + S_{\tau(u)}^+(1) + S_{u-\tau(u)}^-(1) \\ &\geq -\eta + (u - \kappa)v(1) - \eta + \kappa b^-(1) - \eta \\ &\geq \kappa b^-(1) - 3\eta = 2\eta. \end{aligned}$$

On the event $K(\eta, T, \gamma)$, $\Theta_t = 1$ whenever $\tilde{X}_t^\gamma(1) \geq \eta$ so that u cannot be a point of increase of $t - \tau(t)$. The claim is thus true by proof by contradiction.

Thus, for $0 \leq t \leq T$,

$$\begin{aligned} |\tilde{X}_t^\gamma - \sigma_t| &\leq |\tilde{X}_t^\gamma - (x_0 + S_t^+)| + |x_0 + S_t^+ - \sigma_t| \\ &\leq |x_0 - y_0| + \left(\sup_{t-\kappa \leq r \leq t} |S_r^+ - S_t^+| \right) + \left(\sup_{0 \leq r \leq \kappa} |S_r^-| \right) + |tv - S_t^+| \\ &\leq \eta + (|v|\kappa + 2\eta) + (|b^-|\kappa + \eta) + \eta = C\eta, \end{aligned}$$

where the constant C depends only on v and b^-. Taking $\eta = \delta/C$, the event $\mathcal{E}(\sigma, \delta, t)$ is true, and the lemma is proved in the case $(x_0 \in A^+, x_1 \in A^o)$. The proof in the case $(x_0 \in A^o, x_1 \in A^+)$ is similar and is omitted. □

Lemma 4.4.4 *If $x_0, x_1 \in A^o$, and if $0 < \beta < 1$, $v^+ \in A^-$, and $v^- \in A^+$ are such that $v = \beta v^+ + (1-\beta)v^-$, then there exist $\eta = \eta(\delta) \to 0$ and $\kappa = \kappa(\delta) \to 0$ as $\delta \to 0$ so that*

$$\begin{aligned} \mathcal{E}(\sigma, \delta, T) \supset \{|S_t^+ - tv^+| \leq \eta, 0 \leq t \leq \beta T + \kappa\} \cap \{|S_t^- - tv^-| \leq \eta, 0 \leq t \leq (1-\beta)T + \kappa\} \\ \cap \{|y_0 - x_0| \leq \eta\} \cap K(\eta, T, \gamma). \end{aligned} \tag{4.12}$$

Proof. Take $\kappa = 5\eta/(v^-(1) - v^+(1))$, where $\eta = \eta(\delta)$ is yet to be specified, and suppose that the event on the right side of (4.12) is true. We first prove the following claim: $|\tau(t) - \beta t| < \kappa$, or equivalently, $|t - \tau(t) - (1-\beta)t| < \kappa$, for $0 \leq t \leq T$. If this claim is false, let u denote the minimum value of t such that the inequalities are violated. Then either $\tau(u) = \beta u + \kappa$ or $u - \tau(u) = (1-\beta)u + \kappa$. By symmetry we assume without loss of generality that $u - \tau(u) = (1-\beta)u + \kappa$, and hence also $\tau(u) = \beta u - \kappa$. Thus,

$$\begin{aligned} \tilde{X}_u^\gamma(1) &= y_0(1) + S_{\tau(u)}^+(1) + S_{u-\tau(u)}^-(1) \\ &\geq -\eta + (\beta u - \kappa)v^+(1) - \eta + ((1-\beta)u + \kappa)v^-(1) - \eta \\ &= \kappa(v^-(1) - v^+(1)) - 3\eta = 2\eta. \end{aligned}$$

On the event $K(\eta, t, \gamma)$, $\Theta_t = 1$ whenever $\tilde{X}_t^\gamma(1) \geq \eta$ so that u cannot be a point of increase of $t - \tau(t)$. The claim is thus true by proof by contradiction.

Thus, for $0 \le t \le T$,

$$\begin{aligned}
|\tilde{X}_t^\gamma - \sigma_t| &\le |\tilde{X}_t^\gamma - (x_0 + S^+_{\beta t} + S^-_{(1-\beta)t})| + |S^+_{\beta t} - \beta t v^+| + |S^-_{(1-\beta)t} - (1-\beta)tv^-| \\
&\le |y_0 - x_0| + \left(\sup_{|r-\beta t| \le \kappa} |S^+_r - S^+_{\beta t}| \right) + \left(\sup_{|r-(1-\beta)t| \le \kappa} |S^-_r - S^-_{(1-\beta)t}| \right) + \eta + \eta \\
&\le \eta + (2|v^+|\kappa + 2\eta) + (2|v^-|\kappa + 2\eta) + 2\eta = C\eta,
\end{aligned}$$

where the constant C depends only on v^+ and v^-. Taking $\eta = \delta/C$, the event $\mathcal{E}(\sigma, \delta, t)$ is true, and the lemma is proved. □

Proposition 3.4 of [14] shows that the conditions $y^\pm \in \overline{A^\pm}$ and $0 \le \beta \le 1$ in (4.2) can be replaced by the conditions $y^\pm \in A^\pm$ and $0 < \beta < 1$ without changing the value of Λ^o. Thus, Lemma 4.4.4, with its condition that $v^\pm \in A^\pm$ (rather than $v^\pm \in \overline{A^\pm}$) and $0 < \beta < 1$, suffices for the derivation of the required lower large deviations bound for $\mathcal{E}(\sigma, \delta, T)$.

Upper bound: The two lemmas that follow identify events involving (S^+, S^-) that contain the event $\mathcal{E}(\sigma, \delta, T)$ whenever $x_0, x_1 \in \overline{A^+}$. The case $x_0, x_1 \in \overline{A^-}$ can be handled similarly. The large deviations upper bound for the process (S^+, S^-) can then be readily used to provide the required upper bound for $P[\mathcal{E}(\sigma, \delta, T)]$. The first lemma is easily verified and is stated without proof.

Lemma 4.4.5 *If $x_0, x_1 \in \overline{A^+}$, and $\{x_0, x_1\} \not\subset A^o$, then there exist $\eta = \eta(\delta) \to 0$ and $\kappa = \kappa(\delta) \to 0$ as $\delta \to 0$ so that*

$$\mathcal{E}(\sigma, \delta, T) \subset \{|S^+_{T-\kappa} - (T-\kappa)v| \le \eta\} \cup K(\eta, T, \gamma)^c.$$

Lemma 4.4.6 *If $x_0, x_1 \in A^o$, then for δ, κ, $\epsilon > 0$, $\mathcal{E}(\sigma, \delta, T) \subset \{(S^+, S^-) \in F^{2\delta}\}$, where $F^{2\delta}$ and F are the subsets of $C_{[0,T]}(R^d \times R^d)$, defined as follows:*

$$F^{2\delta} = \{(\tilde{\phi}^+, \tilde{\phi}^-) : \sup_{0 \le t \le T} |\tilde{\phi}^+_t - \phi^+_t| \le 2\delta \text{ and } \sup_{0 \le t \le T} |\tilde{\phi}^-_t - \phi^-_t| \le 2\delta \text{ for some } (\phi^+, \phi^-) \in F\},$$

and F denotes the closed set $F_1 \cup F_2 \cup F_3 \cup F_4$, where

$$\begin{aligned}
F_1 &= \{(\phi^+, \phi^-) : \sup_{0 \le t \le \kappa} (|\phi^+_t| + |\phi^-_t|) \ge \epsilon\} \\
F_2 &= \{(\phi^+, \phi^-) : \exists \tau \in [\kappa, T-\kappa] : \phi^+_\tau \in \overline{A^-}, \phi^-_{T-\tau} \in \overline{A^+}, \phi^+_\tau + \phi^-_{T-\tau} = vT\} \\
F_3 &= \{(\phi^+, \phi^-) : |\phi^+_{T-\kappa} - (T-\kappa)v| \le \epsilon + |v|\kappa\}
\end{aligned}$$

$$F_4 \;=\; \{(\phi^+,\phi^-) : |\phi^-_{T-\kappa} - (T-\kappa)v| \le \epsilon + |v|\kappa\}.$$

Proof. Suppose the event $\mathcal{E}(\sigma,\delta,T)$ is true. Since

$$\mathcal{E}(\sigma,\delta,T) = \{|y_0 + S^+_{\tau(t)} + S^-_{t-\tau(t)} - \sigma_t| \le \delta, 0 \le t \le T\},$$

it follows (take $t=0$) that $|y_0 - x_0| \le \delta$ so that

$$|S^+_{\tau(t)} + S^-_{t-\tau(t)} - vt| \le 2\delta,\ 0 \le t \le T. \tag{4.13}$$

To complete the proof of the lemma we consider three cases.

Case 1: Suppose $\kappa \le \tau(T) \le T-\kappa$. Let $\Delta = S^+_{\tau(T)} + S^-_{T-\tau(T)} - vT$, and note that $|\Delta| \le 2\delta$. Note by the construction of the process Y, if $\Theta_T = 0$, then $S^+_{\tau(T)} \in \overline{A^-}$, whereas if $\Theta_T = 1$, then $S^-_{T-\tau(T)} \in \overline{A^+}$. Define (ϕ^+,ϕ^-) by setting

$$(\phi^+_t,\phi^-_t) = \begin{cases} (S^+_t, S^-_t - \Delta(\frac{t}{T-\tau(T)} \wedge 1)) & \text{if } \Theta_T = 0 \\ (S^+_t - \Delta(\frac{t}{\tau(T)} \wedge 1), S^-_t) & \text{if } \Theta_T = 1 \end{cases}$$

for $0 \le t \le T$. Then $(\phi^+,\phi^-) \in F$ and $\sup_{0\le t\le T} |S^\pm_t - \phi^\pm_t| \le 2\delta$ so that $(S^+,S^-) \in F^{2\delta}$.

Case 2: Suppose $\tau(T) > T-\kappa$. Let $t_0 = \min\{t \ge 0 : \tau(t) = T-\kappa\}$, and let $t_1 = t_0 - (T-\kappa)$. Then $T-\kappa \le t_0 \le T$ and $0 \le t_1 \le \kappa$. Also, $\tau(t_0) = T-\kappa$ and $t_0 - \tau(t_0) = t_1$, so by (4.13),

$$|S^+_{T-\kappa} + S^-_{t_1} - (t_1 + T - \kappa)v| \le 2\delta. \tag{4.14}$$

We assume in addition that

$$\sup_{0\le t\le\kappa} |S^-_t| \le \epsilon, \tag{4.15}$$

for otherwise $(S^+,S^-) \in F \subset F^{2\delta}$. Combining (4.14), (4.15) and the fact $0 \le t_1 \le \kappa$ yields that

$$|S^+_{T-\kappa} - (T-\kappa)v| \le 2\delta + \epsilon + |v|\kappa.$$

Therefore, $(S^+,S^-) \in F^{2\delta}$.

Case 3: Suppose $\tau(T) \le \kappa$. This case is the same as case 2 with the roles of S^+ and S^- reversed. Lemma 4.4.6 is thus proved. □

Lemma 4.4.5 immediately implies that if $x_0, x_1 \in \overline{A^+}$, and $\{x_0, x_1\} \not\subset A^o$, then

$$\lim_{\delta \to 0} \limsup_{\gamma \to \infty} \gamma^{-1} \log P[\mathcal{E}(\sigma, \delta, T)] \leq -T\Lambda^+(v).$$

Similarly, Lemma 4.4.6 yields the appropriate large deviations upper bound if $x_0, x_1 \in A^o$:

Lemma 4.4.7 *If $x_0, x_1 \in A^o$, then*

$$\lim_{\delta \to 0} \limsup_{\gamma \to \infty} \gamma^{-1} \log P[\mathcal{E}(\sigma, \delta, T)] \leq -T\Lambda^o(v).$$

Proof. Let $\Gamma^{\pm}$ denote the rate function Γ with $\Lambda \equiv \Lambda^{\pm}$. The process (S^+, S^-) satisfies a large deviations principle with the good rate function $\Gamma^+ + \Gamma^-$, so by Lemma 4.4.6 for each $\kappa, \epsilon > 0$

$$\begin{aligned} \lim_{\delta \to 0} \limsup_{\gamma \to \infty} \gamma^{-1} \log P[\mathcal{E}(\sigma, \delta, T)] &\leq -\lim_{\delta \to 0} \inf_{(\phi^+, \phi^-) \in F^{2\delta}} \{\Gamma^+(\phi^+, 0) + \Gamma^-(\phi^-, 0)\} \\ &= -\inf_{(\phi^+, \phi^-) \in F} \{\Gamma^+(\phi^+, 0) + \Gamma^-(\phi^-, 0)\}. \end{aligned}$$

To complete the proof of the lemma, it suffices to show that for each $\rho > 0$ there exist $\kappa, \epsilon > 0$ such that for each $j \in \{1, 2, 3, 4\}$

$$\inf_{(\phi^+, \phi^-) \in F_j} \{\Gamma^+(\phi^+, 0) + \Gamma^-(\phi^-, 0)\} \geq T\Lambda^o(v) - \rho. \tag{4.16}$$

Note that inequality (4.16) holds for $j = 2$ for all $\rho, \epsilon, \kappa > 0$. Choose $L > T\Lambda^o(v)$. The fact that $\Lambda^{\pm}(y)/|y| \to \infty$ as $|y| \to \infty$ implies the existence of $\kappa(\epsilon) \to 0$ as such that for each ϵ,

$$\inf_{(\phi^+, \phi^-) \in F_1} \{\Gamma^+(\phi^+, 0) + \Gamma^-(\phi^-, 0)\} > L,$$

so that (4.16) holds for $j = 1$. Since

$$\inf_{(\phi^+, \phi^-) \in F_3} \{\Gamma^+(\phi^+, 0) + \Gamma^-(\phi^-, 0)\} = \inf_{|y^+ - v| \leq \epsilon + |v|\kappa(\epsilon)} (T - \kappa(\epsilon))\Lambda^+(y^+) \to T\Lambda^+(v) \text{ as } \epsilon \to 0$$

and $\Lambda^+(v) \geq \Lambda^o(v)$, inequality (4.16) holds for $j = 3$, and similarly for $j = 4$, for sufficiently small ϵ. □

4.5 The Lower Bound

This section establishes the large deviations lower bound for Theorem 4.2.1 roughly as follows. Given $\phi \in C_{[0,T]}(R^d)$, the process X^γ is approximated by a "patchwork" Markov process with a time-varying transition mechanism that for each t is constant on each half-space. The time variation is determined by ϕ and a partition of $[0,T]$. Lemma 4.4.1 is used to prove a local lower bound inequality for the patchwork process. Then by comparing the quantities on each side of this inequality to the corresponding quantities for X^γ, a local lower bound is obtained for X^γ. Following standard techniques, this local lower bound is shown to imply the lower bound for Theorem 4.2.1.

For $T > 0$, a *partition* of the interval $[0,T]$ is a finite sequence $\theta = (\theta_0, \cdots, \theta_{J(\theta)})$ such that $0 = \theta_0 < \theta_1 < \cdots < \theta_{J(\theta)} = T$. Given $\phi \in C_{[0,T]}(R^d)$ and a partition θ of $[0,T]$, let $X^{\gamma,\phi,\theta}$ denote a Markov process with a time-varying transition mechanism: For each $i \in \{0, \cdots, J(\theta)-1\}$, $X^{\gamma,\phi,\theta}$ in the time interval $(\theta_i, \theta_{i+1}]$ is generated by the pair (γ, ν_i), where for each $x \in R^d$ the rate measure $\nu_i(x)$ satisfies

$$\nu_i(x) = \begin{cases} \nu^+(\phi_{\theta_i}) & \text{if } x \in \overline{A^+} \\ \nu^-(\phi_{\theta_i}) & \text{if } x \in A^-. \end{cases}$$

Also let $\Lambda^{\phi_{\theta_i}}$ denote the function Λ defined by Equation (4.3) when $\nu^+(x) \equiv \nu^+(\phi_{\theta_i})$ and $\nu^-(x) \equiv \nu^-(\phi_{\theta_i})$.

Lemma 4.5.1 *(Intermediate Lower Bound) For each $T > 0$, partition $\theta = (\theta_0, \cdots, \theta_{J(\theta)})$ of $[0,T]$, and absolutely continuous $\phi \in C_{[0,T]}(R^d)$,*

$$\lim_{\delta \searrow 0} \lim_{\rho \searrow 0} \liminf_{\gamma \to \infty} \gamma^{-1} \log \inf_{|x-\phi_0|<\rho} P_x(\sup_{0 \le t \le T} |X_t^{\gamma,\phi,\theta} - \phi_t| < \delta) \ge -\sum_{i=0}^{J(\theta)-1} \int_{\theta_i}^{\theta_{i+1}} \Lambda^{\phi_{\theta_i}}(\phi_t, \dot{\phi}_t) dt.$$

Proof. We prove the lemma by induction on $J(\theta)$. Lemma 4.4.1, along with Remark 3.2.1 and Lemma 4.3.2, implies that the statement of the lemma holds whenever $J(\theta) = 1$. As the induction hypothesis, let $k \ge 1$ and suppose that the lemma holds for any $T > 0$ and partition θ of $[0,T]$ such that $J(\theta) = k$. Then $\forall \epsilon > 0$ $\exists \delta_k(\epsilon) > 0$ such that $\forall \delta \in (0, \delta_k(\epsilon))$ $\exists \rho_k(\delta, \epsilon)$ such that $\forall \rho \in (0, \rho_k(\delta, \epsilon))$ $\exists \gamma_k(\rho, \delta, \epsilon)$ such that for $\gamma > \gamma_k(\rho, \delta, \epsilon)$,

$$\gamma^{-1} \log \inf_{|x-\phi_0|<\rho} P_x(\sup_{0 \le t \le \theta_k} |X_t^{\gamma,\phi,\theta} - \phi_t| < \delta) \ge -\sum_{i=0}^{k-1} \int_{\theta_i}^{\theta_{i+1}} \Lambda^{\phi_{\theta_i}}(\phi_t, \dot{\phi}_t) dt - \epsilon. \tag{4.17}$$

By the time-homogeneous Markov property of the pair $(X^{\gamma,\phi,\theta},\phi)$, $\forall\epsilon>0$ $\exists\delta(\epsilon)>0$ such that $\forall\delta\in(0,\delta(\epsilon))$ $\exists\rho(\delta,\epsilon)$ such that $\forall\rho\in(0,\rho(\delta,\epsilon))$ $\exists\gamma(\rho,\delta,\epsilon)$ such that for $\gamma>\gamma(\rho,\delta,\epsilon)$,

$$\gamma^{-1}\log\inf_{|y-\phi_{\theta_k}|<\rho}P_x(\sup_{\theta_k\le t\le\theta_{k+1}}|X_t^{\gamma,\phi,\theta}-\phi_t|<\delta\mid X_{\theta_k}^{\gamma,\phi,\theta}=y)\ge-\int_{\theta_k}^{\theta_{k+1}}\Lambda^{\phi_{\theta_i}}(\phi_t,\dot\phi_t)dt-\epsilon. \quad (4.18)$$

To show that the claim holds for $J(\theta)=k+1$, fix $\epsilon>0$. For all $\delta\in(0,\delta(\epsilon/2))$, $\alpha=(\delta\wedge\delta_k(\epsilon/2)\wedge\rho(\delta,\epsilon/2))/2$, $\rho\in(0,\rho_k(\alpha,\epsilon/2))$, and $\gamma>\gamma(\alpha,\delta,\epsilon/2)\vee\gamma_k(\rho,\alpha,\epsilon/2)$,

$$\begin{aligned}\gamma^{-1}\log\inf_{|x-\phi_0|<\rho}P_x(\sup_{0\le t\le\theta_{k+1}}|X_t^{\gamma,\phi,\theta}-\phi_t|<\delta) &\ge \gamma^{-1}\log\inf_{|x-\phi_0|<\rho}P_x(\sup_{0\le t\le\theta_k}|X_t^{\gamma,\phi,\theta}-\phi_t|<\alpha)\\ &\quad+\gamma^{-1}\log\inf_{|y-\phi_{\theta_k}|<\alpha}P_x(\sup_{\theta_k\le t\le\theta_{k+1}}|X_t^{\gamma,\phi,\theta}-\phi_t|<\delta\mid X_{\theta_k}^{\gamma,\phi,\theta}=y)\\ &\ge -\sum_{i=0}^{k}\int_{\theta_i}^{\theta_{i+1}}\Lambda^{\phi_{\theta_i}}(\phi_t,\dot\phi_t)dt-\epsilon,\end{aligned}$$

where the first step follows by the Markov property of $X^{\gamma,\phi,\theta}$ and the fact that $\alpha\le\delta$, and the second step follows by the statements (4.17) and (4.18) together with the choice of α. This completes the induction step and establishes the lemma. □

Given $T>0$, a partition θ of the interval $[0,T]$, $\phi\in C_{[0,T]}(R^d)$, and $x\in R^d$, let $P_x^{\gamma,T}$ and $P_x^{\gamma,\phi,\theta,T}$ denote, respectively, the probability distributions of $(X_t^\gamma:0\le t\le T)$ and $(X_t^{\gamma,\phi,\theta}:0\le t\le T)$, with $X_0^\gamma\equiv X_0^{\gamma,\phi,\theta}\equiv x$. Note that both measures are concentrated on the space of piecewise constant functions that take values in R^d, equal to x at time 0, are right continuous, and have a finite number of jumps in $[0,T]$. There is a version D of the Radon-Nikodym derivative $dP_x^{\gamma,T}/dP_x^{\gamma,\phi,\theta,T}$, which satisfies for any such function ω

$$D(\omega)=\exp\left(-\gamma\int_0^T(\nu(\omega_t,R^d)-\nu(\omega_{\ell(t)},R^d))dt\right)\prod_{k=1}^{N_T(\omega)}\frac{d\nu(\omega_{\tau_k^-})}{d\nu(\omega_{\ell(\tau_k)})}(\Delta\omega_k), \quad (4.19)$$

where $N_T(\omega)$ denotes the number of jumps of ω in $[0,T]$, τ_k and $\Delta\omega_k$ denote, respectively, the time and size of the k^{th} jump of ω, and $\ell(t)=\max\{\theta_i:\theta_i<t\}$.

Lemma 4.5.2 *(Local Lower Bound) For each $\phi\in C_{[0,T]}(R^d)$ and $x_0\in R^d$,*

$$\lim_{\delta\searrow0}\lim_{\rho\searrow0}\liminf_{\gamma\to\infty}\gamma^{-1}\log\inf_{|x-x_0|<\rho}P_x(\sup_{0\le t\le T}|X_t^\gamma-\phi_t|<\delta)\ge-\Gamma(\phi,x_0).$$

Proof. We may take $\Gamma(\phi, x_0) < \infty$ so that ϕ is absolutely continuous and $\phi_0 = x_0$. Fix $\epsilon > 0$. Since both ν^+ and ν^- satisfy Condition 4.2.3, there exists a $\delta > 0$ such that

$$(1+\epsilon)^{-1} \leq \frac{d\nu^+(x)}{d\nu^+(x')} \leq (1+\epsilon) \quad \text{and} \quad (1+\epsilon)^{-1} \leq \frac{d\nu^-(x)}{d\nu^-(x')} \leq (1+\epsilon)$$

whenever $|x - x'| < 2\delta$. Appeal to the uniform continuity of ϕ on $[0,T]$ to choose a partition $\theta = (\theta_0, \cdots, \theta_{J(\theta)})$ of $[0,T]$ such that $\sup_{\theta_i \leq t \leq \theta_{i+1}} |\phi_t - \phi_{\theta_i}| < \delta$ for each $i \in \{0, \cdots, J(\theta)-1\}$. Then Lemma 4.3.3 applied to the definition of $\Lambda^{\phi_{\theta_i}}$ implies that

$$\begin{aligned} \sum_{i=0}^{J(\theta)-1} \int_{\theta_i}^{\theta_{i+1}} \Lambda^{\phi_{\theta_i}}(\phi_t, \dot{\phi}_t) dt &\leq \sum_{i=0}^{J(\theta)-1} \int_{\theta_i}^{\theta_{i+1}} \Big((1+\epsilon)\Lambda^{\phi_t}(\phi_t, \dot{\phi}_t) + (1+\epsilon)\epsilon em \Big) dt \\ &= (1+\epsilon)\Gamma(\phi, x_0) + (1+\epsilon)\epsilon emT. \end{aligned} \tag{4.20}$$

Let $N_T^{\gamma,\phi,\theta}$ and N_T^{γ} denote, respectively, the number of jumps of $X^{\gamma,\phi,\theta}$ and X^{γ} in the interval $[0,T]$. Appeal to Condition 4.2.1 to choose a B large enough so that

$$\limsup_{\gamma \to \infty} \gamma^{-1} \log P_{x_0}(N_T^{\gamma,\phi,\theta} \geq \gamma B) \leq -\Gamma(\phi, x_0).$$

The choice of θ and Equation (4.19) imply that for each $\gamma > 0$ and $x \in R^d$,

$$\begin{aligned} P_x(\sup_{0 \leq t \leq T} |X_t^{\gamma,\phi,\theta} - \phi_t| < \delta) &\leq P_x(\sup_{0 \leq t \leq T} |X_t^{\gamma,\phi,\theta} - \phi_t| < \delta, N_T^{\gamma,\phi,\theta} < \gamma B) + P_x(N_T^{\gamma,\phi,\theta} \geq \gamma B) \\ &\leq e^{\gamma\epsilon(mT+B)} P_x(\sup_{0 \leq t \leq T} |X_t^{\gamma} - \phi_t| < \delta, N_T^{\gamma} < \gamma B) + P_x(N_T^{\gamma,\phi,\theta} \geq \gamma B) \\ &\leq e^{\gamma\epsilon(mT+B)} P_x(\sup_{0 \leq t \leq T} |X_t^{\gamma} - \phi_t| < \delta) + P_x(N_T^{\gamma,\phi,\theta} \geq \gamma B), \end{aligned} \tag{4.21}$$

where the second step uses the fact that $\log(1+\epsilon) \leq \epsilon$. Inequality (4.21), together with Lemma 4.5.1 and inequality (4.20) on the left-hand side, and the choice of B on the right-hand side imply that

$$\begin{aligned} &-(1+\epsilon)\Gamma(\phi, x_0) - (1+\epsilon)\epsilon emT \leq \\ &\qquad \left(\epsilon(mT+B) + \lim_{\delta \searrow 0} \lim_{\rho \searrow 0} \liminf_{\gamma \to \infty} \gamma^{-1} \log \inf_{|x-\phi_0| < \rho} P_x(\sup_{0 \leq t \leq T} |X_t^{\gamma} - \phi_t| < \delta) \right) \bigvee (-\Gamma(\phi, x_0)). \end{aligned}$$

The lemma follows by the arbitrariness of $\epsilon > 0$. □

Given $\phi \in C_{[0,T]}(R^d)$ and $\delta > 0$, let $B(\phi, \delta)$ continue to denote the open ball of radius δ around ϕ.

Lemma 4.5.3 *(Lower Bound) For any Borel measurable $S \subset C_{[0,T]}(R^d)$, $x_0 \in R^d$, and sequence $(x^\gamma : \gamma > 0)$ such that $\lim_{\gamma\to\infty} x^\gamma = x_0$,*

$$\liminf_{\gamma\to\infty} \gamma^{-1} \log P_{x^\gamma}\left((\tilde{X}_t^\gamma : 0 \le t \le T) \in S\right) \ge -\inf_{\phi\in S^o} \Gamma(\phi, x_0).$$

Proof. Fix $\phi \in S^o$, and let $\delta' > 0$ be such that $B(\phi, \delta)$ is contained in S for all $\delta < \delta'$. Lemma 4.3.2 and Lemma 4.5.2 imply that

$$\begin{aligned}
\liminf_{\gamma\to\infty} \gamma^{-1} \log P_{x^\gamma}\left((\tilde{X}_t^\gamma : 0 \le t \le T) \in S\right) &\ge \lim_{\delta\searrow 0}\lim_{\rho\searrow 0}\liminf_{\gamma\to\infty} \gamma^{-1} \log \inf_{|x-x_0|<\rho} P_x(\sup_{0\le t\le T} |\tilde{X}_t^\gamma - \phi_t| < \delta) \\
&\ge \lim_{\delta\searrow 0}\lim_{\rho\searrow 0}\liminf_{\gamma\to\infty} \gamma^{-1} \log \inf_{|x-x_0|<\rho} P_x(\sup_{0\le t\le T} |X_t^\gamma - \phi_t| < \delta) \\
&\ge -\Gamma(\phi, x_0).
\end{aligned}$$

Since $\phi \in S^o$ is arbitrary, the lemma follows. □

4.6 The Upper Bound

This section establishes the large deviations upper bound for Theorem 4.2.1 by adapting the methods of Section 4.5.

Lemma 4.6.1 *(Intermediate Upper Bound) For each $T > 0$, partition $\theta = (\theta_0, \cdots, \theta_{J(\theta)})$ of $[0, T]$, and absolutely continuous $\phi \in C_{[0,T]}(R^d)$,*

$$\lim_{\delta\searrow 0}\limsup_{\gamma\to\infty} \gamma^{-1} \log \sup_{|x-\phi_0|<\delta} P_x(\sup_{0\le t\le T} |X_t^{\gamma,\phi,\theta} - \phi_t| < \delta) \le -\sum_{i=0}^{J(\theta)-1} \int_{\theta_i}^{\theta_{i+1}} \Lambda^{\phi_{\theta_i}}(\phi_t, \dot{\phi}_t)dt. \qquad (4.22)$$

Furthermore, if ϕ is not absolutely continuous, then the left-hand side of (4.22) equals $-\infty$.

Proof. By induction on $J(\theta)$. Lemma 4.4.1, along with Remark 3.2.1 and Lemma 4.3.2, implies that the statement of the lemma holds whenever $J(\theta) = 1$. As the induction hypothesis, let $k \ge 1$

and suppose that the lemma holds for any $T > 0$ and partition θ of $[0,T]$ such that $J(\theta) = k$. To show that the claim holds for $J(\theta) = k+1$, note that by the Markov property of $X^{\gamma,\phi,\theta}$,

$$\gamma^{-1}\log \sup_{|x-\phi_0|<\delta} P_x(\sup_{0\le t\le\theta_{k+1}} |X_t^{\gamma,\phi,\theta}-\phi_t|<\delta) \le \gamma^{-1}\log \sup_{|x-\phi_0|<\delta} P_x(\sup_{0\le t\le\theta_k} |X_t^{\gamma,\phi,\theta}-\phi_t|<\delta)$$
$$+\gamma^{-1}\log \sup_{|y-\phi_{\theta_k}|<\delta} P_x(\sup_{\theta_k\le t\le\theta_{k+1}} |X_t^{\gamma,\phi,\theta}-\phi_t|<\delta \mid X_{\theta_k}^{\gamma,\phi,\theta}=y).$$

Therefore, if ϕ is absolutely continuous, then the induction hypothesis and the time-homogeneous Markov property of the pair $(X^{\gamma,\phi,\theta},\phi)$ imply

$$\lim_{\delta\searrow 0}\limsup_{\gamma\to\infty}\gamma^{-1}\log \sup_{|x-\phi_0|<\delta} P_x(\sup_{0\le t\le\theta_{k+1}} |X_t^{\gamma,\phi,\theta}-\phi_t|<\delta) \le -\sum_{i=0}^{k}\int_{\theta_i}^{\theta_{i+1}}\Lambda^{\phi_{\theta_i}}(\phi_t,\dot\phi_t)dt. \qquad (4.23)$$

Otherwise, either $(\phi_t : 0 \le t \le \theta_k)$ or $(\phi_t : \theta_k \le t \le \theta_{k+1})$ is not absolutely continuous; hence the left-hand side of (4.23) equals $-\infty$. This completes the induction step and establishes the lemma. □

Lemma 4.6.2 *(Local Upper Bound) For each $\phi \in C_{[0,T]}(R^d)$ and $x_0 \in R^d$,*

$$\lim_{\delta\searrow 0}\limsup_{\gamma\to\infty}\gamma^{-1}\log \sup_{|x-x_0|<\delta} P_x(\sup_{0\le t\le T} |X_t^{\gamma}-\phi_t|<\delta) \le -\Gamma(\phi,x_0). \qquad (4.24)$$

Proof. Without loss of generality, we may assume that ϕ is continuous and $\phi_0 = x_0$, since otherwise the left-hand side of (4.24) equals $-\infty$. Fix $\epsilon > 0$, and choose $\delta > 0$ such that

$$(1+\epsilon)^{-1} \le \frac{d\nu^+(x)}{d\nu^+(x')} \le (1+\epsilon) \quad \text{and} \quad (1+\epsilon)^{-1} \le \frac{d\nu^-(x)}{d\nu^-(x')} \le (1+\epsilon)$$

whenever $|x-x'| < 2\delta$. Appeal to the uniform continuity of ϕ on $[0,T]$ to choose a partition $\theta = (\theta_0,\cdots,\theta_{J(\theta)})$ of $[0,T]$ such that $\sup_{\theta_i\le t\le\theta_{i+1}}|\phi_t-\phi_{\theta_i}| < \delta$ for each $i \in \{0,\cdots,J(\theta)-1\}$.

Let N_T^γ and $N_T^{\gamma,\phi,\theta}$ denote, respectively, the number of jumps of X^γ and $X^{\gamma,\phi,\theta}$ in the interval $[0,T]$. By the choice of θ and Equation (4.19), for each $x \in R^d$, $\gamma > 0$, and $B > 0$,

$$P_x(\sup_{0\le t\le T}|X_t^\gamma-\phi_t|<\delta) \le P_x(\sup_{0\le t\le T}|X_t^\gamma-\phi_t|<\delta, N_T^\gamma<\gamma B) + P_x(N_T^\gamma \ge \gamma B)$$

$$\leq e^{\gamma\epsilon(mT+B)} P_x\Big(\sup_{0\leq t\leq T} |X_t^{\gamma,\phi,\theta} - \phi_t| < \delta, N_T^{\gamma,\phi,\theta} < \gamma B\Big) + P_x(N_T^\gamma \geq \gamma B)$$

$$\leq e^{\gamma\epsilon(mT+B)} P_x\Big(\sup_{0\leq t\leq T} |X_t^{\gamma,\phi,\theta} - \phi_t| < \delta\Big) + P_x(N_T^\gamma \geq \gamma B), \tag{4.25}$$

where the second step uses the fact that $\log(1+\epsilon) \leq \epsilon$. By hypothesis, uniformly for all initial states, N_T^γ is stochastically dominated by a Poisson random variable with mean γmT. Therefore, if ϕ is not absolutely continuous, then inequality (4.25), along with Lemma 4.6.1 and choice of arbitrarily large B on the right-hand side, implies that the left-hand side of (4.24) equals $-\infty$, and the lemma holds.

If ϕ is absolutely continuous, then the choice of θ and Lemma 4.3.3 applied to the definition of $\Lambda^{\phi_{\theta_i}}$ imply

$$\begin{aligned}\sum_{i=0}^{J(\theta)-1} \int_{\theta_i}^{\theta_{i+1}} \Lambda^{\phi_{\theta_i}}(\phi_t, \dot\phi_t) dt &\geq \sum_{i=0}^{J(\theta)-1} \int_{\theta_i}^{\theta_{i+1}} \Big((1+\epsilon)^{-1}\Lambda^{\phi_t}(\phi_t, \dot\phi_t) + \epsilon em\Big)\, dt \\ &= (1+\epsilon)^{-1}\Gamma(\phi, x_0) + \epsilon emT. \end{aligned} \tag{4.26}$$

Appeal to Condition 4.2.1 to choose B large enough so that

$$\limsup_{\gamma\to\infty} \gamma^{-1} \log \sup_x P_x(N_T^\gamma \geq \gamma B) \leq -\Gamma(\phi, x_0).$$

Then inequality (4.25), together with Lemma 4.6.1, inequality (4.26), and the choice of B, implies that

$$\lim_{\delta\searrow 0} \limsup_{\gamma\to\infty} \gamma^{-1} \log \sup_{|x-x_0|<\delta} P_x\Big(\sup_{0\leq t\leq T} |X_t^\gamma - \phi_t| < \delta\Big) \leq$$
$$\Big(-(1+\epsilon)^{-1}\Gamma(\phi, x_0) - \epsilon emT + \epsilon(mT+B)\Big) \bigvee \big(-\Gamma(\phi, x_0)\big).$$

The lemma follows by the arbitrariness of $\epsilon > 0$. □

Lemma 4.6.3 *(Exponential Tightness) Let $(x^\gamma : \gamma > 0)$ be a sequence such that $\lim_{\gamma\to\infty} x^\gamma = x_0$. For each $\alpha > 0$ there exists a compact $K_\alpha \subset C_{[0,T]}(R^d)$ such that*

$$\limsup_{\gamma\to\infty} \gamma^{-1} \log P_{x^\gamma}\Big((\tilde X_t^\gamma : 0 \leq t \leq T) \notin K_\alpha\Big) \leq -\alpha.$$

Proof. The lemma follows by [12, Lemma 5.58] through a straightforward adaptation of [12, Corollary 5.8] so as to incorporate the continuous support of the measures $\nu(x)$, $x \in R^d$. □

Lemma 4.6.4 *(Upper Bound) For any Borel measurable* $S \subset C_{[0,T]}(R^d)$, $x_0 \in R^d$, *and sequence* $(x^\gamma : \gamma > 0)$ *such that* $\lim_{\gamma\to\infty} x^\gamma = x_0$,

$$\limsup_{\gamma\to\infty} \gamma^{-1} \log P_{x^\gamma}\left((\tilde{X}_t^\gamma : 0 \le t \le T) \in S\right) \le -\inf_{\phi\in\overline{S}} \Gamma(\phi, x_0).$$

Proof. Fix $\epsilon > 0$. For each $\phi \in \overline{S}$, appeal to Lemma 4.6.2 and Lemma 4.3.2 to choose a $\delta_\phi > 0$ such that

$$\limsup_{\gamma\to\infty} \gamma^{-1} \log P_{x^\gamma}\left((\tilde{X}_t^\gamma : 0 \le t \le T) \in B(\phi, \delta_\phi)\right) \le -\Gamma(\phi, x_0) + \epsilon,$$

and appeal to Lemma 4.6.3 to choose a compact subset K of $C_{[0,T]}(R^d)$ such that

$$\limsup_{\gamma\to\infty} \gamma^{-1} \log P_{x^\gamma}\left((\tilde{X}_t^\gamma : 0 \le t \le T) \notin K\right) \le -(\frac{1}{\epsilon} \wedge \inf_{\phi\in\overline{S}} \Gamma(\phi, x_0)).$$

By the compactness of $\overline{S} \cap K$, there exists a finite subset $\{\phi^1, \cdots, \phi^I\} \subset \overline{S}$ such that $\overline{S} \cap K \subset \cup_{i=1}^I B(\phi^i, \delta_{\phi^i})$; hence for each $\gamma > 0$,

$$P_{x^\gamma}\left((\tilde{X}_t^\gamma : 0 \le t \le T) \in S\right) \le P_{x^\gamma}\left((\tilde{X}_t^\gamma : 0 \le t \le T) \notin K\right) + \sum_{i=1}^I P_{x^\gamma}\left((\tilde{X}_t^\gamma : 0 \le t \le T) \in B(\phi^i, \delta_{\phi^i})\right).$$

This in turn implies

$$\begin{aligned}
\limsup_{\gamma\to\infty} \gamma^{-1} \log P_{x^\gamma}\left((\tilde{X}_t^\gamma : 0 \le t \le T) \in S\right) &\le \left(-(\frac{1}{\epsilon} \wedge \inf_{\phi\in\overline{S}} \Gamma(\phi, x_0))\right) \bigvee \max_{1\le i\le I}\{-\Gamma(\phi^i, x_0) + \epsilon\} \\
&\le \left(-(\frac{1}{\epsilon} \wedge \inf_{\phi\in\overline{S}} \Gamma(\phi, x_0))\right) \bigvee \left(-\inf_{\phi\in\overline{S}} \Gamma(\phi, x_0) + \epsilon\right).
\end{aligned}$$

The lemma now follows by the arbitrariness of $\epsilon > 0$. □

4.7 Goodness of the Rate Function

This section concludes the proof of Theorem 4.2.1 by establishing the goodness of the rate function Γ.

Lemma 4.7.1 *(Lower Semicontinuity) Given $x_0 \in R^d$, the function $\Gamma(\cdot, x_0) : C_{[0,T]}(R^d) \to R_+ \cup \{+\infty\}$ is lower semicontinuous.*

Proof. Let $(\phi^m : m > 0)$ be a sequence such that $\phi^m \to \phi$ in $C_{[0,T]}(R^d)$. To prove the lemma it is enough to show that $\Gamma(\phi, x_0) \leq \liminf_{m\to\infty} \Gamma(\phi^m, x_0)$. Fix $\epsilon > 0$ and a sequence $x^\gamma \to x_0$. By Lemma 4.6.2, there exists a $\delta > 0$ such that

$$\limsup_{\gamma\to\infty} \gamma^{-1} \log P_{x^\gamma}\left((\tilde{X}_t^\gamma : 0 \leq t \leq T) \in B(\phi, \delta)\right) \leq -\Gamma(\phi, x_0) + \epsilon. \tag{4.27}$$

Let M_ϵ be such that $\sup_{0\leq t\leq T} |\phi_t^m - \phi_t| < \delta$ whenever $m > M_\epsilon$. By Lemma 4.5.3,

$$\begin{aligned} \liminf_{\gamma\to\infty} \gamma^{-1} \log P_{x^\gamma}\left((\tilde{X}_t^\gamma : 0 \leq t \leq T) \in B(\phi, \delta)\right) &\geq -\inf_{\phi' \in B(\phi,\delta)} \Gamma(\phi', x_0) \\ &\geq -\Gamma(\phi^m, x_0) \end{aligned} \tag{4.28}$$

for all $m > M_\epsilon$. Inequalities (4.27) and (4.28) imply that $\Gamma(\phi^m, x_0) \geq \Gamma(\phi, x_0) - \epsilon$ whenever $m > M_\epsilon$, and the lemma is proved. □

Lemma 4.7.2 *(Goodness) The rate function Γ is good.*

Proof. In view of [13, Lemma 1.2.18.b], the lemma is implied by Lemmas 4.5.3, 4.6.3, and 4.7.1. □

CHAPTER 5

CONCLUSION

We concentrated on complexity-performance trade-offs in dynamic resource allocation in load sharing networks. Realizing the difficulties involved in the exact analysis of arbitrary network structures, we first focused on the optimality properties of certain simple allocation policies implied by the corresponding fluid limit approximations. Explicit fluid equations were obtained, and through a characterization of their solutions it was shown that the LLR policy asymptotically achieves the most balanced load in the sense of minimizing a wide class of long-term average costs. LLR is also robust to migration, provided that consumers are reassigned according to LLR whenever their types change. When the resources have finite capacities, the class of LRLR policies asymptotically achieve the minimum possible blocking probability. From a practical point of view, important properties of the considered policies are low computational complexity, decentralized implementation, and robustness to arrival and migration rates.

The fluid limit approximations considered here are essentially process-level laws of large numbers for load sharing networks; thus, they provide only a first-order description of the network behavior. This description does not appear to be accurate enough to contrast certain allocation policies. In particular, the OR, BS, and LLR policies appear to have the same performance in the fluid scale. The second part of the thesis concerned a finer analysis to order the three policies: The theory of large deviations was employed, and network overflow was studied in terms of overflow exponents. The overflow exponents for each policy were obtained in the simple W network, and it was shown that the LLR policy performs as well as the OR policy for small values of capacities, whereas it performs significantly better then the BS policy for the whole range of capacities. The general form of the overflow exponents for arbitrary network topologies is identified for the OR and BS policies and conjectured for the LLR policy.

Obtaining overflow exponents for the LLR policy entails establishing large deviations principles (LDPs) for Markov processes with discontinuous transition mechanisms. An explicit LDP was established in the case when the discontinuity is along a singie hyperplane, and it was used to

obtain the overflow exponents in the W network. The established LDP generalizes the work of Blinovskii and Dobrushin and holds under somewhat more relaxed technical conditions than those required by related results in the literature.

The results of this thesis suggest maximization of overflow exponents as a guiding principle for capacity allocation and policy design in load sharing networks. Although obtaining overflow exponents for general networks appears difficult, rigorous study of simple examples, such as least ratio routing and maximum residual capacity routing in the W network, may yield good heuristic arguments. These issues remain to be explored.

REFERENCES

[1] R. Gibbens, F. P. Kelly, and S. Turner, "Dynamic routing in multiparented networks," *IEEE/ACM Transactions on Networking*, no. 2, pp. 261–270, 1993.

[2] G. M. Chiu, C. S. Raghavendra, and S. S. Ng, "Resource allocation with load balancing consideration in distributed systems," in *Proceedings of IEEE Infocom 89*, 1989, pp. 758–765.

[3] G. R. Ganger, B. L. Worthington, R. Y. Hou, and Y. N. Patt, "Disk subsystem load balancing: Disk striping vs. conventional data placement," in *Proceedings of 26th Hawaii International Conference on System Sciences*, 1993, pp. 40–49.

[4] B. Hajek, "Performance of global load balancing by local adjustment," *IEEE Transactions on Information Theory*, vol. 36, no. 6, pp. 1398–1414, 1990.

[5] H. T. Liu and J. Silvester, "An approximate performance model for load-dependent interactive queues with application to load balancing in distributed systems," in *Proceedings of IEEE Infocom 88*, 1988, pp. 956–965.

[6] M. H. Willebeck-LeMair and A. P. Reeves, "Strategies for dynamic load balancing on highly parallel computers," *IEEE Transactions on Parallel and Distributed Systems*, no. 9, pp. 979–993, 1993.

[7] P. J. Hunt and T. Kurtz, "Large loss networks," *Stochastic Processes and Their Applications*, no. 53, pp. 363–378, 1994.

[8] Y. Azar, A. Broder, and A. Karlin, "On-line load balancing," in *Proceedings of 33rd Annual Symposium on FOCS*, 1992, pp. 218–225.

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

[10] J. G. Dai and R. J. Williams, "Existence and uniqueness of semimartingale reflecting Brownian motion in convex polyhedrons," *Probability Theory and Its Applications*, vol. 40, no. 1, pp. 3–53, 1995.

[11] H. L. Royden, *Real Analysis*, 3rd ed. New York: Macmillan, 1988.

[12] A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis, Queues, Communication and Computing*, London: Chapman & Hall, 1995.

[13] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Boston: Jones and Bartlett, 1992.

[14] V. M. Blinovskii and R. L. Dobrushin, "Process level large deviations for a class of piecewise homogeneous random walks," in *The Dynkin Festschrift: Markov Processes and Their Applications*, Boston: Birkhauser, 1994, pp. 1–59.

[15] P. Dupuis and R. S. Ellis, "The large deviation principle for a general class of queueing systems I," *Transactions of the American Mathematical Society*, vol. 347, no. 8, pp. 2689–2751, 1995.

[16] P. Dupuis and R. S. Ellis, "Large deviations for Markov processes with discontinuous statistics, II: Random walks," *Probability Theory and Related Fields*, no. 91, pp. 153–194, 1992.

[17] V. Malyshev, I. A. Ignatyuk, and V. V. Scherbakov, "Boundary effects in large deviation problems," *Russian Mathematical Surveys*, vol. 49, no. 2, pp. 41–99, 1994.

[18] R. S. Ellis, P. Dupuis, and A. Weiss, "Large deviations for Markov processes with discontinuous statistics, I: General upper bounds," *Annals of Probability*, , no. 19, pp. 1280–1297, 1991.

[19] R.T. Rockafellar, *Convex analysis*, New Jersey: Princeton University Press, 1970.

VITA

Murat Alanyali was born in Ankara, Turkey, on January 2, 1966. He received the B.S. degree from Middle East Technical University, in 1988 and the M.S. degree from Bilkent University, in 1990. He was a part-time researcher at ASELSAN Military Electronic Industries, Inc., Ankara from August 1986 to August 1988, a research assistant at Bilkent University from August 1988 to July 1990, and at the University of Illinois from August 1992 to January 1996.

He has authored and coauthored the following papers:

- M. Alanyali and B. Hajek, "On large deviations of Markov processes with discontinuous statistics," submitted to *The Annals of Applied Probability*, January 1996.
- M. Alanyali and B. Hajek, "On large deviations in load sharing networks," submitted to *The Annals of Applied Probability*, January 1996.
- M. Alanyali and B. Hajek, "On load balancing in Erlang networks," submitted to *Stochastic Networks: Theory and Applications*, F. Kelly, S. Zachary, and I. Ziedins, Eds. Oxford: Oxford University Press, December 1995.
- M. Alanyali and B. Hajek, "Analysis of simple algorithms for dynamic load balancing," submitted to *Mathematics of Operations Research*, June 1995.
- M. Alanyali and B. Hajek, "On simple algorithms for dynamic load balancing," *Proceedings of IEEE INFOCOM 95*, vol. 1, pp. 230–238, April 1995.
- M. Alanyali, "Neural networks in optimization problems," *Proceedings of Yöneylem Araştırması 12. Ulusal Kongresi*, pp. 297-298, June 1989. (In Turkish.)
- M. Alanyali, M. B. Alp and M. Yücel, "Image coding via discrete cosine and Hadamard transforms," *Proceedings of ODTÜ Elektrik-Elektronik Mühendisliği Bölümü 30. Yıl Sempozyumu*, pp. 13-16, February 1989. (In Turkish.)