Heavy traffic limits for join the shortest estimated queue policy using delayed information*

Rami Atar and David Lipshutz

Viterbi Faculty of Electrical Engineering
Technion — Israel Institute of Technology

January 12, 2020

Abstract

We consider a load balancing problem for a network of parallel queues where information on the state of the queues is subject to a delay. In this setting, adopting a routing policy that performs well when applied to the current state of the queues can perform quite poorly when applied to the delayed state of the queues. Viewing this as a problem of control under partial observations, we propose using an *estimate* of the current queue lengths as the input to the join-the-shortest-queue policy. For a general class of estimation schemes, under a heavy traffic condition, we prove convergence of the diffusion scaled process to a solution of a so-called diffusion model, where an important step toward this goal establishes that the estimated queue lengths undergo state space collapse. In some cases, our diffusion model is given by a novel stochastic delay equation with reflection, where the Skorokhod boundary term appears with delay. We illustrate our results with examples of natural estimation schemes, discuss their implementability and compare their relative performance using simulations.

1 Introduction

In this paper we consider the problem of routing a stream of jobs to load balance a network of parallel queues, where information on the state of the queues is subject to delay. The problem is relevant, for example, when there are messaging or routing delays due to the physical separation between the dispatcher that routes the incoming jobs and the locations of the servers. Furthermore, the distance between the dispatcher and the servers may not be uniform, leading to heterogeneous delays. In this setting, a naive approach is to adopt a routing policy that performs well when applied to the current state of the queues and instead apply it with the delayed state of the queues as the input. As explained below, this approach may perform poorly and result in large oscillations of the queue lengths (QLs). Instead, we take a viewpoint from the literature of control under partial observations, which divides the problem into two parts: state estimation and control based on estimation. Specifically, routing is performed in two stages: (i) estimating the current state of the queues using their delayed state and past routing decisions, and (ii) executing a load balancing

*Research supported in part by the ISF (grant 1184/16). The second author was also supported in part at the Technion by a Zuckerman fellowship.

algorithm that normally runs on actual QLs, but using the estimated QLs as the input since the actual QLs are not available. In the first stage we consider a general class of estimation schemes, while in the second stage the load balancing scheme is *join the shortest queue* (JSQ); that is, upon arrival of an incoming job, the dispatcher estimates the current QLs and routes the job to the queue with the shortest estimated QL. We consider convergence of the diffusion scaled system under a heavy traffic condition, which states that the arrival rate of exogenous jobs and the service rates of the servers are asymptotically balanced. Our main result states that for a general class of estimation schemes, the limit is characterized by a so-called diffusion model, in which the *estimates* of the QLs are load balanced. In addition, we show that for a certain subclass of estimation schemes, solutions of the diffusion model satisfy a novel stochastic delay equation with reflection, where the (Skorokhod) boundary term appears with delays. For this subclass of estimators we also obtain asymptotic bounds on the load balance of the QLs that are linear in the coefficient of variation of the scaled service times. We illustrate our results with examples of estimators, we comment on their implementability and, when feasible, compare their relative performance using simulations.

The model consists of a queueing system with multiple single server stations, each with an infinite capacity buffer. There is a single stream of jobs arriving to the system and upon arrival of a job the dispatcher estimates the state of the queues and routes the job to the queue with the shortest estimated length. At each server jobs are served according to the first-in-first-out discipline, and the server remains busy as long as there are customers in its buffer. Information about the QLs reaches the dispatcher with delay, where the delays are possibly heterogeneous as well as time-inhomogeneous. We assume the dispatcher stores information about its past routing decisions and so the number of jobs routed to each queue up until the current time is also available to the dispatcher during the estimation stage. Thus the information used in the estimation stage consists of the QL processes up to delayed times and the (routed) arrival processes up to the current time.

The main goal of this paper is the analysis of the model at the diffusion scale under heavy traffic conditions. More specifically, the exogenous arrival processes and the service processes are accelerated renewal processes that satisfy an asymptotic balance condition. The basic lemma used to prove the diffusion limit results in this paper asserts that, for a general class of estimators, the estimated QL processes undergo *state space collapse* (SSC). This term refers to the property that the differences between these processes vanish in the scaling limit. SSC is broadly known in the heavy traffic literature to play an important role in describing the limit diffusion process as well as in proving convergence. Specifically, for load balancing policies such as JSQ, it constitutes a key step in establishing the limit (of QL and workload) in the form of a one-dimensional reflected Brownian motion (BM). In the present setting, however, this phenomenon does not apply to the actual QLs but rather to their estimators. Building on the SSC result, we prove convergence of the diffusion scaled versions of the estimated and actual QLs, as well as the idleness processes. For general estimation schemes, the limit is characterized by a *diffusion model*. This model is specified in terms of three equations which can be described as follows: (i) a balance equation for QLs; (ii) a balance equation for routing processes; and (iii) an SSC equation for the estimated QLs. In addition, the formal limits of the scaled exogenous arrival process and scaled service processes are BMs. An advantage of the diffusion model characterization of the heavy traffic limits is that the diffusion model is more analytically tractable than the prelimit scaled processes. For example, an interesting future problem is to analyze stationary distributions of the diffusion model (under suitable stability conditions).

We now provide a rough description of our estimation scheme for the QLs. Since the delayed QLs and past routing decisions are known, estimating the QLs is equivalent to estimating the number of jobs served at each queue over its delay interval. In particular, at each queue,

$$\begin{aligned}\text{estimated QL} = \text{delayed QL} &+ \text{routed arrivals over the delay interval} \\ &- \text{estimated jobs served at the queue over the delay interval.}\end{aligned}$$

Our approach is to estimate the number of jobs served at a queue over its delay interval as the estimated *potential* service of the server over the delay interval (assuming it remains busy over the delay interval) minus the estimated *unused* service over the delay interval. With some abuse of terminology, we refer to the unused service process as the *idleness process* since it is a process that increases only when the server is idle. Under the heavy traffic assumption, our estimate of the centered potential service vanishes and as a result the scaled estimated QL is given by

$$\begin{aligned}\text{scaled estimated QL} = \text{scaled delayed QL} &\\ + \text{centered \& scaled routed arrivals over the delay interval} &\\ + \text{scaled estimated idleness over the delay interval.} &\qquad (1)\end{aligned}$$

In this way the problem of estimating the QLs is reduced to a problem of estimating the idleness over the delay interval. We consider a general class of schemes for estimating the idleness over the delay interval. Our main assumptions for the estimators are technical contraction-type conditions. The main theoretical results proved in this generality consist of existence of continuous weak limits of the diffusion scaled processes, their characterization as the unique solution to the diffusion model equations, and the aforementioned SSC.

We illustrate the generality of the permissible estimation schemes with four examples. In our first example (referred to as a zeroth order estimator), we drop the last term on the right hand side of (1); that is, we set the estimated idleness to be identically zero. [Note from (1) that the QL estimator still uses past routing information even if the idleness estimator does not.] This is the simplest estimator to implement that utilizes past routing decisions. It is motivated by the fact that when QLs are large it is unlikely that idleness is incurred during the delay interval. Furthermore, as we elaborate below, under this estimation scheme, solutions of the diffusion model satisfy a novel stochastic delay equation with reflection. In addition, one can derive asymptotic bounds on the load balance of the QLs. Our second and third examples (respectively referred to as first and second order estimators) are refined versions of the first, in which the idleness is estimated using a first order approximation and, respectively, a second order approximation. These two examples have the advantage that they account for the fact that idleness is likely incurred when the QLs are small, while still being easy to implement. We compare the relative performance of our three estimators under various conditions and demonstrate that they significantly outperform the naive estimator that is simply equal the delayed QLs and does not utilize information about past routing decisions. (Interestingly, the zeroth order estimator and the first order estimator yield the same relative ordering of the estimated QLs, despite the fact that the estimates of the QLs are different, and therefore the estimators lead to identical routing decisions.) Our fourth example is a conditional expectation-type estimator, where the scaled service process is replaced by its BM limit. While implementing this schemes does not appear to be practically feasible, it has theoretical significance. It shows that our results cover estimators that are asymptotically equivalent to the conditional expectation given the complete observable information.

When the delay time is constant and identical for all queues (we consider a more general setting in our main results section) and the estimate of the idleness depends continuously and only on the delayed QL (e.g., as in our first estimator), the solution of the diffusion model satisfies a stochastic delay equation with (normal) reflection of the form:

$$Z(t) = Z(0) + W(t)\mathbf{A} + W([t-r]^+)\mathbf{B} + L([t-r]^+)\mathbf{C} + g(t, Z([t-r]^+)) + L(t). \tag{2}$$

Here $K \geq 2$, $r > 0$ is the delay/lag, Z is a continuous K-dimensional process on $[0, \infty)$ taking values in the non-negative orthant $\mathbb{R}^K_+$, W is a multidimensional BM, $[t-r]^+ = \max\{t-r, 0\}$, $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are matrices of compatible dimension, g is a continuous vector-valued function on $[0, \infty) \times \mathbb{R}^K_+$, and L is a K-dimensional continuous non-decreasing regulator process on $[0, \infty)$ starting at zero whose k^{th} component can increase only when Z_k is zero. (Note that throughout this work we interpret vectors as row vectors.) In (2) the term $[t-r]^+$ appears because we assume Z, W and L are processes on $[0, \infty)$. An alternative approach would be to specify initial conditions for Z, W and L over the delay interval $[-r, 0]$ and let Z, W and L be processes on $[-r, \infty)$, in which case $[t-r]^+$ would be replaced by $t - r$ in (2).

The stochastic delay equation with reflection (2) can be contrasted with the form of a *stochastic delay differential equation* (SDDE) with reflection,

$$d\widetilde{Z}(t) = b(t, \widetilde{Z}_t)dt + \sigma(t, \widetilde{Z}_t)dW(t) + d\widetilde{L}(t), \qquad \widetilde{Z}_0 = \xi. \tag{3}$$

Here $\widetilde{Z}$ is a continuous K-dimensional process on $[-r, \infty)$ taking values in the non-negative orthant $\mathbb{R}^K_+$; $\widetilde{Z}_t$ denotes the path segment in $C([-r, 0], \mathbb{R}^K_+)$, the space of continuous functions on $[-r, 0]$ taking values in $\mathbb{R}^K_+$, defined by $\widetilde{Z}_t(s) = \widetilde{Z}(t+s)$ for $s \in [-r, 0]$; b and σ are time inhomogeneous coefficients of compatible dimension defined on these path segments; $\widetilde{L}$ is a K-dimensional regulator process on $[0, \infty)$ satisfying the same conditions as L in (2); and ξ is a random element taking values in $C([-r, 0], \mathbb{R}^K_+)$. While there is a rich literature on SDDEs without reflection (see [12] and [14] for general references), there is relatively limited work on SDDE with reflection. Some exceptions include the work [8] on stationary distributions of SDDE with reflection and the book [9] which treats numerical methods for SDDE with reflection as well as control. It appears that equations of the form (2) have not previously been considered. Furthermore, the equation has some interesting features that are not apparent in the SDDE with reflection setting. For example, in the SDDE with reflection setting, the $C([-r, 0], \mathbb{R}^K_+)$-valued lag process $\{\widetilde{Z}_t, t \geq 0\}$ is Markovian. In contrast, the analogous lag process for the solution Z of (2) is not Markovian because the boundary term L and the Brownian motion W appear with a delay.

There are a few other works that consider load balancing problems in the presence of delays [13, 15, 16]. Each of these works considers the case that routing decisions are based solely on the delayed QLs (and hence could be categorized as using the delayed QL estimator) and it is shown that using natural load balancing schemes with the delayed QLs as inputs can lead to large oscillations in QLs (especially as the length of the delay increases). This is perhaps not surprising as it is well known in the theory of delay differential equations that delayed negative feedback can lead to sustained oscillations (see, e.g., [11]). The work [13] considers a load balancing problem similar to the one considered here, but implements a routing policy that, upon arrival of an incoming job, selects a fixed number of queues uniformly at random and routes the job to the queue with the shortest delayed length (if all the queues are selected then the policy simply routes incoming jobs to the queue with the shortest delayed length). It is shown (using analysis of fluid equations

and numerics) that as the delay increases, the performance of the policy decreases and the QLs can exhibit large oscillations. To counteract this, the author suggests incorporating increased randomness into the routing policy by selecting fewer queues when incoming jobs arrive. The main contrasts between this work and [13] is that we account for past routing decisions in our policy and consider scaling limits, which allows us to prove SSC of the estimated QLs. Another closely related model is analyzed in [16] (see also [15]). In this work arriving customers individually select which service station to join and there is a delay either because only delayed states of the stations are available to the customers and/or because there is a non-negligible travel time for the customer to the stations. Because routing decisions are made by individual customers rather than by a single dispatcher, it is assumed that when a customer selects a service station to join past routing decisions of other customers are not available. This is the most significant difference between the model studied in [16] and the one studied in this paper, as we elaborate on below. There are several additional differences. For instance, in [16] service stations consist of infinite-server queues, and thus the state corresponds to the number of served customers rather than QLs; service times are exponential with the same service rates at all of the stations; and the selection scheme is not according to JSQ but is probabilistic, where the probabilities are a function of the delayed states of the stations.

The main results of [16] are on fluid and diffusion scaling limits as well as the analysis of oscillations of the fluid model. These limits are characterized as unique solutions of two equations. In the case of the fluid limit, the equation is a deterministic delay differential equation. As for the diffusion limit, the characterizing equation is of the form

$$dZ(t) = b(t)Z(t)dt + \sigma(t)dW(t), \qquad Z(0) = \xi(0). \tag{4}$$

Here, Z is a stochastic process taking values in $\mathbb{R}_+^K$, W is a BM, and the coefficients b and σ are deterministic functions given in terms of the aforementioned fluid limit. In particular, these coefficients do not depend on the delayed solution Z [in contrast to (3)], and although (4) is referred to as an SDDE in [16], perhaps the term stochastic differential equation with time inhomogeneous coefficients captures its dynamics more precisely. (Alternatively, if the state is instead taken to be the joint process consisting of the fluid limit and diffusion limit, then the joint process satisfies an SDDE with time homogeneous coefficients and degenerate noise coefficient.)

The difference in the information structure between the model in [16] and the one studied in this paper leads to strikingly different qualitative properties of the fluid limit. The following comparison is presented to demonstrate this point and is partially based on a formal argument. While a probabilistic routing is used in the model of [16], it is possible to obtain JSQ (but using the delayed QLs) as a formal limit of this probabilistic scheme. In particular, as $\theta \to \infty$, where θ is a parameter used there to define the probabilistic scheme, the probability of the customer selecting the queue with the shortest delayed length converges to one. In [16] it is shown that there is a threshold $r_0 > 0$ for the delay, depending on θ, such that if the delay is greater than r_0, the fluid limit exhibits sustained oscillations. (When the delay is less than this threshold, there may still be oscillations, but those oscillations decay in time.) According to the formulas in [16], r_0 converges to zero as $\theta \to \infty$. Formally, this shows that under JSQ oscillations always occur in presence of positive delay. In contrast, in the setting of this paper, no oscillations occur in the fluid limit, simply because this limit is given by the zero trajectory. (This is an immediate corollary of the fact that diffusion limits exists, because, in our setting, the fluid scaled QL is given by $n^{-1/2}$ times the diffusion scaled QL, where n is the usual scaling parameter.) Thus, using information

on past routings, which constitutes the main difference between the two models, gives rise to a significant qualitative difference: in one model oscillations always occur, in the other they never occur. (This comparison is limited by the difference in structure of the stations between these two models. An analogous comparison for same queueing station structure using different policies would be interesting to carry out, but requires further work.)

The remainder of this paper is organized as follows. In Section 2, before defining the routing policy, we introduce the general structure of the queueing network and our heavy traffic assumption. Then in Section 3 we give a precise formulation of the routing policy and state our main assumptions. We present an important class of estimators in this section as well. In Section 4 we present our main results on the convergence of the scaled system to the diffusion model. Here we also introduce the constrained stochastic delay equation. The proofs of our main results are given in Section 5. In Section 6 we prove a load balancing bound on the QLs for a certain class of estimators. In Section 7 we present numerical results on the performance of our policy. Concluding remarks and open problems appear in Section 8. The appendix contains well known results on the one-dimensional Skorokhod problem and some proofs that our estimators satisfy technical conditions.

Before introducing our model, we list some notation that is used throughout this work. Let $\mathbb{N} = \{1, 2, \dots\}$ denote the natural numbers and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. For $d \in \mathbb{N}$ let $\mathbb{R}^d$ denote d-dimensional Euclidean space and $\mathbb{R}^d_+$ denote the non-negative orthant in $\mathbb{R}^d$. When $d = 1$ we drop the superscript 1 and write $\mathbb{R}$ for the real numbers and $\mathbb{R}_+$ for the non-negative axis. Given $s \in \mathbb{R}$ let $s^+ = \max\{s, 0\}$ and $s^- = \max\{-s, 0\}$. Given row vectors $v, u \in \mathbb{R}^d$ we let $\langle v, u\rangle = v_1u_1 + \cdots + v_du_d$ denote their inner product and $|v| = \sqrt{\langle v, v\rangle}$ denote the Euclidean norm of v. Let $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^d$ denote the vector with one in each component. For a subset $E \subset \mathbb{R}^d$, let $\mathbb{D}(E)$ denote the set of functions from $[0, \infty)$ to E that are right continuous with finite left limits (RCLL). Let $\mathbb{D}_0(E)$ denote the subset of functions $f \in \mathbb{D}(E)$ with $f(0) = 0$, and let $\mathbb{D}_+(\mathbb{R})$ denote the subset of functions $f \in \mathbb{D}(\mathbb{R})$ with $f(0) \geq 0$. We let $\mathbb{C}(E)$ and $\mathbb{C}_0(E)$ respectively denote the subsets of continuous functions in $\mathbb{D}(E)$ and $\mathbb{D}_0(E)$. We equip $\mathbb{D}(E)$ and its subsets with the J_1-Skorokhod topology and recall that when restricted to $\mathbb{C}(E)$ the J_1-Skorokhod topology coincides with the topology of uniform convergence on compact intervals in $[0, \infty)$. Given a function $f \in \mathbb{D}(\mathbb{R}^d)$ and $t \geq 0$, we let $f(t-) = \lim_{s \uparrow t} f(s)$ denote the left limit of f at t with the convention that $f(0-) = f(0)$, and we let $f|_t$ denote the restriction of the function to the interval $[0, t]$. In addition, we let

$$f[s, t] = f(t) - f(s), \qquad 0 \leq s < t < \infty.$$

We let $\iota \in \mathbb{C}(\mathbb{R})$ denote the identity function $\iota(t) = t$ for all $t \geq 0$. The following inequality will be useful throughout this work. For $v \in \mathbb{R}^d$,

$$|v - d^{-1}\langle v, \mathbf{1}\rangle \mathbf{1}| \leq |v|. \tag{5}$$

We abbreviate "random variable", "independent and identically distributed", "left hand side", "right hand side" and "right continuous with finite left limits" as RV, IID, LHS, RHS and RCLL, respectively.

2 General description of the model

Before introducing our routing policy in the next section, we first introduce a general description of our model, the heavy traffic assumption and the diffusion scaling regime that we focus on. Fix an integer $K \geq 2$, which denotes the number of servers in the queueing system, and let $\mathbb{K} = \{1, \dots, K\}$.

2.1 A sequence of networks

Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that supports $K+1$ independent renewal processes: $X, S_1, \ldots, S_K$. We assume the first arrival and the interarrival times for X (resp. S_k) are positive IID random variables with mean 1 and finite variance $\alpha^2 > 0$ (resp. $\gamma_k^2 > 0$). Consider a sequence of queueing networks indexed by $n \in \mathbb{N}$ described as follows:

- For $t \geq 0$ and $k \in \mathbb{K}$ let $Q_k^n(t)$ denote the number of jobs in the k^{th} queue at time t, including those currently in service. We refer to $Q^n = \{Q^n(t) = (Q_1^n(t), \ldots, Q_K^n(t)), t \geq 0\}$ as the *QL process*. For convenience, we assume the queues are initially empty, i.e., $Q^n(0) = 0$.

- For $t \geq 0$ let $X^n(t)$ denote the number of exogenous arrivals to the queueing system in the interval $[0, t]$. We refer to $X^n = \{X^n(t), t \geq 0\}$ as the *exogenous arrival process*. We assume X^n is defined as a time acceleration of X. In particular, $X^n(t) = X(\lambda^n t)$ for $t \geq 0$, where $\lambda^n > 0$ denotes the exogenous arrival rate.

- For $t \geq 0$ and $k \in \mathbb{K}$ let $A_k^n(t)$ denote the number of exogenous arrivals that are routed to the k^{th} queue in the interval $[0, t]$. We refer to $A^n = \{A^n(t) = (A_1^n(t), \ldots, A_K^n(t)), t \geq 0\}$ as the *(routed) arrival process*. Note that $A_k^n(t)$ counts the number of jobs that were routed to the k^{th} server in the interval $[0, t]$, and that the number of (routed) arrivals to the queues in the interval $[0, t]$ must sum to the number of exogenous arrivals in $[0, t]$, i.e.,

$$\langle A^n(t), \mathbf{1} \rangle = X^n(t). \tag{6}$$

- For $t \geq 0$ and $k \in \mathbb{K}$ let $S_k^n(t)$ denote the number of service completions at the k^{th} server after the server is busy for a total of t units of time. We refer to $S^n = \{S^n(t) = (S_1^n(t), \ldots, S_K^n(t)), t \geq 0\}$ as the *service process*. For each $k \in \mathbb{K}$ we assume S_k^n is defined as a time acceleration of S_k. In particular, $S_k^n(t) = S_k(\mu_k^n t)$ for $t \geq 0$, where $\mu_k^n > 0$ denotes the service rate of the k^{th} server. Let $\mu^n = (\mu_1^n, \ldots, \mu_K^n)$.

- For $t \geq 0$ and $k \in \mathbb{K}$ let $B_k^n(t)$ denote the amount of the time the k^{th} server is busy in the interval $[0, t]$. Since we assume the servers are non-idling, it follows that

$$B_k^n(t) = \int_0^t 1_{\{Q_k^n(s) > 0\}} ds. \tag{7}$$

  Then $S_k^n(B_k^n(t))$ denotes the total departures from the k^{th} server in the interval $[0, t]$ and the QL process satisfies, for $k \in \mathbb{K}$,

$$Q_k^n(t) = A_k^n(t) - S_k^n(B_k^n(t)), \qquad t \geq 0. \tag{8}$$

Note that the above equations do not uniquely define the model because the routing policy has not been specified. The routing policy, which is the main focus of this work, is introduced in Section 3.

2.2 Diffusion scaling and heavy traffic assumption

For each $n \in \mathbb{N}$ we define the centered and scaled processes as follows: for $t \geq 0$, let

$$\widehat{Q}_k^n(t) = \frac{Q_k^n(t)}{\sqrt{n}}, \qquad k \in \mathbb{K}, \tag{9}$$

$$\widehat{X}^n(t) = \frac{X^n(t) - \langle \mu^n, \mathbf{1} \rangle t}{\sqrt{n}}, \tag{10}$$

$$\widehat{A}_k^n(t) = \frac{A_k^n(t) - \mu_k^n t}{\sqrt{n}}, \qquad k \in \mathbb{K}, \tag{11}$$

$$\widehat{S}_k^n(t) = \frac{S_k^n(t) - \mu_k^n t}{\sqrt{n}}, \qquad k \in \mathbb{K}, \tag{12}$$

$$\widehat{L}_k^n(t) = \frac{\mu_k^n (t - B_k^n(t))}{\sqrt{n}}, \qquad k \in \mathbb{K}. \tag{13}$$

By relations (6) and (8), and the above definitions, we have the following relations for the diffusion scaled processes:

$$\langle \widehat{A}^n(t), \mathbf{1} \rangle = \widehat{X}^n(t), \qquad t \geq 0, \tag{14}$$

and, for $k \in \mathbb{K}$,

$$\widehat{Q}_k^n(t) = \widehat{A}_k^n(t) - \widehat{S}_k^n(B_k^n(t)) + \widehat{L}_k^n(t), \qquad t \geq 0. \tag{15}$$

Remark 2.1. For $k \in \mathbb{K}$ by the definition of $\widehat{L}_k^n$ in (13) and the definition of B_k^n in (7), we see that $\widehat{L}_k^n$ is non-decreasing and can increase only when $\widehat{Q}_k^n(t) = 0$; that is, when the k^{th} server is idle. We refer to $\widehat{L}^n$ as the scaled *idleness process*. Since $\widehat{Q}_k^n$ is non-negative and (15) holds for each $k \in \mathbb{K}$, it follows that $(\widehat{Q}_k^n, \widehat{L}_k^n)$ is the solution of the one-dimensional Skorokhod problem for $\widehat{A}_k^n - \widehat{S}_k^n \circ B_k^n$ (see Definition A.1). It is well known that the solution of the one-dimensional Skorokhod problem is unique and has an explicit form; in particular, $(\widehat{Q}_k^n, \widehat{L}_k^n) = \Gamma(\widehat{A}_k^n - \widehat{S}_k^n \circ B_k^n)$, where $\Gamma = (\Gamma_1, \Gamma_2)$ is the one-dimensional Skorokhod map (SM) defined in (72)–(73) of Proposition A.2.

According to the functional central limit theorem for renewal processes, the centered and scaled service processes $\{\widehat{S}^n\}_{n=1}^{\infty}$ converge in distribution to a K-dimensional BM with zero drift and suitable diffusion coefficients, as $n \to \infty$. The following heavy traffic condition on the (exogenous) arrival and service rates ensures that the centered and scaled exogenous arrival processes $\{\widehat{X}^n\}_{n=1}^{\infty}$ also converge in distribution to a one-dimensional BM with suitable drift and diffusion coefficient, as $n \to \infty$ (see Proposition 2.3 below).

Assumption 2.2. There are positive constants $\bar{\lambda}$ and $\bar{\mu}_k$, $k \in \mathbb{N}$, and real-valued constants $\hat{\lambda}$ and $\hat{\mu}_k$, $k \in \mathbb{K}$, such that

$$\begin{aligned} \lambda^n &= n\bar{\lambda} + \sqrt{n}\hat{\lambda} + o(\sqrt{n}), \\ \mu_k^n &= n\bar{\mu}_k + \sqrt{n}\hat{\mu}_k + o(\sqrt{n}), \qquad k \in \mathbb{K}, \end{aligned}$$

and the following heavy traffic condition holds:

$$\langle \bar{\mu}, \mathbf{1} \rangle = \bar{\lambda}. \tag{16}$$

We assume the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ supports a one-dimensional BM $\widehat{X} = \{\widehat{X}(t), t \geq 0\}$ with drift $\hat{\lambda} - \langle \hat{\mu}, \mathbf{1} \rangle$ and variance $\bar{\lambda}\alpha^2$, and an independent K-dimensional BM $\widehat{S} = \{\widehat{S}(t) = (\widehat{S}_1(t), \ldots, \widehat{S}_K(t)), t \geq 0\}$ with zero drift and diagonal covariance matrix whose $(k,k)^{\text{th}}$ entry is given by $\bar{\mu}_k \gamma_k^2$, for $k \in \mathbb{K}$.

Proposition 2.3. *Under the heavy traffic condition on the arrival and service rates stated in Assumption 2.2, the sequence* $\{(\widehat{X}^n, \widehat{S}^n)\}_{n=1}^{\infty}$ *converges in distribution to* $(\widehat{X}, \widehat{S})$ *as* $n \to \infty$.

Proof. This follows from the functional central limit theorem for renewal processes (see, e.g., [1, Theorem 14.6]). □

Throughout the remainder of this work we assume, without restatement, that the heavy traffic condition stated in Assumption 2.2 holds.

3 The routing policy: Join the shortest estimated queue

We now provide a precise description of our routing policy.

3.1 The delay function

In this work we assume that when incoming jobs arrive, the dispatcher has access to the state of the queues at some delayed time (as well as all past routings). In order to indicate the information the routing policy depends on, we fix a componentwise non-decreasing *delay function* $\tau : [0, \infty) \to [0, \infty)^K$ satisfying $\tau_k(t) \leq t$ for all $k \in \mathbb{K}$ and $t > 0$. At time $t \geq 0$, the routing policy can depend on the number of jobs in queue k only up until time $\tau_k(t)$, for $k \in \mathbb{K}$. We impose the following technical assumption on the delay function.

Assumption 3.1. The delay function τ is continuous and there is an increasing sequence $\{T_m\}_{m=0}^{\infty}$ with $T_0 = 0$ and $T_m \to \infty$ as $m \to \infty$ such that $\tau_k(t) \leq T_m$ for all $t \in [0, T_{m+1}]$ and $k \in \mathbb{K}$.

Remark 3.2. It is also interesting to consider the case that the delay function has jump discontinuities. For example, a piecewise constant delay function with positive jumps is natural in applications where the dispatcher is updated at discrete time points (see [13]). In this work we restrict to the case of continuous delay functions, which yields a continuous diffusion model, and leave the case of discontinuous delay functions for future consideration.

Remark 3.3. The existence of the sequence $\{T_m\}_{m=0}^{\infty}$ is repeatedly used throughout this work to inductively prove our results hold on the intervals $[0, T_m]$, $m \in \mathbb{N}$. Note the function $\tau = (\iota, \ldots, \iota)$, where we recall that $\iota(t) = t$ for all $t \geq 0$, does not satisfy Assumption 3.1.

Remark 3.4. Given a function $f \in \mathbb{D}(\mathbb{R}^K)$, a delay function $\tau : [0, \infty) \to [0, \infty)^K$ and $t \geq 0$, we adopt the notation

$$(f \circ \tau)(t) = f(\tau(t)) = (f_1(\tau_1(t)), \ldots, f_K(\tau_K(t))) \in \mathbb{R}^K, \tag{17}$$

and

$$f[\tau(t), t] = f(t) - f(\tau(t)) \in \mathbb{R}^K. \tag{18}$$

In addition, we let

$$f|_{\tau(t)} = (f_1|_{\tau_1(t)}, \ldots, f_K|_{\tau_K(t)}) \in \mathbb{D}([0, \tau_1(t)], \mathbb{R}) \times \cdots \times \mathbb{D}([0, \tau_K(t)], \mathbb{R}), \tag{19}$$

where $\mathbb{D}([0, T], \mathbb{R})$ denotes the set of RCLL functions from $[0, T]$ to $\mathbb{R}$, for $T \geq 0$.

We close this section with two examples of delay functions that satisfy Assumption 3.1.

Example 1 (Zero information). Suppose $\tau(t) = 0$ for all $t \geq 0$. Then Assumption 3.1 holds with any increasing sequence $\{T_m\}_{m=0}^{\infty}$ satisfying $T_0 = 0$ and $T_m \to \infty$ as $m \to \infty$.

Example 2 (Delayed information). Suppose $r_1, \ldots, r_K \geq r > 0$ and $\tau_k(t) = (t - r_k)^+$ for $t \geq 0$ and $k \in \mathbb{K}$. Then Assumption 3.1 holds with $T_m = mr$ for $m \in \mathbb{N}_0$.

3.2 Estimator functions

In this work we assume the routing policy is based on (a) estimating the state of the queues and (b) selecting the queue with the shortest estimated QL (ties are broken in lexicographical order). To motivate the assumptions we impose on the estimators, recall the balance equation (15), which implies that for $k \in \mathbb{K}$ and $t \geq 0$,

$$\widehat{Q}_k^n(t) = \widehat{Q}_k^n(\tau_k(t)) + \widehat{A}_k^n[\tau_k(t), t] - (\widehat{S}_k^n \circ B_k^n)[\tau_k(t), t] + \widehat{L}_k^n[\tau_k(t), t]. \tag{20}$$

Note that in the setting considered here, the first two terms on the RHS (i.e., the delayed QL and the past routings over the delay interval) are observable whereas the last two are not. Therefore, estimating QL can be achieved by estimating the third and fourth terms in (20). We are motivated by the role that the *conditional expectation* plays in estimation theory. While our approach does not insist that the estimator for QL be given by the conditional expectation (or approximations thereof), it does treat the third term on the RHS of (20) as if conditional expectation is used. A formal limit of this term is given as the increment of a zero drift BM over the delay interval, where this BM increment is independent of the observed data, and therefore its limiting conditional expectation is zero. In our treatment, the estimate of this term is always zero. In view of this, the task of estimating $\widehat{Q}_k^n(t)$ is transformed into that of estimating the incremental idleness term $\widehat{L}_k^n[\tau_k(t), t]$. Consequently, given an estimator for the incremental idleness term, the corresponding estimator for QL is informally given by

$$\text{estimate of } \widehat{Q}_k^n(t) = \widehat{Q}_k^n(\tau_k(t)) + \widehat{A}_k^n[\tau_k(t), t] + \text{estimate of } \widehat{L}_k^n[\tau_k(t), t]. \tag{21}$$

(A precise version of this identity appears in (31) below.) Whereas (20) is a balance equation for scaled QL, (21) can be viewed as a corresponding balance equation for the scaled estimated QL. All of the assumptions that we impose are related to the structure of the two estimators which appear in (21). They are, however, first presented in the context of the unscaled processes, so as to relate them directly to the queueing model, and rescaling is performed in the next section (Section 3.3). We also introduce an important class of estimators at the end of Section 3.3.

Fix measurable *estimator functions*

$$\Phi^n : \mathbb{D}_0(\mathbb{N}_0^K) \times \mathbb{D}_0(\mathbb{N}_0^K) \to \mathbb{D}_0(\mathbb{R}^K), \tag{22}$$

$$\Psi^n : \mathbb{D}_0(\mathbb{N}_0^K) \times \mathbb{D}_0(\mathbb{N}_0^K) \to \mathbb{D}_0(\mathbb{R}_+^K), \tag{23}$$

such that for $(q, a) \in \mathbb{D}_0(\mathbb{N}_0^K) \times \mathbb{D}_0(\mathbb{N}_0^K)$, $k \in \mathbb{K}$ and $t > 0$,

$$\Phi_k^n(q, a)(t) = q_k(\tau_k(t)) + a_k[\tau_k(t), t] - \mu_k^n(t - \tau_k(t)) + \Psi_k^n(q, a)(t) \tag{24}$$

and $(\Phi^n, \Psi^n)(q, a)(t)$ depends only on $(q|_{\tau(t)}, a|_t)$. Define the RCLL processes $R^n = \Phi^n(Q^n, A^n)$ and $M^n = \Psi^n(Q^n, A^n)$ so that for all $k \in \mathbb{K}$ and $t \geq 0$,

$$R_k^n(t) = Q_k^n(\tau_k(t)) + A_k^n[\tau_k(t), t] - \mu_k^n(t - \tau_k(t)) + M_k^n(t). \tag{25}$$

Then $R_k^n(t)$ serves as our estimate of the length of the k^{th} queue at time t, and the terms $\mu_k^n(t-\tau_k(t))$ and $M_k^n(t)$ can respectively be interpreted as the estimated service capacity of the k^{th} server over the delay interval $[\tau(t), t]$ and the expected idleness at the k^{th} queue over the delay interval. This estimate on service capacity, which is independent of the observable data, is in line with our setting to zero the estimate of service fluctuations. We refer to $R^n = \{R^n(t) = (R_1^n(t), \ldots, R_K^n(t)), t \geq 0\}$ as the *estimated QL process* and we refer to $M^n = \{M^n(t) = (M_1^n(t), \ldots, M_K^n(t)), t \geq 0\}$ as the *estimated idleness process*. The fact that $(\Phi^n, \Psi^n)(Q^n, A^n)(t)$ depends only on $(Q^n|_{\tau(t)}, A^n|_t)$ reflects our condition that the estimators depend only on the delayed QLs as well as past routing decisions. Note from (22)–(23) that we allow the estimated QL process to take negative values; however, we require that the estimated idleness be non-negative. In addition, we require that both processes are initialized to be zero.

With an estimate of the QLs in hand, the dispatcher routes an incoming job to the queue with the shortest estimated QL, where ties are broken according to lexicographical order. For $i \in \mathbb{N}$ let $\sigma_i^n = \inf\{t \geq 0 : X^n(t) \geq i\}$ denote the exogenous arrival time of the i^{th} job. Then upon arrival to the queue, the i^{th} job is routed to the k^{th} queue if $R_k^n(\sigma_i^n-) < R_j^n(\sigma_i^n-)$ for all $j < k$ and $R_k^n(\sigma_i^n-) \leq R_l^n(\sigma_i^n-)$ for all $l \geq k$. In other words, the i^{th} job is routed to the queue with the fewest estimated number of jobs at time σ_i^n. We formalize this policy as follows. Define the measurable *routing function*

$$\Theta^n : \mathbb{D}_0(\mathbb{N}_0^K) \times \mathbb{D}_0(\mathbb{N}_0^K) \to \mathbb{D}(\{e_1, \ldots, e_K\}).$$

as follows. Given $t > 0$, $q \in \mathbb{D}_0(\mathbb{N}_0^K)$ and $a \in \mathbb{D}_0(\mathbb{N}_0^K)$, we set $\Theta^n(q, a)(t)$ equal to e_k, where $k \in \mathbb{K}$ is the unique index such that

$$\begin{aligned} \Phi_k^n(q, a)(t) &< \Phi_j^n(q, a)(t) \qquad \text{for all } j < k, \\ \Phi_k^n(q, a)(t) &\leq \Phi_l^n(q, a)(t) \qquad \text{for all } l \geq k. \end{aligned}$$

Then the arrivals process A^n satisfies

$$A^n(t) = \int_0^t \Theta^n(Q^n, A^n)(s-) dX^n(s).$$

Along with the equations introduced in Section 2.1, this uniquely defines the queueing model (since the trajectories are piecewise constant and the routing function immediately prior to time s, $\Theta^n(Q^n, A^n)(s-)$, depends only on $(Q_1^n|_{[0,\tau_1(s))}, \ldots, Q_K^n|_{[0,\tau_K(s))}, A^n|_{[0,s)})$, this can be argued by induction on the jump times).

3.3 Scaled estimator functions

In order to state our main results, that are concerned with the diffusion scaled model, we need scaled versions of the estimator functions introduced in the last section. Such scaled versions are introduced in this section. At the end of this section we also present an important class estimators, along with their scaled versions.

For each $n \in \mathbb{N}$, define the scaled functions

$$\begin{aligned} \widehat{\Phi}^n &: \mathbb{D}_0(\mathbb{R}_+^K) \times \mathbb{D}_0(\mathbb{R}^K) \to \mathbb{D}_0(\mathbb{R}^K), \\ \widehat{\Psi}^n &: \mathbb{D}_0(\mathbb{R}_+^K) \times \mathbb{D}_0(\mathbb{R}^K) \to \mathbb{D}_0(\mathbb{R}_+^K), \end{aligned}$$

by

$$\widehat{\Phi}^n(q,a)(t) = \frac{\Phi^n\left(\sqrt{n}q, \sqrt{n}a + \mu^n\iota\right)(t)}{\sqrt{n}}, \tag{26}$$

$$\widehat{\Psi}^n(q,a)(t) = \frac{\Psi^n\left(\sqrt{n}q, \sqrt{n}a + \mu^n\iota\right)(t)}{\sqrt{n}}, \tag{27}$$

for all $q \in \mathbb{D}_0(\mathbb{R}_+^K)$ and $a \in \mathbb{D}_0(\mathbb{R}^K)$. Then by (24),

$$\widehat{\Phi}^n(q,a)(t) = q(\tau(t)) + a[\tau(t),t] + \widehat{\Psi}^n(q,a)(t). \tag{28}$$

Define, for each $n \in \mathbb{N}$,

$$\widehat{R}_k^n(t) = \frac{R_k^n(t)}{\sqrt{n}}, \qquad k \in \mathbb{K}, \tag{29}$$

$$\widehat{M}_k^n(t) = \frac{M_k^n(t)}{\sqrt{n}}, \qquad k \in \mathbb{K}. \tag{30}$$

It follows from the above definitions that

$$\widehat{R}^n = \widehat{\Phi}^n(\widehat{Q}^n, \widehat{A}^n) \qquad \text{and} \qquad \widehat{M}^n = \widehat{\Psi}^n(\widehat{Q}^n, \widehat{A}^n).$$

Finally, with these definitions, it follows from equation (28) that equation (21) is satisfied with the rescaled estimators that are now well defined, namely

$$\widehat{R}_k^n(t) = \widehat{Q}_k^n(\tau_k(t)) + \widehat{A}_k^n[\tau_k(t),t] + \widehat{M}_k^n(t), \qquad t \geq 0. \tag{31}$$

In particular, $\widehat{R}_k^n(t)$ and $\widehat{M}_k^n(t)$ serve as estimates of the scaled QL $\widehat{Q}_k^n(t)$ and increment of idleness $\widehat{L}_k^n[\tau_k(t),t]$, respectively. In addition, let

$$\widehat{\Phi} : \mathbb{C}_0(\mathbb{R}_+^K) \times \mathbb{C}_0(\mathbb{R}^K) \to \mathbb{C}_0(\mathbb{R}^K),$$
$$\widehat{\Psi} : \mathbb{C}_0(\mathbb{R}_+^K) \times \mathbb{C}_0(\mathbb{R}^K) \to \mathbb{C}_0(\mathbb{R}^K),$$

be measurable functions such that given $(q,a) \in \mathbb{C}_0(\mathbb{R}_+^K) \times \mathbb{C}_0(\mathbb{R}^K)$ and $t > 0$,

$$\widehat{\Phi}(q,a)(t) = q(\tau(t)) + a[\tau(t),t] + \widehat{\Psi}(q,a)(t), \qquad t > 0. \tag{32}$$

and $(\widehat{\Phi},\widehat{\Psi})(q,a)(t)$ depends only on $(q|_{\tau(t)}, a|_t)$. In the next section we will require that the scaled estimator functions $\{(\widehat{\Phi}^n,\widehat{\Psi}^n)\}_{n=1}^\infty$ converge to $(\widehat{\Phi},\widehat{\Psi})$, which we refer to as the limiting estimator functions, in an appropriate sense.

In Section 7 we provide four detailed examples of estimators (Φ^n,Ψ^n) and discuss their motivation and relative performance. For now we present an important class of estimators for the reader to keep in mind throughout the remainder of this work (in fact, the first three examples to be provided in Section 7 lie in this class).

Example 3. This is a generic example in which the estimator $\widehat{\Psi}^n$ is given by

$$\widehat{\Psi}^n(q,a)(t) = h\left(t, q(\tau(t)), a[\tau(t),t]\right), \tag{33}$$

for a fixed function h within a rich class of functions. The unscaled estimator function Ψ^n is defined in such a way that the relation above holds. Moreover, conditions are imposed on h so that it satisfies several assumptions stated in the next subsection.

To this end, let $h : \mathbb{R}_+ \times \mathbb{R}_+^K \times \mathbb{R}^K \to \mathbb{R}_+^K$ be a function satisfying $h(0,0,0) = 0$ and the following continuity property: for each $T < \infty$ there exist a constant $\eta \in [0,1)$ and a non-decreasing continuous function $\kappa : [0,T] \to \mathbb{R}_+$ with $\kappa(0) = 0$ such that for all $t, t' \in [0,T]$, $v, v' \in \mathbb{R}_+^K$ and $u, u' \in \mathbb{R}^K$

$$|h_k(t,v,u) - h_k(t',v',u')| \leq \kappa(|t-t'|) + |v_k - v_k'| + \eta|u_k - u_k'|, \qquad k \in \mathbb{K}. \tag{34}$$

Define the estimator Ψ^n, for $q \in \mathbb{D}(\mathbb{N}_0^K)$ and $a \in \mathbb{D}_0(\mathbb{N}_0^K)$, by

$$\Psi^n(q,a)(t) = \sqrt{n} h\left(t, \frac{q(\tau(t))}{\sqrt{n}}, \frac{a[\tau(t),t] - \nu^n(t))}{\sqrt{n}}\right), \qquad t \geq 0, \tag{35}$$

where $\nu^n(t)$ is the vector in $\mathbb{R}_+^K$ defined by $\nu_k^n(t) = \mu_k^n(t - \tau_k(t))$ and Φ^n is defined so that (24) holds. Then $\widehat{\Psi}^n$ is of the form (33). In this case, the limiting estimator $\widehat{\Psi}$ is also given by

$$\widehat{\Psi}(q,a)(t) = h(t, q(\tau(t)), a[\tau(t),t]), \qquad t \geq 0. \tag{36}$$

This estimator depends in a rather general way on the delayed QLs and the increment of arrivals over the delay interval. However, it does not use information about the timings of those arrivals within the delay interval. As mentioned above, the condition (34) is required to ensure that it satisfies assumptions on the estimators that are introduced in the next section.

3.4 Assumptions on the estimators

We now state three assumptions on the estimator functions. These assumptions are rather technical, but they reflect our attempt to provide a setting that is as general as possible for the main results to hold. At the end of this section we discuss the generality of our assumptions.

Convergence of the scaled estimator functions

The following is our main convergence assumption on the scaled estimator functions, which assumes that the sequence $\{\widehat{\Psi}^n\}_{n=1}^\infty$ converges to the function $\widehat{\Psi}$ in an appropriate sense.

Assumption 3.5. If the following hold:

(i) $\{(q^n, a^n)\}_{n=1}^\infty$ is a sequence in $\mathbb{D}_0(\mathbb{R}_+^K) \times \mathbb{D}_0(\mathbb{R}^K)$,

(ii) (q,a) lies in $\mathbb{C}_0(\mathbb{R}_+^K) \times \mathbb{C}_0(\mathbb{R}^K)$, and

(iii) $\{(q^n, a^n)\}_{n=1}^\infty$ converges to (q,a) in $\mathbb{D}_0(\mathbb{R}_+^K) \times \mathbb{D}_0(\mathbb{R}^K)$ as $n \to \infty$,

then $\{\widehat{\Psi}^n(q^n, a^n)\}_{n=1}^\infty$ converges to $\widehat{\Psi}(q,a)$ in $\mathbb{D}_0(\mathbb{R}_+^K)$ as $n \to \infty$.

Remark 3.6. Recall that the Skorokhod J_1-topology relativized to the continuous functions coincides with the topology of uniform convergence on compact intervals. Therefore, conditions (i)–(iii) imply that the sequence $\{(q^n, a^n)\}_{n=1}^\infty$ converges to (q,a) uniformly on compact intervals.

From (28) and (32) we note that Assumption 3.5 also implies that if (i)—(iii) hold, then $\{\widehat{\Phi}^n(q^n, a^n)\}_{n=1}^\infty$ converges to $\widehat{\Phi}(q,a)$ in $\mathbb{D}_0(\mathbb{R}^K)$ as $n \to \infty$. This assumption is used to show that sub-sequential limits of the scaled estimator processes exist provided that sub-sequential limits of the QL and arrival processes exist.

Modulus of continuity of the scaled estimated idleness

Our next assumption ensures that the sequence of scaled idleness estimators is C-tight, which implies that the sequence has continuous sub-sequential limits. In the following we recall the definition of C-tightness. Given $d \geq 1$, a function $f : [0, \infty) \to \mathbb{R}^d$, $T < \infty$ and $\delta \in (0, T)$, define the modulus of continuity by

$$w(f, \delta, T) = \sup \{|f(t) - f(s)| : s, t \in [0, T], |t - s| \leq \delta\}.$$

Definition 3.7. We say a sequence of RCLL processes $\{Y^n\}_{n=1}^{\infty}$ is C-tight if the following hold:

$$\lim_{C\to\infty} \limsup_{n\to\infty} \mathbb{P}(|Y^n(0)| \geq C) = 0,$$

and, for all $T < \infty$ and $\varepsilon > 0$,

$$\lim_{\delta\to 0} \limsup_{n\to\infty} \mathbb{P}(w(Y^n, \delta, T) \geq \varepsilon) = 0.$$

The assumption states that the modulus of continuity of the scaled estimator function $\widehat{\Psi}^n(q, a)$ is appropriately bounded by the moduli of continuity of certain transformations of the input functions (q, a).

Assumption 3.8. There is a constant $\zeta_1 \in [0, 1)$ such that for all $n \in \mathbb{N}$, $q \in \mathbb{D}_0(\mathbb{R}_+^K)$, $a \in \mathbb{D}_0(\mathbb{R}^K)$, $T < \infty$, $\delta \in (0, T)$ and $k \in \mathbb{K}$,

$$w(\widehat{\Psi}_k^n(q, a), \delta, T) \leq w(q_k \circ \tau_k, \delta, T) + 2w(a_k \circ \tau_k, \delta, T) + \zeta_1 w(a_k, \delta, T) + o_{\delta,n}(1), \tag{37}$$

where $o_{\delta,n}(1)$ denotes a function of n and δ that does not depend on (q, a) and is such that

$$\lim_{\delta\to 0} \lim_{n\to\infty} o_{\delta,n}(1) = 0.$$

The use of this assumption in our treatment is in showing that subsequential limits of various scaled processes, such as estimators and QLs, have continuous sample paths.

Contraction property of the limiting estimator

Under Assumption 3.5 and Assumption 3.8, (sub-sequential) limits of the scaled processes are shown to be solutions of a so-called diffusion model. The following contraction property ensures uniqueness of solutions to the diffusion model.

Assumption 3.9. There exists $\zeta_2 \in [0, 1)$ such that for each $q \in \mathbb{C}_0(\mathbb{R}_+^K)$, $a, a' \in \mathbb{C}_0(\mathbb{R}^K)$, $t > 0$, and $k \in \mathbb{K}$,

$$|\widehat{\Psi}_k(q, a)(t) - \widehat{\Psi}_k(q, a')(t)| \leq \zeta_2 \sup_{\tau_k(t)\leq s\leq t} |a_k[\tau_k(t), s] - a'_k[\tau_k(t), s]|. \tag{38}$$

Discussion of the assumptions

We now discuss the generality of our assumptions. Recall that, for $k \in \mathbb{K}$ and $t \geq 0$, $\widehat{\Psi}_k^n(\widehat{Q}^n, \widehat{A}^n)(t)$ serves as an estimate of the scaled increment of idleness $\widehat{L}_k^n[\tau_k(t), t]$. From Remark 2.1 and the semigroup property of Γ (Proposition A.2), we have

$$\widehat{L}_k^n[\tau_k(t), t] = \Gamma_2 \left(\widehat{Q}_k^n(\tau_k(t)) + \widehat{A}_k^n[\tau_k(t), \tau_k(t) + \cdot] - \widehat{S}_k^n \circ B_k^n[\tau_k(t), \tau_k(t) + \cdot] \right) (t - \tau_k(t)).$$

By an oscillation inequality for the SM (Lemma A.3), for all $0 < \delta < T$,

$$\begin{aligned} w\left(\widehat{L}_k^n[\tau_k(\cdot), \cdot], \delta, T\right) \leq w\left(\widehat{Q}_k^n(\tau_k(\cdot)), \delta, T\right) + 2w\left(\widehat{A}_k^n(\tau_k(\cdot)), \delta, T\right) + 2w\left(\widehat{S}_k^n(B_k^n(\tau_k(\cdot))), \delta, T\right) \\ + w\left(\widehat{A}_k^n, \delta, T\right) + w\left(\widehat{S}_k^n(B_k^n(\cdot)), \delta, T\right), \end{aligned}$$

and by the Lipschitz(1) continuity of the SM (Proposition A.2), for fixed $t \geq 0$, the mapping

$$\left(\widehat{Q}_k^n(\tau_k(t)), \widehat{A}_k^n[\tau_k(t), \tau_k(t) + \cdot], \widehat{S}_k^n \circ B_k^n[\tau_k(t), \tau_k(t) + \cdot]\right) \mapsto \widehat{L}_k^n[\tau_k(t), t]$$

is Lipschitz(1). Since $\widehat{\Psi}^n(\widehat{Q}^n, \widehat{A}^n)(t)$ serves as an estimator for $\widehat{L}^n[\tau_k(t), t]$ and $\widehat{S}_k^n \circ B_k^n$ is C-tight [due to the facts that $\widehat{S}_k^n$ is C-tight and B_k^n is Lipschitz(1)], it is reasonable to assume that its estimator $\widehat{\Psi}^n$ (and its limit $\widehat{\Psi}$) should satisfy Assumptions 3.5, 3.8 and 3.9 with $\zeta_1 = \zeta_2 = 1$. Here we must take $\zeta_1, \zeta_2 < 1$ for technical reasons — specifically, to ensure strict contraction type properties hold in the proofs of C-tightness of the scaled processes (Lemma 5.2) and uniqueness of the diffusion model (Proposition 5.1). Since ζ_1 and ζ_2 can be chosen arbitrarily close to 1, our assumptions (asymptotically) treat any reasonable estimator of the idleness.

We conclude this section with the following lemma, which states that the class of estimators introduced in Example 3 satisfies our main assumptions.

Lemma 3.10. *Let $h : \mathbb{R}_+ \times \mathbb{R}_+^K \times \mathbb{R}^K \to \mathbb{R}_+^K$ satisfy the conditions stated in Example 3. Suppose the estimators $\{\Psi^n\}_{n=1}^\infty$ are defined as in* (35) *and $\widehat{\Psi}$ is defined as in* (36)*. Then Assumptions 3.5, 3.8 and 3.9 hold.*

The proof of Lemma 3.10 is given in the appendix.

Having stated our main assumptions, we are now ready to introduce the diffusion model and state our main result on the convergence of the diffusion scaled processes.

4 Diffusion limits

We start by introducing the diffusion model. Recall the BMs $\widehat{X}$ and $\widehat{S}$ introduced in Section 2.2 and the limiting estimator function $\widehat{\Phi}$ introduced in Section 3.3.

Definition 4.1. A solution of the diffusion model associated with $(\widehat{X}, \widehat{S}, \widehat{\Phi})$ is a pair of processes $(\widehat{Q}, \widehat{A})$ on $(\Omega, \mathcal{F}, \mathbb{P})$ satisfying the following conditions, with $\{\widehat{\mathcal{F}}_t\}$ and $\{\widehat{\mathcal{G}}_t\}$ denoting the filtrations generated by $\{\widehat{X}(t), \widehat{S}(t)\}$ and $\{\widehat{X}(t), \widehat{S}(\tau(t))\}$ respectively:

(i) $\widehat{Q} = \{\widehat{Q}(t) = (\widehat{Q}_1(t), \ldots, \widehat{Q}_K(t)), t \geq 0\}$ is a K-dimensional $\{\widehat{\mathcal{F}}_t\}$-adapted continuous process such that almost surely

$$\widehat{Q}(t) = \widehat{A}(t) - \widehat{S}(t) + \widehat{L}(t) \in \mathbb{R}^K_+, \qquad t \geq 0, \tag{39}$$

where $\widehat{L} = \{\widehat{L}(t) = (\widehat{L}_1(t), \ldots, \widehat{L}_K(t)), t \geq 0\}$ is an auxiliary K-dimensional $\{\widehat{\mathcal{F}}_t\}$-adapted continuous process such that almost surely $\widehat{L}_k(0) = 0$, $\widehat{L}_k$ is non-decreasing and $\widehat{L}_k$ can increase only when $\widehat{Q}_k$ is zero, for $k \in \mathbb{K}$.

(ii) $\widehat{A} = \{\widehat{A}(t) = (\widehat{A}_1(t), \ldots, \widehat{A}_K(t)), t \geq 0\}$ is a K-dimensional $\{\widehat{\mathcal{G}}_t\}$-adapted continuous process such that almost surely

$$\widehat{A}_1(t) + \cdots + \widehat{A}_K(t) = \widehat{X}(t), \qquad t \geq 0. \tag{40}$$

(iii) Almost surely

$$\widehat{R}_1(t) = \cdots = \widehat{R}_K(t), \qquad t \geq 0, \tag{41}$$

where $\widehat{R}$ is the $\{\widehat{\mathcal{G}}_t\}$-adapted process defined by $\widehat{R} = \widehat{\Phi}(\widehat{Q}, \widehat{A})$.

In this diffusion model, equations (39) and (40) are derived from equations (15) and, respectively, (14), by taking formal limits. Equation (41), on the other hand, that expresses state space collapse of the QL estimators, does not have an analogue in the prelimit equations. Rather, it is related to the fact that the routing policy routes jobs to the queue with the shortest estimated QL, and this action is immediately reflected in the estimators (as follows from the fact that the scaled arrivals $\widehat{A}^n_k(t)$ appears in equation (31) for the scaled estimator $\widehat{R}^n_k(t)$). Hence routing tends to drive the estimators toward equalization. Equation (41) asserts that exact equalization of the estimators is achieved in the limit.

Remark 4.2. By condition (i) of Definition 4.1, almost surely $(\widehat{Q}_k, \widehat{L}_k)$ is a solution of the one-dimensional Skorokhod problem for $\widehat{A}_k - \widehat{S}_k$ (see Definition A.1), for $k \in \mathbb{K}$.

Remark 4.3 (A constrained stochastic delay equation). Let $(\widehat{Q}, \widehat{A})$ denote the unique solution of the diffusion model associated with $(\widehat{X}, \widehat{S}, \widehat{\Phi})$. Then, for $k \in \mathbb{K}$,

$$\begin{aligned}\widehat{Q}_k(t) &= K^{-1}\langle \widehat{R}(t), \mathbf{1}\rangle - \widehat{S}_k[\tau_k(t), t] - \widehat{M}_k(t) + \widehat{L}_k[\tau_k(t), t] \\ &= K^{-1}\widehat{X}(t) - K^{-1}\langle \widehat{S}(\tau(t)), \mathbf{1}\rangle + K^{-1}\langle \widehat{L}(\tau(t)), \mathbf{1}\rangle - \widehat{S}_k[\tau_k(t), t] - \widehat{M}_k(t) + \widehat{L}_k[\tau_k(t), t].\end{aligned}$$

Suppose $\widehat{\Psi} = g(t, \widehat{Q}(\tau(t)))$ for some continuous function $g : \mathbb{R}_+ \times \mathbb{R}^K_+ \to \mathbb{R}^K_+$; in particular, the estimated idleness does not depend on the increment of routed arrivals. Let $\mathbf{m} = (K^{-1}, \ldots, K^{-1})$ and $\mathbf{M}$ denote the $K \times K$ matrix whose diagonal entries are all equal to $K^{-1} - 1$ and whose off diagonal entries are equal to K^{-1}. Then $\widehat{Q}$ satisfies the constrained stochastic delay equation:

$$\widehat{Q}(t) = \widehat{X}(t)\mathbf{m} - \widehat{S}(t) - \widehat{S}(\tau(t))\mathbf{M} + \widehat{L}(\tau(t))\mathbf{M} - g(t, \widehat{Q}(\tau(t))) + \widehat{L}(t). \tag{42}$$

Existence and uniqueness of solutions to the constrained stochastic delay equation (42) follows directly from the form of equation (42) and Assumption 3.1 on the delay function $\tau(\cdot)$ using a proof by induction on intervals of the form $[0, T_m]$, $m \in \mathbb{N}$, and the explicit form of the one-dimensional SM shown in the appendix.

We can now state our main diffusion limit result.

Theorem 4.4. *Suppose the delay function $\tau(\cdot)$ satisfies Assumption 3.1, the sequence of scaled estimators $\{\widehat{\Psi}^n\}_{n=1}^{\infty}$ and the limiting estimator $\widehat{\Psi}$ satisfy the convergence condition stated in Assumption 3.5, the modulus of continuity condition stated in Assumption 3.8, and the contraction property stated in Assumption 3.9. Then there exists a unique solution $(\widehat{Q}, \widehat{A})$ of the diffusion model associated with $(\widehat{X}, \widehat{S}, \widehat{\Phi})$ and the sequence $\{(\widehat{Q}^n, \widehat{A}^n)\}_{n=1}^{\infty}$ converges in distribution to $(\widehat{Q}, \widehat{A})$ as $n \to \infty$.*

The next section is devoted to the proof of Theorem 4.4.

5 Convergence of the diffusion scaled processes

In this section we prove Theorem 4.4. The proof is established by arguing that the scaled processes associated with the model are tight, that every subsequential weak limit solves the diffusion model, and that uniqueness holds for solutions of the diffusion model. This is achieved in the following steps. First, in Section 5.1, the diffusion model is analyzed and it is shown that it has at most one solution. In Section 5.2, a certain sequence of stopping times is introduced, and it is proved that the collection of rescaled stopped processes is C-tight. These stopping times are directly related to the SSC property, in that one can deduce from their convergence in law to $+\infty$ that SSC holds. This is carried out in Section 5.3. The SSC property is thus key in showing that C-tightness holds for all processes involved. In addition, it governs one of the equations forming the diffusion model. With these elements at hand, it only remains to show that every subsequential limit satisfies the diffusion model. This is the content of Section 5.4 where the proof of the main result is completed.

5.1 Uniqueness of solutions to the diffusion model

The following proposition establishes uniqueness for solutions of the diffusion model. Note that for the purpose of proving our main result, Theorem 4.4, there is no need to argue directly that existence of solutions holds for this model. Rather, one obtains existence as a consequence of the main result; namely, the scaling limits are solutions of the diffusion model.

Proposition 5.1. *Suppose the delay function $\tau(\cdot)$ satisfies Assumption 3.1 and the limiting estimator $\widehat{\Psi}$ satisfies the contraction property stated in Assumption 3.9. Then there is at most one solution $(\widehat{Q}, \widehat{A})$ of the diffusion model associated with $(\widehat{X}, \widehat{S}, \widehat{\Phi})$.*

Proof. Suppose $(\widehat{Q}, \widehat{A})$ is a solution of the diffusion model. Set $\widehat{M} = \widehat{\Psi}(\widehat{Q}, \widehat{A})$. Let $k \in \mathbb{K}$. By condition (iii) of Definition 4.1, the fact that $\widehat{R} = \widehat{\Phi}(\widehat{Q}, \widehat{A})$ and the relation (32), we see that

$$\widehat{Q}_k(\tau_k(t)) + \widehat{A}_k(t) - \widehat{A}_k(\tau_k(t)) + \widehat{M}_k(t) = K^{-1}\langle \widehat{Q}(\tau(t)) + \widehat{A}(t) - \widehat{A}(\tau(t)) + \widehat{M}(t), \mathbf{1}\rangle.$$

Rearranging and using condition (ii) of Definition 4.1 yields

$$\begin{aligned}\widehat{A}_k(t) = K^{-1}\widehat{X}(t) + \widehat{A}_k(\tau_k(t)) - K^{-1}\langle \widehat{A}(\tau(t)), \mathbf{1}\rangle - \left(\widehat{Q}_k(\tau(t)) - K^{-1}\langle \widehat{Q}(\tau(t)), \mathbf{1}\rangle\right) \\ - \left(\widehat{M}_k(t) - K^{-1}\langle \widehat{M}(t), \mathbf{1}\rangle\right).\end{aligned} \tag{43}$$

Now suppose $(\widehat{Q}', \widehat{A}')$ is also a solution of the diffusion model. As above, set $\widehat{M}' = \widehat{\Psi}(\widehat{Q}', \widehat{A}')$. Fix a sample path $\omega \in \Omega$. We show that

$$(\widehat{Q}(\omega,t), \widehat{A}(\omega,t)) = (\widehat{Q}'(\omega,t), \widehat{A}'(\omega,t)), \qquad t \in [0, T_m], \tag{44}$$

for all $m \in \mathbb{N}_0$, which, along with the fact that $T_m \to \infty$ as $m \to \infty$, will complete the proof. For notational convenience, we omit the ω dependence throughout the remainder of the proof.

In order to prove (44) holds for each $m \in \mathbb{N}$, we proceed with a proof by induction. The base case $m = 0$ follows immediately because $T_0 = 0$ and $\widehat{Q}(0) = \widehat{Q}'(0) = \widehat{A}(0) = \widehat{A}'(0) = 0$. To prove the induction step let $m \in \mathbb{N}_0$ and suppose that (44) holds. Let $t \in [0, T_{m+1}]$. Since $\tau_k(t) \le T_m$ by Assumption 3.1, it follows from the induction hypothesis that

$$(\widehat{Q} \circ \tau, \widehat{A} \circ \tau)(t) = (\widehat{Q}' \circ \tau, \widehat{A}' \circ \tau)(t), \qquad t \in [0, T_{m+1}] \tag{45}$$

and $\widehat{A}[\tau(t), t] - \widehat{A}'[\tau(t), t] = \widehat{A}(t) - \widehat{A}'(t)$ for all $t \in [0, T_{m+1}]$. Thus, by the contraction property for $\widehat{\Psi}$ stated in Assumption 3.9, there exists $\zeta_2 \in [0, 1)$ such that

$$|\widehat{M}(t) - \widehat{M}'(t)| = |\widehat{\Psi}(\widehat{Q}, \widehat{A})(t) - \widehat{\Psi}(\widehat{Q}', \widehat{A}')(t)| \le \zeta_2 \sup_{0 \le s \le t} |\widehat{A}(s) - \widehat{A}'(s)|.$$

Note that the terms on the RHS of (43) involving $\widehat{A}$ and $\widehat{Q}$ appear with delay, and therefore (45) is applicable. Consequently, by (43), (45), the bound (5) and the previous display, we have, for all $t \in [0, T_{m+1}]$,

$$\sup_{0 \le s \le t} |\widehat{A}(s) - \widehat{A}'(s)| \le \sup_{0 \le s \le t} |\widehat{M}(s) - \widehat{M}'(s)| \le \zeta_2 \sup_{0 \le s \le t} |\widehat{A}(s) - \widehat{A}'(s)|.$$

Since $\zeta_2 \in [0, 1)$, it holds that $\widehat{A} = \widehat{A}'$ on $[0, T_{m+1}]$. Then due to the fact that $\widehat{Q} = \Gamma_1(\widehat{A} - \widehat{S})$ and $\widehat{Q}' = \Gamma_1(\widehat{A}' - \widehat{S})$, it follows that $\widehat{Q} = \widehat{Q}'$ on $[0, T_{m+1}]$. This completes the proof of the induction step. Hence, by the principle of mathematical induction (44) holds for all $m \in \mathbb{N}$. □

5.2 C-tightness of stopped processes

The goal of this section is to introduce a sequence of stopping times, θ^n, that are closely related to the SSC property, and show that some of the processes associated with the model (specifically, $\widehat{A}^n$, $\widehat{Q}^n$, $\widehat{L}^n$, $\widehat{M}^n$) stopped at θ^n, form a C-tight sequence.

For $n \in \mathbb{N}$ define the one-dimensional non-negative process $\widehat{\Delta}^n = \{\widehat{\Delta}^n(t), t \ge 0\}$ by

$$\widehat{\Delta}^n(t) = \max_{k \in \mathbb{K}} \widehat{R}^n_k(t) - \overline{R}^n(t), \qquad t \ge 0, \tag{46}$$

where $\overline{R}^n(t) = K^{-1}\langle \widehat{R}^n(t), \mathbf{1} \rangle$ for all $t \ge 0$. Let $\mathcal{Z}$ denote the set of $(0, \infty)$-valued sequences converging to zero. Given a sequence $\{r_n\}_{n=1}^\infty \in \mathcal{Z}$ define the sequence of stopping times $\{\theta^n\}_{n=1}^\infty$, by

$$\theta^n = \inf\{t \ge 0 : \widehat{\Delta}^n(t) \ge r_n\}, \qquad n \in \mathbb{N}. \tag{47}$$

Lemma 5.2. *Suppose the delay function $\tau(\cdot)$ satisfies Assumption 3.1 and the scaled estimator functions $\{\widehat{\Psi}^n\}_{n=1}^\infty$ satisfy the modulus of continuity condition stated in Assumption 3.8. Let $\{r_n\}_{n=1}^\infty \in \mathcal{Z}$ and define the sequence $\{\theta^n\}_{n=1}^\infty$ as in (47). Then the sequence of stopped processes*

$$\{(\widehat{A}^n(\cdot \wedge \theta^n), \widehat{Q}^n(\cdot \wedge \theta^n), \widehat{L}^n(\cdot \wedge \theta^n), \widehat{M}^n(\cdot \wedge \theta^n)\}_{n=1}^\infty$$

in $\mathbb{D}_0(\mathbb{R}^K) \times \mathbb{D}_0(\mathbb{R}^K_+) \times \mathbb{D}_0(\mathbb{R}^K_+) \times \mathbb{D}_0(\mathbb{R}^K_+)$ is C-tight.

Proof. For each $Z \in \{A, Q, L, M\}$, by the definition of C-tightness in Definition 3.7, along with Assumption 3.1 and the fact that $\widehat{Z}^n(0) = 0$, it suffices to show that the following limit holds for each $m \in \mathbb{N}_0$ and $\varepsilon > 0$,

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}\left(w\left(\widehat{Z}^n, \delta, T_m \wedge \theta^n \right) \geq \varepsilon \right) = 0. \tag{48}$$

The proof proceeds by induction on $m \in \mathbb{N}$. The base case $m = 0$ is immediate since $T_0 = 0$.

Next we prove the induction step. Let $m \in \mathbb{N}$ and suppose that (48) holds for all $\varepsilon > 0$ and $Z \in \{A, Q, L, M\}$. By (31), for each $n \in \mathbb{N}$,

$$\widehat{A}^n(t) = \widehat{A}^n(\tau(t)) + \widehat{R}^n(t) - \widehat{Q}^n(\tau(t)) - \widehat{M}^n(t), \qquad t \geq 0.$$

Since $\langle \widehat{A}^n(t), \mathbf{1} \rangle = \widehat{X}^n(t)$ by (14), it follows that

$$\begin{aligned} \widehat{A}^n(t) = K^{-1}\widehat{X}^n(t)\mathbf{1} + \widehat{A}^n(\tau(t)) - K^{-1}\langle \widehat{A}^n(\tau(t)), \mathbf{1} \rangle \mathbf{1} + \widehat{R}^n(t) - K^{-1}\langle \widehat{R}^n(t), \mathbf{1} \rangle \mathbf{1} \\ + \widehat{Q}^n(\tau(t)) - K^{-1}\langle \widehat{Q}^n(\tau(t)), \mathbf{1} \rangle \mathbf{1} - \widehat{M}^n(t) + K^{-1}\langle \widehat{M}^n(t), \mathbf{1} \rangle \mathbf{1}. \end{aligned}$$

Thus, using the bound (5), we have, for $0 \leq s < t < \infty$,

$$\begin{aligned} |\widehat{A}^n(t) - \widehat{A}^n(s)| \leq |\widehat{A}^n(\tau(t)) - \widehat{A}^n(\tau(s))| + \sqrt{2}K^{-1}|\widehat{X}^n(t) - \widehat{X}^n(s)| \\ + 2\sqrt{K} \sup_{u \in [s,t]} \widehat{\Delta}^n(u) + |\widehat{Q}^n(\tau(t)) - \widehat{Q}^n(\tau(s))| + |\widehat{M}^n(t) - \widehat{M}^n(s)|. \end{aligned}$$

Let $\delta > 0$. By the previous display, the fact that $\widehat{M}^n = \widehat{\Psi}(\widehat{Q}^n, \widehat{A}^n)$, and Assumption 3.8 bounding the modulus of continuity of $\widehat{\Psi}^n(q, a)$, there exists $\zeta_1 \in [0, 1)$ such that

$$\begin{aligned} (1 - \zeta_1) w\left(\widehat{A}^n, \delta, T_{m+1} \wedge \theta^n \right) \leq 3w\left(\widehat{A}^n \circ \tau, \delta, T_{m+1} \wedge \theta^n \right) + \sqrt{2}K^{-1} w\left(\widehat{X}^n, \delta, T_{m+1} \wedge \theta^n \right) \\ + 2\sqrt{K} r_n + 2w\left(\widehat{Q}^n \circ \tau, \delta, T_{m+1} \wedge \theta^n \right) + o_{\delta,n}(1). \end{aligned}$$

Thus, by the induction hypothesis (48), the fact that $\tau(t) \in [0, T_m]$ for all $t \in [0, T_{m+1}]$ by Assumption 3.1, and the C-tightness of $\{\widehat{X}^n\}_{n=1}^{\infty}$ that follows from Proposition 2.3, the following limit holds

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}\left(w\left(\widehat{A}^n, \delta, T_{m+1} \wedge \theta^n \right) \geq \varepsilon \right) = 0.$$

Since $(\widehat{Q}_k^n, \widehat{L}_k^n) = \Gamma(\widehat{A}_k^n - \widehat{S}_k^n \circ B_k^n)$ for each $k \in \mathbb{K}$ by Remark 2.1, $\{\widehat{S}_k^n\}_{n=1}^{\infty}$ is C-tight by Proposition 2.3 and $B_k^n(\cdot)$ is Lipschitz(1) by definition, it follows from the oscillation inequality for the SM stated in (74) that the induction hypothesis (48) holds with $Z \in \{Q, L\}$ and $m + 1$ in place of m. In view of Assumption 3.8, we conclude the induction hypothesis (48) also holds with $Z = M$. This completes the proof of the induction step. Hence, by the principle of mathematical induction, (48) holds for all $\varepsilon > 0$, $m \in \mathbb{N}$ and $Z \in \{A, Q, L, M\}$. □

5.3 State space collapse for the estimated QL process

In this section it is shown that the sequence of random times θ^n converges in law to $+\infty$. By the very definition of θ^n, this directly implies that SSC holds. In view of the results of the previous section, this also gives the C-tightness of the rescaled processes $\widehat{A}^n$, $\widehat{Q}^n$ and $\widehat{L}^n$.

Lemma 5.3. *Suppose the delay function $\tau(\cdot)$ satisfies Assumption 3.1 and the sequence of scaled estimators $\{\widehat{\Psi}^n\}_{n=1}^{\infty}$ satisfy the modulus of continuity condition stated in Assumption 3.8. Given $T \in (0,\infty)$ there exists a sequence $\{r_n\}_{n=1}^{\infty} \in \mathcal{Z}$ such that, if the sequence of stopping times $\{\theta^n\}_{n=1}^{\infty}$ is defined as in (47), then $\mathbb{P}(\theta^n < T) \to 0$. Consequently, the sequence $\{(\widehat{A}^n, \widehat{Q}^n, \widehat{L}^n, \widehat{M}^n)\}_{n=1}^{\infty}$ in $\mathbb{D}_0(\mathbb{R}^K) \times \mathbb{D}_0(\mathbb{R}_+^K) \times \mathbb{D}_0(\mathbb{R}_+^K) \times \mathbb{D}_0(\mathbb{R}_+^K)$ is C-tight, and $\widehat{\Delta}^n \to 0$ in probability.*

Proof. The second assertion follows from the first one in view of Lemma 5.2. To prove the first assertion it suffices to show that for each $m \in \mathbb{N}_0$,

$$\text{there exists a sequence } \{r_n\}_{n=1}^{\infty} \in \mathcal{Z} \text{ such that } \mathbb{P}(\theta^n < T_m) \to 0 \text{ as } n \to \infty. \tag{49}$$

The proof of (49) is established in steps 1 and 2 below.

Step 1. We show that for each $m \in \mathbb{N}$ and sequence $\{r_n\}_{n=1}^{\infty} \in \mathcal{Z}$ satisfying

$$\lim_{n\to\infty} \sqrt{n} r_n = \infty, \tag{50}$$

there exists a sequence of $(0, T_m)$-valued RVs $\{\delta^n\}_{n=1}^{\infty}$ such that, for all n sufficiently large, on the event $\{\theta^n < T_m\}$ one has

$$r_n + \sqrt{n}\delta^n \le c\left[w(\widehat{X}^n, \delta^n, T_m) + w(U^n(\cdot \wedge \theta^n), \delta^n, T_m)\right] + o_n(1). \tag{51}$$

where $c < \infty$ is a deterministic constant that does not depend on m, n or $\{r_n\}_{n=1}^{\infty}$, and $U^n = \{U^n(t), t \ge 0\}$ is the K-dimensional RCLL process defined by

$$U_k^n(t) = -\widehat{S}_k^n(B_k^n(\tau_k(t))) + \widehat{L}_k^n(\tau_k(t)) + \widehat{M}_k^n(t), \qquad t \ge 0.$$

To this end, fix $m \in \mathbb{N}$, a sequence $\{r_n\}_{n=1}^{\infty} \in \mathcal{Z}$, and the corresponding sequence of stopping times $\{\theta^n\}_{n=1}^{\infty}$. By the definition of θ^n and the right continuity of $\widehat{R}^n$, on the event $\{\theta^n < T_m\}$ there exists $k_* = k_*(n) \in \mathbb{K}$, fixed in what follows, such that

$$\widehat{R}_{k_*}^n(\theta^n) - \overline{R}^n(\theta^n) \ge r_n.$$

Let

$$\sigma^n = \sup\left\{t \in [0, \theta^n] : \widehat{R}_{k_*}^n(t) - \overline{R}^n(t) \le \frac{r_n}{4}\right\},$$

where we note that the supremum is not over the empty set because $\widehat{R}^n(0) = 0$. Since the jumps of $\widehat{A}^n$ and $\widehat{S}^n$ are of size $n^{-1/2}$ due to the diffusion scaling, it follows from Assumption 3.8 that the jumps of $\widehat{M}^n$ are of order $O(n^{-1/2})$. Hence, by relation (31) for $\widehat{R}^n$, the jumps of $\widehat{R}_k^n$ are also of order $O(n^{-1/2})$. As a result, for all n sufficiently large, we have $\sigma^n < \theta^n$ and $\widehat{R}_{k_*}^n(\sigma^n) - \overline{R}^n(\sigma^n) \le r_n/2$, owing to property (50) of r_n. Thus,

$$\left[\widehat{R}_{k_*}^n(\theta^n) - \overline{R}^n(\theta^n)\right] - \left[\widehat{R}_{k_*}^n(\sigma^n) - \overline{R}^n(\sigma^n)\right] \ge \frac{r_n}{2} \tag{52}$$

and

$$\widehat{R}^n_{k_*}(t) > \overline{R}^n(t) \text{ for } t \in [\sigma^n, \theta^n]. \tag{53}$$

Since the estimated QL for the $k_*{}^{\text{th}}$ queue is larger than the average estimated QL over the interval $[\sigma^n, \theta^n]$ by (53), it follows from the the routing policy that there are no arrivals to the $k_*{}^{\text{th}}$ queue in the interval $[\sigma^n, \theta^n]$. Therefore by (11), letting $\delta^n = \theta^n - \sigma^n$, we have

$$\widehat{A}^n_{k_*}[\sigma^n, \theta^n] = -\frac{\mu^n_{k_*}}{\sqrt{n}}\delta^n = -\left(\sqrt{n}\bar{\mu}_{k_*} + \hat{\mu}_{k_*}\right)\delta^n + o_n(1), \tag{54}$$

where $\bar{\mu}_{k_*} > 0$ and $\hat{\mu}_{k_*} \in \mathbb{R}$ are the constants from Assumption 2.2 and $o_n(1)$ denotes a constant that converges to zero as $n \to \infty$.

Next we write useful equations for $\widehat{R}^n_k$, $k \in \mathbb{K}$, and $\overline{R}^n$. For $k \in \mathbb{K}$ we have

$$\widehat{R}^n_k(t) = \widehat{A}^n_k(t) + U^n_k(t), \qquad t \geq 0. \tag{55}$$

Letting $\overline{U}^n = K^{-1}\langle U^n, \mathbf{1}\rangle$, it follows from (6) that

$$\overline{R}^n(t) = K^{-1}\widehat{X}^n(t) + \overline{U}^n(t). \tag{56}$$

Hence, by (52), (54), (55) and (56),

$$\frac{r_n}{2} \leq -\sqrt{n}\bar{\mu}_{k_*}\delta^n - \hat{\mu}_{k_*}\delta^n + U^n_{k_*}[\sigma^n, \theta^n] - K^{-1}\widehat{X}^n[\sigma^n, \theta^n] - \overline{U}^n[\sigma^n, \theta^n] + o_n(1).$$

Thus, for a suitable constant $c > 0$, independent of m, n and the sequence $\{r_n\}_{n=1}^\infty$, for sufficiently large n, on the event $\{\theta^n < T_m\}$, we see that (51) holds.

Step 2. We use the principle of mathematical induction to prove (49) holds for each $m \in \mathbb{N}_0$. The base case $m = 0$ is immediate. Now suppose $m \in \mathbb{N}$ is such that (49) holds. Fix a deterministic sequence $\{\delta^n_*\}_{n=1}^\infty$ in $(0, \infty)$ such that $\delta^n_* \to 0$ and $\sqrt{n}\delta^n_* \to \infty$ as $n \to \infty$. Recall that the sequence $\{U^n(\cdot \wedge \theta^n)\}_{n=1}^\infty$ is C-tight by Lemma 5.2 and the uniform continuity of B^n_k and τ_k on compact intervals. Since the processes $\widehat{X}^n$ and $\widehat{S}^n$ are C-tight by Proposition 2.3, and $\delta^n_* \to 0$ as $n \to \infty$, the sequence $\{N_n\}_{n=1}^\infty$ in $\mathbb{R}_+$ defined for $n \in \mathbb{N}$ by

$$N_n = w(\widehat{X}^n, \delta^n, T_m) + w(\widehat{U}^n(\cdot \wedge \theta^n), \delta^n, T_m)$$

converges to zero in probability as $n \to \infty$. Thus we can choose $\{r_n\}_{n=1}^\infty \in \mathcal{Z}$ converging to zero sufficiently slowly such that

$$\limsup_{n\to\infty} \mathbb{P}\left(r_n \leq cN_n\right) = 0.$$

For each $n \in \mathbb{N}$,

$$\mathbb{P}(\theta^n < T_{m+1}) \leq P(\Omega^n_1) + P(\Omega^n_2),$$

where

$$\Omega^n_1 = \{(51) \text{ holds and } \delta^n < \delta^n_*\}, \qquad \Omega^n_2 = \{(51) \text{ holds and } \delta^n \geq \delta^n_*\}.$$

By (51), on Ω^n_1, $r_n \leq cN_n$ and so $\limsup_{n\to\infty} \mathbb{P}(\Omega^n_1) = 0$. As for Ω^n_2, on this event we have, by (51),

$$\sqrt{n}\delta^n_* \leq 2c\left[\sup_{t\in[0,T_{m+1}]} |\widehat{X}^n(t)| + \sup_{t\in[0,T_{m+1}]} |\widehat{U}^n(t \wedge \theta^n)|\right].$$

Since, as mentioned above, $\{\widehat{X}^n\}_{n=1}^{\infty}$ and $\{U^n(\cdot \wedge \theta^n)\}_{n=1}^{\infty}$ are C-tight, the RHS above forms a tight sequence of RVs indexed by $n \in \mathbb{N}$, whereas the LHS tends to infinity. This implies $\limsup_{n\to\infty} \mathbb{P}(\Omega_2^n) = 0$. This completes the proof that $\mathbb{P}(\theta^n < T_{m+1}) \to 0$ as $n \to \infty$ and so the induction step holds. Thus, by the principle of mathematical induction (49) holds for all $m \in \mathbb{N}_0$. $\square$

5.4 Proof of Theorem 4.4

To prove the main result, it remains to show that every subsequential limit of the rescaled processes satisfies the diffusion model.

Proof of Theorem 4.4. First of all, by Lemma 5.3, the RVs $\{\widehat{L}^n\}_{n=1}^{\infty}$ are C-tight; particularly, for each T, the sequence of RVs $\{\widehat{L}^n(T)\}_{n=1}^{\infty}$ is tight. By (13) and the fact that $n^{-1/2}\mu_k^n \to \infty$, using also the non-decreasing property of $\widehat{L}_k^n$ the convergence of B_k^n to ι in distribution as $n \to \infty$ follows. This, along with Proposition 2.3 and Lemma 5.2, implies that given any subsequence $\{n_j\}_{j=1}^{\infty}$, there is a further subsequence, also denoted $\{n_j\}_{j=1}^{\infty}$, and K-dimensional continuous processes $\widehat{A}$, $\widehat{Q}$ and $\widehat{L}$ such that, as $j \to \infty$,

$$\left(\widehat{X}^{n_j}, \widehat{S}^{n_j}, B^{n_j}, \widehat{A}^{n_j}, \widehat{Q}^{n_j}, \widehat{L}^{n_j}\right) \Rightarrow \left(\widehat{X}, \widehat{S}, I, \widehat{A}, \widehat{Q}, \widehat{L}\right),$$

where $I \in \mathbb{C}_0(\mathbb{R}_+^K)$ is defined by $I_k = \iota$ for $k \in \mathbb{K}$. According to the Skorokhod representation theorem (see, e.g., [1, Theorem 6.7]), there is a probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$, a sequence of random variables $\{(\widetilde{X}^{n_j}, \widetilde{S}^{n_j}, \widetilde{B}^{n_j}, \widetilde{A}^{n_j}, \widetilde{Q}^{n_j}, \widetilde{L}^{n_j})\}_{j=1}^{\infty}$ on $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ and random variables $(\widetilde{X}, \widetilde{S}, \widetilde{A}, \widetilde{Q}, \widetilde{L})$ on $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ with

$$\left(\widetilde{X}^{n_j}, \widetilde{S}^{n_j}, \widetilde{B}^{n_j}, \widetilde{A}^{n_j}, \widetilde{Q}^{n_j}, \widetilde{L}^{n_j}\right) \stackrel{d}{=} \left(\widehat{X}^{n_j}, \widehat{S}^{n_j}, B^{n_j}, \widehat{A}^{n_j}, \widehat{Q}^{n_j}, \widehat{L}^{n_j}\right), \qquad j \in \mathbb{N}, \tag{57}$$

and

$$\left(\widetilde{X}, \widetilde{S}, \widetilde{A}, \widetilde{Q}, \widetilde{L}\right) \stackrel{d}{=} \left(\widehat{X}, \widehat{S}, \widehat{A}, \widehat{Q}, \widehat{L}\right), \tag{58}$$

such that $\widetilde{\mathbb{P}}$-almost surely

$$(\widetilde{X}^{n_j}, \widetilde{S}^{n_j}, \widetilde{B}^{n_j}, \widetilde{A}^{n_j}, \widetilde{Q}^{n_j}, \widetilde{L}^{n_j}) \to (\widetilde{X}, \widetilde{S}, I, \widetilde{A}, \widetilde{Q}, \widetilde{L}) \tag{59}$$

in $\mathbb{D}_0([0,\infty), \mathbb{R}) \times (\mathbb{D}_0([0,\infty), \mathbb{R}^K))^5$ as $j \to \infty$. By (57), (11) and (14), for each $j \in \mathbb{N}$,

$$\widetilde{A}_1^{n_j}(t) + \cdots + \widetilde{A}_K^{n_j}(t) = \widetilde{X}^{n_j}(t), \qquad t \geq 0.$$

Upon letting $j \to \infty$, it follows from (59) that $\widetilde{\mathbb{P}}$-almost surely

$$\widetilde{A}_1(t) + \cdots + \widetilde{A}_K(t) = \widetilde{X}(t), \qquad t \geq 0.$$

By (57) and (15), for each $j \in \mathbb{N}$,

$$\widetilde{Q}^{n_j}(t) = \widetilde{A}^{n_j}(t) - \widetilde{S}^{n_j}(\widetilde{B}^{n_j}(t)) + \widetilde{L}^{n_j}(t), \qquad t \geq 0.$$

Letting $j \to \infty$, it follows from (59) that $\widetilde{\mathbb{P}}$-almost surely

$$\widetilde{Q}(t) = \widetilde{A}(t) - \widetilde{S}(t) + \widetilde{L}(t), \qquad t \geq 0.$$

By Assumption 3.5 (and the discussion following Assumption 3.5), $\widetilde{R}^n = \widehat{\Phi}^n(\widetilde{Q}^n, \widetilde{A}^n)$ converges in $\mathbb{D}_0(\mathbb{R}^K)$ to $\widetilde{R} = \widehat{\Phi}(\widetilde{Q}, \widetilde{A})$. In addition, by Lemma 5.3, we have, almost surely,

$$\widetilde{R}_1(t) = \cdots = \widetilde{R}_K(t), \qquad t \geq 0.$$

In particular, by Definition 4.1 and Proposition 5.1 we see that $(\widetilde{A}, \widetilde{Q})$ is the unique solution of the diffusion model associated with $(\widehat{\Phi}, \widetilde{X}, \widetilde{S})$, and so by (58), $(\widehat{A}, \widehat{Q})$ is the unique solution of the diffusion model associated with $(\widehat{\Phi}, \widehat{X}, \widehat{S})$. Since this holds for every convergent subsequence $\{n_j\}_{j=1}^{\infty}$, the proof is complete. □

6 Load balancing bounds

One of our main results is the SSC for the estimated QLs, which asserts that these processes equalize asymptotically. More important is the question of the degree to which the QLs themselves are balanced. Whereas the routing processes are fully observable to the dispatcher, the service processes are not, and therefore their stochasticity necessarily affects the accuracy of any estimation procedure. As a result, one does not expect that SSC holds for the QLs. However, one expects that the degree to which QLs are balanced can be estimated in terms of the level of stochasticity of the service time distributions, and that the QLs nearly load balance when the service times become close to deterministic. The goal of this subsection is to justify and quantify this heuristic by deriving a bound on the level of QL imbalance, that specifically can be made as small as desired by making the service times nearly deterministic. For technical reasons, the load balancing bounds we obtain are for estimators that depend only on the delayed QLs and not the routed arrivals over the delay interval; however, we expect these results to hold more broadly.

To state the result we first must be precise about the terms 'stochasticity level of service times' and 'level of QL imbalance'. The role of the former will be played by the coefficients of variation of the service times. Recall from Section 2.1 that the service process at station k in the n^{th} system to be an accelerated version of the process; that is, $S_k^n(t) = S_k(\mu_k^n t)$ for all $t \geq 0$, where S_k is a renewal process with mean 1 and finite variance $\gamma_k^2 > 0$. Thus, the mean service time at station k in the n^{th} system is given by $1/\mu_k^n$ and its variance is given by $(\gamma_k/\mu_k^n)^2$. Whereas these parameters depend on n, the coefficient of variation (defined for a RV ξ as its standard deviation divided by its mean) of this rescaled service time is equal to γ_k, which does not depend on n. Thus given any $\gamma \in (0,\infty)^K$, we denote the dependence on (n,γ) of all the processes involved by indicating (n,γ) in the superscript, e.g., $Q_k^{n,\gamma}$ and $\widehat{Q}_k^{n,\gamma}$. Next, the level of QL imbalance is defined, for a given $T < \infty$, as

$$\Lambda^{n,\gamma}(T) = \sup_{t\in[0,T]} \max_{j,k\in\mathbb{K}} |\widehat{Q}_j^{n,\gamma}(t) - \widehat{Q}_k^{n,\gamma}(t)|.$$

The following result asymptotically bounds the QLs imbalance in terms of the coefficient of variation γ.

Theorem 6.1. *Suppose $\widehat{\Psi}^n$ is defined as in Example 3 for some continuous function $h : [0,\infty) \times \mathbb{R}_+^K \times \mathbb{R}^K \to \mathbb{R}^K$ satisfying the conditions stated in Example 3 with $\eta = 0$; in other words, h does not depend on the increment of arrivals. Assume that for some fixed $r > 0$, $t - \tau_k(t) \leq r$ for all $t \geq 0$ and $k \in \mathbb{K}$. Then given $T > 0$, there exists a $(0,\infty)$-valued RV ξ, that does not depend on n, γ (but may depend on T, r), such that for every $\gamma \in (0,\infty)^K$, the sequence $\Lambda^{n,\gamma}(T)$ is asymptotically*

dominated by $\langle\gamma,\mathbf{1}\rangle\xi$; that is, for all $u>0$,

$$\limsup_{n\to\infty} P(\Lambda^{n,\gamma}(T)>u)\le P(\langle\gamma,\mathbf{1}\rangle\xi>u)$$

Proof. In this proof we suppress the dependence on γ from notation of all processes. By equation (31) and the definition of the estimator,

$$\widehat{R}^n_k(t)=\widehat{Q}^n_k(\tau_k(t))+\widehat{A}^n_k[\tau(t),t]+\widehat{M}^n_k(t).$$

Denote $\widehat{D}^n_k=\widehat{S}^n_k\circ B^n_k$. Then by (15),

$$\widehat{Q}^n_k(t)=\widehat{Q}^n_k(\tau_k(t))+\widehat{A}^n_k[\tau_k(t),t]-\widehat{D}^n_k[\tau_k(t),t]+\widehat{L}^n_k[\tau_k(t),t].$$

Combining these two equations,

$$\widehat{Q}^n_k(t)=U^n_k(t)+\widehat{L}^n_k(t),\qquad U^n_k(t)=\widehat{R}^n_k(t)-\widehat{M}^n_k(t)-\widehat{D}^n_k[\tau_k(t),t]-\widehat{L}^n_k(\tau_k(t)).$$

Given $k\in\mathbb{K}$, by the properties of $\widehat{Q}^n_k$ and $\widehat{L}^n_k$, it is readily seen that

$$(\widehat{Q}^n_k,\widehat{L}^n_k)=\Gamma(U^n_k),\tag{60}$$

where we recall that $\Gamma=(\Gamma_1\Gamma_2)$ is the SM defined in (72)–(73) of Proposition A.2. Let $m\in\mathbb{N}$ be sufficiently large so that $T\le T_m$. Let $e^n=\sup_{t\in[0,T_m]}\max_{j,k\in\mathbb{K}}|\widehat{R}^n_j(t)-\widehat{R}^n_k(t)|$ and recall that by Lemma 5.3, for any γ, $e^n\to0$ in probability as $n\to\infty$. Define $v^n(T)=\sup_{s\in[0,T]}\max_{j,k\in\mathbb{K}}|U^n_j(s)-U^n_k(s)|$ for all $T\in\mathbb{R}_+$. By the definition of U^n, the bound (35) with $\eta=0$, the Lipschitz(1) property of Γ_2 and the Lipschitz(2) property of Γ_1, we have, for all $l\le m$,

$$\begin{aligned}
v^n(T_l)&=\sup_{s\in[0,T_l]}\max_{j,k\in\mathbb{K}}|U^n_j(s)-U^n_k(s)|\\
&\le e^n+\sup_{s\in[0,T_{l-1}]}\max_{j,k\in\mathbb{K}}|\widehat{Q}^n_j(s)-\widehat{Q}^n_k(s)|+2\max_{k\in\mathbb{K}}w(\widehat{D}^n_k,t-\tau_k(t),T_l)+v^n(T_{l-1})\\
&\le e^n+2\max_{k\in\mathbb{K}}w(\widehat{D}^n_k,t-\tau_k(t),T)+3v^n(T_{l-1}).
\end{aligned}\tag{61}$$

Letting

$$\xi^n=2\max_{k\in\mathbb{K}}w(\widehat{S}^n_k,r,T_m),$$

it follows by the Lipschitz(1) property of the sample paths of B^n_k, and the assumption $t-\tau_k(t)\le r$ for all $t\ge0$ and $k\in\mathbb{K}$, that $2\max_{k\in\mathbb{K}}w(\widehat{D}^n_k,r,T_m)\le\xi^n$, and thus

$$v^n(T_l)\le e^n+\xi^n+3v^n(T_{l-1}).$$

Then denoting

$$C=\sum_{j=0}^{m}3^j,$$

we arrive at the bound

$$v^n(T_m)\le C(e^n+\xi^n).$$

By Proposition 2.3, $\widehat{S}^n \Rightarrow \widehat{S}$, where $\widehat{S}$ is a K-dimensional with zero drift and covariance matrix given by $\mathrm{diag}(\gamma_k^2 \bar{\mu}_k)$. Hence if we let W be a fixed K-dimensional BM with zero drift and covariance matrix $\mathrm{diag}(\bar{\mu}_k)$, so that $\widehat{S} \stackrel{d}{=} \mathrm{diag}(\gamma) W$ and $\widehat{S}^n \Rightarrow \mathrm{diag}(\gamma) W$.

Next, the mapping $\psi \mapsto w(\psi, r, T)$ is continuous in the uniform topology on $[0, T]$. Hence

$$\xi^n \Rightarrow 2 \max_{k \in \mathbb{K}} w(\gamma_k W_k, r, T) = 2 \max_{k \in \mathbb{K}} \gamma_k w(W_k, r, T) \leq \langle \gamma, \mathbf{1} \rangle \xi',$$

where $\xi' = 2 \max_{k \in \mathbb{K}} w(W_k, r, T)$. We have thus shown that, for each γ, $v^n(T) \leq v^n(T_m)$ is asymptotically dominated by $\langle \gamma, \mathbf{1} \rangle C \xi'$, where ξ' and C do not depend on γ.

As for $\Lambda^n(T)$, it follows from (60), the Lipschitz(2) property of Γ_1 that, for every γ, the sequence $\{\Lambda^n(T)\}_{n=1}^{\infty}$ is asymptotically bounded by $\langle \gamma, \mathbf{1} \rangle 2C\xi'$. □

7 Examples of estimators and their relative performance

In this section we illustrate the generality of the permissible estimation schemes with four examples. We discuss the relative performance of the three schemes that are straightforward to implement, and show that they significantly outperform the delayed QL estimator. As mentioned in Section 3.2, our approach is inspired by the role that conditional expectation plays in estimation theory. All our examples can be viewed as approximations of conditional expectation. Note that we do not claim that conditional expectation is an optimal or asymptotically optimal estimator with respect to a given criterion; in Section 8 the question of finding an optimal scheme is posed as an open problem.

Recall that the balance equation (21) was obtained by removing service fluctuations process by arguing that, in the formal limit, the conditional expectation of these fluctuations is zero. As a result, (21) expresses the scaled QL estimator in terms of observables (delayed QL and incremental routings) and the scaled incremental idleness estimator. In order to develop an estimator for the scaled incremental idleness, note first that one can write an explicit expression for incremental idleness in terms of the observables and the service process fluctuations. To this end recall Remark 2.1, according to which $\widehat{L}_k^n$ is given by $\Gamma_2(\widehat{A}_k^n - \widehat{S}_k^n \circ B_k^n)$. By the semigroup property of the Skorokhod map (Proposition A.2), for fixed t, letting

$$Z_k^n(s) = \widehat{Q}_k^n(\tau_k(t)) + \widehat{A}_k^n[\tau_k(t), \tau_k(t) + s] - (\widehat{S}_k^n \circ B_k^n)[\tau_k(t), \tau_k(t) + s], \qquad s \geq 0, \tag{62}$$

we have $\widehat{L}_k^n[\tau_k(t), \tau_k(t) + s] = \Gamma_2(Z_k^n)(s)$. Hence we can express the incremental idleness as

$$\widehat{L}_k^n[\tau_k(t), t] = \Gamma_2(Z_k^n)(t - \tau_k(t)). \tag{63}$$

Our estimators are approximations of the following conditional expectation:

$$\mathbb{E}\left[\Gamma_2(Z_k^n)(t - \tau_k(t)) \big| (\widehat{Q}^n|_{\tau(t)}, \widehat{A}^n|_t)\right]. \tag{64}$$

In particular, each estimator replaces Z_k^n with an approximation derived according to certain heuristics.

Our four estimators are explained in detail below, and are arranged in increasing order of complexity. Roughly speaking, the first three estimators can be viewed as zeroth, first and second order approximations of the conditional expectation in (64). In particular, in the zeroth, first and

second order approximations, we respectively replace the difference $\widehat{A}^n_k - \widehat{S}^n_k \circ B^n_k$ in (62) by the zero function, a linear trajectory and a diffusion process. The advantage of these three estimators is that they have explicit formulae and are straightforward to implement. The fourth estimator is the closest approximation to the conditional expectation (64). Indeed, the estimator is equal to the conditional expectation (64), but with two modifications that are made for technical reasons — the term $\widehat{S}^n_k \circ B^n_k$ in (62) replaced by its weak limit $\widehat{S}_k$ and we discount the expectation by a factor that is less than 1. While this estimator is the closest approximation to conditional expectation, it has the significant disadvantage that it is impractical to implement.

Finally, we note that the estimator process is only needed to make a routing decision, so while it is advantageous for the estimators to be close to the actual QL, it is more important that the relative ordering of the estimators reflect the relative ordering of the QLs. Indeed, as we demonstrate below, our first two estimators yield different estimates of the QLs but identical relative orderings, and therefore result in identical routing decisions.

We now provide precise descriptions of the estimators listed above.

7.1 Zeroth order estimator

Our zeroth order estimator is obtained according to the following heuristic. Replacing $\widehat{A}^n_k$ and $\widehat{S}^n_k \circ B^n_k$ by their zeroth order approximations (i.e., the constant functions identically equal to zero) in equation (62) for Z^n_k, we obtain a zeroth order approximation for Z^n_k that is identically equal to $\widehat{Q}^n_k(\tau_k(t))$. Upon substituting this approximation for Z^n_k in the conditional expectation in (64) and using the explicit form for Γ_2 in (73), we obtain an estimate for $\widehat{L}^n_k[\tau_k(t), t]$ that is identically zero. Specifically, for $n \in \mathbb{N}$, we set

$$\Psi^n(q, a) \equiv 0, \qquad (q, a) \in \mathbb{D}_0(\mathbb{R}^K_+) \times \mathbb{D}_0(\mathbb{R}^K).$$

Then Ψ^n clearly lies in the class of estimators introduced in Example 3 with $h \equiv 0$, and therefore satisfies our main assumptions by Lemma 3.10. In this case, the scaled estimator is given, for $n \in \mathbb{N}$, by

$$\widehat{R}^n_k(t) = \widehat{Q}^n_k(\tau_k(t)) + \widehat{A}^n_k[\tau_k(t), t], \qquad t \geq 0. \tag{65}$$

When the system is overloaded and the QLs are expected to be large, it is natural to assume that the idleness process will be close to zero. The zeroth order approximation takes the estimated idleness to be identically zero even if the QLs are small. Note that while the estimated idleness is set to be zero, the estimated QL $\widehat{R}^n$ still uses past routing decisions, as follows from the fact that the arrival process $\widehat{A}^n_k$ appears on the RHS of (65).

Note that the estimated idleness does not use information about the routing process over the delay interval. In this case the solution $(\widehat{Q}, \widehat{A})$ of the diffusion model (see Remark 4.3) satisfies the constrained stochastic delay equation

$$\widehat{Q}(t) = \widehat{X}(t)\mathbf{m} - \widehat{S}(t) - \widehat{S}(\tau(t))\mathbf{M} + \widehat{L}(\tau(t))\mathbf{M} + \widehat{L}(t),$$

where $\mathbf{m}$ and $\mathbf{M}$ are as in Remark 4.3.

7.2 First order estimator

For our first order estimator we replace the processes $\widehat{A}^n_k$ and $\widehat{S}^n_k \circ B^n_k$ with linear trajectories, as follows. Since the routing policy has access to past routing decisions, we approximate the centered

and scaled arrival process $\widehat{A}_k^n[\tau_k(t), \tau_k(t)+\cdot]$ on $[0, t-\tau_k(t)]$ by the linear trajectory

$$\frac{\widehat{A}_k^n[\tau_k(t), t]}{t-\tau_k(t)}\iota(\cdot), \tag{66}$$

where Assumption 3.1 ensures that $\tau_k(t) < t$ for all $t > 0$. We then use a first order (LLN) approximations for $\widehat{S}_k^n \circ B_k^n$, which is simply equal to the zero function. Substituting these approximations into equation (63) for Z_k^n we obtain the following estimate for $\widehat{L}_k^n[\tau_k(t), t]$ when $t > 0$:

$$\Gamma_2\left(\widehat{Q}_k^n(\tau_k(t)) + \frac{\widehat{A}_k^n[\tau_k(t), t]}{t-\tau_k(t)}\iota(\cdot)\right)(t-\tau_k(t)).$$

Finally, in order to ensure the estimator satisfies our main assumptions, we must discount the above by a constant $\eta^{(1)} \in (0,1)$, which can be chosen to be arbitrarily close to 1. Our scaled estimator $\widehat{\Psi}^n$ is then defined, for $q \in \mathbb{D}_0(\mathbb{R}_+^K)$, $a \in \mathbb{D}_0(\mathbb{R}^K)$ and $k \in \mathbb{K}$, by $\widehat{\Psi}_k^n(q,a)(0) = 0$ and

$$\widehat{\Psi}_k^n(q,a)(t) = \eta^{(1)}\Gamma_2\left(q_k(\tau_k(t)) + \frac{a_k[\tau_k(t), t]}{t-\tau_k(t)}\iota(\cdot)\right)(t-\tau_k(t)), \qquad t > 0.$$

From the explicit expression for Γ_2 and the fact that $q(\tau(t)) \in \mathbb{R}_+^K$, it follows that $\widehat{\Psi}^n(q,a)(t) = h^{(1)}(t, q(\tau(t)), a[\tau(t), t])$, where $h^{(1)} : \mathbb{R}_+ \times \mathbb{R}_+^K \times \mathbb{R}^K \to \mathbb{R}_+^K$ is defined, for $k \in \mathbb{K}$, by

$$h_k^{(1)}(t, v, u) = \eta^{(1)}\left(v_k + u_k\right)^-, \qquad (t, v, u) \in \mathbb{R}_+ \times \mathbb{R}_+^K \times \mathbb{R}^K.$$

It then follows from equation (27) for $\widehat{\Psi}^n$ that

$$\Psi_k^n(q,a)(t) = \sqrt{n}h_k^{(1)}\left(t, \frac{q(\tau(t))}{\sqrt{n}}, \frac{a[\tau(t), t] - \nu^n(t)}{\sqrt{n}}\right),$$

where $\nu^n(t)$ denotes the vector in $\mathbb{R}_+^K$ defined by $w_k^n(t) = \mu_k^n(t-\tau_k(t))$. It is readily verified the first order estimator is of the form introduced in Example 3 and therefore satisfies our main assumptions by Lemma 3.10. In addition, by (28) and the explicit expression for $h^{(1)}$, we see that

$$\begin{aligned}\widehat{\Phi}_k^n(q,a)(t) &= q(\tau_k(t)) + a_k[\tau_k(t), t] + \eta^{(1)}\left\{q(\tau_k(t)) + a_k[\tau_k(t), t]\right\}^- \\ &= \left\{q(\tau_k(t)) + a_k[\tau_k(t), t]\right\}^+ - (1-\eta^{(1)})\left\{q(\tau_k(t)) + a_k[\tau_k(t), t]\right\}^-.\end{aligned}$$

Note that the relative ordering of the estimates $\widehat{\Phi}_k^n$ does not depend on $\eta^{(1)} \in [0,1)$. In fact, the zeroth order estimator is a special case of the first order estimator obtained upon setting $\eta^{(1)} = 0$. *Thus, there is no difference in routing when using the zeroth order estimator and the first order estimator, for any* $\eta^{(1)} \in (0,1)$.

7.3 Second order estimator

Our second order estimator accounts for stochastic fluctuations in the service times. To this end, we replace the process $\widehat{S}_k^n \circ B_k^n$ by its second order (CLT) approximation $\widehat{S}_k$. As in the case of the first order estimator, we replace $\widehat{A}_k^n[\tau_k(t), \tau_k(t)+\cdot]$ by a linear trajectory. (Here one could also use a diffusion approximations of the arrivals, such as a Brownian bridge between the points

$\widehat{A}_k^n(\tau_k(t))$ and $\widehat{A}_k(t)$; however, we are unaware of the explicit formulae needed for implementing such an estimator.) Substituting these approximations into equation (63) for Z_k^n we obtain the following estimate for $\widehat{L}_k^n[\tau_k(t),t]$:

$$\Gamma_2\left(\widehat{Q}_k^n(\tau_k(t)) + \frac{\widehat{A}_k^n[\tau_k(t),t]}{t-\tau_k(t)}\iota(\cdot) + \widehat{S}_k(\cdot)\right)(t-\tau_k(t)).$$

Finally, as in the first order estimator, in order to ensure the estimator satisfies our main assumptions, we must discount the above by a constant $\eta^{(2)} \in (0,1)$. Our scaled estimator $\widehat{\Psi}^n$ is then defined, for $q \in \mathbb{C}_0(\mathbb{R}_+^K)$, $a \in \mathbb{C}_0(\mathbb{R}^K)$ and $k \in \mathbb{K}$, by $\widehat{\Psi}_k^n(q,a)(0) = 0$ and

$$\widehat{\Psi}_k^n(q,a)(t) = \eta^{(2)}\mathbb{E}\left[\Gamma_2\left(q_k(\tau_k(t)) + \frac{a_k[\tau_k(t),t]}{t-\tau_k(t)}\iota(\cdot) + \widehat{S}_k(\cdot)\right)(t-\tau_k(t))\right], \qquad t>0.$$

It follows that

$$\Psi_k^n(q,a)(t) = \sqrt{n}h_k^{(2)}\left(t, \frac{q(\tau(t))}{\sqrt{n}}, \frac{a[\tau(t),t]-\nu^n(t)}{\sqrt{n}}\right),$$

where $\nu^n(t)$ denotes the vector in $\mathbb{R}_+^K$ defined by $\nu_k^n(t) = \mu_k^n(t-\tau_k(t))$ for $k \in \mathbb{K}$, and $h^{(2)} : \mathbb{R}_+ \times \mathbb{R}_+^K \times \mathbb{R}^K \to \mathbb{R}_+^K$ is defined, for $k \in \mathbb{K}$, by $h_k^{(2)}(0,v,u) = \eta^{(2)}v_k^-$ and

$$h_k^{(2)}(t,v,u) = \eta^{(2)}\mathbb{E}\left[\Gamma_2\left(v_k + \frac{u_k}{t-\tau_k(t)}\iota(\cdot) + \widehat{S}_k(\cdot)\right)(t-\tau_k(t))\right], \qquad t>0. \tag{67}$$

It follows from properties of the SM Γ_2 that $h^{(2)}$ is in the class of estimator introduced in Example 3, as stated in the following lemma, whose proof is given in the appendix. In addition, we provide an explicit expression for $h_k^{(2)}(t,v,u)$ since it is equal to $\eta^{(2)}$ times the expected local time of a one-dimensional reflected BM with drift u_k, variance $\mu_k\gamma_k^2$, starting at v_k on the interval $[0,t-\tau_k(t)]$, and the transition probabilities for a one-dimensional reflected BM are known (see, e.g., [7, page 48]).

Lemma 7.1. *Suppose $h^{(2)}$ is defined as in* (67)*. Then $h^{(2)}$ satisfies the conditions stated in Example 3 and for each $k \in \mathbb{K}$,*

$$h_k^{(2)}(t,v,u) = \eta^{(2)}g^{(2)}\left(v_k, \frac{u_k}{t-\tau_k(t)}, \sqrt{\mu_k}\gamma_k, t-\tau_k(t)\right), \qquad (t,v,u) \in (0,\infty)\times\mathbb{R}_+^K\times\mathbb{R}^K,$$

where $g^{(2)} : \mathbb{R}\times\mathbb{R}\times(0,\infty)\times(0,\infty) \to \mathbb{R}$ is defined by

$$\begin{aligned} g^{(2)}(x,b,\sigma,\theta) = {} & -x - b\theta + (x+b\theta)\Upsilon\left(\frac{x+b\theta}{\sigma\sqrt{\theta}}\right) + \sigma\sqrt{\theta}\phi\left(\frac{x+b\theta}{\sigma\sqrt{\theta}}\right) \\ & + 1_{\{b\neq 0\}}\frac{\sigma^2}{2b}\left[\exp\left(-\frac{2bx}{\sigma^2}\right)\Upsilon\left(\frac{-x+b\theta}{\sigma\sqrt{\theta}}\right) - \Upsilon\left(\frac{-x-b\theta}{\sigma\sqrt{\theta}}\right)\right] \\ & + 1_{\{b=0\}}\left[\sigma\sqrt{\theta}\phi\left(-\frac{x}{\sigma\sqrt{\theta}}\right) - x\Upsilon\left(-\frac{x}{\sigma\sqrt{t}}\right)\right]. \end{aligned} \tag{68}$$

Here ϕ and Υ are respectively the density function and the cumulative distribution functions for the standard normal RV; that is, for $x \in \mathbb{R}$,

$$\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}, \qquad \Upsilon(x) = \int_{-\infty}^{x}\phi(y)dy. \tag{69}$$

7.4 Conditional expectation-type estimator

For the final estimator the arrivals are used directly without any approximation, whereas the service fluctuations are approximated as a BM as in the second order estimator. This is our closest example to taking the conditional expectation of the incremental idleness expression given in (63), conditioned on the observables. In particular, we would like to use the conditional expectation

$$\mathbb{E}\left[\Gamma_2\left(\widehat{Q}^n_k(\tau_k(t)) + \widehat{A}^n_k[\tau_k(t),\tau_k(t)+\cdot] - (\widehat{S}^n_k \circ B^n_k)[\tau_k(t),\tau_k(t)+\cdot]\right)(t-\tau_k(t))\middle|(\widehat{Q}^n|_{\tau(t)}, \widehat{A}^n|_t)\right] \tag{70}$$

as our estimator for $\widehat{L}^n_k[\tau_k(t),t]$. To ensure it satisfies our assumptions, we replace $\widehat{S}^n_k \circ B^n_k$ by its limit $\widehat{S}_k$, which is independent of $(\widehat{Q}^n, \widehat{A}^n)$, and discount the conditional expectation by a constant $\eta^{(3)} \in (0,1)$, which can be arbitrarily close to 1. In particular, we define

$$\widehat{\Psi}(q,a)(t) = \eta^{(3)}\widehat{F}(q,a)(t), \qquad t>0,$$

where, for $k \in \mathbb{K}$,

$$\widehat{F}_k(q,a)(t) = \mathbb{E}\left[\Gamma_2\left(q(\tau_k(t)) + a[\tau_k(t),\tau_k(t)+\cdot] - \widehat{S}_k(\cdot)\right)(t-\tau_k(t))\right], \qquad t>0,$$

and set $\Psi^n(q,a) = \sqrt{n}\widehat{\Psi}(q,a)$ for each $n \in \mathbb{N}$. Then formally, $\widehat{\Psi}^n(q,a)$ is asymptotically equivalent to the conditional expectation (70) in the limit as $n\to\infty$ and $\eta^{(3)} \to 1$. Note that $\widehat{F}_k(q,a)(0) = 0$ clearly holds for all $(q,a) \in \mathbb{D}_0(\mathbb{R}^K_+) \times \mathbb{D}_0(\mathbb{R}^K)$ and $k \in \mathbb{K}$. The following lemma, whose proof is given in the appendix, states that the estimator satisfies our main assumptions.

Lemma 7.2. *The sequence of scaled estimators $\{\widehat{\Psi}^n\}_{n=1}^\infty$ and limiting estimator $\widehat{\Psi}$ satisfy the convergence condition stated in Assumption 3.5, the C-tightness condition stated in Assumption 3.8, and the contraction condition stated in Assumption 3.9.*

In general, there is no explicit expression for $\widehat{F}_k(q,a)$ when a is an arbitrary function in $\mathbb{D}_0(\mathbb{R}^K)$. However, for the purposes of estimation, we are interested in evaluating the function $\widehat{F}_k$ at $(\widehat{Q}^n, \widehat{A}^n)$, where $\widehat{A}^n$ is our centered scaled routing process (11). The sample paths of this process are piecewise linear (with drift $-\mu^n/\sqrt{n}$) with positive jumps of size $n^{-1/2}$. In this case, we can use the explicit expression for the transition probability (see, e.g., [7, page 48]) for a one-dimensional reflected BM to derive a formula for $\widehat{F}_k(\widehat{Q}^n, \widehat{A}^n)$ in terms of when the jumps of $\widehat{A}^n_k$ occur, as follows.

Lemma 7.3. *Let $q \in \mathbb{D}_0(\mathbb{R}^K_+)$, $a \in \mathbb{D}_0(\mathbb{R}^K)$, $k \in \mathbb{K}$ and $t \geq 0$, and suppose a_k satisfies*

$$a_k(s) = a_k(\tau_k(t)) + b_k(s-\tau_k(t)) + \sum_{j=1}^J c_j 1_{\{s \geq t_j\}}, \qquad s \in [\tau_k(t), t),$$

where $J \in \mathbb{N}_0$, $\{c_j\}_{j=1}^J$ is a sequence in $\mathbb{R}$, and $\{t_j\}_{j=1}^J$ is an increasing sequence in $(\tau_k(t), t)$. Then

$$\begin{aligned}\widehat{F}_k(q,a)(t) = &-q_k(\tau_k(t)) - a_k(t) + a_k(\tau_k(t)) \\ &+ \int_{c_1}^\infty dz_1 \cdots \int_{c_J}^\infty dz_J \prod_{j=1}^J \pi_k(z_{j-1}, t_j - t_{j-1}; z_j - c_j) \\ &\times \int_{a_k(t)-a_k(t-)}^\infty z\pi_k(z_J, t-t_J; z - a_k(t) + a_k(t-))dz,\end{aligned}$$

where $z_0 = q_k(\tau_k(t))$, $t_0 = \tau_k(t)$ and $\pi_k(x, s; \cdot)$ denotes the transition density, at time $s \geq 0$, for a one-dimensional reflected BM starting at $x \geq 0$ with drift b and variance $\mu_k \gamma_k^2$.

The proof of Lemma 7.3 is given in the appendix. Given $(\widehat{Q}^n, \widehat{A}^n)$, $t > 0$ and $k \in \mathbb{K}$, let J_k^n denote the number of jumps $\widehat{A}_k^n$ has in the interval $(\tau_k(t), t)$ and let $\{\rho_j\}_{j=1}^{J_k^n}$ be the jump times for $\widehat{A}_k^n$ in $(\tau_k(t), t)$, listed in increasing order, so that

$$\widehat{A}_k^n(s) = \widehat{A}_k^n(\tau_k(t)) - \frac{\mu_k^n}{\sqrt{n}} s + \frac{1}{\sqrt{n}} \sum_{j=1}^{J_k^n} 1_{\{s \geq \rho_j^n\}}, \qquad s \in (\tau_k(t), t).$$

Then

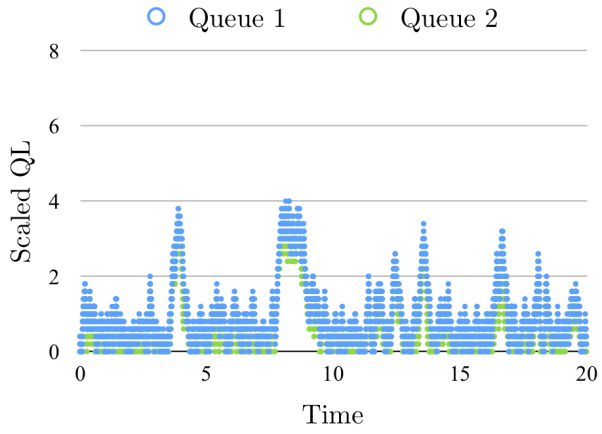
$$\begin{aligned}
\widehat{F}_k(\widehat{Q}^n, \widehat{A}^n)(t) = &-\widehat{Q}_k^n(\tau_k(t)) - \widehat{A}_k^n(t) + \widehat{A}_k^n(\tau_k(t)) \\
&+ \int_{1/\sqrt{n}}^{\infty} dz_1 \cdots \int_{1/\sqrt{n}}^{\infty} dz_J \prod_{j=1}^{J_k^n} \pi_k^n \left(z_{j-1}, \rho_j^n - \rho_{j-1}^n; z_j - \frac{1}{\sqrt{n}} \right) \\
&\times \int_{\widehat{A}_k^n(t) - \widehat{A}_k^n(t-)}^{\infty} z \pi_k^n \left(z_{J_k^n}, t - \tau_k(t) - \rho_J^n; z - \widehat{A}_k^n(t) + \widehat{A}_k^n(t-) \right) dz.
\end{aligned}$$

where $z_0 = \widehat{Q}_k^n(\tau_k(t))$, $\rho_0^n = \tau_k(t)$ and $\pi_k^n(x, s; \cdot)$ denotes the transition density, at time $s \geq 0$, for a one-dimensional reflected BM starting at $x \geq 0$ with drift $-\mu_k^n/\sqrt{n}$ and variance $\mu_k \gamma_k^2$. Thus, the lemma provides an explicit expression for $\widehat{F}(\widehat{Q}^n, \widehat{A}^n)(t)$ in terms of the jumps of $\widehat{A}^n$ in the interval $[\tau_k(t), t]$. However, the expression contains $\sum_{k \in \mathbb{K}} J_k^n + K$ integrals, and $\sum_{k \in \mathbb{K}} J_k^n$ is of order n, so is it impractical to compute the explicit expression for $\widehat{F}(\widehat{Q}^n, \widehat{A}^n)$ each time an exogenous job arrives. Therefore, the significance of this estimator is not practical but rather theoretical. It shows that an approximation of conditional expectation, which is formally equal to conditional expectation in the limit, can be addressed by our theoretical results.

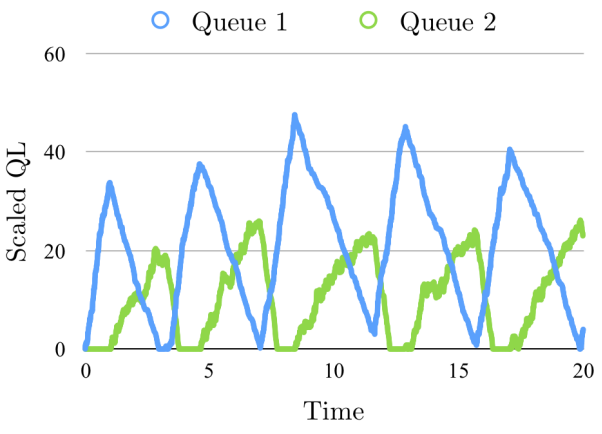
7.5 Relative performance

We now compare the relative performance of the zeroth order estimator and the second order estimator, which we respectively abbreviate as JSEQ(0) and JSEQ(2), and also compare their performance to JSQ and the naive policy that routes jobs to the queue with the shortest delayed QL, abbreviated JSDQ. (Recall there is no difference in routing when using the zeroth order estimator and the first order estimator.)

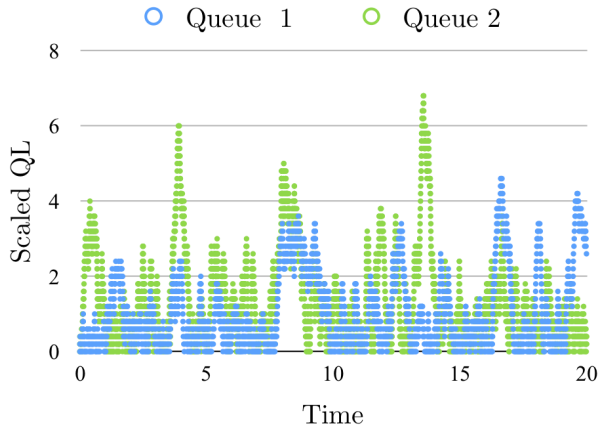
In Figures 1 and 2 we plot sample paths of the scaled QLs $(\widehat{Q}_1^n, \widehat{Q}_2^n)$ for a system of two parallel servers with exponential interarrival times and log normal service times under routing schemes JSQ, JSDQ, JSEQ(0) and JSEQ(2). In the case of JSEQ(0) and JSEQ(2), we also plot sample paths of the scaled estimated QLs $(\widehat{R}_1^n, \widehat{R}_2^n)$. In Figure 1 the difference $b := \hat{\lambda} - \langle \hat{\mu}, \mathbf{1} \rangle$ (i.e., the drift of the BM $\widehat{X}$) is negative whereas in Figure 2 the difference is positive. As observed in Figures 1 and 2, under the JSDQ policy the scaled QLs exhibit large sustained oscillations and both the JSEQ(0) and JSEQ(2) policies significantly outperform the JSDQ policy. (Note that the QLs under JSDQ are plotted on a different scale than the QLs under the other routing policies.) The oscillations under JSDQ arise because JSDQ does not account for routing decisions over the delay interval. In addition, we note that while the scaled estimator under the JSEQ(0) takes negative values when



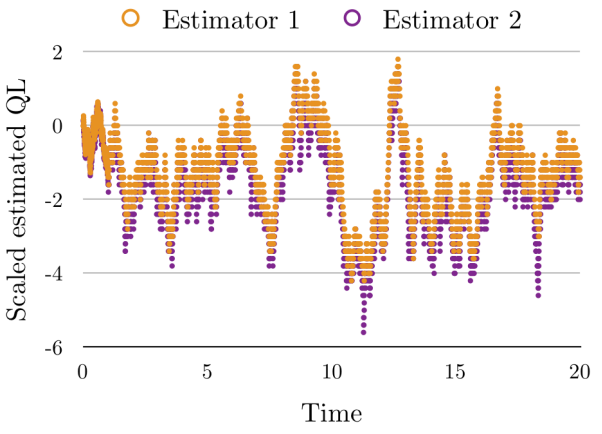
(a) $(\widehat{Q}_1^n, \widehat{Q}_2^n)$ under JSQ



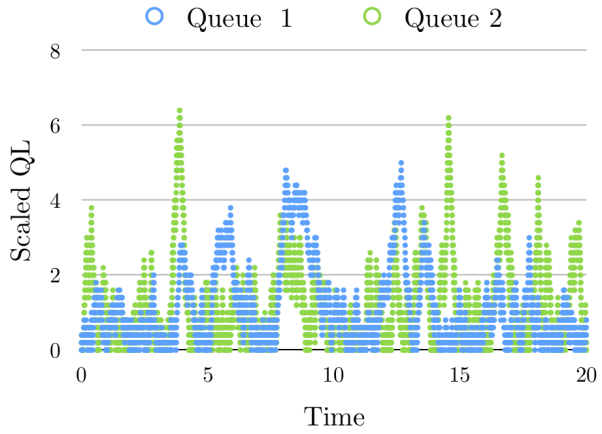
(b) $(\widehat{Q}_1^n, \widehat{Q}_2^n)$ under JSDQ



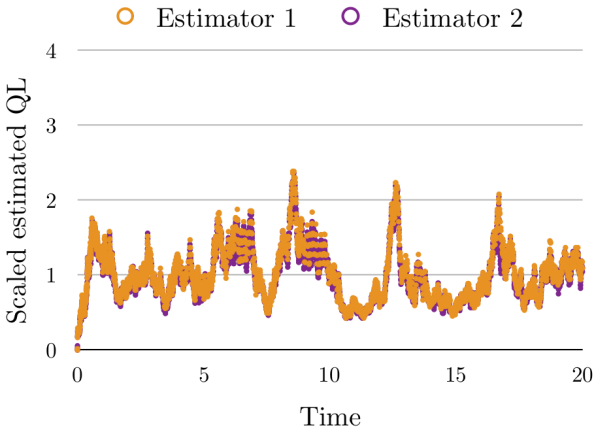
(c) $(\widehat{Q}_1^n, \widehat{Q}_2^n)$ under JSEQ(0)



(d) $(\widehat{R}_1^n, \widehat{R}_2^n)$ under JSEQ(0)

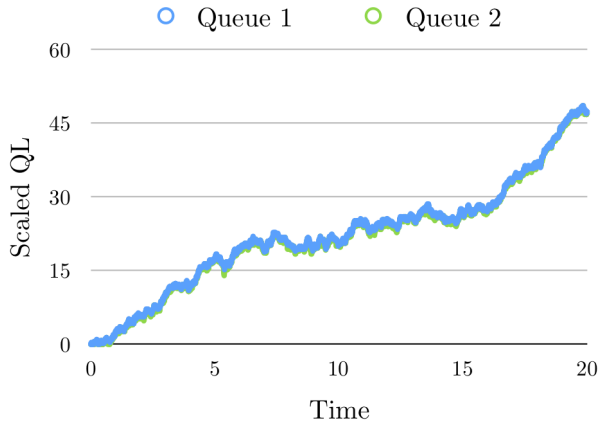


(e) $(\widehat{Q}_1^n, \widehat{Q}_2^n)$ under JSEQ(2)

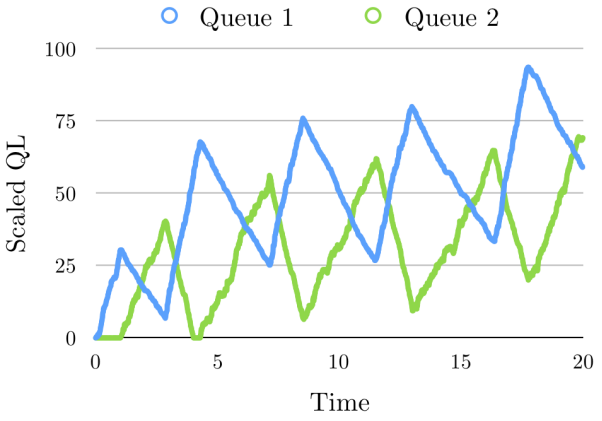


(f) $(\widehat{R}_1^n, \widehat{R}_2^n)$ under JSEQ(2)

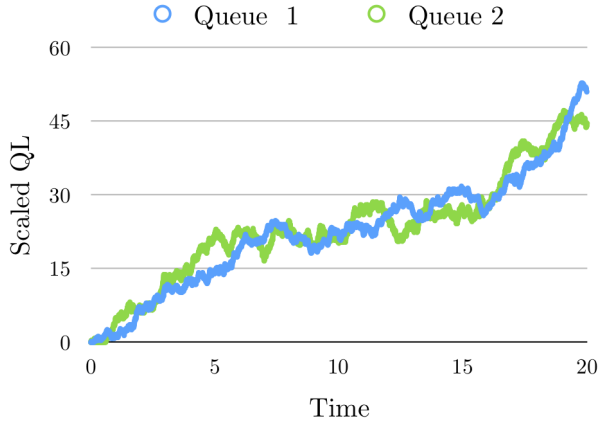
Figure 1: Sample paths for the scaled QLs $(\widehat{Q}_1^n, \widehat{Q}_2^n)$ and scaled estimated QLs $(\widehat{R}_1^n, \widehat{R}_1^n)$ (when applicable) under the following routing policies: JSQ, JSDQ, JSEQ(0), and JSEQ(2) with $\eta^{(2)} = .99$. The interarrival times are exponential (so $\alpha = 1$) and the service times are log normal with $\gamma_1^2 = \gamma_2^2 = 1$. The sample paths are generated using common input RVs for each routing policy. The parameters are $\tau(t) = (t-1)^+$, $\bar{\lambda} = 10$, $\hat{\lambda} = -5$, $\bar{\mu}_1 = 3$, $\hat{\mu}_1 = 0$, $\bar{\mu}_2 = 7$, $\hat{\mu}_2 = 0$ and $n = 25$.



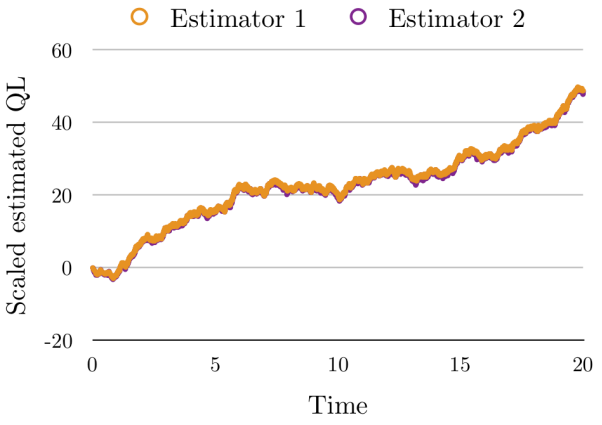
(a) $(\widehat{Q}_1^n, \widehat{Q}_2^n)$ under JSQ



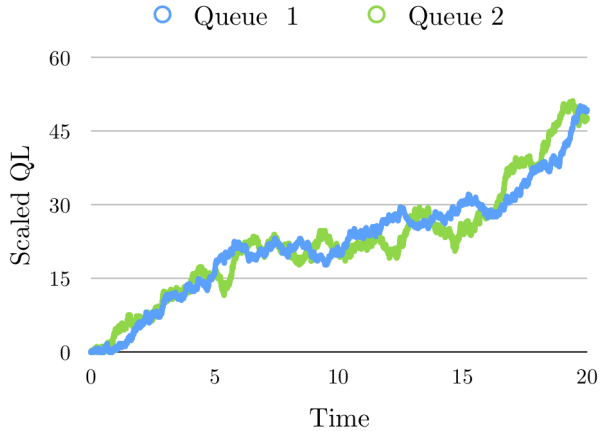
(b) $(\widehat{Q}_1^n, \widehat{Q}_2^n)$ under JSDQ



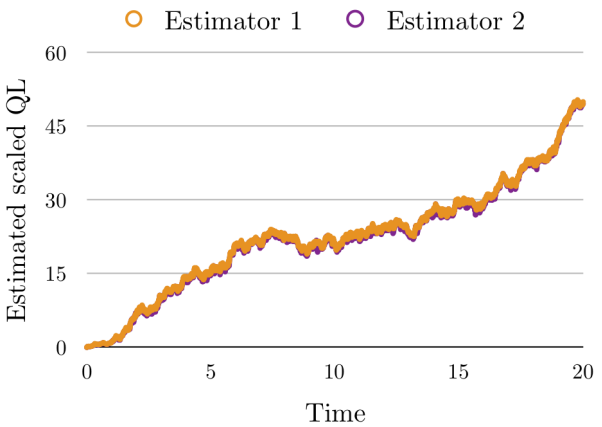
(c) $(\widehat{Q}_1^n, \widehat{Q}_2^n)$ under JSEQ(0)



(d) $(\widehat{R}_1^n, \widehat{R}_2^n)$ under JSEQ(0)



(e) $(\widehat{Q}_1^n, \widehat{Q}_2^n)$ under JSEQ(2)



(f) $(\widehat{R}_1^n, \widehat{R}_2^n)$ under JSEQ(2)

Figure 2: Sample paths for the scaled QLs $(\widehat{Q}_1^n, \widehat{Q}_2^n)$ and scaled estimated QLs $(\widehat{R}_1^n, \widehat{R}_1^n)$ (when applicable) under the following routing policies: JSQ, JSDQ, JSEQ(0), and JSEQ(2) with $\eta^{(2)} = .99$. The interarrival times are exponential (so $\alpha = 1$) and the service times are log normal with $\gamma_1^2 = \gamma_2^2 = 1$. The sample paths are generated using common input RVs for each routing policy. The parameters are $\tau(t) = (t-1)^+$, $\bar{\lambda} = 10$, $\hat{\lambda} = 5$, $\bar{\mu}_1 = 3$, $\hat{\mu}_1 = 0$, $\bar{\mu}_2 = 7$, $\hat{\mu}_2 = 0$ and $n = 25$.

$b < 0$, the scaled estimator under JSEQ(2) remains positive. (One can prove that the estimator under JSEQ(2) remains non-negative if $\eta^{(2)} = 1$. When $\eta^{(2)} < 1$, one can show that the estimator remains non-negative with probability converging to one as $\eta^{(2)} \uparrow 1$.)

To quantify the degree to which two parallel queues are load balanced, define

$$\Theta^{n,\gamma}(T) = \frac{1}{T}\int_0^T |\widehat{Q}_1^n(s) - \widehat{Q}_2^n(s)|^2 ds. \tag{71}$$

Then $\Theta^{n,\gamma}(T)$ serves as a measure of the average squared difference between the scaled QLs over the interval $[0, T]$. In Table 1 we evaluate the sample mean and 95% confidence interval for $\Theta^{25,\gamma}(100)$ obtained via simulation of 100 sample paths in the case of two parallel queues with exponential interarrival times and log normal service times operating under the JSEQ(0) policy or the JSEQ(2) policy, for varying values of b and $\gamma_1^2 = \gamma_2^2 = \sigma^2$, where we recall that γ_k^2 denotes the variance of the interarrival times of the renewal process S_k. We also include the values of $\Theta^{25,\gamma}(100)$ when operating under the JSQ and JSDQ policies for comparison. As seen in the table, both routing policies JSEQ(0) and JSEQ(2) significantly outperform JSDQ; however, there does not appear to be a statistically significant advantage to using the more refined estimator JSEQ(2) over JSEQ(0), even in the case that $b < 0$ and one expects the idleness process to have a larger effect on the dynamics.

8 Conclusion and open problems

We considered the problem of load balancing under delayed information on QL. Borrowing an idea from the theory of control under partial observations, we divided the problem into estimation and control parts. For the estimation part we allowed a broad set of QL estimators, whereas as control we chose to work with a policy that always routes to the shortest estimated queue. Our general results establish that the estimated QLs undergo SSC, but the QLs themselves do not. In the heavy traffic literature, SSC for QLs or workload processes has broadly been used in describing the limit diffusion process and as a tool for establishing convergence (see, e.g., [17]). However, SSC for estimators of QLs has not been considered before, to the best of our knowledge. Our results also prove weak limits of diffusion scaled QLs and estimated QL, identified as the unique solution to a diffusion model. The diffusion model itself builds on the SSC phenomenon, as one of its elements is an equation asserting that the limiting scaled estimated QLs are equal. There is novelty in this model, which accounts for both the QLs and their estimators, and in special cases gives rise to a stochastic delay equation with reflection, in which the reflection term appears with delay.

Although the conditional expectation estimator inspires several aspects of our development, we do not make any claim about its optimality. In fact, optimality with respect to a given criterion is a central open question about the model. We formulate it as follows. Given a positive constant β, consider the criterion

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \sum_{j,k} \mathbb{E}\left[|\widehat{Q}_k^n(t) - \widehat{Q}_j^n(t)|^\beta\right] dt,$$

where $\widehat{Q}^n$ denotes the scaled QL process when using an estimator $\widehat{\Phi}^n$. This criterion reflects an attempt to load balance, or minimize imbalance among the different QLs. Another criterion of interest may be the imbalance of delays associated with the different queues.

Table 1: 95% confidence intervals for $\mathbb{E}[\Theta^{25,\gamma}(100)]$, with exponential interarrival times and log-normal services times, operating under the JSQ, JSDQ, JSEQ(0), and JSEQ(2) with $\eta^{(2)} = .99$ policies, with $\tau(t) = (t-1)^+$, $\bar{\lambda} = 10$, $\bar{\mu}_1 = 3$, $\bar{\mu}_2 = 7$, for varying values of $b = \hat{\lambda} - \langle \hat{\mu}, \mathbf{1} \rangle$ and $\gamma_1 = \gamma_2 = \sigma$.

Estimator	b	σ^2	$\Theta^{25,\gamma}(100)$
JSQ	-5	.5	$.0438 \pm .0026$
JSDQ	-5	.5	295.1639 ± 31.0764
JSEQ(0)	-5	.5	$.6509 \pm .1953$
JSEQ(2)	-5	.5	$.5249 \pm .1793$
JSQ	-5	1	$.0625 \pm .0059$
JSDQ	-5	1	315.1499 ± 38.6355
JSEQ(0)	-5	1	$1.5770 \pm .5786$
JSEQ(2)	-5	1	$1.3080 \pm .5074$
JSQ	-5	2	$.1060 \pm .0162$
JSDQ	-5	2	335.6563 ± 44.9582
JSEQ(0)	-5	2	4.1222 ± 1.7280
JSEQ(2)	-5	2	3.5431 ± 1.8988
JSQ	0	.5	$.0523 \pm .0030$
JSDQ	0	.5	547.7839 ± 34.1452
JSEQ(0)	0	.5	$2.1006 \pm .5700$
JSEQ(2)	0	.5	$2.0976 \pm .5418$
JSQ	0	1	$.0775 \pm .0072$
JSDQ	0	1	576.1120 ± 44.7662
JSEQ(0)	0	1	4.4260 ± 1.1616
JSEQ(2)	0	1	4.3465 ± 1.2267
JSQ	0	2	$.1318 \pm .0167$
JSDQ	0	2	614.6724 ± 55.4677
JSEQ(0)	0	2	9.2948 ± 2.8597
JSEQ(2)	0	2	9.2317 ± 2.6948
JSQ	5	.5	$.0522 \pm .0036$
JSDQ	5	.5	544.2919 ± 38.4926
JSEQ(0)	5	.5	$2.0565 \pm .1428$
JSEQ(2)	5	.5	$2.0705 \pm .5885$
JSQ	5	1	$.0669 \pm .0056$
JSDQ	5	1	703.4523 ± 42.8280
JSEQ(0)	5	1	4.9691 ± 1.2040
JSEQ(2)	5	1	4.9618 ± 1.1919
JSQ	5	2	$.1097 \pm .0123$
JSDQ	5	2	753.4017 ± 49.4041
JSEQ(0)	5	2	10.6035 ± 2.4505
JSEQ(2)	5	2	10.2657 ± 2.5698

Problem 1. *Find a sequence of estimators* $\{\widehat{\Phi}^n\}_{n=1}^{\infty}$ *that asymptotically minimizes the above criterion in the limit* $n \to \infty$.

As a step toward solving Problem 1, we pose an analogous question for the diffusion model. Consider the criterion

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \sum_{j,k} \mathbb{E}\left[|\widehat{Q}_k(t) - \widehat{Q}_j(t)|^\beta\right] dt,$$

where $(\widehat{Q}, \widehat{A})$ denotes the solution to the diffusion model (Definition 4.1) associated with the estimator $\widehat{\Phi}$.

Problem 2. *Find an estimator* $\widehat{\Phi}$ *that minimizes the above criterion for the the diffusion model.*

An additional contribution of this paper is the detailed analysis of concrete estimators. One main conclusion from the simulation results is a clearly seen difference in performance between the delayed QL estimator and estimators which use the routing information; all our three examples of estimators use this information for QL estimation even if some of them do not use it for idleness estimation. It is evident that oscillations occur under delayed QL estimation, but not under the other estimators.

One may consider extending the approach presented in this paper. First, whereas the estimation stage is treated here in general, the load balancing stage is according JSQ. There are numerous other load balancing algorithms treated in the literature, most notably ones that only sample a subset of QLs and select the shortest (for example, as in [13]). One may ask for scaling limits under these various policies. Another valid problem is to view this model within the framework of optimal control under partial state observation. To this end, a criterion must be specified, and it could be the total QL or the degree to which queues are balanced. Particularly, it is interesting to ask if a form of a separation principle holds for the asymptotics of the control problem, asserting that estimation and load balancing can be treated as separate problems without loss of (asymptotic) optimality. The separation principle is well known to hold in other (non-asymptotic) control problems [5]. This question is of course in line with the approach presented in this paper.

There are several additional directions of interest when studying load balancing policies in the presence of delays. For example, while here we consider the case of a fixed number of queues, one may ask about the performance of the "join the shortest estimated queue" policy under heavy traffic scaling when the number of queues scale with n or $\sqrt{n}$ (and the exogeneous arrival rate scales with 1 or $\sqrt{n}$, respectively). There have been a number of recent works analyzing JSQ in these regimes when there are no messaging or routing delays (see [2, 4, 6] and the survey [18]). For further examples of interesting open problems related to load balancing in the presence of delays, see [10].

A One-dimensional Skorokhod problem

In this appendix we briefly review some well known properties of the one-dimensional Skorokhod problem. For proofs of the results here, see [3, Chapter 8].

Definition A.1. Given $x \in \mathbb{D}_+([0,\infty), \mathbb{R})$ we say that a pair $(z, y) \in \mathbb{D}([0,\infty), \mathbb{R}_+) \times \mathbb{D}_0([0,\infty), \mathbb{R}_+)$ satisfies the one-dimensional Skorokhod problem for x if the following conditions hold,

1. $z(t) = x(t) + y(t)$ for all $t \geq 0$;

2. y is non-decreasing and can increase only when z is zero, i.e., $\int_0^t z(s)dy(s) = 0$ for all $t \geq 0$.

Proposition A.2. *Given $x \in \mathbb{D}_+([0,\infty),\mathbb{R})$ there exists a unique solution (z,y) of the one-dimensional Skorokhod problem for x given by $(z,y) = (\Gamma_1,\Gamma_2)(x)$, where, for $t \geq 0$,*

$$\Gamma_1(x)(t) = x(t) + \Gamma_2(x)(t), \tag{72}$$

$$\Gamma_2(x)(t) = \sup_{0\leq s\leq t}(x(s))^-. \tag{73}$$

Consequently, the following properties hold:

1. *Oscillation inequality: Denote $Osc(f,[s,t]) = \sup\{|f(u) - f(r)| : s \leq r, u \leq t\}$. Then given $x \in \mathbb{D}_+([0,\infty),\mathbb{R})$ and $0 \leq s < t < \infty$,*

$$Osc(\Gamma_1(x),[s,t]) \leq Osc(x,[s,t]) \qquad \text{and} \qquad Osc(\Gamma_2(x),[s,t]) \leq Osc(x,[s,t]). \tag{74}$$

2. *Semigroup property: given $x \in \mathbb{D}_+([0,\infty),\mathbb{R})$ and $0 \leq s < t < \infty$,*

$$\Gamma_1(x)(t) = \Gamma_1(z^s)(t-s) \qquad \text{and} \qquad \Gamma_2(x)(t) = \Gamma_2(x)(s) + \Gamma_2(z^s)(t-s),$$

where $z^s(\cdot) = \Gamma_1(x)(s) + x(s+\cdot) - x(s)$.

3. *Lipschitz continuity: for $x_1, x_2 \in \mathbb{D}_+([0,\infty),\mathbb{R})$ and $t \geq 0$,*

$$\sup_{0\leq s\leq t}|\Gamma_1(x_1)(s) - \Gamma_1(x_2)(s)| \leq 2\sup_{0\leq s\leq t}|x_1(s) - x_2(s)|, \tag{75}$$

$$\sup_{0\leq s\leq t}|\Gamma_2(x_1)(s) - \Gamma_2(x_2)(s)| \leq \sup_{0\leq s\leq t}|x_1(s) - x_2(s)|. \tag{76}$$

Lemma A.3. *Let $x \in \mathbb{D}_+(\mathbb{R})$ and $(z,y) = (\Gamma_1,\Gamma_2)(x)$. Let $\tau : \mathbb{R}_+ \to \mathbb{R}_+$ be non-decreasing such that $t > \tau(t)$ for all $t \in \mathbb{R}_+$. Define $H : \mathbb{R}_+ \to \mathbb{R}_+$ by*

$$H(t) = \Gamma_2(z(\tau(t)) + x[\tau(t),\tau(t)+\cdot])(t - \tau(t)).$$

Then for $0 \leq s < t$,

$$Osc(H,[s,t]) \leq Osc(z,[\tau(s),\tau(t)]) + Osc(x,[\tau(s),\tau(t)]) + Osc(x,[\tau(s),\tau(t)]) \vee Osc(x,[s,t]).$$

Proof. Since τ is non-decreasing, we have $\tau(s) \leq \tau(t)$. By the definition of Γ_2 there exists $v_t \in [\tau(t),t]$ such that either

$$H(t) = [z(\tau(v_t)) + x(v_t) - x(\tau(v_t))]^-$$

or

$$H(t) = [z(\tau(v_t-)) + x(v_t-) - x(\tau(v_t-))]^- .$$

Let u_t stand for either v_t or v_t-, so that $H(t) = [z(\tau(u_t)) + x(u_t) - x(\tau(u_t))]^-$. Analogously, define $u_s \in [\tau(s),s]$. Then $x(u_t) \leq x(u)$ for all $u \in [\tau(t),t]$ and $x(u_s) \leq x(u)$ for all $u \in [\tau(s),s]$. It follows that

$$x(\tau(t) \vee u_s) - x(u_s) \leq x(u_t) - x(u_s) \leq x(u_t) - x(s \wedge u_t).$$

Thus,

$$\begin{aligned}|H(t) - H(s)| &= |(z(\tau(t)) + x[\tau(t),u_t])^- - (z(\tau(s)) - x[\tau(s),u_s])^-| \\ &\leq |z(\tau(t)) - z(\tau(s))| + \mathrm{Osc}(x,[\tau(s),\tau(t)]) + \mathrm{Osc}(x,[\tau(s),\tau(t)]) \vee \mathrm{Osc}(x,[s,t]).\end{aligned}$$

□

B Analysis of estimator functions

B.1 Proof of Lemma 3.10

Proof of Lemma 3.10. From the definitions for $\widehat{\Psi}^n$ and $\widehat{\Psi}$ and the condition $h(0,0,0) = 0$, we see that $\widehat{\Psi}^n(q,a)(0) = 0$ for all $(q,a) \in \mathbb{D}_0(\mathbb{R}_+^K) \times \mathbb{D}_0(\mathbb{R}^K)$ and $\widehat{\Psi}(q,a)(0) = 0$ for all $(q,a) \in \mathbb{C}_0(\mathbb{R}_+^K) \times \mathbb{C}_0(\mathbb{R}^K)$. Now let $\{(q^n,a^n)\}_{n=1}^\infty$ and (q,a) be as in parts (i)–(iii) of Assumption 3.5. Let $t < \infty$. By the definitions of $\widehat{\Psi}^n$ and $\widehat{\Psi}$, and the continuity condition (34), for each $s \in [0,t]$,

$$\begin{aligned}|\widehat{\Psi}_k^n(q^n,a^n)(s) - \widehat{\Psi}_k(q,a)(s)| &= |h(s,q^n(\tau(s)),a^n[\tau(s),s]) - h(s,q(\tau(s)),a[\tau(s),s])| \\ &\le |q_k^n(\tau(s)) - q_k(\tau(s))| + 2\eta \sup_{0\le u\le s} |a_k^n(u) - a_k(u)|.\end{aligned}$$

Since $\{(q^n,a^n)\}_{n=1}^\infty$ converges to (q,a) uniformly on compacts (see Remark 3.6), taking supremums over $s \in [0,t]$ on both sides of the previous display and letting $n \to \infty$ we see that Assumption 3.5 holds. Next, fix $T < \infty$ and let $s,t \in [0,T]$ satisfy $|s-t| \le \delta$. From the continuity condition (34), it follows that

$$w(\widehat{\Psi}^n(q,a),\delta,[0,T]) \le \kappa(\delta) + w(q(\tau(\cdot)),\delta,[0,T]) + \eta w(a\circ\tau,\delta,[0,T]) + \eta w(a,\delta,[0,T]),$$

and so Assumption 3.8 holds. Finally, given $q \in \mathbb{D}_0(\mathbb{R}_+^K)$, $a,a' \in \mathbb{D}_0(\mathbb{R}^K)$ and $t \ge 0$, by the Lipschitz condition (34),

$$|\widehat{\Psi}(q,a)(t) - \widehat{\Psi}(q,a')(t)| \le \eta |a[\tau(t),t] - a'[\tau(t),t]|.$$

Thus, Assumption 3.9 holds, which completes the proof. □

B.2 Proof of Lemma 7.1

We first need the following lemma.

Lemma B.1. *For $x \in \mathbb{R}_+$, $b \in \mathbb{R}$ and $\sigma^2 > 0$, let $(Z,Y) = \Gamma(x + b\iota + \sigma W)$. That is, Z be a one-dimensional reflected BM with initial condition x, drift b and variance σ^2, and Y is the associated local time process. Then, for $\theta \ge 0$,*

$$\mathbb{E}[Z(\theta)] = x + b\theta + g^{(2)}(x,b,\sigma,\theta), \tag{77}$$

$$\mathbb{E}[Y(\theta)] = g^{(2)}(x,b,\sigma,\theta), \tag{78}$$

where $g^{(2)}(x,b,\sigma,\theta)$ is defined as in (68).

Proof. By Equation (3.63) on [7, page 48],

$$\mathbb{E}[Z(\theta)] = \int_0^\infty \Upsilon\left(\frac{-u+x+bt}{\sigma\sqrt{t}}\right) du + \int_0^\infty \exp\left(\frac{2bu}{\sigma^2}\right) \Upsilon\left(\frac{-u-x-bt}{\sigma\sqrt{t}}\right) du,$$

where Υ denotes the cumulative distribution function for the standard normal distribution. For $r \in \mathbb{R}$ and $s > 0$, after substituting in with the definition of Υ, interchanging the order of integration

and using the fact that $\phi'(t) = -t\phi(t)$, we see that

$$\begin{aligned}\int_0^\infty \Upsilon\left(r - su\right) du &= \int_{-\infty}^{r} \phi(t) \int_0^{\frac{-t+r}{s}} du dt \\ &= \frac{r}{s}\int_{-\infty}^{r} \phi(t)dt - \frac{1}{s}\int_{-\infty}^{r} t\phi(t)dt \\ &= \frac{r}{s}\Upsilon(r) + \frac{1}{s}\phi(r).\end{aligned}$$

Next, for $q \neq 0$, $r \in \mathbb{R}$ and $s > 0$, after substituting in with the definition of Υ, interchanging the order of integration and a completion of squares argument, we obtain

$$\begin{aligned}\int_0^\infty \exp(qu)\Upsilon\left(-r - su\right) du &= \int_{-\infty}^{-r} \phi(t) \int_0^{-\frac{t+r}{s}} \exp(qu) du dt \\ &= \frac{1}{q}\exp\left(\frac{q^2}{2s^2} - \frac{qr}{s}\right)\int_{-\infty}^{-r} \phi\left(t + qs^{-1}\right) dt - \frac{1}{q}\int_{-\infty}^{-r}\phi(t)dt \\ &= \frac{1}{q}\exp\left(\frac{q^2}{2s^2} - \frac{qr}{s}\right)\Upsilon(-r + qs^{-1}) - \frac{1}{q}\Upsilon(-r).\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[Z(\theta)\right] = (x + b\theta)\Upsilon\left(\frac{x + b\theta}{\sigma\sqrt{\theta}}\right) &+ \sigma\sqrt{\theta}\phi\left(\frac{x + b\theta}{\sigma\sqrt{\theta}}\right) \\ &+ 1_{\{b\neq 0\}}\frac{\sigma^2}{2b}\left[\exp\left(-\frac{2bx}{\sigma^2}\right)\Upsilon\left(\frac{-x + b\theta}{\sigma\sqrt{\theta}}\right) - \Upsilon\left(\frac{-x - b\theta}{\sigma\sqrt{\theta}}\right)\right] \\ &+ 1_{\{b=0\}}\left[\sigma\sqrt{\theta}\phi\left(-\frac{x}{\sigma\sqrt{\theta}}\right) - x\Upsilon\left(-\frac{x}{\sigma\sqrt{t}}\right)\right].\end{aligned}$$

Since $Z(\theta) = x + b\theta + \sigma W(\theta) + Y(\theta)$ by definition and $E[W(\theta)] = 0$, the expressions (77)–(78) follow. □

Proof of Lemma 7.1. The explicit expression for $h^{(2)}$ follows from the fact that $h_k^{(2)}(t, v, u)$ is equal to $\eta^{(2)}$ multiplied by the expected local time of a reflected BM over the interval $[0, t - \tau_k(t)]$ with initial condition v_k, drift $\frac{u_k}{t-\tau_k(t)}$, variance $\bar{\mu}_k\gamma_k^2$.

Fix $T < \infty$. First observe that by definition $h^{(2)}(0, 0, 0) = 0$. Next, let $t, t' \in [0, T]$, $v, v' \in \mathbb{R}_+^K$ and $u, u' \in \mathbb{R}^K$. Then

$$|h_k^{(2)}(t, v, u) - h_k^{(2)}(t', v', u')| \leq |h_k^{(2)}(t, v, u) - h_k^{(2)}(t', v, u)| + |h_k^{(2)}(t', v, u) - h_k^{(2)}(t', v', u')|.$$

We treat the two terms on the RHS of the previous display separately. By the definition of $h_k^{(2)}$ and the Lipschitz(1) continuity of Γ_2,

$$|h_k^{(2)}(t', v, u) - h_k^{(2)}(t', v', u')| \leq \eta^{(2)}|v_k - v_k'| + \eta^{(2)}|u_k - u_k'|.$$

We are left to show there exists $\kappa : [0, T] \to \mathbb{R}_+$ continuous and non-decreasing satisfying $\kappa(0) = 0$ and such that the first term is bounded by $\kappa(|t - t'|)$. By the definition of $h^{(2)}$ and the explicit

expression for Γ_2,

$$
\begin{aligned}
&|h_k^{(2)}(t,v,u) - h_k^{(2)}(t',v,u)| \\
&\quad \le \eta^{(2)}\mathbb{E}\left[\left|\sup_{0\le s\le t-\tau_k(t)}\left(v_k + \frac{u_k s}{t-\tau_k(t)} - \widehat{S}_k(s)\right)^- - \sup_{0\le s\le t'-\tau_k(t')}\left(v_k + \frac{u_k s}{t'-\tau_k(t')} - \widehat{S}_k(s)\right)^-\right|\right]
\end{aligned}
$$

Since ι and $\widehat{S}_k$ are continuous, there are (random) times $\rho \in [0, t-\tau_k(t)]$ and $\rho' \in [0, t'-\tau_k(t')]$ such that

$$
\begin{aligned}
\sup_{0\le s\le t-\tau_k(t)}\left(v_k + \frac{u_k s}{t-\tau_k(t)} - \widehat{S}_k(s)\right)^- &= \left(v_k + \frac{u_k \rho}{t-\tau_k(t)} - \widehat{S}_k(\rho)\right)^-, \\
\sup_{0\le s\le t'-\tau_k(t')}\left(v_k + \frac{u_k s}{t'-\tau_k(t')} - \widehat{S}_k(s)\right)^- &= \left(v_k + \frac{u_k \rho'}{t'-\tau_k(t')} - \widehat{S}_k(\rho')\right)^-.
\end{aligned}
$$

Note that

$$
\frac{\rho'}{t'-\tau_k(t')}(t-\tau_k(t)) \in [0, t-\tau_k(t)] \qquad \text{and} \qquad \frac{\rho}{t-\tau_k(t)}(t'-\tau_k(t')) \in [0, t'-\tau_k(t')],
$$

and so the following inequalities hold:

$$
\begin{aligned}
\left(v_k + \frac{u_k \rho'}{t'-\tau_k(t')} - \widehat{S}_k\left(\frac{t-\tau_k(t)}{t'-\tau_k(t')}\rho'\right)\right)^- &\le \left(v_k + \frac{u_k \rho}{t-\tau_k(t)} - \widehat{S}_k(\rho)\right)^-, \\
\left(v_k + \frac{u_k \rho}{t-\tau_k(t)} - \widehat{S}_k\left(\frac{t'-\tau_k(t')}{t-\tau_k(t)}\rho\right)\right)^- &\le \left(v_k + \frac{u_k \rho'}{t'-\tau_k(t')} - \widehat{S}_k(\rho')\right)^-.
\end{aligned}
$$

Thus, we have the following string of inequalities:

$$
\begin{aligned}
&\left(v_k + \frac{u_k \rho'}{t'-\tau_k(t')} - \widehat{S}_k\left(\frac{t-\tau_k(t)}{t'-\tau_k(t')}\rho'\right)\right)^- - \left(v_k + \frac{u_k \rho'}{t'-\tau_k(t')} - \widehat{S}_k(\rho')\right)^- \\
&\quad \le \left(v_k + \frac{u_k \rho}{t-\tau_k(t)} - \widehat{S}_k(\rho)\right)^- - \left(v_k + \frac{u_k \rho'}{t'-\tau_k(t')} - \widehat{S}_k(\rho')\right)^- \\
&\quad\quad \le \left(v_k + \frac{u_k \rho}{t-\tau_k(t)} - \widehat{S}_k(\rho)\right)^- - \left(v_k + \frac{u_k \rho}{t-\tau_k(t)} - \widehat{S}_k\left(\frac{t'-\tau_k(t')}{t-\tau_k(t)}\rho\right)\right)^-.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\left|\sup_{0\le s\le t-\tau_k(t)}\left(v_k + \frac{u_k s}{t-\tau_k(t)} - \widehat{S}_k(s)\right)^- - \sup_{0\le s\le t'-\tau_k(t')}\left(v_k + \frac{u_k s}{t'-\tau_k(t')} - \widehat{S}_k(s)\right)^-\right| \\
&\quad \le w(\widehat{S}_k, |t-t'|, T) + (\widehat{S}_k \circ \tau_k, |t-t'|, T),
\end{aligned}
$$

where we have used the fact that

$$
\max\left(\left|\frac{t-\tau_k(t)}{t'-\tau_k(t')}\rho' - \rho'\right|, \left|\rho - \frac{t'-\tau_k(t')}{t-\tau_k(t)}\rho\right|\right) \le |t-t'| + |\tau_k(t) - \tau_k(t')|.
$$

It follows that $|h_k^{(2)}(t,v,u) - h_k^{(2)}(t',v,u)| \le \kappa(|t-t'|)$, where, for $\delta \in [0,T]$,

$$
\kappa(\delta) = \eta^{(2)}\mathbb{E}\left[w(\widehat{S}_k, \delta, [0,T])\right] + \eta^{(2)}\mathbb{E}\left[w(\widehat{S}_k \circ \tau, \delta, [0,T])\right], \qquad \delta > 0.
$$

Since $\widehat{S}_k$ and τ_k are continuous, it follows that $\kappa : [0,T] \to \mathbb{R}_+$ is continuous, non-decreasing and satisfies $\kappa(0) = 0$. This completes the proof that (34) holds. □

B.3 Proof of Lemma 7.2

Proof of Lemma 7.2. Let $\{(q^n, a^n)\}_{n=1}^{\infty}$ and (q, a) be as in parts (i)–(iii) of Assumption 3.5. Let $t < \infty$. By the Lipschitz(1) continuity of the SM Γ_2, for each $s \in [0, t]$ and $k \in \mathbb{K}$,

$$|\widehat{\Psi}_k^n(q^n, a^n)(s) - \widehat{\Psi}_k(q, a)(s)| \leq \eta^{(3)}|q^n(\tau(s)) - q(\tau(s))| + 2\eta^{(3)} \sup_{0 \leq u \leq s} |a^n(u) - a(u)|.$$

Taking supremums over $s \in [0, t]$ on both sides and letting $n \to \infty$ we see that Assumption 3.5 holds. Next, fix $T < \infty$ and let $s, t \in [0, T]$ satisfy $|s - t| \leq \delta$. By Lemma A.3,

$$\begin{aligned} w(\widehat{\Psi}^n(q, a), \delta, [0, T]) \leq \eta^{(3)} w(q(\tau(\cdot)), \delta, [0, T]) + 2\eta^{(3)} w(a \circ \tau, \delta, [0, T]) + \eta^{(3)} w(a, \delta, [0, T]) \\ + \eta^{(3)} \mathbb{E}\left[w(\widehat{S} \circ \tau, \delta, [0, T])\right] + \eta^{(3)} \mathbb{E}\left[w(\widehat{S}, \delta, [0, T])\right]. \end{aligned}$$

Thus, Assumption 3.8 holds with $\zeta_1 = \eta^{(3)}$ and

$$\kappa(\delta) = \eta^{(3)} \mathbb{E}\left[w(\widehat{S}_k, \delta, [0, T])\right] + \eta^{(3)} \mathbb{E}\left[w(\widehat{S}_k \circ \tau, \delta, [0, T])\right], \qquad \delta > 0.$$

Finally, given $q \in \mathbb{D}_0(\mathbb{R}_+^K)$, $a, a' \in \mathbb{D}_0(\mathbb{R}^K)$ and $t \geq 0$, by the Lipschitz(1) continuity of the SM Γ_2,

$$|\widehat{\Psi}(q, a)(t) - \widehat{\Psi}(q, a')(t)| \leq \eta^{(3)}|a[\tau(t), t] - a'[\tau(t), t]|.$$

Thus, Assumption 3.9 holds with $\zeta_2 = \eta^{(3)}$, which completes the proof. □

B.4 Proof of Lemma 7.3

Proof of Lemma 7.3. Define the processes Y_k^n and Z_k^n on $[0, t - \tau_k(t)]$ by

$$Y_k^n(s) = q_k(\tau_k(t)) - b_k s - \widehat{S}_k(s), \qquad s \in [0, t - \tau_k(t)],$$

and

$$Z_k^n(s) = Y_k^n(s) + \sum_{j=1}^{\infty} c_j 1_{\{s \geq t_j\}}, \qquad s \in [0, t - \tau_k(t)].$$

Then $F_k(q, a)(t) = \mathbb{E}\left[\Gamma_2(Z_k^n)(t - \tau_k(t))\right]$. Set $t_0 = 0$ and $t_{J+1} = t - \tau_k(t)$. Since the jumps of Z_k^n are positive, for each $j = 1, \ldots, J + 1$,

$$\Gamma_1(Z_k^n)(t_j) = \Gamma_1(Z_k^n)(t_j-) + a_k(t_j) - a_k(t_j-).$$

Furthermore, by the semigroup property of the SM Γ_1,

$$\Gamma_1(Z_k^n)(t_j-) = \Gamma_1\left(\Gamma_1(Z_k^n) + Y_k^n(t_j + \cdot) - Y_k^n(t_j)\right)(t_j - t_{j-1}).$$

Thus

$$\mathbb{P}\left(\Gamma_1(Z_k^n)(t_j-) \in dz_j\right) = \pi_k(\Gamma_1(Z_k^n)(t_{j-1}), t_j - t_{j-1}; z_j) dz_j.$$

It follows that

$$\begin{aligned} \mathbb{P}\left(\Gamma_1(Z_k^n)(t_j) \in dz_j\right) &= \mathbb{P}\left(\Gamma_1(Z_k^n)(t_j-) + c_j \in dz_j\right) \\ &= \pi_k(\Gamma_1(Z_k^n)(t_{j-1}), t_j - t_{j-1}; z - c_j) dz_j. \end{aligned}$$

Therefore, recursively applying the last display, we obtain

$$\begin{aligned}\mathbb{P}(\Gamma_1(Z_k^n)(t-\tau_k(t)) \in dz) &= \mathbb{P}(\Gamma_1(Z_k^n)(t_{J+1}) \in dz) \\ &= \Big[\int_{c_1}^{\infty} dz_1 \cdots \int_{c_J}^{\infty} dz_J \prod_{j=1}^{J} \pi(z_{j-1}, t_j - t_{j-1}; z_j - c_j) \\ &\qquad \times \pi(z_J, t - t_J; z - a(t) + a(t-)) \Big] dz.\end{aligned}$$

The lemma then follows from the facts that $F_k(q,a)(t) = \mathbb{E}\left[\Gamma_2(Z_k^n)(t-\tau_k(t))\right]$ and $\Gamma_2(Z_k^n)(t-\tau_k(t)) = \Gamma_1(Z_k^n)(t-\tau_k(t)) - Z_k^n(t-\tau_k(t))$, and above expression for the transition density of $\Gamma_1(Z_k^n)(\cdot)$ and time $t-\tau_k(t)$. □

Acknowledgements. We thank the anonymous referees for their many helpful comments.

References

[1] P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, Inc., New York, 1999.

[2] A. Braverman. Steady-state analysis of the join the shortest queue model in the halfin-whitt regime. working paper. 2018. arXiv preprint arXiv:1801.05121.

[3] K. L. Chung and R. J. Williams. *Introduction to Stochastic Integration*. Birkhäuser, Boston, second edition, 1990.

[4] P. Eschenfeldt and D. Gamarnik. Join the shortest queue with many servers. the heavy traffic asymptotics. *Mathematics of Operations Research*, 43(3):867–886, 2018.

[5] W. H. Fleming and R. W. Rishel. *Deterministic and Stochastic Optimal Control*. Springer-Verlag, Berlin-New York, 1975. vii+222 pp. Applications of Mathematics, No. 1.

[6] V. Gupta and N. Walton. Load balancing in the nondegenerate slowdown regime. *Operations Research*, 67(1):281–294, 2019.

[7] J. M. Harrison. *Brownian models of performance and control*. Cambridge University Press, 2013.

[8] M. S. Kinnally and R. J. Williams. On existence and uniqueness of stationary distributions for stochastic delay differential equations with positivity constraints. *Electron. J. Probab.*, 15: no. 15, 409–451, 2010. ISSN 1083-6489. URL `https://doi.org/10.1214/EJP.v15-756`.

[9] H. J. Kushner. *Numerical Methods for Controlled Stochastic Delay Systems*. Systems & Control: Foundations & Applications. Birkhäuser Boston, Inc., Boston, MA, 2008. ISBN 978-0-8176-4534-2. xx+281 pp. URL `https://doi.org/10.1007/978-0-8176-4621-9`.

[10] D. Lipshutz. Open problem—load balancing using delayed information. *Stochastic Systems*, 9 (3):305–306, 2019.

[11] D. Lipshutz and R. J. Williams. Existence, uniqueness, and stability of slowly oscillating periodic solutions for delay differential equations with nonnegativity constraints. *SIAM Journal on Mathematical Analysis*, 47(6):4467–4535, 2015.

[12] X. Mao. *Exponential Stability of Stochastic Differential Equations*, volume 182 of *Monographs and Textbooks in Pure and Applied Mathematics*. Marcel Dekker, Inc., New York, 1994. ISBN 0-8247-9080-4. xii+307 pp.

[13] M. Mitzenmacher. How useful is old information? *IEEE Transactions on Parallel and Distributed Systems*, 11:6–20, 2000.

[14] S. E. A. Mohammed. *Stochastic Functional Differential Equations*, volume 99 of *Research Notes in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1984. ISBN 0-273-08593-X. vi+245 pp.

[15] J. Pender, R. H. Rand, and E. Wesson. Queues with choice via delay differential equations. *International Journal of Bifurcation and Chaos*, 27(4):1730016, 2017.

[16] J. Pender, R. H. Rand, and E. Wesson. A stochastic analysis of queues with customer choice and delayed information. Preprint, 2018.

[17] M. I. Reiman. Some diffusion approximations with state space collapse. In *Modelling and performance evaluation methodology*, pages 207–240. Springer, 1984.

[18] M. van der Boor, S. C. Borst, J. S. van Leeuwaarden, and D. Mukherjee. Scalable load balancing in networked systems: A survey of recent advances. arXiv preprint arXiv:1806.05444, 2018.