Electron. Commun. Probab. **21** (2016), no. 2, 1–10.
DOI: 10.1214/16-ECP4370
ISSN: 1083-589X

ELECTRONIC COMMUNICATIONS in PROBABILITY

A note on non-existence of diffusion limits for serve-the-longest-queue when the buffers are equal in size*

Rami Atar† Subhamay Saha‡

Abstract

We consider the serve-the-longest-queue discipline for a multiclass queue with buffers of equal size, operating under (i) the conventional and (ii) the Halfin-Whitt heavy traffic regimes, and show that while the queue length process' scaling limits are fully determined by the first and second order data in case (i), they depend on finer properties in case (ii). The proof of the latter relies on the construction of a *deterministic* arrival pattern.

Keywords: diffusion limits; Halfin-Whitt regime; serve-the-longest queue; heavy-traffic.
AMS MSC 2010: 60F17; 60J60; 60K25.
Submitted to ECP on June 16, 2015, final version accepted on January 4, 2016.

1 Introduction

We analyze the multi-class queue in two different diffusion regimes, namely the *conventional* and the *Halfin-Whitt* (HW) heavy traffic regimes, operating under the serve-the-longest-queue (SLQ) scheduling policy. In both regimes the traffic intensity is asymptotic to unity, where in conventional heavy traffic, the model is based on a single server and the arrival rate and service time distributions are scaled up, while in the HW regime, the arrival rate and number of servers are scaled up and the service time distributions are kept fixed; see [2] and references therein for more on these regimes. Our goal is to demonstrate that if the buffers are finite and of equal size, then, perhaps counterintuitively, the first and second order data of the underlying primitive processes do not uniquely determine the queue length asymptotics in the HW regime (the term 'first and second order data of the underlying primitive processes' informally means their Law of Large Numbers and Central Limit Theorem limit laws; it is rigorously defined in Section 2). As a result, a diffusion limit does not always exist under the 'usual' set of assumptions. This stands in contrast to the conventional regime where, as we show, the limit is fully determined by the first and second order data.

Our motivation to study systems with finite buffers stems from a recent treatment [3], where they arise in a game-theoretic setting of customers that act strategically, and avoid joining the queue if they expect that the delay will exceed a threshold. In that setting, determining the diffusion-scale asymptotics of the queue length provides a

*Research supported in part by the ISF (grant 1315/12)
†Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel. E-mail: atar@ee.technion.ac.il
‡Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel. E-mail: subhamay585@gmail.com

crucial step in the analysis of a Nash equilibrium. The usual role played by finite buffers, namely to model finite storage room, provides, of course, an additional motivation.

In Sections 2 and 3 we treat the HW and the conventional regimes, respectively, where in the former we provide a counterexample to existence of limits, and in the latter we determine the limit. The aforementioned counterexample is based on the construction of a certain deterministic arrival pattern; the problem of whether existence of limits fails under a more common model for arrivals, such as renewal processes with no fixed times of discontinuity, is left open (see Problem 2.2).

We use the following notation. For $a, b \in \mathbb{R}$, the maximum [resp., minimum] is denoted by $a \vee b$ [resp., $a \wedge b$], and $a^+ = a \vee 0$, $a^- = (-a) \vee 0$. For $x, y \in \mathbb{R}^k$ (k a positive integer), $x \cdot y$ and $\|x\|$ denote the usual scalar product and ℓ_2 norm, respectively. Write $\{\mathbf{e}_i\}$, $i = 1, \ldots, k$ for the standard basis in $\mathbb{R}^k$ and 1 for $\sum_{i=1}^k \mathbf{e}_i$. Denote $\mathbb{R}_+ = [0, \infty)$, and let $\iota : \mathbb{R}_+ \to \mathbb{R}_+$ the identity. For $f : \mathbb{R}_+ \to \mathbb{R}^k$, $\|f\|_T = \sup_{t \in [0,T]} \|f(t)\|$, and, for $\theta > 0$, $w_T(f, \theta) = \sup_{0 \le s < u \le s+\theta \le T} \|f_u - f_s\|$. For a Polish space $\mathcal{S}$, let $\mathbb{C}_{\mathcal{S}}([0,T])$ and $\mathbb{D}_{\mathcal{S}}([0,T])$ denote the set of continuous and, respectively, cadlag functions $[0,T] \to \mathcal{S}$. Write $\mathbb{C}_{\mathcal{S}}$ and $\mathbb{D}_{\mathcal{S}}$ for the case where $[0,T]$ is replaced by $\mathbb{R}_+$. Endow $\mathbb{D}_{\mathcal{S}}$ with the Skorohod J_1 topology. Write $X_n \Rightarrow X$ for convergence in distribution. A sequence of processes X_n with sample paths in $\mathbb{D}_{\mathcal{S}}$ is said to be *C-tight* if it is tight and every subsequential limit has, with probability 1, sample paths in $\mathbb{C}_{\mathcal{S}}$. For a positive integer k, $m \in \mathbb{R}^k$ and a symmetric, positive matrix $A \in \mathbb{R}^{k \times k}$, *an (m, A)-Brownian motion* (BM) is a k-dimensional BM starting from zero, having drift m and infinitesimal covariance matrix A.

2 A counterexample to existence of limits in the Halfin-Whitt regime

A sequence of queueing models, indexed by $n \in \mathbb{N}$, and defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, has n identical servers and a fixed number, $N \ge 2$, of buffers dedicated to customers of N classes. For $i = 1, 2, \ldots, N$, class-i customers arrive according to an arrival process E_i^n and upon arrival go directly for service on the event that any of the servers is available, and otherwise are queued in buffer i if the buffer is not fully occupied. Arrivals are lost when the corresponding buffer is full. When a server becomes available and the buffers are non-empty, it picks a customer from the buffer with most customers, and, in case of equal maximal queue lengths, a fair N-coin is tossed to determine which buffer to pick from. Class-i jobs take exponential time to process, with parameter μ_i^n, where

$$\mu_i^n = \mu_i + n^{-1/2}\hat{\mu}_i + o(n^{-1/2}), \tag{2.1}$$

and $\mu_i > 0$ and $\hat{\mu}_i \in \mathbb{R}$ are constants. The arrival counting processes, E_i^n, are assumed to satisfy the Law of Large Numbers,

$$\bar{E}_i^n := n^{-1} E_i^n \Rightarrow \lambda_i \iota, \tag{2.2}$$

where $\lambda_i > 0$ are constants, and the Central Limit Theorem,

$$\hat{E}_i^n := n^{-1/2}(E_i^n - \lambda_i^n \iota) \Rightarrow W_i^{\text{arr}}, \tag{2.3}$$

where $\lambda_i^n = \lambda_i n + n^{1/2}\hat{\lambda}_i + o(n^{1/2})$, and W_i^{arr} is a $(0, \lambda_i \sigma_i^2)$-BM, for constants $\hat{\lambda}_i \in \mathbb{R}$, $\sigma_i^2 \ge 0$. It is also assumed that arrival processes are independent. The resulting asymptotic traffic intensity is given by $\sum_i \rho_i$, where $\rho_i = \lambda_i/\mu_i$, assumed to satisfy the critical load condition $\sum_i \rho_i = 1$. The queue length processes are denoted by $Q^n = (Q_1^n, \ldots, Q_N^n)$. The number of class-$i$ customers in the system (resp., in the buffer, in service) at time

t is denoted by $X_i^n(t)$ (resp., $Q_i^n(t)$, $\Psi_i^n(t)$). Note that $X^n = Q^n + \Psi^n$, and $1 \cdot \Psi^n \leq n$. Diffusion scaled versions of these processes are denoted by

$$\hat{X}^n = n^{-1/2}(X^n - n\rho), \qquad \hat{Q}^n = n^{-1/2}Q^n, \qquad \hat{\Psi}^n = n^{-1/2}(\Psi^n - n\rho).$$

It is assumed that the initial condition $\hat{X}^n(0)$ satisfies

$$\hat{X}^n(0) \Rightarrow X_0,$$

where X_0 is an $\mathbb{R}^N$-valued r.v., whose distribution is denoted by m_0, and, for simplicity, the queue lengths are assumed to start at zero, that is, $Q^n(0) = (Q_1^n(0), \ldots, Q_N^n(0)) = 0$. We will assume that the buffer sizes, denoted throughout by $\{\beta_i^n\}$, are asymptotic to $\{\beta_i n^{1/2}\}$, where $\beta_i > 0$ are constants, namely $\beta_i^n = \beta_i n^{1/2} + o(n^{1/2})$.

The tuples (μ_i, λ_i) and $(\hat{\mu}_i, \hat{\lambda}_i, \sigma_i^2, m_0)$ are often referred to as *first* and *second* order data, respectively. We denote them jointly by

$$\boldsymbol{\delta} = (\mu_i, \lambda_i, \hat{\mu}_i, \hat{\lambda}_i, \sigma_i^2, m_0).$$

Given $k \in \{1, \ldots, N\}$, consider a stochastic differential equation (SDE) with reflection, for a process X that lives in

$$G_k = \{x \in \mathbb{R}^N : 1 \cdot x \leq N\beta_k\},$$

and reflects on the boundary of G_k in the direction $-\mathbf{e}_k$. Let $\{W(t)\}$ be a $(\hat{\lambda}, A)$-BM, where $A = \mathrm{diag}(\lambda_i(\sigma_i^2 + 1))$. Let $b : \mathbb{R}^N \to \mathbb{R}^N$ be given by

$$b(x) = -(\mu_1(x_1 - N^{-1}(1 \cdot x)^+), \ldots, \mu_N(x_N - N^{-1}(1 \cdot x)^+)). \tag{2.4}$$

Let $(X, L) = (X^{(k)}, L^{(k)})$ be the unique pair of processes that is adapted to the filtration $\sigma\{X_0\} \vee \sigma\{W(u), u \leq t\}$, where X has sample paths in $\mathbb{C}(\mathbb{R}_+ : G_k)$, L has nondecreasing sample paths in $\mathbb{C}(\mathbb{R}_+ : \mathbb{R}_+)$, and the pair satisfies a.s.,

$$\begin{aligned} &X(t) = X_0 + W(t) + \int_0^t b(X(u))du - L(t)\mathbf{e}_k, \qquad t \geq 0, \\ &\int_{[0,\infty)} 1_{\{1 \cdot X(t) < N\beta_k\}} dL(t) = 0\,. \end{aligned} \tag{2.5}$$

The existence and uniqueness of such a pair follows from Proposition 3 of [1] on noting that b is Lipschitz continuous. We denote by $X^{(k)}$ the solution to the SDE (2.5).

It follows from the results of [3] that the limits of $(\hat{X}^n, \hat{Q}^n, \hat{\Psi}^n)$ are not uniquely determined by $\boldsymbol{\delta}$ when the buffer sizes are asymptotically equal, i.e., $\beta_i = \beta_1$ for all i. More precisely, the following result appears in [3] (Proposition 4.3):

Assume that for some k and all $i \neq k$, $\beta_k < \beta_i$. Then $(\hat{X}^n, \hat{Q}^n, \hat{\Psi}^n) \Rightarrow (X, Q, \Psi)$, where $X = X^{(k)}$ is the unique solution of (2.5)*, and Q and Ψ are recovered from it via $Q = N^{-1}(1 \cdot X)^+$ and $\Psi = X - Q$.*

One can draw from this result the following conclusions regarding the case $\beta_i = \beta_1$ for all $i \in \{1, \ldots, N\}$:

(i) For every k, one can choose $\{\beta_i^n\}$ asymptotic to $\{\beta_i n^{1/2}\}$, in such a way that $\hat{X}^n \Rightarrow X$, where $X = X^{(k)}$. *Thus the first and second order data do not determine the limits.*

(ii) One can choose $\{\beta_i^n\}$ asymptotic to $\{\beta_i n^{1/2}\}$ in such a way that $\hat{X}^n$ do not converge in distribution. *Thus limits need not exist.*

Indeed, (i) follows because, given k, we have $\hat{X}^n \Rightarrow X^{(k)}$ when $\beta_k = c - \varepsilon$ and $\beta_i = c$ for all $i \neq k$, with $c, \varepsilon > 0$ fixed; hence by a diagonal argument, the same is true with ε replaced by $\varepsilon_n > 0$, for some $\varepsilon_n \to 0$. Of course, (ii) is immediate from (i).

In the present note, we are interested in the case where β_i^n are *exactly equal to each other*, for every n. Assuming in what follows that for a constant $\beta_1 > 0$,

$$\beta_i^n = \beta^n := \lfloor \beta_1 n^{1/2} \rfloor, \qquad i \in \{1, \dots, N\},\ n \in \mathbb{N}, \tag{2.6}$$

we ask whether, in this situation, the first and second order data still fall short of determining the limit behavior. More precisely, we aim at addressing the following assertions:

(i') For every k one can choose $\{E_i^n\}$ that satisfy (2.2) and (2.3), in such a way that $\hat{X}^n \Rightarrow X$, where $X = X^{(k)}$.

(ii') One can choose $\{E_i^n\}$ that satisfy (2.2) and (2.3), in such a way that $\hat{X}^n$ do not converge.

An affirmative answer will confirm that the first and second order data do not determine the limits even when the buffers are exactly equal in size. We address these questions in the special case where $N = 2$, but it will be clear from the proof that analogous treatment is possible in general.

Theorem 2.1. *Consider $N = 2$ and assume that the buffer sizes are given by* (2.6)*. Fix $k \in \{1, 2\}$. Then one can find $\{E_1^n\}$ and $\{E_2^n\}$ satisfying* (2.2) *and* (2.3)*, so that $\hat{X}^n \Rightarrow X^{(k)}$, the solution of the SDE* (2.5)*.*

Note that the domain G_k does not depend on k in this case, since $\beta_1 = \beta_2$. However, the SDEs still differ in terms in the direction of reflection, and in this situation the solutions $X^{(1)}$ and $X^{(2)}$ are not equal in law. Hence the validity of (i') and (ii') is an immediate consequence of the above result.

As mentioned earlier, the proof of the result will be based on the construction of a deterministic arrival pattern. It is natural to ask whether the result remains valid under the additional requirement that the arrivals follow a more common model, such as renewals. More precisely, we formulate the following problem, that we leave open.

Problem 2.2. Determine whether existence of limits may fail when the arrivals are given by accelerated versions of independent renewal processes (namely, $E_i^n(t) = E_i(\mu_i^n t)$, $t \geq 0$) with inter-renewal distributions that have density.

Proof of Theorem 2.1: The construction will be with the parameters $\lambda_i = 1$, $\hat{\lambda}_i = 0$, $\mu_i = 2$, $\hat{\mu}_i = 0$, $\rho_i = 1/2$. The arrival processes we construct are deterministic, and satisfy (2.2), as well as (2.3) with $\sigma_i = 0$. In particular, the driving BM in (2.5) is a $(0, A)$-BM where $A = \mathrm{diag}(\lambda_i) = \mathrm{diag}(1, 1)$. The construction is presented for $k = 1$; the case $k = 2$ is obtained by interchanging the roles of class 1 and class 2.

Fix a sequence $m_n = \lfloor n^a \rfloor$, $n \in \mathbb{N}$, where $a \in (0, \frac{1}{2})$ is constant. For ease of notation we suppress the index n in m_n and β^n (of (2.6)) and write m and β, respectively.

First, we construct E_i^n on the interval $[0, \tau]$, where $\tau = \frac{m}{n}$, by letting

$$E_1^n(t) = \begin{cases} 0, & t \in [0, \tau), \\ m, & t = \tau, \end{cases}$$

$$E_2^n(t) = \begin{cases} 0, & t \in [0, \frac{\tau}{2}), \\ \lfloor 2n(t - \frac{\tau}{2}) + 2 \rfloor, & t \in [\frac{\tau}{2}, \tau), \\ m, & t = \tau. \end{cases} \tag{2.7}$$

Thus, for each class, m arrivals occur during $[0,\tau]$, where class-1 customers all arrive at time τ, whereas class-2 arrivals are at $\frac{\tau}{2}, \frac{\tau}{2}+\frac{1}{2n}, \frac{\tau}{2}+\frac{2}{2n}, \ldots, \frac{\tau}{2}+\frac{m-1}{2n}$. Beyond $[0,\tau]$, the pattern defined on $(0,\tau]$ repeats itself with period τ. Namely, E^n is given by

$$E^n(t+j\tau) = E^n(j\tau) + E^n(t), \qquad t \in (0,\tau],\, j \in \mathbb{N}.$$

Note that for both $i=1,2$, (2.2) holds with $\lambda_i = 1$, and (2.3) holds with $\lambda_i^n = n$ and $\sigma_i^2 = 0$ (thus $W_i^{\mathrm{arr}} = 0$ a.s.). The parameters μ_i^n are given by $\mu_i^n = 2$ for $n \in \mathbb{N}$, $i = 1,2$.

We need some additional notation. Denote by R_i^n the counting process for class-i losses since time 0, by B_i^n the counting process for class-i customers sent to the service pool since time 0, by D_i^n the counting process for class-i departures from service, and by S a unit-rate Poisson process representing potential service. Namely,

$$1 \cdot D^n(t) = S\Big(2\int_0^t 1 \cdot \Psi^n(u)du\Big)\,. \tag{2.8}$$

We have the following balance equations

$$Q_i^n(t) = Q_i^n(0) + E_i^n(t) - B_i^n(t) - R_i^n(t)\,, \tag{2.9}$$

$$\Psi_i^n(t) = \Psi_i^n(0) + B_i^n(t) - D_i^n(t)\,. \tag{2.10}$$

Denote $\hat{R}^n = n^{-1/2}R^n$. The main estimate will be to show that $\hat{R}_2^n \Rightarrow 0$. Fix T and note that

$$\mathbb{E}R_2^n(T) \le \sum_{j=0}^{\lfloor T/\tau\rfloor} \mathbb{E}[R_2^n(j\tau+\tau) - R_2^n(j\tau)] \le \sum_{j=0}^{\lfloor T/\tau\rfloor} m\mathbb{P}[\sup_{t\in[j\tau+\frac{\tau}{2}, j\tau+\tau)} Q_2^n(t) = \beta], \tag{2.11}$$

where we used the fact that a class-2 loss can only occur if a customer arrives when the buffer is full (that is, $Q_2^n = \beta$), that class-2 arrivals occur only within $[j\tau+\frac{\tau}{2}, j\tau+\tau)$, and that the total number of losses over each such interval is bounded by m.

Towards bounding the RHS of (2.11), note that, by construction, for each n, the tuple $\Sigma^n := (Q_1^n, Q_2^n, \Psi^n)$ forms an inhomogeneous Markov process on the state space

$$\mathcal{S}^n := \{(q_1,q_2,\psi) \in \mathbb{Z}_+^3 : q_1 \vee q_2 \le \beta, \psi \le n, (q_1+q_2)\wedge(n-\psi) = 0\},$$

where the first constraint expresses the buffer limit, the second states that the number of jobs in service does not exceed the number of servers, and the last corresponds to the non-idling condition (the inhomogeneity is due to the structure of arrivals). Denote by $\mathbb{P}_x^n$, $x \in \mathcal{S}^n$, the corresponding Markov family, where x serves as the initial condition, i.e., $\mathbb{P}_x^n(\Sigma^n(0) = x) = 1$. Although Σ^n is not a homogeneous Markov process, The path-valued Markov chain $\{\Sigma^n|_{(j\tau, j\tau+\tau]}\}$, $j \in \mathbb{N}$ is homogeneous by construction, and in particular,

$$\mathbb{P}[\sup_{t\in[j\tau+\frac{\tau}{2}, j\tau+\tau)} Q_2^n(t) = \beta | \Sigma^n(j\tau) = x] = \mathbb{P}_x^n(\sup_{t\in[\frac{\tau}{2},\tau)} Q_2^n(t) = \beta). \tag{2.12}$$

Below, we show that

$$\sup_{x\in\mathcal{S}^n} \mathbb{P}_x^n(\sup_{t\in[\frac{\tau}{2},\tau)} Q_2^n(t) = \beta) \le c_1 e^{-c_2 m}, \tag{2.13}$$

where $c_1, c_2 > 0$ are constants that do not depend on n or k. (Note that the initial condition x could have $q_2 = \beta$, but this does not contradict (2.13) which is a statement regarding the times $[\tau/2, \tau)$.) Combining (2.12) with the estimates (2.11) and (2.13) gives

$$\mathbb{E}R_2^n(T) \le c_1 \frac{T}{\tau} m e^{-c_2 m} = c_1 T n e^{-c_2 m}.$$

Recalling that $m = \lfloor n^a \rfloor$, where $a > 0$, gives $R_2^n(T) \Rightarrow 0$ as $n \to \infty$, and therefore $\hat{R}_2^n(T) \Rightarrow 0$.

The intuitive explanation of (2.13) is simple. During the first half of the period, $(0, \tau/2)$, there are no arrivals, and both queue lengths drop dramatically below the buffer size β, regardless of their initial condition. On $[\tau/2, \tau)$, there are still no class-1 arrivals, and so if Q_2^n comes near β, it is necessarily the longer among the two queues. At these times, class-2 jobs receive all service effort, which again causes Q_2^n to drop.

To prove (2.13), fix $x \in \mathcal{S}^n$. Denote $\theta = \inf\{t \in [\tau/2, \tau) : Q_2^n(t) = \beta\}$. The event indicated in (2.13) can be written as $\{\theta < \infty\}$ (equivalently, $\{\theta \leq \tau\}$). Note first that on that event, it is impossible to have $1 \cdot \Psi^n(s) < n$ for some $s \in [0, \theta]$, when n is sufficiently large. Namely, if n is large then $m = m_n = \lfloor n^a \rfloor < \beta = \beta^n = \lfloor \beta_1 n^{1/2} \rfloor$. Note that non-idling condition can be expressed as

$$\text{for every } t,\ 1 \cdot Q^n(t) > 0 \text{ implies } 1 \cdot \Psi^n(t) = n.$$

Hence the existence of such s implies $Q_2^n(s) = 0$, and thus by (2.9),

$$\beta - 0 = Q_2^n(\theta) - Q_2^n(s) \leq E_2^n(\theta) - E_2^n(s) \leq m,$$

that contradicts $m < \beta$. As a result, using also (2.8), on the event $\{\theta < \infty\}$, one has

$$1 \cdot D^n(t) = S(2nt), \qquad t \leq \theta. \tag{2.14}$$

Next, on the time interval $[0, \tau]$, all class-1 arrivals occur at time τ, thus if there are any losses at this class, they also occur at that time. Thus, by (2.9),

$$Q_1^n(t) = Q_1^n(0) - B_1^n(t), \qquad t \in [0, \tau). \tag{2.15}$$

As for Q_2^n, the same is true regarding the interval $[0, \frac{\tau}{2})$. Thus

$$Q_2^n(t) = Q_2^n(0) - B_2^n(t), \qquad t \in [0, \frac{\tau}{2}).$$

Hence

$$1 \cdot Q^n(\frac{\tau}{2}-) = 1 \cdot Q^n(0) - 1 \cdot B^n(\frac{\tau}{2}-) = 1 \cdot Q^n(0) - S(m),$$

where we used (2.10) and (2.14). Now, using the fact that each queue length is bounded above by β, it follows from the property of the policy to always offer service to the longer queue that, for any $\ell \in \mathbb{N}$, once 2ℓ jobs are removed from the buffers and sent to service, each of the queue lengths is bounded above by $\beta - \ell$. As we have just argued, on the event $\{\theta < \infty\}$ there are $S(m)$ such removals during $[0, \tau/2)$, hence

$$Q_1^n(\frac{\tau}{2}-) \vee Q_2^n(\frac{\tau}{2}-) \leq \beta - \Big\lfloor \frac{S(m)}{2} \Big\rfloor.$$

If indeed $\theta < \infty$, namely, Q_2^n reaches β during $[\frac{\tau}{2}, \tau)$, then there must exist a time $u \in [\tau/2, \theta]$ such that

$$Q_2^n(u-) = \beta - \Big\lfloor \frac{S(m)}{2} \Big\rfloor, \qquad \beta - \Big\lfloor \frac{S(m)}{2} \Big\rfloor < Q_2^n(t) < \beta,\ t \in [u, \theta).$$

Using (2.9), noting there are no losses on this interval,

$$\Big\lfloor \frac{S(m)}{2} \Big\rfloor = Q_2^n(\theta) - Q_2^n(u-) = E_2^n(\theta) - E_2^n(u-) - B_2^n(\theta) + B_2^n(u-).$$

Also $Q_2^n > Q_1^n$ must hold on the interval $[u, \theta)$, since by (2.15), Q_1^n can only decrease from $Q_1^n(\frac{\tau}{2}-)$. Thus the increment of B_2^n equals that of $1 \cdot B^n$. In turn, using (2.10) and the

fact that $1 \cdot \Psi^n = n$ on this interval, this increment is equal to the increment of $1 \cdot D^n$, which, by (2.14) is given by $S(2n\theta) - S(2nu-)$. We thus obtain

$$\Big\lfloor \frac{S(m)}{2} \Big\rfloor = E_2^n(\theta) - E_2^n(u-) - S(2n\theta) + S(2nu-).$$

By (2.7),

$$|E_2^n(\theta) - E_2^n(u-) - 2n(\theta - u)| \le 3.$$

Recalling that, on $\theta < \infty$, $u, \theta \in [\tau/2, \tau]$, it follows that

$$\mathbb{P}_x^n(\theta < \infty) \le \mathbb{P}\Big(\sup_{s,t \in [m,2m]} |S(t) - S(s) - (t-s)| \ge \frac{S(m)}{2} - 5 \Big).$$

Denoting $\bar{S}^m(t) = \frac{S(mt) - mt}{m}$, we have

$$\begin{aligned}\mathbb{P}_x^n(\theta < \infty) &\le \mathbb{P}\Big(\frac{S(m)}{2} - 5 < \frac{m}{4} \Big) + \mathbb{P}\Big(\sup_{s,t \in [1,2]} |\bar{S}^m(t) - \bar{S}^m(s)| \ge \frac{1}{4} \Big) \\ &\le \mathbb{P}\Big(\bar{S}^m(1) < -\frac{1}{2} + \frac{10}{m} \Big) + \mathbb{P}\Big(\sup_{t \in [0,2]} |\bar{S}^m(t)| \ge \frac{1}{8} \Big).\end{aligned}$$

Note that the expression on the RHS does not depend on x. Moreover, by the sample path large deviations principle satisfied by $\bar{S}^m$, each of the two terms above is bounded by $c_1 e^{-c_2 m}$, for constants $c_1, c_2 > 0$ that do not depend on m. This completes the proof of (2.13). As we have argued above, this gives $\hat{R}_2^n \Rightarrow 0$.

Based on the above, the completion of the proof follows closely along the lines of Section 4 of [3]. Thus, for this part, we only provide a sketch. First, the model (2.8) for departures, based on the primitive data S, can alternatively be represented in terms of a pair of potential service processes, namely two rate-1 Poisson processes S_1 and S_2, that are mutually independent, and independent of the system's initial condition:

$$D_i^n(t) = S_i\Big(\mu_i \int_0^t \Psi_i^n(u) du \Big).$$

Next, the balance equations (2.9) and (2.10) translate to the diffusion scale as

$$\hat{Q}_i^n(t) = \hat{Q}_i^n(0) + \hat{E}_i^n(t) - \hat{B}_i^n(t) - \hat{R}_i^n(t)\,, \tag{2.16}$$

$$\hat{\Psi}_i^n(t) = \hat{\Psi}_i^n(0) + \hat{B}_i^n(t) - \hat{S}_i^n\Big(\mu_i \int_0^t \bar{\Psi}_i^n(u) du \Big) - \mu_i \int_0^t \hat{\Psi}_i^n(u) du\,, \tag{2.17}$$

where

$$\bar{\Psi}_i^n = n^{-1} \Psi_i^n, \qquad \hat{S}_i^n = n^{-1/2}(S_i(n\iota) - n\iota), \qquad \hat{B}_i^n = n^{-1/2}(B_i^n - n\lambda_i \iota).$$

Hence

$$\hat{X}_i^n = \hat{Q}_i^n + \hat{\Psi}_i^n = \hat{X}_i^n(0) + \hat{W}_i^n - \mu_i \int_0^{\cdot} (\hat{X}_i^n(u) - \hat{Q}_i^n(u)) du - \hat{R}_i^n, \tag{2.18}$$

where

$$\hat{W}_i^n = \hat{E}_i^n - \hat{S}_i^n\Big(\mu_i \int_0^{\cdot} \bar{\Psi}_i^n(u) du \Big). \tag{2.19}$$

Fix a sequence k_n, $n \in \mathbb{N}$, such that $\lim n^{-1/2} k_n = \infty$ and $\lim n^{-1} k_n = 0$, and, given $T < \infty$, define $T_n = \inf\{t : 1 \cdot R^n(t) \ge k_n\} \wedge T$. Lemma 4.2 of [3] states that, for $i = 1, 2$, $\|\hat{Q}_i^n - N^{-1}(1 \cdot \hat{X}^n)^+\|_{T_n} \to 0$, and $\|\bar{\Psi}_i^n(t) - \rho_i\|_{T_n} \to 0$, in probability, as $n \to \infty$. In the

proof of Proposition 4.3 of [3] it is shown that $\mathbb{P}(T_n < T) \to 0$ as $n \to \infty$. As a result, in the above two statements, T_n can be replaced by T, namely, for any $T < \infty$, for $i = 1, 2$,

$$\|\hat{Q}_i^n - N^{-1}(1 \cdot \hat{X}^n)^+\|_T \to 0, \qquad \|\bar{\Psi}_i^n - \rho_i\|_T \to 0, \qquad \text{in probability, as } n \to \infty. \quad (2.20)$$

By the central limit theorem, $(\hat{S}_1^n, \hat{S}_2^n) \Rightarrow W$, where W is a $(0, A)$-BM, with $A = \mathrm{diag}(1, 1)$. Since $\mu_i = 2$ and $\rho_i = 1/2$, it follows that $\hat{W}^n \Rightarrow W$.

Define $\Gamma : \mathbb{D}_{\mathbb{R}^2}([0, T]) \to \mathbb{D}_{\mathbb{R}^2}([0, T])$ by

$$\Gamma(f)(t) = f(t) - g(t)\mathbf{e}_1\,, \qquad g(t) = \sup_{0 \le u \le t} (2\beta - 1 \cdot f(u))^-\,. \quad (2.21)$$

The following two properties follow directly from the definition, namely there exists a constant C such that

$$\|\Gamma(f) - \Gamma(\tilde{f})\|_T \le C\|f - \tilde{f}\|_T, \qquad f, \tilde{f} \in \mathbb{D}_{\mathbb{R}^2}([0, T]), \quad (2.22)$$

and

$$w_T(\Gamma(f), \cdot) \le C w_T(f, \cdot), \qquad f \in \mathbb{D}_{\mathbb{R}^2}([0, T]). \quad (2.23)$$

Given $z \in \mathbb{D}_{\mathbb{R}^2}$, $z(0) \in G := \{x \in \mathbb{R}^2 : 1 \cdot x \le 2\beta\}$, we say that $(y, \ell) \in \mathbb{D}_{\mathbb{R}^2} \times \mathbb{D}_{\mathbb{R}}$ solves the Skorohod problem (SP) in G, with reflection in the direction $-\mathbf{e}_1$, for data z, if $y(t) \in G$ for all t, ℓ is nonnegative and nondecreasing, and

$$y = z - \ell\mathbf{e}_k, \qquad \int_{[0,\infty)} 1_{\{1 \cdot y < 2\beta\}} d\ell = 0.$$

It is well known that for z as above, a necessary and sufficient condition for (y, ℓ) to be a solution is that $y = \Gamma(z)$.

Based on the fact that $\hat{R}_2^n \Rightarrow 0$ and (2.20), there exists a process $\tilde{X}^n$ such that $\tilde{X}^n - \hat{X}^n \Rightarrow 0$, $\tilde{X}^n(t) \in G$ for all t, and

$$\tilde{X}^n = \hat{X}^n(0) + \hat{W}^n + \int_0^{\cdot} b(\tilde{X}^n(u))du - \hat{R}_1^n\mathbf{e}_1 + \varepsilon^n,$$

where ε^n is a sequence of processes converging to 0 in probability, and $\int 1_{\{1 \cdot \tilde{X}^n < 2\beta\}} d\hat{R}_1^n = 0$. As a result,

$$\tilde{X}^n = \Gamma\Big(\hat{X}^n(0) + \hat{W}^n + \int_0^{\cdot} b(\tilde{X}^n(u))du + \varepsilon^n\Big). \quad (2.24)$$

Taking limits, using properties (2.22) and (2.23) gives the convergence result. □

3 A limit result in conventional heavy traffic

In this section we show that in conventional heavy traffic, the first and second order data of the primitives fully determine the diffusion-scale behavior, and in particular, the diffusion limit exists. The purpose of presenting this result is mainly to contrast it with the previous section's counterexample. An important distinction between the two regimes is that the HW regime gives rise to a nondegenerate N-dimensional diffusion process (such as (2.5)), whereas in the conventional regime the limit is a 1-dimensional diffusion. It therefore comes as no surprise that the reflection due to the buffer size constraint can only occur according to the 1-dimensional Skorohod map. While the result appears to be standard, we have not been able to find it in the literature.

The model is similar to the one considered in Section 2, but has only one sever. The probabilistic assumptions regarding arrivals are as before, namely they satisfy (2.2) and

(2.3). The service time distribution is general. With S_i^n denoting the potential service counting process for class-i customers, it is assumed, analogously to (2.2) and (2.3), that $n^{-1}S_i^n \Rightarrow \mu_i\iota$, and $\hat{S}_i^n := n^{-1/2}(S_i^n - n\mu_i\iota) \Rightarrow W_i^{\rm ser}$, where $W_i^{\rm ser}$ is a $(0, \mu_i\gamma_i^2)$-BM, and $\mu_i > 0$, $\gamma_i \geq 0$ are constants. For each n, the $2N$ processes (A_i^n, S_i^n) are mutually independent.

As before, the sequence of queueing networks approaches heavy traffic, i.e., the limiting traffic intensity $\sum \rho_i = 1$, where $\rho_i = \lambda_i/\mu_i$, the scheduling is according to SLQ, and server is non-idling. We also assume that the system is initially empty. The number of class-i customers in the system at time t is denoted by $X_i^n(t)$. If $T_i^n(t)$ is the service time devoted to class-i customers up to time t and $R_i^n(t)$ counts the number of lost arrivals up to time t then we have

$$X_i^n(t) = E_i^n(t) - S_i^n(T_i^n(t)) - R_i^n(t)\,. \tag{3.1}$$

The ith buffer size is given by $\beta_i^n = \beta n^{1/2} + \varepsilon_i^n n^{1/2}$, where $\varepsilon_i^n \to 0$ for each i, and $\beta > 0$ is a constant. Denote the diffusion-scale versions of the processes by $\hat{X}_i^n = n^{-1/2}X_i^n$ and $\hat{R}_i^n = n^{-1/2}R_i^n$. Straightforward calculation gives

$$\hat{X}_i^n = \hat{W}_i^n + \hat{Y}_i^n - \hat{R}_i^n, \tag{3.2}$$

where

$$\hat{W}_i^n(t) = \hat{E}_i^n(t) - \hat{S}_i^n(T_i^n(t)) + \hat{\lambda}_i^n t\,, \qquad \hat{Y}_i^n = \mu_i n^{1/2}(\rho_i\iota - T_i^n)\,, \tag{3.3}$$

and $\hat{\lambda}_i^n := (\lambda_i^n - n\lambda_i)n^{-1/2} \to \hat{\lambda}_i$, by the assumption made following equation (2.3). The following is often referred to as a *state space collapse* result.

Lemma 3.1. *The scaled number of customers in the various classes are asymptotically equal. Namely,* $\max_{i,j} \|\hat{X}_i^n - \hat{X}_j^n\|_T \Rightarrow 0$*, for any* $T < \infty$*.*

Proof. The proof follows along the lines of Proposition 1 in [5], with minor modifications for finite buffers. □

For $a > 0$, the Skorohod map on the interval $[0, a]$ will be denoted by $\Gamma_{[0,a]}$. It maps $\mathbb{D}_{\mathbb{R}}$ to itself, and is characterized as the first component of the solution map $\psi \to (\varphi, \eta_1, \eta_2)$ to the problem of finding, for a given ψ, a triplet $(\varphi, \eta_1, \eta_2)$, such that

$$\begin{aligned}
&\varphi = \psi + \eta_1 - \eta_2, \qquad \varphi(t) \in [0, a] \text{ for all } t,\\
&\eta_i \text{ are nonnegative and nondecreasing, } \eta_i(0-) = 0, \text{ and}\\
&\int_{[0,\infty)} 1_{(0,a]}(\varphi)d\eta_1 = \int_{[0,\infty)} 1_{[0,a)}(\varphi)d\eta_2 = 0.
\end{aligned}$$

Existence and uniqueness of solutions are well-known (see eg. [4]).

Denote $\alpha = \big(\sum_{i=1}^N \mu_i^{-1}\big)^{-1}$. Let $\tilde{W}$ be a (one-dimensional) $(\tilde{m}, \tilde{A})$-BM, where $\tilde{m} = \alpha\sum_{i=1}^N \frac{\hat{\lambda}_i}{\mu_i}$ and $\tilde{A} = \alpha^2\sum_{i=1}^N \frac{\lambda_i}{\mu_i^2}(\sigma_i^2 + \gamma_i^2)$. Then the process $\tilde{X} := \Gamma|_{[0,\beta]}(\tilde{W})$ is a reflected BM on $[0, \beta]$.

Theorem 3.2. *We have* $(\hat{X}_1^n, \ldots, \hat{X}_N^n) \Rightarrow (\tilde{X}, \ldots, \tilde{X})$.

Proof. Define $\tilde{X}^n = \alpha\sum_{i=1}^N \mu_i^{-1}\hat{X}_i^n$. It follows from Lemma 3.1 that there exists a sequence $\bar{\delta}_n \to 0$, such that, with

$$\Omega_n = \{\max_i \|\hat{X}_i^n - \tilde{X}^n\|_T < \bar{\delta}_n\},$$

one has $\mathbb{P}(\Omega_n) \to 1$ as $n \to \infty$. Now, by (3.2), $\tilde{X}^n = \tilde{W}^n + \tilde{Y}^n - \tilde{R}^n$, where

$$\tilde{W}^n = \alpha\sum_{i=1}^N \frac{\hat{W}_i^n}{\mu_i}, \qquad \tilde{Y}^n = \alpha n^{1/2}\Big(\iota - \sum_{i=1}^N T_i^n\Big), \qquad \tilde{R}^n = \alpha\sum_{i=1}^N \frac{\hat{R}_i^n}{\mu_i}.$$

Note that $t - \sum_{i=1}^{N} T_i^n(t)$ gives the cumulative idleness time of the server by time t. As a result, the process $\tilde{Y}^n$ is non-decreasing, $\tilde{Y}^n(0) = 0$, and by the non-idling condition, increases only when $\tilde{X}^n = 0$. Moreover, $\tilde{R}^n$, is non-decreasing, starts from 0 and since arriving jobs are lost only when the corresponding buffer is full, this process increases only when $\max_i \hat{X}_i^n \geq \beta - \delta_n$, where we denote $\delta_n = \max_i |\varepsilon_i^n|$. As a result, on the event Ω_n, $\tilde{R}^n$ increases only when $\tilde{X}^n \geq a_n := \beta - \delta_n - \bar{\delta}_n$. On Ω_n we have

$$\tilde{X}^n(t) = \alpha \sum_{i=1}^{N} \frac{\hat{X}_i^n(t)}{\mu_i} \leq \alpha \frac{1}{\alpha}(\beta + \delta_n) = \beta + \delta_n \,.$$

Defining $X^{*,n} = \tilde{X}^n \wedge a_n$, we have $X^{*,n} = \tilde{X}^n + e_1^n$, where e_1^n is a process that satisfies $|e_1^n(t)| \leq 2\delta_n + \bar{\delta}_n$ for all t, on Ω_n. Since $\mathbb{P}(\Omega_n) \to 1$, e_1^n converges to zero in probability. By the discussion above, we also have on Ω_n,

$$X^{*,n} = e_1^n + \tilde{W}^n + \tilde{Y}^n - \tilde{R}^n, \qquad X^{*,n}(t) \in [0, a_n] \text{ for all } t,$$

$$\int_{[0,\infty)} 1_{(0,a_n]}(X^{*,n}) d\tilde{Y}^n = \int_{[0,\infty)} 1_{[0,a_n)}(X^{*,n}) d\tilde{R}^n = 0.$$

As a result, $X^{*,n} = \Gamma_{[0,a_n]}(e_1^n + \tilde{W}^n)$ on Ω_n. It follows from the explicit expression for the Skorohod map, provided in [4], that $\|\Gamma_{[0,a_1]}(\psi) - \Gamma_{[0,a_2]}(\psi)\|_T \leq a_2 - a_1$, for any $T < \infty$, $0 < a_1 < a_2 < \infty$ and ψ. As a result, $X^{*,n} = \Gamma_{[0,\beta]}(e_1^n + \tilde{W}^n) + e_2^n$, holds on Ω_n, where $\|e_2^n\|_T \leq \delta_n + \bar{\delta}_n$. Hence, on all of Ω,

$$\tilde{X}^n = \Gamma_{[0,\beta]}(e_1^n + \tilde{W}^n) + e_3^n, \tag{3.4}$$

where e_3^n converges to zero in probability. By (3.3) and the assumed convergence of the processes $\hat{E}_i^n$, $\hat{S}_i^n$ and constants $\hat{\lambda}_i^n$, it follows that $\tilde{W}^n$ is a C-tight sequence of processes. As a result of relation (3.4) and the continuity of $\Gamma_{[0,\beta]}$ as a map from $\mathbb{D}_{\mathbb{R}}([0,T])$ (for arbitrary T), equipped with the uniform topology, to itself, $(\tilde{X}^n, \tilde{Y}^n, \tilde{R}^n)$ is also a C-tight sequence. Hence we obtain from (3.3) that $T_i^n \Rightarrow \rho_i \iota$. It follows that $\tilde{W}^n \Rightarrow \tilde{W}$. Arguing again by the continuity of the Skorohod map, we obtain $\tilde{X}^n \Rightarrow \Gamma_{[0,\beta]}(\tilde{W})$. The result now follows. □

References

[1] Anderson, R. and Orey, S.: Small random perturbations of dynamical systems with reflecting boundary. *Nagoya Math. J.* **60**, (1976), 189–216. MR-0397893

[2] Atar, R.: A diffusion regime with nondegenerate slowdown. *Oper. Res.* **60**, (2012), 490–500. MR-2935073

[3] Atar, R. and saha, S.: An ε-nash equilibrium for strategic customers in heavy traffic, arXiv:1505.01328

[4] Kruk, L., Lehoczky, J., Ramanan, K. and Shreve, S.: An explicit formula for the Skorokhod map on $[0, a]$. *Ann. Probab.* **35**, (2007), 1740–1768. MR-2349573

[5] van Mieghem, J. A.: Due-date scheduling: asymptotic optimality of generalized longest queue and generalized largest delay rules. *Oper. Res.* **51**, (2003), 113–122. MR-1956199

Acknowledgments. The authors are grateful to the two referees for their valuable comments.