CCIT Report #456
December 2003

Asymptotically optimal policies for treelike parallel server stations in heavy traffic*

Rami Atar
Department of Electrical Engineering
Technion–Israel Institute of Technology
Haifa 32000, Israel

December 5, 2003

Abstract

We study a multiclass queueing system operating in the heavy traffic regime proposed by Halfin and Whitt, a regime that models systems with large number of servers working independently. An optimal control problem is considered, where the control corresponds to scheduling of jobs and the cost is a cumulative discounted functional of the system's state. Under the scaling limit a control problem for a diffusion is obtained. The dynamic programming PDE was proved in [1] to uniquely characterize the value function for the diffusion control problem. In this paper we show that the solution to the PDE can be used to construct policies for the queueing system that are asymptotically optimal.

1 Introduction

In [1] we studied the dynamic programming PDE of Hamilton-Jacobi-Bellman (HJB) type for a diffusion control problem associated with a family of multiclass queueing systems, and characterized the control problem's value as the unique solution to the PDE. The diffusion control problem was obtained by parametrizing the queueing system in a central limit theorem (CLT) regime, and taking *formal* weak limits of the processes involved. In the current paper we establish the validity of the diffusion control problem as the correct asymptotic description of the queueing problem in this regime, by showing that the optimal solution to the queueing problem converges to that of the diffusion problem. In addition, we use the PDE solution to construct scheduling policies for the queueing system that are asymptotically optimal.

The queueing system has a fixed number of customer classes arriving according to renewal processes, and a fixed number of service stations, where each service station has many servers with the same capabilities (see Figure 1(a)). Each customer requires service exactly once. The CLT point

*Research supported in part by the Israel Science Foundation (grant no. 126/02)

of view taken here is the one proposed by Halfin and Whitt [7], where the system is parametrized by taking, in an appropriate fashion, the arrival rates and the number of servers at each station to grow without bound. See [6] for motivation for this model and parametric regime. The precise probabilistic model and scaling assumptions, as well as additional assumptions, are described in Section 2. One then considers scheduling of jobs as control, and attempts to minimize an expected cumulative discounted functional of rather general performance criteria, including queue lengths of different classes, number of customers in each of the stations, and number of servers that are idle at each station. Taking the parametrization limit is meant to simplify the problem. In particular, the non-Markovian scheduling problem Markovianizes, with the advantage of a dynamic programming equation, characterizing the diffusion problem's value, being available.

As explained in [1], because the number of servers is large, the problem of optimally controlling such a system is very different depending on whether only nonpreemptive scheduling is possible, or whether it is allowed to use preemptive policies (where service to a customer can be stopped and resumed at a different station). In general, these two problems give rise to two different diffusion control problems, and it is the one associated with preemptive scheduling that was analyzed in [1]. However, for reasons explained in detail in [1], it is expected that under a certain structural condition, the diffusion control problem associated with nonpreemptive scheduling, that in general lies in higher dimension, degenerates to the one associated with preemptive scheduling. The structural condition states that the graph, having classes and stations as nodes and having class-station pairs as edges if the station can serve the class, is a tree (as in Figure 1(b)). In this paper, the treelike assumption is made so as to make it possible to treat nonpreemptive and preemptive scheduling under the same umbrella of the PDE studied in [1]. Thus the asymptotic approach simplifies the problem also in reducing the nonpreemptive problem to the diffusion model of the preemptive problem.

The main results are the following. Under appropriate assumptions, the scheduling control problem's value under preemption converges to the value of the diffusion problem. Moreover, preemptive and nonpreemptive scheduling policies are constructed, that are asymptotically optimal in the sense that they asymptotically achieve the diffusion problem's value. The proof requires two different types of arguments. One has to do with tightness of the processes involved. To this end the approach from [2] is adopted and extended. The other regards estimates on the large time behaviour of the prelimit system, where results from [1] are crucially used.

The organization of the paper is as follows. In Section 2 we introduce the probabilistic queueing model under study and the assumptions regarding scaling as well as work conservation. We then describe the diffusion control problem and the HJB equation, propose scheduling control policies for the queueing model based on the HJB equation, and state the main result regarding asymptotic optimality of these policies. Section 3 contains the proofs.

Notation. Vectors in $\mathbb{R}^k$ are considered as column vectors. For $x \in \mathbb{R}^k$ [resp., $x \in \mathbb{R}^{k\times l}$] let $\|x\| = \sum_i |x_i|$ [resp., $\sum_{i,j} |x_{i,j}|$]. For two column vectors v, u, $v \cdot u$ denotes their scalar product. The symbols e_i denote the unit coordinate vectors and $e = (1, \dots, 1)'$. The dimension of e may change from one expression to another, and for example $e \cdot a + e \cdot b = \sum_i a_i + \sum_j b_j$ even if a and b are of different dimension. Write $\mathbb{N} = \{1, 2, \dots\}$, $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$, $\mathbb{R}_+ = [0, \infty)$. $C^{m,\varepsilon}(D)$ [respectively, $C^m(D)$] denotes the class of functions on $D \subset \mathbb{R}^k$ for which all derivatives up to

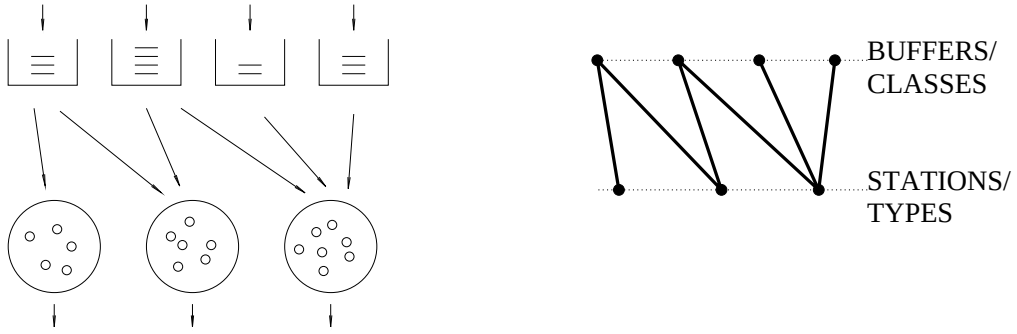

Figure 1: Parallel server station system

order m are Hölder continuous uniformly on compact subsets of D [continuous on D]. $C_{\rm pol}(\mathbb{R}^k)$ denotes the class of continuous functions f on $\mathbb{R}^k$, satisfying a polynomial growth condition: there are constants c and r such that $|f(x)| \leq c(1+\|x\|^r)$, $x \in \mathbb{R}^k$. We let $C^{m,\varepsilon}_{\rm pol} = C_{\rm pol} \cap C^{m,\varepsilon}$. For E a metric space, we denote by $\mathbb{D}(E)$ the space of all cadlag functions (i.e., right continuous and having left limits) from $\mathbb{R}_+$ to E. We endow $\mathbb{D}(E)$ with the usual Skorohod topology. If X^n, $n \in \mathbb{N}$ and X are processes with sample paths in $\mathbb{D}(E)$, we write $X^n \Rightarrow X$ to denote weak convergence of the measures induced by X^n (on $\mathbb{D}(E)$) to the measure induced by X. If X is an $\mathbb{R}^k$- or an $\mathbb{R}^{k\times l}$-valued process (or function on $\mathbb{R}_+$) and $0 \leq s \leq t < \infty$, $\|X\|^*_{s,t} = \sup_{s\leq u\leq t}\|X(u)\|$, and if X takes real values, $|X|^*_{s,t} = \sup_{s\leq u\leq t}|X(u)|$. Also, $\|X\|^*_t = \|X\|^*_{0,t}$ and $|X|^*_t = |X|^*_{0,t}$. The notations $X(t)$ and X_t are used interchangeably. For a locally integrable function $f : \mathbb{R}_+ \to \mathbb{R}$ denote $If = \int_0^{\cdot} f(s)ds$. In case that f is vector- or matrix-valued, If is understood elementwise. For $y \in \mathbb{R}^k_+$ denote by $[\![y]\!]$ the element $y' \in \mathbb{Z}^k_+$ having $y'_i = \lfloor y_i \rfloor$ for $i = 1 \ldots k-1$, and $y'_k = \lfloor y_k \rfloor + \sum_{i=1}^k (y_i - \lfloor y_i \rfloor)$. Note that $e\cdot [\![y]\!] = e\cdot y$, and

$$\|y - [\![y]\!]\| \leq 2k. \tag{1}$$

The symbol c denotes a deterministic positive constant whose value may change from line to line.

2 Setting and results

2.1 Queueing model

The queueing model under study has $\bar{\imath}$ customer classes and $\bar{\jmath}$ server stations. At each service station there are many independent servers of the same type. Each customer requires service only once and can be served indifferently by any server at the same station, but possibly at different rates at different stations. Only some stations can offer service to each class. When referring to the physical location of customers we say that they are in the buffer, or in the queue if they are not being served, and we say that they are in a certain station if they are served by a server of the corresponding type. There is one buffer per each customer class and one station per each server type (see Figure 1(a)).

A complete probability space (Ω, F, P) is given, supporting all stochastic processes defined below. Expectation with respect to P is denoted by E. Since the set of all classes and all types constitutes the vertex set of graphs, it is convenient to label the classes as $1, \ldots, \bar{\imath}$ and the types as $\bar{\imath}+1, \ldots, \bar{\imath}+\bar{\jmath}$:

$$\mathcal{I} = \{1, \ldots, \bar{\imath}\}, \quad \mathcal{J} = \{\bar{\imath}+1, \ldots, \bar{\imath}+\bar{\jmath}\}.$$

The buffers [resp., stations] are labeled as the corresponding classes [resp., types]. For $j \in \mathcal{J}$ let N_j^n be the number of servers at station j. Thus the total number of servers is $e \cdot N^n$. Let $X_i^n(t)$ denote the total number of class-i customers in the system at time t. Let $Y_i^n(t)$ denote the number of class-i customers in the queue at time t. Let $Z_j^n(t)$ denote the number of idle servers in station j at time t. And let $\Psi_{ij}^n(t)$ denote the number of class-i customers in station j at time t. We have $X^n = (X_i^n)_{i\in\mathcal{I}}$, $Y^n = (Y_i^n)_{i\in\mathcal{I}}$, $Z^n = (Z_j^n)_{j\in\mathcal{J}}$, $\Psi^n = (\Psi_{ij}^n)_{i\in\mathcal{I}, i\in\mathcal{J}}$. Straightforward relations are expressed by the following equations:

$$Y_i^n + \sum_{j\in\mathcal{J}} \Psi_{ij}^n = X_i^n, \quad i \in \mathcal{I}, \tag{2}$$

$$Z_j^n + \sum_{i\in\mathcal{I}} \Psi_{ij}^n = N_j^n, \quad j \in \mathcal{J}, \tag{3}$$

$$Y_i^n(t), Z_j^n(t) \geq 0, \quad i \in \mathcal{I}, j \in \mathcal{J}, t \geq 0. \tag{4}$$

$$X_i^n \geq 0, \Psi_{ij}^n \geq 0. \tag{5}$$

To define arrival processes, let, for each $i \in \mathcal{I}$, $\{\check{U}_i(k), k \in \mathbb{N}\}$ be a sequence of strictly positive i.i.d. random variables with $E\check{U}_i(1) = 1$ and squared coefficient of variation $(E\check{U}_i(1))^{-2}\mathrm{Var}(\check{U}_i(1)) = C_{U,i}^2 \in [0, \infty)$. Assume also that the sequences are independent. Let

$$U_i^n(k) = \frac{1}{\lambda_i^n}\check{U}_i(k), \tag{6}$$

where $\lambda_i^n > 0$. With $\sum_1^0 = 0$, define

$$A_i^n(t) = \sup\{l \geq 0 : \sum_{k=1}^{l} U_i^n(k) \leq t\}, \quad t \geq 0. \tag{7}$$

The renewal processes A_i^n are used to model arrivals: The number of arrivals up to time t is equal to $A_i^n(t)$. Note that the first class-i customer arrives at $U_i^n(1)$, and the time between the $(k-1)$st and kth arrival of class-i customers is $U_i^n(k)$.

To model service times as exponential independent random variables, let S_{ij}^n, $i \in \mathcal{I}, j \in \mathcal{J}$ be Poisson processes with rate $\mu_{ij}^n \in [0, \infty)$ (where a zero rate Poisson process is the zero process). These processes are assumed to be mutually independent, and independent of the arrival processes. Let $T_{ij}^n(t)$ denote the time up to t devoted to a class-i customer by a server, summed over all type-j servers and note that

$$T_{ij}^n(t) = \int_0^t \Psi_{ij}^n(s)ds, \quad i \in \mathcal{I}, j \in \mathcal{J}, t \geq 0.$$

The number of service completions of class-i customers by all type-j servers by time t is $S_{ij}^n(T_{ij}^n(t))$. Thus, with $X_i^{0,n} := X_i^n(0)$, we have

$$X_i^n(t) = X_i^{0,n} + A_i^n(t) - \sum_j S_{ij}^n \left(\int_0^t \Psi_{ij}^n(s) ds \right), \quad i \in \mathcal{I}, t \geq 0. \tag{8}$$

To introduce a basic assumption on the structure of the system we consider the graph $\mathcal{G}$ having vertex set $\mathcal{I} \cup \mathcal{J}$ with a node per each class and a node per each type, and an edge set $\mathcal{E}$, with an edge joining a class and a type if the corresponding service rate is nonzero:

$$\mathcal{E} = \{(i,j) \in \mathcal{I} \times \mathcal{J} : \mu_{ij}^n > 0\}.$$

Under assumptions that we introduce below (especially (13)) there will be no loss of generality assuming that $\mathcal{E}$ does not depend on the parameter n. We assume throughout that *the graph $\mathcal{G}$ is a tree*. The reader is referred to [1] for explanation on how the problem is different in nature with regard to asymptotic behaviour depending on whether $\mathcal{G}$ is a tree or a connected graph other than a tree. We denote $i \sim j$ and $j \sim i$ if $(i,j) \in \mathcal{E}$. By assumption we have

$$\Psi_{ij}^n = 0, \quad i \not\sim j. \tag{9}$$

A pair (i,j) is said to be an *activity* if $(i,j) \in \mathcal{E}$.

2.2 Scheduling

Scheduling decisions are made by continuously selecting Ψ^n, subject to appropriate constraints. Scheduling is regarded as *preemptive* if service to a customer can be stopped and resumed at a later time, possibly in a different station. Formally this is expressed by stating that the process Ψ may be selected subject only to equations (2)–(9) holding. Note that according to this definition, customers can be moved instantaneously not only between a service station and the buffer, but also between different service stations that offer service to the corresponding class. Scheduling is regarded as *nonpreemptive* if every customer completes service with the server it is first assigned. More precisely, consider the processes $B_{ij}^n(t)$, $i \in \mathcal{I}, j \in \mathcal{J}$, where $B_{ij}^n(0) = 0$, and B_{ij}^n increases by k each time k class-i jobs are moved to station j from the buffer or from another station (to start or resume service), and decreases by k each time k such jobs are moved from station j back to the buffer or to another station. Then B_{ij}^n can be expressed as

$$B_{ij}^n(t) = \Psi_{ij}^n(t) - \Psi_{ij}^{0,n} + S_{ij}^n \left(\int_0^t \Psi_{ij}^n(s) ds \right). \tag{10}$$

Under nonpreemptive scheduling, each of these processes is nondecreasing. Thus to define nonpreemptive scheduling in terms of the model equations (2)–(9), we will require that Ψ is selected subject to (2)–(9) and such that B_{ij}^n are nondecreasing processes. This is summarized in the following.

Definition 1 *Let initial data $X^{0,n}$ be given.*

i. We say that Ψ^n *is a* preemptive resume scheduling control policy *(P-SCP) if* Ψ_{ij} *have cadlag paths (*$i \in \mathcal{I}, j \in \mathcal{J}$*), and there exist processes* X^n, Y^n *and* Z^n *such that* (2)–(9) *are met.* X^n *is said to be the controlled process associated with initial data* $X^{0,n}$ *and P-SCP* Ψ^n*.*

ii. We say that Ψ^n *is a* non preemptive scheduling control policy *(N-SCP) if it is a P-SCP and* B^n_{ij} *of* (10) *have non-decreasing paths.*

We collectively refer to P-SCPs and N-SCPs as *scheduling control policies* (SCPs) (although the class of SCPs is simply the class of P-SCPs).

Let $\zeta = (X^{0,n}; n \in \mathbb{N})$ and $p = (\Psi^n, n \in \mathbb{N})$ denote a sequence of initial conditions, and respectively, SCPs. We denote by P^p_ζ the measure under which, for each n, X^n is the controlled process associated with $X^{0,n}$ and Ψ^n. E^p_ζ denotes expectation under P^p_ζ.

We need a notion of SCPs that do not anticipate the future. Unlike in a Markovian setting, in presence of renewal processes such a notion has to take into account that the time of the next arrival is correlated with information from the past, and hence should not be regarded as innovative information. Denote

$$\tau^n_i(t) = \inf\{u \geq t : A^n_i(u) - A^n_i(u-) > 0\} \quad i \in \mathcal{I}.$$

Set

$$\mathcal{F}^n_t = \sigma\{A^n_i(s), S^n_{ij}(T^n_{ij}(s)), \Psi^n_{ij}(s), X^n_i(s), Y_i(s), Z_j(s) : i \in \mathcal{I}, j \in \mathcal{J}, s \leq t\}, \tag{11}$$

and

$$\mathcal{G}^n_t = \sigma\{A^n_i(\tau^n_i(t) + u) - A^n_i(\tau^n_i(t)), S^n_{ij}(T^n_{ij}(t) + u) - S^n_{ij}(T^n_{ij}(t)) : i \in \mathcal{I}, j \in \mathcal{J}, u \geq 0\}. \tag{12}$$

Definition 2 *We say that a scheduling control policy is* admissible *if*

i. for each t, $\mathcal{F}^n_t$ *is independent of* $\mathcal{G}^n_t$*;*

ii. for each i, j *and* t*, the process* $S^n_{ij}(T^n_{ij}(t) + \cdot) - S^n_{ij}(T^n_{ij}(t))$ *is equal in law to* $S^n_{ij}(\cdot)$*.*

2.3 Fluid and diffusion scaling

A sequence of systems as described above is considered, of which the symbol n is the index. See [1] for more information on the relation of this scaling to Halfin and Whitt [7] and to Harrison and López [9].

<u>*Scaling of parameters:*</u> There are constants $\lambda_i > 0$, $i \in \mathcal{I}$, $\mu_{ij} > 0$, $(i,j) \in \mathcal{E}$, and $\nu_j > 0$, $j \in \mathcal{J}$ such that

$$n^{-1}\lambda^n_i \to \lambda_i, \quad \mu^n_{ij} \to \mu_{ij} \tag{13}$$

and

$$n^{-1}N_j^n \to \nu_j.$$

Setting

$$\bar{\mu}_{ij} = \nu_j \mu_{ij},$$

a central assumption on the limit parameters indicating that the sequence of systems is asymptotically critically loaded is introduced below (cf. [8], [9]).

Linear program: Minimize ρ subject to

$$\sum_{j \in \mathcal{J}} \bar{\mu}_{ij} \xi_{ij} = \lambda_i, \quad i \in \mathcal{I}, \tag{14}$$

$$\sum_{i \in \mathcal{I}} \xi_{ij} \leq \rho, \quad j \in \mathcal{J}, \tag{15}$$

$$\xi_{ij} \geq 0, \quad i \in \mathcal{I}, j \in \mathcal{J}. \tag{16}$$

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

In the rest of this paper, ξ_{ij}^* denotes the quantities from the above condition, and $x^* = (x_i^*)$, $\psi^* = (\psi_{ij}^*)$, where

$$x_i^* = \sum_j \xi_{ij}^* \nu_j, \quad \psi_{ij}^* = \xi_{ij}^* \nu_j. \tag{17}$$

We refer to the quantities ξ_{ij}^*, x_i^* and ψ_{ij}^* as *the static fluid model.*

An additional assumption will be the following (see [1] for more details).

Complete resource pooling condition: $\xi_{ij}^* > 0$ for $(i,j) \in \mathcal{E}$.

Assumption 1 *The graph $\mathcal{G}$ is a tree. The heavy traffic and complete resource pooling conditions hold.*

Second order scaling assumptions are as follows.

Scaling of parameters, continued: There are constants $\hat{\lambda}_i, \hat{\mu}_{ij} \in \mathbb{R}$, $i \in \mathcal{I}, j \in \mathcal{J}$, such that

$$n^{1/2}(n^{-1}\lambda_i^n - \lambda_i) \to \hat{\lambda}_i, \qquad n^{1/2}(\mu_{ij}^n - \mu_{ij}) \to \hat{\mu}_{ij}, \tag{18}$$

$$n^{1/2}(n^{-1}N_j^n - \nu_j) \to 0. \tag{19}$$

Scaling of initial conditions: There are constants x_i, y_i, z_j, ψ_{ij} such that the deterministic initial conditions satisfy

$$\hat{X}_i^{0,n} := n^{-1/2}(X_i^{0,n} - nx_i^*) \to x_i, \quad \hat{Y}_i^{0,n} := n^{-1/2}Y_i^{0,n} \to y_i, \tag{20}$$

$$\hat{Z}_j^{0,n} := n^{-1/2}Z_j^{0,n} \to z_j, \quad \hat{\Psi}_{ij}^{0,n} := n^{-1/2}(\Psi_{ij}^{0,n} - \psi_{ij}^* n) \to \psi_{ij}, \tag{21}$$

where, due to (2), (3) and (4) we assume $y_i + \sum_j \psi_{ij} = x_i$, $z_j + \sum_i \psi_{ij} = 0$, $y_i \geq 0$, and $z_j \geq 0$, $i \in \mathcal{I}$, $j \in \mathcal{J}$.

The processes rescaled at the fluid level are defined as

$$\bar{X}_i^n(t) = n^{-1}X_i^n(t), \quad \bar{Y}_i^n(t) = n^{-1}Y_i^n(t),$$
$$\bar{Z}_j^n(t) = n^{-1}Z_j^n(t), \quad \bar{\Psi}_{ij}^n(t) = n^{-1}\Psi_i^n(t).$$

The primitive processes are centered about their means and rescaled at the diffusion level as

$$\hat{A}_i^n(t) = n^{-1/2}(A_i^n(t) - \lambda_i^n t), \qquad \hat{S}_{ij}^n(t) = n^{-1/2}(S_{ij}^n(nt) - n\mu_{ij}^n t). \tag{22}$$

Similarly, the state processes are centered about the static fluid model and rescaled:

$$\hat{X}_i^n(t) = n^{-1/2}(X_i^n(t) - nx_i^*), \tag{23}$$

$$\hat{Y}_i^n(t) = n^{-1/2}Y_i^n(t), \qquad \hat{Z}_j^n(t) = n^{-1/2}Z_j^n(t), \tag{24}$$
$$\hat{\Psi}_{ij}^n(t) = n^{-1/2}(\Psi_{ij}^n - \psi_{ij}^* n).$$

The relations (2), (3) and (4) take the new form

$$\hat{Y}_i^n + \sum_j \hat{\Psi}_{ij}^n = \hat{X}_i^n, \quad i \in \mathcal{I}, \tag{25}$$

$$\hat{Z}_j^n + \sum_i \hat{\Psi}_{ij}^n = 0, \quad j \in \mathcal{J}, \tag{26}$$

$$\hat{Y}_i^n, \hat{Z}_j^n \geq 0, \quad i \in \mathcal{I}, j \in \mathcal{J}. \tag{27}$$

Using the definitions above of the rescaled processes and the relation $\lambda = \mu(\xi^*)$, one finds that (8) takes the form

$$\hat{X}_i^n(t) = \hat{X}_i^{0,n} + r_i\hat{W}_i^n(t) + \ell_i^n t - \sum_j \mu_{ij}^n \int_0^t \hat{\Psi}_{ij}^n(s)ds \tag{28}$$

where

$$r_i\hat{W}_i^n(t) = \hat{A}_i^n(t) - \sum_j \hat{S}_{ij}^n(\int_0^t \bar{\Psi}_{ij}^n(s)ds), \tag{29}$$

$$\ell_i^n = n^{1/2}(n^{-1}\lambda_i^n - \lambda_i) - \sum_i n^{1/2}(\mu_{ij}^n - \mu_{ij})\psi_{ij}^*.$$

With (18) we have

$$\lim_n \ell_i^n = \ell_i := \hat{\lambda}_i - \sum_j \hat{\mu}_{ij}\psi_{ij}^*. \tag{30}$$

One is free to choose the values of r_i, and it is convenient to choose them so that, with the formal substitution $\bar{\Psi}_{i,j}^n = \psi_{i,j}^*$, one has

$$\lim_n E[(\hat{W}_i^n(1))^2] = 1. \tag{31}$$

Namely, $r_i = (\lambda_i C_{U,i}^2 + \lambda_i)^{1/2}$. Denote also $\ell = (\ell_1, \ldots, \ell_{\bar{\imath}})'$, $r = \mathrm{diag}(r_i)$.

The results of this paper are concerned with constructing sequences of SCPs that, in an appropriate sense, minimize the limit as $n \to \infty$ of cost of the following form:

$$E\int_0^\infty e^{-\gamma t}\tilde{L}(\hat{X}_t^n, \hat{\Psi}_t^n)dt. \tag{32}$$

2.4 Joint work conservation

A policy is said to be *work conserving* if it does not allow for a server to idle while a customer that it can serve is in the queue. In the current context one can consider a stronger condition for preemptive policies. Recall that if preemption is allowed, customers of each class can be moved between the queue and the various stations that offer service to them. Let t be given, and recall that the components of $X^n(t)$ denote the number of customers of each class present in the system at time t. Since t is fixed and its value will be immaterial in the following discussion, we omit it from the notation and write e.g., X^n for $X^n(t)$. In a preemptive policy, the set of controls Ψ^n that can be applied at time t correspond to different rearrangements of the customers X^n in the stations and buffers. Let $\mathcal{X}^n$ denote the set of all possible values of X^n for which there is a rearrangement of customers with the property: *either there are no customers in the queue, or no server in the system is idle.* This property is expressed as

$$e\cdot Y^n \wedge e\cdot Z^n = 0. \tag{33}$$

We shall say that a preemptive policy is *jointly work conserving* if it is work conserving and, in addition, for every s, if $X^n(s) \in \mathcal{X}^n$ then customers are arranged according to (33) (see also [1]).

In the heavy traffic regime considered here, it is anticipated that it is nearly always the case that a rearrangement of customers according to (33) is possible. This element will be justified in the process of proving our main result.

2.5 The diffusion control problem

We take limits as $n \to \infty$ in (25), (26), (28) and (33). The process $\hat{W}^n$ converges to a standard Brownian motion. Denoting the weak limit of $(\hat{X}^n, \hat{Y}^n, \hat{Z}^n, \hat{W}^n, \hat{\Psi}^n)$ by (X, Y, Z, W, Ψ), we get the

equation below (at this point this is meant as a formal step only; however, see Proposition 1)

$$X_i(t) = x_i + \tilde{W}_i(t) - \sum_j \mu_{ij} \int_0^t \Psi_{ij}(s)ds, \quad i \in \mathcal{I} \tag{34}$$

where $\tilde{W}_i(t) = r_i W_i(t) + \ell_i t$, W is a standard Brownian motion, and

$$\sum_j \Psi_{ij} = X_i - Y_i, \quad i \in \mathcal{I}, \tag{35}$$

$$\sum_i \Psi_{ij} = -Z_j, \quad j \in \mathcal{J}, \tag{36}$$

$$e\cdot Y \wedge e\cdot Z = 0. \tag{37}$$

Relations (34)–(37) above can be written in the convenient form $dX = b(X,U)dt + rdW$. To this end, note that by (35) and (36), $e\cdot X = e\cdot Y - e\cdot Z$, and thus by (37)

$$e\cdot Y = (e\cdot X)^+, \quad e\cdot Z = (e\cdot X)^-.$$

Hence Y and Z can be represented in terms of the process $e\cdot X$ and an additional process U as

$$Y_i(t) = (e\cdot X(t))^+ u_i(t), \quad Z_j(t) = (e\cdot X(t))^- v_j(t), \quad i \in \mathcal{I}, j \in \mathcal{J} \tag{38}$$

where $U(t) = (u(t), v(t))$ takes values in

$$\mathbb{U} := \{(u,v) \in \mathbb{R}^{\bar{\imath}+\bar{\jmath}} : u_i, v_j \geq 0,\ i \in \mathcal{I}, j \in \mathcal{J},\ e\cdot u = e\cdot v = 1\}.$$

The following is shown in [1].

Lemma 1 *Let Assumption 1 hold. Then given $\alpha_i, \beta_j \in \mathbb{R}$, $i \in \mathcal{I}, j \in \mathcal{J}$ satisfying $e\cdot\alpha = e\cdot\beta$ there exists a unique solution ψ_{ij} to the set of equations*

$$\sum_j \psi_{ij} = \alpha_i, \quad i \in \mathcal{I}, \qquad \sum_i \psi_{ij} = \beta_j, \quad j \in \mathcal{J}, \tag{39}$$

where $\psi_{i,j} = 0$ for $i \not\sim j$.

As a result there is a map, denoted throughout by $G : D_G \to \mathbb{R}^{\bar{\imath}+\bar{\jmath}-1}$,

$$D_G := \{(\alpha,\beta) \in \mathbb{R}^{\bar{\imath}+\bar{\jmath}} : e\cdot\alpha = e\cdot\beta\},$$

such that $\psi = G(\alpha,\beta)$. Let also

$$C_G = \sup\{\max_{ij} |G(\alpha,\beta)_{ij}| : (\alpha,\beta) \in D_G, \|\alpha\| \vee \|\beta\| \leq 1\}. \tag{40}$$

Clearly the map is linear on D_G. Applying Lemma 1 to (35), (36), and by (38) it follows that

$$\Psi = G(X - Y, -Z) = G(X - (e\cdot X)^+ u, -(e\cdot X)^- v) =: \hat{G}(X, U). \tag{41}$$

For $\psi = (\psi_{ij})$ denote by $\mu \circ \psi$ a column vector with $(\mu \circ \psi)_i = \sum_j \mu_{ij}\psi_{ij}$. With

$$b = -\mu \circ \hat{G} + \ell, \tag{42}$$

we can now write (34) as

$$X(t) = x + rW(t) + \int_0^t b(X(s), U(s))ds, \quad 0 \le t < \infty. \tag{43}$$

We summarize the equivalence between the two representations in what follows.

Lemma 2 *If equations* (34)–(37) *hold then* (43) *holds with some* $\mathbb{U}$*-valued* U*. Conversely, if* (43) *holds (with* U *taking values in* $\mathbb{U}$*) then one can find* Ψ, Y, Z *such that* (34)–(37) *hold. In both cases,* $\tilde{W}(t) = rW(t) + \ell t$*.*

Proof: We have already proved the first statement of the result. To see the converse, write $U = (u, v)$, let $Y = (e\cdot X)^+ u$, $Z = (e\cdot X)^- v$, and $\Psi = G(X - Y, -Z)$. (35), (36) automatically hold, and (41) implies (34). □

Definition 3 *We call* $\pi = (\Omega, F, (F_t), P, U, W)$ *an* admissible system *if* $(\Omega, F, (F_t), P)$ *is a complete filtered probability space,* U *is a* $\mathbb{U}$*-valued,* (F_t)*-progressively measurable process, and* W *is a standard* $\bar{\imath}$*-dimensional* (F_t)*-Brownian motion. The process* U *is said to be a* control *associated with* π*.* X *is said to be a* controlled process *associated with initial data* $x \in \mathbb{R}^{\bar{\imath}}$ *and an admissible system* π*, if it is a continuous sample paths,* (F_t)*-adapted such that* $\int_0^t |b(X(s), U(s))|ds < \infty$ $t \ge 0$*,* P*-a.s., and* (43) *holds* P*-a.s.*

For any $x \in \mathbb{R}^{\bar{\imath}}$ and any admissible system π there exists a controlled process X, unique in the strong sense (cf. [1]). With an abuse of notation we sometimes denote the dependence on x and π by writing P_x^π in place of P and E_x^π in place of E. We denote by Π the class of all admissible systems.

Given a constant $\gamma > 0$ and a function $\tilde{L}$, the cost of interest for the queueing system is given by (32). It is convenient to perform change of variables from (X, Ψ) to (X, U). To this end, define L as

$$L(X, U) = \tilde{L}(X, G(X - (e\cdot X)^+ u, -(e\cdot X)^- v)), \quad X \in \mathbb{R}^{\bar{\imath}}, U = (u, v) \in \mathbb{U}. \tag{44}$$

Our conditions on $\tilde{L}$ (given mostly via conditions on L) are as follows.

Assumption 2 *i.* $L(x, U) \ge 0$*,* $(x, U) \in \mathbb{R}^{\bar{\imath}} \times \mathbb{U}$*.*

ii. The mapping $(x,\psi) \mapsto \tilde{L}(x,\psi)$ is continuous. In particular, the mapping $(x,U) \mapsto L(x,U)$ is continuous.

iii. There is $\varrho \in (0,1)$ such that for any compact $A \subset \mathbb{R}^{\bar{\imath}}$,

$$|L(x,U) - L(y,U)| \leq c\|x-y\|^{\varrho}$$

holds for $U \in \mathbb{U}$ and $x, y \in A$, where c depends only on A.

iv. There are constants $c > 0$ and $m_L \geq 1$ such that $L(x,U) \leq c(1 + \|x\|^{m_L})$, $U \in \mathbb{U}$, $x \in \mathbb{R}^{\bar{\imath}}$.

Consider the cost

$$C(x,\pi) = E_x^{\pi} \int_0^{\infty} e^{-\gamma t} L(X(t), U(t)) dt, \quad x \in \mathbb{R}^{\bar{\imath}}, \pi \in \Pi.$$

Define the value function as

$$V(x) = \inf_{\pi \in \Pi} C(x,\pi).$$

The HJB equation for the problem is (cf. [5])

$$\mathcal{L}f + H(x, Df) - \gamma f = 0, \tag{45}$$

where $\mathcal{L} = (1/2)\sum_i r_i^2 \partial^2 / \partial x_i^2$, and

$$H(x,p) = \inf_{U \in \mathbb{U}} [b(x,U) \cdot p + L(x,U)]. \tag{46}$$

The equation is considered on $\mathbb{R}^{\bar{\imath}}$ with the growth condition

$$\exists C, m, \quad |f(x)| \leq C(1 + \|x\|^m), \quad x \in \mathbb{R}^{\bar{\imath}}. \tag{47}$$

We say that f is a solution to (45) if it is of class C^2, and the equation is satisfied everywhere in $\mathbb{R}^{\bar{\imath}}$.

Definition 4 *Let $x \in \mathbb{R}^{\bar{\imath}}$ be given. We say that a measurable function $h : \mathbb{R}^{\bar{\imath}} \to \mathbb{U}$ is a* Markov control policy *if there is an admissible system π and a controlled process X corresponding to x and π, such that $U_s = h(X_s)$, $s \geq 0$, P-a.s. We say that an admissible system π is* optimal *for x, if $V(x) = C(x,\pi)$. We say that a Markov control policy is optimal for x if the corresponding admissible system is.*

Assumption 3 *Either (i) or (ii) below holds.*

i. For $(i,j) \in \mathcal{E}$, μ_{ij} depends only on i; or for $(i,j) \in \mathcal{E}$, μ_{ij} depends only on j.

ii. The tree $\mathcal{G}$ is of diameter 3 at most.

The following is proved in [1].

Theorem 1 *Let Assumptions 1, 2 and 3 hold. Then the value V solves (45), (47), uniquely in $C^{2,\varrho}_{\rm pol}(\mathbb{R}^{\bar{\imath}})$, and there exists a Markov control policy $h : \mathbb{R}^{\bar{\imath}} \to \mathbb{U}$ that is optimal for all $x \in \mathbb{R}^{\bar{\imath}}$.*

To state our main result we need to introduce SCPs defined via the function h of Theorem 1. Write $h = (h_1, h_2)$ where $(u, v) = h(x) \Leftrightarrow u = h_1(x), v = h_2(x)$. Let $\alpha_0 > 0$ denote the constant

$$\alpha_0 = (4C_G)^{-1} \min_{i,j:i\sim j} \psi^*_{ij}. \tag{48}$$

Let

$$\check{Y}^n = (e\cdot \hat{X}^n)^+ h_1(\hat{X}^n), \quad \check{Z}^n = (e\cdot \hat{X}^n)^- h_2(\hat{X}^n) \tag{49}$$

$$\check{\Psi}^n = G(\hat{X}^n - \check{Y}^n, -\check{Z}^n). \tag{50}$$

The quantities $\check{Y}^n, \check{Z}^n$ will be used to propose a P-SCP while $\check{\Psi}^n$ will be used to propose an N-SCP.

The proposed sequence of P-SCPs is defined as follows. If $\|X^n(t) - nx^*\| \le \alpha_0 n$, set

$$Y^n = [\![n^{1/2}\check{Y}^n]\!], \quad Z^n = [\![n^{1/2}\check{Z}^n]\!], \quad \Psi^n = G(X^n - Y^n, N^n - Z^n). \tag{51}$$

Setting Ψ^n this way is in accordance with joint work conservation in the sense that (33) always holds on $\{\|X^n(t) - nx^*\| \le \alpha_0 n\|\}$. To see this note that by (49), $\check{Y}^n \wedge \check{Z}^n = 0$, hence by (51), $Y^n(t) \wedge Z^n(t) = 0$. Also, since by definition X^n, Y^n and Z^n are integer valued, so is Ψ^n as follows from the proof of Proposition 1 of [1]. It remains to show that setting $\Psi^n(t)$ as in (51) meets relation (5) of the model, namely that $\Psi^n_{ij}(t) \ge 0$ for all i, j. The proof of this fact is deferred to Section 3.1 (cf. Lemma 3 and the remark that follows). Finally, if $\|X^n(t) - nx^*\| > \alpha_0 n$, set $\Psi^n = F_n(X^n(t))$ where, for each n, F_n is a fixed function chosen in such a way that the resulting process Ψ^n is jointly work conserving (as defined in Section 2.4). Other than that the choice of F_n is immaterial. Denote the resulting sequence of P-SCPs by p^*.

The proposed sequence of N-SCPs is described in what follows. The idea is to keep $\hat{\Psi}^n$ close to $\check{\Psi}^n$ of (50) by declaring activity (i, j) as "blocked" at any time when $\hat{\Psi}^n_{ij} > \check{\Psi}^n_{ij}$, and not routing customers through blocked activities. More precisely, given an activity (i, j) and a time interval $[s, t)$, if $\hat{\Psi}^n_{ij} > \check{\Psi}^n_{ij}$ holds on $[s, t)$ then no routings take place on the activity throughout this interval:

$$\hat{\Psi}^n_{ij} > \check{\Psi}^n_{ij} \text{ on } [s,t) \qquad \text{implies} \qquad B^n_{ij} \text{ is constant on } [s,t). \tag{52}$$

On the other hand, when there is a class-i customer in the queue and there are stations $j \sim i$ with idle servers and such that (i, j) is not blocked, the customer is instantaneously routed to one of these stations. If there are no such stations, the customer stays in queue. It is not hard to see that this property can be expressed as follows. For every activity (i, j),

$$\hat{\Psi}^n_{ij}(t) \le \check{\Psi}^n_{ij}(t) \qquad \text{implies} \qquad Y^n_i(t) \wedge Z^n_j(t) = 0. \tag{53}$$

The selection of an activity among the non-blocked activities through which to route a customer does not turn out to be significant, except that care must be taken because it is possible that two

(or more) customers will be routed instantaneously. Therefore it remains to show that one can always perform instantaneous routings meeting (53). This is deferred to Section 3.1 (cf. Lemma 4).

A sequence of N-SCPs satisfying the two properties will be denoted by p'.

Our assumption on the interarrival times is as follows. The two parts of the assumption correspond to different parts of the result.

Assumption 4 *There is a constant m_A such that $E(\check{U}_i(1))^{m_A} < \infty$, $i \in \mathcal{I}$, satisfying either of the following.*

(a) $m_A > 2m_L$ (where m_L is as in Assumption 2);

(b) $m_A(m_A-2)(5m_A-2)^{-1} > m_L$.

Our main result is the following.

Theorem 2 *Let Assumptions 1, 2 and 3 hold. Let ζ be a sequence of initial conditions $(X^{0,n}; n \in \mathbb{N})$ such that $\hat{X}^{0,n} = n^{-1/2}(X^{0,n} - nx^*) \to x \in \mathbb{R}^{\bar{\imath}}$. Then items (i) and (ii) below hold under Assumption 4(a), and item (iii) holds under Assumption 4(b).*

i. For any sequence p of jointly work conserving admissible P-SCPs,

$$\liminf_{n\to\infty} E_\zeta^p \int_0^\infty e^{-\gamma t}\tilde{L}(\hat{X}_t^n, \hat{\Psi}_t^n)dt \geq V(x);$$

ii. The sequence p^* of jointly work conserving admissible P-SCPs satisfies

$$\lim_{n\to\infty} E_\zeta^{p^*} \int_0^\infty e^{-\gamma t}\tilde{L}(\hat{X}_t^n, \hat{\Psi}_t^n)dt = V(x);$$

iii. Provided that h (of Theorem 1) is locally Hölder on $\{x \in \mathbb{R}^{\bar{\imath}} : e\cdot x \neq 0\}$ and globally Hölder on $\{x \in \mathbb{R}^{\bar{\imath}} : |e\cdot x| \geq 1\}$, the sequence p' of admissible N-SCPs satisfies

$$\limsup_{n\to\infty} E_\zeta^{p'} \int_0^\infty e^{-\gamma t}\tilde{L}(\hat{X}_t^n, \hat{\Psi}_t^n)dt \leq V(x).$$

Remarks. (a) The Hölder assumption on h required in part (iii) can be shown to hold under some strict convexity properties of L (similar to Proposition 3 of [2]).
(b) As explained in Section 3.3, there is a particular aspect of the large time estimates that we are unable to prove in general, and this aspect alone is stopping us from relaxing Assumption 3. We comment however that Assumption 3 can be removed at the price of assuming $\gamma > \gamma_0$, where γ_0 depends on the parameters of the diffusion model alone (by using (89) along with the idea of Section 3.3 so as to estimate moments of $\hat{X}^n$ in terms of $ce^{\gamma_0 t}$ uniformly in n). We do not pursue this direction here.

3 Proofs

3.1 Preliminary results

In Section 2 we obtained the convenient representation (43) of the controlled process from equations (34)–(37). We begin by showing how an equation analogous to (43) is obtained for the prelimit model. To this end, let

$$M^n := e\cdot Y^n \wedge e\cdot Z^n \geq 0. \tag{54}$$

Denote also $\hat{M}^n = n^{-1/2}M^n$. By (25) and (26) we can write $e\cdot \hat{X}^n = e\cdot \hat{Y}^n - e\cdot \hat{Z}^n$ and therefore by (54),

$$e\cdot \hat{Y}^n = \hat{M}^n + (e\cdot \hat{X}^n)^+, \quad e\cdot \hat{Z}^n = \hat{M}^n + (e\cdot \hat{X}^n)^-. \tag{55}$$

Let $u^n = Y^n/(e\cdot Y^n)$ when $e\cdot Y^n > 0$, and arbitrarily set $u^n = e_1$ otherwise. Similarly, $v^n = Z^n/(e\cdot Z_n)$ if $e\cdot Z^n > 0$ and otherwise set $v^n = e_{\bar{\ell}+1}$. Letting $U^n = (u^n, v^n)$, noting that U^n takes values in $\mathbb{U}$, and using Lemma 1 and linearity of G on D_G, we have

$$\hat{Y}^n = (\hat{M}^n + (e\cdot \hat{X}^n)^+)u^n, \quad \hat{Z}^n = (\hat{M}^n + (e\cdot \hat{X}^n)^-)v^n, \tag{56}$$

$$\begin{aligned}\hat{\Psi}^n &= G(\hat{X}^n - \hat{Y}^n, -\hat{Z}^n)\\ &= G(\hat{X}^n - (e\cdot \hat{X}^n)^+u^n, -(e\cdot \hat{X}^n)^-v^n) + \hat{M}^n G(u^n, v^n)\\ &= \hat{G}(\hat{X}^n, U^n) + \hat{M}^n G(u^n, v^n).\end{aligned} \tag{57}$$

Defining $b^n = -\mu^n \circ \hat{G} + \ell^n$ we obtain from (28)

$$\hat{X}^n(t) = \hat{X}^{0,n} + \int_0^t b^n(\hat{X}^n(s), U^n(s))ds + r\hat{W}^n(t) - \int_0^t \hat{M}^n(s)\mu^n \circ G(u^n(s), v^n(s))ds. \tag{58}$$

It is useful to note that

$$\|b^n(x,U) - b^n(y,U)\| \leq c_1\|x-y\|, \quad \|b^n(x,U)\| \leq c_1(1+\|x\|), \quad n\in\mathbb{N}, U\in\mathbb{U}, x, y\in\mathbb{R}^{\bar{\ell}}, \tag{59}$$

where c_1 does not depend on n, U, x and y.

The following result states that under preemptive scheduling, joint work conservation can be maintained whenever $\|X^n(t) - nx^*\| \leq \alpha_0 n$. In particular it shows that the p^* is a well defined sequence of P-SCPs.

Lemma 3 *Fix t. Assume the following relations hold: equations* (2) *and* (3) *(equivalently, $\Psi^n(t) = G(X^n(t) - Y^n(t), N^n - Z^n(t))$), $e\cdot Y^n(t) \wedge e\cdot Z^n(t) = 0$, and $\|X^n(t) - nx^*\| \leq \alpha_0 n$, where $\alpha_0 > 0$ is the constant from* (48). *Then $\Psi^n_{ij}(t) \geq 0$ (in particular* (5) *holds).*

Remark: The lemma has two implications.
(a) p^* is a legitimate sequence of jointly work conserving P-SCPs. As argued in Section 2, one only has to show that when $\|X^n(t) - nx^*\| \le \alpha_0 n$, $\Psi^n_{ij}(t) \ge 0$ holds. This follows from the construction of p^* and Lemma 3.
(b) Let p be any sequence of jointly work conserving P-SCPs. Let $\tau^n = \inf\{s : M^n(s) > 0\}$. By Lemma 3, if $\tau^n \le t$ then $\|X^n - nx^*\|^*_t \ge \alpha_0 n$. In particular, $\|X^n - nx^*\|^*_{\tau^n} \ge \alpha_0 n$ on $\{\tau^n < \infty\}$.

Proof of Lemma 3: By Lemma 1 and equations (2), (3),

$$\Psi^n = G(X^n - Y^n, N^n - Z^n), \tag{60}$$

and by (17),

$$\psi^* = G(x^*, \nu). \tag{61}$$

Since it is assumed that $M^n = 0$, (55) implies that $\|Y^n\|, \|Z^n\| \le \|X^n - nx^*\|$. Hence by linearity of the map G on the domain D_G and by (19), (40) and (48),

$$\begin{aligned}
\Psi^n_{ij}(t) &= G(nx^*, n\nu) + G(X^n - nx^* - Y^n, N^n - n\nu - Z^n)_{ij} \\
&\ge n\psi^*_{ij} - C_G\|X^n - nx^* - Y^n\| \vee \|N^n - n\nu - Z^n\| \\
&\ge n\psi^*_{ij} - 2C_G\|X^n - nx^*\| - C_G\|N^n - n\nu\| \\
&\ge n\psi^*_{ij} - 2C_G\alpha_0 n - cn^{1/2} \\
&\ge 0,
\end{aligned}$$

where the last inequality holds for all n large enough. □

The following lemma refers to instantaneous routing through non-blocked activities in the construction of the N-SCPs p' of Section 2.

Lemma 4 *Let $(\tilde\Psi, \tilde X, \tilde Y, \tilde Z)$ satisfy*

$$\sum_j \tilde\Psi_{ij} = \tilde X_i - \tilde Y_i,\ i \in \mathcal{I}, \quad \sum_i \tilde\Psi_{ij} = -\tilde Z_j,\ j \in \mathcal{J}, \quad \tilde\Psi_{ij} = 0,\ i \not\sim j \tag{62}$$

$$\tilde Y_i \ge 0, \quad \tilde Z_j \ge 0, \quad i \in \mathcal{I}, j \in \mathcal{J}. \tag{63}$$

Assume that all components of $\tilde X, \tilde Y, \tilde Z$ and $\tilde\Psi$ are integers. Let a subset $\mathcal{E}_1 \subset \mathcal{E}$ ("non-blocked" activities) be given. Then one can find (Ψ, X, Y, Z) satisfying relations analogous to (62), (63), and

$$X = \tilde X, \quad Y \le \tilde Y, \quad Z \le \tilde Z, \quad \Psi \ge \tilde\Psi, \tag{64}$$

$$(i,j) \in \mathcal{E}_1 \quad \text{implies} \quad Y_i \wedge Z_j = 0. \tag{65}$$

Proof: Define inductively a sequence $(X^{(k)}, Y^{(k)}, Z^{(k)}, \Psi^{(k)})$, $k = 0, \ldots k_1$ as follows. Let

$$(X^{(0)}, Y^{(0)}, Z^{(0)}, \Psi^{(0)}) = (\tilde{\Psi}, \tilde{X}, \tilde{Y}, \tilde{Z}).$$

Let $k \geq 0$ be given, for which $(X^{(k)}, Y^{(k)}, Z^{(k)}, \Psi^{(k)})$ is defined. If $Y^{(k)}$, $Z^{(k)}$ satisfy (65) then terminate, declaring $k_1 = k$. Otherwise, define $(X^{(k+1)}, Y^{(k+1)}, Z^{(k+1)}, \Psi^{(k+1)})$ as follows. Let i_0 be the smallest $i \in \mathcal{I}$ such that there is j with $(i,j) \in \mathcal{E}_1$ and $Y_i^{(k)} \wedge Z_j^{(k)} > 0$. Let j_0 be the smallest such j. For $i \in \mathcal{I}, j \in \mathcal{J}$ define

$$X^{(k+1)} = X^{(k)}, \quad Y_i^{(k+1)} = Y_i^{(k)} - 1_{i=i_0}, \quad Z_j^{(k+1)} = Z_j^{(k)} - 1_{j=j_0}, \quad \Psi_{ij}^{(k+1)} = \Psi_{ij}^{(k)} + 1_{(i,j)=(i_0,j_0)}.$$

Since by construction $0 \leq e \cdot Y^{(k)} = e \cdot \tilde{Y} - k$, the procedure must terminate. Defining $(X, Y, Z, \Psi) = (X^{(k_1)}, Y^{(k_1)}, Z^{(k_1)}, \Psi^{(k_1)})$ completes the proof. □

The following lemma will be useful in analyzing the N-SCPs p'.

Lemma 5 *Let (ψ, x, y, z) satisfy*

$$\sum_j \psi_{ij} = x_i - y_i, \ i \in \mathcal{I}, \quad \sum_i \psi_{ij} = -z_j, \ j \in \mathcal{J}, \quad \psi_{ij} = 0, \ i \not\sim j,$$

and let $(\check{\psi}, \check{x}, \check{y}, \check{z})$ satisfy analogous relations. In addition, assume

$$\text{if } i \sim j \text{ and } \psi_{ij} < \check{\psi}_{ij} \text{ then } y_i \wedge z_j = 0, \tag{66}$$

and $\check{y}_i \geq 0, \ i \in \mathcal{I}, \ \check{z}_j \geq 0, \ j \in \mathcal{J}$. Then

$$\sum_{i,j} |\psi_{ij} - \check{\psi}_{ij}| \leq c \sum_{i,j} (\psi_{ij} - \check{\psi}_{ij})^+ + c\|x - \check{x}\|,$$

where c does not depend on $\psi, x, y, z, \check{\psi}, \check{x}, \check{y}$ or $\check{z}$.

Proof of Lemma 5: Let ε be an upper bound on $\psi_{ij} - \check{\psi}_{ij}$ for all i, j, and on $|x_i - \check{x}_i|$ for all i. Let j_0 be such that $z_{j_0} = 0$. Then

$$\sum_i \psi_{ij_0} = \sum_i \check{\psi}_{ij_0} + \check{z}_{j_0} \geq \sum_i \check{\psi}_{ij_0},$$

and since $\psi_{ij_0} \leq \check{\psi}_{ij_0} + \varepsilon$, $\psi_{ij_0} - \check{\psi}_{ij_0} \geq -c\varepsilon$ for every $i \sim j_0$. Let i_0 be such that $y_{i_0} = 0$. Then

$$\sum_j \psi_{i_0 j} = x_{i_0} \geq \check{x}_{i_0} - \varepsilon = \sum_j \check{\psi}_{i_0 j} + \check{y}_{i_0} - \varepsilon \geq \sum_j \check{\psi}_{i_0 j} - \varepsilon.$$

Since $\psi_{i_0 j} \leq \check{\psi}_{i_0 j} + \varepsilon$ for every j, we have $\psi_{i_0 j} - \check{\psi}_{i_0 j} \geq -c\varepsilon$ for every $j \sim i_0$. Thus we have shown that $|\psi_{ij} - \check{\psi}_{ij}| \leq c\varepsilon$ for every (i,j), $i \sim j$ with either $y_i = 0$ or $z_j = 0$. In view of (66), we have shown that $|\psi_{ij} - \check{\psi}_{ij}| \leq c\varepsilon$ for every (i,j), $i \sim j$ such that $\psi_{ij} < \check{\psi}_{ij}$. On the other hand, if $\psi_{ij} \geq \check{\psi}_{ij}$, then simply $|\psi_{ij} - \check{\psi}_{ij}| = \psi_{ij} - \check{\psi}_{ij} \leq \varepsilon$ by assumption. □

Denote

$$J_t^n = \|\hat{Y}_t^n - \check{Y}_t^n\| + \|\hat{Z}_t^n - \check{Z}_t^n\|, \tag{67}$$

$$Q_t^n = \int_0^t b^n(\hat{X}_s^n, U_s^n)ds, \quad R_t^n = \int_0^t e^{-\gamma s} L(\hat{X}_s^n, U_s^n)ds. \tag{68}$$

Throughout, let p, p^*, p' and ζ be as in Theorem 2, and let f denote the unique $C^2_{\rm pol}$ solution to (45).

Proposition 1 *Items (i)–(iii) below hold under p (in particular, under p^*) and under p'.*
(i) $(\bar{X}^n, \bar{Y}^n, \bar{Z}^n, \bar{\Psi}^n) \Rightarrow (x^*, 0, 0, \psi^*)$.
(ii) $(\hat{W}^n, I\hat{M}^n, \hat{X}^n, Q^n, R^n)$ is tight.
(iii) $(\hat{W}^n, I\hat{M}^n) \Rightarrow (W, 0)$, where W is a standard Brownian motion on $\mathbb{R}^{\bar{\imath}}$.
Moreover, the following holds under p^* and under p':
(iv) $|J^n|^*_{s,t} \to 0$ and $|\hat{M}^n|^*_{s,t} \to 0$ in distribution, for every $0 < s < t < \infty$.

Proof: See Section 3.2.

Lemma 6 *Under p and under p' one has the following. Denote by (X, Q, R, W) a limit point of $(\hat{X}^n, Q^n, R^n, \hat{W}^n)$ along a subsequence. Let (F_t) denote the filtration generated by (X, Q, W). Then W is an (F_t)-standard Brownian motion, X, Q and R have continuous sample paths, and Q has sample paths of bounded variation over finite time intervals. Moreover, $\int e^{-\gamma s} Df(\hat{X}_s^n)\cdot dQ_s^n \Rightarrow \int e^{-\gamma s} Df(X_s)\cdot dQ_s$ along the subsequence, where f is the solution to (45).*

Proof of Lemma 6: Based on Proposition 1, the proof of Lemma 6 is identical to that of Lemma 6 of [2] and is therefore omitted. □

Proposition 2 *For either $q = p$, with fixed $m_0 \in (m_L, m_A/2)$, or for $q = p'$, with fixed $m_0 \in (m_L, m_A(m_A-2)(5m_A-2)^{-1})$, one has*

$$E^q_\zeta \|\hat{X}_t^n\|^{m_0} \le C(1+t)^{m_1} \tag{69}$$

where C, m_1 do not depend on t.

Proof: See Section 3.3.

The method of [2], that we adopt here, is based on estimating the process

$$K_t^n = b(\hat{X}_t^n, U_t^n)\cdot Df(\hat{X}_t^n) + L(\hat{X}_t^n, U_t^n) - H(\hat{X}_t^n, Df(\hat{X}_t^n)) \ge 0, \tag{70}$$

where the inequality above follows from (46).

Lemma 7 *Let the assumptions of Theorem 2 hold. For every sequence p of admissible jointly work conserving P-SCPs*

$$\liminf_{n\to\infty} E_\zeta^p \int_0^\infty e^{-\gamma t} L(\hat{X}_t^n, U_t^n)dt \geq V(x).$$

If q is an admissible SCP under which

$$\int_0^\cdot e^{-\gamma s} K_s^n ds \Rightarrow 0, \tag{71}$$

then

$$\limsup_{n\to\infty} E_\zeta^q \int_0^\infty e^{-\gamma t} L(\hat{X}_t^n, U_t^n)dt \leq V(x).$$

Proof of Lemma 7: Equipped with Proposition 1, Lemma 6 and Proposition 2, the proof is identical to that of Theorem 4 of [2], and is therefore omitted. □

Proof of Theorem 2: In view of Lemma 7, it suffices to show that under p^* and under p', $\int_0^\cdot e^{-\gamma s} K_s^n ds \Rightarrow 0$. The function h satisfies the following (see the proof of Theorem 1 of [1]):

$$H(x, Df(x)) = b(x, h(x))\cdot Df(x) + L(x, h(x)), \quad x \in \mathbb{R}^{\bar{\ell}}. \tag{72}$$

Combining (70) and (72),

$$K_t^n = (b(\hat{X}_t^n, U_t^n) - b(\hat{X}_t^n, h(\hat{X}_t^n)))\cdot Df(\hat{X}_t^n) + L(\hat{X}_t^n, U_t^n) - L(\hat{X}_t^n, h(\hat{X}_t^n)). \tag{73}$$

By definition of b (42), and by (67),

$$\|b(\hat{X}_t^n, U_t^n) - b(\hat{X}_t^n, h(\hat{X}_t^n))\| \leq cJ_t^n. \tag{74}$$

By (44) and (50),

$$L(\hat{X}_t^n, U_t^n) - L(\hat{X}_t^n, h(\hat{X}_t^n)) = \tilde{L}(\hat{X}_t^n, \Psi_1^n(t)) - \tilde{L}(\hat{X}_t^n, \check{\Psi}^n(t)) \tag{75}$$

where, using (56),

$$\begin{aligned}\Psi_1^n(t) &:= G(\hat{X}_t^n - (e\cdot \hat{X}_t^n)^+ u_t^n, -(e\cdot \hat{X}_t^n)^- v_t^n) \\ &= G(\hat{X}_t^n - \hat{Y}_t^n + \hat{M}_t^n u_t^n, -\hat{Z}_t^n + \hat{M}_t^n v_t^n) \\ &= \hat{\Psi}_t^n + \hat{M}_t^n G(u_t^n, v_t^n).\end{aligned} \tag{76}$$

Note that

$$\|\hat{\Psi}_t^n - \check{\Psi}_t^n\| = \|G(\hat{X}_t^n - \hat{Y}_t^n, -\hat{Z}_t^n) - G(\hat{X}_t^n - \check{Y}_t^n, -\check{Z}_t^n)\| \leq cJ_t^n. \tag{77}$$

By Assumption 2, $\tilde{L}$ is uniformly continuous on compacts, hence there are functions $\alpha_k(\delta)$ with $\lim_{\delta\to 0}\alpha_k(\delta) = 0$ such that

$$|\tilde{L}(x,\psi) - \tilde{L}(x,\psi')| \leq \alpha_k(\delta), \quad \|x\| \vee \|\psi\| \vee \|\psi'\| \leq k, \|\psi - \psi'\| \leq \delta. \tag{78}$$

Combining (73)–(78), $\|\hat{X}^n_t\| \vee \|\hat{\Psi}^n_t\| \leq k$ implies

$$K^n_t \leq cJ^n_t \|Df(\hat{X}^n_t)\| + \alpha_{ck}(J^n_t + \hat{M}^n_t) \leq cJ^n_t \beta_k + \alpha_{ck}(J^n_t + \hat{M}^n_t), \tag{79}$$

where β_k depends only on k. Since by (49), $e\cdot \check{Y}^n \wedge e\cdot \check{Z}^n = 0$, we have

$$\hat{M}^n = e\cdot \hat{Y}^n \wedge e\cdot \hat{Z}^n \leq |e\cdot \hat{Y}^n - e\cdot \check{Y}^n| + |e\cdot \hat{Z}^n - e\cdot \check{Z}^n| \leq J^n. \tag{80}$$

Moreover, by (25), (26) and (49),

$$J^n \leq c(\|\hat{X}^n\| + \|\hat{\Psi}^n\|). \tag{81}$$

Fix T and let $\Omega^{n,k,\varepsilon,\delta}$ denote the event that $\|\hat{X}^n\|^*_T \vee \|\hat{\Psi}^n\|^*_T \leq k$ and $|J^n + \hat{M}^n|^*_{\varepsilon,T} \leq \delta$. By Proposition 1,

$$\lim_k \liminf_{\delta\to 0^+} \liminf_{\varepsilon\to 0^+} \liminf_n P^q(\Omega^{n,k,\varepsilon,\delta}) = 1, \tag{82}$$

for $q = p^*$ and for $q = p'$, $\varepsilon > 0$. Combining (79)–(81), on $\Omega^{n,k,\varepsilon,\delta}$ we have

$$0 \leq \int_0^T e^{-\gamma t} K^n_t dt \leq c\varepsilon(k\beta_k + \alpha_{ck}(ck)) + cT(\beta_k\delta + \alpha_{ck}(\delta)).$$

Taking $n \to \infty$, then $\varepsilon \to 0$, $\delta \to 0$ and finally $k \to \infty$, using (82), it follows that $\int_0^T e^{-\gamma t} K^n_t dt \to 0$ in distribution. Since T is arbitrary, $\int_0^\cdot e^{-\gamma t} K^n_t dt \Rightarrow 0$. □

Lemma 8 *Let Assumption 4 hold. Then*

$$E(\|\hat{A}^n\|^*_t)^{m_A} \leq c(1+t)^{m_A/2},$$

where c does not depend on n or t.

Proof: See Lemma 2 of [2]. □

3.2 Tightness estimates

We prove Proposition 1. Most involved is the treatment of the nonpreemptive case. The main idea is a "bootstrap" argument (a variation of which is also used in the next subsection), where one first establishes tightness of the processes up to a certain stopping time, and then uses this to show that the probability that the stopping time is incurred in arbitrarily fixed finite time approaches zero. The proof is established in a number of steps.

Step 1: $n^{-1/2}\hat{W}^n \Rightarrow 0$.
Step 2: Under p, $\hat{M}^n \Rightarrow 0$ and $\bar{X}^n \Rightarrow x^*$; Under p', $\bar{X}^n(\cdot \wedge \sigma^n) \Rightarrow x^*$, where $\sigma^n = \inf\{s > 0 : I\hat{M}^n(s) \geq 1\}$.
Step 3: Under p, $(\bar{X}^n, \bar{Y}^n, \bar{Z}^n, \bar{\Psi}^n) \Rightarrow (x^*, 0, 0, \psi^*)$, and $(\hat{X}^n, \hat{W}^n, Q^n, R^n)$ is tight.

Step 4: Under p^*, $J^n \Rightarrow 0$
Step 5: Under p', conclusions of Step 3 hold, upon stopping all processes involved at σ^n. As a result, $I\hat{M}^n \Rightarrow 0$, and conclusions analogous to those of Step 3 hold under p'.

Step 1. Let A_i, $i \in \mathcal{I}$, S_{ij}, $i \sim j$ be independent Brownian motions with zero mean and variance given by $EA_i^2(1) = \lambda_i C_{U,i}^2$, $ES_{ij}^2(1) = \mu_{ij}$. Set $S_{ij} = 0$ for $i \not\sim j$, $A = (A_i)$ and $S = (S_{ij})$. For the fact $(\hat{A}^n, \hat{S}^n) \Rightarrow (A, S)$ see [2], Lemma 4(i). Note that by (3) and (4), $\Psi_{ij}^n(t) \leq N_j$ for every i, j and t. Hence by (29),

$$\|\hat{W}^n\|_t^* \leq c\|\hat{A}^n\|_t^* + c\|\hat{S}^n\|_{ct}^*. \tag{83}$$

As a result,

$$n^{-1/2}\|\hat{W}^n\|_t^* \to 0 \quad \text{in distribution, as } n \to \infty,\ t \geq 0. \tag{84}$$

Step 2. We show first that under p, $\hat{M}^n \Rightarrow 0$. Let $\tau^n = \inf\{s : \hat{M}_s^n > 0\}$. We shall show that, for every T, $P(\tau^n \leq T) \to 0$ as $n \to \infty$; this implies that, for every T, $\|\hat{M}^n\|_T^* \to 0$ in distribution, and as a result $\hat{M}^n \Rightarrow 0$. Indeed, by (58),

$$\hat{X}^n(t \wedge \tau^n) = \hat{X}^{0,n} + \int_0^{t\wedge\tau^n} b^n(\hat{X}^n(s), U^n(s))ds + r\hat{W}^n(t \wedge \tau^n).$$

By (59),

$$\|\hat{X}^n(t \wedge \tau^n)\| \leq c + c\int_0^{t\wedge\tau^n} (1 + \|\hat{X}^n(s)\|)ds + c\|\hat{W}^n(t \wedge \tau^n)\|,$$

and it follows from Gronwall's inequality that

$$\|\bar{X}^n - x^*\|_{t\wedge\tau^n}^* = n^{-1/2}\|\hat{X}^n\|_{t\wedge\tau^n}^* \leq cn^{-1/2}e^{ct}\|\hat{W}^n\|_t^*. \tag{85}$$

Hence by Remark (b) following Lemma 3 and (84),

$$P(\tau^n \leq T) \leq P(\|\bar{X}^n - x^*\|_{T\wedge\tau^n}^* > \alpha_0) \to 0. \tag{86}$$

As a result, $\hat{M}^n \Rightarrow 0$. By (84), (85) and (86) it follows that for every t, $\|\bar{X}^n - x^*\|_t^* \to 0$ in distribution as $n \to \infty$. As a result $\bar{X}^n \Rightarrow x^*$.

Next, under p', let

$$\sigma^n = \inf\{t > 0 : I\hat{M}^n(t) \geq 1\}. \tag{87}$$

By (58) and (59),

$$\|\hat{X}^n(t \wedge \sigma^n)\| \leq c + c\int_0^{t\wedge\sigma^n} (1 + \|\hat{X}^n(s)\|)ds + c\|\hat{W}^n(t \wedge \sigma^n)\|.$$

Using again Gronwall's lemma and (84), we have

$$\|\bar{X}^n - nx^*\|_{t\wedge\sigma^n}^* \to 0 \quad \text{in distribution, } t > 0. \tag{88}$$

Step 3. This step refers to p only. By (55),

$$e\cdot \bar{Y}^n = n^{-1/2}(\hat{M}^n + (e\cdot \hat{X}^n)^+) \le n^{-1/2}\hat{M}^n + \|\bar{X}^n - x^*\|.$$

Since $\bar{Y}_i \ge 0$, $i \in \mathcal{I}$, it follows that $\bar{Y}^n \Rightarrow 0$. By a similar argument, $\bar{Z}^n \Rightarrow 0$. By (60), (61) and linearity of the map G on D_G,

$$\begin{aligned}\bar{\Psi}^n &= n^{-1}G(X^n - Y^n, N^n - Z^n)\\ &= G(\bar{X}^n - \bar{Y}^n, n^{-1}N^n - \bar{Z}^n)\\ &= G(\bar{X}^n - x^* - \bar{Y}^n, n^{-1}N^n - \nu - \bar{Z}^n) + G(x^*, \nu)\\ &= G(\bar{X}^n - x^* - \bar{Y}^n, n^{-1}N^n - \nu - \bar{Z}^n) + \psi^*.\end{aligned}$$

Since $\bar{Y}^n, \bar{Z}^n, (\bar{X}^n - x^*) \Rightarrow 0$ and by (19) and continuity of G, we obtain $\bar{\Psi}^n \Rightarrow \psi^*$.

We have now shown that $\|\bar{X}^n - x^*\|_t^* + \|\bar{Y}^n\|_t^* + \|\bar{Z}^n\|_t^* + \|\bar{\Psi}^n - \psi^*\|_t^*$ converges to zero in distribution, for every t. Hence $(\bar{X}^n, \bar{Y}^n, \bar{Z}^n, \bar{\Psi}^n) \Rightarrow (x^*, 0, 0, \psi^*)$.

Next we show that the sequence $(\hat{X}^n, \hat{W}^n, Q^n, R^n)$ is tight in $(\mathbb{D}(\mathbb{R}^k))^3 \times \mathbb{D}(\mathbb{R})$. We have shown already that $\hat{S}^n \Rightarrow S$ and $\bar{\Psi}^n \Rightarrow \psi^*$. An application of the time change lemma [3] shows that $\hat{S}^n_{ij}(\int_0^\cdot \bar{\Psi}^n_{ij}(s)ds) \Rightarrow S_{ij}(\psi^*\cdot)$. By (29) and (31) it follows that $\hat{W}^n \Rightarrow W$, a standard Brownian motion in $\mathbb{R}^{\bar{\imath}}$.

By (58), (59) (recall that $If = \int_0^\cdot f$),

$$\|\hat{X}^n(t)\| \le \|\hat{X}^{0,n}\| + c\|\hat{W}^n(t)\| + cI\hat{M}^n(t) + c\int_0^t (1 + \|\hat{X}^n(s)\|)ds$$

and therefore by Gronwall's inequality,

$$\|\hat{X}^n\|_t^* \le ce^{ct}(1 + I\hat{M}^n_t + \|\hat{W}^n\|_t^*). \tag{89}$$

Using tightness of $\hat{W}^n$ and $\hat{M}^n$, it follows that for every t,

$$\lim_{m\to\infty} \limsup_{n\to\infty} P(\|\hat{X}^n\|_t^* \ge m) = 0. \tag{90}$$

Fix T. By (58) and (59), for $s, t \in [0, T]$, $s < t$,

$$\|\hat{X}^n(t) - \hat{X}^n(s)\| \le c\|\hat{W}^n(t) - \hat{W}^n(s)\| + c(t-s)(1 + |\hat{M}^n|_T^* + \|\hat{X}^n\|_T^*). \tag{91}$$

Let $w(x, S) = \sup_{s,t\in S} \|x(s) - x(t)\|$ where $S \subset [0, T)$, and let

$$w'_T(x, \delta) = \inf \max_{1\le i\le v} w(x, [t_{i-1}, t_i)),$$

where the infimum is over all decompositions $[t_{i-1}, t_i)$, $1 \le i \le v$ of $[0, T)$ such that $t_i - t_{i-1} > \delta$, $1 \le i \le v$ (cf. [3], p. 171). The notation

$$w_T(x, \delta) = \sup_{0\le s<t\le (s+\delta)\wedge T} w(x, [s, t]) \tag{92}$$

will also be useful. It follows from Theorem 16.8 of [3] and tightness of $\hat{W}^n$ that for each t and ε, $\lim_{\delta\to 0}\limsup_{n\to\infty} P(w'_t(\hat{W}^n,\delta)\geq\varepsilon)=0$. Hence by tightness of $\hat{M}^n$ and using (91), for each $t\leq T$ and ε,

$$\lim_{\delta\to 0}\limsup_{n\to\infty} P(w'_t(\hat{X}^n,\delta)\geq\varepsilon)=0. \tag{93}$$

Since T is arbitrary, (90) and (93) imply tightness of $\hat{X}^n$, using Theorem 16.8 of [3]. By (59), (68) and Assumption 2,

$$\|Q^n\|_t^* \vee |R^n|_t^* \leq ct(1+\|\hat{X}^n\|_t^*)^{m_L},$$

and for $s,t\in[0,T]$, $s<t$,

$$\|Q^n(t)-Q^n(s)\| \vee |R^n(t)-R^n(s)| \leq c(t-s)(1+\|\hat{X}^n\|_T^*)^{m_L}.$$

Hence, using Theorem 16.8 of [3], tightness of Q^n and R^n follows from (90). We have shown that $\hat{X}^n$, $\hat{W}^n$, Q^n and R^n are tight, and using Proposition 3.2.4 of [4], it follows that $(\hat{X}^n,\hat{W}^n,Q^n,R^n)$ is tight.

Step 4. Fix T. By (1) and (51), under p^* one has $|J^n|_T^*\leq 2(\bar{\imath}+\bar{\jmath})n^{-1/2}$ on the event $\{\|\bar{X}^n-x^*\|_T^*\leq\alpha_0\}$. Since $\|\bar{X}^n-x^*\|_T^*$ converges to zero in distribution, so does $|J^n|_T^*$, and since T is arbitrary, $J^n\Rightarrow 0$.

Step 5. This step refers to p'. Let σ^n be as in (87) and recall (88). Reviewing Step 3 shows that all its conclusions still hold under p' in place of p, upon replacing $\bar{X}^n$ by $\bar{X}^n(\cdot\wedge\sigma^n)$, $\bar{Y}^n$ by $\bar{Y}^n(\cdot\wedge\sigma^n)$, and similar substitutions for the processes $\bar{\Psi}^n$, $\hat{W}^n$, $\hat{X}^n$, Q^n and R^n. As a result,

$$(\bar{X}^n_{\cdot\wedge\sigma^n},\bar{Y}^n_{\cdot\wedge\sigma^n},\bar{Z}^n_{\cdot\wedge\sigma^n},\bar{\Psi}^n_{\cdot\wedge\sigma^n})\Rightarrow(x^*,0,0,\psi^*), \tag{94}$$

$$(\hat{X}^n_{\cdot\wedge\sigma^n},\hat{W}^n_{\cdot\wedge\sigma^n},Q^n_{\cdot\wedge\sigma^n},R^n_{\cdot\wedge\sigma^n}) \text{ is tight.} \tag{95}$$

Let T be fixed and denote $T^n=T\wedge\sigma^n$. For $i\sim j$ let $\Lambda^n_{ij}=\hat{\Psi}^n_{ij}-\check{\Psi}^n_{ij}$. Let $\Omega^{n,k}$ denote the event

$$\{\|\hat{X}^n\|_{T^n}^*\vee\|\hat{Y}^n\|_{T^n}^*\vee\|\hat{Z}^n\|_{T^n}^*\vee\|\hat{\Psi}^n\|_{T^n}^*\leq k\}.$$

By tightness of $\hat{X}^n(\cdot\wedge\sigma^n)$ and by (55), (57),

$$\lim_{k\to\infty}\liminf_{n\to\infty} P(\Omega^{n,k})=1. \tag{96}$$

We will show that

$$\lim_{\varepsilon_0\to 0+}\limsup_n P(\sigma^n\leq\varepsilon_0)=0, \tag{97}$$

and that for every $i\sim j$ and small enough $\varepsilon_0>0$,

$$\limsup_n P(\sigma^n>\varepsilon_0,\ \sup_{[\varepsilon_0,T^n]}\Lambda^n_{ij}>c\varepsilon)=0. \tag{98}$$

To this end let $\varepsilon_0 \in (0,T)$. Fix i,j, $i \sim j$. If $\Lambda^n_{ij} > 0$ on $[s,r)$, no customers are routed on activity (i,j) on this time interval and by (52) and (10),

$$\Psi^n_{ij}(t) = \Psi^n_{ij}(s) - \Delta^n_{ij}(s,t), \quad t \in [s,r),$$

where

$$\Delta^n_{ij}(s,t) = S^n_{ij}(I\Psi^n_{ij}(t)) - S^n_{ij}(I\Psi^n_{ij}(s)). \tag{99}$$

Therefore

$$\hat{\Psi}^n_{ij}(t) = \hat{\Psi}^n_{ij}(s) - n^{-1/2}\Delta^n_{ij}(s,t).$$

Now,

$$P(\sigma^n > \varepsilon_0, \sup_{t\in[\varepsilon_0,T^n]} \Lambda^n_{ij}(t) > 3\varepsilon) \le P((\Omega^{n,k})^c) + P(\Omega^{n,k}_1) + P(\Omega^{n,k}_2), \tag{100}$$

where

$$\Omega^{n,k}_1 = \Omega^{n,k} \cap \{\sigma^n > \varepsilon_0\} \cap \{\exists 0 \le s \le r \le T^n : \Lambda^n_{ij}(s) \le \varepsilon, \ \inf_{t\in(s,r)} \Lambda^n_{ij}(t) \ge \varepsilon, \ \Lambda^n_{ij}(r) \ge 3\varepsilon\},$$

$$\Omega^{n,k}_2 = \Omega^{n,k} \cap \{\sigma^n > \varepsilon_0\} \cap \{\inf_{t\in[0,\varepsilon_0]} \Lambda^n_{ij} \ge \varepsilon\}.$$

Let k be fixed. By the Hölder assumption on h_1, h_2 away from $e\cdot\hat{X}^n = 0$, there are positive constants c_k, p_k (depending on k and ε but not on n) such that the following holds on $\Omega^{n,k}$:

$$\begin{aligned}
|\check{\Psi}^n_{ij}(t) - \check{\Psi}^n_{ij}(s)| &= |G(\hat{X}^n(t) - \check{Y}^n(t), -\check{Z}^n(t))_{ij} - G(\hat{X}^n(s) - \check{Y}^n(s), -\check{Z}^n(s))_{ij}| \\
&\le c(\|(\hat{X}^n(t) - \check{Y}^n(t)) - (\hat{X}^n(s) - \check{Y}^n(s))\| + \|\check{Z}^n(t) - \check{Z}^n(s)\|) \\
&\le c'_k(\|\hat{X}^n(t) - \hat{X}^n(s)\| + \|\hat{X}^n(t) - \hat{X}^n(s)\|^{p_k}) + (\varepsilon/4)1_{|e\cdot\hat{X}^n_s|<\varepsilon/8} + (\varepsilon/4)1_{|e\cdot\hat{X}^n_t|<\varepsilon/8} \\
&\le c_k\|\hat{X}^n(t) - \hat{X}^n(s)\|^{p_k} + \varepsilon/2.
\end{aligned} \tag{101}$$

By (22), for $0 \le s \le t \le r \le T^n$ as above,

$$\begin{aligned}
n^{-1/2}\Delta^n_{ij}(s,t) = \hat{S}^n_{ij}(I\bar{\Psi}^n_{ij}(t)) - \hat{S}^n_{ij}(I\bar{\Psi}^n_{ij}(s)) + n^{1/2}\mu^n_{ij}\int_s^t (\bar{\Psi}^n_{ij}(\theta) - \psi^*_{ij})d\theta \\
+ n^{1/2}(\mu^n_{ij} - \mu_{ij})\psi^*_{ij}(t-s) + n^{1/2}\mu_{ij}\psi^*_{ij}(t-s).
\end{aligned} \tag{102}$$

Since $\Psi^n_{ij} \le N^n_j$, it follows that $\bar{\Psi}^n_{ij}$ are all bounded uniformly by a constant. Since the jumps of the process $\hat{S}^n_{ij}$ are all of size $n^{-1/2}$, a use of (12.9) of [3] shows that $|\hat{S}^n_{ij}(I\bar{\Psi}^n_{ij}(t)) - \hat{S}^n_{ij}(I\bar{\Psi}^n_{ij}(s))| \le 2w'_{c^1T^n}(\hat{S}^n_{ij}, c^1(t-s)) + n^{-1/2}$, where $c^1 > 0$ is a constant. Since $\|\bar{\Psi}^n - \psi^*\|^*_{T^n} \le kn^{-1/2}$ on $\Omega^{n,k}$, using (18), and assuming n is large enough we therefore have

$$n^{-1/2}\Delta^n_{ij}(s,t) \ge -2w'_{c^1T^n}(\hat{S}^n_{ij}, c^1(t-s)) - n^{-1/2} + c^2n^{1/2}(t-s), \tag{103}$$

where $c^2 = \mu_{ij}\psi^*_{ij}/2$. Hence on $\Omega^{n,k}_1$, noting that $\Lambda^n_{ij}(s) \le \varepsilon$,

$$\begin{aligned}
\Lambda^n_{ij}(t) &= \Lambda^n_{ij}(s) + (\hat{\Psi}^n_{ij}(t) - \hat{\Psi}^n_{ij}(s)) - (\check{\Psi}^n_{ij}(t) - \check{\Psi}^n_{ij}(s)) \\
&\le -n^{-1/2}\Delta^n_{ij}(s,t) + c_k\|\hat{X}^n(t) - \hat{X}^n(s)\|^{p_k} + 2\varepsilon \\
&\le 2w'_{c^1T^n}(\hat{S}^n_{ij}, c^1(t-s)) - c^2n^{1/2}(t-s) + c_k\|\hat{X}^n(t) - \hat{X}^n(s)\|^{p_k} + 2\varepsilon.
\end{aligned}$$

Since on $\Omega_1^{n,k}$, $\Lambda_{ij}^n(r) \geq 3\varepsilon$,

$$\varepsilon \leq 2w'_{c^1T^n}(\hat{S}_{ij}^n, c^1(r-s)) - c^2 n^{1/2}(r-s) + c_k\|\hat{X}^n(r) - \hat{X}^n(s)\|^{p_k} \tag{104}$$

$$P(\Omega_1^{n,k}) \leq P(\Omega_3^{n,k}) + P(\Omega_4^{n,k}),$$

where

$$\Omega_3^{n,k} = \{\exists 0 \leq s \leq r \leq T^n : r - s \leq an^{-1/2},\ (104) \text{ holds}\},$$
$$\Omega_4^{n,k} = \{\exists 0 \leq s \leq r \leq T^n : r - s > an^{-1/2},\ (104) \text{ holds}\}.$$

On $\Omega_3^{n,k}$,

$$\varepsilon \leq 2w'_{c^1T^n}(\hat{S}_{ij}^n, c^1 an^{-1/2}) + c_k w'_{T^n}(\hat{X}^n, an^{-1/2})^{p_k}$$

and since $(\hat{S}^n, \hat{X}^n(\cdot \wedge \sigma^n))$ are tight, for every a,

$$\lim_n P(\Omega_3^{n,k}) = 0.$$

On $\Omega_4^{n,k}$,

$$c^2 a \leq 2\|\hat{S}_{ij}^n\|^*_{c^1T^n} + c_k(2\|\hat{X}^n\|^*_{T^n})^{p_k}$$

and by tightness of $(\hat{S}^n, \hat{X}^n(\cdot \wedge \sigma^n))$,

$$\lim_{a\to\infty} \limsup_{n\to\infty} P(\Omega_4^{n,k}) = 0.$$

As a result we have, for every k,

$$\lim_{n\to\infty} P(\Omega_1^{n,k}) = 0. \tag{105}$$

Regarding $\Omega_2^{n,k}$, substituting $s = 0$ and $t \in [0, \varepsilon_0]$ in (101), (102) and (103),

$$\Lambda_{ij}^n(t) \leq \Lambda_{ij}^n(0) + 2\|\hat{S}_{ij}^n\|^*_{c^1\varepsilon_0} - c^2 n^{1/2} t + c_k\|\hat{X}^n(t) - \hat{X}^{0,n}\|^{p_k} + \varepsilon/2.$$

Since $\Lambda_{ij}^n(\varepsilon_0) \geq \varepsilon$, we have on $\Omega_2^{n,k}$

$$\varepsilon/2 \leq \Lambda_{ij}^n(0) + 2\|\hat{S}_{ij}^n\|^*_{c^1\varepsilon_0} - c^2 n^{1/2}\varepsilon_0 + c_k\|\hat{X}^n(\varepsilon_0) - \hat{X}^{0,n}\|^{p_k}.$$

By tightness of the random variables $\xi^n := 2\|\hat{S}_{ij}^n\|^*_{c^1\varepsilon_0} + c_k\|\hat{X}^n(\varepsilon_0 \wedge \sigma^n) - \hat{X}^{0,n}\|^{p_k}$, the fact that $\sigma^n > \varepsilon_0$ on $\Omega_2^{n,k}$, and since $\Lambda_{ij}^n(0)$ are bounded (as follows from (20), (21)),

$$\limsup_{n\to\infty} P(\Omega_2^{n,k}) \leq \limsup_{n\to\infty} P(\xi^n \geq c\varepsilon_0 n^{1/2}) = 0. \tag{106}$$

Combining (100), (105), (106) and (96), we obtain (98).

Now, on $\Omega^{n,k}$, $\hat{M}^n(\sigma^n \wedge \varepsilon_0) \leq e \cdot \hat{Y}^n(\sigma^n \wedge \varepsilon_0) \leq k$. Hence by (87),

$$P(\sigma^n \leq \varepsilon_0) \leq P(I\hat{M}^n(\sigma^n \wedge \varepsilon_0) \geq 1) \leq P((\Omega^{n,k})^c) + 1_{\varepsilon_0 k \geq 1}.$$

Therefore, for every k,

$$\limsup_{\varepsilon_0 \to 0^+} \limsup_n P(\sigma^n \le \varepsilon_0) \le \limsup_n P((\Omega^{n,k})^c),$$

and (97) follows using (96).

Having established (97) and (98), we argue as follows. Lemma 5 and the property (53) of the policy p' imply that

$$\|\Lambda^n(t)\| \le c \sum_{i \sim j} (\Lambda^n_{ij}(t))^+. \tag{107}$$

By (50) and the uniqueness statement in Lemma 1, $\hat{X}^n_i - \check{Y}^n_i = \sum_j \check{\Psi}^n_{ij}$ for every i. This and (25), along with an analogous argument for $\check{Z}^n$ imply that

$$\|\hat{Y}^n - \check{Y}^n\| + \|\hat{Z}^n - \check{Z}^n\| \le c\|\Lambda^n\|. \tag{108}$$

Combining (80), (107) and (108), on $\Omega^{n,k}$, for every $t \in (\varepsilon_0, T]$,

$$I\hat{M}^n(t) \le ck\varepsilon_0 + ct \sum_{i \sim j} \sup_{[\varepsilon_0, t]} (\Lambda^n_{ij})^+. \tag{109}$$

Hence

$$\begin{aligned} P(\sigma^n \le T) &\le P((\Omega^{n,k})^c) + P(\sigma^n \le \varepsilon_0) + P((\Omega^{n,k})^c \cap \{\sigma^n > \varepsilon_0, I\hat{M}^n(T^n) \ge 1\}) \\ &\le P((\Omega^{n,k})^c) + P(\sigma^n \le \varepsilon_0) + P(\sigma^n > \varepsilon_0, cT \sum_{i \sim j} \sup_{[\varepsilon_0, T^n]} (\Lambda^n_{ij})^+ \ge 1 - ck\varepsilon_0). \end{aligned} \tag{110}$$

Taking ε_0 small enough and using (96), (97) and (98), we have that

$$\lim_n P(\sigma^n \le T) = 0. \tag{111}$$

Since T is arbitrary, we finally have from (94), (95) and (111) that

$$(\bar{X}^n, \bar{Y}^n, \bar{Z}^n, \bar{\Psi}^n) \Rightarrow (x^*, 0, 0, \psi^*), \quad (\hat{X}^n, \hat{W}^n, Q^n, R^n) \text{ is tight.}$$

Also, with (111), the relations (98) and (107) show that

$$\lim_n P(\sup_{[\varepsilon_0, T]} \|\Lambda^n\| > c\varepsilon) = 0. \tag{112}$$

In view of (67) and (108), we have from (112) that $|J^n|^*_{s,t}$ converges to zero in distribution, for every $0 < s < t < \infty$. As follows from (80), a similar statement holds for $|\hat{M}^n|^*_{s,t}$. Moreover, using again (109), now equipped with (112), letting $n \to \infty$, then $\varepsilon_0 \to 0^+$ and finally $k \to \infty$ we obtain that $I\hat{M}^n(T) \to 0$ in distribution, and since T is arbitrary, $I\hat{M}^n \Rightarrow 0$. This completes the proof of Proposition 1. $\square$

3.3 Large time estimates

In this section we prove Proposition 2. A key ingredient is the following estimate from [1].

Proposition 3 *Let Assumption 3 hold. Let* (25)–(28) *hold. Then*

$$\|\hat{X}^n(t)\| \le C(1+t)^m(\|x\| + \|\hat{W}^n\|_t^* + |\hat{M}^n|_t^*). \tag{113}$$

Proof of Proposition 3: We use the following result from [1]: If

$$x_i = w_i - \sum_j \mu_{ij} I\psi_{ij}, \quad \psi = G(x-y,-z), \quad e\cdot y \wedge e\cdot z = 0, \tag{114}$$

then

$$\|x\| \le c(1+t)^m \|w\|_t^*, \tag{115}$$

where the constants c, m do not depend on ψ, x, y, z: Under Assumption 3(i), (115) follows from Propositions 3 and 4 of [1]. Under Assumption 3(ii), (115) follows from Proposition 3, Theorem 3 and Lemma 4 of [1].

Let

$$W_1^n(t) = \hat{W}^n(t) - r^{-1}\int_0^t \hat{M}^n(s)\mu^n \circ G(u^n(s), v^n(s))ds, \tag{116}$$

we have from (58) $\hat{X}_t^n = \hat{X}^{0,n} + \int_0^t b^n(\hat{X}_s^t, U_s^n)ds + rW_1^n(t)$. Lemma 2 implies that there are Ψ_1^n, Y_1^n and Z_1^n such that (34)–(37) hold. Hence $\hat{X}^n, Y_1^n, Z_1^n, \Psi_1^n, W_1^n$ satisfy relations analogous to (114). As a result, a relation as in (115) holds, and using (116) we obtain (113). □

Remark. The only two places in this paper where Assumption 3 is used are in Theorem 1 and in obtaining (115) from (114). In fact, also the proof of Theorem 1 (that was carried out in [1]) uses this assumption only in order to establish (115). In other words, if the implication "(114) implies (115)" holds true in greater generality then so do Theorems 1 and 2.

Proof of Proposition 2: First consider policies of the form p. By (3), (19) and (29),

$$\|\hat{W}^n\|_t^* \le c(\|\hat{A}^n\|_t^* + \|\hat{S}^n\|_{ct}^*).$$

By Assumption 4, applying Lemma 8 shows that $E(\|\hat{A}^n\|_t^*)^{m_A} \le c(1+t)^{m_2}$, and a similar estimate holds for $\hat{S}^n$. As a result,

$$E(\|\hat{W}^n\|_t^*)^{m_A} \le c(1+t)^{m_2}. \tag{117}$$

As in the proof of Proposition 1, let $\tau^n = \inf\{t : \hat{M}_t^n > 0\}$. Under $\{\tau^n > t\}$, $|\hat{M}^n|_t^* = 0$ and therefore by Proposition 3,

$$\|\hat{X}_t^n\| \le C(1+t)^m(\|x\| + \|\hat{W}^n\|_t^*), \quad \{\tau^n > t\}. \tag{118}$$

On $\{\tau^n \le t\}$, by (3), (5) and (19),

$$|\hat{M}^n|_t^* = n^{-1/2}|M^n|_t^* \le n^{-1/2}|e\cdot Z^n|_t^* \le n^{-1/2} e\cdot N^n \le cn^{1/2},$$

and by Proposition 3,

$$\|\hat{X}_t^n\| \le C(1+t)^m(\|x\| + \|\hat{W}^n\|_t^* + n^{1/2}), \quad \{\tau^n \le t\}. \tag{119}$$

Combining (117), (118) and Lemma 3 (in particular, Remark (b) that follows),

$$P(\tau^n \le t) \le P(\|\hat{X}^n\|_{t\wedge\tau^n}^* \ge \alpha_0 n^{1/2}) \le cn^{-m_A/2} E(\|\hat{X}^n\|_{t\wedge\tau^n}^*)^{m_A} \le cn^{-m_A/2}(1+t)^{m_3}. \tag{120}$$

Therefore by (119), (120) and Hölder inequality, with $q^{-1} + \bar{q}^{-1} = 1$ and $qm_0 = m_A$,

$$\begin{aligned} E\|\hat{X}_t^n\|^{m_0} &\le E[\|\hat{X}_t^n\|^{m_0} 1_{\tau^n > t}] + (E\|\hat{X}_t^n\|^{m_A})^{1/q}(P(\tau^n \le t))^{1/\bar{q}} \\ &\le c(1+t)^{m_4} + cn^{(2q)^{-1}m_A} n^{-(2\bar{q})^{-1}m_A}(1+t)^{m_3/\bar{q}} \\ &\le c(1+t)^{m_5}, \end{aligned} \tag{121}$$

where the inequality $q^{-1} - \bar{q}^{-1} \le 0$, used on the last line above, follows from $m_0 < m_A/2$.

Next consider the policy p'. Let $b > 0$ be a constant that does not depend on n or t, whose actual value is determined later. Let

$$\theta_n = \inf\{t \ge 0 : \max_{i\sim j} \Lambda_{ij}^n \ge b\}.$$

By (80), (107) and (108),

$$\hat{M}^n \le J^n \le c\|\Lambda^n\| \le c \max_{i\sim j}(\Lambda_{ij}^n)^+.$$

Letting $T_n = T \wedge \theta_n$, it follows that

$$I\hat{M}^n(T_n) \le cT_n,$$

(c depends on b but not on n, T). Hence by Proposition 3, we have

$$\|\hat{X}^n\|_{T_n}^* \le c(1+T_n)^m(\|x\| + \|\hat{W}^n\|_{T_n}^* + cT_n) \le c(1+T)^{m+1}(1 + \|\hat{W}^n\|_T^*), \tag{122}$$

where c depends on x and b but not on T or n. We establish below the estimate

$$P(\theta_n \le T) \le c_0 n^{1/4 - m_A/8}(1+T)^m, \tag{123}$$

where c_0 and m are constants that do not depend on T or n. Repeating the argument of (121), with θ_n in place of τ^n, and (122) [resp., (123)] in place of (118) [resp., (120)], shows that $E\|\hat{X}_T^n\|^{m_0} \le c(1+T)^{m_6}$, with $qm_0 = m_A$, provided that $m_0 < m_A(m_A-2)(5m_A-2)^{-1}$.

In the rest of this section we prove (123). The argument is similar to that used to prove tightness in Section 3.2, but since the estimates are uniform in time, a more careful analysis is required. By (55) and (122),

$$\|\hat{X}^n\|_{T_n}^*, \|\hat{Y}^n\|_{T_n}^*, \|\hat{Z}^n\|_{T_n}^*, \|\Psi^n\|_{T_n}^* \le c(1+T)^{m+1}(1 + \|\hat{W}^n\|_T^*) =: \xi(n,T).$$

Clearly

$$P(\theta_n \le T) \le \sum_{i\sim j} P(\sup_{t\le T} \Lambda_{ij}^n(t) \ge b).$$

Let $i \sim j$ be fixed. Then

$$P(\sup_{t\le T} \Lambda_{ij}^n(t) \ge b) \le P(\exists 0 \le s \le r \le T_n : \Lambda_{ij}^n(s) \le b/2, \Lambda_{ij}^n(t) > 0, t \in (s,r), \Lambda_{ij}^n(r) \ge b). \quad (124)$$

Here b is chosen so that $1 + \max_{i\sim j} \max_n \Lambda_{ij}^n(0) < b/2 < \infty$. Note that with Δ^n as in (99), (102) still holds. Arguing as in (101), using the global Hölder assumption,

$$|\check{\Psi}_{ij}^n(t) - \check{\Psi}_{ij}^n(s)| \le c\xi(n,T) g(\hat{X}^n(t) - \hat{X}^n(s)) + 1,$$

where $g(x) = \|x\| + \|x\|^c$. Also, $\bar{\Psi}^n - \psi^* = n^{-1/2}\hat{\Psi}^n$ and therefore $\|\Psi^n - \psi^*\|_{T_n}^* \le n^{-1/2}\xi(n,T)$. Thus for s,t as in (124), recalling the notation (92) for $w_T(x,\delta)$,

$$n^{-1/2}\Delta_{ij}^n(s,t) \ge -2w_{cT_n}(\hat{S}_{ij}^n, c(t-s)) - c\xi(n,T)(t-s) + cn^{1/2}(t-s).$$

As a result,

$$b/2 \le \Lambda_{ij}^n(t) \le 2w_{cT_n}(\hat{S}_{ij}^n, c(t-s)) + \xi(n,T)(t-s) - cn^{1/2}(t-s) + c\xi(n,T) g(\hat{X}^n(t) - \hat{X}^n(s)). \quad (125)$$

Hence

$$P(\theta_n \le T) \le P(\Omega_1^n) + P(\Omega_2^n),$$

where

$$\Omega_1^n = \{\exists 0 \le s \le t \le T_n : t - s \le n^{-1/4},\ (125) \text{ holds}\},$$
$$\Omega_2^n = \{\exists 0 \le s \le t \le T_n : t - s > n^{-1/4},\ (125) \text{ holds}\}.$$

On Ω_1^n,

$$b/2 \le 2w_{cT_n}(\hat{S}_{ij}^n, n^{-1/4}) + \xi(n,T)n^{-1/4} + c\xi(n,T) g(\hat{X}_t^n - \hat{X}_s^n),$$

where $t - s \le n^{-1/4}$. By Lemma 9 below, if one chooses $b > 12$,

$$P(2w_{cT}(\hat{S}_{ij}^n, n^{-1/4}) \ge b/6) \le cTn^{1/4 - m_A/8}.$$

By (58), and the bound b on $\hat{M}^n$, we have $\|\hat{X}^n(t) - \hat{X}^n(s)\| \le c\xi(n,T)(t-s) + c\|\hat{W}^n(t) - \hat{W}^n(s)\| + b(t-s)$. Arguing again by Lemma 9, choosing β_0 large enough,

$$P(c\xi(n,T) \max\{g(\hat{X}_t^n - \hat{X}_s^n) : 0 \le t - s \le n^{-1/4}, t \le T_n\} \ge b/6) \le cT^{\beta_1} n^{1/4 - m_A/8}.$$

Also, by (83) and Lemma 8,

$$P(\xi(n,T)n^{-1/4} \ge b/6) \le P(\|\hat{W}^n\|_T^* \ge cn^{1/4}) \le c(1+T)^{m_A/2} n^{-m_A/4}.$$

Hence

$$P(\Omega_1^n) \le cT^{\beta_1} n^{1/4 - m_A/8} + c(1+T)^{m_A/2} n^{-m_A/4}. \quad (126)$$

On Ω_2^n,

$$n^{1/4} \le \|\hat{S}_{ij}^n\|_{cT_n}^* + \xi(n,T)T_n + c\xi(n,T)g(\hat{X}_t^n - \hat{X}_s^n) \le \|\hat{S}_{ij}^n\|_{cT_n}^* + c(1+T)^{m_1}(1+\|\hat{W}^n\|_T^*)^2,$$

for an appropriate m_1. Hence by (83), $n^{1/8} \le c_3(1+T)^{m_2}(1+\|\hat{A}^n\|_T^* + \|\hat{S}^n\|_{cT}^*)$. Using Lemma 8,

$$P(\|\hat{A}^n\|_T^* \ge n^{1/8}(1+T)^{-m_2}) \le c(1+T)^{m_A/2+m_2m_A}n^{-m_A/8}$$

A similar bound holds for $\hat{S}^n$. Hence

$$P(\Omega_2^n) \le c(1+T)^{m_A/2+m_2m_A}n^{-m_A/8}. \tag{127}$$

Combining (126) and (127) we obtain (123). □

Lemma 9 *Given $\beta_0 \ge 0$ there are constants c_1, β_1, independent of n and T such that*

$$P(w_T(\hat{A}_i^n, n^{-1/4}) \ge T^{-\beta_0}) \le c_1 T^{\beta_1} n^{1/4-m_A/8}, \quad n \in \mathbb{N}, T \ge 1.$$

A similar estimate holds for $\hat{S}_{ij}^n$ in place of $\hat{A}_i^n$.

Proof of Lemma 9: Fix i and suppress it from the notation. By (22),

$$\hat{A}^n(t) - \hat{A}^n(s) = n^{-1/2}(A^n(t) - A^n(s)) - n^{-1/2}\lambda^n(t-s).$$

Recall that $EA^n(T) = \lambda^n T$. Thus

$$P(w_T(\hat{A}^n, n^{-1/4}) \ge T^{-\beta_0}) \le P(A^n(T) > 2\lambda^n T) + P(\Omega_{0,+}^n) + P(\Omega_{0,-}^n), \tag{128}$$

where

$$\Omega_{0,+}^n = \{A^n(T) \le 2\lambda^n T, \exists s,t \in [0, 2\lambda^n T], 0 \le t-s \le n^{-1/4}, A^n(t) - A^n(s) \ge n^{1/2}T^{-\beta_0} + \lambda^n(t-s)\},$$

$$\Omega_{0,-}^n = \{A^n(T) \le 2\lambda^n T, \exists s,t \in [0, 2\lambda^n T], 0 \le t-s \le n^{-1/4}, A^n(t) - A^n(s) \le -n^{1/2}T^{-\beta_0} + \lambda^n(t-s)\}.$$

Recall that $E\check{U}(k) = 1$. Letting $\bar{U}^n(k) = U^n(k) - (\lambda^n)^{-1}$, by (6) we have $E\bar{U}^n(k) = 0$. Let $M_j^n = \sum_{k=1}^j \bar{U}^n(k)$ and note that it is a martingale. For a real-valued function X on $\mathbb{Z}_+$ let $\mathrm{osc}(X,i,j) = \max\{|X(k) - X(l)| : k,l \in [i+1, j]\}$. By (7), using $\lambda^n \le c_2 n$, denoting $\rho = c_2 n^{3/4}$, we have

$$\begin{aligned}
P(\Omega_{0,+}^n) &\le P(\exists j \le 2\lambda^n T, \exists r \le n^{-1/4}\lambda^n, \sum_{k=j+1}^{j+n^{1/2}T^{-\beta_0}+r} U^n(k) \le (\lambda^n)^{-1} r) \\
&\le (\exists j \le 2c_2 nT, \exists r \le c_2 n^{3/4}, \sum_{k=j+1}^{j+n^{1/2}+r} \bar{U}^n(k) \le -(\lambda^n)^{-1} n^{1/2} T^{-\beta_0}) \\
&\le P(\exists j \le 2c_2 nT, \mathrm{osc}(M^n, j, j + 2c_2 n^{3/4}) \ge c_2^{-1} n^{-1/2} T^{-\beta_0}) \\
&\le P(\exists j \le 2c_2 nT, j = 0, \rho, 2\rho, \dots, \mathrm{osc}(M^n, j, j+\rho) \ge c_2^{-1} n^{-1/2} T^{-\beta_0}/3) \\
&\le 1 - (1 - P(|M^n|_\rho^* \ge c_2^{-1} n^{-1/2} T^{-\beta_0}/6))^{2n^{1/4}T}
\end{aligned}$$

Burkholder's inequality shows that

$$E(|M^n|^*_\rho)^{m_A} \le c_3 E(\rho|\bar{U}(1)|^2)^{m_A/2} \le c_4 n^{3m_A/8}(\lambda^n)^{-m_A} \le c_5 n^{-5m_A/8}.$$

Hence

$$P(\Omega^n_{0,+}) \le 1 - (1 - c_6 n^{-m_A/8} T^{m_A\beta_0})^{2n^{1/4}T} \le c_7 n^{1/4-m_A/8} T^{1+m_A\beta_0},$$

for an appropriate constant c_7. By a similar argument one shows that a similar bound holds for $P(\Omega^n_{0,-})$. As follows from Lemma 8,

$$P(A^n(T) > 2EA^n(T)) \le cn^{-m_A/2}(1+T)^{-m_A/2}.$$

Hence by (128),

$$P(w_T(\hat{A}^n, n^{-1/4}) \ge 1) \le c_1 n^{1/4-m_A/8} T^{1+m_A\beta_0}$$

for an appropriate constant c_1 independent of n and $T \ge 1$. Since the Poisson processes S^n_{ij} are in particular renewal processes (with finite m_Ath moment for interarrival times), a similar result holds for $\hat{S}^n_{ij}$. □

References

[1] R. Atar. Treelike parallel server stations in heavy traffic. Preprint

[2] R. Atar, A. Mandelbaum and M. Reiman. Scheduling a multi-class queue with many exponential servers: asymptotic optimality in heavy-traffic. *Ann. Appl. Probab.*, to appear

[3] P. Billingsley. *Convergence of probability measures. Second edition.* Wiley, New York, 1999.

[4] S. N. Ethier and T. G. Kurtz. *Markov processes. Characterization and convergence.* Wiley, New York, 1986.

[5] W. H. Fleming and H. M. Soner. *Controlled Markov processes and viscosity solutions.* Springer-Verlag, New York, 1993.

[6] N. Gans, G. Koole, A. Mandelbaum. Telephone Call Centers: Tutorial, Review and Research Prospects. MSOM 5(2), 2003.

[7] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29 (1981), no. 3, 567–588.

[8] J. M. Harrison. Brownian models of open processing networks: canonical representation of workload. *Ann. Appl. Probab.* 10 (2000), no. 1, 75–103.

[9] J.M. Harrison and M.J. López. Heavy traffic resource pooling in parallel-server systems. *Queueing Systems*, 33:339-368, 1999.