The Annals of Applied Probability
2014, Vol. 24, No. 2, 760–810
DOI: 10.1214/13-AAP935

SCHEDULING PARALLEL SERVERS IN THE NONDEGENERATE SLOWDOWN DIFFUSION REGIME: ASYMPTOTIC OPTIMALITY RESULTS[1]

BY RAMI ATAR AND ITAI GURVICH

Technion–Israel Institute of Technology and Northwestern University

We consider the problem of minimizing queue-length costs in a system with heterogenous parallel servers, operating in a many-server heavy-traffic regime with nondegenerate slowdown. This regime is distinct from the well-studied heavy traffic diffusion regimes, namely the (single server) conventional regime and the (many-server) Halfin–Whitt regime. It has the distinguishing property that waiting times and service times are of comparable magnitudes. We establish an asymptotic lower bound on the cost and devise a sequence of policies that asymptotically attain this bound. As in the conventional regime, the asymptotics can be described by means of a Brownian control problem, the solution of which exhibits a state space collapse.

1. Introduction. Many-server approximations are ubiquitous in the modeling of large-scale service systems. A prevalent mode of analysis in this context is the *Halfin–Whitt* heavy traffic diffusion regime [18], also called the *quality and efficiency driven* (QED) regime [16]. For the $M/M/N$ queue, a sequence of systems in heavy traffic (HT), indexed by n, is constructed by letting the number of servers, N^n, and the arrival rate, λ^n, grow with n while the service rate μ remains fixed, so that the utilization in the nth system, $\rho^n := \lambda^n/(N^n\mu^n)$, behaves like

$$\rho^n = 1 - O\left(\frac{1}{\sqrt{\lambda^n}}\right) = 1 - O\left(\frac{1}{\sqrt{N^n}}\right).$$

Customer waiting times in this regime are of the order $1/\sqrt{\lambda^n}$ and are thus order of magnitudes smaller than the service times. It has been argued that this order of magnitude relation renders the analysis in this regime relevant for some call centers and certain health-care systems to whose study it has been applied; see, for example, [1, 21]. The Halfin–Whitt regime is typically contrasted with the so-called *conventional* HT diffusion regime. Conventional limits are obtained by fixing the number of servers (typically 1) and letting both the arrival and service

Received November 2011; revised December 2012.

[1]Supported in part by the US–Israel BSF (Grant 2008466), the ISF (Grant 1349/08) and the Technion's fund for promotion of research.

MSC2010 subject classifications. 60K25, 60J60, 60F17, 90B22, 68M20.

Key words and phrases. The parallel server model, many-server queues, heavy traffic, diffusion limits, asymptotically optimal control, nondegenerate slowdown regime.

rate scale so that

$$\rho^n = 1 - O\left(\frac{1}{\sqrt{\lambda^n}}\right) = 1 - O\left(\frac{1}{\sqrt{\mu^n}}\right).$$

In this regime, the waiting time is of the order of $1/\sqrt{\mu^n}$ so that, in perfect contrast with the Halfin–Whitt regime, the service time is negligible compared to the waiting time.

From a modeling viewpoint, it is desirable to allow for these two important performance measures to be comparable under the scaling. Gurvich, Mandelbaum and Shaikhet [17, 20, 22] and, independently, Whitt [25] have identified a many-server regime with this property. Limits for the $M/M/N$ queue in this regime are obtained by scaling the parameters so that μ^n and N^n are of the order of $\sqrt{\lambda^n}$ and

$$\rho^n = 1 - O\left(\frac{1}{\sqrt{\lambda^n}}\right) = 1 - O\left(\frac{1}{\sqrt{\mu^n N^n}}\right) = 1 - O\left(\frac{1}{N^n}\right). \tag{1}$$

One defines the *slowdown* of a queueing system as the ratio between the sojourn time and the service time experienced by a typical customer. By the foregoing discussion among the three regimes alluded to above, regime (1) is unique in that the slowdown does not degenerate, in the sense that it does not converge to one of the extreme values, 1 or ∞. We therefore refer to it as the *nondegenerate slowdown* (NDS) diffusion regime. This term was coined in [3], where a queue with heterogenous servers was analyzed in this regime, and the limiting joint law of waiting time and service time was identified. Both the conventional and the NDS regimes are often referred to as *efficiency driven* (ED). We refer the reader to [3] for further discussion of the three regimes and to [16] for the distinction between QED and ED regimes. The relevance of the NDS diffusion regime in real-world applications has been argued in [3] by demonstrating that some call centers do operate with comparable delays and service times; particularly, this is the case for the detailed empirical study of a call center performed in [11]. This makes a strong case for analyzing these systems by NDS diffusion approximations (see [3] for further discussion on this modeling issue, and Whitt [26] for an alternative (ED) regime with comparable delays and service times).

Control of queueing networks under both the conventional and Halfin–Whitt diffusion regimes (as well as fluid regimes) has been an active research area in recent years. Particularly, the *parallel server model* has been studied in this context, where customers of a number of classes are served in parallel by servers of various types, and a system administrator dynamically controls the routing. (For the conventional diffusion regime see [7, 8, 19, 23] and for the Halfin–Whitt regime see, e.g., [24] and references therein.) In this paper we study the problem of minimizing queue-length costs in a parallel server model with renewal arrival processes and exponential service times operating in the NDS regime. From a control standpoint, a distinctive property of the NDS regime is that *sojourn time* cost criteria are meaningful as neither the service time or waiting-time degenerate asymptotically.

Having solved the queue-length problem, we argue (heuristically) how in a simple case, the sojourn time problem can be reduced to a queue-length problem. This provides further motivation for the latter. We leave open the rigorization of this argument and the question of how general this reduction is, as well as the extension of this work to general service-time distributions.

In terms of the asymptotic behavior, the NDS control problem is close to the one in the conventional regime. In particular, the Brownian control problem (BCP) which describes the asymptotics is the same as the one studied in [19] for the conventional regime. This BCP, under a *complete resource pooling* (CRP) condition, undergoes a reduction to a one-dimensional problem. This reduction is often called a *state space collapse*. Bell and Williams [7, 8] and Mandelbaum and Stolyar [23] have studied the parallel server model in conventional heavy traffic and obtained asymptotic optimality results under the CRP condition. Bell and Williams address linear costs and construct certain threshold type policies. Mandelbaum and Stolyar consider separable convex (and flat at the origin) costs and work with policies that obey the generalized-$c\mu$ rule.

In this paper we aim at a relatively general cost of the form $\int_0^u C(\hat{Q}^n(t))\,dt$, where we denote by $\hat{Q}^n$ a properly scaled version of queue-length, and use the term "cost" to mean a random variable that is to be minimized stochastically. The function C, referred to as the cost function, satisfies an assumption slightly weaker than convexity (Assumption 2.3), as well as an assumption regarding the existence of a continuous minimizer (Assumption 5.1). The first main result (Theorem 2.1) asserts that the BCP value constitutes an asymptotic lower bound on the cost under any sequence of policies. The second main result (Theorem 5.1) shows that this lower bound is tight. The price paid for the generality of C is that $c\mu$ type policies must be abandoned, and a more complicated policy has to be used to attain the lower bound. We introduce a *tracking* type policy, in which a certain target process [denoted by $\check{X}^n$ in (82)] is computed, and routing is performed so as to keep the difference between the actual queue-length process $\hat{Q}^n$ and the target, small. It turns out that the techniques required here are quite different from those in the conventional HT, due both to the general cost and the difference between the regimes.

Theorem 2.1 provides a lower bound that is weaker than what, in the conventional regime, is referred to as a *pathwise* bound (as, e.g., in [23]). As we discuss in Section 2, in the NDS regime a corresponding pathwise lower bound does not hold, and this is the main reason for the complicated proof of the result as compared to that of the pathwise lower bound in the conventional regime.

The main part of the proof of asymptotic optimality of the proposed policy is showing that the difference $\hat{Q}^n - \check{X}^n$ is small. (From that, asymptotic optimality follows rather easily because the process $\hat{Q}^n$ can then be shown to behave like the explicit solution to the BCP.) This proof is based on showing that re-balancing of the workload among the queues can occur quickly on the relevant time scale. This

is facilitated by the fact that the workload is evenly divided (in the sense of order of magnitude) between the queues and the servers and short service times allow to move significant workload from one queue to the other before the total-workload process changes considerably. (This explains the similarity to the behavior in conventional heavy traffic, where the same quick response is possible. On the other hand, such nearly instantaneous re-balancing cannot be performed in the Halfin–Whitt regime in which most of the workload is in service and, to re-direct a non-negligible fraction of workload from one class to the other requires a large number of service completions. Indeed, the resulting Brownian control problem there is different [2].)

An analysis of the problem in the case of homogenous servers and interruptible service policies was carried out in [6]. The asymptotics of the queueing control problem were shown to be governed by a BCP that is a special case of the one identified in this paper. Thanks to the more special model and the (easier to handle) interruptible service assumption, it was possible to attain an analogous result for the *headcount* process as well as queue-length (in this paper, Theorem 2.1 is proved for the queuelength process only).

We use the following notation throughout the paper. For a positive integer d and $x \in \mathbb{R}^d$, we write $\|x\|$ for $\sum_{l=1}^{d} |x_l|$, and for $f : [0, \infty) \to \mathbb{R}^d$, $\|f\|_t = \sup_{0 \le s \le t} \|x(s)\|$. We denote by $\mathcal{D}_{\mathbb{R}^d}$ the space of functions from $[0, \infty)$ to $\mathbb{R}^d$ that are right continuous with left limits (RCLL), and equip it with the usual Skorohod topology. We remove the subscript when $d = 1$. For a sequence of r.v.'s $\{X^n\}$, X, with values in $\mathbb{R}^d$, or processes with sample paths in $\mathcal{D}_{\mathbb{R}^d}$, $X^n \Rightarrow X$ denotes convergence in distribution.

The remainder of the paper is organized as follows. Section 2 describes the model and states the result regarding the lower bound. It also contains a discussion of sojourn time costs as well as an aspect of the lower bound related to pathwise dominance. The definitions of various diffusion-scaled processes and some preliminary lemmas appear in Section 3. This section also contains a formulation and solution of the underlying Brownian control problem. The proof of the lower bound appears in Section 4. Section 5 contains the second main result, asserting that the lower bound is tight. This is shown by constructing a sequence of policies that asymptotically achieves the bound.

2. The model and a lower bound on performance.

2.1. *Model*, *scaling*, *heavy traffic assumptions*. We consider a sequence, indexed by $n \in \mathbb{N}$, of parallel server systems with $\mathbf{I}$ classes of customers and $\mathbf{J}$ pools of servers, an example of which is depicted in Figure 1. The index set for the classes is denoted by $\mathcal{I}$ and that for the pools by $\mathcal{J}$ (thus $|\mathcal{I}| = \mathbf{I}$ and $|\mathcal{J}| = \mathbf{J}$).

Arrivals are modeled as independent renewal processes, denoted by $(A_i^n, i \in \mathcal{I})$ so that $A_i^n(t)$ is the number of class-i arrivals by time t. To construct these processes, let $(A_i, i \in \mathcal{I})$ be independent (undelayed) renewal processes, where,

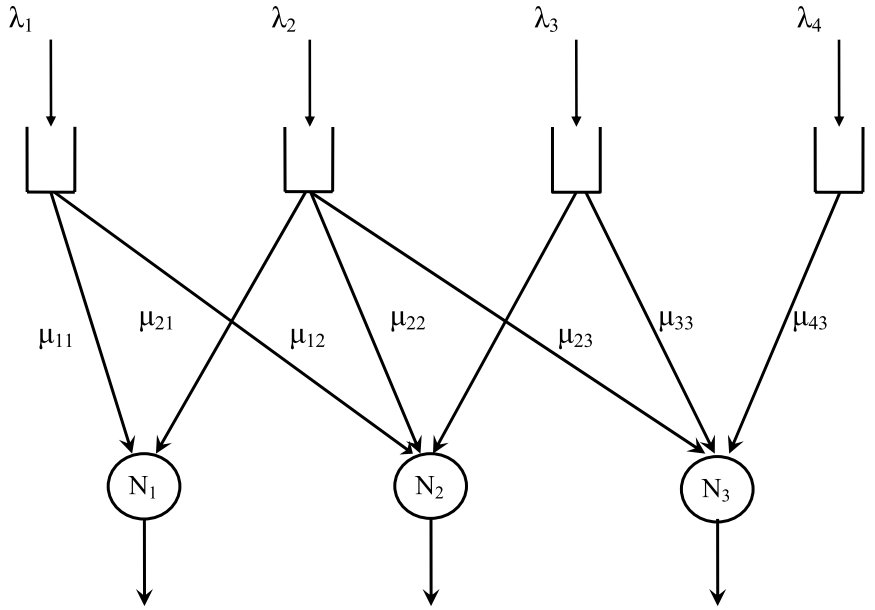

FIG. 1. *A parallel server system.*

for each i, the time of the first arrival and the inter-arrival times are positive i.i.d. random variables with mean 1, variance $(C_{i,IA})^2 \geq 0$. We assume that the inter-arrival times have a finite moment of some order $r > 2$. The processes A_i^n are defined as time accelerations of the above, namely

$$A_i^n(t) = A_i(\lambda_i^n t), \qquad t \geq 0, i \in \mathcal{I},$$

where the acceleration parameters satisfy $\lim_n \lambda_i^n/n = \lambda_i > 0$, and

$$\hat{\lambda}_i^n := n^{-1/2}(\lambda_i^n - n\lambda_i) \to \hat{\lambda}_i \in (-\infty, \infty), \qquad i \in \mathcal{I} \tag{2}$$

as $n \to \infty$.

For $j \in \mathcal{J}$, N_j^n denotes the number of servers in pool j and is assumed to satisfy

$$N_j^n = \nu_j n^{1/2} + O(n^{1/4}), \qquad j \in \mathcal{J} \tag{3}$$

for some constants $\nu_j > 0$. We assume that service times are exponentially distributed and denote by $1/\mu_{ij}^n$ the mean service time of a class-i customer with a server from pool j (so that μ_{ij}^n is interpreted as the corresponding rate). If servers from pool j cannot serve customers from class i, write $\mu_{ij}^n = 0$. This property is assumed to be independent of n. Write $i \sim j$ or $j \sim i$ if $\mu_{ij}^n > 0$ (for all, equivalently, one n). This information is encoded in a graph $\mathcal{G}$ whose vertex set is $\mathcal{I} \cup \mathcal{J}$ and has an edge connecting $i \in \mathcal{I}$ and $j \in \mathcal{J}$ if and only $i \sim j$. The edge set of the graph is denoted by $\mathcal{E}$, and thus $(i, j) \in \mathcal{E}$ if and only if $i \sim j$. Denote by $\mathbf{K}$ the

cardinality of $\mathcal{E}$. When all servers from pool j are occupied with class-i customers, they jointly serve this class at rate $N_j^n \mu_{ij}^n$. Further assume that

$$\mu_{ij}^n = \mu_{ij} n^{1/2} + O(n^{1/4}), \tag{4}$$

so that

$$\bar{\mu}_{ij}^n := n^{-1} N_j^n \mu_{ij}^n \to \bar{\mu}_{ij} := \mu_{ij} \nu_j \qquad \text{as } n \to \infty, \tag{5}$$

and assume that $\bar{\mu}_{ij} > 0$ whenever $i \sim j$ (clearly, $\bar{\mu}_{ij} = 0$ otherwise). Also, assume that, for every $i \in \mathcal{I}$ and $j \in \mathcal{J}$,

$$\hat{\mu}_{ij}^n := n^{1/2}(\bar{\mu}_{ij}^n - \bar{\mu}_{ij}) \to \hat{\mu}_{ij} \in (-\infty, \infty) \qquad \text{as } n \to \infty. \tag{6}$$

Thus, the nominal joint processing rate of pool-j servers for class-i customers (namely $N_j^n \mu_{ij}^n$) is asymptotic to $n\bar{\mu}_{ij}$. At the same time, the rate of an individual server (namely μ_{ij}^n) is asymptotic to $n^{1/2}\mu_{ij}$. The quantities $\bar{\mu} = (\bar{\mu}_{ij})$ will appear in the fluid model, while $\mu = (\mu_{ij})$ will show in the diffusion model.

Let Σ denote the set of $\mathbf{I} \times \mathbf{J}$ matrices for which the (i, j) entry is zero whenever $i \nsim j$. Let Ξ be the subset of Σ of "column-substochastic" matrices. That is, members ξ of Ξ satisfy $\xi_{ij} \geq 0$ for every i, j, $\sum_i \xi_{ij} \leq 1$ for every j and $\xi_{ij} = 0$ for $(i, j) \notin \mathcal{E}$. Following [23], for $\xi \in \Sigma$, write $\bar{\mu}(\xi)$ for the column vector $(\bar{\mu}(\xi)_1, \bar{\mu}(\xi)_2, \ldots, \bar{\mu}(\xi)_{\mathbf{I}})'$, where

$$\bar{\mu}(\xi)_i = \sum_{j \in \mathcal{J}} \bar{\mu}_{ij} \xi_{ij}, \qquad i \in \mathcal{I}. \tag{7}$$

The first order parameters $\lambda = (\lambda_i)$ and $\bar{\mu} = (\bar{\mu}_{ij})$ are assumed to satisfy a critical load condition. To specify it, consider a *static fluid model*, consisting of $\mathbf{I}$ classes of fluid and $\mathbf{J}$ processing stations. For $i \in \mathcal{I}$, fluid of class i enters at rate λ_i. Each station may divide its processing effort so as to process fluids of different classes simultaneously. A member $\xi \in \Xi$ is said to be an *allocation matrix* for the model, representing how the effort is distributed among classes. When the system operates under a given allocation matrix ξ, the element ξ_{ij} represents the fraction of effort devoted at station j to processing class-i fluid. Consequently, station j processes class-i fluid at rate $\bar{\mu}_{ij}\xi_{ij}$. A system operating under ξ is balanced if the *balance equation* $\bar{\mu}(\xi) = \lambda$ holds so that the stations process the incoming fluid at the rate at which it enters the system.

Consider now the following linear program:

$$\text{Minimize } \rho \text{ over } \xi \in \Xi \text{ subject to } \bar{\mu}(\xi) = \lambda \text{ and } \sum_i \xi_{ij} \leq \rho,\ j \in \mathcal{J},\ \rho \geq 0. \tag{8}$$

The following condition asserts that the static fluid model is critically loaded.

ASSUMPTION 2.1 [Heavy traffic (HT)]. There exists a unique optimal solution (ξ^*, ρ^*) to the linear program (8). Moreover, $\sum_{i \in \mathcal{I}} \xi_{ij}^* = 1$ for all $j \in \mathcal{J}$ (and consequently, $\rho^* = 1$).

Following terminology from [19], a pair (i, j) where $i \sim j$, is called an *activity*, and an activity (i, j) is said to be *basic* if $\xi^*_{ij} > 0$ and is said to be *nonbasic* otherwise. In the static fluid model operating under ξ^*, class-i fluid is processed at a positive rate by station j if and only if (i, j) is a basic activity. Let $\mathcal{G}_b$ be the sub-graph of $\mathcal{G}$, having $\mathcal{I} \cup \mathcal{J}$ as a vertex set, and an edge connecting $i \in \mathcal{I}$ and $j \in \mathcal{J}$ if and only if (i, j) is a basic activity. The edge set of $\mathcal{G}_b$ will be denoted by $\mathcal{E}_b$. We will write $\mathcal{E}_{nb} = \mathcal{E} \setminus \mathcal{E}_b$ for the set of nonbasic activities. For $i \in \mathcal{I}$ we let $\mathcal{J}(i) := \{j \in \mathcal{J} : (i, j) \in \mathcal{E}_b\}$ be the set of server pools that are connected to class i via basic activities and, similarly, for $j \in \mathcal{J}$, we let $\mathcal{I}(j) := \{i \in \mathcal{I} : (i, j) \in \mathcal{E}_b\}$ be the set of customer classes connected to server pool j via basic activities.

ASSUMPTION 2.2 (Complete resource pooling (CRP) [19]). The graph $\mathcal{G}_b$ is connected.

Both Assumptions 2.1 and 2.2 will be in force throughout the paper. In the context of conventional heavy-traffic, the connectedness of the stations via basic activities leads to a high level of cooperation in that the system asymptotically behaves as if it operates under a single super-server [8, 23]. As explained in [19] (see also [7, 8, 28]), the CRP condition is related to the so-called *workload process* being one-dimensional, and allows for the corresponding Brownian control problem to admit a one-dimensional solution. It is known from Williams [28] that, under Assumption 2.1, $\mathcal{G}_b$ is *connected* if and only if it is a *tree*. Both the lower bound and the asymptotic optimality results build on the tree structure of $\mathcal{G}_b$.

It will be useful to state an equivalent form of the above assumptions, given by Mandelbaum and Stolyar [23]. To this end, denote

$$\mathscr{M} = \{\bar{\mu}(\xi) : \xi \in \Xi\}.$$

Note that $\mathscr{M}$ is a convex polygonal domain and a subset of $\mathbb{R}^{\mathbf{I}}_+$. It is argued in [23] that the conjunction of the HT and the CRP conditions is equivalent to the following statement: λ *is a maximal element of* $\mathscr{M}$ *w.r.t. the usual partial order on* $\mathbb{R}^{\mathbf{I}}_+$; *the unit outward normal to* $\mathscr{M}$ *at* λ *is unique*; *and the matrix* $\xi \in \Xi$ *for which* $\bar{\mu}(\xi) = \lambda$ *is unique.* (Note that, because $\lambda_i > 0$ for all $i \in \mathcal{I}$, it follows that $\mathscr{M}$ is an $\mathbf{I}$-dimensional set, as assumed a priori in [23].) We will denote the unit outward normal to $\mathscr{M}$ at λ by θ. As argued in [23], θ must satisfy $\theta_i > 0$, $i \in \mathcal{I}$. These facts will be used in our analysis.

We continue the description of the probabilistic model. We let a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ be given, supporting all random variables and stochastic processes defined below. We write $\mathbb{E}$ for expectation w.r.t. $\mathbb{P}$. Let B^n_{ij} denote the process representing the number of pool-j servers working on class-i customers [note that $B^n_{ij} \equiv 0$ for $(i, j) \notin \mathcal{E}$]. Let X^n_i, Q^n_i and I^n_j denote, respectively, the number of class-i customers in the system (the "headcount" process), the number of

class-i customers in the queue and the number of pool-j servers that are idle. Note that

$$X_i^n = Q_i^n + \sum_j B_{ij}^n, \qquad i \in \mathcal{I} \tag{9}$$

and

$$N_j^n = I_j^n + \sum_i B_{ij}^n, \qquad j \in \mathcal{J}. \tag{10}$$

We are given standard (unit rate) Poisson processes $(S_{ij}, (i, j) \in \mathcal{E})$. The number of service completions of class-i customers by pool-j servers is constructed by setting

$$D_{ij}^n(t) = S_{ij}\big(T_{ij}^n(t)\big), \tag{11}$$

where

$$T_{ij}^n(t) = \mu_{ij}^n \int_0^t B_{ij}^n(s)\, ds. \tag{12}$$

We then have

$$X_i^n(t) = X_i^n(0) + A_i^n(t) - \sum_j D_{ij}^n(t). \tag{13}$$

Naturally, it is required that for all $t \geq 0$,

$$X_i^n(t), B_{ij}^n(t), Q_i^n(t) \in \mathbb{Z}_+, \qquad i \in \mathcal{I}, j \in \mathcal{J}. \tag{14}$$

For each n, the processes $(A_i, i \in \mathcal{I})$, $(S_{ij}, (i, j) \in \mathcal{E})$ and the initial condition $((B_{ij}^n(0), (i, j) \in \mathcal{E}))$, $Q_i^n(0))$ are assumed to be mutually independent. We refer to $(A, S, B^n(0), Q^n(0))$ as the *primitives*.

When routing decisions are made in a causal manner based on the observed histories of the processes involved, namely

$$\Pi^n := \big(X^n, Q^n, B^n, I^n, D^n, T^n\big), \tag{15}$$

the construction of the departure process via (11) and (12) assures that the customers' service times are independent, exponential random variables (in particular, this follows from [10], Theorem 16, page 41; see also the discussion in Section 2.1 of [14]). For the treatment of this paper, it will not be necessary to require any nonanticipating property of the class of policies we consider (although the exponential property will be lost when the policy does not satisfy a suitable nonanticipating property). Instead, we use an elaborate definition of the term "policy" that merely relies on the *equations* presented thus far. More precisely, *any process* Π^n*, of the form* (15) *and possessing RCLL sample paths will be referred to as a policy for the* n*th system, provided that equations* (9)–(14) *hold, and that the stochastic primitives satisfy our probabilistic assumptions mentioned above.* Given n, the collection of all policies for the nth system is denoted by $\mathcal{P}^n$. Note that policies need not satisfy any work conservation condition.

2.2. *Cost functional and asymptotic lower bound.* Our results will be concerned with asymptotically minimizing a cost associated to the diffusion-scaled queueing process, defined by

$$\hat{Q}_i^n(t) = n^{-1/2} Q_i^n(t), \qquad i \in \mathcal{I}. \tag{16}$$

Let a cost function $C : \mathbb{R}_+^{\mathbf{I}} \to \mathbb{R}_+$ be given, that is continuous and nondecreasing with respect to the usual partial order on $\mathbb{R}_+^{\mathbf{I}}$. Fix $u > 0$. The cost criterion of interest will be

$$\int_0^u C\big(\hat{Q}^n(s)\big)\, ds. \tag{17}$$

Note that this criterion is a r.v. for each n. We do not formulate the problem in terms of minimizing the *expected value* of (17). Considering (17) allows us to state a result on the asymptotic behavior of these r.v.'s that is stronger than one about their expectation; see Remark 2.1.

ASSUMPTION 2.3. The function $C_*(\cdot)$ defined by

$$C_*(a) = \inf\big\{C(q) : q \in \mathbb{R}_+^{\mathbf{I}}, \theta' q = a\big\}, \qquad a \geq 0, \tag{18}$$

is convex.

Clearly, a sufficient condition for the convexity of C_* is the convexity of the function C. However, it is not necessary. Consider, for example, $\mathcal{I} = \{1, 2\}$ and the cost function $C(x_1, x_2) = 2(x_1 + x_2)^2 - (x_1 - x_2)^2$ for $x \in \mathbb{R}_+^2$ and assume $\theta = (1, 1)$. Then $C_*(y) = y^2$ is convex while C is not.

To state our first main result we introduce additional notation: for $x \in \mathbb{R}$, $x^+ = \max(x, 0)$ and $x^- = \max(-x, 0)$. The *Skorohod map* $\Gamma : \mathcal{D}([0, \infty) : \mathbb{R}) \to \mathcal{D}([0, \infty), \mathbb{R}_+)$ is defined by

$$\Gamma[\zeta](t) = \zeta(t) + \sup_{s \leq t} \big(-\zeta(s)\big)^+, \qquad t \geq 0. \tag{19}$$

The process

$$\hat{X}_i^n(t) = n^{-1/2}\left(X_i^n(t) - \sum_j \xi_{ij}^* N_j^n\right) \tag{20}$$

represents the diffusion-scale deviation of the headcount process from the quantities dictated by the static fluid model. Throughout, we assume

$$\hat{X}^n(0) \Rightarrow X_0 \qquad \text{as } n \to \infty, \tag{21}$$

where X_0 is a.s. finite r.v. Finally,

$$\ell_i := \hat{\lambda}_i - \sum_j \hat{\mu}_{ij} \xi_{ij}^* \quad \text{and} \quad \sigma_i := \left(\lambda_i C_{i,IA}^2 + \sum_j \bar{\mu}_{ij} \xi_{ij}^*\right)^{1/2}, \qquad i \in \mathcal{I}.$$

THEOREM 2.1. *Fix an arbitrary sequence* $\{\Pi^n = (X^n, Q^n, B^n, I^n, D^n, T^n), n \in \mathbb{N}\}$ *of policies. Then* $\{\Pi^n\}$ *can be coupled on a common probability space with the r.v.* X_0 *and an* $\mathbf{I}$*-dimensional Brownian motion* W (*with drift vector* ℓ *and diffusion coefficient* σ) *so that* W *and* X_0 *are mutually independent and, w.p.*1,

$$\liminf_{n\to\infty} \int_0^u C\big(\hat{Q}^n(t)\big)\,dt \geq \int_0^u C_*\big(Q^*(t)\big)\,dt,$$

where Q^* *is the* (*one-dimensional*) *reflected Brownian motion given by* $\Gamma[\theta' X_0 + \theta' W]$.

REMARK 2.1. A more standard control theoretic setting is one where an *expected* cost, such as $\mathbb{E}[\int_0^u C(\hat{Q}^n(t))\,dt]$, is to be minimized. An asymptotic lower bound on the expected cost follows immediately from the above result, using Fatou's lemma, namely $\liminf \mathbb{E}[\int_0^u C(\hat{Q}^n(t))\,dt] \geq \mathbb{E}[\int_0^u C_*(Q^*(t))\,dt]$. The result stated in the theorem is, of course, stronger.

REMARK 2.2. The family $\mathcal{P}^n$ includes both preemptive and nonpreemptive policies. The policy that we will construct in Section 5 is nonpreemptive, but we will prove that it is asymptotically optimal within the larger family $\mathcal{P}^n$.

2.3. *Discussion. On sojourn time costs.* In the NDS regime, unlike in the conventional regime, sojourn time costs lead to control policies that are distinct from those designed to minimize waiting times. We provide here a heuristic argument, showing that sojourn time costs can be expressed as queue-length (or waiting time) costs. Rigorizing and extending this viewpoint is subject for future work. This heuristic argument provides further motivation to study queue-length costs.

For this discussion we remove the superscript n from the notation. For $t \geq 0$, $i \in \mathcal{I}$, denote by $\Delta_i(t)$ [resp., $\Sigma_i(t)$, $\mathrm{SOJ}_i(t)$] the properly scaled waiting time (resp., service time, sojourn time) experienced by the class-i customer to first arrive after time t. The scaling is obtained by multiplying the original quantities by $\sqrt{n}$ (see [3]). We have

$$\mathrm{SOJ}_i(t) = \Delta_i(t) + \Sigma_i(t).$$

Consider a cost of the form

$$J = \int_0^u \sum_i C_i\big(\mathrm{SOJ}_i(t)\big)\,dt.$$

One expects that under reasonable policies the fraction of class i customers routed to servers in pool j be roughly $p_{ij} = \bar{\mu}_{ij}\xi^*_{ij}/\lambda_i$. Moreover, one expects that $\Sigma_i(t)$ is approximately a mixture of exponentials where it is an exponential with rate μ_{ij} with probability p_{ij} (this is consistent with the result for the so-called

inverted V model in [3]; see Theorem 2.2 there). Denote by F_{Σ_i} the corresponding distribution function. We have $\mathbb{E}[C_i(\mathrm{SOJ}_i(t))] = E[R_i(t)]$, where

$$R_i(t) = \mathbb{E}\big[C_i\big(\Delta_i(t) + \Sigma_i(t)\big)|\Delta_i(t)\big].$$

By the above discussion one expects that $R_i(t) \approx \tilde{C}_i(\Delta_i(t))$, where

$$\tilde{C}_i(y) = \int_0^\infty C_i(y+x)\,dF_{\Sigma_i}(x), \qquad y \geq 0.$$

Using $\Delta_i(t) \approx \lambda_i Q_i(t)$ (by the *snapshot principle*) thus shows

$$J \approx \mathbb{E}\bigg[\int_0^u \sum_i \tilde{C}_i\big(\lambda_i Q_i(t)\big)\,dt\bigg].$$

On pathwise lower bounds. Note that Theorem 2.1 does not assert that, under the coupling, w.p.1,

$$\liminf_{n\to\infty} C\big(\hat{Q}^n(t)\big) \geq C_*\big(Q^*(t)\big), \qquad t \geq 0. \tag{22}$$

It only provides an integral version of this inequality. In conventional HT, (22) is often called a pathwise lower bound, and specifically, for the parallel server model, it is shown to hold in [23].

However, in the NDS regime, particularly, under the setting of Theorem 2.1, (22) is a false statement. In fact, *given suitable initial conditions* (*e.g., zero*) *one can find constants* $t_0 > 0$ *and* $c > 0$ *such that, under a suitable policy,* $\hat{Q}^n(t) = 0$ *for all* $t \in [t_0, t_0 + cn^{-1/2}]$*, with probability tending to one.* [This clearly shows that under no coupling can (22) hold.] We do not prove this statement here. A detailed study of an analogous phenomenon in the Halfin–Whitt regime, referred to as *null-controllability*, has been studied in detail [4, 5]. Briefly, this phenomenon is described as follows. Under suitable algebraic conditions on the system parameters, the *critically loaded* parallel server model in the Halfin–Whitt regime can be controlled so that all queue-lengths vanish for $O(1)$ units of time, with probability tending to one. When the mechanisms described in these works are applied in the NDS regime, they yield vanishing of all queue-lengths, though only for $O(n^{-1/2})$ units of time. [Clearly, such a property could not hold for $O(1)$ units of time in the NDS regime, since this would contradict Theorem 2.1.]

This complication explains why the proof of Theorem 2.1 is more involved than the analogous lower bound in the conventional regime (as well as the need for tools such as Proposition 4.1 and Lemma 4.1).

On other heavy traffic regimes. Atar [3] emphasizes the viewpoint that there exists a whole spectrum of heavy traffic diffusion approximations. For any $\alpha \in [0, 1]$ one obtains a distinct heavy traffic regime by setting the quantities λ, μ and N proportional to n, $n^{1-\alpha}$ and, respectively, n^α, while maintaining the critical load condition of having $\lambda - N\mu$ at the order of $n^{1/2}$. The conventional, NDS and Halfin–Whitt regimes can then be identified with the cases $\alpha = 0, 1/2$ and,

respectively, 1. Whether results analogous to those of this paper hold for all values of $\alpha \in (0, 1)$ is left as an important open question. In view of the so-called null-controllability results in [4, 5], $\alpha = 1$ should be excluded since the lower bound is not expected to hold in the Halfin–Whitt regime.

3. Preliminaries.

3.1. *Diffusion-scale processes*. In the present subsection we define some diffusion-scale processes and develop relations that they satisfy. Let

$$\bar{T}_{ij}^n(t) = n^{-1} T_{ij}^n(t) \equiv \frac{\mu_{ij}^n}{n} \int_0^t B_{ij}^n(s)\, ds, \tag{23}$$

$$\hat{A}_i^n(t) = n^{-1/2}\big(A_i^n(t) - \lambda_i^n t\big), \qquad i \in \mathcal{I}, \tag{24}$$

$$\hat{S}_{ij}^n(t) = n^{-1/2}\big(S_{ij}(nt) - nt\big) \qquad (i, j) \in \mathcal{E}, \tag{25}$$

$$\tilde{B}_{ij}^n = B_{ij}^n - \xi_{ij}^* N_j^n, \qquad \hat{B}_{ij}^n = n^{-1/2} \tilde{B}_{ij}^n, \tag{26}$$

$$V_{ij}^n = n^{-1/2}\big(D_{ij}^n - T_{ij}^n\big) \equiv \hat{S}_{ij}^n \circ \bar{T}_{ij}^n, \tag{27}$$

$$\ell_i^n = \hat{\lambda}_i^n - \sum_j \hat{\mu}_{ij}^n \xi_{ij}^* \tag{28}$$

and

$$W_i^n(t) = \ell_i^n t + \hat{A}_i^n(t) - \sum_j V_{ij}^n(t). \tag{29}$$

Since $\sum_i \xi_{ij}^* = 1$, we have by (10) that

$$I_j^n + \sum_i \tilde{B}_{ij}^n = 0, \qquad j \in \mathcal{J}, \tag{30}$$

and by (9) and (20) that

$$\hat{X}_i^n = \hat{Q}_i^n + \sum_j \hat{B}_{ij}^n. \tag{31}$$

Using (11), (12), (13) and (29), we get

$$\begin{aligned}
\hat{X}_i^n(t) &= \hat{X}_i^n(0) + n^{-1/2} A_i^n - n^{-1/2} \sum_j D_{ij}^n(t) \\
&= \hat{X}_i^n(0) + W_i^n(t) + n^{1/2} \lambda_i t - n^{1/2} \sum_j \bar{\mu}_{ij} \xi_{ij}^* t \\
&\quad - n^{-1/2} \sum_j \mu_{ij}^n \int_0^t \tilde{B}_{ij}^n(s)\, ds.
\end{aligned}$$

Since $\bar{\mu}(\xi^*) = \lambda$, the third and forth terms above cancel out so that letting

$$\varepsilon_{ij}^n := n^{-1/2}\mu_{ij}^n - \mu_{ij}, \tag{32}$$

we arrive at the following identity:

$$\hat{X}_i^n(t) = \hat{X}_i^n(0) + W_i^n(t) - \sum_j (\mu_{ij} + \varepsilon_{ij}^n) \int_0^t \tilde{B}_{ij}^n(s)\, ds. \tag{33}$$

Using (31), we obtain

$$\hat{Q}_i^n(t) = \hat{X}_i^n(0) + W_i^n(t) - \sum_j (\mu_{ij} + \varepsilon_{ij}^n) \int_0^t \tilde{B}_{ij}^n(s)\, ds - \sum_j \hat{B}_{ij}^n(t). \tag{34}$$

Identities (33) and (34) will be used in the sequel.

3.2. *Auxiliary results.*

LEMMA 3.1. (i) *The rescaled primitive processes* $(\hat{A}_i^n, i \in \mathcal{I})$ *and* $(\hat{S}_{ij}^n, (i,j) \in \mathcal{E})$ *and initial condition* $\hat{X}^n(0)$, *jointly converge in law, uniformly on compacts, to processes denoted* $(W_{A,i}, i \in \mathcal{I})$ *and* $(W_{S_{ij}}, (i,j) \in \mathcal{E})$, *and the r.v.* X_0, *where* $W_{A,i}$ (*resp.*, $W_{S_{ij}}$) *is a zero mean Brownian motion with diffusion coefficient* $\lambda_i^{1/2} C_{i,IA}$ (*resp.*, 1). *Moreover, the* $\mathbf{I} + \mathbf{K}$ *Brownian motions and the r.v.* X_0 *are mutually independent.*

(ii) *The parameters defined in* (28) *and* (32) *satisfy*

$$\ell_i^n \to \ell_i := \hat{\lambda}_i - \sum_j \hat{\mu}_{ij}\xi_{ij}^* \quad \text{and} \quad \varepsilon_{ij}^n = O\big(n^{-1/4}\big) \qquad \text{as } n \to \infty. \tag{35}$$

(iii) *Consequently, the processes*

$$\hat{W}_i^n(t) := \ell_i^n t + \hat{A}_i^n(t) - \sum_j \hat{S}_{ij}^n(\bar{\mu}_{ij}\xi_{ij}^* t), \qquad i \in \mathcal{I}, \tag{36}$$

and the initial condition X_0^n *jointly converge in law to mutually independent processes* $(W_i, i \in \mathcal{I})$ *and r.v.* X_0, *where* W_i *is a Brownian motion starting from zero, with drift* ℓ_i [*cf.* (35)] *and diffusion coefficient*

$$\sigma_i := \left(\lambda_i C_{i,IA}^2 + \sum_j \bar{\mu}_{ij}\xi_{ij}^*\right)^{1/2}, \qquad i \in \mathcal{I}. \tag{37}$$

Invoking the Skorohod representation theorem, we can, and will, assume throughout that the convergence statements of the above lemma occur in an a.s. sense.

REMARK 3.1. Note that Lemma 3.1 deals with convergence of processes that depend only on the primitives, and thus the same a.s. limit is attained under any policy.

PROOF OF LEMMA 3.1. (i) It is well known that a renewal process, scaled in the fashion of (24) and (25), converges in law, uniformly on compacts, to a Brownian motion with zero mean and diffusion coefficient as stated [9], Section 17. The mutual independence of the processes and the independence from the initial conditions follows the validity of this property for the pre-limit objects. (ii) The first statement follows by (2), (6) and (28). The second follows by (3), (6) and (32). □

3.3. *Diffusion model formulations*. In this subsection we present two diffusion models, originating from [19] and [23]. We associate to these models a control problem analogous to the one used above for the queuing network, and provide a complete solution. The optimum is exactly analogous to the lower bound from Theorem 2.1. The main point of this analysis is that the problem of identifying a minimizing control (Proposition 3.1 below) can later be mimicked to construct a policy for the queueing network that achieves the lower bound in an asymptotic sense; see Section 5. Along the way we establish equivalence of the two diffusion models.

Harrison and López [19] present and analyze a model of controlled diffusion, which stands for the formal limit of diffusion-scaled processes associated with the queueing model (in the conventional HT regime). The diffusion model was later used by Bell and Williams [8] as the basis for the construction of asymptotically optimal policies for the queueing model. This diffusion model, to which we refer as *Model* I, consists of the r.v.'s $X_{0,i}$ and BMs W_i alluded to in Section 2, and in addition, processes $(X_i, i \in \mathcal{I})$, $(I_j, j \in \mathcal{J})$ and $(Y_{ij}, (i,j) \in \mathcal{E})$, possessing RCLL sample paths, which satisfy, in addition, the following relations:

$$X_i(t) = X_{0,i} + W_i(t) + \sum_{j:i\sim j} \mu_{ij} Y_{ij}(t) \geq 0, \qquad t \geq 0, i \in \mathcal{I}, \tag{38}$$

$$I_j := \sum_i Y_{ij} \text{ is nondecreasing and } I_j(0) \geq 0, \qquad j \in \mathcal{J}, \tag{39}$$

$$Y_{ij} \text{ is nonincreasing and } Y_{ij} \leq 0 \qquad (i,j) \in \mathcal{E}_{nb}. \tag{40}$$

(Y_{ij} are further required in [19] to be adapted, but here we take the viewpoint of [8] where this requirement is dropped.) Taking formal limits of the scaled processes from our queueing model gives rise to the same diffusion model. Indeed, assuming that W^n converges to W and that $\hat{B}^n$ converge to zero, noting that $\hat{X}^n(0)$ converge to X_0 and $\varepsilon^n \to 0$, and writing Y for a limit of $-\int_0^{\cdot} \tilde{B}^n$, equation (38) arises as a limit form of (33) [also of (34)], where X stands for the limit of $\hat{X}^n$ (also of $\hat{Q}^n$). The nonnegativity constraint on X_i is clear from that of Q^n, observing (16). Similarly, if I_j corresponds to a limit of $\int_0^{\cdot} I_j^n$, then I_j should be nondecreasing and, because of (30), satisfy (39). Finally, (40) represents the fact that $\tilde{B}_{ij}^n \geq 0$ for any nonbasic activity (i,j) due to (26) and the fact that $\xi_{ij}^* = 0$ for such an activity. Making the above formal statements rigorous will be one of the main issues dealt with in Sections 4 and in 5.

Mandelbaum and Stolyar [23] construct asymptotically optimal control policies for the parallel server model in the conventional HT regime, without explicitly alluding to a diffusion model (or a Brownian control problem). However, their verbal discussion and mathematical treatment of the diffusion scaled processes suggest the following diffusion model, to which we refer here as *Model* II. In addition to the random vector X_0 and the $\mathbf{I}$-dimensional process W from Section 2, the model consists of $\mathbf{I}$-dimensional processes X and Z. These are assumed to have RCLL sample paths and satisfy

$$X_i(t) = X_{0,i} + W_i(t) + Z_i(t) \geq 0, \qquad t \geq 0, i \in \mathcal{I}, \tag{41}$$

$$\theta' Z \text{ is nondecreasing and } \theta' Z(0) \geq 0. \tag{42}$$

The interpretation of X is the same as in Model I, while Z corresponds to the term $\sum_j \mu_{ij} Y_{ij}$ of (38). Below we claim that Models I and II are equivalent in a suitable sense, and that they both achieve the lower bound of Theorem 2.1.

PROPOSITION 3.1. *Suppose that Assumptions 2.1 and 2.2 hold and fix an initial state X_0 and a Brownian motion W as in Lemma 3.1.*

(i) *Given a pair (X, Y) that satisfies (38)–(40), there exists Z such that the pair (X, Z) satisfies (41)–(42).*

(ii) *Given a pair (X, Z) that satisfies (41)–(42), there exists Y such that (X, Y) satisfies (38)–(40).*

(iii) *Let (X, Y, Z) be such that (X, Y) satisfy (38)–(40) and (X, Z) satisfy (41)–(42). Then, with probability 1, $C(X(t)) \geq C_*(Q^*(t))$ for all t, where Q^* is as in Theorem 2.1. Moreover, the lower bound is attainable: there exist stochastic processes (X, Y, Z) such that (X, Y) [resp., (X, Z)] satisfies (38)–(40) [resp. (41)–(42)] and with probability 1,*

$$C\big(X(t)\big) = C_*\big(Q^*(t)\big), \qquad t \geq 0.$$

Below, the notation Σ, Ξ is as in Section 2 and, for $\xi \in \Sigma$, $\bar{\mu}(\xi)$ is as defined in (7).

LEMMA 3.2. *Let $x \in \Sigma$ be such that $x_{ij} \leq 0$ for all $(i, j) \in \mathcal{E}_{nb}$, and $\sum_i x_{ij} \geq 0$ for all $j \in \mathcal{J}$. Then $\theta' \bar{\mu}(x) \geq 0$. Also, if $x_{ij} = 0$ for all $(i, j) \in \mathcal{E}_{nb}$ and $\sum_i x_{ij} = 0$ for all $j \in \mathcal{J}$ then $\theta' \bar{\mu}(x) = 0$.*

PROOF. Let us first show that $\xi := \xi^* - \varepsilon x$ is an element of Ξ, provided that $\varepsilon > 0$ is sufficiently small. For $(i, j) \in \mathcal{E}_{nb}$, $\xi^*_{ij} = 0$, and so $\xi_{ij} \geq 0$ by the assumptions of the lemma. For $(i, j) \in \mathcal{E}_b$, $\xi^*_{ij} > 0$, hence $\xi_{ij} \geq 0$ for all sufficiently small $\varepsilon > 0$. Finally, since we assumed that $\sum_i x_{ij} \geq 0$ for all $j \in \mathcal{J}$ and since $\xi^* \in \Xi$, we have that

$$\sum_i \xi_{ij} \leq \sum_i \xi^*_{ij} \leq 1, \qquad j \in \mathcal{J},$$

so that $\xi \in \Xi$. Next, since θ is an outward normal to the convex set $\mathscr{M}$ at λ, we have that $\theta'(m-\lambda) \le 0$ for every $m \in \mathscr{M}$. Since $\bar{\mu}(\xi^*) = \lambda$ and $\xi \in \Xi$, we have

$$\theta'\bar{\mu}(\varepsilon x) = \theta'\big(\bar{\mu}(\xi^*) - \bar{\mu}(\xi)\big) = \theta'\big(\lambda - \bar{\mu}(\xi)\big) \ge 0. \tag{43}$$

Since $\varepsilon > 0$, the first claim follows.

For the second part, following [23], we claim that there exist constants z_j^*, $j \in \mathcal{J}$, such that $\theta_i \bar{\mu}_{ij} = z_j^*$ for all $(i,j) \in \mathcal{E}_b$. Indeed, by (43) $\theta'\lambda = \theta'\bar{\mu}(\xi^*) = \sup_{\xi\in\Xi} \theta'\bar{\mu}(\xi)$ and, in turn, we must have that $\theta_i\bar{\mu}_{ij} = \max_k \theta_k \bar{\mu}_{kj}$ for all $(i,j) \in \mathcal{E}_b$. Define

$$z_j^* = \max_k \theta_k \bar{\mu}_{kj}. \tag{44}$$

Thus, $\theta_i \mu_{ij} = z_j^*$ for all $(i,j) \in \mathcal{E}_b$. If x is such that $x_{ij} = 0$ for all $(i,j) \in \mathcal{E}_{nb}$ and $\sum_i \xi_{ij} = 1$ for all $j \in \mathcal{J}$, then $\xi := \xi^* - \varepsilon x$ satisfies $\xi_{ij} = 0$ for all $(i,j) \in \mathcal{E}_{nb}$ and $\sum_i \xi_{ij} = 1$ for all $j \in \mathcal{J}$ and, in turn,

$$\theta'\bar{\mu}(\varepsilon x) = \theta'\big(\bar{\mu}(\xi^*) - \bar{\mu}(\xi)\big) = \sum_{(i,j)\in\mathcal{E}_b} \theta_i \mu_{ij}\big(\xi_{ij}^* - \xi_{ij}\big)$$

$$= \sum_j z_j^* \Big(\sum_i \xi_{ij}^* - \sum_i \xi_{ij} \Big) = 0.$$

□

PROOF OF PROPOSITION 3.1. (i) Suppose that relations (38)–(40) hold. Letting $Z_i = \sum_j \mu_{ij} Y_{ij}$, the relation (41) is immediate. To show (42), we must prove that, for $0 \le s \le t$,

$$\sum_{ij} \theta_i \mu_{ij}\big(Y_{ij}(t) - Y_{ij}(s)\big) \ge 0 \qquad \Big[\text{resp.}, \sum_{ij} \theta_i \mu_{ij} Y_{ij}(t) \ge 0\Big],$$

which can alternatively be stated as $\theta'\bar{\mu}(x) \ge 0$, where $x_{ij} = \nu_j^{-1}(Y_{ij}(t) - Y_{ij}(s))$ [resp., $x_{ij} = \nu_j^{-1} Y_{ij}(t)$]. Properties (39) and (40) of Y and an application of Lemma 3.2 imply $\theta'\bar{\mu}(x) \ge 0$, in both cases, and the claim is proved.

(ii) Assume we are given a solution (X, Z) to (41) and (42). We will construct a process Y so that the pair (X, Y) satisfies (38)–(40). Fix an arbitrary $l \in \mathcal{J}$ and set $I_j = 0$ for all $j \ne l$ and $I_l(t) = \frac{\nu_l}{z_l^*}\theta' Z(t)$ [where z^* is as in (44)]. Then, given X_0, X, Z and W, let us show there exists a solution Y to the set of equations

$$\sum_{j:(i,j)\in\mathcal{E}_b} \mu_{ij} Y_{ij} = X_i - X_{i,0} - W_i, \qquad i \in \mathcal{I}, \tag{45}$$

$$\sum_{i:(i,j)\in\mathcal{E}_b} Y_{ij} = 0 \qquad \forall j \in \mathcal{J} \setminus \{l\}, \tag{46}$$

$$\sum_{i:(i,l)\in\mathcal{E}_b} Y_{il} = \frac{\nu_l}{z_l^*}\theta' Z, \tag{47}$$

and $Y_{ij} = 0$ for $(i, j) \notin \mathcal{E}_b$. To this end, we first remove one on the $I + J$ equations from the system. Fix $i_0 \in \mathcal{I}$ and assume that (45) holds for all $i \neq i_0$. Then

$$\begin{aligned}\sum_{i\neq i_0} \theta_i X_i(t) &= \sum_{i\neq i_0} \theta_i X_{i,0} + \sum_{i\neq i_0} \theta_i W_i(t) + \sum_{i\neq i_0} \theta_i \sum_j \bar{\mu}_{ij} Y_{ij}/\nu_j \\ &= \sum_{i\neq i_0} \theta_i X_{i,0} + \sum_{i\neq i_0} \theta_i W_i(t) - \sum_j \frac{\bar{z}^*_j}{\nu_j} Y_{i_0 j} + \theta' Z(t),\end{aligned} \tag{48}$$

where we used $\theta_i \bar{\mu}_{ij} = z^*_j$ for all $(i, j) \in \mathcal{E}_b$ (see the proof of Lemma 3.2) and equations (46) and (47). On the other hand, since (X, Z) solves (41) and (42) we have that $\theta' X(t) = \theta' X_0 + \theta' W(t) + \theta' Z(t)$ and, together with (48), that

$$\theta_{i_0} X_{i_0}(t) = \theta_{i_0} X_{i_0} + \theta_{i_0} W_{i_0}(t) + \sum_j \frac{\bar{z}^*_j}{\nu_j} Y_{i_0 j}.$$

Substituting $\theta_i \bar{\mu}_{ij} = z^*_j$ we have thus shown that (45) holds for i_0 provided that it holds for all $i \neq i_0$ and provided that (46) and (47) hold. Hence we can remove the equation for $i = i_0$ from (45).

The reduced system of equations has $I + J - 1$ variables (one for each basic activity) and the same number of equations. By the discussion on page 348 of [19], these equations are linearly independent. As a result, given X_0, X, Z and W, there exists a unique solution Y.

Note that (39) and (40) hold by construction.

To establish the lower bound in item (iii), let $M(t) := \sup_{s\leq t}(\theta' X_0 + \theta' W(t))^-$ and note that, by the minimality of Skorohod problem (see, e.g., [12], Section 2)

$$\theta' X(t) \geq \theta' X_0 + \theta' W(t) + M(t) \quad \text{and} \quad \theta' \tilde{X}(t) \geq \theta' X_0 + \theta' W(t) + M(t).$$

Since, $C(x) \geq C_*(\theta' x)$ for all $x \in \mathbb{R}^I$, the lower bound is established.

Finally, to show that the lower bound is attained we explicitly construct a pair (X, Z) that attains the lower bound. A process (X, Y) that attains the lower bound will then be constructed from (X, Z) as above. To that end, for $t \geq 0$, let $Q^*(t) = \theta' X_0 + \theta' W(t) + M(t)$ where $M(t)$ is as above. Set $X_i(t) = f^*_i(Q^*(t))$ where, given $a \in \mathbb{R}_+$, $f^*(a) := (f^*_1(a), \ldots, f^*_I(a))$ satisfies

$$f^*(a) \in \arg\min_{q\geq 0}\{C(q) : q \in \mathbb{R}^{\mathbf{I}}_+, \theta' q = a\}.$$

f^* can be selected to be measurable, as follows from Corollary 10.3 in the Appendix of [15], using the continuity of $C(\cdot)$. Thus $\theta' X(t) = Q^*(t)$ and $C(X(t)) = C(f^*(\theta' X(t))) = C(f^*(Q^*(t))) = C_*(Q^*(t))$. Setting $Z_i := X_i - X_{0,i} - W_i$, we have that $\theta' Z(t) = \theta' X - \theta' X_0 - \theta' W = Q^*(t) - \theta' X_0 - \theta' W = M(t)$ so that $\theta' Z$ is nonnegative and nondecreasing and the pair (X, Z) satisfies (41) and attains the lower bound. $\square$

4. Proof of the lower bound. In this section we prove Theorem 2.1. The main estimate on which the proof is based, Proposition 4.1, is stated in Section 4.1 where it is also used to prove Theorem 2.1. Proposition 4.1 is then proved in Section 4.2.

4.1. *Proof of Theorem 2.1.* An outline of the proof is as follows. Fix $u > 0$, which will serve as a time horizon. Given a sequence of policies $(\Pi^n, n \in \mathbb{N})$ we show that up to a certain random time $\tau_n \wedge u$, the cumulative process $\int_0^t \theta' \hat{Q}^n(s)\,ds$ is asymptotically bounded from below by the integrated RBM $\int_0^t Q^*(s)\,ds$. We then show that if $\tau_n < u$, then $\theta' \hat{Q}^n$ is large on a subinterval of $(\tau_n, u]$ and thus, with high probability, is bounded from below by the integrated RBM. The convexity of $C_*(\cdot)$ is then used in translating these bounds to bounds on the cost.

We turn to the proof. Denote $\varepsilon_M^n = \max_{i,j} |\varepsilon_{ij}^n|$, and recall that $\varepsilon_M^n = O(n^{-1/4})$ (Lemma 3.1). Let us fix a sequence ϱ_n such that

$$n^{-1/8}\varrho_n \to \infty \qquad \text{while } n^{-1/4}\varrho_n \to 0. \tag{49}$$

In particular, ϱ_n satisfies

$$n^{-1/2}\varrho_n \to 0, \qquad \varepsilon_M^n \varrho_n \to 0, \qquad n^{-1/2}(\varepsilon_M^n)^{-1}\varrho_n^2 \to \infty. \tag{50}$$

Let a sequence of policies $(\Pi^n, n \in \mathbb{N})$ be given, and define

$$\tau_n = \inf\left\{ t \geq 0 : \max_{i,j} \left| \int_0^t \tilde{B}_{ij}^n(s)\,ds \right| \geq \varrho_n \right\} \wedge u. \tag{51}$$

Since the conclusion of Theorem 2.1 clearly holds when the function C_* is constant, we henceforth assume that C_* is not constant. Denote $E^n(t) = \theta' \hat{Q}^n(t)$, $t \geq 0$. Below Q^* is as defined in Theorem 2.1.

PROPOSITION 4.1. *There exist constants c, $\tilde{c}$, $\bar{c}$, a strictly positive sequence $\{t_n\}$ satisfying $t_n\varrho_n \to \infty$, a sequence of events Ω^n satisfying $1_{\Omega^n} \to 1$ a.s., and processes $P^{*,n}$ and $H^{*,n}$, such that, with*

$$\tilde{\tau}_n := (\tau_n + t_n) \wedge u,$$

the following statements hold:

(i) *$P^{*,n}$ converges to Q^* uniformly on $[0, u]$, a.s.*

(ii) *$|H^{*,n}(t)| \leq \bar{c}$, for every n and $t \in [0, u]$, and*

$$\left| \int_0^t H^{*,n}(s)\,ds \right| \leq \tilde{c} n^{-1/2} \varrho_n, \qquad t \in [0, u]. \tag{52}$$

(iii) *On Ω^n one has*

$$E^n(t) \geq P^{*,n}(t) + H^{*,n}(t) \qquad \text{for all } t \in [0, \tau_n); \tag{53}$$

$$E^n(t) \geq c\varrho_n \qquad \text{for all } t \in [\tau_n, \tilde{\tau}_n), \text{ whenever } \tau_n < u; \tag{54}$$

$$E^n(t) \geq 0 \qquad \text{for all } t \in [\tilde{\tau}_n, u], \text{ whenever } \tilde{\tau}_n < u. \tag{55}$$

The proof is deferred to Section 4.2. Theorem 2.1 will be deduced from the Proposition 4.1, with the aid of the following lemma for which we define for $-\infty < a \le b < \infty$, for $x : [a, b] \to \mathbb{R}$ and $\delta > 0$,

$$\bar{w}_{[a,b]}(x, \delta) = \sup_{s,t\in[a,b];|s-t|\le\delta} \big|x(s) - x(t)\big|. \tag{56}$$

LEMMA 4.1. *Let $C_1 : \mathbb{R} \to \mathbb{R}_+$ be a nondecreasing, convex function. Let $T > 0$, $0 < \Delta < T/2$, $r > 0$, and functions $p, h : \mathbb{R}_+ \to \mathbb{R}$ be given such that*

$$\left|\int_0^t h(s)\,ds\right| \le \varepsilon, \qquad t \in [0, T]$$

and

$$\big|p(t)\big| + \big|h(t)\big| \le r, \qquad t \in [0, T].$$

Then

$$\int_0^T C_1\big(p(t) + h(t)\big)\,dt \ge \int_0^T C_1\big(p(t)\big)\,dt - \gamma_1 T - \gamma_2 T - \gamma_3,$$

where

$$\gamma_1 = \bar{w}_{[-r,r]}\left(C_1, \frac{2\varepsilon}{\Delta}\right), \qquad \gamma_2 = \bar{w}_{[-r,r]}\big(C_1, \bar{w}_T(p, \Delta)\big), \qquad \gamma_3 = 2C_1(r)\Delta.$$

PROOF. For $t \in [0, T - \Delta]$, using Jensen's inequality,

$$\begin{aligned}\frac{1}{\Delta}\int_t^{t+\Delta} C_1\big(p(t) + h(t)\big)\,dt &\ge C_1\left(\frac{1}{\Delta}\int_t^{t+\Delta}\big(p(s) + h(s)\big)\,ds\right)\\ &\ge C_1\left(\frac{1}{\Delta}\int_t^{t+\Delta} p(s)\,ds\right) - \gamma_1\\ &\ge C_1\big(p(t)\big) - \gamma_1 - \gamma_2.\end{aligned}$$

Thus

$$\int_0^T C_1\big(p(s) + h(s)\big)\,ds \ge \int_\Delta^{T-\Delta} C_1\big(p(t)\big)\,dt - \gamma_1 T - \gamma_2 T.$$

The result follows. □

PROOF OF THEOREM 2.1. Extend C^* to $\mathbb{R}$ by setting $C^* = C^*(0)$ on $(-\infty, 0)$. Fix $\Delta \in (0, u/2)$. Let $P^{*,n}$, $H^{*,n}$, Ω^n, $\tilde{\tau}_n$ be as in Proposition 4.1. Combining Proposition 4.1 and Lemma 4.1 we have, on Ω^n,

$$\begin{aligned}\int_0^u C_*\big(E^n(t)\big)\,dt &\ge \int_0^{\tau_n} C_*\big(P^{*,n}(t) + H^{*,n}(t)\big)\,dt + C_*(c\varrho_n)(\tilde{\tau}_n - \tau_n)\\ &\ge \int_0^{\tau_n} C_*\big(P^{*,n}(t)\big)\,dt + C_*(c\varrho_n)(\tilde{\tau}_n - \tau_n) - \gamma_1^n u - \gamma_2^n u - \gamma_3^n,\end{aligned}$$

where

$$\gamma_1^n = \bar{w}_{[-r_n,r_n]}\biggl(C_*, \frac{2\tilde{c}n^{-1/2}\varrho_n}{\Delta}\biggr), \qquad \gamma_2^n = \bar{w}_{[-r_n,r_n]}\bigl(C_*, \bar{w}_u\bigl(P^{*,n}, \Delta\bigr)\bigr),$$
$$\gamma_3^n = 2C_*(r_n)\Delta, \qquad r_n = \bigl\|P^{*,n}\bigr\|_u + \bar{c}.$$

Since $P^{*,n}$ converge uniformly, a.s., we have that r_n converge to a finite-valued r.v. Since C_* is a continuous function and $n^{-1/2}\varrho_n \to 0$ (50), $\gamma_1^n \to 0$ a.s. as $n \to \infty$. Moreover, by the uniform convergence of $P^{*,n}$ to a process with continuous sample paths, we have $\lim_{\Delta\to 0} \limsup_{n\to\infty} \gamma_2^n = 0$. It follows that for some $\gamma_4^n = \gamma_4^n(\Delta) \geq 0$ satisfying $\lim_{\Delta\to 0} \limsup_{n\to\infty} \gamma_4^n = 0$,

$$\begin{aligned}\int_0^u C_*\bigl(E^n(t)\bigr)\,dt \geq{}& \int_0^u C_*\bigl(P^{*,n}(t)\bigr)\,dt + \bigl[C_*(c\varrho_n) - K^n\bigr](\tilde{\tau}_n - \tau_n)\\ &- K^n 1_{\{\tilde{\tau}^n < u\}} - \gamma_4^n,\end{aligned} \tag{57}$$

holds on Ω^n, where

$$K^n = \sup_{t\in[0,u]} C_*\bigl(P^{*,n}(t)\bigr).$$

Now, $\limsup_n K^n$ is a finite r.v. On the other hand, one has $\liminf_{r\to\infty} C_*(r)/r > 0$ due to the fact that C_* is convex, nondecreasing and nonconstant whence (recalling that $\varrho_n \to \infty$ as $n \to \infty$) $C_*(c\varrho_n) \to \infty$ grows without bound. Thus the second term on the RHS of (57) is negative for only finitely many n. Also, if $\tilde{\tau}^n < u$, then we have, by definition, that $\tilde{\tau}^n - \tau^n = t_n$ and (recalling that $t_n\varrho_n \to \infty$) we have that the second term on the RHS of (57) grows to infinity so that the sum of the second and third terms in (57) is negative for only finitely many n. Since $1_{\Omega^n} \to 1$ a.s., we have thus shown that a.s.,

$$\liminf_{n\to\infty} \int_0^u C_*\bigl(E^n(t)\bigr)\,dt \geq \liminf_{n\to\infty} \int_0^u C_*\bigl(P^{*,n}(t)\bigr)\,dt = \int_0^u C_*\bigl(Q^*(t)\bigr)\,dt.$$

Noting that, by definition of C_* and E^n, $C(\hat{Q}^n(t)) \geq C_*(E^n(t))$, and recalling that u is arbitrary completes the proof. □

4.2. *Proof of Proposition* 4.1. The result is proved in three major steps, where the first two establish the lower bounds (53) and (54) for suitably defined processes $P^{*,n}$ and $H^{*,n}$ [note that (55) is immediate from the nonnegativity of $\hat{Q}^n$]. The third step verifies the statements of Proposition 4.1 with regard to convergence.

Step 1: *The interval* $[0, \tau_n)$. In this step we analyze the time interval $[0, \tau_n)$, introduce processes $P^{*,n}$ and $H^{*,n}$ [cf. (71) and (72)] and argue that they satisfy the bound (53).

We will next need a result that is similar to the minimality property of the Skorohod map Γ (19) but that allows for a certain kind of perturbation. The minimality

property of the Skorohod map is stated as follows: Let $\zeta \in \mathcal{D}$. Let $\eta \in \mathcal{D}$ be nondecreasing and satisfy $\eta(0) \geq 0$. Assume $\zeta(t) + \eta(t) \geq 0$, for all $t \geq 0$. Then

$$\zeta(t) + \eta(t) \geq \Gamma[\zeta](t) \equiv \zeta(t) + \sup_{s \leq t}[\zeta(s)^-], \qquad t \geq 0. \tag{58}$$

The following lemma provides the variant that we need.

LEMMA 4.2. *Let $u > 0$ and $\varepsilon > 0$, $\varepsilon < u$, be given. Let $\zeta \in \mathcal{D}$ and assume $\zeta(0) \geq 0$. Let*

$$\alpha = \zeta + \eta + \beta,$$

where $\eta \in \mathcal{D}$ is nondecreasing and satisfies $\eta(0) \geq 0$, $\beta \in \mathcal{D}$ satisfies

$$-\varepsilon^2 \leq \int_0^t \beta(s)\, ds \leq \varepsilon^2, \qquad t \in [0, u] \tag{59}$$

and $\alpha(t) \geq 0$, $t \in [0, u]$. Then

$$\alpha(t) \geq \Gamma[\zeta](t) + \beta(t) - \bar{w}_u(\zeta, \varepsilon) - 3\varepsilon, \qquad t \in [0, u]. \tag{60}$$

PROOF. By (58), we have $\alpha \geq \Gamma[\zeta + \beta]$. Thus

$$\alpha(t) \geq \zeta(t) + \beta(t) + \sup_{s \leq t}[(\zeta(s) + \beta(s))^-].$$

Denote $\delta = \bar{w}_u(\zeta, \varepsilon) + 3\varepsilon$. To prove the claim, it suffices to show that

$$\sup_{[0,t]}[(\zeta + \beta)^-] \geq \sup_{[0,t]}[\zeta^-] - \delta, \tag{61}$$

because then $\alpha \geq \zeta + \beta + \sup_{[0,\cdot]}[\zeta^-] - \delta = \Gamma[\zeta] + \beta - \delta$. To this end, consider first $t \in [0, \varepsilon]$. Since $\zeta(0) \geq 0$ by assumption, we have

$$\sup_{[0,\varepsilon]}[\zeta^-] - \delta \leq \bar{w}_u(\zeta, \varepsilon) - \delta = -3\varepsilon.$$

Thus (61) is immediate from the nonnegativity of the LHS of that inequality.

Next, fix $t \in [\varepsilon, u]$. Toward showing that (61) holds in this case as well, note that, for any $s \in [\varepsilon, u]$,

$$\inf_{[s-\varepsilon,s]} \beta \leq 3\varepsilon \tag{62}$$

for otherwise we would have, by (59),

$$3\varepsilon^2 \leq \int_{s-\varepsilon}^{s} \beta(\tau)\, d\tau \leq 2\varepsilon^2.$$

(note that (59) is imposed for all $t \in [0, u]$ and, in particular, for $t = s - \varepsilon$—we are using that here). We consequently have

$$\inf_{[s-\varepsilon,s]}(\zeta + \beta) \leq \sup_{[s-\varepsilon,s]} \zeta + \inf_{[s-\varepsilon,s]} \beta \leq \sup_{[s-\varepsilon,s]} \zeta + 3\varepsilon \leq \inf_{[s-\varepsilon,s]} \zeta + \delta,$$

where in the second inequality we used (62) and our choice of δ above. Taking the infimum over $s \in [\varepsilon, t]$, we obtain

$$\inf_{[0,t]}(\zeta + \beta) \le \inf_{[0,t]} \zeta + \delta. \tag{63}$$

Let us deduce from the above that

$$\inf_{[0,t]} 0 \wedge (\zeta + \beta) \le \inf_{[0,t]} 0 \wedge \zeta + \delta. \tag{64}$$

Indeed, if $\inf_{[0,t]} \zeta \ge -\delta$, then the RHS of (64) is nonnegative, and hence this inequality is valid. If $\inf_{[0,t]} \zeta < -\delta$, then $\inf \zeta = \inf 0 \wedge \zeta$, and hence the claim follows (63). In both cases, (64) holds. Note that (64) is equivalent to (61). We have thus shown that (61) holds for any $t \in [0, u]$. This completes the proof. □

We proceed with the proof of Theorem 2.1. Recall equation (34) for $\hat{Q}^n$ and that $E^n = \theta' \hat{Q}^n$. Writing

$$F^n = \theta' \hat{X}^n(0) + \theta' W^n - \int_0^{\cdot} \sum_{i,j} \theta_i \varepsilon_{ij}^n \tilde{B}_{ij}^n(s)\, ds, \tag{65}$$

$$G^n = -\int_0^{\cdot} \sum_{i,j} \theta_i \mu_{ij} \tilde{B}_{ij}^n(s)\, ds, \qquad H^n = -\sum_{i,j} \theta_i \hat{B}_{ij}^n, \tag{66}$$

we have

$$E^n = F^n + G^n + H^n. \tag{67}$$

We will apply Lemma 4.2, substituting $E^n(t)$, $F^n(t)$, $G^n(t)$ and $H^n(t)$, $t \in [0, \tau_n]$, for α, ζ, η and β, respectively. To this end, let us verify the assumptions on these processes. First, by (24), (27) and (29), we have $W^n(0) = 0$. Since $\hat{X}^n(0)$ is assumed to have values in $\mathbb{R}_+^{\mathbf{I}}$, we have by (65) that $F^n(0) \ge 0$.

Let us show that G^n is nondecreasing. It suffices to show that, for fixed n and t,

$$\sum_{i,j} \theta_i \mu_{ij} \tilde{B}_{ij}^n(t) \le 0. \tag{68}$$

We do this by invoking Lemma 3.2. Let $x \in \Sigma$ be defined by $x_{ij} = -\nu_j^{-1} \tilde{B}_{ij}^n(t)$. By (26) and the fact that $\xi_{ij}^* = 0$ for $(i, j) \in \mathcal{E}_{nb}$, we have $x_{ij} \le 0$ for $(i, j) \in \mathcal{E}_{nb}$. Also, for any j, $\sum_i x_{ij} = -\nu_j^{-1} \sum_i \tilde{B}_{ij}^n(t) = I_j^n(t)$, by (30). Hence $\sum_i x_{ij} \ge 0$. Thus by Lemma 3.2, $\theta' \bar{\mu}(x) \ge 0$. Recalling that $\bar{\mu}_{ij} = \nu_j \mu_{ij}$, we obtain (68).

Clearly, $G^n(0) = 0$.

Let us show that H^n satisfies a bound of the form (59). By the definition of τ_n, $|\int_0^t \tilde{B}_{ij}^n(s)\, ds| \le \varrho_n$, for all $t \le \tau_n$. Hence by (26),

$$\left| \int_0^t \hat{B}_{ij}^n(s)\, ds \right| \le n^{-1/2} \varrho_n, \qquad t \in [0, \tau_n]$$

and

$$\left|\int_0^t H^n(s)\,ds\right| \le c_0 n^{-1/2}\varrho_n, \qquad t \in [0, \tau_n], \tag{69}$$

some constant c_0.

Finally, note that $E^n(t) \ge 0$ for all t, since $\hat{Q}_i^n(t) \ge 0$ for every i and all t.

Having verified the hypotheses of Lemma 4.2, we obtain $E^n(t) \ge \Gamma[F^n](t) - \delta_n + H^n(t)$, for all $t \in [0, \tau_n]$, where

$$\delta_n = \bar{w}_{\tau_n}(F_n, \bar{\varepsilon}_n) + 3\bar{\varepsilon}_n, \qquad \bar{\varepsilon}_n = c_0^{1/2} n^{-1/4} \varrho_n^{1/2}. \tag{70}$$

Set

$$P^{*,n}(t) = \big(\Gamma\big[F^n\big](t) - \delta_n\big)1_{\{t<\tau_n\}} + \Gamma\big[\theta' \hat{X}^n(0) + \theta' \hat{W}^n\big](t)1_{\{t\ge\tau_n\}}, \tag{71}$$

$$H^{*,n}(t) = H^n(t)1_{\{t<\tau_n\}}. \tag{72}$$

Then $(P^{*,n}, H^{*,n})$ agree with $(\Gamma[F^n] - \delta_n, H^n)$ on the interval $[0, \tau_n)$, and we have shown

$$E^n(t) \ge P^{*,n}(t) + H^{*n}(t), \qquad t \in [0, \tau_n). \tag{73}$$

Step 2: *The interval* $[\tau_n, \tilde{\tau}_n)$. In this step we show that (54) holds on a suitably defined event Ω^n. The argument is based on the following lemma. For $x \in \Sigma$ denote $\|x\|^2 = \sum_{i,j} x_{ij}^2$. For the lemma below recall that Σ is the set of $\mathbf{I} \times \mathbf{J}$ matrices for which the (i, j) entry is zero whenever $i \nsim j$. Also, note that in this lemma the uniqueness of ξ^* is used in a crucial manner.

LEMMA 4.3. *Let* $x \in \Sigma$ *be such that* $x_{ij} \le 0$ *for all* $(i, j) \in \mathcal{E}_{nb}$, *and* $\sum_i x_{ij} \ge 0$ *for all* $j \in \mathcal{J}$. *Then*

$$\max_i \bar{\mu}_i(x) \ge c_1 \|x\|,$$

where $c_1 > 0$ *is a constant that does not depend on* x.

PROOF. Let

$$K = \left\{\xi^* - x : \|x\| = \varepsilon, x_{ij} \le 0, (i, j) \in \mathcal{E}_{nb}, \sum_i x_{ij} \ge 0, j \in \mathcal{J}\right\}.$$

A review of the proof of Lemma 3.2 shows that $K \subset \Xi$, provided that $\varepsilon > 0$ is sufficiently small. Let such ε be fixed. Recall that $\lambda = \bar{\mu}(\xi^*)$, and note that $\xi^* \notin K$. Thus the uniqueness of ξ^*, stated in Assumption 2.1, implies that there is no $\xi \in K$ for which $\bar{\mu}(\xi) = \lambda$. Hence $\lambda \notin \bar{\mathcal{M}}$, where $\bar{\mathcal{M}}$ is the image $\bar{\mu}(K)$ of K under $\bar{\mu}$. Recall that λ is a maximal element in $\mathcal{M}$ with respect to the usual partial order in $\mathbb{R}^{\mathbf{I}}$. Since $\lambda \notin \bar{\mathcal{M}}$, this says

$$\max_i\big[\lambda_i - \big(\bar{\mu}(\xi)\big)_i\big] > 0 \qquad \text{for every } \xi \in K.$$

Note that K is a compact set, and that the LHS of the above display depends continuously on ξ. Hence there exists $\delta > 0$ such that

$$\max_i \bigl[\lambda_i - \bigl(\bar{\mu}(\xi)\bigr)_i\bigr] \geq \delta \qquad \text{for every } \xi \in K.$$

Noting that the conclusion of the lemma holds for $x = 0$, consider any nonzero member x of Σ, satisfying the lemma's assumptions. Then $\xi^* - \varepsilon \|x\|^{-1} x \in K$. Hence

$$\begin{aligned}
\varepsilon \|x\|^{-1} \max_i \bar{\mu}_i(x) &= \max_i \bar{\mu}_i\bigl(\varepsilon \|x\|^{-1} x\bigr) \\
&= \max_i \bigl[\bar{\mu}\bigl(\xi^*\bigr) - \bar{\mu}\bigl(\xi^* - \varepsilon \|x\|^{-1} x\bigr)\bigr] \\
&= \max_i \bigl[\lambda_i - \bar{\mu}\bigl(\xi^* - \varepsilon \|x\|^{-1} x\bigr)\bigr] \\
&\geq \delta.
\end{aligned}$$

The claim follows with $c_1 = \delta/\varepsilon$. $\square$

We have already argued that if $x \in \Sigma$ is defined by setting $x_{ij} = -\nu_j^{-1} \tilde{B}_{ij}^n(t)$, then x satisfies the assumptions of Lemma 3.2, equivalently Lemma 4.3. As a consequence, so does $y := \int_0^{\tau_n} x(t)\,dt$. By (51), on the event $\tau_n < u$, we have that, for some $(i, j) \in \mathcal{E}$, $|y_{ij}| = \nu_j^{-1} \varrho_n \geq \varrho_n$, and so $\|y\| \geq \varrho_n$. Applying Lemma 4.3 yields that there exists $i^* \in \mathcal{I}$ such that

$$\bar{\mu}_{i^*}(y) \geq c_1 \|y\| \geq c_1 \varrho_n.$$

Namely,

$$\sum_j \mu_{i^* j} \int_0^{\tau_n} \tilde{B}_{i^* j}^n(t)\,dt \leq -c_1 \varrho_n.$$

Invoking (34) and using the nonnegativity of $\hat{Q}_i^n(0)$, we obtain

$$\hat{Q}_{i^*}^n(\tau_n) \geq W_{i^*}^n(\tau_n) + c_1 \varrho_n - \sum_j \varepsilon_{i^* j}^n \varrho_n - \sum_j \hat{B}_{i^* j}^n(\tau_n), \tag{74}$$

where we used the fact that $|\int_0^{\tau_n} \tilde{B}_{ij}^n(t)\,dt| \leq \varrho_n$, by (51). By (3), $\sum_j N_j^n \leq c_2 n^{1/2}$, for some constant $c_2 > 0$. Hence

$$\sum_j \hat{B}_{ij}^n(t) \leq n^{-1/2} \sum_j \tilde{B}_{ij}^n(t) \leq n^{-1/2} \sum_j B_{ij}^n(t) \leq n^{-1/2} \sum_j N_j^n \leq c_2, \qquad i \in \mathcal{I}. \tag{75}$$

Also, by (50), for $n \geq n_0$, the third term on the RHS of (74) is bounded by 1, some deterministic n_0. As a result, for a suitable constant $c_3 > 0$, we have, for $n \geq n_0$,

$$\hat{Q}_{i^*}^n(\tau_n) \geq W_{i^*}^n(\tau_n) + c_3 \varrho_n.$$

Recall $\varepsilon_M^n = \max_{i,j} |\varepsilon_{ij}^n|$ and let $t_n = c_3 \rho_n c_2^{-1} (\varepsilon_M^n)^{-1} n^{-1/2}/2$ (where $t_n = \infty$ if $\varepsilon_M^n = 0$). Using (34), and the bound

$$\sum_j \varepsilon_{ij}^n \tilde{B}_{ij}^n \le c_2 \varepsilon_M^n n^{1/2},$$

we have

$$\begin{aligned} \hat{Q}_{i*}^n(t) &\ge W_{i*}^n(\tau_n) + c_3 \varrho_n - c_2 \varepsilon_M^n n^{1/2}(t - \tau_n) \\ &\ge W_{i*}^n(\tau_n) + \frac{c_3}{2} \varrho_n, \qquad t \in [\tau_n, \tilde{\tau}_n], \end{aligned}$$

where, as in the statement of Proposition 4.1, $\tilde{\tau}_n = (\tau_n + t_n) \wedge u$. By the nonnegativity of $\hat{Q}_i^n$ and positivity of θ_i, $i \in \mathcal{I}$, we have, with $\theta_m = \min_i \theta_i$,

$$E^n(t) \ge \theta_{i*} \hat{Q}^n(t) \ge \theta_m \left(\frac{c_3}{2} \varrho_n - \| W^n(\tau_n) \| \right), \qquad t \in [\tau_n, \tilde{\tau}_n]. \tag{76}$$

This shows that (54) is valid on $\Omega^n := \{\|W^n(\tau_n)\| < c_3 \varrho_n/4\} \cap \{n \ge n_0\}$.

Step 3: *Convergence*. We are now in a position to prove all statements of the proposition. By (69), (52) holds up to τ^n; by (72), this estimate remains valid up to u. Moreover, it is shown in (75) that $\hat{B}^n$ are bounded above, and it can similarly be shown that they are bounded below. Hence $H^{*,n}(t)$ are bounded uniformly in n and t. Note that $\varrho_n t_n = c \varrho_n^2 (\varepsilon_M^n)^{-1} n^{-1/2}$, for some constant c. By (50), this product converges to ∞. Thus to complete the proof, it remains to argue that $1_{\Omega^n} \to 1$ a.s. and $P^{*,n} \to Q^*$ a.s.

Toward this end, denote

$$\bar{T}_{ij}(t) = \bar{\mu}_{ij} \xi_{ij}^* t, \qquad t \ge 0, (i,j) \in \mathcal{E}. \tag{77}$$

Let us first show that

$$|\bar{T}_{ij}^n - \bar{T}_{ij}|_{\tau_n} \to 0 \qquad \text{in probability.} \tag{78}$$

Indeed, by (23), (26) and (5),

$$\bar{T}_{ij}^n(t) = n^{-1} \mu_{ij}^n \xi_{ij}^* N_j^n t + n^{-1} \mu_{ij}^n \int_0^t \tilde{B}_{ij}^n(s)\, ds = \bar{\mu}_{ij}^n \xi_{ij}^* t + n^{-1} \mu_{ij}^n \int_0^t \tilde{B}_{ij}^n(s)\, ds.$$

Hence by (51) and (77),

$$|\bar{T}_{ij}^n(t) - \bar{T}_{ij}(t)| \le \hat{\delta}_n := |\bar{\mu}_{ij}^n - \bar{\mu}_{ij}| \xi_{ij}^* u + n^{-1} \mu_{ij}^n \varrho_n, \qquad t \in [0, \tau_n]. \tag{79}$$

Using the convergence (5) and recalling that μ_{ij}^n is asymptotic to $\mu_{ij} n^{1/2}$, it follows by (50) that $\hat{\delta}_n$, defined in the above display, converges to zero. This shows (78).

Recall the process W from Lemma 3.1(iii). We now argue that

$$\|W^n - W\|_{\tau_n} \to 0 \qquad \text{a.s.} \tag{80}$$

By (27), (29) and (36),

$$W_i^n(t) - \hat{W}_i^n(t) = -\sum_j [\hat{S}_{ij}^n(\bar{T}_{ij}^n(t)) - \hat{S}_{ij}^n(\bar{\mu}_{ij}\xi_{ij}^* t)]$$
$$= -\sum_j [\hat{S}_{ij}^n(\bar{T}_{ij}^n(t)) - \hat{S}_{ij}^n(\bar{T}_{ij}(t))].$$

Hence by (79),

$$|W_i^n(t) - \hat{W}_i^n(t)| \le \sum_j \bar{w}_{u_1}(\hat{S}_{ij}^n, \hat{\delta}_n), \qquad t \in [0, \tau_n],$$

where $u_1 = \max_{(i,j)\in\mathcal{E}} \bar{T}_{ij}(u) + 1$. By Lemma 3.1(i), the processes $\hat{S}_{ij}^n$ converge, uniformly on compacts, to processes with continuous sample paths. Hence the RHS of the above display converges to zero. Applying Lemma 3.1(iii), we conclude (80).

It follows from (80) that $1_{\Omega^n} \to 1$ a.s.

Next, by definition of Γ, for any $v > 0$, the mapping $x|_{[0,v]} \mapsto \Gamma[x]|_{[0,v]}$ is Lipschitz continuous in the sup norm, with constant 2. Recalling the definition of $P^{*,n}$ (71), we obtain, for some constant c,

$$\|P^{*,n} - Q^*\|_u \le |\delta_n| + c\|F^n - \theta' X_0 - \theta' W\|_{\tau_n} + c\|\hat{X}^n(0) - X_0\| + c\|\hat{W}^n - W\|_u.$$

We have already argued that the last two terms above converge to zero a.s. [Lemma 3.1 and Remark 3.1(a)]. By (65), for some constant c_1,

$$\|F^n - \theta' X_0 - \theta' W\|_{\tau_n} \le c_1\|\hat{X}^n(0) - X_0\| + c_1\|W^n - W\|_u + c_1\varepsilon_M^n \varrho_n \to 0 \qquad \text{a.s.},$$

where we used (50) and (80). By (50) and (70), $\bar{\varepsilon}_n \to 0$, hence the convergence in the above display implies that $\delta_n \to 0$ a.s. This shows the convergence of $P^{n,*}$ to Q^*, and completes the proof.

5. Asymptotic optimality in heavy-traffic.

5.1. *The tracking policy.* In this section we devise a sequence of controls that asymptotically achieve the lower bound in Theorem 2.1. To construct them and prove their asymptotic optimality we need some further assumptions. Recall from (18) that

$$C_*(a) = \inf\{C(q) : q \in \mathbb{R}_+^{\mathbf{I}}, \theta' q = a\}. \tag{81}$$

ASSUMPTION 5.1 (Continuous minimizer). There exists a locally Lipschitz function $f : \mathbb{R}_+ \to \mathbb{R}_+^I$ such that $\theta' f(a) = a$ and $C(f(a)) = C_*(a)$ for all $a \ge 0$.

We extend the function f to the real line by setting it to zero on $(-\infty, 0)$ (note that the extended function is continuous). In fact, the actual implementation of the

policy will be based on small perturbations of f, that we denote by f^n. These perturbations are explicitly provided below; see (84).

The policies we construct will not use the nonbasic activities at all. It will be convenient, in terms of notation, to disregard these activities by assuming they do not exist. Thus we assume (w.l.o.g., as far as the results of this section are concerned) that *all activities in the model are basic*. In particular, $\mathcal{G} = \mathcal{G}_b$, and $i \sim j$ means (i, j) is an activity, equivalently, a basic activity. Since Assumptions 2.1 and 2.2 are still in force, the graph is a tree.

Based on the structure of the control which achieves the lower bound for the *diffusion* control problem (Proposition 3.1), we seek control policies for the queueing model having two main properties. Namely, (i) that the sequence $\theta' \hat{Q}^n$ converges in law to the RBM Q^*, and (ii) that, given $\theta' \hat{Q}^n = a$, $\hat{Q}^n$ itself is close to the minimizing f in (81). To be more precise about (ii), denote

$$\check{X}^n = f^n(\theta' \hat{X}^n) \quad \text{and} \quad \check{Q}^n = f^n(\theta' \hat{Q}^n). \tag{82}$$

Further, note by (31) that $\|\hat{X}^n - \hat{Q}^n\| \le \|\hat{B}^n\|$. Then property (ii) corresponds to having $\hat{Q}^n - \check{Q}^n \Rightarrow 0$. If $\hat{B}^n \Rightarrow 0$, then the above can be achieved by having $\hat{Q}^n - \check{X}^n \Rightarrow 0$. It turns out to be more convenient to work with the latter, that is, to prove $\hat{B}^n \Rightarrow 0$ and $\hat{Q}^n - \check{X}^n \Rightarrow 0$.

The proposed policy, to which we refer as *the tracking policy*, seeks to achieve the convergence $\hat{Q}^n - \check{X}^n \Rightarrow 0$ by letting $\hat{Q}^n$ track $\check{X}^n$. That is, upon service completion in pool j at time t, the policy assigns to the newly-available server a customer from a class within the set

$$\{i : i \sim j, \hat{Q}^n_i(t-) > \check{X}^n_i(t-)\}, \tag{83}$$

so as to decrease the difference. This, however, is not a precise description of the policy, as the choice of the class for service in (83) will not be arbitrary. It will rely on the structure of the tree. For a precise statement we need some additional notation.

Let $\mathcal{V} = \mathcal{I} \cup \mathcal{J}$ denote the vertex set of $\mathcal{G}$. Pick any $i_0 \in \mathcal{I}$ and designate it as the root. Denote by $d(k)$ the graph distance of a node $k \in \mathcal{V}$ from the root. For $i \in \mathcal{I} \setminus \{i_0\}$ denote by $\bar{\jmath}(i)$ the neighbor $j \sim i$ that is closer to the root than i, and by $\mathcal{J}(i)$ the (possibly empty) set of neighbors $j \sim i$ that are farther from the root than i. We sometimes refer to $\bar{\jmath}(i)$ and $\mathcal{J}(i)$ as the nodes right *above* and, respectively, *below* i, thinking of the tree as being depicted with the root at the top. For nodes $j \in \mathcal{J}$, define analogously $\bar{\imath}(j)$ and $\mathcal{I}(j)$. Next, fix a labeling of all graph nodes by distinct numbers between 1 and $\mathbf{I} + \mathbf{J}$, so that members in $\mathcal{I}$ have the labels $\{1, 2, \ldots, \mathbf{I}\}$ and those in $\mathcal{J}$ have the labels $\{\mathbf{I} + 1, \mathbf{I} + 2, \ldots, \mathbf{I} + \mathbf{J}\}$. Identify the set $\mathcal{V}$ with $\{1, \ldots, \mathbf{I} + \mathbf{J}\}$ accordingly. The labeling should satisfy the following additional condition, namely for $i_1, i_2 \in \mathcal{I}$:

$$d(i_1) < d(i_2) \text{ implies } i_1 > i_2,$$

and for $j_1, j_2 \in \mathcal{J}$,

$$d(j_1) < d(j_2) \text{ implies } j_1 > j_2.$$

We let $j_0 = \max\{j : j \sim i_0\}$. Some of our statements preclude the root i_0. We write $\mathcal{I}_{-i_0} = \mathcal{I} \setminus \{i_0\}$ and, for a vector $x \in \mathbb{R}^{\mathbf{I}}$, we write $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{\mathbf{I}})$.

Perturbed functions: As alluded to earlier, we will work with a perturbed version, f^n, of the function f. Fix a function f that satisfies Assumption 5.1. Let sequences $\{\kappa_n\}$ and $\{\bar{\kappa}_n\}$ with $\kappa_n/\bar{\kappa}_n \to 0$ and $\bar{\kappa}_n \to 0$ be given. For $i \in \mathcal{I}_{-i_0}$ we set

$$f_i^n(x) = \begin{cases} (\mathbf{I}\theta_i)^{-1}x, & x \in [0, \kappa_n), \\ (\mathbf{I}\theta_i)^{-1}\kappa_n, & x \in [\kappa_n, \bar{\kappa}_n), \\ f_i(x)(1 - \bar{\kappa}_n/x) + (\mathbf{I}\theta_i)^{-1}\kappa_n, & x \in [\bar{\kappa}_n, \infty). \end{cases} \tag{84}$$

Also set $f_{i_0}^n(x) = (x - \sum_{i \neq i_0} \theta_i f_i^n(x))/\theta_{i_0}$.

REMARK 5.1 (Properties of f^n). The perturbed functions f^n are such that, for $i \in \mathcal{I}_{-i_0}$, f_i^n is small but strictly positive in the vicinity of 0. Thus, if one can guarantee that $\hat{Q}_i^n \approx f_i^n(\theta' \hat{X}^n)$, then $\hat{Q}_i^n > 0$ for all $i \in \mathcal{I}$ whenever $\theta' \hat{X}^n > 0$, thus there are no idling servers. This property is used in the proof. Additional observations regarding f^n will be useful in what follows: $f_i^n(x) \geq \kappa_n$ for $x \geq \kappa_n$ and $i \in \mathcal{I}_{-i_0}$. Also, $f_{i_0}^n(x) \geq (\mathbf{I}\theta_{i_0})^{-1}x$ for all $x \in [0, \bar{\kappa}_n)$ and $f_{i_0}^n(x) \geq (\theta_{i_0})^{-1}\bar{\kappa}_n \geq (\mathbf{I}\theta_{i_0})^{-1}\bar{\kappa}_n$ for all $x \geq \bar{\kappa}_n$. Since $\theta_i f_i^n(x) \leq x$ for all $i \in \mathcal{I}$ and $x \geq 0$, we have that

$$\sup_{x \geq 0} |\theta_i f_i^n(x) - \theta_i f_i(x)| \leq \bar{\kappa}_n + \sup_{x \geq \bar{\kappa}_n} |\theta_i f_i^n(x) - \theta_i f_i(x)| \leq 2\bar{\kappa}_n \to 0, \quad i \in \mathcal{I}. \tag{85}$$

Finally, it is easy to verify that f^n are locally Lipschitz uniformly in n.

Recall that $I_j^n(t)$ is the number of idle servers in pool j at time t.

The tracking policy:

(i) Upon each arrival of a class-i customer, say at time t, if there are idle servers in one of the pools $j \in \mathcal{J}(i)$, then it is routed to the pool

$$\min\{j \in \mathcal{J}(i) : I_j^n(t-) > 0\}.$$

Otherwise it is queued [even if there are idle servers in pool $\bar{\jmath}(i)$].

(ii) Upon each service completion in pool j, say at time t, the server admits to service a customer from class

$$\min \mathcal{K}_j(t-),$$

where

$$\mathcal{K}_j(t-) := \{k \in \mathcal{I}(j) : \hat{Q}_k^n(t-) > \check{X}_k^n(t-)\},$$

provided this set is nonempty. If $\mathcal{K}_j(t-) = \varnothing$ but $\hat{Q}^n_{\bar{\imath}(j)}(t-) > 0$ the server admits to service a customer from class $\bar{\imath}(j)$. Otherwise, the server remains idle.

REMARK 5.2 (Work conservation). It is clear that the policy is not work conserving. However, it is not hard to see that at any given time t, all servers in pool j must be busy at t, provided $Q^n_i(t) > 0$ for *all* $i \sim j$. Note that this means $\sum_i B^n_{ij}(t) = N_j$ [and, in turn, $\sum_i \tilde{B}^n_{ij}(t) = 0$]. This property will be useful in what follows.

The main result of this section is the following. For simplicity, the initial conditions for $\hat{B}^n$ and $\hat{Q}^n$ are assumed to vanish.

THEOREM 5.1. *Suppose that Assumptions 2.1, 2.2 and 5.1 hold and that* $\hat{B}^n(0) = \hat{Q}^n(0) = 0$ *for all* n. *Then, under the tracking policy,*

$$(\hat{Q}^n - \check{Q}^n, \hat{B}^n, \theta'\hat{Q}^n) \Rightarrow (0, 0, Q^*), \tag{86}$$

where Q^* *is as in Theorem 2.1. Consequently,*

$$\int_0^u C(\hat{Q}^n(t))\,dt \Rightarrow \int_0^u C_*(Q^*(t))\,dt,$$

where $C_*(\cdot)$ *is as in* (18).

REMARK 5.3 (Special cost structures). *Separable convex costs*: Consider, in terms of costs, the setting of [23]. This is the case that $C(q) = \sum_{i\in\mathcal{I}} C_i(q_i)$ where C_i, $i \in \mathcal{I}$, are twice continuously differentiable strictly increasing and strictly convex functions with $C_i'(0) = 0$ for all $i \in \mathcal{I}$. Fixing the constant a, under the Kuhn–Tucker conditions for (81), the unique solution $f(a)$ must satisfy

$$C_i'(f_i(a)) = -y(a)\theta_i,$$

where $y(a)$ is the Lagrange multiplier of the constraint $\theta'q = a$. Thus $f_i(a) = (C_i')^{-1}(-y(a)\theta_i)$. It can be verified that $f(a)$ is locally Lipschitz and, in particular, that it satisfies Assumption 5.1. Moreover, since C_i is strictly increasing and strictly convex, $f_i(a)$ is a strictly increasing function.

Recall from (44) that $\theta_i\mu_{ij} = z^*_j$ for all $(i, j) \in \mathcal{E}_b$. The Kuhn–Tucker condition is equivalently written as

$$\mu_{ij}C_i'(f_i(a)) = -y(a)z^*_j \qquad \text{for all } (i, j) \in \mathcal{E}_b,$$

and, in particular, for all $a \geq 0$,

$$\mu_{ij}C_i'(f_i(a)) = \mu_{kj}C_k'(f_k(a)) \qquad \text{for all } j \text{ and } i, k \in \mathcal{I}(j) \cup \bar{\imath}(j).$$

Thus, our policy is consistent with the $Gc\mu$ policy in [23] in that it aims at setting

$$\mu_{ij}C_i'(\hat{Q}^n_i(t)) \approx \mu_{kj}C_k'(\hat{Q}^n_k(t)) \qquad \text{for all } j \text{ and } i, k \in \mathcal{I}(j) \cup \bar{\imath}(j).$$

The actual implementation is, however, different. Our service mechanism follows the tree structure which is contrasted with the $Gc\mu$ rule that would serve upon service completion a class in the set

$$\arg\max_{i\in\mathcal{I}(j)\cup\bar{i}(j)} \mu_{ij}C_i'\big(\hat{Q}_i^n(t)\big).$$

Linear costs: Suppose that $C(q)=\sum_i c_i q_i$ where $c_i, i\in\mathcal{I}$, are positive constants with $c_1\geq c_2\geq\cdots\geq c_{\mathbf{I}}$, and let us designate class $\mathbf{I}$ as the root of the tree. In this case, (81) has a trivial solution $f_i(x)=0$ for all $i<\mathbf{I}$ and $f_{\mathbf{I}}(x)=x/\theta_{\mathbf{I}}$. Since $\kappa_n\to 0$ as $n\to\infty$, Theorem 5.1 guarantees that $\hat{Q}_i^n\Rightarrow 0$ for all $i<\mathbf{I}$ and $\theta_{\mathbf{I}}\hat{Q}_{\mathbf{I}}^n-\theta'\hat{Q}^n\Rightarrow 0$, so that all queues except for the lowest-cost queue are close to zero at diffusion scale.

Our tracking policy is here a tree-based threshold policy as is the one studied in [8]. If t is such that $\theta'\hat{X}^n(t)>\kappa_n$, then available servers give priority to classes $i<\mathbf{I}$ that exceed their threshold, that is, with $\hat{Q}_i^n(t)\geq(\mathbf{I}\theta_i)^{-1}\kappa_n$.

In the rest of this section we prove Theorem 5.1. All symbols with superscript n, such as $\hat{Q}^n$ and $\hat{X}^n$, denote the respective processes under the tracking policy.

For sequences a_n and b_n of positive numbers, satisfying $b_n/a_n>n^c$ for some $c>0$ and all large n, we write $a_n\ll b_n$. Set $\varrho^n=n^{3/16}$ (satisfying the conditions we put on ϱ^n in the previous section). Recall that $\varepsilon_M^n=\max_{i,j}|\varepsilon_{ij}^n|=O(n^{-1/4})$ (Lemma 3.1), by which $\varepsilon_M^n\varrho^n\ll 1$.

By assumption, the inter-arrival times of all the processes $(A_i, i\in\mathcal{I})$ have finite moments of order $r>2$. Let $\alpha_A=(1/2-1/r)$. Fix sequences p_n, q_n, r_n, s_n of positive numbers, satisfying

$$\varepsilon_M^n\varrho^n\vee n^{-\alpha_A}\ll p_n\ll q_n\ll r_n\ll s_n\ll 1. \tag{87}$$

Assume, without loss of generality, that p_n is given by $n^{-\alpha_g}$ where α_g is a constant. Let $\omega:[0,\infty)\to[0,\infty)$ be given by $\omega(x)=x^{\alpha_\omega}$, where $\alpha_\omega\in(1/3,1/2)$ is a constant. We further assume, without loss of generality, that $\alpha_\omega>2\alpha_g$.

In (84) κ_n and $\bar{\kappa}_n$ are such that

$$p_n\ll\kappa_n\ll q_n\quad\text{and}\quad s_n\ll\bar{\kappa}_n\ll 1.$$

For the remainder of this section we fix the time horizon u. Recall the processes $\hat{A}^n$, V^n and W^n, defined in (24), (27) and (29), respectively.

LEMMA 5.1. *For every $\varepsilon>0$ there exist $K, L>0$ and n_0 such that for $n\geq n_0$,*

$$\mathbb{P}\big\{\text{there exist } 0\leq s\leq t\leq u \text{ such that } \big\|\Lambda^n(t)-\Lambda^n(s)\big\|>p_n+L\omega(t-s)\big\}\leq\varepsilon$$

and

$$\mathbb{P}\big\{\|\Lambda^n\|_u>K\big\}\leq\varepsilon,$$

where Λ^n is any one of the processes $\hat{A}^n$, V^n and W^n.

PROOF. For the processes $\hat{A}^n$ and $\hat{S}^n$, the result follows from strong approximations for renewal processes (see, e.g., Theorem 2.1.2 in [13]) and the Hölder continuity of Brownian motion paths. For V^n (and consequently for W^n), the result thus follows from (27), using the uniform Lipschitz property of the processes $\bar{T}^n_{ij}$ (note that $\mu^n_{ij} B^n_{ij} \le cn$ where c is constant). $\square$

When fixing $\varepsilon > 0$ we will, for simplicity of presentation, assume that $n \ge n_0(\varepsilon)$.

Given $\varepsilon > 0$, let $L = L_\varepsilon$ be as in Lemma 5.1. Recall the process G^n (66). We define the following random times:

$$\sigma^n := \inf\{t \ge 0 : \|\hat{B}^n(t)\| \ge s_n\} \wedge u$$

and

$$\zeta^n := \inf\{t \ge 0 : G^n(s_2) - G^n(s_1) \ge r_n + 4L\omega(s_2 - s_1) \text{ for some } 0 \le s_1 \le s_2 \le t\} \wedge u.$$

Finally, let τ^n be as in (51) and define $T^n = T^n(\varepsilon)$ by

$$T^n := \sigma^n \wedge \tau^n \wedge \zeta^n. \tag{88}$$

PROPOSITION 5.1. *Under the assumptions of Theorem 5.1, given $\varepsilon > 0$ there exists a constant $\bar{K}$ such that*

$$\limsup_{n\to\infty} \mathbb{P}\{\|\hat{Q}^n_{-i_0} - \check{X}^n_{-i_0}\|_{T^n} > \bar{K} p_n\} < \varepsilon \tag{89}$$

and

$$\limsup_{n\to\infty} \mathbb{P}\{|H^n|_{T^n} + \|\hat{B}^n\|_{T^n} + \|\hat{Q}^n - \check{Q}^n\|_{T^n} > \bar{K} q_n\} < \varepsilon. \tag{90}$$

This result is proved in the next subsection.

PROOF OF THEOREM 5.1. For $x : [0, u] \to \mathbb{R}$ denote

$$\mathrm{Osc}(x, [s, t]) := \sup_{s \le t_1 \le t_2 \le t} |x(t_1) - x(t_2)|.$$

Fix $\varepsilon > 0$, and let

$$\begin{aligned} \check{\Omega}^n = {} & \{|H^n|_{T^n} + \|\hat{B}^n\|_{T^n} + \|\hat{Q}^n - \check{Q}^n\|_{T^n} \le \bar{K} q_n\} \\ & \cap \{\|\hat{Q}^n_{-i_0} - \check{X}^n_{-i_0}\|_{T^n} \le \bar{K} p_n\} \\ & \cap \{\mathrm{Osc}(\theta' W^n, [s, t]) \le p_n + L\omega(t - s), 0 \le s < t \le u\} \\ & \cap \{\|W^n\|_u \le K\}. \end{aligned} \tag{91}$$

Using Proposition 5.1 and Lemma 5.1,

$$\mathbb{P}\{\check{\Omega}^n\} \geq 1 - 4\varepsilon, \tag{92}$$

provided n is sufficiently large. We begin by showing that

$$T^n = u \qquad \text{on } \check{\Omega}^n. \tag{93}$$

Let $M = \min_{i\in\mathcal{I}}(\mathbf{I}\theta_i)^{-1}$. Given $t \geq 0$, we argue that, on $\check{\Omega}^n$, for $t < T^n$, and all n sufficiently large,

$$\theta'\hat{Q}^n(t) \geq \varepsilon_n := \bigl(1 + 2M^{-1}\bigr)\bar{K}q_n \text{ implies } \sum_i \tilde{B}^n_{ij}(t) = 0 \qquad \text{for all } j. \tag{94}$$

To see this note that if $\theta'\hat{Q}^n(t) \geq \varepsilon_n$, then on $\check{\Omega}^n$,

$$\hat{Q}^n_{i_0}(t) \geq \check{Q}^n_{i_0}(t) - \bar{K}q_n = f^n_{i_0}\bigl(\theta'\hat{Q}^n(t)\bigr) - \bar{K}q_n \geq 2\bar{K}q_n - \bar{K}q_n > 0,$$

where we use the fact that $f^n_{i_0}(x) \geq M(x \wedge \bar{\kappa}_n)$ for all $x \geq 0$; see Remark 5.1. Further, on $\check{\Omega}^n$, $\theta'\hat{X}^n(t) = \theta'\hat{Q}^n(t) - H^n(t) \geq \varepsilon_n - \bar{K}q_n = 2M^{-1}\bar{K}q_n$ and $\|\hat{Q}^n_{-i_0} - \check{X}^n_{-i_0}\|_{T^n} \leq \bar{K}p_n$ so that, for $i \in \mathcal{I}_{-i_0}$,

$$\hat{Q}^n_i(t) \geq \check{X}^n_i(t) - \bar{K}p_n = f^n_i\bigl(\theta'\hat{X}^n(t)\bigr) - \bar{K}p_n \geq M\kappa_n - \bar{K}p_n > 0,$$

where the inequalities follow from the fact that $p_n/\kappa_n \to 0$, and $f^n_i(x) \geq M\kappa_n$ for all $x \geq \kappa_n$.

Thus, we have that $\hat{Q}^n_i(t) > 0$ for all $i \in \mathcal{I}$ and the claim (94) follows by Remark 5.2.

We now argue that G^n (66) remains constant when $\sum_i \tilde{B}^n_{ij}(t) = 0$ for all j. Given such t, let $x \in \Xi$ be defined by $x_{ij} = -\nu_j^{-1}\tilde{B}^n_{ij}(t)$. Then $\sum_i x_{ij} = 0$ for all $j \in \mathcal{J}$. Recalling that we do not have nonbasic activities, we conclude by the second part of Lemma 3.2 that $\frac{d}{dt}G^n(t) = -\sum_{i\sim j}\theta_i\mu_{ij}\tilde{B}^n_{ij}(t) = 0$.

Combining the above argument with (94), we have on $\check{\Omega}^n$,

$$\int_0^{T^n} 1_{(\varepsilon_n,\infty)}\bigl(\theta'\hat{Q}^n(t)\bigr)\,dG^n(t) = 0. \tag{95}$$

By (34),

$$\theta'\hat{Q}^n(t) = \tilde{W}^n(t) + G^n(t), \tag{96}$$

where

$$\tilde{W}^n(t) = \theta'\hat{X}^n(0) + \theta'W^n(t) - R^n(t) + H^n(t), \tag{97}$$

$$R^n(t) = \sum_i \theta_i \sum_j \varepsilon^n_{ij}\int_0^t \tilde{B}^n_{ij}(s)\,ds \tag{98}$$

and H^n is as in (66).

Thus, on $\check{\Omega}^n$ and for $t \in [0, T^n]$, the triplet $(\theta' \hat{Q}^n, \tilde{W}^n, G^n)$ satisfies the following relations: namely, $G^n(0) = 0$; as shown in the proof of Proposition 4.1, G^n is nondecreasing; and since $\hat{Q}^n$ takes values in the positive orthant, $\theta' \hat{Q}^n \geq 0$. With these properties, along with (95) and (96), we are in a position to apply the oscillation inequalities from [27], Theorem 5.1, to conclude that

$$\text{Osc}(\theta' \hat{Q}^n, [s,t]) + \text{Osc}(G^n, [s,t]) \leq 4\big(\text{Osc}(\tilde{W}^n, [s,t]) + \varepsilon_n\big) \tag{99}$$

for all $0 \leq s \leq t \leq T^n$. The precise constant in the inequality follows the proof of the one-dimensional case [27], page 15.

To prove (93) we use the oscillation inequality to show first that $T^n = \tau^n$. We then show that $\tau^n = u$. By the definition of T^n, using (51), we have

$$\|R^n\|_{T^n} \leq J\|\theta\|\varepsilon_M^n \varrho^n \leq p_n, \tag{100}$$

and using the definition of $\check{\Omega}^n$, we have on $\check{\Omega}^n$,

$$|H^n|_{T^n} \leq \bar{K} q_n.$$

By Lemma 5.1, $\text{Osc}(\theta' W^n, [s,t]) \leq p_n + L\omega(t-s)$. Hence on $\check{\Omega}^n$,

$$\text{Osc}(\tilde{W}^n, [s,t]) \leq (2 + \bar{K}) q_n + L\omega(t-s). \tag{101}$$

Using (99), we then have

$$\text{Osc}(G^n, [s,t]) \leq 4(2 + \bar{K}) q_n + 4L\omega(t-s)) + 4\varepsilon_n$$

for $s < t \leq T^n$. Recalling that $r_n/q_n \to \infty$, we then must have $\zeta^n \geq \tau^n$ on $\check{\Omega}^n$. Further, since $\|\hat{B}^n\|_{T^n} \leq \bar{K} q_n$ on $\check{\Omega}^n$ and, recalling that $q_n/s_n \to 0$, clearly $\sigma^n \geq T^n$. We conclude that $T^n = \tau^n$ on $\check{\Omega}^n$.

To prove (93) it then remains to show that $\tau^n = u$. Following the same arguments leading to (76), on the event that $\{\tau^n < u\}$ one has $\theta' \hat{Q}^n(\tau^n) \geq c(\varrho_n - K)$, for some constant $c > 0$. Recall that $\varrho_n \to \infty$. On the other hand, using (96), (99), and the bounds on R^n and H^n,

$$|\theta' \hat{Q}^n|_{\tau^n} \leq |\tilde{W}^n|_{\tau^n} + |G^n|_{\tau^n} \leq 5\big(\|\theta\| K + (1 + \bar{K}) q_n\big) + 4\varepsilon_n \tag{102}$$

on $\check{\Omega}^n(\varepsilon)$. Thus $|\theta' \hat{Q}^n|_{\tau^n}$ is bounded on this event. This shows that $\tau^n = u$ on $\check{\Omega}^n$ for all sufficiently large n. We have thus proved that $T^n = \tau^n = u$ on $\check{\Omega}^n$, for large n.

Since ε is arbitrary, using Proposition 5.1, (92) and (100), we obtain

$$|H^n|_u + \|\hat{B}^n\|_u + \|\hat{Q}^n - \check{Q}^n\|_u + |R^n|_u \to 0 \qquad \text{in probability.} \tag{103}$$

Using Lemma 3.1(iii) and arguing along the lines of step 3 of the proof of Proposition 4.1 (showing, in particular, that $\bar{T}^n \to \bar{T}$ in probability) gives $W^n \Rightarrow W$. Using (97), (103) and recalling that we assumed zero initial conditions, we have $\tilde{W}^n \Rightarrow \theta' W$.

We conclude that the process $\theta' \hat{Q}^n$ satisfies $\theta' \hat{Q}^n = \tilde{W}^n + G^n \geq 0$, where $\tilde{W}^n \Rightarrow \theta' W$, and G^n is nondecreasing and satisfies, with probability arbitrarily close to one,

$$\int_0^\infty 1_{(\varepsilon_n, \infty)}\big(\theta' \hat{Q}^n(t)\big)\, dG^n(t) = 0.$$

It is a standard fact that these properties suffice to characterize the limit behavior, namely that $\theta' \hat{Q}^n \Rightarrow \Gamma[\theta' W] = Q^*$; see, for example, the proof of [27], Theorem 4.1. By (82), the uniform convergence of f^n to f, (86) and the continuous mapping theorem, $\check{Q}^n \Rightarrow q(Q^*)$. Hence by (103), $\hat{Q}^n \Rightarrow q(Q^*)$, thus $C(\hat{Q}^n) \Rightarrow C(q(Q^*)) = C_*(Q^*)$. Another application of the continuous mapping theorem gives

$$\int_0^u C\big(\hat{Q}^n(t)\big)\, dt \Rightarrow \int_0^u C_*\big(Q^*(t)\big)\, dt,$$

which completes the proof the theorem. □

5.2. *Proof of Proposition* 5.1. The key idea in the proof is to identify an event occurring with high probability on which the policy self tunes the balance between $\check{X}^n$ and $\hat{Q}^n$: when the process $\hat{X}^n$ goes "out of balance," namely, when $\|\hat{Q}^n_{-i_0}(t) - \check{X}^n_{-i_0}(t)\| > cp_n$, the occupancy process $\hat{B}^n$ re-adjusts quickly so as to pull the process $\hat{Q}^n$ back toward $\check{X}^n$.

Throughout the remainder of the analysis we fix $\varepsilon > 0$; T^n is as in (88). Define

$$\begin{aligned} \Omega_1^n &= \big\{|\theta' \hat{X}^n|_{T^n} \leq K\big\}, \\ \Omega_2^n &= \Big\{ \max_{\Lambda^n = \hat{A}^n, V^n, W^n} \big\|\Lambda^n(t) - \Lambda^n(s)\big\| \leq p_n + L\omega(t-s), 0 \leq s \leq t \leq u \Big\}, \end{aligned} \tag{104}$$

where, with an abuse of notation, $K = K(\varepsilon)$ and $L = L(\varepsilon)$ will be chosen (possibly) larger than the values from Lemma 5.1. Let

$$\Omega^n = \Omega_1^n \cap \Omega_2^n. \tag{105}$$

LEMMA 5.2. *Suppose that the assumptions of Theorem* 5.1 *hold. Then K and L can be chosen so that $\mathbb{P}\{\Omega^n\} \geq 1 - \varepsilon$. Moreover, on Ω^n, and for $0 \leq s \leq t \leq T^n$,*

$$\big\|\check{X}^n(t) - \check{X}^n(s)\big\| \leq c\big(\min\big[n^{1/2} s_n(t-s) + p_n, r_n\big] + \omega(t-s)\big), \tag{106}$$

where c is a constant not depending on n, s, t.

Lemma 5.2 is proved in Section 5.5.

Throughout what follows, K and L are as in the above lemma, and fixed, Ω^n is as in (105) and T^n as in (88). Given strictly positive constants $\{c_k^1, k \in \mathcal{I}\}$ define the following times:

$$\tau_{2,k}^n := \inf\big\{s \geq 0 : |\Delta_k^n(s)| > c_k^1 p_n\big\} \wedge T^n, \tag{107}$$

$$\tau_{1,k}^n := \sup\big\{s \leq \tau_{2,k}^n : |\Delta_k^n(s)| \leq c_k^1 p_n/2\big\}, \tag{108}$$

where, throughout,

$$\Delta_k^n = \hat{Q}_k^n - \check{X}_k^n.$$

Let

$$\Omega_{k,U}^n = \Omega^n \cap \{\tau_{2,k}^n < T^n\} \cap \{\Delta_k^n(\tau_{2,k}^n) > c_k^1 p_n\},$$

where U is mnemonic for "up," and define analogously $\Omega_{k,D}^n$, with "$> c_k^1 p_n$" replaced by "$< -c_k^1 p_n$," where D is mnemonic for "down." Note that the jumps of $\hat{Q}^n$ and $\check{X}^n$ are of order $n^{-1/2}$ while $p_n \gg n^{-1/2}$. Moreover, the initial condition is assumed to be zero. As a result, we have on the event $\Omega_{k,U}^n$ (resp., $\Omega_{k,D}^n$) that $\tau_{1,k}^n \in [0, \tau_{2,k}^n)$, and

$$\Delta_k^n(s) \geq c_k^1 p_n/2 \qquad (\text{resp., } \leq -c_k^1 p_n/2) \text{ for all } \tau_{1,k}^n \leq s < \tau_{2,k}^n. \tag{109}$$

The proof of Proposition 5.1 will be based on showing that $\tau_{2,k}^n \geq T^n$ on Ω^n for all $k \in \mathcal{I}_{-i_0}$. This statement is proved inductively.

PROPOSITION 5.2. *Suppose that the assumptions of Proposition 5.1 hold. Then the following holds on the event Ω^n. Let $k \in \mathcal{I}_{-i_0}$ be either 1 or such that, for all $l < k$,*

$$\left|\Delta_l^n\right|_{T^n} \leq c_l^1 p_n \tag{110}$$

for some constants c_l^1, $l < k$. Then there exists a constant c_k^1, such that if $\tau_{2,k}^n$ (107) is defined with c_k^1, then $\tau_{2,k}^n \geq T^n$. Consequently, there exist constants c_k^1, $k \in \mathcal{I}_{-i_0}$ such that, for all $k \in \mathcal{I}_{-i_0}$,

$$\left|\Delta_k^n\right|_{T^n} \leq c_k^1 p_n. \tag{111}$$

Recall (87) and assume, without loss of generality, that $q_n = n^\delta p_n$ for some $\delta > 0$ such that $q_n \ll r_n$. Recall that $\hat{B}_{ij}^n = B_{ij}^n = 0$ if $i \nsim j$.

The next proposition is where the perturbation f^n of the function f is used in an important way.

PROPOSITION 5.3. *Suppose that the assumptions of Proposition 5.1 hold. Then, there exists a constant γ such that, on the event Ω^n,*

$$\left\|\hat{B}^n\right\|_{T^n} \leq \gamma q_n.$$

The proofs of Propositions 5.2 and 5.3 appear in Sections 5.3 and 5.4, respectively. Henceforth we let M_x be a uniform Lipschitz constant of f^n on $[0, x]$. Throughout the proofs we use $c_1, c_2, \ldots$ to denote strictly positive constants that do not depend on n.

PROOF OF PROPOSITION 5.1. Equation (89) follows directly from Proposition 5.2 and the definition of Ω^n.

To prove (90), we will first prove that

$$\|\hat{Q}^n - \check{X}^n\|_{T^n} \le c_1 q_n. \tag{112}$$

To this end, by Proposition 5.3 and identity (9) we have that

$$\|\hat{X}^n - \hat{Q}^n\|_{T^n} \le \|\hat{B}^n\|_{T^n} \le c_2 q_n. \tag{113}$$

By Proposition 5.2 and the fact that $p_n \ll q_n$ we have that

$$\|\hat{Q}^n_{-i_0} - \check{X}^n_{-i_0}\|_{T^n} \le c_3 q_n. \tag{114}$$

By definition $\theta'\hat{X}^n = \theta'\check{X}^n$ whenever $\theta'\hat{X}^n \ge 0$ and $\theta\check{X}^n = 0$ otherwise. Using (113) and the nonnegativity of $\hat{Q}^n$, we get that $\theta'\hat{X}^n(t) \ge -c_4 q_n$ for all $t \le T^n$ and, in turn, that $|\theta'\hat{X}^n - \theta'\check{X}^n|_{T^n} \le c_5 q_n$. Thus

$$|\hat{X}^n_{i_0} - \check{X}^n_{i_0}|_{T^n} \le |\theta'\hat{X}^n - \theta'\check{X}^n|_{T^n} + \|\hat{X}^n_{-i_0} - \check{X}^n_{-i_0}\|_{T^n} \le c_6 q_n.$$

Applying (113) we conclude that $|\hat{Q}^n_{i_0} - \check{X}^n_{i_0}|_{T^n} \le c_7 q_n$ and, together with (114), that (112) holds.

Next, since $|\theta'\hat{X}^n|_{T^n} \le K$ on Ω^n, we have $|\theta'\hat{Q}^n|_{T^n} \le |\theta'\hat{X}^n|_{T^n} + |H^n|_{T^n} \le 2K$. Hence, by Assumption 5.1 and the definition of $\check{Q}^n$ and $\check{X}^n$,

$$\|\check{X}^n - \check{Q}^n\|_{T^n} \le M_{2K}|\theta'\hat{Q}^n - \theta'\hat{X}^n|_{T^n} \le c_8 q_n,$$

where we used (113). Using this along with (112), we have on Ω^n,

$$\|\hat{Q}^n - \check{Q}^n\|_{T^n} \le c_9 q_n. \tag{115}$$

Finally, recall that $-H^n = \sum_{i,j} \theta_i \hat{B}^n_{ij}$ so that

$$|H^n|_{T^n} \le c_{10} q_n, \tag{116}$$

by Proposition 5.3. Combining (115), (116) and Proposition 5.3 we conclude that for any ε there exists a constant c_{11} independent of n, such that

$$\mathbb{P}\{|H^n|_{T^n} + \|\hat{B}^n\|_{T^n} + \|\hat{Q}^n - \check{Q}^n\|_{T^n} > c_{11} q_n\} < \varepsilon.$$

This proves (90). □

5.3. *Proof of Proposition* 5.2. We begin by stating a sequence of lemmas that provide estimates on various properties of the dynamics. They are proved in Section 5.5, along with Lemma 5.2 above. Throughout this and the next subsection, ε is fixed, and the assumptions of Theorem 5.1 are in force. Moreover, the statements of the lemmas are understood to be *on the event* Ω^n. For $f : [0, \infty) \to \mathbb{R}^k$ and $0 \le s \le t$, denote $f[s, t] := f(t) - f(s)$.

The proposition provides estimates concerning the r.v.'s $\tau^n_{2,k}$, that are based on the lemmas below. At the same time, the proof of the proposition involves choosing

the constants c_k^1, used to define these r.v.'s. It will therefore be important to specify which of the estimates, stated in the lemmas (at least those that involve $\tau_{2,k}^n$), depend on c_k^1, and which do not.

Define the processes

$$\mathcal{B}_{ij}^n(t) := \hat{B}_{ij}^n(t) + \mu_{ij}^n \int_0^t \hat{B}_{ij}^n(s)\,ds, \qquad i \in \mathcal{I},\ j \in \mathcal{J}. \tag{117}$$

LEMMA 5.3. *Fix $k \in \mathcal{I}_{-i_0}$. Suppose that $|\Delta_l^n|_{T^n} \le c_l^1 p_n$ for all $l < k$ for some constants c_l^1, $l < k$. Then there exists a constant γ_1, not depending on the constant c_k^1 with which $\tau_{2,k}^n$ is defined, such that, for all $0 \le s \le t \le T^n$,*

$$\big|\mathcal{B}_{kj}^n[s,t]\big| \le \gamma_1\big(n^{1/2}s_n(t-s) + p_n + \omega(t-s)\big), \qquad j \in \mathcal{J}(k). \tag{118}$$

Moreover, there exists a constant γ_2 (that may depend on c_k^1) such that (118) holds for $j = \bar{\jmath}(k)$ and $0 \le s \le t \le \tau_{2,k}^n$, with γ_1 replaced by γ_2.

LEMMA 5.4. *Fix $k \in \mathcal{I}_{-i_0}$. Then there exists a constant γ (that may depend on c_k^1) such that for all $0 \le s \le t \le \tau_{2,k}^n$,*

$$\big|\hat{Q}_k^n[s,t]\big| \le \gamma\big(n^{1/2}s_n(t-s) + p_n + \omega(t-s)\big).$$

To state the last preliminary lemma, let

$$\mathcal{L}(k) = \big\{i \in \mathcal{I} : i \le k,\ i \sim \bar{\jmath}(k)\big\} \tag{119}$$

be the set of customer classes that are not higher than k in the hierarchy and are connected to the parent node $\bar{\jmath}(k)$. Recall that if $\tau_{2,k}^n < T_n$, then one of the two events $\Omega_{k,U}^n$, $\Omega_{k,D}^n$ must occur, and consequently, one of the two inequalities specified in (109) holds. In the former case, any service completion in pool $\bar{\jmath}(k)$ during $[\tau_{1,k}^n, \tau_{2,k}^n)$ is followed, under our tracking policy, with an admission of a customer from one of the queues in the set $\mathcal{L}(k)$. In the latter case, no class-k customers are admitted to pool $\bar{\jmath}(k)$ on $[\tau_{1,k}^n, \tau_{2,k}^n)$. The following lemma is based on these two properties.

LEMMA 5.5. *Fix $k \in \mathcal{I}_{-i_0}$, and let $j = \bar{\jmath}(k)$. Then there exists a constant $\gamma > 0$, not depending on c_k^1, such that the following holds*:

(i) *On $\Omega_{k,U}^n$ one has*

$$\sum_{l \in \mathcal{L}(k)} \mathcal{B}_{lj}^n[s,t] \ge \gamma n^{1/2}(t-s) - L\omega(t-s) - p_n, \qquad \tau_{1,k}^n \le s \le t \le \tau_{2,k}^n. \tag{120}$$

(ii) *On* $\Omega^n_{k,D}$ *one has*

$$\mathcal{B}^n_{kj}[s,t] \le -\gamma n^{1/2}(t-s) + L\omega(t-s) + p_n, \qquad \tau^n_{1,k} \le s \le t \le \tau^n_{2,k}. \tag{121}$$

PROOF OF PROPOSITION 5.2. Note that all statements regard the event Ω^n (105).

It is required to show that there exists a constant c^1_k, with which $\tau^n_{2,k}$ is defined, such that $\tau^n_{2,k} \ge T_n$.

To this end, let us analyze the event $\Omega^n \cap \{\tau^n_{2,k} < T^n\}$, considering separately the two sub-events $\Omega^n_{k,U}$ and $\Omega^n_{k,D}$. The goal is to show that one can choose c^1_k so that the two events are empty, provided n is sufficiently large.

We start with the former. Fix $k \in \mathcal{I}$ [and note that in the case $k = 1$, the set $\mathcal{L}(k)$ is simply $\{1\}$], and denote $j = \bar{\jmath}(k)$. The goal of showing that $\Omega^n_{k,U}$ is empty (for suitable c^1_k and large n) is achieved by arguing that there exists a constant c, not depending on c^1_k or n, such that on $\Omega^n_{k,U}$,

$$\Delta^n_k(t) \le \Delta^n_k\bigl(\tau^n_{1,k}\bigr) + cp_n \qquad \text{for all } t \in \bigl[\tau^n_{1,k}, \tau^n_{2,k}\bigr). \tag{122}$$

To this end, consider any $\tau^n_{1,k} \le s \le t < \tau^n_{2,k}$. Using (34) and (117) we have

$$\begin{aligned}\sum_{l\in\mathcal{L}(k)} \hat{Q}^n_l[s,t] = {} & \sum_{l\in\mathcal{L}(k)} W^n_l[s,t] \\ & - \sum_{l\in\mathcal{L}(k)} \sum_{m\in\mathcal{J}(l)} \mathcal{B}^n_{lm}[s,t] \\ & - \sum_{l\in\mathcal{L}(k)} \mathcal{B}^n_{lj}[s,t],\end{aligned} \tag{123}$$

where we used the fact that $\bar{\jmath}(l) = j$ for any $l \in \mathcal{L}(k)$. By the assumption of the proposition, $|\Delta^n_l|_{T^n} \le c^1_l p_n$, $l < k$, by which $\tau^n_l = T^n$ for $l < k$. Thus in view of Lemma 5.3, estimate (118) is valid for all $l < k$, and for all $0 \le s \le t \le T^n$. Moreover, the constants in this estimate do depend on c^1_l, $l < k$, but not on c^1_k. As a result,

$$\sum_{l\in\mathcal{L}(k)} \sum_{m\in\mathcal{J}(l)} \bigl|\mathcal{B}^n_{lm}[s,t]\bigr| \le c_1\bigl(n^{1/2}s_n(t-s) + p_n + \omega(t-s)\bigr), \tag{124}$$

where c_1 does not depend on c^1_k. We bound the first line of (123) using (104), the second line using (124) and the third line using Lemma 5.5. Here we recall that the constants in Lemma 5.5 do not depend on c^1_k. In turn, we have

$$\begin{aligned}\sum_{l\in\mathcal{L}(k)} \hat{Q}^n_l[s,t] &\le c_2\bigl(p_n + \omega(t-s) + n^{1/2}s_n(t-s)\bigr) - c_3 n^{1/2}(t-s) \\ &\le -c_4 n^{1/2}(t-s) + c_5\bigl(p_n + \omega(t-s)\bigr),\end{aligned} \tag{125}$$

or positive constants $c_2, \dots, c_5$ that do not depend on c_k^1. In the second inequality above we used the fact that $s_n \to 0$. Thus, using Lemma 5.4

$$\sum_{l<k} |\hat{Q}_l^n[s,t]| \le c_8\big(n^{1/2}s_n(t-s) + p_n + \omega(t-s)\big).$$

Above, the constant may depend on c_l^1, $l<k$, but not on c_k^1. Thus by (126) and using the fact that $s_n \to 0$ and applying Lemma 5.2 to bound $|\check{X}_k^n[s,t]|$, we obtain

$$\Delta_k^n(t) = \hat{Q}_k^n(t) - \check{X}_k^n(t) \le \Delta_k^n(s) - c_{11}n^{1/2}(t-s) + c_{12}\big(p_n + \omega(t-s)\big).$$

Note that we have

$$-c_{11}n^{1/2}(t-s) + c_{12}\omega(t-s) \le c_{13}p_n \tag{126}$$

for all sufficiently large n. Indeed, the function $\nu \to -c_{11}n^{1/2}\nu + c_{12}\omega(\nu)$ is concave and it is easily verified that the unique maximum is bounded by $c_{13}n^{-\alpha_\omega/(2(1-\alpha_\omega))}$. Since $\alpha_\omega \in (1/3, 1/2)$ the maximum is further bounded by $c_{14}n^{-1/4} \ll p_n$. Thus, we conclude that (126) holds. Choosing c_k^1 sufficiently large, we then have that $\tau_{2,k}^n \ge T^n$ must hold on the event $\Omega_{k,U}^n$. In other words, $\Omega_{k,U}^n$ is empty.

Next we consider the event $\Omega_{k,D}^n$. Arguing as above, using the second part of Lemma 5.5, we have

$$\hat{Q}_k^n(t) \ge \hat{Q}_k^n(s) - c_{15}\big(\omega(t-s) + n^{1/2}s_n(t-s)\big) + c_{16}n^{1/2}(t-s).$$

Bounding $\check{X}_k^n[s,t]$ and using (106) as before we have

$$\Delta_k^n(t) \ge \Delta_k^n(s) - c_{17}\big(\omega(t-s) + p_n\big) + c_{18}n^{1/2}(t-s).$$

Similarly to the above, $\Delta_k^n(t) \le \Delta_k^n(\tau_{1,k}^n) + c_{19}p_n$ for all $t \in [\tau_{1,k}^n, \tau_{2,k}^n)$. Re-choosing $c_k^1 \ge 4c_{20}$ we then conclude that $\tau_{2,k}^n \ge T^n$ on $\Omega_{k,D}^n$.

This proves the first part of the proposition. The second part is argued inductively using the above. If (110) holds for all $l<k$ (or for the induction basis $k=1$) and since we proved that $\tau_{2,k}^n \ge T^n$, the definition of $\tau_{2,k}^n$ implies that $|\Delta_k^n(s)| \le c_k^1 p_n$ for all $s < T_n$. It is not hard to see that the jumps of both $\hat{Q}^n$ and $\check{X}^n$ are $O(n^{-1/2})$. Recalling that $p_n \gg n^{-1/2}$, it follows that (110) holds for k (with a suitable constant c_k^1). We conclude that there exist constants, that with abuse of notation we still denote by $\{c_k^1, k \in \mathcal{I}\}$, such that, on Ω^n, for all $k \in \mathcal{I}_{-i_0}$, $|\Delta_k^n|_{T^n} \le c_k^1 p_n$. This concludes the proof. □

5.4. *Proof of Proposition 5.3.* We begin by stating a sequence of auxiliary lemmas that are proved in Section 5.5. As before, the statements of the lemmas are understood to be *on the event* Ω^n and the assumptions of Theorem 5.1 are in force. Fixing throughout δ such that $q_n = p_n n^\delta \ll r_n$ we let $\vartheta_n = n^{-(1/2-\delta)}$. Below the constants z_j^*, $j \in \mathcal{J}$, are as in (44).

The following relates the process $\mathcal{B}^n$ to the idleness process.

LEMMA 5.6. *Fix $j \in \mathcal{J}$ with $j \neq j_0$. Suppose that there exists a constant γ_1 such that*

$$\sum_{k \in \mathcal{I}(j)} \left|\mathcal{B}^n_{kj}[s,t]\right| \leq \gamma_1 \big(n^{1/2} s_n(t-s) + p_n + \omega(t-s)\big)$$

for all $0 \leq s \leq t \leq T^n$. Then there exists a constant γ_2 such that $|\hat{I}^n_j|_{T^n} \leq \gamma_2 p_n$.

Consequently, if $|\Delta^n_k|_{T^n} \leq c^1_k p_n$ for $k \in \mathcal{I}_{-i_0}$ and constants c^1_k, $k \in \mathcal{I}_{-i_0}$, then there exists a constant γ_2 such that $|\hat{I}^n_j|_{T^n} \leq \gamma_2 p_n$, for all $j \neq j_0$.

LEMMA 5.7. *There exists a constant γ so that for all $s, t \leq T^n$ with $|t-s| \leq \vartheta_n$,*

$$\left|\theta' \hat{X}^n[s,t] - z^*_{j_0} n^{1/2} \int_s^t \hat{I}^n_{j_0}(u)\, du\right| \leq \gamma q_n$$

and

$$\theta' \hat{X}^n[s,t] \geq -\gamma p_n.$$

LEMMA 5.8. *Fix $k \in \mathcal{I}_{-i_0}$. Then there exists a constant γ such that for all $s, t \leq T^n$ with $|t-s| \leq \vartheta_n$*

$$\left|\check{X}^n_k[s,t]\right| \leq \gamma q_n.$$

LEMMA 5.9. *Fix $k \in \mathcal{I}_{-i_0}$ and $j \neq j_0$. Then there exists a constant γ such that*

$$\left|\hat{B}^n_{kj}\right|_{T^n} \leq \gamma q_n.$$

LEMMA 5.10. *There exists a constant γ such that the following holds. For $s, t \leq T^n$ with $|t-s| \leq \vartheta_n$ and such that $\hat{I}^n_{j_0}(u) > 0$ for all $u \in [s,t)$, we have*

$$\hat{I}^n_{j_0}[s,t] \leq -\check{X}^n_{i_0}[s,t] + \gamma q_n \leq -\frac{1}{2} z^*_{j_0} \int_s^t I^n_{j_0}(u)\, du + \gamma q_n.$$

PROOF OF PROPOSITION 5.3. By Lemma 5.9, $\sum_{j \neq j_0} \sum_{i \neq i_0} |\hat{B}^n_{ij}|_{T^n} \leq c_1 q_n$. By Proposition 5.2 and Lemma 5.6 we have that $\sum_{j \neq j_0} |\hat{I}^n_j|_{T^n} \leq c_2 p_n$. Thus, using the identity $\hat{I}^n_j = -\sum_i \hat{B}^n_{ij}$ and since $p_n \ll q_n$, we have

$$\sum_{j \neq j_0} |\hat{B}^n_{i_0 j}|_{T^n} \leq \sum_{j \neq j_0} |\hat{I}^n_j|_{T^n} + \sum_{j \neq j_0} \sum_{i \in \mathcal{I}(j)} |\hat{B}^n_{ij}| \leq c_3 q_n.$$

To prove the proposition it only remains to show that

$$\left|\hat{I}^n_{j_0}\right|_{T^n} \leq c_4 q_n, \tag{127}$$

in which case we will have by the same argument that $|\hat{B}^n_{i_0 j_0}|_{T^n} \le c_5 q_n$. Together with Lemma 5.9, this would allow us to conclude that $\|\hat{B}^n\|_{T^n} \le c_6 q_n$ as required.

The remainder of the argument is dedicated to the proof of (127). To that end, fix $\zeta > 2\gamma$ with γ as in Lemma 5.10, and let

$$\tau_1^n = \inf\{t \ge 0 : \hat{I}^n_{j_0}(t) > 2\zeta q_n\} \wedge T^n$$

and

$$\tau_0^n = \sup\{t \le \tau_1^n : \hat{I}^n_{j_0}(t) \le \zeta q_n\} \wedge T^n.$$

Argue by contradiction and assume that $\tau_1^n < T^n$ and, in particular, $\tau_0^n < T^n$. Consider the interval $[\tau_0^n, (\tau_0^n + \vartheta_n) \wedge \tau_1^n)$. By Lemma 5.10 it holds, for $s, t \in [\tau_0^n, (\tau_0^n + \vartheta_n) \wedge \tau_1^n)$ that

$$\hat{I}^n_{j_0}[s,t] \le -\frac{1}{2} z^*_{j_0} n^{1/2} \int_s^t \hat{I}^n_{j_0}(u)\, du + \gamma q_n. \tag{128}$$

In particular, $\hat{I}^n_j(u) \le \zeta q_n + \gamma q_n$ for all $u \in [\tau_0^n, (\tau_0^n + \vartheta_n) \wedge \tau_1^n)$, and it must be the case that $\tau_1^n \ge \tau_0^n + \vartheta_n$. Since $\hat{I}^n_{j_0} \ge \zeta q_n$ on $[\tau_0^n, \tau_1^n)$ and $n^{-1/2} \ll \vartheta_n$ we also have by (128) that

$$\hat{I}^n_{j_0}\big(\tau_0^n + 4(z^*_{j_0})^{-1} n^{-1/2}\big) \le 2\zeta q_n - 2\zeta q_n + \gamma q_n \le \gamma q_n.$$

Since $\tau_0^n + 4(z^*_{j_0})^{-1} n^{-1/2} \in (\tau_0^n, \tau_1^n)$ this contradicts the definition of τ_0^n. We conclude that $\tau_1^n \ge T^n$ and, since the jumps of $\hat{I}^n$ are of size $O(n^{-1/2})$ and $n^{-1/2} \ll q_n$, that $|\hat{I}^n_{j_0}|_{T^n} \le 3\zeta q_n$. This establishes (127) and completes the proof of the proposition. □

5.5. *Proofs of auxiliary lemmas.*

PROOF OF LEMMA 5.2. It follows directly from Lemma 5.1 that $\mathbb{P}\{\Omega_2^n\} \ge 1 - \varepsilon/2$ for a sufficiently large L. To treat Ω_1^n, recall that

$$\theta' \hat{X}^n(t) = \theta' \hat{X}^n(0) + \theta' W^n(t) + G^n(t) - R^n(t). \tag{129}$$

By the definition of T^n, it is easy to see that $|R^n|_{T^n}$ and $|G^n|_{T^n}$ are uniformly bounded. Hence using Lemma 5.1, K can be chosen so that $\mathbb{P}\{\Omega_1^n\} \ge 1 - \varepsilon/2$. This shows the first assertion of the lemma.

By (33)

$$\|\hat{X}^n[s,t]\| \le \|W^n[s,t]\| + \sum_{i,j} n^{1/2}(\mu_{ij} + \varepsilon^n_{ij}) \int_s^t |\hat{B}^n_{ij}(s)|\, ds.$$

By the definition of Ω_2^n and the fact that $\|\hat{B}^n\|_{T^n} \le s_n$ and using Assumption 5.1 to write

$$\|\check{X}^n[s,t]\| \le M_K |\theta' \hat{X}^n[s,t]|, \tag{130}$$

it follows that

$$\|\check{X}^n[s,t]\| \le c\big(\sqrt{n}s_n(t-s) + p_n + L\omega(t-s)\big).$$

To prove the result, it remains to show that

$$\|\check{X}^n[s,t]\| \le c\big(r_n + L\omega(t-s)\big). \tag{131}$$

We use (129) and (130). The increment of W^n can be bounded as before, while that of G^n is bounded using the definition of T^n, specifically ζ^n. Thus

$$\|\check{X}^n[s,t]\| \le c\big(r_n + L\omega(t-s)\big) + \big|R^n[s,t]\big|.$$

Moreover, by the definition of τ^n (51) and R^n (98), we have $|R^n| \le c\varepsilon_M^n \varrho^n < p_n < r_n$. As a result (131) holds and the result follows. □

For the proof of Lemma 5.3 we define $A_{kl}^n(t)$ to be the number of class-k customers entering pool l by time t and let its centered and scaled version be given by

$$\hat{A}_{kl}^n(t) = n^{-1/2}\big(A_{kl}^n(t) - \mu_{kl}^n \xi_{kl}^* \nu_l n^{1/2} t\big). \tag{132}$$

PROOF OF LEMMA 5.3. The proof of the lemma proceeds by induction on the class number.

Induction base, $k = 1$. For this class all server pools $j \in \mathcal{J}(k)$ (if there are any) are necessarily leafs of the tree. Thus if j is such a pool, one has $\hat{B}_{kj}^n = -\hat{I}_j^n$ so that, by Lemma 5.6, $|\hat{B}_{kj}^n|_{T^n} \le c_1 p_n$. Thus for $s < t \le \tau_{2,k}^n$,

$$\begin{aligned} \big|\mathcal{B}_{kj}^n[s,t]\big| &\le \big|\hat{B}_{kj}^n\big|_{T^n}\big(2 + \mu_{kj}^n(t-s)\big) \le c_2\big(p_n + n^{1/2}p_n(t-s)\big) \\ &\le c_2\big(p_n + n^{1/2}s_n(t-s)\big), \end{aligned} \tag{133}$$

where we used the fact that $s_n/p_n \to \infty$. Note that c_2 does not depend on c_k^1.

Consider next $j = \bar{\jmath}(k)$. Using (34) we have

$$\begin{aligned} \big|\mathcal{B}_{kj}^n[s,t]\big| &\le \big|\hat{Q}_k^n[s,t]\big| + \sum_{l \in \mathcal{J}(k)} \big|\mathcal{B}_{kl}^n[s,t]\big| + \big\|W^n[s,t]\big\| \\ &\le c_3\big(p_n + n^{1/2}s_n(t-s) + \omega(t-s)\big). \end{aligned}$$

For $s < t \le \tau_{2,k}^n$, we used Lemma 5.4 and (133). Note that c_3 does depend on c_k^1.

Induction step, $k > 1$. Assume that the result of the lemma holds for all $m < k$. Namely, if $m < k$ is such that $|\Delta_l^n|_{T^n} \le c_l^1 p_n$ for all $l < m$, then (118) holds with m replacing k and for all $0 \le s \le t \le T^n$ and all $j \in \mathcal{J}(m)$ with a constant γ_1 that does not depend on c_m^1. It holds for $j = \bar{\jmath}(m)$ up to $\tau_{2,k}^n$ with a constant γ_2 that does depend on c_m^1.

We will show that this holds for k. We thus assume that $|\Delta_m^n|_{T^n} \le c_m^1$ for all $m < k$. By the induction assumption we have the existence of a constant c_1, depending

on $(c_m^1, m<k)$ but not depending on c_k^1, such that (118) holds for all $m<k$ and all $l \sim m$. By the argument leading to (141) we have for all $s,t \le T^n$ that

$$\sum_{m,j<k} |\hat{A}^n_{mj}[s,t]| \le c_4\big(p_n + n^{1/2}s_n(t-s) + \omega(t-s)\big). \tag{134}$$

Considering a pool $l \in \mathcal{J}(k)$, the idleness process satisfies

$$\begin{aligned}\hat{I}^n_l(t) &= \hat{I}^n_l(s) - n^{-1/2}\sum_{m\le k} A^n_{ml}[s,t] + n^{-1/2}\sum_{m\le k} D^n_{ml}[s,t] \\ &= \hat{I}^n_l(s) - \sum_{m\le k}\hat{A}^n_{ml}[s,t] + \sum_{m\le k}\mu^n_{ml}\int_s^t \hat{B}^n_{ml}(v)\,dv + \sum_{m\le k} V^n_{ml}[s,t].\end{aligned}$$

In turn,

$$\begin{aligned}|\hat{A}^n_{kl}[s,t]| \le{}& 2|\hat{I}^n_l|_{T^n} + \sum_{m<k}|\hat{A}^n_{ml}[s,t]| + \sum_{m\le k}|V^n_{ml}[s,t]| \\ &+ \sum_{m\le k}\Big|\mu^n_{ml}\int_s^t \hat{B}^n_{ml}(v)\,dv\Big|.\end{aligned} \tag{135}$$

By the induction assumption (118) holds for all classes $m \in \mathcal{I}(l)$ so that, by Lemma 5.6,

$$|\hat{I}^n_l|_{T^n} \le c_5 p_n \tag{136}$$

for a constant that does not depend on c_k^1. Also, $|\hat{B}^n_{ml}|_{T^n} \le s_n$ by definition so that

$$\sum_{m\le k}\Big|\mu^n_{ml}\int_s^t \hat{B}^n_{ml}(v)\,dv\Big| \le c_7 n^{1/2}s_n(t-s). \tag{137}$$

Thus using (134), (136) and (137) in (135) and applying the definition of Ω^n to bound the increments of V^n, we conclude that

$$|\hat{A}^n_{kl}[s,t]| \le c_8\big(p_n + n^{1/2}s_n(t-s) + \omega(t-s)\big),$$

where c_8 does not depend on c_k^1. By (140) we then have that

$$\begin{aligned}|\mathcal{B}^n_{kl}[s,t]| &\le |\hat{A}^n_{kl}[s,t]| + |V^n_{kj}[s,t]| \\ &\le c_9\big(p_n + \omega(t-s) + n^{1/2}s_n(t-s)\big),\end{aligned}$$

where, as required, the constants do not depend on c_k^1 in the definition of $\tau^n_{2,k}$ (107). This argument is repeated for each $l \in \mathcal{L}(k)$. The argument for the pool $\bar{J}(k)$ then follows exactly as in the induction basis and this concludes the proof. □

PROOF OF LEMMA 5.4. By the definition of $\tau^n_{2,k}$, and since the jumps of both $\hat{Q}^n$ and $\check{X}^n$ are $O(n^{-1/2})$ and recalling that $p_n \gg n^{-1/2}$, we have that $|\Delta^n_k|_{\tau^n_{2,k}} \le 2c_k^1 p_n$. The result is thus an immediate consequence of Lemma 5.2. □

PROOF OF LEMMA 5.5. Recall that k is fixed and $j = \bar{j}(k)$. Consider first the event $\Omega^n_{k,U}$. By (109), Δ^n_k remains positive on the interval $\mathcal{T}_n := [\tau^n_{1,k}, \tau^n_{2,k})$. By part (ii) of the definition of the policy, during this interval all service completions in pool j are followed by admission to service of customers from the classes in $\mathcal{L}(k)$. Also note that part (i) of this definition is irrelevant during this time interval because there are no idle servers at pool j (indeed, the set $\mathcal{K}_j$ is not empty on this interval; if there were any idle servers in pool j then they would be immediately assigned to customers of classes in the set $\mathcal{K}_j$). Thus, for $s, t \in \mathcal{T}_n$, $s < t$, we have

$$\sum_{l\in\mathcal{L}(k)} B^n_{lj}[s,t] = -\sum_{l\in\mathcal{L}(k)} D^n_{lj}[s,t] + \sum_{l:l\sim j} D^n_{lj}[s,t].$$

Using (12) and (27) we rewrite this as

$$\begin{aligned}\sum_{l\in\mathcal{L}(k)} B^n_{lj}[s,t] = &-\sum_{l\in\mathcal{L}(k)} \mu^n_{lj}\int_s^t B^n_{lj}(v)\,dv - \sum_{l\in\mathcal{L}(k)} n^{1/2}V^n_{lj}[s,t]\\ &+\sum_{l:l\sim j} \mu^n_{lj}\int_s^t B^n_{lj}(v)\,dv + \sum_{l:l\sim j} n^{1/2}V^n_{lj}[s,t].\end{aligned}$$

Denote $\mathcal{L}^c(k) = \{l \in \mathcal{I} : l \sim j, l \notin \mathcal{L}(k)\}$. After centering and scaling we have

$$\begin{aligned}\sum_{l\in\mathcal{L}(k)} \mathcal{B}^n_{lj}[s,t] = &\sum_{l\in\mathcal{L}^c(k)} \mu^n_{lj}\xi^*_{lj}\nu_l(t-s) + \sum_{l:l\sim j} \mu^n_{lj}\int_s^t \hat{B}^n_{lj}(v)\,dv\\ &+\sum_{l\in\mathcal{L}^c(k)} V^n_{lj}[s,t].\end{aligned} \tag{138}$$

Note that $i := \bar{\imath}(j) \in \mathcal{L}^c(k)$ so that the first term on the RHS of (138) is bounded below by

$$\mu^n_{ij}\xi^*_{ij}\nu_j(t-s) \geq c_1 n^{1/2}(t-s). \tag{139}$$

Since $\|\hat{B}^n\|_{T^n} \leq s_n$, the second term is bounded, in absolute value, by $c_2 n^{1/2} s_n (t-s)$. Since $s_n \to 0$, this gives

$$\sum_{l\in\mathcal{L}(k)} \mathcal{B}^n_{lj}[s,t] \geq c_3 n^{1/2}(t-s) - \|V^n[s,t]\|.$$

Equation (120) now follows by using the definition of Ω^n to bound the increment of V^n.

Let us now consider $\Omega^n_{k,D}$. To prove (121), note by (109) that $\Delta^n_k < 0$ on the time interval $\mathcal{T}^n$. Note that by the first part of the policy definition, new class-k arrivals are not sent to pool j even if there are idle servers. Moreover, by the second part, upon each service completion, no class-k customers are admitted into service in pool j during this time (since $\Delta^n_k < 0$). Hence

$$B^n_{kj}[s,t] = -D^n_{kj}[s,t] = -\mu^n_{kj}\int_s^t B^n_{kj}(v)\,dv + n^{1/2}V^n_{kj}[s,t],$$

or, after scaling and centering,

$$\mathcal{B}^n_{kj}[s,t] = -\mu^n_{kj}\xi^*_{kj}\nu_j(t-s) + V^n_{kj}[s,t].$$

Exploiting again the bounds for V^n on Ω^n, we have (121). This completes the proof. $\square$

PROOF OF LEMMA 5.6. Fix $j \in \mathcal{J}$, $j \neq j_0$. For $i = i_0$ we will use here, with some abuse of notation, $\mathcal{J}(i_0) = \mathcal{J}(i_0) \setminus \{j_0\}$.

We start with an observation that relates the condition of the lemma to the processes $\hat{A}^n_{kj}$, $k \in \mathcal{I}(j)$. To that end, note that the process $\hat{B}^n_{kj}$ satisfies the relation

$$\begin{aligned}\hat{B}^n_{kj}(t) &= \hat{B}^n_{kj}(0) + n^{-1/2}\big(A^n_{kj}(t) - D^n_{kj}(t)\big)\\ &= \hat{B}^n_{kj}(0) + \hat{A}^n_{kj}(t) - \mu^n_{kj}\int_s^t \hat{B}^n_{kj}(v)\,dv + V^n_{kj}(t).\end{aligned} \tag{140}$$

Hence, using the definition of Ω^n to bound the increment of V^n,

$$\big|\hat{A}^n_{kj}[s,t]\big| \le \big|\mathcal{B}^n_{kj}[s,t]\big| + c_1\big(p_n + \omega(t-s)\big). \tag{141}$$

In view of this and the assumption of the lemma,

$$\big|\hat{A}^n_{kj}[s,t]\big| \le c_2\big(p_n + n^{1/2}s_n(t-s) + \omega(t-s)\big), \qquad k \in \mathcal{I}(j). \tag{142}$$

Let $i = \bar{\imath}(j)$ [note that $j \in \mathcal{J}(i)$] and define

$$\tau^n_2 := \inf\Big\{s \ge 0 : \sum_{l\in\mathcal{J}(i)} \hat{I}^n_l(s) \ge \gamma p_n\Big\} \wedge T^n,$$

$$\tau^n_1 := \sup\Big\{s \le \tau^n_2 : \sum_{l\in\mathcal{J}(i)} \hat{I}^n_l(s) \le \gamma p_n/2\Big\}.$$

Note that for all $s < t$, and each $l \in \mathcal{J}$,

$$\begin{aligned}\hat{I}^n_l[s,t] &= \frac{1}{n^{1/2}}\sum_{k:k\sim l}\big(-A^n_{kl}[s,t] + D^n_{kl}[s,t]\big)\\ &= \sum_{k:k\sim l}\Big(-\hat{A}^n_{kl}[s,t] - \mu^n_{kl}\int_s^t \hat{B}^n_{kl}(v)\,dv - V^n_{kl}[s,t]\Big).\end{aligned} \tag{143}$$

On $[\tau^n_1, \tau^n_2)$ the tracking policy routes all class-i arrivals to pools in the set $\mathcal{J}(i)$. Hence on $[\tau^n_1, \tau^n_2)$, $\sum_{l\in\mathcal{J}(i)} A^n_{il}[s,t] = A^n_i[s,t]$ and

$$\begin{aligned}\sum_{l\in\mathcal{J}(i)} \hat{A}^n_{il}[s,t] &= n^{-1/2}\lambda^n_i(t-s) - \sum_{l\in\mathcal{J}(i)} \mu^n_{il}\xi^*_{il}\nu_j(t-s) + \hat{A}^n_i[s,t]\\ &\ge \frac{c_3}{2}n^{1/2}(t-s) + \hat{A}^n_i[s,t]\end{aligned} \tag{144}$$

for a positive constant c_3. The last inequality follows from the following observation: by (2) and (8) we have that $\lambda_i^n = \lambda_i n + O(n^{1/2}) = \sum_{l\in\mathcal{J}(i)} \mu_{il}\xi_{kl}^* \nu_l n + \sum_{l\notin\mathcal{J}(i)} \mu_{il}\xi_{il}^* \nu_l n + O(n^{1/2})$. If $i \neq i_0$, then by Assumption 2.2, $\xi_{i\bar{J}(i)}^* > 0$. If $i = i_0$, then $\xi_{ij_0}^* > 0$. In either case, there exists $l \notin \mathcal{J}(i)$ with $\xi_{il}^* > 0$. Hence $n^{-1/2}(\lambda_i^n - \sum_{l\in\mathcal{J}(i)} \mu_{il}\xi_{il}^* \nu_j n) \geq c_3 n^{1/2}$ for a positive constant c_3 as required.

Using $\|\hat{B}\|_{T^n} \leq s_n$, (144), (142) and the assumption of the lemma in (143), we have

$$\sum_{l\in\mathcal{J}(i)} \hat{I}_l^n[s,t] \leq c_4\bigl(n^{1/2}s_n(t-s) - n^{1/2}(t-s)\bigr) + \bigl|\hat{A}_i^n[s,t]\bigr| + \bigl\|V^n[s,t]\bigr\|$$

$$\leq -c_5 n^{1/2}(t-s) + c_6\bigl(p_n + \omega(t-s)\bigr)$$

for all $\tau_1^n \leq s \leq t \leq \tau_2^n$. As in (126) we have that $-c_5 n^{1/2}(t-s) + c_6\omega(t-s) + c_6 g(n) \leq c_6 p_n$ for all $s \leq t$ and all sufficiently large n and, in turn, that $\sum_{l\in\mathcal{J}(i)} \hat{I}_l^n[\tau_1^n, t] \leq c_6 p_n$, for $t \in [\tau_1^n, \tau_2^n)$. Note that c_6 does not depend on the constant γ. Hence γ can be chosen in such a way that $\tau_2^n \geq T^n$. Since $\hat{I}^n \geq 0$, $|\hat{I}_j^n|_{T^n} \leq |\sum_{l\in\in\mathcal{J}(i)} \hat{I}_l^n|_{T^n} \leq c_6 p_n$. This completes the proof. □

PROOF OF LEMMA 5.7. Recall that

$$\theta'\hat{X}^n(t) = \theta'\hat{X}^n(0) + \theta' W^n(t) + G^n(t) - R^n(t). \tag{145}$$

We treat separately each of the elements on the right-hand side above. First, by (104) we have on Ω^n that $|\theta' W^n(t) - \theta' W^n(s)| \leq c_1(p_n + \omega(t-s))$. Recall that $\alpha_\omega > 2\alpha_g$ so that $\omega(\vartheta_n) < n^{-(1/2\alpha_\omega - \delta\alpha_\omega)} \leq n^{-\alpha_g+\delta} = p_n n^\delta = q_n$. Thus

$$\bigl|\theta' W^n(t) - \theta' W^n(s)\bigr| \leq c_2 q_n. \tag{146}$$

Next, by the definition of T^n, using (51), we have that

$$\bigl|R^n\bigr|_{T^n} \leq J\|\theta\|\varepsilon_M^n \varrho^n \leq p_n. \tag{147}$$

Recalling that $\theta_i \mu_{ij} = z_j^*$ for all $i \sim j$ and that $\sum_i \tilde{B}_{ij}^n = -I_j^n$ for all $j \in \mathcal{J}$ we have that

$$G^n(t) = -\sum_{i,j} \theta_i \mu_{ij} \int_0^t \tilde{B}_{ij}^n(u)\,du = \sum_j z_j^* \int_0^t I_j^n(u)\,du, \tag{148}$$

from which we also see that G^n is nondecreasing. By Proposition 5.2, $|\Delta_k^n|_{T^n} \leq c_k^1 p_n$ for all $k \in \mathcal{I}_{-i_0}$ so that, by Lemma 5.6, $|\hat{I}_j^n|_{T^n} \leq c_3 p_n$ for all $j \neq j_0$. In turn, $\int_s^t I_j^n(u)\,du \leq c_4 p_n n^{1/2}(t-s)$ for all $j \neq j_0$ and for all $s, t \leq T^n$,

$$\left|G^n(t) - G^n(s) - z_{j_0}^* \int_s^t I_{j_0}^n(u)\,du\right| \leq \sum_{j\neq j_0} z_j^* \int_s^t I_j^n(u)\,du \leq c_5 n^{1/2} p_n (t-s).$$

Letting $t-s=\vartheta^n$ proves the first part of the lemma. The second part then follows immediately from (146), (147) and the fact that G^n is nondecreasing. $\square$

PROOF OF LEMMA 5.8. Throughout we fix $s,t\le T^n$ as in the statement of the lemma. We argue separately for two cases according to whether there exists $u\in[s,t)$ such that $\theta'\hat{X}^n(u)\ge\bar{\kappa}_n$.

Suppose first that $\theta'\hat{X}^n(u)<\bar{\kappa}_n$ for all $u\in[s,t)$. By the properties of the functions f^n (see Remark 5.1) and since $\kappa_n\ll q_n$ we have here that $\check{X}^n_i[s,t]\le(\mathbf{I}\theta_i)^{-1}\kappa_n\ll q_n$ as required.

To treat the other case we establish first the following claim. Fix $\beta>0$, then for all sufficiently large n, if $u\le T^n$ has $\theta'\hat{X}^n(u)\ge\beta\bar{\kappa}_n$, then $\hat{Q}^n_i(u)>0, i\in\mathcal{I}$.

To see this, note by Proposition 5.2 and the properties of f^n (see Remark 5.1) that $\theta_i\hat{Q}^n_i(u)\ge\theta_i f^n_i(\theta'\hat{X}^n(u))-c_1p_n\ge(\mathbf{I}\theta_i)^{-1}\kappa_n-c_1p_n>0$ for all $i\in\mathcal{I}_{-i_0}$ where we use the fact that $p_n\ll\kappa_n$. For $i=i_0$,

$$\begin{aligned}\theta_{i_0}\hat{Q}^n_{i_0}(u)&=\theta'\hat{X}^n(u)-\sum_{i\neq i_0}\theta_i\hat{Q}^n_i(u)-\sum_{i,j}\theta_i\hat{B}^n_{ij}(u)\\&\ge\theta'\hat{X}^n(u)-\sum_{i\neq i_0}\theta_i\check{X}^n_i(u)-c_2s_n-c_3p_n\\&=\theta_{i_0}\check{X}^n_{i_0}(u)-c_4s_n>0.\end{aligned}$$

The first inequality follows from Proposition 5.2 and the fact that $\|\hat{B}^n\|_{T^n}\le s_n$ by the definition of T^n. The second inequality follows from $p_n\ll s_n$ and from the fact that $\theta'\hat{X}^n=\theta'\check{X}^n$ whenever $\theta'\hat{X}^n\ge0$. The last inequality then follows from the definition of $\check{X}^n$, the fact that $f^n_{i_0}(x)\ge(\mathbf{I}\theta_{i_0})^{-1}(\bar{\kappa}_n\wedge x)$ for all $x\ge0$ (see Remark 5.1) and recalling that $s_n\ll\bar{\kappa}_n$.

Having the above we proceed to consider the case in which $\theta'\hat{X}^n(u)\ge\bar{\kappa}_n$ for some $u\in[s,t)$. Let

$$\tau^n_1=\sup\{\eta\le u:\theta'\hat{X}^n(u)\le\tfrac{1}{2}\bar{\kappa}_n\}$$

and

$$\tau^n_2=\inf\{\eta\ge u:\theta'\hat{X}^n(u)\le\tfrac{1}{2}\bar{\kappa}_n\}\wedge t,$$

where we set $\tau^n_1=s$ if $\theta'\hat{X}^n(\eta)\ge\bar{\kappa}_n$ for all $\eta\in[s,u)$. Since the jumps of $\theta'\hat{X}^n$ are of size $O(n^{-1/2})$ we must have that, if $\tau^n_2<t$ or $\tau^n_1>s$, then $\tau^n_1<u<\tau^n_2$. Setting $\beta=1/2$ in the argument above we have that $\hat{Q}^n_i>0$ on $[\tau^n_0,\tau^n_1)$ so that, by Remark 5.2 $\hat{I}^n_j=0, j\in\mathcal{J}$ on this interval. By equation (148) we then have that $G^n(\tau^n_2)-G^n(\tau^n_1)=0$. Using (145) and (147) we then have that

$$\begin{aligned}\bigl|\theta'\hat{X}^n(\tau^n_1)-\theta'\hat{X}^n(\tau^n_2)\bigr|&\le\bigl|\theta'W^n(\tau^n_2)-\theta W^n(\tau^n_1)\bigr|+c_4p_n\\&\le c_5\bigl(p_n+\omega(\tau^n_2-\tau^n_1)\bigr).\end{aligned}\tag{149}$$

It then follows, as in the beginning of the proof of Lemma 5.7, that

$$|\theta' \hat{X}^n[\tau_1^n, \tau_2^n]| \le c_6 q_n. \tag{150}$$

Since $q_n \ll \bar{\kappa}_n$ we conclude that $\theta' \hat{X}^n \ge \frac{3}{4}\bar{\kappa}_n$ for all $u \in [\tau_1, \tau_2^n)$ which contradicts $s < \tau_1^n < \tau_2^n < t$. We conclude that $\tau_1^n = s$ and $\tau_2^n = t$ and, by (161), that $\theta' \hat{X}^n[s,t] \le c_7 q_n$. From the local Lipschitz continuity of f^n it then follows that $\check{X}_i^n[s,t] \le c_8 q_n$ as required. □

PROOF OF LEMMA 5.9. We use the notation $\bar{w}_t := \bar{w}_{[0,t]}$ with the latter defined in (56). Recall also that the index set $\mathcal{I}$ is identified with $\{1, \dots, \mathbf{I}\}$ and define

$$b_k(n) = n^{k\delta/\mathbf{I}} p_n, \qquad k \in \{0, 1, \dots, \mathbf{I}\}. \tag{151}$$

Note that $b_{\mathbf{I}}(n) = n^{\delta} p_n = q_n$.

Fix $k \in \mathcal{I}_{-i_0}$ and let $j = \bar{\jmath}(k)$. By (34) and using $\hat{X}^n(0) = 0$ we can write

$$\hat{B}_{kj}^n(t) = -\mu_{kj}^n \int_0^t \hat{B}_{kj}^n(v)\, dv + F^n(t), \qquad t \ge 0, \tag{152}$$

where $F^n = F_1^n + F_2^n$,

$$F_1^n(t) = -\hat{Q}_k^n(t) + W_k^n(t), \qquad F_2^n(t) = -\sum_{l \in \mathcal{J}(k)} \mu_{kl}^n \int_0^t \hat{B}_{kl}^n(s)\, ds - \sum_{l \in \mathcal{J}(k)} \hat{B}_{kl}^n(t).$$

The proof is based on the following estimate; see Lemma 3.4 of [6] and its proof. Let X be the unique solution to the integral equation

$$X(t) = -\mu \int_0^t X(s)\, ds + F(t), \qquad t \ge 0,$$

with $\mu > 0$ and data $F : [0, \infty) \to \mathbb{D}$. Then given $u > 0$ and $\vartheta \in (0, u)$,

$$|X|_u \le 2|F|_u e^{-\mu\vartheta} + \bar{w}_u(F, \vartheta). \tag{153}$$

Thus in view of (152), one can bound $\hat{B}_{kj}^n$ by suitably estimating F^n.

If $\sum_{l \in \mathcal{J}(k)} |\hat{B}_{kl}^n|_{T^n} \le \beta b_{k-1}(n)$ for some β, then given $\vartheta > 0$,

$$\bar{w}_{T^n}\big(F_2^n, \vartheta\big) \le c_1 b_{k-1}(n)\big(1 + \vartheta n^{1/2}\big), \tag{154}$$

where c_1 does not depend on n or ϑ. By Proposition 5.2 we have that $|\hat{Q}_k^n - \check{X}_k^n|_{T^n} \le c_2 p_n$. Combined with Lemma 5.8, and letting $\vartheta = \vartheta_n$ this gives

$$\bar{w}_{T^n}\big(F_1^n, \vartheta_n\big) \le c_3 p_n$$

for some constant c_3 (not depending on n). Noting that $n^{1/2}\vartheta_n = n^{\delta}$, we get

$$\bar{w}_{T^n}\big(F^n, \vartheta_n\big) \le c_4 b_{k-1}(n) n^{\delta} = c_4 b_k(n), \tag{155}$$

where we used the fact that $n^{1/2}b_{k-1}(n)\vartheta_n = n^{\delta}b_{k-1}(n) = b_k(n)$ [recall that $b_k(n) = n^{(|k|/\mathbf{I})\delta}p_n$]. Noting $F^n(0) = 0$ by our assumptions and using (106) and (104), we also have

$$|F^n|_{T^n} \le c_5(r_n + 1 + b_{k-1}(n)n^{1/2}). \tag{156}$$

Thus

$$\bar{w}_{T^n}(F^n, \vartheta_n) \le c_6 b_k(n), \tag{157}$$

and, in turn,

$$|\hat{B}^n_{kj}|_T \le |F^n|_{T^n} e^{-\mu^n_{kj}\vartheta_n} + \bar{w}_t(F^n, \vartheta_n)(1 - e^{-\mu^n_{kj}\vartheta_n}).$$

Since $n^{1/2}\vartheta_n = n^{\delta}$ we have that $\mu^n_{kj}\vartheta_n \ge c_7 n^{\delta}$. Further, since $b_{k-1}(n) \to 0$, we have that $n^{1/2}b_{k-1}(n) \le c_8 n^{1/2}$ so that $n^{1/2}z(n)e^{-n^{\delta}} \to 0$. We conclude that

$$\|\hat{B}^n_{kj}\|_{T^n} \le 2\bar{w}_{T^n}(F^n, \vartheta_n) \le c_9 b_k(n), \tag{158}$$

provided that

$$\sum_{l\in\mathcal{J}(k)} |\hat{B}^n_{kl}|_{T^n} \le \beta b_{k-1}(n) \tag{159}$$

for some β.

The requirement (159) holds trivially if k is a leaf of the tree, in which case the set $\mathcal{J}(k)$ is empty. It also holds if all pools in $\mathcal{J}(k)$ are leafs of the tree by Proposition 5.2 and Lemma 5.6 because in that case $|\hat{B}_{kj}|_{T^n} = |\hat{I}^n_j|_{T^n}$ for all $j \in \mathcal{J}(k)$. Thus the fact that it holds for all $k \ne i_0$ and $j = \bar{j}(k)$ now follows by induction on the class number using (159) and (158).

To bound $\hat{B}^n_{kj}$ for $j \ne \bar{j}(k)$, by identity (30) we have that $|\hat{B}^n_{kj}|_{T^n} \le |\hat{I}^n_j|_{T^n} + \sum_{l\in I(j)} |\hat{B}_{lj}|_{T^n}$. The first part of this proof guarantees that $\sum_{l\in I(j)} |\hat{B}_{lj}|_{T^n} \le c_{10}q_n$. From Proposition 5.2 and Lemma 5.6 it follows that $|\hat{I}^n_j|_{T^n} \le c_{10}p_n$. This completes the proof. □

PROOF OF LEMMA 5.10. Let $s, t \le T^n$ be an interval as in the statement of the lemma. In particular, $\hat{I}^n_{j_0} > 0$ on $[s, t)$ so that, by Remark 5.2, $\hat{Q}^n_{i_0} = 0$ on $[s, t)$ and $\theta_{i_0}\hat{X}^n_{i_0}[s, t] = \sum_j \hat{B}^n_{i_0 j}[s, t]$. Thus

$$\theta_{i_0}\hat{B}^n_{i_0 j_0}[s, t] = \theta_{i_0}X^n_{i_0}[s, t] - \theta_{i_0}\sum_{j\ne j_0} \hat{B}^n_{i_0 j}[s, t] \ge \theta_{i_0}\hat{X}^n_{i_0}[s, t] - c_1 q_n,$$

where the inequality follows from Lemma 5.9.

By Lemma 3.1 and the identity $\hat{I}^n_{j_0} = -\sum_i \hat{B}^n_{ij_0}$ it further holds that

$$|\hat{I}_{j_0}[s, t] + \hat{B}^n_{i_0 j_0}[s, t]| \le \sum_{i\ne i_0} |\hat{B}^n_{ij_0}|_{T^n} \le c_2 q_n,$$

so that

$$\theta_{i_0}\hat{I}_{j_0}[s,t] \le -\theta_{i_0}X^n_{i_0}[s,t] + c_4 q_n. \tag{160}$$

The lemma would then follow from Lemma 5.7 provided that

$$\theta_{i_0}\hat{X}^n_{i_0}[s,t] \ge \theta'\hat{X}^n[s,t] - c_5 q_n. \tag{161}$$

We next prove (161). By Lemma 5.9, $\|\hat{Q}^n_{-i_0} - \hat{X}^n_{i_0}\|_{T^n} = \sum_{i\in\mathcal{I}_{-i_0}, j\in\mathcal{J}} |\hat{B}^n_{ij}|_{T^n} \le c_6 q_n$ and by Proposition 5.2, $\|\hat{Q}^n_{-i_0} - \check{X}^n_{-i_0}\|_{T^n} \le c_7 p_n$. In turn,

$$\|\hat{X}^n_{-i_0} - \check{X}^n_{-i_0}\|_{T^n} \le \|\hat{Q}^n_{-i_0} - \hat{X}^n_{-i_0}\|_{T^n} + \|\hat{Q}^n_{-i_0} - \check{X}^n_{-i_0}\|_{T^n} \le c_8 q_n.$$

Using the identity $\theta_{i_0}\hat{X}^n_{i_0}[s,t] = \theta'\hat{X}^n[s,t] - \sum_{i\ne i_0}\hat{X}^n_i[s,t]$ we then have that

$$\left|\theta_{i_0}\hat{X}^n_{i_0}[s,t] - \theta'\hat{X}^n[s,t] + \sum_{i\ne i_0}\check{X}^n_i[s,t]\right| \le c_9 q_n. \tag{162}$$

By Lemma 5.8, $\sum_{i\ne i_0}|\check{X}^n_i[s,t]| \le c_{10}q_n$ for all $s,t \le T^n$ with $|t-s| \le \vartheta_n$ which proves (161) and thus completes the proof of the lemma. □

REFERENCES

[1] ARMONY, M. and WARD, A. R. (2010). Fair dynamic routing in large-scale heterogeneous-server systems. *Oper. Res.* **58** 624–637. MR2680568

[2] ATAR, R. (2005). Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *Ann. Appl. Probab.* **15** 2606–2650. MR2187306

[3] ATAR, R. (2012). A diffusion regime with nondegenerate slowdown. *Oper. Res.* **60** 490–500. MR2935073

[4] ATAR, R., MANDELBAUM, A. and SHAIKHET, G. (2006). Queueing systems with many servers: Null controllability in heavy traffic. *Ann. Appl. Probab.* **16** 1764–1804. MR2288704

[5] ATAR, R. and SHAIKHET, G. (2009). Critically loaded queueing models that are throughput suboptimal. *Ann. Appl. Probab.* **19** 521–555. MR2521878

[6] ATAR, R. and SOLOMON, N. (2011). Asymptotically optimal interruptible service policies for scheduling jobs in a diffusion regime with nondegenerate slowdown. *Queueing Syst.* **69** 217–235. MR2886469

[7] BELL, S. L. and WILLIAMS, R. J. (2001). Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: Asymptotic optimality of a threshold policy. *Ann. Appl. Probab.* **11** 608–649. MR1865018

[8] BELL, S. L. and WILLIAMS, R. J. (2005). Dynamic scheduling of a parallel server system in heavy traffic with complete resource pooling: Asymptotic optimality of a threshold policy. *Electron. J. Probab.* **10** 1044–1115. MR2164040

[9] BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York. MR0233396

[10] BRÉMAUD, P. (1981). *Point Processes and Queues*: *Martingale Dynamics*. Springer, New York. MR0636252

[11] BROWN, L., GANS, N., MANDELBAUM, A., SAKOV, A., SHEN, H., ZELTYN, S. and ZHAO, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.* **100** 36–50. MR2166068

[12] CHEN, H. and MANDELBAUM, A. (1991). Leontief systems, RBVs and RBMs. In *Applied Stochastic Analysis* (*London*, 1989). *Stochastics Monogr.* **5** 1–43. Gordon and Breach, New York. MR1108415

[13] CSÖRGŐ, M. and HORVÁTH, L. (1993). *Weighted Approximations in Probability and Statistics*. Wiley, Chichester. MR1215046

[14] DAI, J. G. and TEZCAN, T. (2011). State space collapse in many-server diffusion limits of parallel server systems. *Math. Oper. Res.* **36** 271–320. MR2828761

[15] ETHIER, S. N. and KURTZ, T. G. (1986). *Markov Processes*: *Characterization and Convergence*. Wiley, New York. MR0838085

[16] GARNETT, O., MANDELBAUM, A. and REIMAN, M. (2002). Designing a call center with impatient customers. *Manufacturing and Service Operations Management* **4** 208–227.

[17] GURVICH, I. (2004). Design and control of the $M/M/N$ queue with multi-class customers and many servers. M.Sc. thesis, Technion.

[18] HALFIN, S. and WHITT, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29** 567–588. MR0629195

[19] HARRISON, J. M. and LÓPEZ, M. J. (1999). Heavy traffic resource pooling in parallel-server systems. *Queueing Syst.* **33** 339–368. MR1742575

[20] MANDELBAUM, A. (2003). QED Q's. Notes from a lecture delivered at the Workshop on Heavy Traffic Analysis and Process Limits of Stochastic Networks, EURANDOM.

[21] MANDELBAUM, A., MOMČILOVIĆ, P. and TSEYTLIN, Y. (2012). On fair routing from emergency departments to hospital wards: QED queues with heterogeneous servers. *Management Science* **58** 1273–1291.

[22] MANDELBAUM, A. and SHAIKHET, G. Private communication.

[23] MANDELBAUM, A. and STOLYAR, A. L. (2004). Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Oper. Res.* **52** 836–855. MR2104141

[24] STOLYAR, A. L. and TEZCAN, T. (2010). Control of systems with flexible multi-server pools: A shadow routing approach. *Queueing Syst.* **66** 1–51. MR2674107

[25] WHITT, W. (2003). How multiserver queues scale with growing congestion-dependent demand. *Oper. Res.* **51** 531–542. MR1991969

[26] WHITT, W. (2004). Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science* **50** 1449–1461.

[27] WILLIAMS, R. J. (1998). An invariance principle for semimartingale reflecting Brownian motions in an orthant. *Queueing Syst.* **30** 5–25. MR1663755

[28] WILLIAMS, R. J. (2000). On dynamic scheduling of a parallel server system with complete resource pooling. In *Analysis of Communication Networks*: *Call Centres*, *Traffic and Performance* (*Toronto*, *ON*, 1998). *Fields Inst. Commun.* **28** 49–71. Amer. Math. Soc., Providence, RI. MR1788708

DEPARTMENT OF ELECTRICAL ENGINEERING
TECHNION–ISRAEL INSTITUTE OF TECHNOLOGY
HAIFA 32000
ISRAEL

KELLOGG SCHOOL OF MANAGEMENT
NORTHWESTERN UNIVERSITY
EVANSTON, ILLINOIS 60201
USA