Adv. Appl. Prob. **43,** 545–571 (2011)
Printed in Northern Ireland
© *Applied Probability Trust* 2011

LARGE DEVIATIONS FOR A FEED-FORWARD NETWORK

LEILA SETAYESHGAR [*] AND
HUI WANG,[*] [**] *Brown University*

Abstract

We consider a feed-forward network with a single-server station serving jobs with multiple levels of priority. The service discipline is preemptive in that the server always serves a job with the current highest level of priority. For this system with discontinuous dynamics, we establish the sample path large deviation principle using a weak convergence argument. In the special case where jobs have two different levels of priority, we also explicitly identify the exponential decay rate of the total population overflow probabilities by examining the geometry of the zero-level sets of the system Hamiltonians.

Keywords: Large deviation; weak convergence; discontinuous dynamics

2010 Mathematics Subject Classification: Primary 60F10; 65C05
Secondary 49N90

1. Introduction

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

The theory of large deviations is concerned with the asymptotic behavior of tails of sequences of probability distributions. Let S be a Polish space equipped with the Borel σ-algebra, and let $\{X^n\}$ be a sequence of S-valued random variables. A lower semicontinuous function $I\colon S \to [0, \infty]$ with compact level sets is said to be a large deviation *upper bound rate function* if, for every closed subset F of S,

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

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

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

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

Received 25 January 2010; revision received 18 December 2010.

[*] Postal address: Division of Applied Mathematics, Brown University, 182 George Street, Providence, RI 02912, USA. This research was supported in part by the National Science Foundation (NSF-DMS-0706003).

[**] Email address: huiwang@dam.brown.edu

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

Large deviations analysis for stable stochastic systems with continuous dynamics has been a classical topic in probability theory [15]. However, the general methodologies and techniques therein cannot be applied to models with discontinuous dynamics that arise naturally in a variety of applications (notably queueing networks). In the last two decades, research on the large deviations properties of such models has become more and more popular and many interesting results have been obtained [4], [6], [14], [20], [22]. With minor regularity conditions, it is possible to establish an explicit large deviation upper bound rate function [8] for stochastic systems with very general discontinuous dynamics. However, this upper bound rate function is *not* a lower bound rate function in general [1], [17]. The reason for this gap lies in the so called 'stability-about-the-interface' condition. To give an intuitive explanation, let us consider a simple model of random walk in $\mathbb{R}^d$ where the dynamics are constant in the two half-spaces $\Lambda_1 = \{x \in \mathbb{R}^d : x_1 \leq 0\}$ and $\Lambda_2 = \{x \in \mathbb{R}^d : x_1 > 0\}$. Denote by L_i the large deviation local rate function for the dynamics in the region Λ_i, $i = 1, 2$. The upper bound rate function suggested by Dupuis *et al.* [8] on the interface $\Sigma = \{x \in \mathbb{R}^d : x_1 = 0\}$ is the inf-convolution of L_1 and L_2. That is, for every $x \in \Sigma$ and $\beta \in \mathbb{R}^d$,

$$L(x; \beta) = \inf\, [\rho_1 L_1(\nu) + \rho_2 L_2(\theta)], \tag{1.1}$$

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

$$\nu \in \mathbb{R}^d, \qquad \theta \in \mathbb{R}^d, \qquad \rho_1 \geq 0, \qquad \rho_2 \geq 0, \qquad \rho_1 + \rho_2 = 1, \qquad \rho_1 \nu + \rho_2 \theta = \beta. \tag{1.2}$$

This upper bound rate function L is not a lower bound rate function in general. Indeed, it was shown in [7, Chapter 7] that the large deviation rate function is defined exactly as in (1.1)–(1.2) but with the extra constraints (i.e. the stability-about-the-interface condition)

$$\nu_1 \geq 0 \quad \text{and} \quad \theta_1 \leq 0$$

in (1.2). The reason for these extra constraints is that in order to prove a large deviation lower bound, we need to analyze the cost associated with a piece of trajectory that travels on the interface Σ. This is usually achieved by a change-of-measure argument so that the state process closely tracks the trajectory under the new probability distribution. The vital role of this stability-about-the-interface condition is to characterize all those changes of measures that lead to the desired tracking behavior; see [7, Chapter 7] for more details.

The current paper consists of two parts. In the first part we establish the sample path large deviation principle for the feed-forward network under consideration. It turns out that the stability-about-the-interface condition is *implicitly* built into the upper bound rate function [8]. Consequently, the upper bound rate function is indeed the rate function. Similar results have been obtained by Atar and Dupuis [3], whose analysis used the techniques of the Skorokhod problem and, therefore, does not apply here. We also wish to point out that the analysis in [18] can be applied to the current system to establish a sample path large deviation principle. However, in [18] the rate function is only implicitly defined in terms of the convergence parameters of the transform semigroup. Furthermore, we use a different approach based on weak convergence, which seems to be very powerful, especially in dealing with discontinuous dynamics; see also [10].

The simple form of the upper bound rate function (or the rate function) allows us to characterize through partial differential equations the asymptotic behavior of various types of buffer overflow probabilities. In the second part, we illustrate this connection by explicitly

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

identifying the exponential decay rate of the total population overflow probabilities when the exogenous jobs have two levels of priority. The form of the decay rate is motivated by examining the geometry of the zero-level sets of the system Hamiltonians, and then rigorously verified by constructing suitable subsolutions to the related partial differential equation.

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

This paper is organized as follows. In Section 2, the model setup and system dynamics are introduced. The rate function and the large deviation principle are stated in Section 3. In Section 4 we specialize to the two-dimensional case and explicitly identify the exponential decay rate of the total population overflow probabilities. A brief summary is given in Section 5. Some of the technical proofs are deferred to the appendices.

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

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

2. The model setup and system dynamics

We consider a single-server station serving d classes of exogenous jobs. Jobs of class i, $i = 1,\dots,d$, arrive according to a Poisson process with rate $\lambda_i > 0$, and are buffered at queue i. The service time for a class-i job is exponentially distributed with rate $\mu_i > 0$. The arrival processes and service times are assumed to be mutually independent. The system adopts a service discipline such that a job of class i has preemptive priority over a job of class j whenever $i > j$, and the server always serves a job with the current highest level of priority. Jobs with the same priority level are served according to the first-in–first-out policy. See Figure 1.

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

$$\Pi(x) = \max\{i : x_i > 0\} \quad \text{with the convention that } \Pi(0) = 0. \tag{2.1}$$

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

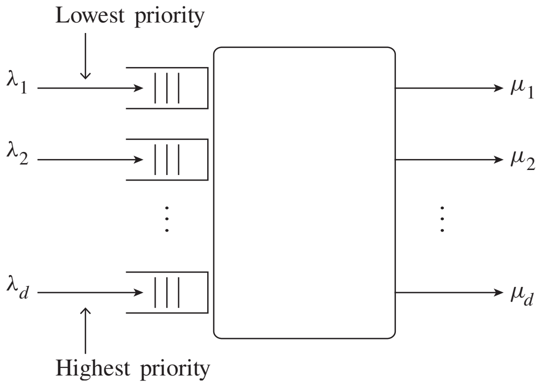

FIGURE 1: Feed-forward network with preemptive priority service policy.

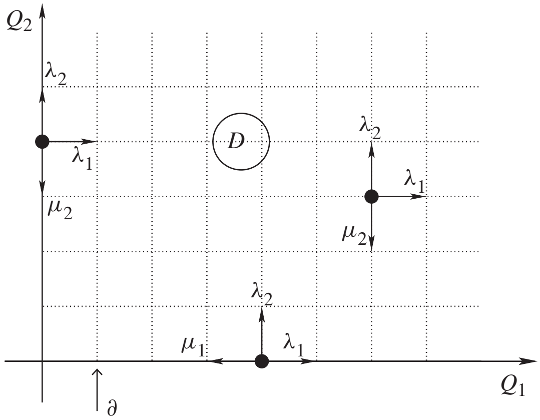

FIGURE 2: System dynamics for $d = 2$.

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

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

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

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

The dynamics of the system are discontinuous at the interface $\{x : \Pi(x) = i\}$ for each $0 \leq i \leq d - 1$. Thus, there are in total d interfaces of discontinuity whose dimensions range from 0 to $d - 1$. These interfaces are also boundaries of the state space. See Figure 2.

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

3. The large deviations analysis

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

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

The processes $\{X^n : n \in \mathbb{N}\}$ are again continuous-time pure-jump Markov processes.

3.1. Hamiltonians and rate functions

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

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

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

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

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

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

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

$$\begin{aligned} \bar{L}_i(\beta) &= (L_i \oplus L_{i+1} \oplus \cdots \oplus L_d)(\beta) \\ &= \inf\Bigg\{ \sum_{j=i}^{d} \rho_j L_j(\beta_j) \colon \beta_j \in \mathbb{R}^d,\ \rho_j \ge 0,\ \sum_{j=i}^{d} \rho_j = 1,\ \sum_{j=i}^{d} \rho_j \beta_j = \beta \Bigg\}. \end{aligned} \tag{3.1}$$

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

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

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

3.2. Sample path large deviations

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

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

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

if $\phi(0) = x$, $\phi(t) \in \mathbb{R}_+^d$ for all t, and ϕ is absolutely continuous, and set $I_x(\phi) = \infty$ otherwise. It was established in [8] that the rate function $\{I_x : x \in \mathbb{R}_+^d\}$ is an upper bound rate function and has compact level sets on compacts in the sense that the set

$$\bigcup_{x \in C} \{\phi \colon I_x(\phi) \leq M\}$$

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

As pointed out in the introduction, the stability-about-the-interface condition is automatically built into the inf-convolution definition of the local rate function; thus, this upper bound rate function is tight. More precisely, we have the following main result regarding the sample path large deviation properties of $\{X^n\}$.

Recall that the large deviation principle and the Laplace principle are equivalent for probability measures on a Polish space [7, Theorem 1.2.1 and Theorem 1.2.3]. Let E_{x_n} denote the expectation conditional on $X^n(0) = x_n$.

Theorem 3.1. *The processes $\{X^n(t) \colon t \in [0, T]\}$ satisfy the uniform Laplace principle with rate functions $\{I_x : x \in \mathbb{R}_+^d\}$. That is, for any sequence $\{x_n\} \subseteq \mathbb{R}_+^d$ such that $x_n \to x$ and any bounded continuous function $h \colon \mathcal{D}([0, T] \colon \mathbb{R}^d) \to \mathbb{R}$, we have*

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

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

3.3. Proof of Theorem 3.1

Throughout the proof, we will assume without loss of generality that $T = 1$. The uniform Laplace principle upper bound is implied by the uniform large deviation upper bound [8, Theorem 1.1] through an argument analogous to [7, Theorem 1.2.1]. Therefore, it suffices to show the uniform Laplace principle lower bound. That is, to show that

$$\liminf_{n} \frac{1}{n} \log \mathrm{E}_{x_n}\{\exp[-nh(X^n)]\} \geq - \inf_{\phi \in \mathcal{D}([0,1] \colon \mathbb{R}^d)} [I_x(\phi) + h(\phi)]. \tag{3.2}$$

Since the above inequality holds trivially if $I_x(\phi) = \infty$, we can a priori assume that $I_x(\phi)$ is finite, which dictates that ϕ is absolutely continuous.

For the convenience of the reader, we divide the long proof into four steps. In step 1, an alternative representation for the left-hand side of (3.2) is established, which turns the analysis of the lower bound (3.2) into that of a stochastic control problem. The construction of nearly optimal controls is given in step 2. The analysis of the limit controlled process is carried out in step 3 via the weak convergence approach. The desired lower bound (3.2) is finally established in step 4.

Step 1: relative entropy representation. The proof utilizes the relative entropy representation for exponential integrals [7, Proposition 1.4.2] and requires the construction of an appropriately controlled process. To this end, it is often convenient to restrict ϕ to a more analytically tractable class $\mathcal{N}$, which consists of those absolutely continuous functions $\phi^* \colon [0, 1] \to \mathbb{R}_+^d$ such that there exists a positive integer K and a partition $0 = t_0 < t_1 < \cdots < t_{k-1} < t_K = 1$ where on each open interval (t_{i-1}, t_i), $i = 1, \ldots, K$, both $\dot{\phi}^*$ and $\Pi(\phi^*)$ take constant values. The following lemma states that any trajectory ϕ with finite cost can be approximated by a trajectory in class $\mathcal{N}$. The proof of this lemma is deferred to Appendix A.

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

Lemma 3.1. *Given any $\phi \in \mathcal{D}([0,1]\colon \mathbb{R}^d)$ such that $I_x(\phi) < \infty$ and any $\delta > 0$, there exists a $\phi^* \in \mathcal{N}$ such that $\|\phi - \phi^*\|_\infty < \delta$ and $I_x(\phi^*) \le I_x(\phi)$.*

Thanks to this lemma and the continuity of h, it is easy to see that in order to show the lower bound (3.2), we only need to prove that

$$\liminf_n \frac{1}{n} \log \mathrm{E}_{x_n}\{\exp[-nh(X^n)]\} \ge -[I_x(\phi^*) + h(\phi^*)] \tag{3.3}$$

for every $\phi^* \in \mathcal{N}$. Denote by P_n the probability measure induced by X^n on the Polish space $\mathcal{D}([0,1]\colon \mathbb{R}^d)$. Then, by the relative entropy representation of exponential integrals [7, Section 1.4],

$$-\frac{1}{n}\log \mathrm{E}_{x_n}\{\exp[-nh(X^n)]\} = \inf\biggl[\frac{1}{n}R(\mathrm{Q}\,\|\,\mathrm{P}_n) + \int_{\mathcal{D}([0,1]\colon \mathbb{R}^d)} h\,\mathrm{dQ}\biggr],$$

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

$$\bar{\mathcal{L}}^n f(x,t) = n\sum_{v\in\mathbb{V}} \bar{r}(x,t;v)\biggl[f\biggl(x+\frac{v}{n}\biggr) - f(x)\biggr]. \tag{3.4}$$

Here $\bar{r}(x,t;v)$ is nonnegative and uniformly bounded, and also satisfies $\bar{r}(x,t;v) = 0$ whenever $r(x;v) = 0$ (in other words, $\bar{X}^n$ is the scaled version of a jump Markov process with $\bar{r}(x,t;v)$ as the jump intensity from state x to $x+v$ at time t). If we restrict the infimum to such probability measures, for which the explicit evaluation of the relative entropy $R(\cdot\,\|\,\mathrm{P}_n)$ is available [21, Theorem B.6], we arrive at the inequality

$$\begin{aligned}&-\frac{1}{n}\log \mathrm{E}_{x_n}\{\exp[-nh(X^n)]\}\\ &\quad\le \inf_{\bar{r}} \mathrm{E}_{x_n}\biggl\{\int_0^1 \sum_{v\in\mathbb{V}} r(\bar{X}^n(t);v)\ell\biggl(\frac{\bar{r}(\bar{X}^n(t),t;v)}{r(\bar{X}^n(t);v)}\biggr)\,\mathrm{d}t + h(\bar{X}^n)\biggr\},\end{aligned}$$

where ℓ is defined by

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

with the convention that $0\cdot\ell(0/0) = 0$. Therefore, in order to prove (3.3), it suffices to construct, for an arbitrarily fixed positive constant ε, an alternative jump intensity function $\bar{r}$ (dependent on ε) such that

$$\begin{aligned}&\limsup_n \mathrm{E}_{x_n}\biggl\{\int_0^1 \sum_{v\in\mathbb{V}} r(\bar{X}^n(t);v)\ell\biggl(\frac{\bar{r}(\bar{X}^n(t),t;v)}{r(\bar{X}^n(t);v)}\biggr)\,\mathrm{d}t + h(\bar{X}^n)\biggr\}\\ &\quad\le I(\phi^*) + h(\phi^*) + \varepsilon.\end{aligned} \tag{3.5}$$

We now set forth to prove this inequality.

Step 2: construction of $\bar{r}$. The construction of $\bar{r}$ is based on the representation of the rate function $\bar{L}_i$ in terms of the function ℓ. More precisely, we have the following lemma, whose proof is very similar to that given in [10, Section 4.3]. For the sake of completeness, we include the proof in Appendix B.

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

Lemma 3.2. *Given $\beta \in \mathbb{R}^d$ and $i = 0, 1, \dots, d$, we have the representation*

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

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

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

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

Since $\phi^* \in \mathcal{N}$, there exists a partition $0 = t_0 < t_1 < \cdots < t_{K-1} < t_K = 1$ such that on the open interval (t_j, t_{j+1}), both $\dot{\phi}^*(t)$ and $\Pi(\phi^*(t))$ take constant values, say $\dot{\phi}^*(t) = \beta_j$ and $\Pi(\phi^*(t)) = I_j$. Thanks to Lemma 3.2, we can define a collection $\{\rho_k^j, \bar{\mu}_k^j, \bar{\lambda}_k^j\}_{k \geq 0}$ such that the following statements hold.

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

$$\sum_{k=I_j}^{d} \rho_k^j = 1, \qquad -\sum_{k=1}^{d} \rho_k^j \bar{\mu}_k^j e_k + \sum_{k=1}^{d} \bar{\lambda}_k^j e_k = \beta_j, \tag{3.6}$$

$$\sum_{k=1}^{d} \rho_k^j \mu_k \ell\left(\frac{\bar{\mu}_k^j}{\mu_k}\right) + \sum_{k=1}^{d} \lambda_k \ell\left(\frac{\bar{\lambda}_k^j}{\lambda_k}\right) \leq \bar{L}_{I_j}(\beta_j) + \varepsilon. \tag{3.7}$$

The alternative jump intensity $\bar{r}$ is defined as follows. For every $t \in [t_j, t_{j+1})$, let

$$\bar{r}(x, t; v) = \begin{cases} \bar{\lambda}_k^j & \text{if } v = e_k, \\ \bar{\mu}_k^j & \text{if } v = -e_k \text{ and } \Pi(x) = k \geq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.8}$$

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

$$\beta^{j,0} = \sum_{k=1}^{d} \bar{\lambda}_k^j e_k, \qquad \beta^{j,i} = -\bar{\mu}_i^j e_i + \sum_{k=1}^{d} \bar{\lambda}_k^j e_k, \quad i = 1, \dots, d. \tag{3.9}$$

It is trivial from definitions (3.8) and (3.9) that, for every $t \in [t_j, t_{j+1})$,

$$\beta^{j,\Pi(x)} = \sum_{v \in \mathbb{V}} \bar{r}(x, t; v) \cdot v. \tag{3.10}$$

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

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

Remark 3.1. The probability measures induced by $\bar{X}^n$ and X^n are absolutely continuous with respect to each other. This is because, for any given jump size v, the corresponding jump intensities $\bar{r}(x, t; v)$ and $r(x; v)$ are either both zero or strictly positive.

Step 3: weak convergence analysis of the limit process. The goal of this step is to argue that $\{\bar{X}^n\}$ converges in distribution to ϕ^*. We first show that $\{\bar{X}^n\}$ is tight and, thus, has a subsequence converging in distribution, and then identify the weak limit to be ϕ^*. The proof of tightness is standard. It is in the identification of the weak limit that the structure of the model, namely the stability-about-the-interface condition, plays a crucial role; see Remark 3.2.

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

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

Here, for every $\omega \in \Omega$, $1_A(\omega)$ is defined to be 1 if $\omega \in A$ and 0 otherwise. Each γ_k^n is a random variable taking values in the Polish space of subprobability measures on the interval $[0, 1]$, equipped with the topology of weak convergence.

Lemma 3.3. *Given any subsequence of $(\gamma^n, \bar{X}^n)$, there exists a subsubsequence and a collection of random measures $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_d)$ on $[0, 1]$ such that*

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

$$h_k(t) = \sum_{j=0}^{K-1} \rho_k^j 1_{(t_j, t_{j+1})}(t). \tag{3.11}$$

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

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

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

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

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

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

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

Note that $\bar{X}^n$ is a scaled version of Z^n. Since the generator of $\bar{X}^n$ takes the form (3.4), it is clear that the generator of Z^n, denoted by $\mathcal{L}^n$, is such that

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

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

$$\lambda_n(z,t;v) = \bar{r}\left(\frac{z}{n}, \frac{t}{n}; v\right). \tag{3.12}$$

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

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

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

$$Z^n(t) = Z^n(0) + \sum_{v\in\mathbb{V}} Y^{n,v}(t)\cdot v = nx_n + \sum_{v\in\mathbb{V}} Y^{n,v}(t)\cdot v, \tag{3.13}$$

and the instantaneous intensity function for $Y^{n,v}$ is $\lambda_n(Z^n(t), t; v)$; see also [21, Appendix B] for a more detailed discussion on counting processes.

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

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

Thanks to (3.10) and (3.12), it follows that

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

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

Combined with (3.13), we have

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

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

$$\Pr_{x_n}\Bigl(\sup_{t \in [0,1]} \| \bar{X}^n(t) - S^n(t) \| > \varepsilon \Bigr) \leq \frac{1}{\varepsilon^2} \mathrm{E}_{x_n} \| \bar{X}^n(1) - S^n(1) \|^2.$$

Thanks to (3.12) and the definition of $\bar{r}$ in (3.8), λ_n is uniformly bounded by $\|r\|_\infty$. Therefore, for some constant C [13, Theorem 2.5.3],

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

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

By Prohorov's theorem [12, Chapter 3], there exists a subsubsequence, still denoted by $(\gamma^n, \bar{X}^n)$, that converges in distribution to say $(\gamma, \bar{X})$, where $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_d)$. Note that $\bar{X}$ is continuous since it is also the weak limit of S^n. By the Skorokhod representation theorem [7, Theorem A.3.9], we can assume that the convergence is almost-sure convergence when everything is defined on some probability space, say $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{\Pr})$. Again, since $\{\gamma_k^n\}$ is absolutely continuous with respect to the Lebesgue measure on [0, 1] with the density uniformly bounded by 1, the limit γ_k also enjoys the same property. Furthermore, it follows that, for every t, $S^n(t)$ converges almost surely to

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

Therefore, $S(t) = \bar{X}(t)$ almost surely for every t. Since both S and $\bar{X}$ are continuous, $S = \bar{X}$ with probability 1. In particular, if we denote by h_k the density of γ_k,

$$\frac{\mathrm{d}\bar{X}(t)}{\mathrm{d}t} = \sum_{k=0}^{d} \beta^{I(t),k} h_k(t) \tag{3.14}$$

for almost every t.

It remains to show (3.11) and that $\bar{X} = \phi^*$. In doing so, we first establish a useful property of $\{h_k\}$, namely, that, with probability 1,

$$\sum_{k=0}^{d} h_k(t) = 1 = \sum_{k=\Pi(\bar{X}(t))}^{d} h_k(t) \quad \text{for almost every } t \in [0, 1]. \tag{3.15}$$

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

The first equality is trivial since $\sum_{k=0}^{d}\gamma_k^n$ equals the Lebesgue measure on [0, 1] for every n. The second equality follows from a standard argument [7, Theorem 7.4.4(c)]. Note that, for almost every $\omega\in\Omega$, $\bar{X}(t,\omega)$ is continuous with respect to t, $\gamma^n(\omega)$ converges weakly to $\gamma(\omega)$, and $\bar{X}^n(\cdot,\omega)$ converges to $\bar{X}(\cdot,\omega)$ in the Skorokhod metric. Arbitrarily fix such an ω. Define $A_i=\{t\in[0,1]\colon \Pi(\bar{X}(t,\omega))=i\}$. Since $\bar{X}^n(\cdot,\omega)$ also converges to $\bar{X}(\cdot,\omega)$ in the supremum norm [21, Theorem A.6.5] and Π is lower semicontinuous, it follows that, for every $t\in A_i$, there exists an open interval (a_t,b_t) containing t and an $N\in\mathbb{N}$ such that $\Pi(\bar{X}^n(s,\omega))\geq i$ for all $s\in(a_t,b_t)$ and $n\geq N$. Therefore, $\sum_{k<i}\gamma_k^n(\omega)\{(a_t,b_t)\}=0$ for all $n\geq N$. Letting $n\to\infty$, it follows that $\sum_{k<i}\gamma_k(\omega)\{(a_t,b_t)\}=0$ for every $t\in A_i$. Since $A_i\subseteq\bigcup_{t\in A_i}(a_t,b_t)$, there exists a countable subcover [19, Lindelöf Theorem, p. 49], that is, there exists $\{t_j\}\subseteq A_i$ such that

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

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

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

This completes the proof (3.15). Combining (3.9), (3.14), and (3.15), we obtain the identity

$$\frac{\mathrm{d}\bar{X}(t)}{\mathrm{d}t}=\sum_{k=0}^{d}\beta^{j,k}h_k(t)=\sum_{k=1}^{d}\bar{\lambda}_k^j e_k-\sum_{k=\max\{\Pi(\bar{X}(t)),1\}}^{d}\bar{\mu}_k^j h_k(t)e_k \tag{3.16}$$

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

We will now use induction to argue (3.11). It is trivial that (3.11) holds for almost every $t\in[0,t_j]$ with $j=0$ since $t_0=0$. Assume that (3.11) holds for almost every $t\in[0,t_j]$. The goal is to show that it holds for almost every $t\in[0,t_{j+1}]$, or, equivalently, $h_k(t)=\rho_k^j$ for almost every $t\in(t_j,t_{j+1})$.

It is not difficult to verify that $\bar{X}(t)=\phi^*(t)$ for all $t\in[0,t_j]$. Indeed, by the induction hypothesis that (3.11) holds for almost every $t\in[0,t_j]$, and (3.6) and (3.9), we have

$$\begin{aligned}\frac{\mathrm{d}\bar{X}(t)}{\mathrm{d}t}&=\sum_{k=0}^{d}\beta^{I(t),k}\sum_{j=0}^{K-1}\rho_k^j 1_{(t_j,t_{j+1})}(t)\\&=\sum_{j=0}^{K-1}\sum_{k=0}^{d}\rho_k^j\beta^{j,k}1_{(t_j,t_{j+1})}(t)\\&=\sum_{j=0}^{K-1}\beta_j 1_{(t_j,t_{j+1})}(t)\\&=\frac{\mathrm{d}\phi^*(t)}{\mathrm{d}t}.\end{aligned}\tag{3.17}$$

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

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

Define $I_j = \Pi(\phi^*(t))$ and $\beta_j = \dot{\phi}^*(t)$ for every $t \in (t_j, t_{j+1})$. Observing that $[\beta_j]_k = 0$ for all $I_j < k \leq d$, we can uniquely determine the value of $\{\rho_k^j\}$ based on the definition of $\{\rho_k^j\}$ and (3.6), namely,

$$\rho_k^j = \begin{cases} \dfrac{\bar{\lambda}_k^j}{\bar{\mu}_k^j} & \text{if } I_j < k \leq d, \\ 1 - \displaystyle\sum_{k=I_j+1}^{d} \dfrac{\bar{\lambda}_k^j}{\bar{\mu}_k^j} & \text{if } k = I_j, \\ 0 & \text{if } k < I_j. \end{cases} \tag{3.18}$$

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

The key step in this inductive argument is to prove that $\Pi(\bar{X}(t)) = I_j$ for every $t \in (t_j, t_{j+1})$. To this end, note that $\Pi(\bar{X}(t))$ can only take finitely many possible values; hence, the maximum of $\Pi(\bar{X}(t))$ on (t_j, t_{j+1}) must be attained at some $t^* \in (t_j, t_{j+1})$. Since Π is lower semicontinuous, there exists an open interval that is contained in (t_j, t_{j+1}) such that, for all t in this interval, $\Pi(\bar{X}(t)) \geq \Pi(\bar{X}(t^*))$. Denote by $(a, b) \subseteq (t_j, t_{j+1})$ the largest of such intervals. By the definition of t^*, $\Pi(\bar{X}(t)) = \Pi(\bar{X}(t^*)) = i$ (say) for every $t \in (a, b)$. It follows from (3.16) that, on the interval (a, b),

$$\frac{\mathrm{d}\bar{X}(t)}{\mathrm{d}t} = \sum_{k=1}^{d} \bar{\lambda}_k^j e_k - \sum_{k=\max\{i,1\}}^{d} \bar{\mu}_k^j h_k(t) e_k. \tag{3.19}$$

Furthermore, since clearly $[\mathrm{d}\bar{X}(t)/\mathrm{d}t]_k = 0$ for all $k > i$ and $t \in (a, b)$, we can directly compute h_k from (3.15) and (3.19) to obtain a formula analogous to (3.18):

$$h_k(t) = \begin{cases} \dfrac{\bar{\lambda}_k^j}{\bar{\mu}_k^j} & \text{if } i < k \leq d, \\ 1 - \displaystyle\sum_{k=i+1}^{d} \dfrac{\bar{\lambda}_k^j}{\bar{\mu}_k^j} & \text{if } k = i, \\ 0 & \text{if } k < i, \end{cases} \tag{3.20}$$

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

We will argue by contradiction that $i \leq I_j$. Assume otherwise, namely, $i > I_j$. Then, by comparing (3.18) and (3.20), it follows easily that $h_i(t) > \rho_i^j$ and, thus,

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

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

$$\Pi(\bar{X}(t)) \leq I_j \quad \text{for all } t \in (t_j, t_{j+1}). \tag{3.21}$$

In order to show the reverse inequality, we exclude the trivial case by assuming that $I_j \geq 1$. Note that (3.21) implies that $[\mathrm{d}\bar{X}(t)/\mathrm{d}t]_k = 0$ for all $k > I_j$. Thanks to (3.16), this is equivalent

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

to $h_k(t) = \bar{\lambda}_k^j / \bar{\mu}_k^j = \rho_k^j$ for all $k > I_j$. It follows that

$$h_{I_j}(t) \le 1 - \sum_{k=I_j+1}^{d} h_k(t) = 1 - \sum_{k=I_j+1}^{d} \rho_k^j = \rho_{I_j}^j,$$

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

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

Since $\bar{X}(t_j) = \phi^*(t_j)$, we have $[\bar{X}(t)]_{I_j} \ge [\phi^*(t)]_{I_j} > 0$, or $\Pi(\bar{X}(t)) \ge I_j$ for all $t \in (t_j, t_{j+1})$. Therefore, taking (3.21) into consideration we arrive at

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

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

The desired equality $h_k(t) = \rho_k^j$ for every $t \in (t_i, t_{j+1})$ is now trivial. Indeed, the two formulae (3.18) and (3.20) are identical when $i = \Pi(\bar{X}(t)) = I_j$. This completes the proof of (3.11).

It remains to show that $\bar{X}(t) = \phi^*(t)$ for all $t \in [0, 1]$. This can be done by repeating the steps in (3.17) for every $t \in (0, 1)$. The proof of Lemma 3.3 is now complete.

Step 4: analysis of the cost. Along the convergent subsubsequence (still denoted by $(\gamma^n, \bar{X}^n)$), Lemma 3.3 and (3.7) imply that

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

This completes the proof of (3.5) as well as the proof of Theorem 3.1.

Remark 3.2. The stability-about-the-interface condition manifests itself in the monotonicity of the value $h_k(t)$ with respect to the value of $\Pi(\bar{X}(t))$; see (3.20). Loosely speaking, this monotonicity property implies that the change of measure (control) defined by the upper bound rate function automatically pushes the trajectory back to the discontinuous interface if it ever wanders off. This guarantees the desired tracking behavior.

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

4. A case study

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

$$p_n = \Pr(\text{total population } Q_1 + Q_2 \text{ reaches } n \text{ before coming back to } 0, \text{ starting from } Q = (0, 0)).$$

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

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

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

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

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

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

4.1. Three important roots of the Hamiltonians

The quantity of interest p_n is just the probability of the scaled process X^n reaching the exit boundary $\partial_e = \{x = (x_1, x_2) : x_i \geq 0,\ x_1 + x_2 = 1\}$ before coming back to the origin, starting from the origin itself. Thanks to Theorem 3.1, it is reasonable to expect that the exponential decay rate of p_n equals the value of the calculus of variations problem

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

where the infimum is taken over all absolutely continuous functions $\phi : [0, \infty) \to \mathbb{R}^2_+$ and $\tau \geq 0$ such that $\phi(0) = 0$ and $\phi(\tau) \in \partial_e$. It is not difficult to see that an optimal trajectory ϕ^*, if it exists, should be a straight line due to the convexity of the local rate function and the homogeneity of the system dynamics. See Figure 3.

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

If the optimal trajectory travels through the interior of the state space then the dynamic programming principle implies that the value function V satisfies the Hamiltonian–Jacobi–Bellman (HJB) equation

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

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

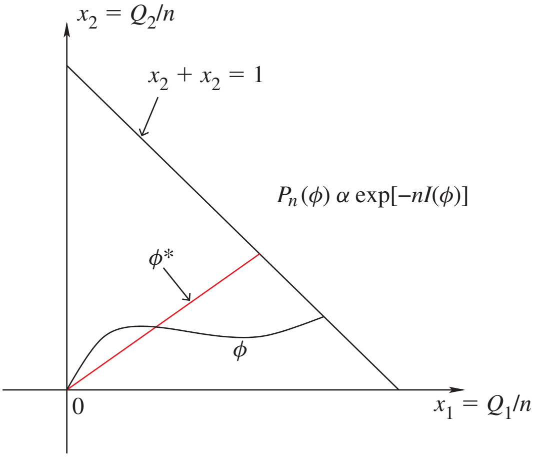

FIGURE 3: Representative limit sample path ϕ.

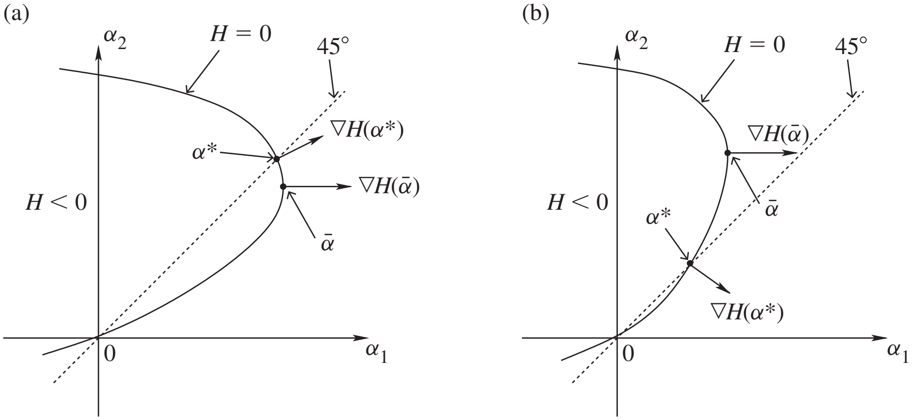

FIGURE 4: Geometry and trajectory (I).

Furthermore, the boundary condition $V(x) = 0$ for $x \in \partial_{\mathrm{e}}$ should hold. This suggests that $\nabla V(x) = -\alpha^*$, where $H(\alpha^*) = 0$, and that α^* is orthogonal to ∂_{e}, or, equivalently, $\alpha_1^* = \alpha_2^*$. In this case, the exponential decay rate of p_n is just α_1^*, and the optimal trajectory leaves the domain in a straight line with slope $\beta^* = \nabla H(\alpha^*)$ (β^* is the minimizer in the HJB equation). We wish to make an important cautionary comment, namely that the geometry of the zero-level set of H has to be taken into consideration in order for these heuristics to determine a possible optimal trajectory. For illustration, consider the following two scenarios (see Figure 4). In both cases, $\bar{\alpha}$ denotes the point on the level set $\{H = 0\}$ with the maximal first component, whence $\nabla H(\bar{\alpha}) = ae_1$ for some nonnegative constant a. In Figure 4(a) the 45° line intersects with the level set at point α^* which is above $\bar{\alpha}$. The corresponding $\beta^* = \nabla H(\alpha^*)$ has nonnegative components. Therefore, the root α^* determines a candidate optimal trajectory $\phi^*(t) = \beta^* t$ that lives in the nonnegative orthant and hits ∂_{e} in finite time. In contrast, in Figure 4(b) the

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

45° line intersects with the level set at point α^* which is below $\bar{\alpha}$ and $\beta^* = \nabla H(\alpha^*)$ has a negative second component. It is clear that this root α^* does *not* associate with any physically meaningful trajectory since $\phi^*(t) = \beta^* t$ will not live in the nonnegative orthant.

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

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

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

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

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

Again, when this heuristic is used to determine a possible optimal trajectory, the geometry of the zero-level sets of H and H_∂ has to be incorporated. For illustration, consider the following two scenarios (see Figure 5). As before, $\bar{\alpha}$ denotes the point on the level set $\{H = 0\}$ with the maximal first component. In Figure 5(a), the intersection of the two zero-level sets, $\hat{\alpha}$, is below $\bar{\alpha}$. The corresponding β^* does determine a possible optimal trajectory $\phi^*(t) = \beta^* t$,

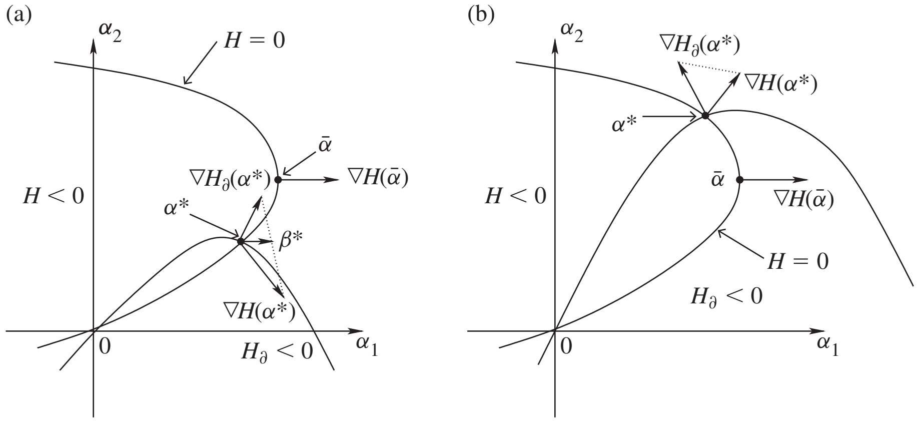

FIGURE 5: Geometry and trajectory (II).

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

which 'pushes into' the boundary ∂. In Figure 5(b), however, $\hat{\alpha}$ is above $\bar{\alpha}$. In this case, since both $\nabla H(\hat{\alpha})$ and $\nabla H_\partial(\hat{\alpha})$ have positive second components, none of their convex combinations will yield a horizontal velocity β^*. Therefore, this root $\hat{\alpha}$ does *not* represent any meaningful trajectory traveling along the boundary ∂. Indeed, the root that will determine such a trajectory is $\bar{\alpha}$. It corresponds to a trajectory that 'glides' along the boundary ∂ with velocity $\beta^* = \nabla H(\bar{\alpha})$, a horizontal vector.

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

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

and define the constant

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

Lemma 4.1. *The constant z satisfies* $\max\{0, 1 - 1/\theta_3\} < z < 1$. *Define vectors α^*, $\bar{\alpha}$, and $\hat{\alpha}$ to be*

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

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

4.2. The exponential decay rate of p_n

It is now intuitively clear what the exponential decay rate of p_n should be. For example, if $\alpha_2^* > \bar{\alpha}_2$ and $\hat{\alpha}_2 < \bar{\alpha}_2$, then it corresponds to Figure 3(a) and Figure 4(a). Therefore, α^* determines a trajectory leaving the domain through the interior with cost α_1^*, whereas $\hat{\alpha}$ determines a trajectory leaving the domain by 'pushing into' the boundary ∂ with cost $\hat{\alpha}_1$. The optimal trajectory should be the one with a smaller cost and the minimal cost is $\min(\alpha_1^*, \hat{\alpha}_1^*)$. More generally, we have Theorem 4.1, which can be shown by constructing suitable subsolutions and invoking Lemma 4.2 below.

Lemma 4.2. *Suppose that $W : \mathbb{R}_+^2 \to \mathbb{R}$ is a twice continuously differentiable function satisfying*

$$\begin{aligned} -H(-\nabla W(x)) &\ge 0 \quad \textit{for } x = (x_1, x_2) \in \mathbb{R}_+^2 \textit{ such that } x_2 > 0, \\ -H_\partial(-\nabla W(x)) &\ge 0 \quad \textit{for } x \in \partial, \\ W(x) &\le 0 \quad \textit{for } x \in \partial_{\mathrm{e}}. \end{aligned}$$

Then

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

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

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

Theorem 4.1. *The exponential decay rate of p_n is*

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

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

$$\liminf_n -\frac{1}{n}\log p_n \geq \gamma. \tag{4.1}$$

Thanks to Lemma 4.2, it suffices to construct a sequence of subsolutions whose values at the origin approach γ. To this end, we define two vectors

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

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

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

We claim that $v_2^* > \hat{v}_2$. Indeed, letting $\nu = (0, \bar{\alpha}_2)$, where, by Lemma 4.1, $\bar{\alpha}_2 = -\log\sqrt{\theta_2}$, a straightforward calculation yields

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

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

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

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

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

The above inequality is impossible since $H(\hat{\alpha}) = 0$. Observing that $\bar{\alpha}_2 < \alpha_1^*$ by assumption, and $\bar{\alpha}_1 > \alpha_1^*$ by Lemma 4.1, we have

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

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

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

where $b = (v_2^* - \hat{v}_2)^{-1} > 0$. Let $W^{\varepsilon,\delta}$ be the classical mollification of W^δ [16, Section 7.2],

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

namely,

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

where ρ is a smooth symmetric kernel defined by

$$\rho(y) = \begin{cases} c \exp\left[\dfrac{1}{\|y\|^2 - 1}\right] & \text{if } \|y\| \le 1, \\ 0 & \text{if } \|y\| \ge 1, \end{cases} \qquad \int_{\mathbb{R}^2} \rho(y)\,\mathrm{d}y = 1.$$

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

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

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

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

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

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

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

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

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

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

Applying Lemma 4.2, we arrive at

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

for all $\delta > 0$. Letting δ tend to 0, we complete the proof of the upper bound (4.1).

It remains to show the lower bound

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

We first observe that the sample path large deviation principle (i.e. Theorem 3.1) implies that

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

where the infimum is taken over all absolutely continuous sample paths $\phi\colon [0,\infty) \to \mathbb{R}^2_+$ such that $\phi(0) = 0$, $\phi(\tau) \in \partial_{\mathrm{e}}$. The proof of this inequality is standard and almost verbatim

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

to that of Equation (8.5) of [11]—the only major difference is that '$L^A(\mathbf{1})$ is finite for any $A \subset \{1, 2, \dots, d\}$' should be replaced by '$\bar{L}_i(\mathbf{1})$ is finite for any $i = 0, 1, \dots, d$'.

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

$$\int_0^{\tau^*} L(\phi^*(t), \dot{\phi}^*(t))\,\mathrm{d}t \le \gamma.$$

We consider the following two cases.

Case 1: $\alpha_1^* \le \hat{\alpha}_1$. Define $\beta^* = \nabla H(\alpha^*) = (\lambda_1 \mathrm{e}^{\alpha_1^*}, \lambda_2 \mathrm{e}^{\alpha_2^*} - \mu_2 \mathrm{e}^{-\alpha_2^*})$. That is, β^* is the conjugate of α^* through the convex duality of H and L. Clearly, $\beta_1^* > 0$. Since $\alpha_2^* = \alpha_1^* > \bar{\alpha}_2 = -\log\sqrt{\theta_2}$ (Lemma 4.1), it follows that

$$\beta_2^* > \lambda_2 \mathrm{e}^{\bar{\alpha}_2} - \mu_2 \mathrm{e}^{-\bar{\alpha}_2} = \sqrt{\lambda_2\mu_2} - \sqrt{\lambda_2\mu_2} = 0. \tag{4.2}$$

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

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

Therefore,

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

Since $\alpha_1^* = \alpha_2^*$ and $\beta^*\tau^* \in \partial_\mathrm{e}$, we have $\langle \alpha^*, \beta^*\tau^* \rangle = \alpha_1^* = \gamma$.

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

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

Since $\hat{\alpha}_2 < \bar{\alpha}_2$, it follows that $\bar{\beta}_2 < 0$ by an argument analogous to (4.2). Define

$$\rho_1 = \hat{\beta}_2(\hat{\beta}_2 - \bar{\beta}_2)^{-1} \quad \text{and} \quad \rho_2 = -\bar{\beta}_2(\hat{\beta}_2 - \bar{\beta}_2)^{-1}.$$

Then ρ_1 and ρ_2 are both nonnegative, $\rho_1 + \rho_2 = 1$, and

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

We claim that $\beta_1^* > 0$. Indeed, since H and L are both strictly convex and $L(\beta) = 0$ if and only if $\beta = \nabla H(0)$, it follows from the conjugacy of $\bar{\beta}$ and $\hat{\alpha}$ that

$$\langle \hat{\alpha}, \bar{\beta} \rangle = \langle \hat{\alpha}, \bar{\beta} \rangle - H(\hat{\alpha}) = L(\bar{\beta}) > 0. \tag{4.3}$$

Similarly,

$$\langle \hat{\alpha}, \hat{\beta} \rangle = \langle \hat{\alpha}, \hat{\beta} \rangle - H_\partial(\hat{\alpha}) = L_\partial(\bar{\beta}) > 0. \tag{4.4}$$

Therefore,

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

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

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

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

However, by the definition of inf-convolution, (4.3), and (4.4),

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

It follows that

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

This completes the proof.

Remark 4.1. The proof of Theorem 4.1 actually shows that the decay rate γ equals the value of the calculus of variation problem

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

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

5. Summary

In this paper we used a weak convergence approach to establish the sample path large deviation principle for a single-server system with preemptive priority service policy. The difficulty in the analysis is due to the discontinuity of the system dynamics. We showed that the general upper bound rate function [8] is indeed tight since the stability-about-the-interface condition is automatically built into the upper bound rate function. This simple form of the rate function proves to be useful when studying the asymptotic behavior of various buffer overflow probabilities. For illustration, in the two-dimensional case the exponential decay rate of the total population overflow probabilities was explicitly identified. This was done by studying the geometry of the zero-level sets of the system Hamiltonians and by constructing appropriate subsolutions to the related partial differential equation.

Appendix A. Proof of Lemma 3.1

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

Lemma A.1. *Given an arbitrary interval* $[\![a, b]\!]$ *and any* $\sigma > 0$, *there exists a finite partition*

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

such that, for each j,

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

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

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

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

It suffices to show that there exists a finite collection of closed intervals $\{[\bar a_i, \bar b_i]\}$ with nonoverlapping interiors such that $0 \le \bar b_i - \bar a_i \le \sigma$, $\Pi(\phi(\bar a_i)) = \Pi(\phi(\bar b_i)) = k^* = k$, and

$$\min\biggl\{\Pi(\phi(t)) \colon t \in [a,b] \biggm\backslash \bigcup_i [\bar a_i, \bar b_i]\biggr\} \ge k+1.$$

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

The values of $\bar a_i$ and $\bar b_i$ are defined recursively as follows. Let

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

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

$$\begin{aligned}\bar a_1 &= \bar a,\\ \bar b_1 &= \sup[t \in [\bar a_1, (\bar a_1 + \sigma) \wedge b] \colon \Pi(\phi(t)) = k],\end{aligned}$$

and, for $i \ge 1$,

$$\begin{aligned}\bar a_{i+1} &= \inf[t \in [\bar a_i + \sigma, b] \colon \Pi(\phi(t)) = k],\\ \bar b_{i+1} &= \sup[t \in [\bar a_{i+1}, (\bar a_{i+1} + \sigma) \wedge b] \colon \Pi(\phi(t)) = k].\end{aligned}$$

The recursion will end if $\bar b_N = \bar b$ for some N. It is clear that N is finite since $\bar a_{i+1} - \bar a_i \ge \sigma$. Furthermore, the collection $\{[\bar a_i, \bar b_i] \colon i = 1,2,\dots,N\}$ clearly has the desired property. This completes the proof.

Since $I_x(\phi) < \infty$, ϕ is absolutely continuous and, hence, uniformly continuous on $[0,1]$. Therefore, there exists $\sigma > 0$ such that, for $s,t \in [0,1]$,

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

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

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

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

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

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

Observing that the rate functions $\{\bar{L}_i\}$ are monotonically nondecreasing in that $\bar{L}_0 \le \bar{L}_1 \le \cdots \le \bar{L}_d$, we have

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

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

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

This completes the proof.

Appendix B. Proof of Lemma 3.2

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

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

for every $\beta \in \mathbb{R}^d$. It is now an immediate consequence of [7, Corollary D.4.2] that L_i, the Legendre transform of H_i, has the following alternative representation. That is, for every $\beta \in \mathbb{R}^d$,

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

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

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

We are now in a position to prove the alternative representation for $\bar{L}_i$. With loss of generality, we assume that $i = 0$. The proof for $i \ge 1$ is similar and thus omitted. Thanks to the definition of $\bar{L}_i$, (3.1), and (B.1)–(B.2), we have

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

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

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

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

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

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

Then the constraints (B.3) become

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

which are exactly the constraints in the statement of Lemma 3.2. Observe that, by the convexity of ℓ,

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

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

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

Appendix C. Proof of Lemma 4.2

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

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

$$\mathrm{E}\biggl\{\exp\biggl[-n\biggl[W\biggl(\frac{Z(k+1)}{n}\biggr) - W\biggl(\frac{Z(k)}{n}\biggr)\biggr]\biggr] \biggm| Z(k) = z, Z(k-1), \dots, Z(0)\biggr\} \leq \mathrm{e}^{M/n} \tag{C.1}$$

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

$$\lambda_1 \exp\biggl[-n\biggl[W\biggl(x + \frac{e_1}{n}\biggr) - W(x)\biggr]\biggr] + \lambda_2 \exp\biggl[-n\biggl[W\biggl(x + \frac{e_2}{n}\biggr) - W(x)\biggr]\biggr]$$
$$+ \mu_2 \exp\biggl[-n\biggl[W\biggl(x - \frac{e_2}{n}\biggr) - W(x)\biggr]\biggr].$$

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

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

for every vector v and some constant M. Therefore, the conditional expectation is bounded from above by

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

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

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

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

Define a nonnegative process

$$Y^n(k) = \exp\biggl[-\frac{Mk}{n} - nW\biggl(\frac{Z(k)}{n}\biggr)\biggr], \qquad k = 0, 1, 2, \dots.$$

It follows from inequality (C.1) that the stopped process $\{Y^n(k \wedge T_n)\}$ is a supermartingale with respect to the natural filtration generated by Z. Let T_0 be the return time to the origin:

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

Owing to the optional sampling theorem and the nonnegativity of Y^n, we have

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

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

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

and, thus,

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

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

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

Taking the logarithm on both sides, it follows easily that

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

This completes the proof.

References

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

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

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

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press

[4] Borovkov, A. A. and Mogul'skiĭ, A. A. (2001). Large deviations for Markov chains in the positive quadrant. *Russian Math. Surveys* **56,** 803–916.

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

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

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

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

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

[10] Dupuis, P., Leder, K. and Wang, H. (2008). On the large deviations properties of the weighted-serve-the-longest-queue policy. In *In and Out of Equilibrium 2*, eds V. Sidoravicius and M. E. Vares, Birkhäuser, Basel, pp. 229–256.

[11] Dupuis, P., Leder, K. and Wang, H. (2009). Importance sampling for weighted-serve-the-longest-queue. *Math. Operat. Res.* **34,** 642–660.

[12] Ethier, S. N. and Kurtz, T. G. (1986). *Markov Processes.* John Wiley, New York.

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

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

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

[16] Gilbarg, D. and Trudinger, N. S. (1983). *Elliptic Partial Differential Equations of Second Order*, 2nd edn. Springer, Berlin.

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

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

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

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

[21] Shwartz, A. and Weiss, A. (1995). *Large Deviations for Performance Analysis*. Chapman and Hall, London.

[22] Stolyar, A. L. and Ramanan, K. (2001). Largest weighted delay first scheduling: large deviations and optimality. *Ann. Appl. Prob.* **11,** 1–48.

https://doi.org/10.1239/aap/1308662492 Published online by Cambridge University Press