J. Appl. Prob. **45,** 347–362 (2008)
Printed in England

DIFFUSION APPROXIMATION OF STATE-DEPENDENT G-NETWORKS UNDER HEAVY TRAFFIC

SAUL C. LEITE * AND
MARCELO D. FRAGOSO,* ** *LNCC*

Abstract

This paper is concerned with the characterization of weak-sense limits of state-dependent G-networks under heavy traffic. It is shown that, for a certain class of networks (which includes a two-layer feedforward network and two queues in tandem), it is possible to approximate the number of customers in the queue by a reflected stochastic differential equation. The benefits of such an approach are that it describes the transient evolution of these queues and allows the introduction of controls, *inter alia*. We illustrate the application of the results with numerical experiments.

Keywords: Heavy-traffic approximation; G-network

2000 Mathematics Subject Classification: Primary 60K25
Secondary 93E05

1. Introduction

Queueing systems that receive signals, in addition to customers, are called G-networks and were first introduced in [22]. Signals may come from outside or from other queues within the network and cause different types of effects on the receiving queue. A common type of signal, which is called the 'negative customer', forces the receiving queue to remove a customer from the waiting line. Other examples of signals include: 'triggers', which moves a customer from one queue to another [23]; 'disasters', which completely cleans the waiting line of the receiving queue [15]; and 'resets', which sets the length of the receiving queue to a random value distributed according to the stationary distribution for that queue [26]. Thus, every queue in the system may exert some sort of control over the network through the signals. These models have been extensively studied (some examples include [14], [17], [18], [29], [33], [34], [36], [37], [47], [48], and [52]) and are motivated by a series of practical applications. One of the most successful applications, which was also the initial motivation for G-networks, is neural network modeling [21], [24], [30]. In this context, each queue represents a neuron and positive and negative customers are interpreted as excitatory and inhibitory signals, respectively (see also, [6], [7], and [27]). Other applications include computer networks with virus infection, load balancing networks, and synchronization signaling in parallel computation (see [5] for an extensive list of references). Another more recent application is modeling genetic regulatory systems [3], [25].

Received 21 December 2007; revision received 5 March 2008.
* Postal address: Departamento de Sistemas e Controle, LNCC, CEP 25651-075, Quitandinha, Petrópolis, RJ, Brazil.
** Email address: frag@lncc.br

Although G-networks generally have some pleasing mathematical properties, such as product-form stationary distributions, the transient evolution of these systems is not easily (or conveniently) described and is rarely treated. The interaction among several different queues and the discrete nature of the system contribute to making it a complex problem and often the only resource available are simulations, which are time consuming and computationally expensive. Moreover, problems such as the optimum choice of signal or customer scheduling are impractical in this setting. Thus, a mathematical model is sought, even if approximative, that can give a reasonable degree of accuracy.

There exists two common types of approximations that describe the transient evolution of queueing networks: fluid and diffusion (or heavy-traffic) approximations. Usually, fluid models describe the dynamics of the system 'average' by a differential equation. Diffusion approximations differ from the fluid model in the fact that the 'randomness' usually found in queueing systems is not averaged out and it appears in the model as a Wiener process (or in some cases as an Itô integral). Hence, diffusion approximations are more faithful to the dynamics of the system when compared to fluid approximations. However, this comes with the addition of the heavy-traffic assumption, which requires the rate of customers entering a queue to be close to the rate of customers leaving this queue. This is a common scenario in many applications of interest, most notably in modern computer systems.

The problem of describing the transient evolution of a queueing network with negative customers has been dealt with in some recent works using fluid approximations [4], [35]. In the former article, transient evolution of a state-dependent network with negative customers was considered using a fluid approximation together with a heavy-traffic assumption. However, as discussed in the above paragraph, diffusion approximations are more suited for systems under this condition. To our knowledge, G-networks have not yet been treated under a diffusion analysis. Such an approximation is useful in practical problems in which G-networks are applicable. For example, one could use the heavy-traffic approximation to construct a stochastic optimal control problem for synchronization of signals in parallel computer systems. In addition, the diffusion model can help us gain insights into the connections among some of the model parameters and the general behavior of queueing networks with signals.

Diffusion approximations for queueing systems have been studied since the pioneering works of Kingman [41], Prohorov [50], and Borovkov [10], [11] in the early 1960s. Other early papers on the subject include [19], [20], [39], [51], and [53], to cite a few. One of the interesting aspects of diffusion approximations is that they offer a 'macroscopic view' [54] of the complex interactions that are present in queueing networks, and synthesize the general behavior of the system in a simple time-dependent equation. In addition, the approximation has been observed to give a good estimate for systems under heavy, or only moderately heavy, traffic (see e.g. [43, Chapter 1]). Hence, it is no surprise that it has been successfully applied to several practical problems, most of them in computer systems, where heavy traffic is common. Some examples include [1], [2], [13], [28], [31], [32], and [45].

In this paper we will consider G-networks with *state-dependent* arrival rates (of customers and signals), service rates, and routeing probabilities. Besides [4], discussed above, other results regarding state-dependent G-queues and networks treat the system under a stationary regime [9], [17], [38]. As mentioned previously, the benefits of the diffusion approximation is that it describes the transient evolution of these networks via a stochastic model. In addition, the state dependence allows for the introduction of feedback controls [43, Chapter 9]. We also consider that the network is *under heavy traffic*, in the sense that every queue in the system is

https://doi.org/10.1239/jap/1214950352 Published online by Cambridge University Press

operating at nearly maximum capacity. Under this condition, it will be shown that the number of customers in each queue in the network can be approximated by a *reflected stochastic differential equation*. The model is an adaptation of the models presented in [43, Section 8.2] and [49] with the introduction of negative customers. The result presented here is for a certain class of queueing networks which satisfy Assumption 3(a), which will be presented later in the development of the model. Two examples of networks are given which satisfy this condition: two queues in tandem and a two-layer feedforward network.

The layout of the paper is as follows: in the following section the queueing model treated here will be described in more detail. In Section 3 the heavy-traffic theorem for the number of customers will be stated and proved. In Section 4 we show two examples of networks which satisfy Assumption 3(a). Finally, in Section 5 we illustrate the application of the model with a numerical example.

2. Queueing model

We will restrict ourselves to queues with one server, first-come–first-served (FCFS) service discipline, and signals of the 'negative customer' type. Hence, any queue that receives this signal is forced to remove a customer from the system. If the queue is empty, the negative customer will have no effect on the system. Although being denominated a 'customer', this signal does not receive service and leaves the receiving queue immediately after its arrival. Signals coming from within the network are regular customers that have finished work at a queue and were routed as negative customers.

The queue length process, X_i, for a network of K queues takes the form

$$X_i(t) = X_i(0) + A_i(t) - D_i(t) - S_i(t) + \sum_{j \le K} (D_{ji}^+ - D_{ji}^-(t)) - U_i(t), \tag{1}$$

$$\begin{aligned}
A_i(t) &= N_i^a\left(\int_0^t \Lambda_i^a(X(s))\,\mathrm{d}s\right), & B_i(t) &= N_i^s\left(\int_0^t \Lambda_i^s(X(s))\,\mathrm{d}s\right),\\
C_i(t) &= N_i^d\left(\int_0^t \Lambda_i^d(X(s))\,\mathrm{d}s\right), & S_i(t) &= \int_0^t \mathbf{1}_{\{X_i(s-)>0\}}\,\mathrm{d}B_i(s),\\
D_i(t) &= \int_0^t \mathbf{1}_{\{X_i(s-)>0\}}\,\mathrm{d}C_i(s), & \tilde{D}_{ij}(t) &= \int_0^t \mathbf{1}_{\{X_i(s-)>0, X_j(s-)>0\}}\,\mathrm{d}C_i(s),\\
D_{ij}^+(t) &= \int_0^t \mathbf{1}_{ij}^+(s)\,\mathrm{d}D_i(s), & D_{ij}^-(t) &= \int_0^t \mathbf{1}_{ij}^-(s)\,\mathrm{d}\tilde{D}_{ij}(s),
\end{aligned} \tag{2}$$

where N_i^α are standard Poisson processes with càdlàg sample paths (those that are continuous from the right with left limits) and $\Lambda_i^\alpha\colon \mathbb{R}_+^K \to \mathbb{R}_+$, $i, j \in \{1, \dots, K\}$, $\alpha \in \{a, s, d\}$, are measurable functions.

The processes $\mathbf{1}_{ji}^+(t)$ and $\mathbf{1}_{ji}^-(t)$ are defined as the indicator functions of the events that a customer leaving queue j at time t is routed to queue i as a positive or negative customer, respectively. The process $U_i(t)$ denotes the cumulative number of customers not allowed to enter the queue due to the buffer being full by time t. If the buffer size is infinite for queue i, the process $U_i(t)$ can be considered as the 'zero' process.

The interpretation of the counting processes in (1) is the following: $A_i(t)$ is the cumulative number of exogenous clients that arrived at queue i by time t, $D_i(t)$ is the number of service completions at queue i by time t, and $S_i(t)$ is the number of removed customers due to an exogenous signal by time t. The process $D_{ji}^+(t)$ denotes the total number of customers that left

https://doi.org/10.1239/jap/1214950352 Published online by Cambridge University Press

queue j and joined queue i as a regular customer by time t, and $D_{ji}^{-}(t)$ is the total number of customers removed from queue i due to a negative arrival that originated from queue j.

All stochastic processes given above are defined on the same probability space $(\Omega, \mathcal{F}, \mathrm{P})$. References to it are not necessary and will be omitted henceforth. Let $\mathcal{F}_t$ be the minimal σ-algebra that measures all driving processes defined above up to time t (i.e. $\{\mathcal{F}_t,\ t \geq 0\}$ is a filtration). In addition, the following assumption will be used. Amongst other things, it guarantees that the counting processes defined above are nonexplosive and have a martingale representation, which will be given below. The condition on the continuity and boundedness of the rates can be relaxed and that will be discussed in the next section.

Assumption 1. (a) *The random quantities $X_i(0)$ and N_i^{α}, $i \in \{1, \ldots, K\}$, $\alpha \in \{a, s, d\}$, are mutually independent.*

(b) *The functions $\Lambda_i^{\alpha}(\cdot)$, $i \in \{1, \ldots, K\}$, $\alpha \in \{a, s, d\}$, given in (2), are continuous and bounded.*

(c) $\mathrm{E}[\mathbf{1}_{ij}^{\alpha}(t) \mid \mathcal{F}_t^r] = Q_{ij}^{\alpha}(X(t-))$ *for $i, j \in \{1, \ldots, K\}$ and $\alpha \in \{+, -\}$, where $Q_{ij}^{\alpha} \colon \mathbb{R}_+^K \to [0, 1]$ is a measurable function and $\mathcal{F}_t^r$ is the minimal σ-algebra that measures all driving processes up to time t, not including the current routeing decision.*

Owing to Assumption 1, the jump processes A_i, D_i, S_i, and $\tilde{D}_{ij}$ have the following martingale decompositions (see [12, Theorem 8] and [49, p. 625]):

$$
\begin{aligned}
A_i(t) &= M_i^a(t) + \int_0^t \Lambda_i^a(X(s))\,\mathrm{d}s,\\
D_i(t) &= M_i^d(t) + \int_0^t \mathbf{1}_{\{X_i(s)>0\}}\,\Lambda_i^d(X(s))\,\mathrm{d}s,\\
S_i(t) &= M_i^s(t) + \int_0^t \mathbf{1}_{\{X_i(s)>0\}}\,\Lambda_i^s(X(s))\,\mathrm{d}s,\\
\tilde{D}_{ij}(t) &= \tilde{M}_{ij}^d(t) + \int_0^t \mathbf{1}_{\{X_i(s)>0,\,X_j(s)>0\}}\,\Lambda_i^d(X(s))\,\mathrm{d}s,
\end{aligned}
$$

where M_i^a, M_i^s, M_i^d, and $\tilde{M}_{ij}^d$ are $\mathcal{F}_t$-martingales. In order to have a martingale decomposition for D_{ij}^{+} and D_{ij}^{-}, define

$$
\begin{aligned}
M_{ij}^{+}(t) &:= \int_0^t (\mathbf{1}_{ij}^{+}(s) - Q_{ij}^{+}(X(s-)))\,\mathrm{d}D_i(s),\\
M_{ij}^{-}(t) &:= \int_0^t (\mathbf{1}_{ij}^{-}(s) - Q_{ij}^{-}(X(s-)))\,\mathrm{d}\tilde{D}_{ij}(s).
\end{aligned}
$$

The same argument used in [49, p. 626] can be used to show that M_{ij}^{+} and M_{ij}^{-} are $\mathcal{F}_t$-martingales. Now it is possible to write

$$
\begin{aligned}
D_{ij}^{+}(t) &= \int_0^t (\mathbf{1}_{ij}^{+}(s) - Q_{ij}^{+}(X(s-)))\,\mathrm{d}D_i(s) + \int_0^t Q_{ij}^{+}(X(s-))\,\mathrm{d}D_i(s)\\
&= M_{ij}^{+}(t) + \int_0^t Q_{ij}^{+}(X(s-))\,\mathrm{d}M_i^d(s) + \int_0^t Q_{ij}^{+}(X(s))\,\mathbf{1}_{\{X_i(s)>0\}}\,\Lambda_i^d(X(s))\,\mathrm{d}s,\\
D_{ij}^{-}(t) &= M_{ij}^{-}(t) + \int_0^t Q_{ij}^{-}(X(s-))\,\mathrm{d}\tilde{M}_{ij}^d(s) + \int_0^t Q_{ij}^{-}(X(s))\,\mathbf{1}_{\{X_i(s)>0,\,X_j(s)>0\}}\,\Lambda_i^d(X(s))\,\mathrm{d}s.
\end{aligned}
$$

https://doi.org/10.1239/jap/1214950352 Published online by Cambridge University Press

Hence, the process $X = (X_i, i = 1, \ldots, K)^\top$ accepts the following representation:

$$X(t) = X(0) + \int_0^t B(X(s))\,\mathrm{d}s + M(t) - U(t),$$

where

$$\begin{aligned} B_i(x) &:= \Lambda_i^a(x) - \Lambda_i^d(x)\,\mathbf{1}_{\{x_i>0\}} - \Lambda_i^s(x)\,\mathbf{1}_{\{x_i>0\}} \\ &\quad + \sum_{j\le K} (Q_{ji}^+(x)\,\mathbf{1}_{\{x_j>0\}} - Q_{ji}^-(x)\,\mathbf{1}_{\{x_j>0,\,x_i>0\}})\Lambda_j^d(x), \\ M_i(t) &:= M_i^a(t) - M_i^d(t) - M_i^s(t) \\ &\quad + \sum_{j\le K}\left(M_{ji}^+(t) - M_{ji}^-(t) + \int_0^t Q_{ji}^+(X(s-))\,\mathrm{d}M_j^d(s) - \int_0^t Q_{ji}^-(X(s-))\,\mathrm{d}\tilde{M}_{ji}^d(s)\right), \end{aligned}$$

and M_i is an $\mathcal{F}_t$-martingale.

3. Heavy-traffic limit

As it is usual in heavy-traffic analysis, we consider a sequence of queueing networks $(X^n,\ n > 0)$ indexed by the parameter n. As n increases, the system approaches heavy traffic in the sense that the rate of customers entering the system approaches that of customers leaving the system. The following scale is usually employed:

$$x^n(t) := \frac{X^n(nt)}{\sqrt{n}}.$$

Let any mathematical object defined in the previous section with respect to X^n now be indexed with an upper script n (e.g. $\mathcal{F}_t^n$, $\Lambda_i^{a,n}$, $Q_{ij}^{+,n}$, etc.). Similarly, any counting process defined in the previous section (e.g. A_i, S_i, D_i, D_{ij}^+, D_{ij}^-, and U_i) is now replaced by its scaled equivalent (e.g. A_i^n, S_i^n, D_i^n, $D_{ij}^{+,n}$, $D_{ij}^{-,n}$, and U_i^n). For example, $A_i^n(t)$ now denotes $1/\sqrt{n}$ times the number of exogenous customers that arrived at queue i by time nt, that is,

$$A_i^n(t) := \frac{1}{\sqrt{n}} N_i^a\left(\int_0^{nt} \lambda_i^{a,n}\left(x^n\left(\frac{s}{n}\right)\right)\mathrm{d}s\right)$$

with $\lambda_i^{a,n}(\xi) := \Lambda_i^{a,n}(\sqrt{n}\xi)$, $\xi \in \mathbb{R}^K$, for each $n > 0$, and the martingale decomposition becomes

$$\begin{aligned} A_i^n(t) &= M_i^{a,n}(t) + \frac{1}{\sqrt{n}}\int_0^{nt} \lambda_i^{a,n}\left(x^n\left(\frac{s}{n}\right)\right)\mathrm{d}s \\ &= M_i^{a,n}(t) + \sqrt{n}\int_0^t \lambda_i^{a,n}(x^n(s))\,\mathrm{d}s, \end{aligned}$$

after a change of variable. Likewise, let

$$q_{ij}^{\alpha,n}(\xi) := Q_{ij}^{\alpha,n}(\sqrt{n}\xi)$$

https://doi.org/10.1239/jap/1214950352 Published online by Cambridge University Press

for $\alpha \in \{+, -\}$, $i, j \in \{1, \ldots, K\}$, and $\xi \in \mathbb{R}^K$. Also, let the size of the buffer for the scaled process be B_i for the ith queue (B_i may be infinite).

In Assumption 2(a), below, we will define how state dependence is introduced. Even though the dependence is very small for large n, it has a significant effect in the limit. As it will be seen, the functions $f_i^\alpha(x)$ and $f_{ij}^\beta(x)$ will appear in the drift term of the limit equation.

Assumption 2. *For $o(\cdot)$ uniformly in x, we make the following assumptions.*

(a) *There exist nonnegative constants r_i^α and r_{ij}^β, and bounded and continuous functions $f_i^\alpha(x)$ and $f_{ij}^\beta(x)$ for $j, i \in \{1, \ldots, K\}$, $\alpha \in \{a, d, s\}$, and $\beta \in \{+, -\}$ such that*

$$\lambda_i^{\alpha,n}(x) = r_i^\alpha + \frac{f_i^\alpha(x)}{\sqrt{n}} + o_\alpha^n\left(\frac{1}{\sqrt{n}}\right),$$
$$q_{ij}^{\beta,n}(x) = r_{ij}^\beta + \frac{f_{ij}^\beta(x)}{\sqrt{n}} + o_q^{n,\beta}\left(\frac{1}{\sqrt{n}}\right).$$

(b) *For any $i \in \{1, \ldots, K\}$,*

$$r_i^a + \sum_{j \le K} r_j^d r_{ji}^+ = r_i^d + r_i^s + \sum_{j \le K} r_j^d r_{ji}^-,$$

which is usually called the heavy-traffic condition. This condition tells us that, for large n, the rate of customers joining a queue in the network is very close to the rate of customers leaving this queue.

Note that Assumption 2 tells us that, for each $x \in \mathbb{R}_+^K$,

$$\sqrt{n}\left(\lambda_i^{a,n}(x) - \lambda_i^{d,n}(x) - \lambda_i^{s,n}(x) + \sum_{j \le K}(q_{ji}^+(x) - q_{ji}^-(x))\lambda_j^{d,n}(x)\right) =: b_i^n(x) \to b_i(x),$$

where $b_i(\cdot)$ is defined in Theorem 1, below.

Assumption 3, below, is on the *reflection directions* and it will become more clear throughout the development of the proof of Theorem 1. Define the following matrices in $\mathbb{R}^{K \times K}$:

$$\begin{aligned} I_r &:= \operatorname{diag}(r_i^d + r_i^s)_{i=1,\ldots,K}, \\ \Theta_{ij} &:= r_i^d(r_{ij}^+ - r_{ij}^-), \\ \Omega^S &:= \operatorname{diag}\left(\sum_{j \in Z \setminus S \cup \{i\}} r_{ji}^- r_j^d\right)_{i=1,\ldots,K}, \\ R^S &:= I_r - \Theta^\top + \Omega^S, \qquad R := R^\varnothing, \end{aligned} \tag{3}$$

where $S \subseteq Z := \{1, \ldots, K\}$ and

$$\operatorname{diag}(a_i)_{i=1,\ldots,m} \in \mathbb{R}^{m \times m}$$

is a diagonal matrix with entries a_i.

https://doi.org/10.1239/jap/1214950352 Published online by Cambridge University Press

Assumption 3. (a) *For any* $S \subseteq \{1, \ldots, K\}$ *with* $|S| \geq 2$, *there exists an* $\alpha = (\alpha_1, \ldots, \alpha_{|S|})^\top \neq 0$, *where* $|S|$ *denotes the number of elements of* S, *such that* $\alpha_i \geq 0$ *and*

$$R_{\{i \in S\}}\alpha = R^S_{\{i \in S\}}e \quad \text{with } e = (1, \ldots, 1)^\top.$$

The subscript $\{i \in S\}$ *on any matrix* A *indicates that* $A_{\{i \in S\}} \in \mathbb{R}^{K \times |S|}$ *is formed by the columns of* A *with indices in* S.

(b) *The matrix* R *satisfies the completely-*$\mathcal{S}$ *condition (see [43, p. 121]).*

The usual interpretation for the matrix R is that its columns, denoted by d_i, $i \in \{1, \ldots, K\}$, are reflection directions. That is, whenever the process $x^n(\cdot)$ tries to cross a boundary $\partial G_i := \{\xi \in \mathbb{R}^K_+ \mid \xi_i = 0\}$, it is 'pushed' back into the state space in the direction of d_i. When $x^n(t) \in \partial G_i \cap \partial G_j$, $i \neq j$, the usual assumption is that the reflection direction at this instant is a positive linear combination of the directions d_i and d_j, and similarly at the intersection of more than two boundaries. This allows us to write the reflection term $z(\cdot)$, that appears in Theorem 1, below, in the usual form $z(t) = \sum_i d_i y_i(t) = Ry(t)$ (omitting the u term). This condition that the reflection directions at 'corners' or 'edges' of the state space are positive linear combinations of the directions at the adjacent faces appears naturally in most queueing systems [43]. However, this is not the case for G-networks. The actual reflection directions that appear at the corners or edges of the state space are given by positive linear combinations of the columns of $R^{S(x^n(t))}$, defined in (3), where $S(\xi) := \{i \in \{1, \ldots, K\} \mid \xi_i = 0\}$ for $\xi \in \mathbb{R}^K$, and that is the reason for introducing Assumption 3(a). Although this condition seems, *prima facie*, restrictive, we will show in Section 4 two important queueing networks that satisfy this assumption.

We are now able to present the heavy-traffic limit in the theorem below.

Theorem 1. *Let* $x^n(0)$ *converge weakly to* $x(0)$. *With Assumptions 1, 2, and 3,* $\{x^n(\cdot)\}$ *is tight and any weakly convergent subsequence satisfies*

$$x(t) = x(0) + \int_0^t b(x(s))\,\mathrm{d}s + M(t) + z(t), \tag{4}$$
$$z(t) = Ry(t) - u(t),$$

where $0 \leq x_i(t) \leq B_i$, $i \in \{1, \ldots, K\}$, *and*

$$b_i(x) = f^a_i(x) - f^d_i(x) - f^s_i(x) + \sum_{j=1}^K (f^d_j(x)(r^+_{ji} - r^-_{ji}) + r^d_j(f^+_{ji}(x) - f^-_{ji}(x))), \tag{5}$$

$$M_i(t) = M^a_i(t) - M^d_i(t) - M^s_i(t) + \sum_{j=1}^K ((r^+_{ji} - r^-_{ji})M^d_j(t) + M^+_{ji}(t) - M^-_{ji}(t)).$$

The M^α_i, $\alpha \in \{a, d, s, r\}$, $i \in \{1, \ldots, K\}$, *are mutually independent Wiener processes, where* $M^r_i(t) := (M^+_{i1}(t), \ldots, M^+_{iK}(t), M^-_{i1}(t), \ldots, M^-_{iK}(t))^\top$. *The* M^α_i *have variances* r^α_i *for* $\alpha \in \{a, d, s\}$ *and* M^r_i *has covariance matrix*

$$(\Sigma_i)_{jk} = r^d_i \begin{pmatrix} (\Sigma^+_i) & (\Sigma^{+-}_i) \\ (\Sigma^{+-}_i)^\top & (\Sigma^-_i) \end{pmatrix} \quad \text{with} \quad (\Sigma^\alpha_i)_{jk} = \begin{cases} (1 - r^\alpha_{ij})r^\alpha_{ij} & \text{if } j = k, \\ -r^\alpha_{ij}r^\alpha_{ik} & \text{otherwise,} \end{cases}$$

for $\alpha \in \{+, -\}$ *and* $(\Sigma^{+-}_i)_{jk} = -r^+_{ij}r^-_{ik}$, *where* Σ^{+-}_i, $\Sigma^\alpha_i \in \mathbb{R}^{K \times K}$.

https://doi.org/10.1239/jap/1214950352 Published online by Cambridge University Press

The process $z(\cdot)$ is the reflection term ($y_i(0) = 0$ and $y_i(\cdot)$ are continuous, nondecreasing, and can increase only at t where $x_i(t) = 0$; similarly, if $B_i < \infty$, $u_i(0) = 0$ and $u_i(\cdot)$ are continuous, nondecreasing, and increase only when $x_i(t) = B_i$).

Proof. The proof follows the ideas of Theorem 8.2.1 of [43]. By the discussion in the last section, we know that $x^n(t)$ has the following representation:

$$x_i^n(t) = x_i(0) + B_i^n(x(t)) + M_i^n(t) - U^n(t),$$

where

$$\begin{aligned} B_i^n(x^n(t)) &= \sqrt{n}\int_0^t \lambda_i^{a,n}(x^n(s)) - \lambda_i^{d,n}(x^n(s))\,\mathbf{1}_{\{x_i^n(s)>0\}} - \lambda_i^{s,n}(x^n(s))\,\mathbf{1}_{\{x_i^n(s)>0\}} \\ &\qquad + \sum_{i\le K}(q_{ji}^{+,n}(x^n(s)) - q_{ji}^{-,n}(x^n(s))\,\mathbf{1}_{\{x_i^n(s)>0\}})\,\mathbf{1}_{\{x_j^n(s)>0\}}\,\lambda_j^{d,n}(x^n(s))\,\mathrm{d}s, \\ M_i^n(t) &= M_i^{a,n}(t) - M_i^{d,n}(t) - M_i^{s,n}(t) \\ &\qquad + \sum_{j\le K}\Bigg(M_{ji}^{+,n}(t) - M_{ji}^{-,n}(t) \\ &\qquad\qquad + \int_0^t q_{ji}^{+,n}(x^n(s-))\,\mathrm{d}M_j^{d,n}(s) - \int_0^t q_{ji}^{-,n}(x^n(s-))\,\mathrm{d}\tilde{M}_{ji}^{d,n}(s)\Bigg), \end{aligned}$$

and $M_i^n(t)$ is a martingale. Define $y_i^n(t) = \sqrt{n}\int_0^t \mathbf{1}_{\{x_i^n(s)=0\}}\,\mathrm{d}s$, which is the total server idle time by time nt for queue i. Similarly, define $y_{ij}^n(t) = y_{ji}^n(t) = \sqrt{n}\int_0^t \mathbf{1}_{\{x_i^n(s)=0,\,x_j^n(s)=0\}}\,\mathrm{d}s$. Using Assumption 2(a) and the heavy-traffic condition (i.e. Assumption 2(b)), we can expand B_i^n as

$$\begin{aligned} B_i^n(x^n(t)) &= \int_0^t b_i(x^n(s))\,\mathrm{d}s + \Bigg(r_i^d + r_i^s - (r_{ii}^+ - r_{ii}^-)r_i^d + \sum_{j\le K(j\ne i)} r_{ji}^- r_j^d\Bigg)y_i^n(t) \\ &\quad - \sum_{j\le K(j\ne i)}(r_{ji}^+ - r_{ji}^-)r_j^d\,y_j^n(t) - \sum_{j\le K(j\ne i)} r_{ji}^- r_j^d\,y_{ji}^n(t) + \sum_{j\le K} O\Bigg(\frac{y_j^n(t)}{\sqrt{n}}\Bigg) \\ &\quad + \sqrt{n}\,o\Bigg(\frac{1}{\sqrt{n}}\Bigg). \end{aligned} \tag{6}$$

The martingales $M_i^{\alpha,n}$, $\alpha \in \{a, d, s\}$, and $\tilde{M}_{ij}^{d,n}$ have the following associated Doob–Meyer processes:

$$\begin{aligned} \langle M_i^{a,n}\rangle(t) &= \int_0^t \lambda_i^{a,n}(x^n(s))\,\mathrm{d}s, \\ \langle M_i^{d,n}\rangle(t) &= \int_0^t \mathbf{1}_{\{x_i^n(s)>0\}}\,\lambda_i^{d,n}(x^n(s))\,\mathrm{d}s, \\ \langle M_i^{s,n}\rangle(t) &= \int_0^t \mathbf{1}_{\{x_i^n(s)>0\}}\,\lambda_i^{s,n}(x^n(s))\,\mathrm{d}s, \\ \langle \tilde{M}_{ij}^{d,n}\rangle(t) &= \int_0^t \mathbf{1}_{\{x_i^n(s)>0,\,x_j^n(s)>0\}}\,\lambda_i^{d,n}(x^n(s))\,\mathrm{d}s. \end{aligned}$$

https://doi.org/10.1239/jap/1214950352 Published online by Cambridge University Press

These processes can be obtained using the result in [43, p. 62]. Also, the martingale $M_i^{r,n} = (M_{i1}^{+,n}, \dots, M_{iK}^{+,n}, M_{i1}^{-,n}, \dots, M_{iK}^{-,n})^\top$ has the following associated Doob–Meyer process:

$$\langle M_{ij}^{+,n}\rangle(t) = \int_0^t \mathbf{1}_{\{x_i^n(s)>0\}}(1 - q_{ij}^{+,n}(x^n(s)))q_{ij}^{+,n}(x^n(s))\lambda_i^{d,n}(x^n(s))\,\mathrm{d}s,$$

$$\langle M_{ij}^{+,n}, M_{ik}^{+,n}\rangle(t) = -\int_0^t \mathbf{1}_{\{x_i^n(s)>0\}}\, q_{ij}^{+,n}(x^n(s))q_{ik}^{+,n}(x^n(s))\lambda_i^{d,n}(x^n(s))\,\mathrm{d}s,$$

$$\langle M_{ij}^{-,n}\rangle(t) = \int_0^t \mathbf{1}_{\{x_i^n(s)>0,\, x_j^n(s)>0\}}(1 - q_{ij}^{-,n}(x^n(s)))q_{ij}^{-,n}(x^n(s))\lambda_i^{d,n}(x^n(s))\,\mathrm{d}s,$$

$$\langle M_{ij}^{-,n}, M_{ik}^{-,n}\rangle(t) = -\int_0^t \mathbf{1}_{\{x_i^n(s)>0,\, x_j^n(s)>0,\, x_k^n(s)>0\}}\, q_{ij}^{-,n}(x^n(s))q_{ik}^{-,n}(x^n(s))\lambda_i^{d,n}(x^n(s))\,\mathrm{d}s,$$

$$\langle M_{ij}^{+,n}, M_{ik}^{-,n}\rangle(t) = -\int_0^t \mathbf{1}_{\{x_i^n(s)>0,\, x_k^n(s)>0\}}\, q_{ij}^{+,n}(x^n(s))q_{ik}^{-,n}(x^n(s))\lambda_i^{d,n}(x^n(s))\,\mathrm{d}s,$$

and analogously for $\langle M_{ij}^{-,n}, M_{ik}^{+,n}\rangle$. The proof of this characterization is omitted since it is straightforward and cumbersome. Also, using Assumption 1(a) and (c), we have

$$\langle M_i^{\alpha,n}, M_i^{\gamma,n}\rangle(t) = 0 \quad \text{and} \quad \langle M_i^{\alpha,n}, M_{ij}^{\beta,n}\rangle(t) = 0,$$

where $\alpha, \gamma \in \{a, d, s\}$, $\alpha \neq \gamma$, and $\beta \in \{+, -\}$. By Theorem 2.8.3 of [43], the martingales are tight, and since each term has discontinuities of the order of $1/\sqrt{n}$, they are asymptotically continuous. Therefore, using the expression in Assumption 2(a), we can write the following:

$$\int_0^t q_{ji}^{+,n}(x^n(s-))\,\mathrm{d}M_j^{d,n}(t) = r_{ji}^+ M_j^{d,n}(t),$$

$$\int_0^t q_{ji}^{-,n}(x^n(s-))\,\mathrm{d}\tilde{M}_{ji}^{d,n}(t) = r_{ji}^- \tilde{M}_{ji}^{d,n}(t),$$

modulo a negligible error that goes to zero as $n \to \infty$.

Define the column vector $y^n(\cdot) := (y_i^n(\cdot),\ i = 1, \dots, K)^\top$. Observe that we can write the terms in (6) with respect to the idle time (i.e. $y_i^n(t)$ and $y_{ij}^n(t)$) in matrix notation as

$$\int_0^t R^{S(x^n(s))}\,\mathrm{d}y^n(s) = \int_0^t R^{S(x^n(s))} I^n(s; e)\,\mathrm{d}s,$$

where we define the column vector $I^n(s;\alpha) := (I_i^n(s;\alpha),\ i = 1, \dots, K)^\top$ for $\alpha \in \mathbb{R}^K$ as

$$I_i^n(s;\alpha) := \alpha_i\sqrt{n}\,\mathbf{1}_{\{x_i^n(s)=0\}},$$

$e = (1, \dots, 1)^\top \in \mathbb{R}^K$, $S(\xi) := \{i \in \{1, \dots, K\} \mid \xi_i = 0\}$ for $\xi \in \mathbb{R}^K$, and R^S is defined in (3). Using Assumption 3(a), we find that, for each $t \geq 0$, there exists $\alpha(t)$ such that

$$R^{S(x^n(t))} I^n(t; e) = R I^n(t; \alpha(t)).$$

Hence,

$$\int_0^t R^{S(x^n(s))}\,\mathrm{d}y^n(s) = R\int_0^t I^n(s;\alpha(s))\,\mathrm{d}s =: R\bar{y}^n(t).$$

Using Assumption 3(b), we can now apply Theorem 3.6.1 of [43] to show that $\{z^n(\cdot), x^n(\cdot)\}$, where $z^n(t) := R\bar{y}^n(t) - U^n(t)$, is tight and $y^n(\cdot)$ and $U^n(\cdot)$ are asymptotically continuous.

https://doi.org/10.1239/jap/1214950352 Published online by Cambridge University Press

Let $z(\cdot) = Ry(\cdot) - u(\cdot)$ denote any weak-sense limit of $z^n(\cdot)$. The tightness of $\{y_i^n(\cdot)\}$ implies that $y_i^n(\cdot)/\sqrt{n} = \int_0^{\cdot} \mathbf{1}_{\{x_i^n(s)=0\}}\, \mathrm{d}s$ converges weakly to the zero process. Therefore, the indicator functions in the Doob–Meyer processes can be dropped without a change in the limit. Also, the processes $\tilde{M}_{ij}^{d,n}$ and $M_i^{d,n}$ converge weakly to the same limit process M_i^d for any i and j, and $M_i^{d,n}$ can be used in place of $\tilde{M}_{ij}^{d,n}$ without affecting the limit.

Now, all that needs to be done is to apply Theorem 2.8.2 of [43] with the fact that the size of the discontinuities are of order $1/\sqrt{n}$ to show that the martingales converge to the asserted Wiener processes.

Remark 1. As in Theorem 8.4.1 of [43], the continuity of the functions $f_i^{\alpha}(\cdot)$ and $f_{ij}^{\beta}(\cdot)$ in Assumption 2(a) can be replaced by measurability as long as the functions $\phi(\cdot) \mapsto \int_0^{\cdot} f_i^{\alpha}(\phi(s))\, \mathrm{d}s$ and $\phi(\cdot) \mapsto \int_0^{\cdot} f_{ij}^{\beta}(\phi(s))\, \mathrm{d}s$ are continuous on $D(\mathbb{R}; 0, \infty)$, with probability 1, with respect to the measure induced by any weak-sense limit $x(\cdot)$. This is verified with the application of Theorem 5.1 of [8].

Remark 2. Theorem 3.5.4 of [43] tells us that there exists a weak-sense solution to (4) under the assumptions of Theorem 1. If we add Assumption (A.3.5.2) of [43] (which is satisfied if R is an M-matrix (see [16, p. 164])) to Assumption 3(b) and suppose that $b(\cdot)$, defined in (5), is Lipschitz continuous, then, by Theorem 3.5.2 of [43], there exists a unique strong-sense solution to (4).

Remark 3. The assumption on the boundedness of the functions $f_i^{\alpha}(\cdot)$ and $f_{ij}^{\beta}(\cdot)$ can be replaced by boundedness when restricted to G. Alternatively, this condition on the functions $f_i^{\alpha}(\cdot)$ and $f_{ij}^{\beta}(\cdot)$ can be replaced by supposing that they have at most linear growth in x, and adding Assumption (A.3.5.2) of [43] to Assumption 3(b). As in Theorem 8.2.1 of [43], this is verified by the truncation technique together with the fact that any solution $x(\cdot)$ of (4) satisfies

$$\lim_{K\to\infty} \mathrm{P}\left(\sup_{s\le t} |x(s)| \ge K\right) = 0,$$

which can be verified with the aid of Theorem 3.5.1 of [43] and the fact that the drift term $b(\cdot)$ will have at most linear growth.

4. Networks satisfying Assumption 3(a)

Assumptions 1, 2, and 3(b) are usual assumptions in heavy-traffic approximations for state-dependent queueing systems [43, Chapter 8]. Loosely speaking, that is also true for Assumption 3(a), which essentially requires that any reflection direction appearing on the edge or corner of the state space be a positive linear combination of the reflections at the adjacent boundaries. In fact, if we consider a network where queues can receive signals from outside but not from within the network, the condition is automatically satisfied. However, it is not valid for any G-network. In this section we show two types of network topologies that satisfy Assumption 3(a).

4.1. Two queues in tandem

Consider two queues in tandem where each queue may send regular or negative customers to each other, as seen in Figure 1. Both queues receive customers and signals from exogenous sources, and there is no feedback, in the sense that a customer that has just left queue i may not be routed immediately to queue i.

https://doi.org/10.1239/jap/1214950352 Published online by Cambridge University Press

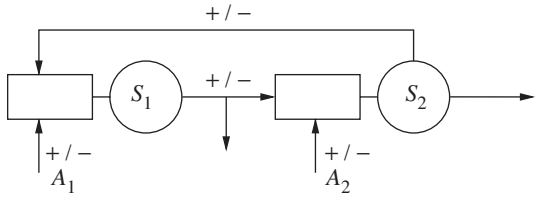

FIGURE 1: Two queues in tandem. The symbols '+' and '−' indicate the arrival of regular or negative customers, respectively.

For this case, the matrices I_r, Θ, Ω^S, and R are defined as follows:

$$I_r = \begin{pmatrix} r_1^d + r_1^s & 0 \\ 0 & r_2^d + r_2^s \end{pmatrix},$$

$$\Omega^{\varnothing} = \begin{pmatrix} r_{21}^- r_2^d & 0 \\ 0 & r_{12}^- r_1^d \end{pmatrix}, \qquad \Omega^{\{1,2\}} = 0,$$

$$\Theta = \begin{pmatrix} 0 & r_1^d (r_{12}^+ - r_{12}^-) \\ r_2^d (r_{21}^+ - r_{21}^-) & 0 \end{pmatrix},$$

$$R = \begin{pmatrix} r_1^d + r_1^s + r_{21}^- r_2^d & -r_2^d (r_{21}^+ - r_{21}^-) \\ -r_1^d (r_{12}^+ - r_{12}^-) & r_2^d + r_2^s + r_{12}^- r_1^d \end{pmatrix},$$

$$R^{\{1,2\}} = \begin{pmatrix} r_1^d + r_1^s & -r_2^d (r_{21}^+ - r_{21}^-) \\ -r_1^d (r_{12}^+ - r_{12}^-) & r_2^d + r_2^s \end{pmatrix}.$$

Hence, the condition is verified if there is an $\alpha = (\alpha_1, \alpha_2)^\top$ with positive components such that $R\alpha = R^{\{1,2\}} e$, which is true as long as $r_1^d, r_2^d > 0$.

4.2. Two-layer feedforward network

Let us now consider a feedforward network with two layers, in the sense that the queues on the first layer may send customers to queues in the second layer, but not vice versa; see Figure 2 for reference. Each queue can also receive regular and negative exogenous arrivals and there is no feedback. Suppose that there are K_1 queues on the first layer and K_2 queues on the second layer. Define $K = K_1 + K_2$. We index the queues starting on the first layer in such a way that if $K_1 < i \le K$, the ith queue is in the second layer.

Define $Z_1 = \{1, \dots, K_1\}$. For this scenario, the matrix R^S is given by

$$R^S = \begin{pmatrix} \operatorname{diag}(r_i^d + r_i^s)_{i=1,\dots,K_1} & 0 \\ -\tilde{\Theta}^\top & \operatorname{diag}(r_i^d + r_i^s + \sum_{j \in Z_1 \setminus S} r_{ji}^- r_j^d)_{i=K_1+1,\dots,K} \end{pmatrix},$$

where $\tilde{\Theta} \in \mathbb{R}^{K_1 \times K_2}$ is defined as

$$\tilde{\Theta} = \begin{pmatrix} r_1 (r_{1(K_1+1)}^+ - r_{1(K_1+1)}^-) & \cdots & r_1 (r_{1K}^+ - r_{1K}^-) \\ \vdots & & \vdots \\ r_{K_1} (r_{K_1(K_1+1)}^+ - r_{K_1(K_1+1)}^-) & \cdots & r_{K_1} (r_{K_1K}^+ - r_{K_1K}^-) \end{pmatrix}.$$

https://doi.org/10.1239/jap/1214950352 Published online by Cambridge University Press

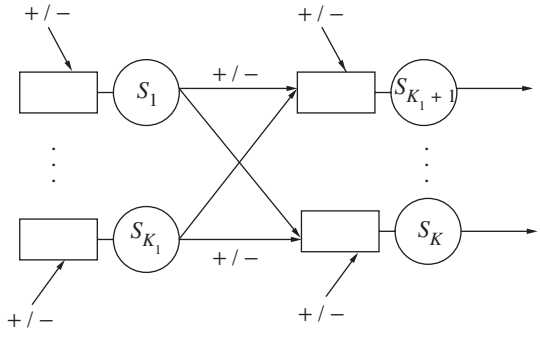

FIGURE 2: Two-layer feedforward network.

In order to verify Assumption 3(a), let $S \subseteq \{1, \ldots, K\}$ be an ordered set with $|S| \geq 2$. Define $S_1 \subseteq \{1, \ldots, K_1\}$ and $S_2 \subseteq \{K_1 + 1, \ldots, K\}$ as ordered sets such that $S_1 \cup S_2 = S$. Then choose $\alpha = (\alpha_1, \ldots, \alpha_{|S_1|}, \beta_1, \ldots, \beta_{|S_2|})^\top$ such that $\alpha_i = 1$ for $i = 1, \ldots, |S_1|$ and

$$\beta_j = \frac{r_k^d + r_k^s + \sum_{l \in Z_1 \setminus S_1} r_{lk}^- r_l^d}{r_k^d + r_k^s + \sum_{l \in Z_1} r_{lk}^- r_l^d} \quad \text{for } j = 1, \ldots, |S_2|,$$

where k is the jth element of S_2. Now we can verify that $R_{\{i \in S\}}\alpha = R^S_{\{i \in S\}}e$. Since this works for any choice of S, Assumption 3(a) is satisfied.

5. Numerical experiments

In order to illustrate an application of Theorem 1, let us suppose that we have the system of Subsection 4.1, given by Figure 1. It will be assumed that every customer leaving queue 1 joins queue 2 as a regular customer, and queue 2 does not receive exogenous clients. Also, both queues have finite buffers. Suppose that queue 2 needs to reduce customer loss due to buffer overflow and it does that by sending signals to queue 1. Hence, every time queue 2 has its buffer almost full, it will start sending signals to the first queue. In this example, it will be shown how we can use the result derived here to choose the *optimal routeing* strategy for a *system operating under heavy traffic*.

As it is common in application (see, e.g. [43], [45], and [46]), we do not have a sequence of queues indexed by the parameter n. Rather, we have one queueing system that we want to approximate. Hence, we need to choose a large N such that the rates for our problem satisfy

$$\Lambda_i^\alpha(\sqrt{N}x) = \lambda_i^{\alpha,N}(x) \approx r_i^\alpha + \frac{f_i^\alpha(x)}{\sqrt{N}}, \qquad \alpha = a, d, s,$$

$$Q_{ij}^\beta(\sqrt{N}x) = q_{ij}^{\beta,N}(x) \approx r_{ij}^\beta + \frac{f_{ij}^\beta(x)}{\sqrt{N}}, \qquad \beta = +, -,$$

and the heavy-traffic condition holds (i.e. Assumption 2(b)). Now, we can approximate the distribution of the number of customers in each queue at time Nt using the distribution $\sqrt{N}x(t)$, where $x(\cdot)$ is the limit process given by (4).

For our example, let us suppose that

$$\lambda_1^{a,N}(x) = \lambda, \qquad \lambda_2^{a,N}(x) = 0, \qquad \lambda_1^{s,N}(x) = 0, \qquad \lambda_2^{s,N}(x) = 0,$$

https://doi.org/10.1239/jap/1214950352 Published online by Cambridge University Press

and that $\lambda_1^{d,N}(x) = \mu_1$ and $\lambda_2^{d,N}(x) = \mu_2$, where μ_1, μ_2, and λ are positive constants. By the heavy-traffic assumption, there exist ('small') constants b_1 and b_2 such that $b_1 = \sqrt{N}(\mu_1 - \lambda)$ and $b_2 = \sqrt{N}(\mu_2 - \mu_1)$. That is, the rate of customers entering each queue is close to the rate of departing customers. Hence, $\lambda_1^{d,N}(x) = \lambda + b_1/\sqrt{N}$ and $\lambda_2^{d,N}(x) = \lambda + (b_1 + b_2)/\sqrt{N}$. Also, we suppose that

$$q_{12}^{+,N}(x) = 1, \qquad q_{21}^{+,N}(x) = 0, \qquad q_{12}^{-,N}(x) = 0, \qquad q_{21}^{-,N}(x) = \frac{g(x)}{\sqrt{N}},$$

where $g\colon \mathbb{R}^2 \to [0, 1]$, and that the size of the (unscaled) buffers are $\sqrt{N}B_1$ for the first queue and $\sqrt{N}B_2$ for the second queue. The heavy-traffic limit is given by

$$\mathrm{d}x(t) = \begin{pmatrix} -b_1 - \lambda g(x(t)) \\ -b_2 \end{pmatrix} \mathrm{d}t + \begin{pmatrix} 2\lambda & -\lambda \\ -\lambda & 2\lambda \end{pmatrix}^{1/2} \mathrm{d}W(t) + \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \mathrm{d}y(t) - \mathrm{d}u(t),$$

where $A^{1/2}(A^{1/2})^\top = A$. It is perhaps noteworthy to mention that the function $g(\cdot)$ only acts upon the first component of $x(\cdot)$, even though we are interested in controlling the second. However, the second queue will be affected indirectly by the control through the reflection term.

Now we want to choose a $g(\cdot)$ which will reduce buffer overflow in queue 2. Let us suppose that $\lambda = 1$, $b_1 = b_2 = 0.1$, and $B_1 = B_2 = 25.6$, and define the step function

$$s(x) = \begin{cases} 1 & \text{if } x_2 > 15, \\ 0 & \text{otherwise.} \end{cases}$$

Figure 3 compares the same realization of $x(t)$ with $g(\cdot)$ set to $g(x) = s(x)$ and $g(x) = 0$ for initial condition $x(0) = (B_1, B_2)^\top$. Observe that the sample path with no control hits buffer overflow more frequently. The sample paths were constructed with the Euler method (e.g. [42, p. 110]). The time discretization parameter was set to $h = 0.01$. The reflection term was implemented by pushing the process back (in the direction of the reflection vector) into the state space every time it crossed a boundary.

We can also find the optimal choice of $g(\cdot)$ with respect to a cost function. For this example, we will use the following discounted cost:

$$W(x, g) = \mathrm{E}_x^g\left[\int_0^\infty \mathrm{e}^{-\beta t}(cg(x(t))\,\mathrm{d}t + v\,\mathrm{d}u_2(t))\right],$$

where c and v are constants associated with the cost of routeing negative customers (or losing customers at the first queue) and the cost of losing customers due to buffer overflow at queue 2, respectively.

We use the Markov chain approximation method [40], [44] to find the optimal control numerically. We set $\beta = 0.01$, and the discretization parameter is set to $h = 0.1$. Plots of the control for different choices of c and v are shown in Figure 4. Note that the optimal control is of the switching type (i.e. after a given threshold, the control is used at maximum rate). This type of optimal control has also been found in different situations for the control of queueing systems [46].

It is interesting to see the shape of these switching curves. Note that the curves move upwards at the right side of the state space. This can be explained by the delay of the control action, since the control at queue 2 is done indirectly. When queue 1 and queue 2 are almost full, there most likely will be buffer overflow loss at queue 2 even if it sends signals to queue 1.

https://doi.org/10.1239/jap/1214950352 Published online by Cambridge University Press

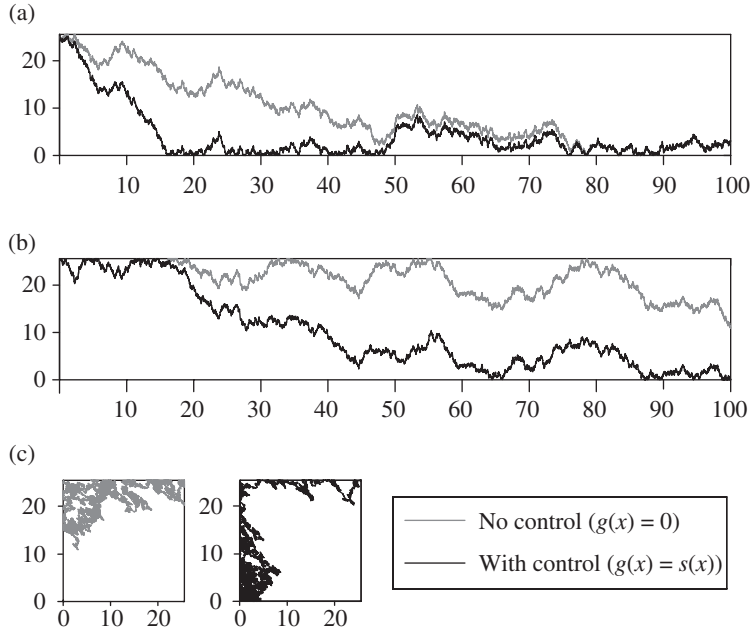

FIGURE 3: Sample path of $x(t)$ with $g(x) = s(x)$ (*black*) and $g(x) = 0$ (*gray*). (a) Plot of $x_1(t)$ versus time. (b) Plot of $x_2(t)$ versus time. (c) Plot of $x_1(t)$ versus $x_2(t)$.

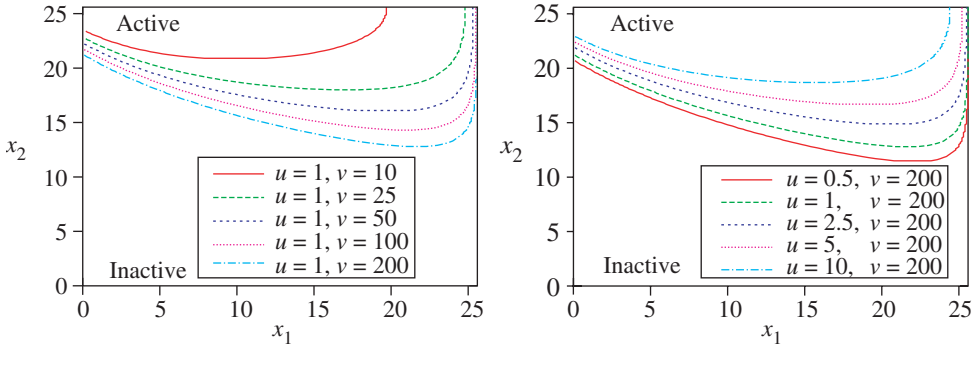

FIGURE 4: Switching curves for the optimal controls with varying values for c and v.

6. Conclusions

We have presented heavy-traffic limits for a class of state-dependent G-networks which satisfy Assumption 3(a). Two examples were given which satisfied this condition. Our current work concentrates on extending the results to any class of G-networks, and for networks with different kinds of signals.

In addition, we are interested in the extension of the ideas considered in [20] to the network case treated in this paper. The mentioned article raises a relevant problem with respect to

https://doi.org/10.1239/jap/1214950352 Published online by Cambridge University Press

diffusion approximations with reflecting boundaries: in the ergodic scenario, no probability mass concentrates at the state space boundaries. The addition of the boundary conditions in [20] could improve the approximations for cases when the traffic intensity is not very high and the distribution at the boundaries is important.

Acknowledgements

We would like to express our gratitude to the anonymous referee for the suggestions that contributed to the improvement of the paper. This research was partially supported by the Brazilian National Research Council-CNPq, under the grant numbers 140687/2005-0, 301740/2007-0, and 470527/2007-2, and by FAPERJ, grant number E-26/100.579/2007.

References

[1] Altman, E. and Kushner, H. J. (1999). Admission control for combined guaranteed performance and best effort communications systems under heavy traffic. *SIAM J. Control Optimization* **37,** 1780–1807.
[2] Altman, E. and Kushner, H. J. (2002). Control of polling in presence of vacations in heavy traffic with applications to satellite and mobile radio systems. *SIAM J. Control Optimization* **41,** 217–252.
[3] Arazi, A., Ben-Jacob, E. and Yechiali, U. (2004). Bridging genetic networks and queueing theory. *Physica A* **332,** 585–616.
[4] Arazi, A., Ben-Jacob, E. and Yechiali, U. (2005). Controlling an oscillating Jackson-type network having state-dependent service rates. *Math. Meth. Operat. Res.* **62,** 453–466.
[5] Artalejo, J. R. (2000). G-networks: a versatile approach for work removal in queueing networks. *Europ. J. Operat. Res.* **126,** 233–249.
[6] Atalay, V. and Gelenbe, E. (1992). Parallel algorithm for color texture generation using the random neural network model. *Internat. J. Pattern Recognition Artificial Intelligence* **6,** 437–446.
[7] Atalay, V., Gelenbe, E. and Yalabik, N. (1992). The random neural network model for texture generation. *Internat. J. Pattern Recognition Artificial Intelligence* **6,** 131–141.
[8] Billingsley, P. (1968). *Convergence of Probability Measures*. John Wiley, New York.
[9] Bocharov, P. P., Gavrilov, E. V. and Pechinkin, A. V. (2004). Exponential queuing network with dependent servicing, negative customers, and modification of the customer type. *Automation Remote Control* **65,** 35–59.
[10] Borovkov, A. (1964). Some limit theorems in the theory of mass service. I. *Theory Prob. Appl.* **9,** 550–565.
[11] Borovkov, A. (1965). Some limit theorems in the theory of mass service. II. *Theory Prob. Appl.* **10,** 375–400.
[12] Brémaud, P. (1981). *Point Processes and Queues, Martingale Dynamics*. Springer, New York.
[13] Buche, R. and Kushner, H. J. (2002). Control of mobile communications with time-varying channels in heavy traffic. *IEEE Trans. Automatic Control* **47,** 992–1003.
[14] Chakka, R. and Do, T. V. (2007). The $mm \sum_{k=1}^{k} cpp_k/ge/c/l$ G-queue with heterogeneous servers: steady state solution and an application to performance evaluation. *Performance Evaluation* **64,** 191–209.
[15] Chao, X. (1995). A queueing network model with catastrophes and product form solution. *Operat. Res. Lett.* **18,** 75–79.
[16] Chen, H. and Yao, D. D. (2001). *Fundamentals of Queueing Networks*. Springer, New York.
[17] Fourneau, J. and Verchère, D. (1995). G-networks with triggered batch state-dependent movement. In *MASCOTS '95: Proceedings of the 3rd International Workshop on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems* (Washington, DC), IEEE, New York, pp. 33–37.
[18] Fourneau, J., Gelenbe, E. and Suros, R. (1996). G-networks with multiple classes of negative and positive customers. *Theoret. Comput. Sci.* **155,** 141–156.
[19] Gaver, D. P. (1968). Diffusion approximations and models for certain congestion problems. *J. Appl. Prob.* **5,** 607–623.
[20] Gelenbe, E. (1975). On approximate computer system models. *J. Assoc. Comput. Mach.* **22,** 261–269.
[21] Gelenbe, E. (1989). Random neural networks with negative and positive signals and product form solution. *Neural Computation* **1,** 502–511.
[22] Gelenbe, E. (1991). Product-form queueing networks with negative and positive customers. *J. Appl. Prob.* **28,** 656–663.
[23] Gelenbe, E. (1993). G-networks with triggered customer movement. *J. Appl. Prob.* **30,** 742–748.
[24] Gelenbe, E. (1994). G-networks: a unifying model for neural and queueing networks. *Ann. Operat. Res.* **48,** 433–461.
[25] Gelenbe, E. (2007). Steady-state solution of probabilistic gene regulatory networks. *Phys. Rev. E* **76,** 031903.
[26] Gelenbe, E. and Fourneau, J. (2002). G-networks with resets. *Performance Evaluation* **49,** 179–191.

[27] Gelenbe, E. and Hussain, K. F. (2002). Learning in the multiple class random neural network. *IEEE Trans. Neural Networks* **13,** 1257–1267.
[28] Gelenbe, E. and Pujolle, G. (1976). The behaviour of a single queue in a general queueing network. *Acta Informatica* **7,** 123–136.
[29] Gelenbe, E. and Schassberger, R. (1992). Stability of product form G-networks. *Prob. Eng. Inf. Sci.* **6,** 271–276.
[30] Gelenbe, E. and Stafylopatis, A. (1991). Global behavior of homogeneous random neural systems. *Appl. Math. Modelling* **15,** 534–541.
[31] Gelenbe, E., Mang, X. and Onvural, R. (1996). Diffusion based statistical call admission control in atm. *Performance Evaluation* **27,** 411–436.
[32] Gelenbe, E., Mang, X. and Onvural, R. (1997). Bandwidth allocation and call admission control in high-speed networks. *IEEE Commun. Mag.* **35,** 122–129.
[33] Gòmez-Corral, A. (2002). On a tandem G-network with blocking. *Adv. Appl. Prob.* **34,** 626–661.
[34] Gòmez-Corral, A. and Martos, M. E. (2006). Performance of two-stage tandem queues with blocking: the impact of several flows of signals. *Performance Evaluation* **63,** 910–938.
[35] Guffens, V., Gelenbe, E. and Bastin, G. (2006). Qualitative dynamical analysis of queueing networks with inhibition. In *INTERPERF '06: Proceedings from the 2006 Workshop on Interdisciplinary Systems Approach in Performance Evaluation and Design of Computer & Communications Sytems*, ACM, New York.
[36] Harrison, P. G. (2004). Compositional reversed Markov processes, with applications to G-networks. *Performance Evaluation* **57,** 379–408.
[37] Harrison, P. G. and Pitel, E. (1996). The M/G/1 queue with negative customers. *Adv. Appl. Prob.* **28,** 540–566.
[38] Henderson, W., Nothcote, B. S. and Taylor, P. G. (1994). State dependent signalling in queueing networks. *Adv. Appl. Prob.* **26,** 436–455.
[39] Iglehart, D. L. and Whitt, W. (1970). Multiple channel queues in heavy traffic. I. *Adv. Appl. Prob.* **2,** 150–177.
[40] Jarvis, D. and Kushner, H. J. (1996). Codes for optimal stochastic control: documentation and users guide. Tech. Rep. 96-3, Brown University.
[41] Kingman, J. F. C. (1961). The single server queue in heavy traffic. *Proc. Camb. Philos. Soc.* **57,** 902–904.
[42] Kloeden, P. E., Platen, E. and Schurz, H. (1994). *Numerical Solution of SDE Through Computer Experiments*. Springer, New York.
[43] Kushner, H. J. (2001). *Heavy Traffic Analysis of Controlled Queueing and Communication Networks*. Springer, New York.
[44] Kushner, H. J. and Dupuis, P. G. (1992). *Numerical Methods for Stochastic Control Problems in Continuous Time*. Springer, New York.
[45] Kushner, H. J. and Martins, L. F. (1993). Heavy traffic analysis of a data transmission system with many independent sources. *SIAM J. Appl. Math.* **53,** 1095–1122.
[46] Kushner, H. J., Yang, J. and Jarvis, D. (1995). Controlled and optimally controlled multiplexing systems: a numerical exploration. *Queueing Systems* **20,** 255–291.
[47] Leite, S. C. and Fragoso, M. D. (2007). On the analysis of G-queues under heavy traffic. Submitted.
[48] Li, Q. and Zhao, Y. Q. (2004). A MAP/G/1 queue with negative customers. *Queueing Systems* **47,** 5–43.
[49] Mandelbaum, A. and Gennady, P. (1998). State-dependent stochastic networks. Part I. Approximations and applications with continuous diffusion limits. *Ann. Appl. Prob.* **8,** 569–646.
[50] Prohorov, Yu. (1963). Transient phenomena in process of mass service. *Litovsk. Mat. Sb.* **3,** 199–205 (in Russian).
[51] Reiman, M. I. (1984). Open queueing networks in heavy traffic. *Math. Operat. Res.* **9,** 441–458.
[52] Shin, Y. W. (2007). Multi-server retrial queue with negative customers and disasters. *Queueing Systems* **55,** 223–237.
[53] Whitt, W. (1972). Complements to heavy traffic limit theorems for the $GI/G/1$ queue. *J. Appl. Prob.* **9,** 185–191.
[54] Whitt, W. (1974). Heavy traffic limit theorems for queues: a survey. In *Mathematical Methods in Queueing Theory* (Lecture Notes in Econom. Math. Systems **98**), eds M. Beckmann and H. P. Kunzi, Springer, New York, pp. 307–350.